https://dx.doi.org/10.22161/ijaers.4.3.41 *ISSN: 2349-6495(P) | 2456-1908(O)*

Breast Cancer Diagnostic System Based on MR images Using KPCA-Wavelet Transform and Support Vector Machine

Mustafa Zuhaer AL-Dabagh, Dr. Firas H. AL-Mukhtar

Department of Computer Science, Knowledge University, Erbil, Iraq.

Abstract*—Automated detection and accurate classification of breast tumors using magnetic resonance image (MRI) are very important for medical analysis and diagnostic fields. Over the last ten years, numbers of methods have been proposed, but only few methods succeed in this field. This paper presents the design and the implementation of CAD system that has the ability to detect and classify the tumor of the breast in the MR images. To achieve this, k-mean clustering methods and morphological operators are applied to segment the tumor. The gray scale, Texture and symmetrical features as well as discrete wavelet transform (DWT) are used in feature extracted stage to obtain the features from MR images. Kernel principle components analysis (K-PCA) are also applied as a feature reduction technique and support vectors machine (SVM) are used as a classifier. Finally, the experiments results have confirmed the robustness and accuracy of proposed system*

Keywords— k-mean clustering, Morphological operators, gray scale, Texture, symmetrical, kernel principle components analysis and Support vectors machine.

I. INTRODUCTION

After the evaluation of technology, the detection and classification of the tumor in magnetic resonance images (MRI) became very significant because it gives important information for each of medical diagnosis and surgical [1]. Nevertheless, detection and classification of tumor is a very hard task because of the size, the shape and the location of tumors is very different and any wrong diagnostic can lead to serious collateral damage [2]. MRI is considered as the most accurate technique used to study the tumors in soft tissues for two reasons. The first one, the MRI images gives many details about the tumor and, the second one is, there is no known side effects related to radiation exposure [3].

In recent years and after the fast evolution of computer technology, computers are used to support the medical decision systems and spread in various medical fields such as breast cancer, gastroenterology, Breast tumors etc. There are many image processing techniques that have been proposed to detect and classify MRI Breast tumor. The complexity of the pathology of the human Breast and the high quality of the images are required in the diagnosis are remain an obstacle in front of any new method [4].

Different types of techniques have been modified for extracting the features from MR images. Some of these techniques are simple such as mean and variance of the gray level of the image hut it is not enough to get a good recognition rate. Therefore, other gray-scale statistics techniques such as the gray-level co-occurrence matrices (GLCM) are applied [5] to solve the problem. GLCM is considered as one of the popular techniques used in this field and very popular in various other applications of texture analysis. GLCM depends on statistical methods for extracting number of textural features from the images [6]. Co-occurrence matrices can provide significant information about patterning of the texture, and it is applied to define the textural features from them [7].

The main problem of most diagnostic systems is how the systems can enhance the performance automatically with the increase of experience. For this reason, different types of machine learning methods were developed to solve these problems like random forests, artificial neural networks, decision tree and support vector machine (SVM). SVM is one of a popular classifier that depends on small sample learning and gives a good generalization capability. It is used in many of the real-time applications such as text mining, face recognition, and image processing, and these make SVMs considered as one of the most developed tools in the field of machine learning and data mining [8].

Various methods have been used for feature extraction and to reduce the data dimensionality such as Principal Components Analysis (PCA), Linear Discriminate Analysis (LDA), Independent Components Analysis (ICA) and Kernel Principal Components Analysis (k-PCA). Kernel-Principal Components Analysis (k-PCA) is considered as a dimensions reduction and feature extraction method which has been widely and effectively used in different types of applications [9-10].

https://dx.doi.org/10.22161/ijaers.4.3.41 *ISSN: 2349-6495(P) | 2456-1908(O)*

This paper is organized as follows: Section II introduces the proposed system architecture. In Section III, experiments and results are discussed. Finally, conclusions are documented in Section IV.

