

# GAUSSIAN MIXTURE MODEL BASED CLASSIFICATION OF MICROCALCIFICATION IN MAMMOGRAMS USING DYADIC WAVELET TRANSFORM

<sup>1</sup>Suman Mishra and <sup>2</sup>Hariharan Ranganathan

<sup>1</sup>Department of Electronics Engineering, Sathyabama University, Chennai, India

<sup>2</sup>Department of Electronics and Communication Engineering,  
Rajiv Gandhi College of Engineering, Chennai, India

Received 2013-08-17, Revised 2013-08-24; Accepted 2013-08-28

## ABSTRACT

Breast cancer is a serious health related issue for women in the world. Cancer detected at premature stages has a higher probability of being cured, whereas at advanced stages chances of survival are bleak. Screening programs aid in detecting potential breast cancer at early stages of the disease. Among the various screening programs, mammography is the proven standard for screening breast cancer, because even small tumors can be detected on mammograms. In this study, a novel feature extraction technique based on dyadic wavelet transform for classification of microcalcification in digital mammograms is proposed. In the feature extraction module, the high frequency sub-bands obtained from the decomposition of dyadic wavelet transform is used to form innovative sub-bands. From the newly constructed sub-bands, the features such as energy and entropy are computed. In the classification module, the extracted features are fed into a Gaussian Mixture Model (GMM) classifier and the severity of given microcalcification; benign or malignant are given. A classification accuracy of 95.5% is obtained using the proposed approach on DDSM database.

**Keywords:** Dyadic Wavelet Transform, Gaussian Mixture Model, Digital Mammograms, Microcalcification, Benign and Malignant, DDSM Database

## 1. INTRODUCTION

According to World Cancer Research Fund International, the cancer cases around the world was estimated as 12.7 million in 2008 and is expected to rise to 21 million by 2030 (WCRF, 2013). The maximum cancer cases were found in Denmark. In Denmark, the number of people affected per 100,000 was 326, in 2008. The age standardized rate was at least 300 per 100,000 for Denmark, Ireland, Australia, New Zealand, Belgium, France and United States of America. The top listed ten countries for occurrences of breast cancer are from the countries of Europe, Oceania and North America (WCRF, 2013). The causes of cancers are primarily due to an environmental disease

with 90-95% of cases attributed to environmental factors and 5-10% to genetics (WCRF, 2013). In this study, a novel technique to classify the microcalcifications into benign or malignant in digital mammograms is discussed.

El-Naqa *et al.* (2002) presented the approach based on Support Vector Machines (SVMs) for the detection of Microcalcification (MC) clusters in digital mammograms. The input pattern to the SVM classifier is a small pixel window placed centered at the location of interest. Wei *et al.* (2005a) implemented an algorithm based on Relevance Vector Machine (RVM) for the detection of MCs in digital mammograms. RVM classifier could greatly reduce the computational complexity. Karahaliou *et al.* (2008) reviewed the breast

**Corresponding Author:** Suman Mishra, Department of Electronics Engineering, Sathyabama University, Chennai, India

cancer diagnosis based on tissue surrounding MC clusters on mammograms. Gray-level texture and wavelet coefficient texture features at three decomposition levels are extracted from surrounding tissue regions of interest.

Peng *et al.* (2009) described the detection of microcalcifications using Stochastic Resonance (SR) noise. Initially, suitable dose of noise is added to the abnormal mammograms. Further, more SR noise based detection approach is presented to improve some suboptimal detectors. Wei *et al.* (2005b) classified the malignant and benign clustered microcalcifications based on several machine learning methods such as SVM, kernel fisher discriminant, RVM and committee machines. Receiver operating characteristic is used to evaluate and to compare classification performance by the different methods. Nakayama *et al.* (2006) developed a computerized scheme for detecting early stage microcalcification clusters in mammograms based on filter bank. Hessian matrix approach is used for classifying nodular structures and linear structures.

Sameti *et al.* (2009) reviewed the early detection of breast cancer based on image feature extraction in the last screening mammograms. The malignant mass developed region and a region which appeared similar to region one on the same mammogram is marked to evaluate the breast cancer. Sixty two texture and photometric image features are calculated for these marked regions. Samulski and Karssemeijer (2011) presented the optimizing case based detection performance in a multi view computer aided diagnosis system for digital mammography. A single view lesion detection system is developed. The method is applied to the problem of detecting malignant masses and architectural distortions.

