*British Journal of Applied Science & Technology*
*19(1): 1-14, 2017; Article no.BJAST.30420*
*ISSN: 2231-0843, NLM ID: 101664541*

**SCIENCEDOMAIN *international***
*www.sciencedomain.org*

# Mammogram Classification Using Discrete Wavelet Transform Features and a Novel Vector Quantization Technique for Breast Cancer Detection

**Ahmad M. Sarhan[1*] and Radaan A. Al-Dosari[2]**

*[1]Department of Computer Engineering, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia.*
*[2]Saudi Electricity Company, Riyadh, Saudi Arabia.*

***Authors' contributions***

*This work was carried out in collaboration between both authors. Author AMS designed the study, performed the statistical analysis, wrote the protocol and wrote the first draft of the manuscript and managed literature searches. Author RAAD managed the analyses of the study and literature searches. Both authors read and approved the final manuscript.*

***Article Information***

DOI: 10.9734/BJAST/2017/30420
Editor(s):
(1) Vitaly Kober, Department of Computer Science, CICESE, Mexico.
Reviewers:
(1) Giovanni Lucca França da Silva, Universidade Federal do Maranhão, Brazil.
(2) Heydy Castillejos Fenández, National Polytechnic Institute, Mexico.
(3) Valliappan Raman, Swinburne University of Technology, Malaysia.
Complete Peer review History: http://www.sciencedomain.org/review-history/17858

***Original Research Article***

***Received 8th November 2016***
***Accepted 18th December 2016***
***Published 16th February 2017***

## ABSTRACT

In this paper, a digital mammogram classification system is presented. The proposed system uses the Discrete Wavelet Transform (DWT) to obtain features from the input mammogram image. The proposed system suggests a new algorithm for generating the codebook used by the vector quantization (VQ) algorithm to classify the input mammogram (malignant, benign, or normal). The obtained results on the DDSM database indicate the significant performance and superiority of the proposed method in comparison with the state of the art approaches. Simulation results show that the proposed system achieves a high accuracy and sensitivity.

*Keywords: Medical images; breast cancer; discrete wavelet transform (DWT); vector quantization (VQ); mammogram.*

**Corresponding author: E-mail: asarhan@hotmail.com;*

## 1. INTRODUCTION

Breast cancer is the most common cancer among women worldwide. According to the American Cancer Society (ACS), one in nine women will develop breast cancer in her lifetime, with an average age of 62 at diagnosis. Breast cancer is the leading cause of cancer death among females, accounting for 23% of the total cancer cases and 14% of the cancer deaths [1]. Breast cancer is any form of malignant tumor that develops from breast cells. The causes of breast cancer are still unknown and there are no generally accepted measures for preventing it. Therefore, the best way to improve the prognosis for breast cancer is early detection. Fortunately, early detection of this type of cancer allows treatment that could lead to a high survival rate.

Conventional cancer classification methods rely on the subjective interpretation of clinical and histopathological information. Many tumors, however, lack morphologic features that are essential for differential diagnosis [2-3]. Moreover, clinical information can be misleading or incomplete as the subjective interpretation by the observer is often affected by fatigue and other human-related factors. Obviously, errors in the pathology laboratory, especially in distinguishing actual cancer lesions from benign abnormalities, can have serious consequences for cancer patients. Digital mammography, on the other hand, is a computer-aided method and is considered to be one of the most effective imaging modality for breast cancer screening. A mammogram is an x ray of the breast to reveal internal structures and screen for breast cancer. Image processing algorithms performed on mammograms have confirmed the effectiveness of obtaining cancer information using mammogram image analysis.

Many classification and clustering algorithms have been proposed and tested on mammograms. The main difficulty in processing mammogram images, however, is the ultra-high dimensionality of these images. A typical mammogram image may have a spatial resolution of about 3000 x 5000 pixels and a 16-bit intensity level. Therefore, in order to make processing mammograms feasible, it is necessary to first reduce their dimensionality. The Discrete Wavelet Transform (DWT) exhibits one of the highest compression capabilities among other transforms such as the discrete cosine transform (DCT) and the discrete Fourier transform (DFT). In this paper, we reduce the ultra-high dimensionality of the mammogram images by exploiting the superior energy-compactness capabilities of the DWT.

To classify the features extracted from the input mammogram, we propose a new Vector Quantization (VQ)-based technique. Specifically, a novel codebook generation method is introduced. In order to show the robustness and the superiority of the proposed system, its performance is compared to that of the well-known Artificial Neural Network (ANN) system, a state-of-the-art classifier in mammogram classification. To test the proposed system, we use the Digital Database for Screening Mammography (DDSM) [4], a publicly available mammography database representing human breast cancer. The DDSM, a collaborative effort between Massachusetts General Hospital and the University of South Florida, is commonly used by mammographic image analysis researchers. Another publically available mammography database which is also widely used by researchers is the Mammographic Image Analysis Society (MIAS) database [5]. Most of the other mammographic databases [6] are not publicly available.

## 2. MAMMOGRAPHY

### 2.1 Overview

Mammography uses low-dose X-rays, detectors that have high contrast and high-resolution, and an X-ray system designed to image the breasts. Mammograms can detect tumors of the breast that are too small to feel. They also detect about 80% of ductal carcinomas-in-situ (DCIS) which are abnormal cells in the lining of the milk ducts. DCIS cannot usually be detected with a breast self-exam or a clinical breast exam. Mammography can also detect micro-calcifications (small calcium deposits that develop in a woman's breast tissue and are usually benign). In spite of its lower sensitivity in detecting cancers in dense breasts and its use of ionizing radiation, which can be potentially harmful if used repeatedly, mammography is still the most recommended procedure for breast cancer screening and detection [7].

There are two types of mammography systems: screen film mammography (SFM), in which the recording device is a film screen [8] and full-field digital mammography (FFDM) which uses digital detectors as the end recording device [9]. Unlike the SFM, the FFDM provides digital images and

therefore enjoys the advantages offered by digital systems including ease of image processing, manipulation, enhancement, archiving, compression, and transmission. The FFDM also uses less radiation and provides faster results. The screening program often generates a large number of mammograms which are to be carefully examined by the physician. According to the ACS, women should begin annual mammography at age 40 [10]. Younger women have denser breasts making the test less diagnostic. On the other hand, as women age, the breasts become less dense which makes the tissue easier to compress and the potential cancer easier to see by the radiologist. For men, the lifetime risk of getting breast cancer is about 1 in 1,000 [11]. Mammography has a false negative rate of 10%, meaning that 10% of breast cancers are not seen on mammograms. Other tests such as sonograms and magnetic resonance imaging (MRI) can sometimes detect those cases, but currently, there is no proven benefit for routine screening with these modalities.

