An Improved automated content-based image retrieval based on SVD for screening digital mammography

N.Laxmi[1], Dr. D. Satyanarayana[2]

[1]*Associate Professor of ECE, Guru Nanak Institute of Engg.& Tech., Ibrahimpatnam, Hyderabad, Telegana State*

[2]*Professor & HOD of ECE, RGM College of Engineering & Technology, Nandyal, A.P. State*

Abstract - Many diseases are curable, when they are diagnosed at the earlier stage. Diagnosis of disease depends on how efficiently the query image abnormality is accurately detected. The computer aided automated system such as content based medical image retrieval technique is used to retrieve query based images in the large databases. Breast cancer is the fifth most common disease leading to the cancer deaths around the world and the second common type of cancer leading to the death among the women. Nowadays, the technique called as mammography is one of the leading technique used for the breast cancer detection among the women. In this paper, automated content-based image retrieval (CBIR) system is presented based on SVD that supports the classification of breast tissues for lesion segmentation and classification. Singular value decomposition (SVD) is used for the breast density characterization by the selection of the first singular values, in order to represent texture along with the dimensionality reduction. The comparative experiments conducted with reference to benchmark mammographic images analysis society (MIAS) database confirmed the effectiveness of the proposed work concerning average precision of 74% and 72.30% for normal & abnormal classes of mammograms, respectively.

Keywords:- Breast cancer, mammography, CBIR, SVD, classification, features etc.

I. INTRODUCTION

Many diseases are curable, when they are diagnosed at the earlier stage. Diagnosis of disease depends on how efficiently the query image abnormality is accurately detected. Medical images are important for diagnosis purposes as they are related to patient’s medical historic and pathology. The computer aided automated diagnostic systems (CAADS) are mostly used by the medical experts as such nowadays huge number of mammograms has been generated in hospitals for the diagnosis of breast cancer. Content-based image retrieval (CBIR) can contribute more reliable diagnosis by classifying the query mammograms and retrieving similar mammograms already annotated by diagnostic descriptions and treatment results. These content based medical image retrieval methods are used to retrieve query based images from the voluminous databases by extracting the significant features and applying the similarity matching methods. Mammography uses low x-ray doses to produce images of breasts and it is an efficient and largely used method to the early detection of breast cancer. Breast cancer represents one of the main causes of death among women in occidental countries (Brazilian National Cancer Institute, Ref:- http://www.inca.gov.br).

Breast density has been shown to be related with the risk of the development of breast cancer [25] since women with a dense breast density can hide lesions and so cancer is detected at later stages. A density scale named Breast Imaging Reporting Data System (BIRADS) developed by the American College of Radiology (http://www.acr.org) informs radiologists about the decline in sensitivity of mammography mammography with increasing breast density. In this fatty densities are defined as 1.almost entirely fatty 2. as scattered fibro-glandular tissue 3. as heterogeneously dense tissue and density 4 as extremely dense tissue.

Radiologists evaluate and report breast density on the basis of the visual analysis of mammographies. Computer aided diagnosis (CAD) and content-based retrieval (CBIR) systems appear as a real possibility to help radiologists in reducing the variability of their analysis. CBIR systems use visual information extracted from images to retrieve similar images to one query image and this system does not need to provide diagnosis information of the retrieved images but just present similar images according to a certain pattern.

Considering a CBIR system based on the breast density, from a clinical point of view, such a system can guide the radiologist for the detection of a lesion and its classification.

Moreover, from a technical point of view, this system is the first step, and a very important one, for the development of a CAD system. With this intent of developing reliable systems for accurate diagnosis a set of image which includes shape; histogram based statistical, Singular Features (SVD based method) Gabor, wavelet, and Gray Level Co-occurrence Matrix (GLCM) features, was computed from the segmented region. In order to select the optimal features, a minimum redundancy maximum relevance (mRMR) feature selection method was then applied. Finally, similar images were retrieved using Euclidean distance similarity measure.