A new method of feature extraction from wavelet coefficients for classification of digital mammograms is proposed by Faye and Samir (2009). A matrix is constructed by putting Wavelet coefficients of each image of a building set as a row vector. It consists of selecting the threshold, the columns which will maximize the Euclidian distances between the different class representatives. The selected columns are then used as features for classification. Classification of microcalcification using dual tree complex wavelet transform and SVM is proposed by Tirtajaya and Santika (2010). It consists of two phases. At the offline phase, training for the SVM is conducted using some training data to find the support vectors. At

the online phase, a mammogram to be classified inputted into the system and then classified by the SVM.

In this study, an efficient method is proposed to effectively classify the microcalcification in digital mammograms into benign and malignant categories. Dyadic wavelet transform is used for feature extraction and Gaussian mixture model is used for classification.

## 2. SYSTEM DESIGN AND PREPROCESSING

The framework for the proposed classification system is depicted in “**Fig. 1**”. A sequence of preprocessing steps is performed before feature extraction and classification. From the preprocessed image, the proposed dyadic wavelet transform based features are extracted and then fed into the classifier. The classification accuracies are computed finally.

In order to eliminate the unwanted portions such as label on the image and background, a series of preprocessing steps is performed on all the images in the database. It starts with reducing the spatial resolution of the image to 1024×1024 pixels by bicubic interpolation technique. The other preprocessing steps include border correction, morphological dilation, extraction of breast region by connected components and zero padding. It is essential to remove the areas that are irrelevant to breast tissue for better performance.

The border correction is performed in order to remove the non x-ray film region in the given image. As most of the images have non x-ray film regions, border correction is performed by removing it; that is the gray intensities in the top, bottom, left and right side of the mammogram image are removed. About 25 pixels are removed by black pixels of zero intensity. The remaining non x-ray file region is removed by the subsequent steps if exists. After border correction, morphological dilation is applied for expanding the shape of the breast as well as label on the image by using disk shaped structuring element. As the size of the breast region is bigger than other regions, the biggest region labeled by the connected components is extracted and considered as the breast area for further feature extraction technique. “**Figure 2**” shows the various stages involved in the proposed preprocessing of mammogram images.