## 2.2 State of the Art in Digital Mammographic Image Analysis

Most of the work in the literature on mammogram classification has aimed at detecting one or more of the three abnormal structures in mammograms: circumscribed masses, speculated lesions, and microcalcifications. A historical perspective of the evolution in breast imaging is provided by Joe et al. [12]. A detailed assessment of the state of the art in automated techniques for the analysis of digital mammogram images is provided by Bowyer et al. [13]. The development of some derivative digital technologies, mainly digital breast tomosynthesis, was covered by Patterson [14]. Digital tomosynthesis creates a 3D picture of the breast using X-rays. Patterson discussed several aspects of digital breast tomosynthesis such as their strengths, weaknesses, and patient impact.

Sundaram et al. [15] used decision tree and the Fuzzy Association Rule Mining (FARM) to classify mammogram images taken from the MIAS database. They reported that the FARM method gives better performance compared to existing methods. Don et al. proposed an approach to classify mammogram images based on fractal features. They used for classification a multi-layered back propagation algorithm [16]. Jog et al. [17] used Gray Level Difference Method (GLDM) for feature extraction and Support Vector Machine (SVM) for classification. They conducted their experiments on the MIAS database. They reported that the combination of GLDM feature extractor with SVM classifier gives appropriate results [17].

Rao et al. [18] worked on the problem of breast cancer diagnosis using the MIAS database and fuzzy rules. Lowis et al. [19] used a Dual-Tree Complex Wavelet Transform (DTCWT) for feature extraction and neural networks for classification. They tested their method using the MIAS database and reported a 96% accuracy [19]. Gardezi et al. [20] investigated the use of morphological component analysis (MCA) for classifying normal and abnormal tissues in mammograms. They used the Local discrete cosine transform (LDCT) and Curvelet transform via wrapping (CURVwrap) with varying iterations to obtain a morphological decomposition [20]. Pak et al. [21] proposed an algorithm for breast cancer detection and classification based on Non Subsampled Contourlet Transform (NSCT) and Super Resolution (SR). Their algorithm included a pre-processing stage to determine the region of interest (ROI), a feature extraction stage, and a classification stage which used the AdaBoost algorithm. They used the MIAS database to test their algorithm and reported a 43% success rate [21].

We conclude this survey with some interesting findings by researchers. Levenson et al. [22] observed the birds' remarkable capabilities to classify. They reported that some pigeons (Columba livia) have an extraordinary ability to distinguish benign from malignant human breasts. After being trained with differential food reinforcement, the pigeons were able to generalize what they had learned when confronted with new images [22]. The relationship between breast cancer and the eating habits was recently studied by Marianac et al. [23]. They reported that eating late at night could raise the risk of breast cancer. As for the latest advances in the treatment of breast cancer, a recent study by Bundred et al. has found a drug combination that can allow some women to avoid chemotherapy [24]. The study reveals that using Herceptin in combination with another drug before surgery shrinks and may even destroy tumors in less than two weeks in women with HER2, an aggressive form of breast cancer.

## 3. MATERIALS AND METHODS

### 3.1 System Block Diagram

A block diagram representing the proposed system is depicted in Fig. 1.

The input image M to the system is a digital mammogram that represents human breast tissue (normal, benign, and malignant). The raw image dataset used in this paper was obtained from the DDSM database. The original DDSM data is stored in a special binary format. A more convenient format is provided through the Image Retrieval in Medical Applications (IRMA) database [25] which offers 9,852 mammograms from the DDSM database in a 16-bit PNG format. Specifically, the IRMA dataset includes12 collections of normal cases, 15 collections of cancer cases, and 14 collections of benign cases. To reduce the overwhelming computational complexity, we have limited our dataset to 900 random samples; 300 samples of each of the normal, benign and cancer cases. About 600 mammograms were used for training and the rest were used for testing. Fig. 2 depicts three samples from the IRMA dataset showing a normal, benign, and cancer cases.

In the proposed system, the input mammogram image M is first preprocessed. Here, the image M is normalized so that the pixel values have magnitudes in the range between zero and one. The original input image has 16-bit intensity levels, giving pixel values between 0 and 65,535. Next, the preprocessed image X is partitioned into M×N non-overlapping blocks. An example showing the proposed block -partitioning scheme is depicted in Fig. 3 which shows the standard image Lena divided into non-overlapping blocks. The DWT is then applied to each block independently.

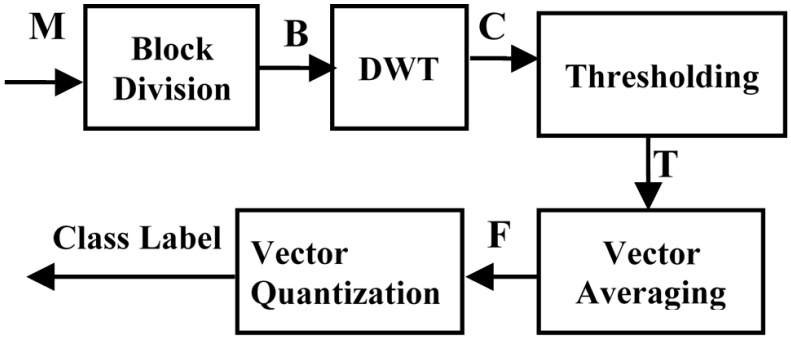
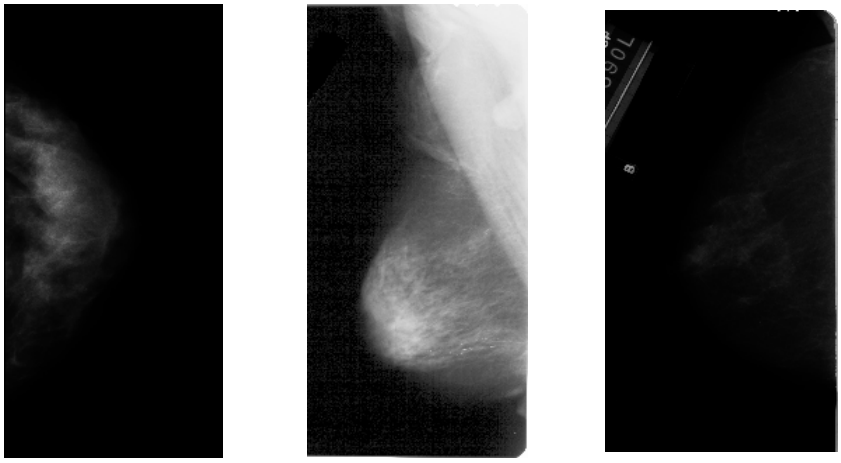