II. PROPOSED SYSTEM ARCHITECTURE

In this proposed system, six stages will be used to detect and classify the tumor in MR image. In the first stage, preprocessing techniques (histogram equalization and median filter) will be applied to improve the images and to eliminate the noise. In the second stage, K-mean clustering and morphological operators will be used to segment the tumor. Features such as gray scale, texture, symmetrical and DWT features will be extracted from the images to use it in the training and the classification stages. K-PCA is applied in the reduction stage to reduce the size of data. Finally, SVM has been applied to determine the types of tumor (Malignant or Benign).

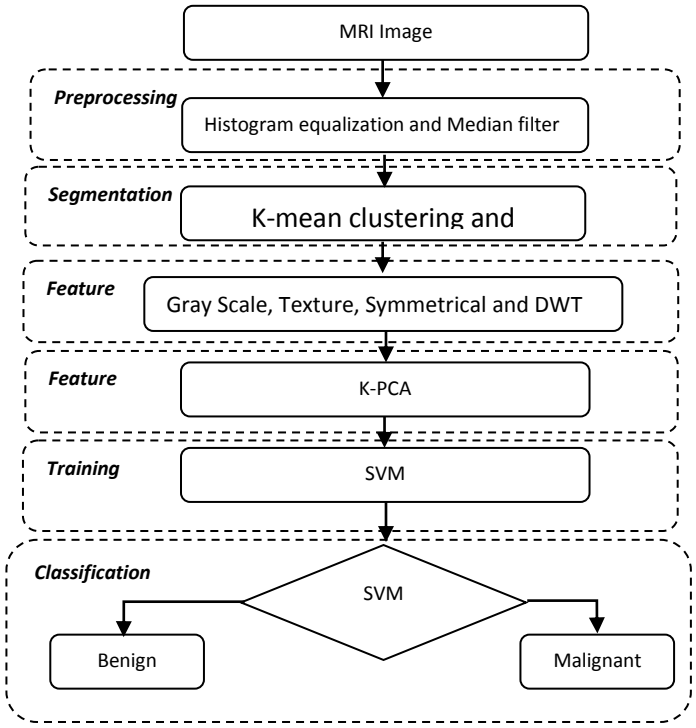

Fig.1: Six Stage of Proposed System

2.1 PRE-PROCESSING

In this stage, the input image is passed into two processing steps (Median filtering and histogram) before it enters to the next processing stage to enhance the quality of the input image as well as remove noise in the image. Median filtering and histogram equalization are used to make the image in the best quality and minimum noise.

2.1.1 Median filter

It is a nonlinear filter and it is applied to remove the noise of an image. It is work depends on window slide idea. The windows slide scans the whole image depending on its central pixel through scanning each part of the image. Through this scan, the neighboring pixels are ordered depending on the intensity value of the pixel and then the median value is replaced with the central pixel value. The benefit of using a median filter is to remove certain kind of noise especially when some pixels of the image contain extreme values. In this way, new pixels value is not generated because the output pixel value is one of the neighboring pixels values. In the median filter, the edges are not affected during removing the noise and it is possible to use it more than one time [11].

2.1.2 Histogram Equalization

It is one of a popular image pre-processing techniques for adjustment the contrast of the image. During acquisition stage, the distribution of intensity in the image may be affected and this leads to a bad contrast and low quality of the image, therefore histogram equalization is employed to enhance the appearance of an image. The main idea of histogram equalization is based on creating a new gray scale from the old gray scale of the input image [12]. The equation used to calculate the histogram equalization is shown below:

$$k_0 = round\left(\frac{c_i(2^k-1)}{w.h}\right) \quad (6)$$

Where K_0 represents a gray level value of histogram equalization, c_i refers to the cumulative distribution of i^{th} gray scale from original image; round represents rounding to the nearest value, while w and h refer to the width and height of the image [13].