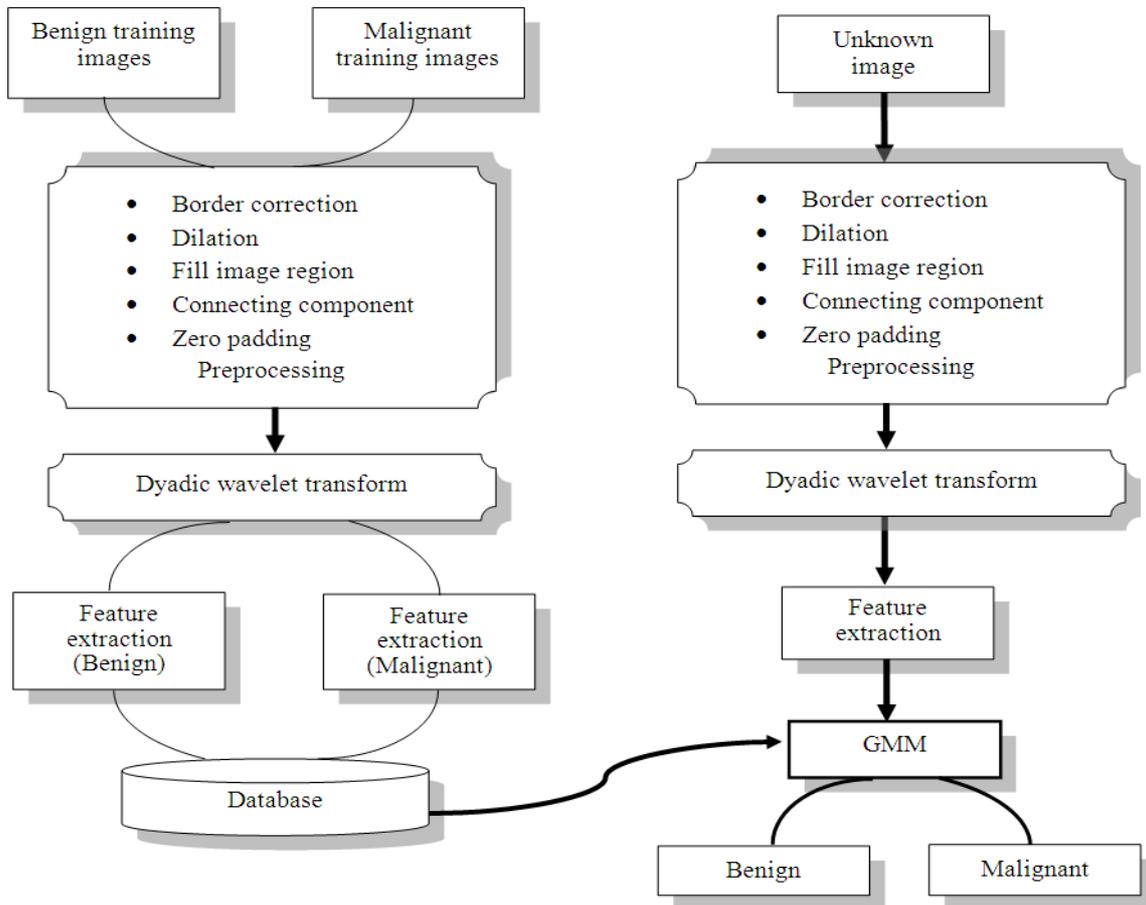


Fig. 1. Automated system for microcalcification classification in digital mammogram

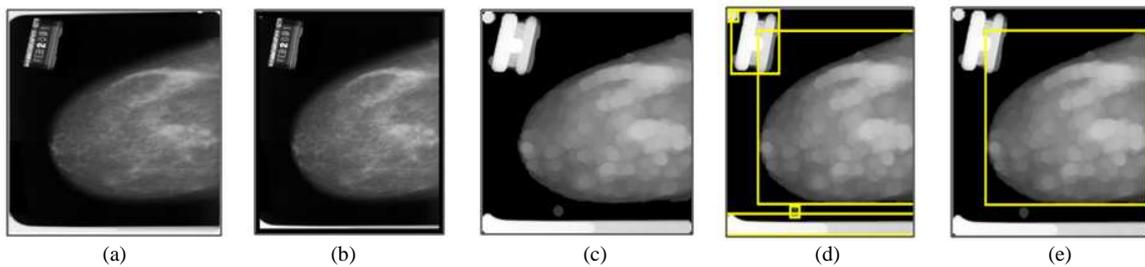


Fig. 2. Preprocessing Stages (a) Original Image (b) Resized Image (c) Morphological Dilated Image (d) Labeled Image (e) Biggest Region of the Image.

### 3. DYADIC WAVELET TRANSFORM BASED FEATURE EXTRACTION

The discrete wavelet transform is obtained by iterating an M-channel multirate filter bank on its low-

pass output. The wavelet transform with symmetric basis functions has recently emerged as the next-generation standard for various image application areas. Wavelet decomposition arises from iteration of the low-pass filtering and decimation steps of a multirate filter bank. For a true wavelet, one iterates on the low-pass output

only, whereas for wavelet packet decomposition, one may iterate on any output. A finite number of iterations will lead to a discrete time multiresolutional analysis with low-pass frequency response  $\prod_{k=1}^n H_0\left(\frac{\omega}{2^k}\right)$ . If the low-pass filter  $h_0$  satisfies the orthonormality constraint and “one vanishing moment” (approximation of order 1):  $\sum_k h_0[k] = \frac{1}{\sqrt{2}}$ , then the infinite product  $\lim_{n \rightarrow \infty} \prod_{k=1}^n H_0\left(\frac{\omega}{2^k}\right)$  converges to a continuous-time function  $\hat{\phi}(\omega)$ , whose inverse Fourier transform  $\phi(t)$  is called the scaling function. The scaling function  $\phi(t)$  and the wavelet  $\psi(t)$  are solution of the dilation equations in “Equation (1)”:

$$\begin{aligned} \phi(t) &= \sum_k h_0(k) \phi(2t-k), \\ \psi(t) &= \sum_k h_1(k) \phi(2t-k) \end{aligned} \tag{1}$$

and it is orthogonal to its integer translates (under mild additional assumptions). If the filter  $h_0(n)$  is FIR, then  $\phi(t)$  and  $\psi(t)$  have compact support. The set of dilates and translates in “Equation (2)”:

$$\left\{ \psi(2^k t - \ell) \right\}_{k, \ell \in \mathbb{Z}} \tag{2}$$

which forms a tight frame (and in most cases an orthonormal basis) for  $L^2(\mathbb{R})$  (Steffen *et al.*, 1993). The span of integer translates the scaling function  $\phi(t)$  is the “low-pass” space  $V_0$ , the set of scale limited signals (Gopinath *et al.*, 1994). Any continuous-time function  $f(t)$  in  $V_0$  can be expanded as a linear combination of “Equation (3)”:

$$f(t) = \sum_n v_n^{(0)} \phi(t-n) \tag{3}$$

The superscript (0) denotes an expansion “at scale level 0”,  $f(t)$  is completely described by the sequence  $\{v_n^{(0)}\}$ . Given such a sequence, its coarse approximation (component in  $V_1$ ) is computed with low-pass filter of the wavelet filter bank in “Equation (4)”:

$$v_n^{(1)} = \left( (v^{(0)} * h_0) \downarrow 2 \right) [n] \tag{4}$$

Analogously, the details  $\omega_n^{(1)}$  in  $V_0 \ominus V_1$  are computed with the high-pass filter  $h_1(n)$ . Hence, a discrete sequence  $v_n$  to be the coefficients of a signal  $f(t)$  at some fixed scale, the discrete wavelet transform of  $v_n$  will decompose the underlying signal  $f$  into a coarse-scale component and detail at several intermediate scales. This decomposition matches multiresolutional models of human and computer vision (Mallat, 1989). In the proposed approach, the dyadic wavelet transform is used to represent the mammograms in multi-scale.

The preprocessed image is decomposed by using the dyadic WT at  $n$ -level decomposition. The decomposed image contains a low frequency sub-band (approximation coefficient) and  $3 \cdot n$  high frequency sub-bands (detailed coefficients). A novel approach is applied to form a new sub-band by retaining the maximum response of detailed coefficients of each level. For simplicity, let us consider 1-level decomposition of dyadic WT which produces 1 low frequency and 3 high frequency sub-bands. As high frequency sub-bands provide the detailed information about the given image, only the high frequency sub-bands are considered for feature extraction. A new sub-band is formed by selecting the maximum coefficients value of the 3-high frequency sub-bands at a particular position.

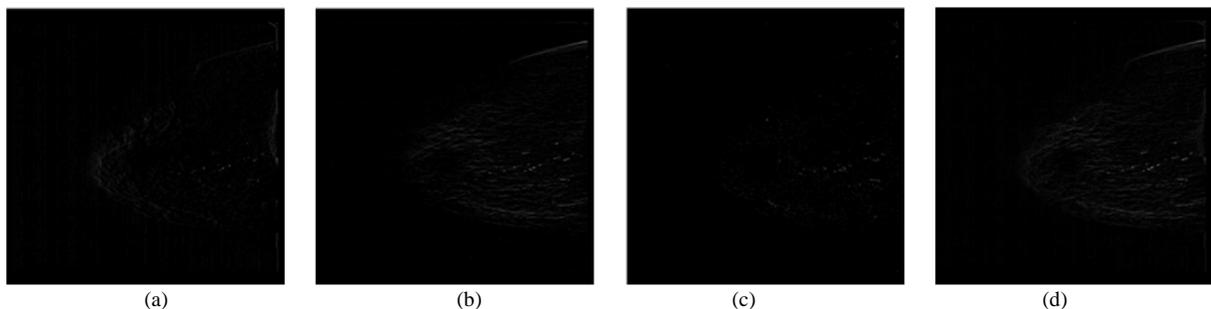


Fig. 3. (a-c) High Frequency Sub-bands (d) Newly Formed Sub-band

“Figure 3” shows the high frequency sub-bands and the newly formed sub-band. After forming the new sub bands, the energy and entropy features are computed. This method is applied to all the training images and the feature vectors are stored in the database. Based on these two features, the proposed approach is evaluated by using GMM based classifier.