**Fig. 1. Block diagram of the proposed system**



(a) (b) (c)

**Fig. 2. Sample mammograms from the IRMA database: (a) normal, (b) benign, and (c) cancer**

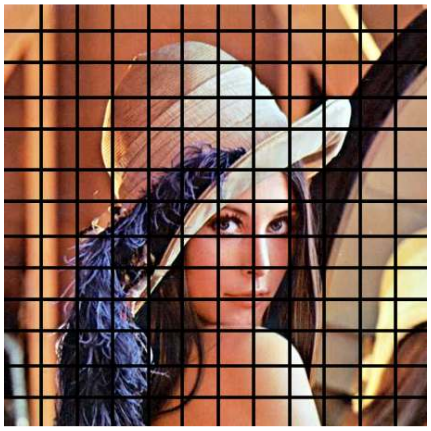


**Fig. 3. The famous Lena image divided into blocks, courtesy of The Mathworks company**

### 3.2 Image Compression Using the Discrete Wavelet Transform

There are two approaches to using the DWT in image compression. The first approach applies the DWT to the whole image. The second approach, followed by the Joint Photographic Experts Group (JPEG) format, processes the input image by dividing it into blocks and transforming each block separately. Both schemes keep only a few coefficients in the transformed region (image/block) while zeroing out the rest of the coefficients in that region [26]. The transform of a signal is just another way of representing the signal. It does not change the energy or information content of the signal. Appropriately chosen transforms can be successfully used in various image processing applications including compression, restoration, and coding. For decades, the Fourier transform (FT) has been the primary transform used in various digital signal processing (DSP) applications. The main weakness of the FT, however, is that it has only frequency resolution and no time resolution. The wavelet transform (WT), on the other hand, has both time and frequency resolutions. Unlike the FT, whose basic functions are sinusoids, the WT is based on filter banks, called wavelets, which have a varying frequency and limited duration. The most commonly used wavelets are the Daubechies (dB) and Haar wavelets. In many cases, the most important part of a signal is contained in the lower-frequency components of the signal spectrum. In wavelet analysis, approximations and details terminology are often used. The approximations are the low-frequency, high-scale components of the signal. The details are the high frequency, low-scale components.

The DWT has proven to be very effective in many DSP applications. Of particular interest to our application at hand is the outstanding energy compactness capability of the DWT which has found many practical applications in signal and image compression. For example, the JPEG format whose compression scheme was originally based on the DCT, currently uses the DWT in their JPEG-2000 image format. In this paper, we exploit the powerful capability of the DWT to effectively compress the huge input mammogram. The strong energy compactness property of the DWT is evident in the small number of high-magnitude coefficients found in the transformed image. To clarify this property, Fig. 4 shows the histogram of the approximation coefficients for a transformed benign mammogram at level 5 using. It is clear from Fig. 4 that the there is only a limited faction of high- magnitude approximation coefficients in the transformed mammogram. For example, only 365 approximation coefficient have magnitudes greater than 15.

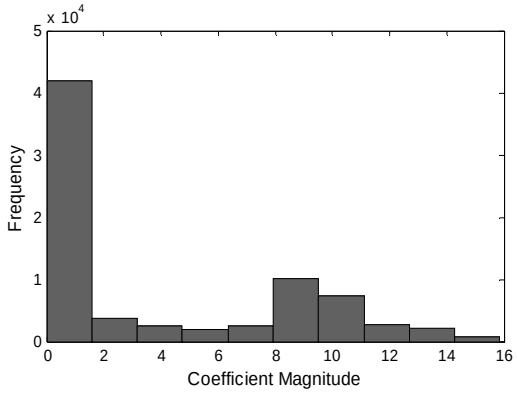



**Fig. 4. Histogram of the approximation coefficients for a transformed benign mammogram at level 5 using the db1 wavelet**

The powerful capability of the DWT to compress the signal energy makes it a good candidate for feature extraction applications. As shown by Fig. 4, the DWT compresses most of the energy of the input image and concentrates it in a few high-magnitude coefficients in the transformed image. Keeping only a number of these high-magnitude coefficients while discarding the rest of the coefficients in the transformed image can produce a valid feature vector representing the input image.

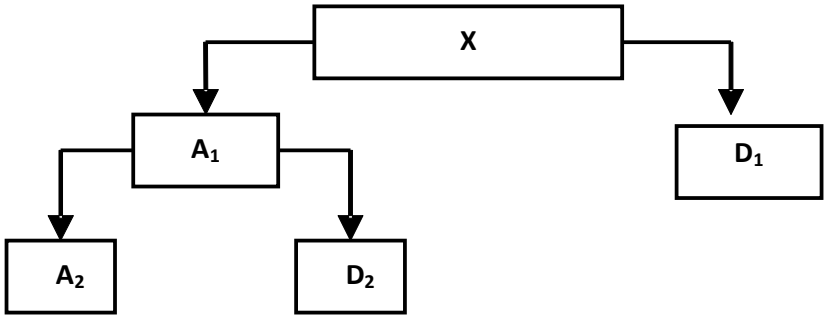



**Fig.5. Wavelet decomposition tree. Variables X, $A_1$ and $D_1$ represent the original image, approximation, and detail coefficients at level 1, respectively**

In the proposed system, the approximation coefficients are extracted from each transformed block. High-magnitude coefficients are then scanned from the approximation coefficients using a novel thresholding technique. The vector average (centroid) of the vectors extracted from all the blocks of the test image is used to form the feature vector representing the test image. The decomposition process of the DWT is illustrated in Fig. 5 above which depicts the Wavelet decomposition tree. As shown in Fig. 5 above, the first level of decomposition yields two vectors of coefficients: approximation and detail coefficients, which represent the characteristics of the image spectrum in the low and high frequency bands, respectively. At level 2, the vector of approximation coefficients undergoes a further decomposition, which reduces its size by half. With each higher decomposition level, the size of the approximation vector is reduced to half of its previous size.