In this work, we propose, implement, and evaluate a CBIR system based on SVD of mammographic images of the breast. The breast density is characterized through singular value decomposition (SVD) [11] and the support vector machine (SVM) [22] is used as a classifier for the retrieving task.In the context of mammography and breast density, some works explored the use of CBIR and CAD systems. Kinoshita et al [14] used breast density as a pattern to retrieve 1,080 mammographies from the Clinical Hospital from the University of S˜ao Paulo, Ribeir˜ao Preto, Brazil.

Shape descriptors, texture features and histograms were used to characterize the breast density, and the Kohonen & Gross Berg self-organizing map (SOM) neural network was used for the retrieval task. Precision rates between 83% and 79% were obtained for 50% and 25% of recall. Despite the fact that these results indicate that through certain types of features, such as histograms and shape, retrieval concerning the breast density can be effective, and additional studies are needed to improve all the process of retrieval.

Regarding only the breast density classification, mammographies were automatically divided between fatty tissue and dense tissue [19] using the Fuzzy-C means algorithm. From all images, texture and morphological features were extracted and then mammographies were classified using decision trees and k-Nearest Neighbor algorithm. Comparing with radiologists' classification,the results are of 86% of correct classification for MIAS (The Mammographic Image Analysis Society Digital Mammogram Database) database [21] and 73% for DDSM (The Digital Database for Screening Mammography) database [13].

The remainder of this paper is broken into six sections. Section 2 gives idea about some of the related work of CBIR in medical domain. Section 3 the introduces the texture characterization of the breasts through SVD. In Section 4, we expose the basic principles of the SVM classifier used for the retrieval task. Section 5 presents the experiments. In Section 6, we present and discuss the results, and in Section 7, we state the conclusion of the work.

CBIR methods are proposed for the retrieval of the images from the mammogram databases. The various types of technique are required to fetch the images from the database depending upon the requirements. The ultimate goal of the retrieval system is to show the results to the radiologists in the form of display.

II. RELATED WORK

V.Nath et.al.(2019) suggested a technique of resolution enhancement is based on wavelet enhanced the resolution of an image as well as do not loss the edge information but it flops to produce the better contrast image. The SVD can also use for enhancement of image

contrast. The leading cause of using SVD, it holds information of illumination.

J´ulia E. E. de Oliveira et al. (2009) presented a content-based image retrieval (CBIR) system called MammoSVD. This CBIR system is developed based on breast density – fatty or dense, and the database used, from the IRMA project, provides images with the ground truth already set. Singular value decomposition (SVD) is proposed for the breast density characterization

by the selection of the first singular values, in order to represent texture along with the dimensionality reduction. Support-vector machine (SVM) is used to perform the retrieval operation.

Júlia E.E. de Oliveira et al. (2010) presented a content-based image retrieval system designed to retrieve mammographies from large medical image database. The system is developed based on breast density, according to the four categories defined by the American College of Radiology, and is integrated to the database of the Image Retrieval in Medical Applications (IRMA) project that provides images with classification ground truth. 2DPCA is used in breast density texture characterization, in order to effectively represent texture and allow for dimensionality reduction. A support vector machine is used to perform the retrieval process. Average precision rates are in the range from 83% to 97% considering a data set of 5024 images. The results indicate the potential of the system as the first stage of a computer-aided diagnosis framework.

Thomas M. Deserno (2018) in their study suggested , a CBIR system was presented that uses breast density together with the existence of a breast lesion as pattern for image retrieval. They have continued comprehensive system evaluation based on a significantly enlarged database of, so far, 3,375 images of 12 classes. This work found to be contributed to CBIR-CAD of mammographies, providing a system able to aid radiologists in their diagnosis or a system that is useful as preprocessing stage for computer-aided systems for breast lesions classification.