2.2 SEGMENTATION

This method is considered as the one of most important stage in the CAD system. It aims to partition the image into a number of segments to make the analysis of the image. In this stage, combinations of K-Means clustering and morphological operators along with basic image processing techniques are applied to implement this function. K-Means clustering aims to detect tumor and shows some abnormality while morphological operators with another basic image processing techniques aims to rectify this detection through separating the tumor cells from the normal cells. The steps of segmentation are as the follows:

2.2.1 Threshold Operation

It is one of widely known techniques. It is used to create a binary format from gray-scale images depending on a particular intensity level. The main idea of thresholding operation is to turning all the pixels that have value larger than intensity level to 1 and all pixels that have value less than the intensity level to 0. The operation of thresholding is described in the following equation:

$$g(x,y) = \begin{cases} 1 & if\ f(X,Y) > T \\ 0 & if\ f(X,Y) < T \end{cases} \quad (1)$$

Where $f(x,y)$ represents the image of x rows and y columns and T refers to the intensity level [14].

https://dx.doi.org/10.22161/ijaers.4.3.41 *ISSN: 2349-6495(P) | 2456-1908(O)*

2.2.2 Watershed Transformation

It is one of the widely used techniques to collect pixels of an image that have similar intensity levels. It is considered as a region-based segmentation technique. The essential idea of this technique based on segregates the image to various intensity portions and then it starts to fill the deeper gradient and then fills the lighter gradient. In a gray scale image the tumor cells have high-intensity values comparing with other parts, therefore watershed segmentation is considered as one of the best tools to classify tumors and high-intensity tissues in MR image [15].

2.2.3 K-means Clustering

It is a high computational efficiency unsupervised method and it is a widely used in many applications. The letter "k" refers to the number of clusters used. Clustering process is grouping the data points that have similar features into the same cluster and the data points of dissimilar features into different clusters. To understand the K-means clustering method, let us consider that, X= {x_1, x_2,……,x_N} is a set of data points and we want to separate into a number of clusters C= {c_1, c_2,……,c_k}. K-means method is aimed to compute the center of data, and then assign the data of the same features to a cluster center. K-means iterates this method until clustering all the data. The center of all clusters is calculated as displayed in the equation:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} ||X_N - C_K||^2 \qquad (2)$$

Where $||\boldsymbol{X_N} - \boldsymbol{C_K}||^2$ refers to the distance between a data point $\boldsymbol{X_N}$ and the center of cluster $\boldsymbol{C_K}$, while $\boldsymbol{J}$ represents the distance of $\boldsymbol{n}$ points from their respective cluster centers [16].

2.2.4 Morphological Operators

Morphological operations are used to reconstructing the structure or shape of an object. They can be employed for both pre-processing and post-processing operations, like obtaining a representation or description of the shape of objects or can be used for filtering, thinning or pruning. The main morphological operations are erosion, dilation, opening and closing. Opening and closing operations are applied in this stage. At first, Opening is applied by implementing erosion to remove unwanted pixels from the image and after that dilation is implemented to focus on the area of interest. Secondly, closing is applied by implementing dilation and then erosion to fill holes. The mathematical expressions of closing and opening are explained in equations 3 and 4:

$$A \circ B = (A \ominus B) \oplus B$$
$$A \bullet B = (A \oplus B) \ominus B \qquad (3)$$

After applying these techniques, the pixels which have high intensity value are grouped together into one cluster while other pixels grouped into other cluster. Figure 2. shows the output image from this stage [17].