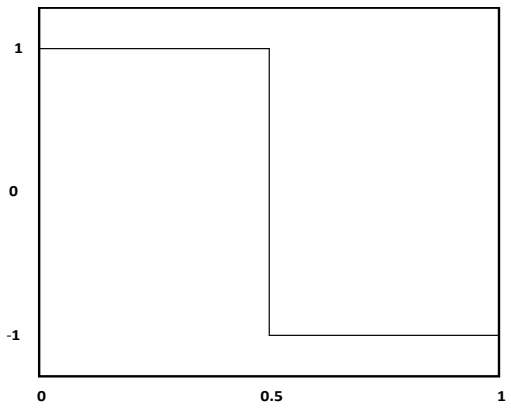



**Fig. 6. The db1 wavelet**

The DWT implemented in the proposed system uses the db1 wavelet. The db1 wavelet is the same as the Haar wavelet which is considered the first and simplest wavelet. The db1 wavelet, depicted in Fig. 6, is discontinuous and resembles a step function.

**Definition 1:** Let the preprocessed image **X** be divided into non-overlapping blocks $\mathbf{B}_i$. Each $\mathbf{B}_i$ is then transformed using the DWT to produce the matrix $\mathbf{W}_i$ which is then converted to the vector $\mathbf{w}_i$. It can be seen that the length or norm of **w**i is given by the following expression:

$$\| \boldsymbol{w} \| = \frac{S^2}{2^{2L}} \quad (1)$$

where S×S denote the block size and L denote the decomposition level.

Eq. (1) gives the number of approximation coefficients in each of the transformed blocks. After dividing the input image into blocks and transforming the blocks, the approximation coefficients are extracted from $\mathbf{w}_i$ to produce the vector $\mathbf{c}_i$. which contains a few high-magnitude coefficients. The limited number of high-amplitude approximation coefficients contain most of the energy of the original image. These high-energy coefficients (positive and negative) are extracted from ci using a masking technique that keeps their values and locations while discarding the rest of the coefficients. Fig. 7 illustrates the masking operation performed on the approximation vector**,** using a threshold value of 9**,** to produce the thresholded coefficient vector $\mathbf{t}_i$.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| -16.2 | 7.5 | -8.9 | -20.1 | 3.2 | 8.4 | -9.1 | 10.7 | 98.2 | -5.2 |

**(a)**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 16.2 | 0 | 0 | 20.1 | 0 | 0 | 9.1 | 10.7 | 98.2 | 0 |

**(b)**

**Fig. 7. Illustration of the masking operation using a threshold value of 9: (a) Input approximation coefficients c and (b) output after thresholding t**

Notice that the proposed thresholding operation not only retains the highest–magnitude coefficients but also keeps track of their locations. The next stage in the block diagram of Fig. 3 is finding the feature vector f representing the input image M. Here f is calculated using a novel approach based on the centroid of all the $t_i$ vectors corresponding to the transformed and thresholded blocks. For example, if the test image M is divided into *k* blocks, then the feature vector f is given by:

$$\mathbf{f} = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{t}_i \qquad (2)$$

where $\mathbf{t}_i$ are the thresholded approximation vectors.

### 3.3 Vector Quantization

In the proposed system, the feature vector **f** is classified by presenting it to a novel VQ-based technique. Traditionally, VQ has been used as a technique in applications involving data compression. In recent years, VQ has been successfully used as a classifier in pattern recognition applications [27]. The property that motivates the use of VQ as a partitioning scheme is its preservation of correlation between vector components; and therefore, the ability to remove spatial redundancies between blocks of samples.

**Definition 2:** A z-level vector quantizer **Q** is a mapping of each input vector $\mathbf{x} \in \mathbb{R}^M$ given by

$$\boldsymbol{x} = [x_1, x_2, x_3, x_4, x_5, \ldots, x_M]^T \qquad (3)$$

to a reproduction vector $\mathbf{z} \in \mathbb{R}^M$ called a codeword, drawn from a set $C = \{\mathbf{z}_i, i=1,\ldots,z\}$ called codebook. A unique index $i = 1, 2, \ldots, z$ is associated with each codeword. Obviously, the codewords $\mathbf{z}_i$, $i$=1, ,…, z, partition the space $\mathbb{R}^M$ into $z$ partitions $s_i$, $i$ = 1,2,…,z such that

$$\mathbf{s}_i = \{ \mathbf{y}: \mathbf{Q}(y)=i\}. \qquad (4)$$

The partitions formed here are called the Voroni regions. Fig. 8 depicts a Voroni diagram.

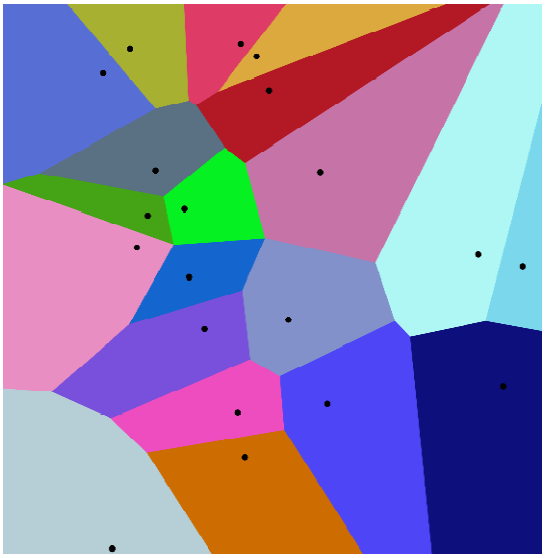


**Fig. 8. Voroni diagram [28]**

The Voroni diagram depicts the partitioning of the plane into regions based on the distance to points in a specific subset of the plane. A famous application of the Voroni diagram was the analysis of the 1854 cholera epidemic in London, in which physician Snow determined a strong correlation of deaths with proximity to a particular (infected) water pump on Broad Street [28].

To encode an input vector, or a feature vector representing an input sample (cancer, benign, or normal), the quantizer measures the distortion (distance) between the input vector and every reproduction vector from the codebook. The codeword yielding the minimum distortion is used as the output vector (class). In other words, determining the class membership of an unknown pattern vector **f** representing a mammogram is done by computing the following distance measure

$$D_i(\boldsymbol{f}) = \| \boldsymbol{f} - \boldsymbol{z}_i \| \quad i = 1,2,3 \qquad (5)$$

Then, **f** is assigned to class $\omega_k$ if $D_k(\boldsymbol{f})$ is the smallest value.