Issam El-Naqa et al.(2002) explored the use of a learning-based kamework for retrieval of relevant mammogram images kom a database, for purposes of aiding diagnoses. A fundamental issue is how to characterize the notion of similarity between images for use in assessing relevance of images in the database. With this intent they have investigated the use of several learning algorithms. namely. neural networks and support vector machines, in a two-stage hierarchical learning network for predicting the perceptual similarity from similarity scores collected in human-observer studies. The proposed approach is demonstrated using micmcalcilication clusters extracted from a database consisting of 76 mammograms. Initial results demonstrate that the proposed two-stage hierarchical learning network outperforms a single-stage learning network.

Vibhav Prakash Singh et al. (2018) developed an efficient and automated CBIR system of mammograms. In this first, they have done the pre-processing steps including automatic labelling-artifact suppression, automatic pectoral muscle removal, and image enhancement using the adaptive median filter were applied. Next, pre-processed images were segmented using the co-occurrence thresholds based seeded region growing algorithm.Furthermore, a set of image features including shape, histogram based statistical, Gabor, wavelet, and Gray Level Cooccurrence Matrix (GLCM) features, was computed from the segmented region. In order to select the optimal features, a minimum redundancy maximum relevance (mRMR) feature selection method was then applied. Finally, similar images were retrieved using Euclidean distance similarity measure. The comparative experiments conducted with reference to benchmark mammographic images analysis society (MIAS) database confirmed the effectiveness of the proposed work concerning average precision of 72% and 61.30% for normal & abnormal classes of mammograms, respectively.

Demirel et al.(2010) presented a technique in which input image having low contrast decomposes into the frequency sub-bands by using DWT and after the decomposition, in next step, the matrix of SVD can be calculated for the low-low (LL) sub-band image. This method is called (DWT–SVD) reforms the enhanced output image by performing the IDWT.

Demirel et al. (2011) proposed a new strategy for improving the resolution of satellite images based on DWT and through interpolated output of given image and sub-bands with high frequency. In this method, given image decomposes into sub-bands of different frequencies after that sub-bands having high frequency and given image having low-resolution interpolated and then perform IDWT for generating image of high resolution.

Srinivas et al.(2013) performed a study of SWT, DWT, DWT and SWT, and DTCWT-based resolution enhancement of satellite image and got better enhanced and sharper image using these wavelet-based enhancement techniques. As a result, the enhancement by DWT-SWT has high resolution than enhancement using DWT.

Sharma et al. (2014) proposed a modified algorithm proposed in [28]. This technique based on gamma correction for enhancement in contrast of satellite images using SVD and DWT. In this method, intensity transformation done by gamma correction improves illumination by using SVD. This presented technique confirmed the effectiveness of its method by comparing with Demirel's [28] by calculating entropy, PSNR and EME (measure of enhancement).

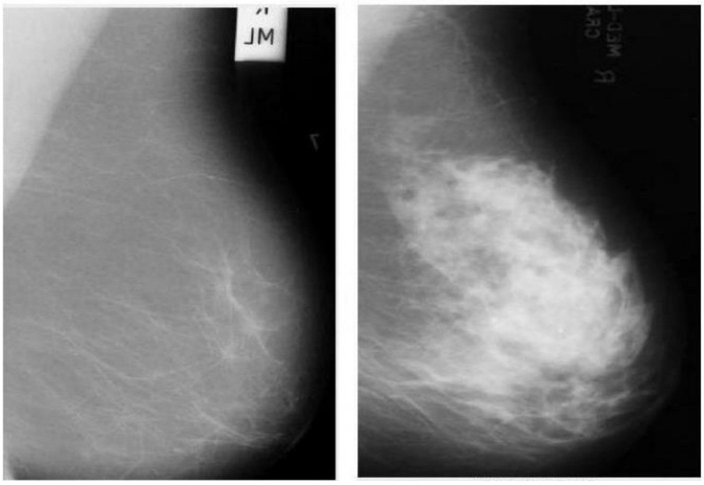
In analysis of the literature a main challenge for the development of CBIR systems is the appropriate characterization of images and the storage and management of the big amount of images produced by hospitals and medical centers. The IRMA (Image Retrieval in Medical Applications) project deals with this kind of problems, as it aims at developing and implementing high-level methods for CBIR systems with prototypal application to medico-diagnostic tasks on radiological image archive [23]. There are currently more than 30,000 diagnostic images with available ground truth information in the IRMA database. They are used for image retrieval and computer-aided diagnosis [15,21]. Regarding mammography, there are more than 10,000 images in the database [14], all of them also with available ground truth information. This database offers invaluable support to the validation of the method proposed in this work.