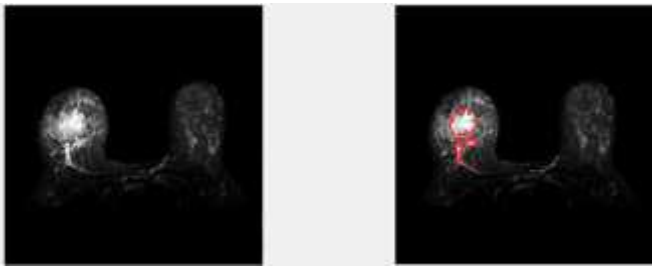

Fig.2: The output image of segmentation stage

2.3 Feature Extraction

It is considered as one of the main parts of the diagnostic system. It is applied to extract unique features from the input image to use it in the classification stage. It aims to reduce the amount of data that can be used to represent a large set of accurate data by computing the properties or features that can be used to recognize different samples. There are many methods available for feature extraction. In this stage, the gray scale, Texture and symmetrical features as well as discrete wavelet transform (DWT) method are applied to implement this function.

2.3.1 Discrete wavelet transform

Discrete wavelet transform is one of the efficient features extraction and decomposition techniques that is widely used in different applications. It is applied to convert the images to the frequency domain by using each of a low pass and a high pass filter. It is aimed to split the input image to the four sun-bands (LL, LH, HL, and HH). The mathematical expression of DWT is shown in an equation below [18]:

A. $$\varphi_{(i,j)}(x) = 2^{\frac{i}{2}} \varphi\,(2^{-i} - x - l) \qquad (4)$$

B. Where x refers to the variable, s and l represent the integers that scale and stretch the function ∅ to produce wavelets. Three stage of DWT are applied to extract features from input image.

2.3.2 Gray Scale features

In this step, five types of Gray Scale features are extracted. These features are included as: mean variance, standard deviation, skewness and kurtosi [19]. These are explained as follows:

- **Variance:** is defined as the difference in intensity of gray levels.

$$Variance = \frac{1}{N} \sum_{i=1}^{N} (|x_i - \mu|^2) \qquad (5)$$

Where, $\boldsymbol{x}$ refers to the individual data point, $\boldsymbol{\mu}$ represents the mean of data points and $\boldsymbol{N}$ indicates to the total number of data points

International Journal of Advanced Engineering Research and Science (IJAERS) *[Vol-4, Issue-3, Mar- 2017]*
https://dx.doi.org/10.22161/ijaers.4.3.41 *ISSN: 2349-6495(P) | 2456-1908(O)*

- **Standard Deviation:** defined as the measure of difference asset of data from its mean value.

$$SD = \sqrt{\text{Variance}} \quad (6)$$

- **Skewness:** is defined as the measure of the symmetries of gray level.

$$Skewness = Variance^{-3} \sum_{x=1}^{m} \sum_{y=1}^{n} (f(x,y) - \mu)^{-3} \quad (7)$$

- **Kurtosis:** is defined as the measure of the flatness of the histogram gray levels.

$$Kurtosis = (Variance^{-4}) \sum_{x=1}^{m} \sum_{y=1}^{n} (f(x,y) - \mu)^{-4} \quad (8)$$

2.3.3 Texture Features

They are the second kind of features that are used. In this stage, thirteen features are taken from co-occurrence matrices which are calculated for each input image. Some of these features are described below [18]:

- $$Entropy = -\sum_{i=1}^{n} \sum_{j=1}^{n} p(i,j)\, log(p(i,j)) \quad (9)$$

- $$Dissimalrity = \sum_{i=1}^{n} \sum_{j=1}^{n} p(i,j) * |(i-j)| \quad (10)$$

- $$Inverse = \sum_{i,j=1}^{n} \frac{p(i,j)}{(i-j)^2} \quad (11$$

- $$Energy = \sum_{i=1}^{n} \sum_{j=1}^{n} (p(i,j))^2 \quad (12)$$

- $$Contrast = \sum_{i=1}^{n} \sum_{j=1}^{n} p(i,j) * (i-j)^2 \quad (13$$

- $$IDM = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{p(i,j)}{1+(i-j)^2}$$

2.3.4 Symmetrical feature

$$Exterior\ Symmetry = \frac{\sum_{I=1}^{n} (m-M)^2}{n} \quad (15)$$

2.4 Feature Reduction

The feature reduction stage aims to minimize the size of features set to reduce the time required to do the mathematic operations, as well as to reduce the size of storage without affecting the efficiency of the system. For this reason, feature reduction stage is very impotent to any system. K-PCA is aimed to extract principal components by mapping a set of data in high-dimensional feature space to a low-dimensional feature space. By this way, kernel PCA reduces the complicated coefficient dependencies that otherwise might not be simply reduced in a linear subspace and this gives k-PCA preference over traditional PCA methods [20].