Several distortion measures have been proposed in the literature including the Euclidean distance, the Manhattan distance, the Holder norm, the Hausdorff distance, the Hamming Distance, the Mahalanobis distance, the Chebyshev distance, and the Minkowski distance [29]. The most commonly used distortion measure, however, is the squared-error or Euclidean, distance. The Euclidean distance between two n-dimensional (row or column) vectors x and y is defined as the scalar

$$d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\| = \|\boldsymbol{y}-\boldsymbol{x}\| = [(x_1-y_1)^2+\ldots+(x_n-y_n)^2]^{1/2} \quad (6)$$

This expression is simply the norm of the difference between the two vectors. A simple way to generate the codebook is by using the Linde-Buzo-Grey (LBG) algorithm [30]. The LBG algorithm is commonly used in the design of an optimal codebook from empirical data. In the proposed modified vector quantization (MVQ) algorithm, we propose a novel method for calculating the codebook. Let the codewords refers to the cancer classes (normal, benign, cancer). Thus, the codebook has three codewords. The proposed MVQ algorithm makes use of the inherent averaging nature of the block processing scheme implemented by the proposed system. Specifically, we define the codeword of a class to be the vector average of all the feature vectors comprising the training set of that class. For example, if the pattern class $\omega_k$ has n feature vectors, then the codeword $z_k$ is given by:

$$\boldsymbol{z}_k = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{f}_i, \quad k = 1, 2, 3, \quad (7)$$

where $f_i$ are the feature vectors composing the training set of the class $\omega_k$.

To prove the validity of the proposed MVQ system, its performance is compared to that of an ANN classifier, a state-of-the-art classifier in mammogram classification. ANNs, originally introduced by McCulloch and Pitts in 1943, are trainable algorithms that can "learn" to solve complex problems from training data that consists of pairs of inputs and targets (desired outputs). They can be trained to perform specific tasks such as prediction and classification. ANNs have been successfully used in many applications including pattern recognition, image processing, regression, and adaptive control. All the ANNs examined in this study are two-layer feed-forward neural networks using Backpropagation as the learning algorithm.

## 4. RESULTS AND DISCUSSION

The proposed MVQ system has several parameters including the decomposition level, threshold value, block size, and Wavelet type. In the following experiments, we aim to discover the optimum values for these parameters. In the first experiment, we explore the separability of the output classes. Fig. 9 shows the centroids of the feature vectors $\mathbf{V} \in \mathbb{R}^{64}$ for each of the three classes (normal, benign, an d cancer). Here each centroid was based on the 300 corresponding vectors.

Fig. 9 clearly indicates that the processed feature vectors clearly divide the three classes into three disjoint partitions, with geometric mean values of 19452, 13840, and 7037 for the normal, benign and cancer classes, respectively. In this experiment, the decomposition level L was 3 and the value of the block size S was 64×64. Substituting these values into Eq. (1) produces a length of 64 coefficients for the feature vector, as follows:

$$\|\boldsymbol{V}\| = \frac{64^2}{2^{2x3}} = 64$$

This result is evident from Fig. 9 which also shows that the feature vectors of the normal and benign classes lie closer to each other while both set of features lie farther away from the cancer class. This implies that distinguishing normal cases from benign abnormalities is actually more difficult than distinguishing normal and benign cases from cancer cases. In the following experiments, we investigate the success rates of the MVQ system as a function of threshold value, decomposition level, and block size. In order to show the robustness and the superiority of the proposed system, its performance is compared throughout the experiments, to that of the well-known ANN classifier. Both the MVQ system and the ANN system received the same input features. Simulation results have proved that the proposed MVQ system always produces higher success rates than the ANN system.

In the second experiment, depicted by Fig. 10, the success rates of both the MVQ and the ANN systems versus threshold values, are examined using a 64×64 block size and a decomposition level of 3. Fig. 10 indicates that the success rate is very sensitive to the choice of threshold value, which is depicted by the several local minima and maxima in the success rates. Fig. 10 also clearly shows that the MVQ system outperforms

the ANN system and yields a maximum success rate of 83.4% at a threshold value of 3 whereas the maximum success rate achieved by the ANN system is only 74.1% and occurs at a threshold value of 3.9.

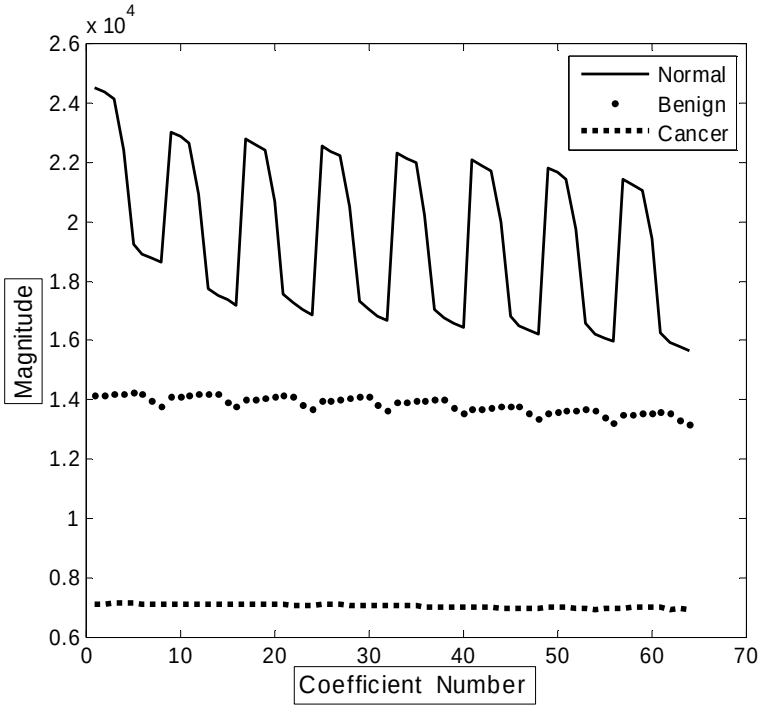



**Fig. 9. Class average of the feature vectors using a decomposition level of 3 and 300 samples for each of the three classes**