III. BREAST DENSITY CHARACTERIZATION

In CBIR systems, the access to information is performed by the visual attributes extracted from images. The definition of a set of features, capable to describe effectively each region contained in an image, is one of the most complex tasks in the analysis of images. In addition, the process of characterization affects all the subsequent process of a CBIR system [7].

An image can be numerically represented by a feature vector, which should reduce the dimensionality of the image and emphasize aspects of this image [11]. Visually, breasts of fatty and dense densities differ through gray level intensity in mammographies, as can be seen in Figure 1. Since texture contains information about the spatial distribution of gray levels and variations in brightness, its use for the representation of breast density becomes appropriate [13].

The high dimensionality of the feature vector is one of the difficulties in the use of the texture attribute, so it is desirable to choose a technique that combines the representation of this texture with the reduction of dimensionality, in a way to turn the retrieval algorithm more effective and computationally treatable.



(a) Fatty density **(b) Dense Density**

Fig. 1. Mammograhy density

The method of SVD consists in decomposing a matrix, whose elements can be composed of the intensity of the pixels belonging to a certain texture, in a matrix multiplication operation [15,23]. The singular values obtained as results of this decomposition provide useful information of the texture, and for purposes of reduction of dimensionality only the first k singular values are kept. The goal is to find the best rank k that would improve the image characterization [15], and this rank k must be no more than the minimum value between the sizes of the matrix.

IV. CBIR USING SVM

Image retrieval has the purpose to retrieve, from a database, images that are relevant for one query. The query image goes through the process of extraction of attributes and the obtained feature vector is submitted to a search for similarity together with the structure containing the feature vector of all images stored in the database. The identities of the resulting images from the search are used to retrieve these images from the database, thus they can be presented to the radiologist.

SVD is applied on mammographs to deals with a binary classification of data: fatty breast density or dense breast density. The support vector machine (SVM) method is considered a good classifier since it is able to predict correctly the class of the new data from the same domain where the learning occurred [20,24].

SVM can be described for a binary classification as follows: given two classes and a set of points that belong to these classes, the SVM classifier determines the hyperplane in the feature space that separates the points in order to place the highest number of points of the same class on the same side, while maximizing the distance of each class to that hyperplane. The hyperplane generated is determined by a subset of items from the two classes, called support vectors.

When the sets of data are linearly separable by a straight line, it is called a linear case of separation. But in most of the cases, this linear case is a restrictive hypothesis to be used in practice. So, instead of a straight line, it is used one function called kernel. The most commonly used kernels are the polynomial and Gaussian ones [22].

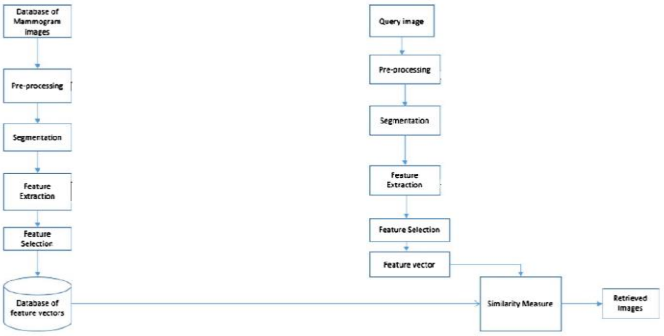


Fig-2. Block Diagram for CBIR in Mammograms

V. EXPERIMENTS

The SVD based system for Mammography images is used from the database of radiological images from the IRMA project [8]. In the project, all images are coded according to a mono-hierarchical, multi-axial coding scheme [18], and this codification provides the ground truth of all mammographies, as all the images were previously verified by an experienced radiologist.