#### 4. GAUSSIAN MIXTURE MODEL

Gaussian Mixture Model (GMM) is commonly used in pattern recognition and machine learning algorithms. This classifier is able to approximate the distribution of the patterns representing the characteristics of a texture in an image. During training, the induction algorithm estimates the mixture of Gaussian models that best approximates the distribution of the given values. Formally, a texture pattern is described by a mixture of M-Gaussian models  $\Gamma = \{\gamma_1, \dots, \gamma_M\}$ : the mixture density is a weighted sum of the M component densities. Given an input feature vector  $\vec{\chi}$ , the conditional probability is computed from the mixture as follows in “Equation (5)”:

$$p(\vec{\chi}|\Gamma) = \sum_{m=1}^M c_m \cdot \gamma_m(\vec{\chi}) \tag{5}$$

where,  $c_m$  are the mixture weights and  $\gamma_m(\vec{\chi})$  an N-variate Gaussian function. The dimensionality of the Gaussian function (N) coincides with the dimensionality of the feature vector  $\vec{\chi}$  while the M models and relative weights are estimated from the training data using a special case of the Expectation-Maximization (EM) algorithm (Reynolds and Rose, 1995). A set of S textures  $\{s_1, \dots, s_S\}$  is represented by S Gaussian mixture models  $\{\Gamma_1, \dots, \Gamma_S\}$ . A given observation sequence  $X = \{x_1, \dots, x_T\}$  is tested by finding the pattern in a texture which has maximum a posteriori probability. By applying Bayes’ rule and using the logarithm (Reynolds and Rose, 1995), the probability can be computed as in “Equation (6)”:

$$\hat{S} = \arg \max_{1 \leq s \leq S} \sum \log p(\vec{x}_t | \Gamma_s) \tag{6}$$

The GMM can assume different forms depending on the choice of covariance matrix used in the estimation of the N-variate Gaussian functions. In this proposed approach GMM is used as a classifier for the

classification of microcalcification into benign or malignant categories.

#### 5. EXPERIMENTS AND RESULTS

The performance of the proposed system is evaluated by using DDSM database of the University of South Florida (Heath *et al.*, 1998). Over the last decade, DDSM has been a standard test bench for mammographic image analysis research community. It has 2620 cases available in 43 volumes. The test set consists of 257 mammograms, containing benign (57) and malignant (200) microcalcifications. The selected mammograms included lesions with different breast densities and different degree of subtlety.

Among the samples 40% of samples in each category are used to test the accuracy of the system and the remaining are used to build the classifier. Initially, all the selected mammograms are rescaled to 1024×1024 pixels by using bicubic interpolation technique in order to preserve the spatial resolution uniformity in the images.

An automated segmentation of breast tissue is developed based on morphological operation. The objective of this segmentation is to extract the proposed features from the breast tissue only. In the feature extraction stage, the dyadic wavelet transform is used to represent the mammogram image characteristics at different resolution level. Then the proposed features are extracted. In order to find the most suitable resolution level at which the feature set may provide best discrimination among the classes such as benign, malignant different levels of resolution are tested with GMM classifier. The number of Gaussians used in GMM classifier also varied to find the better classification accuracy. “Figure 4” shows the performance comparison based on the number of Gaussian used in GMM for the proposed features such as energy, entropy and feature fusion independently. “Figure 5” shows the comparison of features used in the proposed approach by varying the number of Gaussian used. Experimental result demonstrates that the proposed system is not very sensitive to the extracted features. However, the maximum accuracy can always be obtained as long as the number of Gaussian used is 64 and the fusion of energy and entropy features produces 95.5% accuracy slightly over than their individual performance. In this study, K-fold cross validation experiments are used to assess the performance of the proposed method.