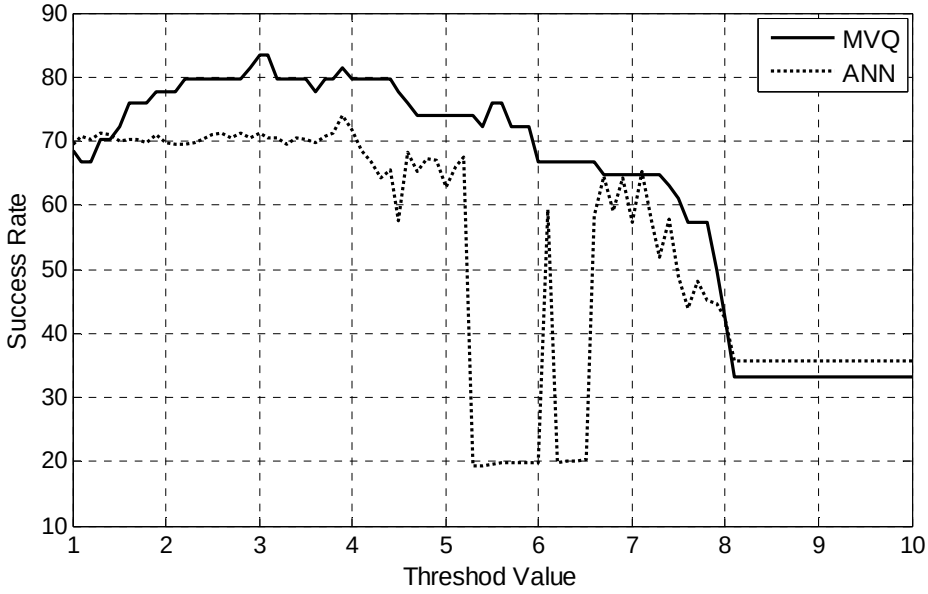



**Fig. 10. Success rates for the MVQ and ANN systems using a block size of 64×64 and a decomposition level of 3**

In the third experiment (Fig. 11), we explore the performance of the two systems when the decomposition level is 4 and the block size is 128×128. Fig. 11 clearly shows that the MVQ system outperforms the ANN system and produces a maximum success rate of 83% at a threshold of 7.5.

Investigated in the fourth experiment, which is depicted by Fig. 12, is the success rate of the MVQ system against decomposition level. Fig. 12 shows that the proposed system achieves a maximum success rate of 83% when the decomposition level is 3, 4, 5, or 6, using different threshold values and block sizes.

Next, we investigate the effects of using different wavelets on the success rate. Fig. 13 depicts the success rates of the MVQ system using the Daubechies (dB) wavelets.

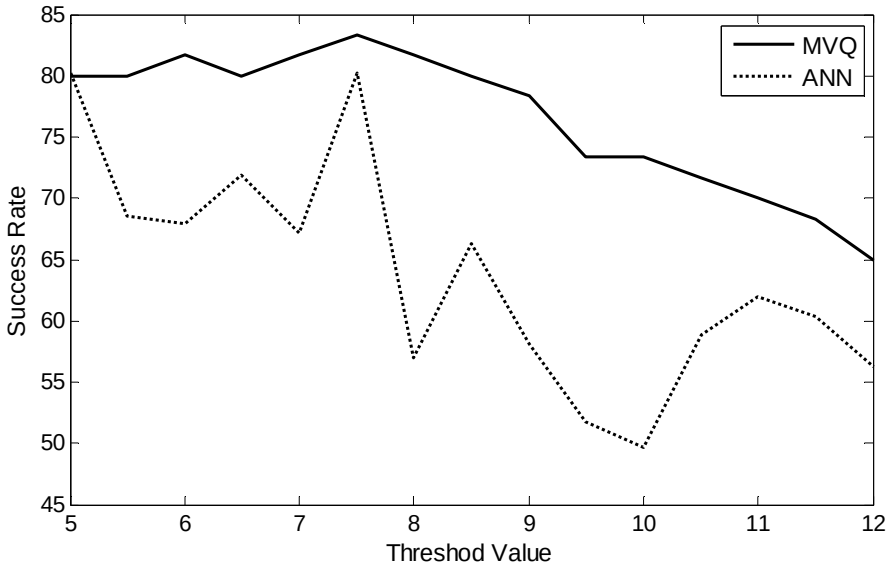



**Fig. 11. Success rates for the MVQ and ANN systems versus threshold values using a block size of 128×128 and a decomposition level of 4**