The images, which have approximately 1024x500 pixels of size and are from both medio-lateral and craniocaudal projections, were grouped in mammographies of fatty density – 200 mammographies from BIRADS 1 and 200 mammographies from ACR BI-RADS 2 – and mammographies of dense density – 200 mammographies from ACR BI-RADS 3 and 200 mammographies from ACR BI-RADS 4.

For all the images, in a way to remove noises such as black areas and exams labels, it was performed a segmentation of the breast region. After this segmentation, the steps followed for the development of the CBIR system were:

Step1 ! Extraction of singular values: the following first k singular values were kept for the composition of the feature vector: 25, 50, 75, 100, 150, 200 and 250. These values were chosen empirically according to [15].

Step2 ! Measurement of similarity between images: SVM computes the similarity between images through the indication of relevance of the image to a certain query. The set of 800 feature vectors was used in the following manner:

- Training: 240 feature vectors of fatty breast density and 240 feature vectors of dense breast density.
- Test: 160 feature vectors of fatty breast density and 160 feature vectors of dense breast density.

The selection of the feature vectors used for training and the ones used for test was done randomly, and the two sets are disjointed. Moreover, tests were done using the linear case and the polynomial and Gaussian kernels. Of the three cases, the polynomial kernel was the one capable of separating more efficiently the two classes.

Step3 ! Evaluation of the CBIR system: measures of precision and recall were obtained and all the 320 mammographies that were not used for the training of the SVM classifier were used as query. We considered values of precision for 10% of recall since radiologists pay more attention to the top returned images.

VI. RESULTS AND DISCUSSION

Figure 3 presents the precision and recall curve for the best result obtained, the one using the first 200 singular values for breast density characterization and SVM for image retrieval. This selection allows representing the texture of the mammographies together with the reduction of dimensionality, as storing only few values results in significant computational savings over storing the whole vector.

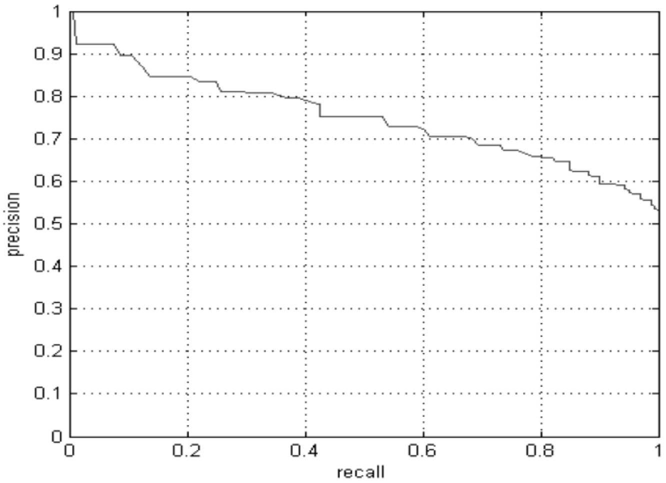


Figure 3. Precision and recall curve for the first 200 singular values for breast density characterization.

A value of 90% of precision for 10% of recall means that from 32 mammographies returned by the SVD system, 29 images are relevant for the query user. The SVD method was able to capture the difference between the gray level intensities of the breast densities and characterize them. Also, because of the high generalization ability of the SVM classifier, the training with the polynomial kernel was able to separate the two classes and indicate the most relevant images to the query one.