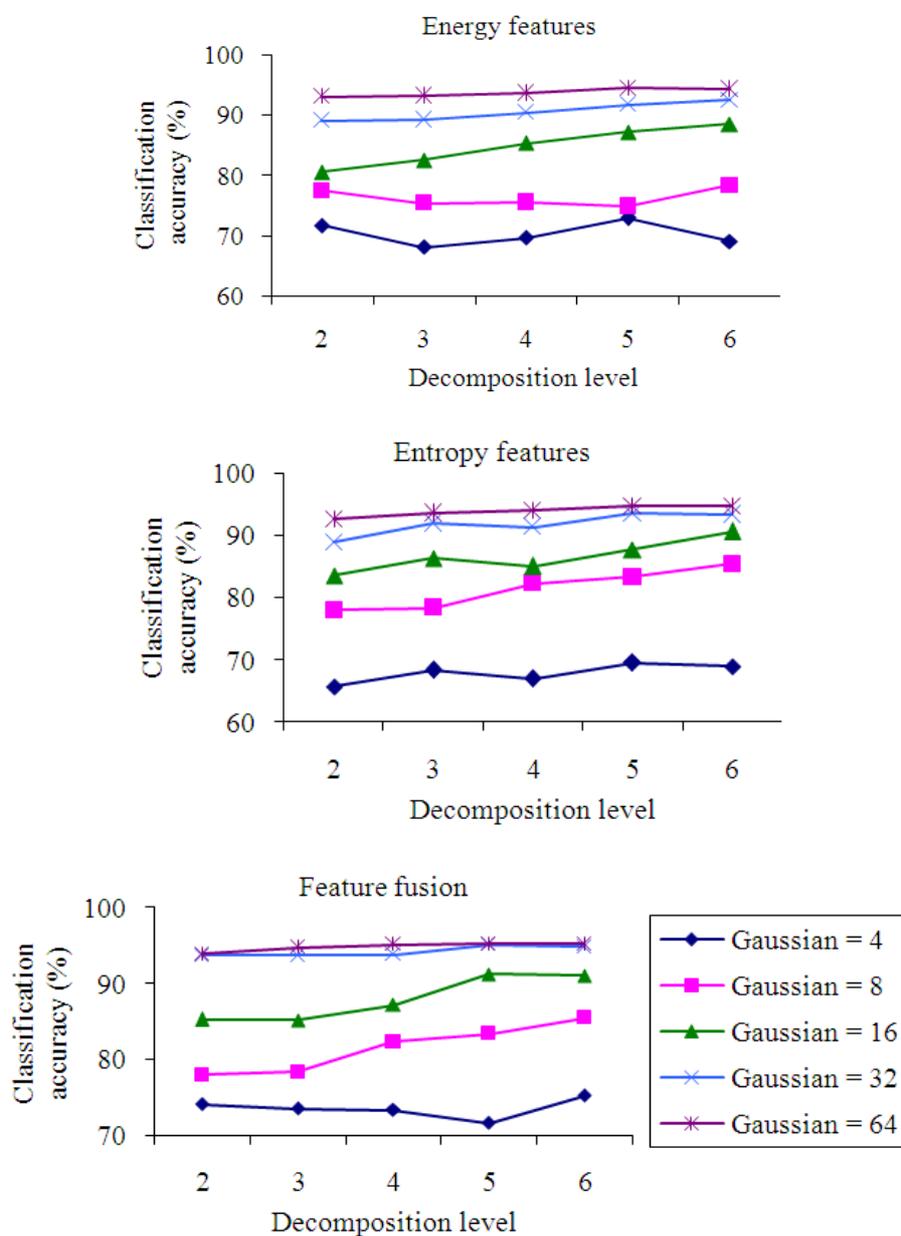


Fig. 4. Performance Comparison based on the Number of Gaussian used in GMM for the proposed features

Table 1. Performance based on the combination of energy and entropy features using 64 Gaussian in GMM

Level of decomposition	Accuracy	Precision	Recall	F-score	Specificity
2	0.939	0.847	0.960	0.895	0.932
3	0.947	0.857	0.979	0.909	0.936
4	0.951	0.865	0.983	0.916	0.941
5	0.952	0.867	0.983	0.917	0.942
6	0.952	0.867	0.983	0.917	0.942