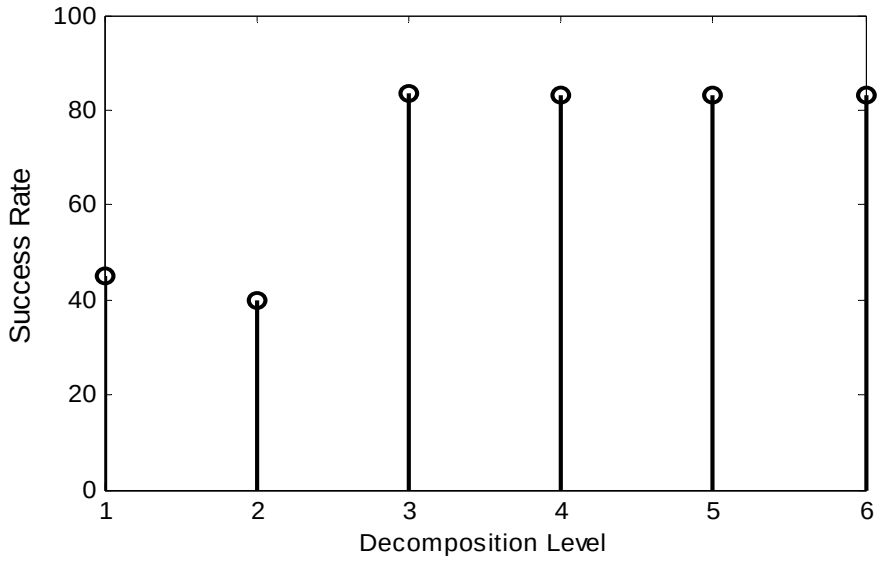



**Fig. 12. Maximum success rates versus decomposition level**

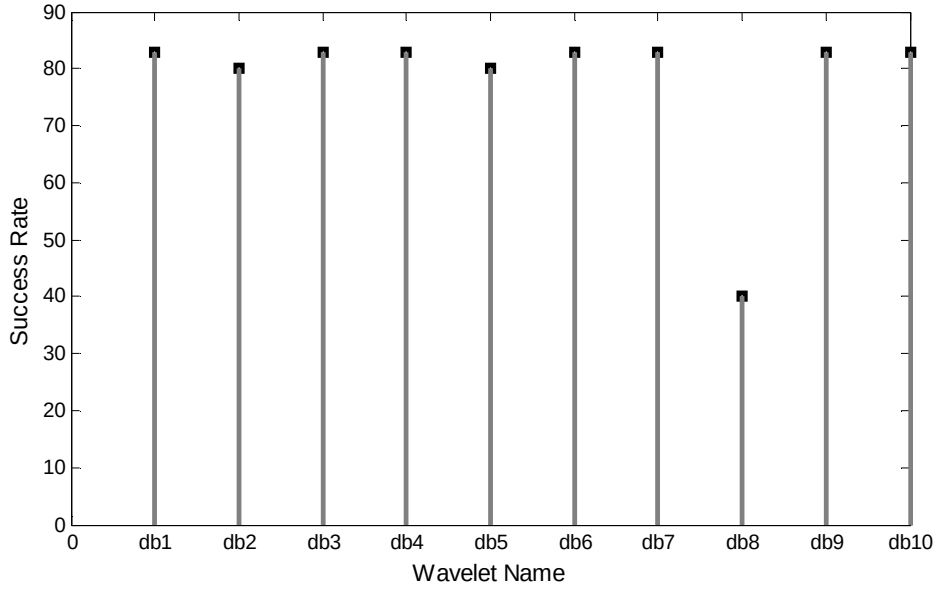



**Fig. 13. Success rate of the MVQ system using the Daubechies wavelets**

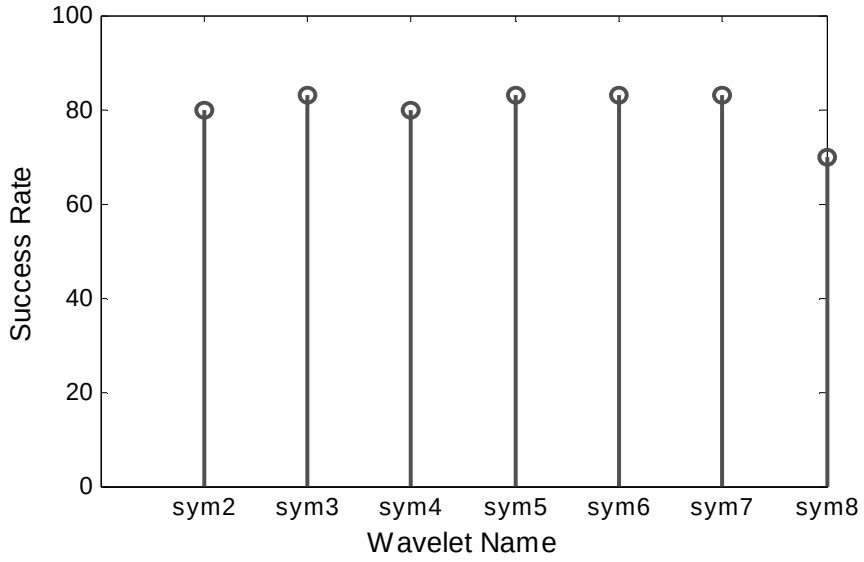



**Fig. 14. Success rate of the MVQ system using the Symmetric (sym) wavelets**

.

Fig. 13 indicates that the maximum success rate occurs when any of the following Daubechies wavelets is used: dB1, dB3, dB4, dB6, dB7, dB9, or dB10. The performance of the proposed MVQ system using the Symmetric (sym) wavelets is shown in Fig. 14.

Fig. 14 shows that the maximum success rate occurs when any of the following Symmetric wavelets is used: sym3, sym5, sym6, or sym7.

The performance of the proposed algorithm is evaluated by computing the percentages of sensitivity (SE), specificity (SP) and accuracy (AC) as follows:

- The sensitivity of a cancer diagnostic test can be defined as its ability to detect cancer when it is present in patients. It is the fraction of real events that are correctly detected among all real events and is given by:

$$SE = \frac{TP\ X\ 100}{(TP+FN)} \quad (8)$$

- The specificity of a cancer diagnostic test can be defined as its ability to show no cancer in healthy patients. It is the fraction of nonevents that are correctly rejected and is given by:

$$SP = \frac{TN\ X\ 100}{(TN + FP)} \quad (9)$$

where,

FP: number of false positive specimens (predicts non-tumor as tumor).
TP: number of true positive specimens (predicts tumor as tumor).
FN: number of false negative specimens (predicts tumor as non-tumor).
TN: number of true negative specimens (predicts non-tumor as non-tumor).

- The prevalence may be determined from the sensitivity, specificity, and accuracy using the equation:

Accuracy = (sensitivity) (prevalence) + (specificity) (1 − prevalence) (10)

The calculated SE, SP, AC, and prevalence are given in Table 1.

**Table 1. Performance metrics of the proposed MVQ system**

| No. of cases | SE | SP | AC | Prevalence |
|---|---|---|---|---|
| 150 | 81% | 79 % | 83% | 2 |

Table 1 shows that the proposed MVQ system produces high sensitivity and specificity values, indicating that the system is robust and reliable.

## 5. CONCLUSION

In this paper, a new method for mammogram classification using VQ and DWT is proposed. Classification features are extracted from the input mammogram using a novel technique that involves image block division, transformation, scanning, and averaging. Specifically, the feature extraction stage utilizes the 2-D DWT to compress each block after dividing the input image into non-overlapping blocks. The approximation coefficients are extracted from the transformed blocks. High-magnitude coefficients are then selected from the approximation coefficients of each transformed block using a threshold-masking method. The vector average of the selected coefficients from all the blocks is used to form the feature vector representing the input image. Also introduced in the proposed MVQ system is a novel method for generating the codebook used by the vector VQ algorithm to classify the input mammogram (tumor, benign, or normal).

The paper investigates through simulations optimal parameters such as the optimal threshold value, block size, decomposition level, and Wavelet type. Experimental tests on the DDSM database achieved 83.4% recognition accuracy using 64×64 blocks, threshold value of 3.9, decomposition level of three, and the dB 1 wavelet.

To show the robustness and the validity of the proposed system, its performance is compared to that of the well-known ANN classifier, the state of the art classifier in this field. Both the proposed MVQ system and the ANN classifier received the same input features. Simulation results have proved that the proposed MVQ system always produces higher success rates than the ANN system.

The following list summarizes the advantages of the proposed MVQ system over the ANN system:

1. The MVQ system always produces higher success rates.
2. Other than the requirement for representative feature vectors to form the codebook, the MVQ system, unlike all ANN system, does not require any training. Consequently, the computational complexity of the MVQ algorithm is much lower than that of the ANN algorithm. Furthermore, to classify an input vector, the MVQ algorithm calculates only the norm of the difference between two vectors, the input vector and each codeword. On the other hand, the ANN algorithm classifies an input vector by calculating weighted linear combinations of the input vector and some weights for all the neurons across all layers, a function which clearly requires many more operations than the simple calculation of the norm. Clearly this classification function further widens the computational complexity gap between the two systems.
3. The ANN system, just like all systems that require training, does not usually exhibit

the same performance on data that it has never seen before.