Although radiologists look for breast lesions like masses and calcifications in mammographies, a CBIR system for mammography should include all possibilities. SVD is the first stage of a CBIR system, as the breast density plays an important role in the diagnostic process.

An important characteristic of the proposed CBIR system is the use of a priori breast density classification, as all the images contained in the IRMA database have their ground truth already set by an experienced radiologist.

Future works will focus on the retrieval of four classes of breast density according to ACR BI-RADS scale and on the visual presentation of the retrieved images.

VII. CONCLUSION

The research on CBIR still has some challenges. One is which approach to use, more efficiently, to characterize images through a small sequence of numerical values, thus reducing the dimensionality of the original images and how to represent these features properly. In this paper we presented a CBIR system, called MammoSVD, which uses the breast density as standard for image retrieval as this can hide lesions indicative of breast cancer. The mammographies used, belonging to IRMA database, are already classified, setting the ground truth, in order to provide for the retrieval process, beyond similar images, also diagnosis information of the mammographies.

In this system, the segmented areas of the breast are characterized through SVD and only the first singular values are kept, in a way to represent texture together with the reduction of dimensionality of the feature vector. This characterization allied with SVM classification for image retrieval enables the development of a CBIR system of mammographies that can really aid radiologists in their diagnosis.

VIII. REFERENCES

[1]. V. Nath and J. K. Mandal (eds.), Proceeding of the Second International Conference on Microelectronics, Computing & Communication Systems (MCCS 2017), Lecture Notes in Electrical Engineering 476, © Springer Nature Singapore Pte Ltd. 2019 https://doi.org/10.1007/978-981-10-8234-4_30

[2] J´ulia E. E. de Oliveira, Ana Paula B. Lopes, Guillermo C´amara-Chavez, Arnaldo de A. Ara´ujo MammoSVD: a Content-Based Image Retrieval System Using a Reference Database of Mammographies Federal University of Minas Gerais - Department of Computer Science Horizonte, MG, Brazil

[3] Júlia E.E. de Oliveiraa,∗, Alexei M.C. Machadob, Guillermo C. Chaveza,Ana Paula B. Lopesa, Thomas M. Desernoc, Arnaldo de A. Araújoa " MammoSys: A content-based image retrieval system using breast density patterns computer methods and programs in biomedicine" (2 0 1 0) 289–297 journal homepage: www.intl.elsevierhealth.com/journals/cmpb

[4] Thomas M. Deserno, Michael Soiron J´ulia E.E. de Oliveira, Arnaldo de A. Ara´ujo " Towards computer-aided diagnostics of screening mammography using content-based image retrieval" RWTH Aachen University Aachen, Germany

[5] Issam El-Naqa. Yongyi Yang, Nikolas P. Galatsanos, and Miles N. Wemick "content-based image retrieval for digital mammography Dept. of Electrical and Computer Engineering, Illinois Institute of Technology 3301 S. Dearborn Street. Chicago, IL 60616 0-7803-7622-6/02/$17.00 02002 IEEE IEEE ICIP 2002

[6] Vibhav Prakash Singh " Automated and effective content-based image retrieval for digital mammography Journal of X-Ray Science and Technology 26 (2018) 29–49 DOI 10.3233/XST-17306 IOS Press

[7] [1] R. Baeza-Yates and B. R. Neto. Modern Information Retrieval. Addison-Wesley Professinal, 1999.

[8] [2] J. E. E. de Oliveira, M. G¨uld, A. de Albuquerque Ara´ujo, B. Ott, and T. Deserno. Towards a standard reference database for computer-aided mammography. In Proceedings of SPIE Medical Imaging, volume 6915, page 69151Y,USA, 2008.

[9] [3] T. Deselaers, H.M¨uller, P. Clough, H. Ney, and T. Lehmann The CLEF 2005 automatic medical annotation task. International Journal of Computer Vision, 74(1):51–58, 2007.