2.5 Training and Classification

This stage aims to classify the features that came from previous stage using SVM. SVM is a binary classifier that depends on supervised learning method. It is used to classify tumor to abnormal or normal. The accuracy of SVM classifier is based on the kernels function. There are different types of kernel functions that can be used with SVM classifier like linear, polynomial and radial basis function [21-22]. The corresponding equations of the widely used kernels functions are shown below:

- *Linear kernel*

$$f_k = f(x, x') \quad (16)$$

Where ***x*** and ***x'*** are represented as the sample vectors and $\boldsymbol{f_k}$ as the linear kernel.

- *Polynomial kernel*

$$k(x, x') = (1 + x.x')^2 \quad (17)$$

Where ***x*** and ***x***' represents the sample vectors, while $\|\boldsymbol{x}-\boldsymbol{x'}\|^2$ represents the Euclidean distance between these sample vectors.

- *RBF kernel*

$$k(x, x') = e^{(-||x-x'||)^2} \quad (18)$$

Where ***x*** and ***x'*** represents the sample vectors.

III. EXPERIMENTS AND RESULTS

The proposed CAD system is developed and simulated using MATLAB® 2016b environment. Graphic USER interface (GUI) is designed also to help any user to use the proposed system easily. Figure 2 is displays the GUI of proposed system. This GUI consists of inputs and outputs for the GUI. The inputs are the loaded image and segmentation pushbutton.

https://dx.doi.org/10.22161/ijaers.4.3.41 *ISSN: 2349-6495(P) | 2456-1908(O)*

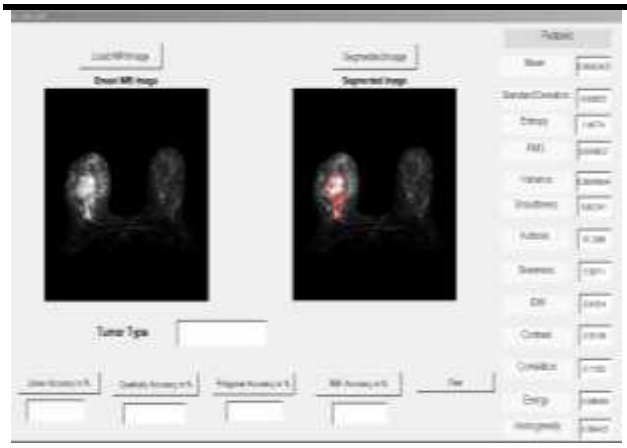

Fig.2: The GUI of the proposed system

The outputs are feature parameters, type of tumor and the performance of the kernel function of the SVM classifier (Linear, Quadratic, and polynomial and RBF kernels). Our database [23-24] consists of 38 cases, 15 of these cases are benign and 23 are malignant. These cases are employed to evaluate the performance of the system. Three other parameters (sensitivity, specificity, and accuracy) are also applied. The sensitivity parameter indicates the proportion of actual positives that are correctly identified, while Specificity parameter represents the proportion of negatives that are correctly identified; and Accuracy parameter, which is the proportion of both true positives and true negatives. The equations of these three parameters are shown in equations below [25]:

$$Sensitivity = \frac{TP}{(TP + FN)} * 100\% \qquad (19)$$

$$Specificity = \frac{TN}{(TN + FP)} * 100\% \qquad (20)$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100\% \qquad (21)$$

True Positives (TP) refers to the correctly identified positive cases, True Negatives (TN) refers to the correctly identified negatives, False Positives (FP) represents the incorrectly identified positives and False Negatives (FN) refers to the incorrectly identified negatives. The performances of the four SVM kernels are shown in table I.