4. The ANN system produces little or insignificant enhancement in performance for a considerable increase in complexity (number of weights and number of layers).
5. As can be seen from Fig. 13, the design of the ANN may produce several local minima/maxima and there is no straightforward way to specify the location (layer and neuron numbers) of the global minima of the error curve.

## COMPETING INTERESTS

Authors have declared that no competing interests exist.

## REFERENCES


1. Siegel R, Deepa N, Ahmedin J. Cancer statistics, 2013. CA Cancer Journal for Clinicians. 2013;63(1):11–30.
2. Ramaswamy S, Osteen RT, Shulman LN, Lenhard RE, Gansler T. Clinical Oncology. eds. R. E. Lenhard, R. T. Osteen, and T. Gansler. Am. Cancer Soc., Atlanta. 2001;711–719.
3. Sarhan AM. Wavelet-based feature extraction for DNA microarray classification. Artificial Intelligence Review (AIR). 2013; 39(3):237-249.
4. University of South Florida Digital Mammography. Available:http://marathon.csee.usf.edu/Mammography/Database.html
5. The mini-MIAS database of mammograms. Available:http://peipa.essex.ac.uk/info/mias.html
6. Mammographic Image Analysis Homepage. Available:http://www.mammoimage.org/databases
7. Patterson SK, Roubidoux MA. Update on new technologies in digital mammography. International Journal of Women's Health. 2015;6(2014):781–788.
8. van Ravesteyn NT, van Lier L, Schechter CB, Ekwueme DU, Royalty J, Miller JW, Near AM, Cronin KA, Heijnsdijk EA, Mandelblatt JS, de Koning HJ. Transition from film to digital mammography: Impact for breast cancer screening through the national breast and cervical cancer early detection program. Am J Prev Med. 2015;48(5):535-42.
9. Bansal GJ, Lyburn I, Jones T. Is full field digital mammography equivalent in performance to screen film mammography? UK perspective. International Journal of Diagnostic Imaging. 2014;1(2):106-113.
10. Smith AR, Manassaram-Baptiste D, Brooks D, Doroshenk M, Fedewa S, Saslow D, Brawley OW, Wender R. Cancer screening in the United States, 2015: A review of current American Cancer Society guidelines and current issues in cancer screening. CA: A Cancer Journal for Clinicians. 2015;65(1):30–54.
11. Eggemann H, Ignatov A, Smith BJ, Altmann U, von Minckwitz G, Röhl FW, Jahn M. Adjuvant therapy with tamoxifen compared to aromatase inhibitors for 257 male breast cancer patients. Breast Cancer Research and Treatment. 2013; 137(2):465-470.
12. Joe BN, Sickles EA. The evolution of breast imaging: Past to present. Radiology. 2014;273(2 Suppl):23-44.
13. Bowyer KW, Astley S. State of the art in digital mammographic image analysis. Series in Machine Perception and Artificial World Scientific. 1994;9.
14. Patterson SK, Roubidoux MA. Update on new technologies in digital mammography. International Journal of Women's Health. 2014;6(1):781-788.
15. Sundaram KM, Rani PA, Sasikala D. An enhanced mammogram image classification using fuzzy association rule mining. International Journal of Innovative Research in Science, Engineering and Technology. 2014;3:3.
16. Don S, Chung D, Revathy K, Choi E, Min D. A new approach for mammogram image classification using fractal properties. Cybernetics and Information Technologies. 2012;12(2):71-83.
17. Jog NV, Mahadik SR. Implementation of classification technique for mammogram image. IOSR Journal of Electronics and Communication Engineering. 2014;9(1): 76-78.
18. Rao TVN, Govardhan A. Efficient segmentation and classification of mammogram images with fuzzy filtering. Indian Journal of Science & Technology. 2015;8(15):17–19.
19. Lowis, Hendra, Lavinia. The use of Dual-Tree Complex Wavelet Transform (DTCWT) based feature for mammogram classification. International Journal of Signal Processing, Image Processing and Pattern Recognition. 2015;8(3):87-96.

20. Gardezi SJS, Faye I. Mammogram Classification Based on Morphological Component Analysis (MCA) and curvelet decomposition. Neuroscience and Biomedical Engineering. 2015;3:27-33.
21. Pak F, Kanan HR, Alikhassi A. Breast cancer detection and classification in digital mammography based on Non-Subsampled Contourlet Transform (NSCT) and Super Resolution. Comput Methods Programs Biomed. 2015;122(2):89-107.
22. Levenson RM, Krupinski EA, Navarro VM, Wasserman EA. Pigeons (Columba livia) as trainable observers of pathology and radiology breast cancer images. PLoS ONE. 2015;10(11).
23. Marinac CR, Natarajan L, Sears DD, Gallo LC, Hartman SJ, Arredondo E, Patterson RE. Prolonged nightly fasting and breast cancer risk: Findings from NHANES (2009–2010). Cancer Epidemiol Biomarkers Prev. 2015;24:783-789.
24. The European CanCer Organization (ECCO). Combination of lapatinib and trastuzumab shrinks HER2 positive breast cancer significantly in 11 days after diagnosis. ScienceDaily, 10 March 2016. Available:www.sciencedaily.com/releases/2016/03/160310125513.htm
25. Image Retrieval in Medical Applications (IRMA).
Available:http:// irma-project.org
26. Kiran Bala Er, Varinderjit K. Advance digital compression using fast Wavelet transforms comparative analysis with DWT. International Journal of Engineering Sciences and Research Technology. 2016; 5(7):1062–1069.
27. Sarhan AM. A Comparison of vector quantization and artificial neural network techniques in typed arabic character recognition. International Journal of Applied Engineering Research (IJAER). 2009;4(5):805:817.
28. Available:https://en.wikipedia.org/wiki/Voronoi_diagram
29. Ahmad M. Sarhan. Cancer classification based on DNA microarray data using cosine transform and vector quantization. International Journal of Computers and Their Applications (IJCA). 2010;17(4):212-223.
30. Linde Y, Buzo A, Gray RM. An algorithm for vector quantization. EEE Trans. Commun. Theory. 1980;COMM-28:84–95.



*Peer-review history:*
*The peer review history for this paper can be accessed here:*
*http://sciencedomain.org/review-history/17858*