[10] [4] R. O. Dudda, P. E. Hart, and D. G. Stork. Pattern Classification.John Wiley Sons, 2001.

[11] [5] G. H. Golub. Matrix computations. Johns Hopkins series in the matematicals sciences, 1983.

[12] [6] R. C. Gonzalez, R. E. Woods, and S. L. Eddins. Digital Image Processing using Matlab. Prentice-Hall, 2003.

[13] [7] M. Heath, K. Bowyer, and D. K. et al. Current status of the digital database for screening mammography. In: Digital Mammography, Kluwer Academic Publishers, pages 457–460, 1998.

[14] [8] S. K. Kinoshita, P. M. de Azevedo Marques, R. R. P. Jr, J. A. H. Rodrigues, and R. M. Rangayyan. Contentbased retrieval of mammograms using visual features related to breast density patterns. Journal of Digital Imaging,20(2):172–190, 2007.

[15] [9] L. ´ Elden. Numerical linear algebra in data mining. Acta Numerica, pages 327–384, 2006.

[16] [10] T. M. Lehmann, M. O. G¨uld, T. Deselaers, D. Keysers,H. Schubert, K. Spitzer, H. Ney, and B. Wein. Automatic categorization of medical images for content-based image retrieval and data mining. Computerized Medical Imaging and Graphics, 29(2):143–155, 2005

[17] [11] T. M. Lehmann, M. O. G¨uld, C. Thies, B. Fischer,K. Spitzer, D. Keysers, H. Ney, M. Kohnen, H. Schubert, and B. Wein. Content-based image retrieval in medical applications.

Methods of Information in Medicine, 43(4):354–361,2004.

[18] [12] T. M. Lehmann, H. Schubert, D. Keysers, M. Kohnen, and B. Wein. The IRMA code for unique classification of medical images. In Proceedings of SPIE, volume 5033, pages 440–451, 2003.

[19] [13] A. Oliver, J. Freixenet, R. Mart´ı, J. Pont, E. P´erez, E. R.Denton, and R. Zwiggelaar. A novel breast tissue density classification methodology. IEEE Transactions on Information Technology in Biomedicine, 12(1):55–65, 2008.

[20] [14] M. Rahman, B. C. Desai, and P. Bhattacharya. Supervised machine learning based medical image annotation and retrieval.In Image CLEFmed, pages 692–701, 2005.

[21] [15] J. Suckling. The mammographic image analysis society digital datagram database. Exerpta Medica International Congress Series, 1069:375–378, 1994.

[22] [16] V. N. Vapnik. The nature of statistical learning theory. Springer-Verlag, New York, 1995.

[23] [17] D. S. Watkins. Fundamentals of matrix computations. John Wileys Sons, 1991.

[24] [18] L. Wei, Y. Yang, R. M. Nishikawa, and M. N. Wernick. Mammogram retrieval by similarity learning from experts.In IEEE International Conference on Image Processing,pages 2517–2520. IEEE, October 2006.

[25] [19] J. N. Wolfe. Breast patterns as an index of risk for developing breast cancer. American Journal of Roentgenology,126:1130–1139, 1976.

[26] [20] 15. H. Demirel, C. Ozcinar, G. Anbarjafari, "Satellite image contrast enhancement using discrete wavelet transform and singular value decomposition", IEEE Geosci. Remote Sens. Lett. 7 (2010) 333–337.

[27] [21] 16. H. Demirel and G. Anbarjafari, "Discrete wavelet transform-based satellite image resolution enhancement," IEEE Trans. Geoscience and Remote Sensing, vol. 49, no. 6, (2011) pp. 1997–2004.

[28] [22] 12. P. Bala srinivas B. Venkatesh, "Comparative Analysis of DWT SWT, DWT & SWT and DTCWT Based Satellite Image Resolution Enhancement", IJECT Vol. 5, Issue 4, (2014).