TABLE I. SVM KERNELS PERFORM

Seq.	***Kernel Type***	***Sensitivity***	***Specificity***	***Accuracy***
1.	Linaer	*91.304%*	*84%*	*88.421%*
2.	Quadratic	*90.869%*	*88%*	*89.736%*
3.	Polynomial	*95.217%*	*93.333%*	*94.210%*
4.	RBF	*93.913%*	*92%*	*93.157%*

IV. CONCLUSION

In this paper, CAD to segment and classify breast tumor is designed and implemented. The system consists of six stages (preprocessing, segmentation, features extraction, features reduction, training and classification). The segmentation of tumor is implemented using k-mean clustering and morphological Operators and it is successfully applied for most of the images of the database. The gray scale, Texture and symmetrical features as well as discrete wavelet transform (DWT) method are applied in feature extracted stage to obtain the features from MR images. The classification is implemented using SVM method with four kernel functions (Linear, Quadratic, polynomial and RBF). The system is tested using database consists of 34 cases, 9 of these cases are benign and 24 are malignant. Three parameters (sensitivity, specificity, and accuracy) are also applied to compute the performance of the system. From the observation of the results of these three parameters, it can be noticed that polynomial kernel give the best performance comparing with other SVM kernels.

REFERENCES

[1] Xiao Xuan and Qingmin Liao, "Statistical Structure Analysis in MRI Brain Tumor Segmentation", Fourth International Conference on Image and Graphics,pp.421-426,2007.

[2] Hassan Khotanlou, OlivierColliot, JamalAtif, IsabelleBloch , "3D brain tumor segmentation in MRI using fuzzy classification, symmetry analysis and spatially constrained deformable models", Fuzzy Sets and Systems, Volume 160, Issue 10, pp.1457-1473, 16 May 2009.

[3] Prakash Tunga P, Vipula Singh," Extraction and Description of Tumour Region from the Brain MRI Image using Segmentation Techniques", IEEE International Conference On Recent Trends In Electronics Information Communication Technology, pp.1571-1576,May 20-21, 2016.

[4] Mahmoud Khaled Abd-Ellah, Ali Ismail Awad, Ashraf A. M. Khalaf, and Hesham F. A. Hamed," Design and Implementation of a Computer-Aided Diagnosis System for Brain Tumor Classification ", 2016 28th International Conference on Microelectronics (ICM),pp.73-76, 17-20 Dec. 2016.

[5] A. Latif-Amet, A. Ertzn, and A. Eril, "An efficient method for texture
defect detection: sub-band domain co-occurrence matrices," Image and Vision Computing, vol. 18, no. 67, pp. 543–553, 2000.

[6] Anna N. Karahaliou, et. al."Breast Cancer Diagnosis: Analyzing Texture of Tissue Surrounding Microcalcifications", IEEE Transactions on

https://dx.doi.org/10.22161/ijaers.4.3.41 *ISSN: 2349-6495(P) | 2456-1908(O)*

Information Technology in Biomedicine, vol. 12, no. 6, pp.731–738,2008.

[7] Arnau Oliver et. al., "A Novel Breast Tissue Density Classification Methodology", IEEE Trans. on Info. Tech. in Biomedicine, vol.12, no.1, pp55-65, 2008.

[8] M. P. Arakeri and G. R. M. Reddy, "Computer-aided diagnosis system for tissue characterization of brain tumor on magnetic resonance images," Signal, Image and Video Processing, vol. 9, no. 2, pp. 409–425, 2015.

[9] Mathieu Fauvel ; Jocelyn Chanussot ; Jon Atli Benediktsson," Kernel Principal Component Analysis for Feature Reduction in Hyperspectrale Images Analysis", Proceedings of the 7th Nordic Signal Processing Symposium - NORSIG , 7-9 June 2006.

[10] Sidhu GS, Asgarian N, Greiner R and Brown MRG (2012) Kernel Principal Component Analysis for dimensionality reduction in fMRI-based diagnosis of ADHD. Front. Syst. Neurosci, Vol. 6, article 74, November, 2012.

[11] Padmakant Dhage, M. R. Phegade and S. K. Shah, "Watershed segmentation brain tumor detection", 2015 International Conference on Pervasive Computing (ICPC),pp.1-5, 16 April 2015.

[12] Natarajan.P, Krishnan.N, "MRI Brain Image Edge Detection with Windowing and Morphological Erosion", IEEE International Conference on Computational Intelligence and computing Research, pp: 94-97, 2011.

[13] W.Gonzalez, "Digital Image Processing", 2nd ed. Prentice Hall, Year of Publication 2008.

[14] P. Natarajan; N. Krishnan; Natasha Sandeep Kenkre; Shraiya Nancyand et.al.," Tumor detection using threshold operation in MRI brain images", 2012 IEEE International Conference on Computational Intelligence and Computing Research, 18-20 Dec. 2012.

[15] L. Vincent and P. Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," IEEE Trans. Pattern and Machine Intelligence. vol. 13, no. 6, pp. 583-598, August 1999.

[16] Hengjin Tang; Tatsushi Matsubayashi and Hiroshi Sawada," Blocked Time Step Algorithm for Accelerating k-eans and Fuzzy c-Means", 2015 IEEE International Conference on Systems, Man, and Cybernetics,pp.2561-2566, 9-12 Oct. 2015.

[17] Muzni Sahar; Hanung Adi Nugroho, Tianur; Igi Ardiyanto and et.al." Automated detection of breast cancer lesions using adaptive thresholding and morphological operation", 2016 International Conference on Information Technology Systems and Innovation (ICITSI), 24-27 Oct. 2016.

[18] M. Abo-Zahhad, R. R. Gharieb, S. M. Ahmed, and M. K. Abd Ellah,"Huffman image compression incorporating DPCM and DWT," Journal of Signal and Information Processing, vol. 6, pp. 123–135, 2015.

[19] Rasel Ahmmed and Md. Foisal Hossain," Tumor detection in brain MRI image using template based K-means andFuzzy C-means clustering algorithm", 2016 International Conference on Computer Communication and Informatics (ICCCI), 7-9 Jan. 2016.

[20] Syed Z. Rizvi ; Javad Mohammadpour ; Roland Tóth ; Nader Meskin," A Kernel-Based PCA Approach to Model Reduction of Linear Parameter-Varying Systems", IEEE Transactions on Control Systems Technology ,Vol. 24, Issue: 5, Sept. 2016 .

[21] Abdul Qayyum and A. Basit," Automatic breast segmentation and cancer detection via SVM in mammograms", 2016 International Conference on Emerging Technologies (ICET), 18-19 Oct. 2016.

[22] Y. Zhang and L. Wu, AN MR BRAIN IMAGES CLASSIFIER VIA PRINCIPAL COMPONENT ANALYSIS AND KERNEL SUPPORTVECTOR MACHINE, Progress In Electromagnetics Research, Vol. 130,pp. 369-388, 2012.

[23] Lingle, W., Erickson, B. J., Zuley, M. L., Jarosz, R., Bonaccio, E., Filippini, J., … Gruszauskas, N. (2016). Radiology Data from The Cancer Genome Atlas Breast Invasive Carcinoma [TCGA-BRCA] collection.

[24] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maffitt D, Pringle M, Tarbox L, Prior F. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository, Journal of Digital Imaging, Volume 26, Number 6, pp 1045-1057, December, 2013.

[25] M. G. Sumithra and B. Deepa," Performance analysis of various segmentation techniques for detection ofbrain abnormality", 2016 IEEE Region 10 Conference (TENCON),pp.2056-2061, 22-25 Nov. 2016.