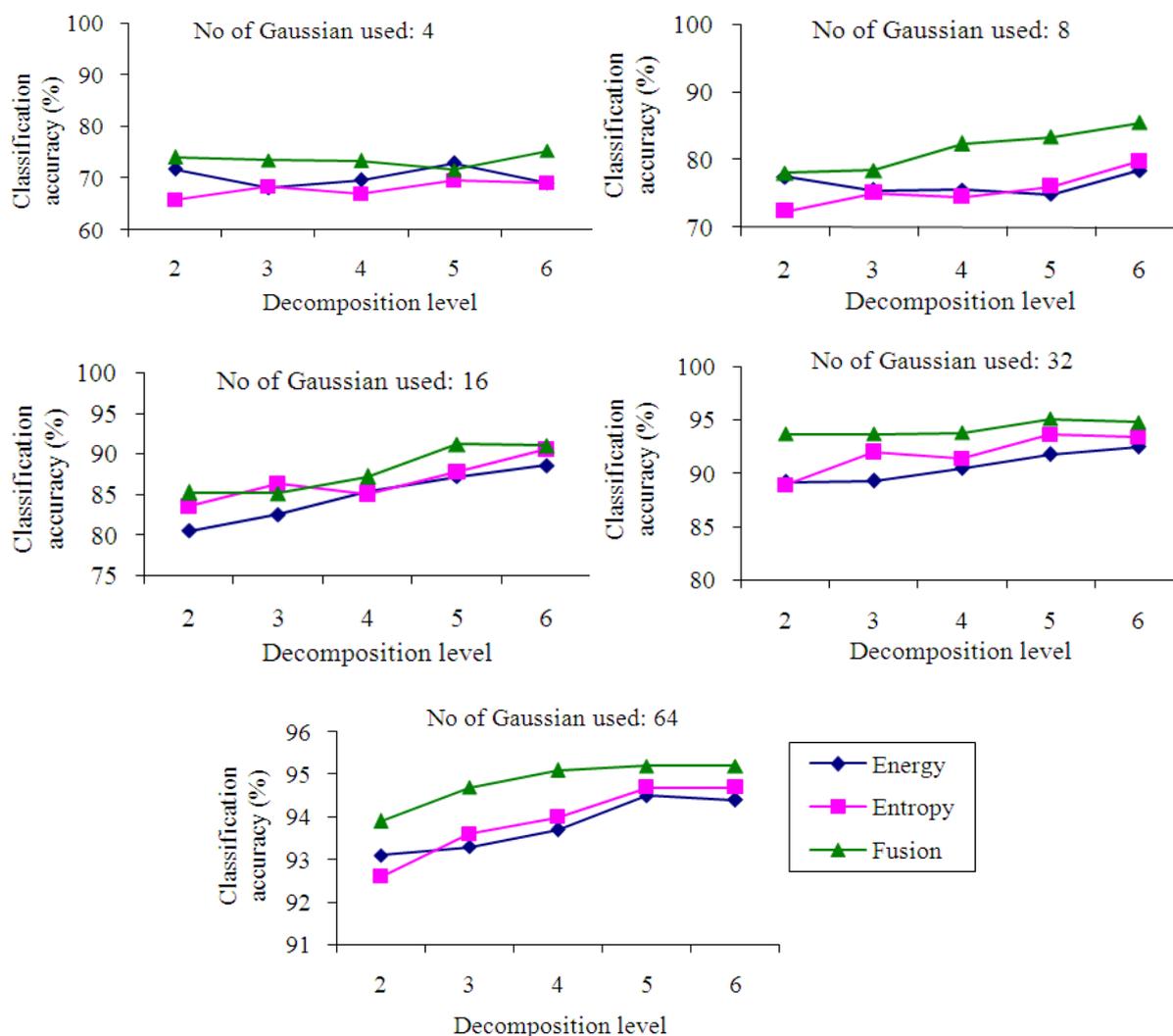


Fig. 5. Performance comparisons of the proposed features

K-fold cross validation is widely used to analyze the classification performance of a classification system. The results shown in the graphs are averaged from 10 fold cross validation experiments. The other metrics used to analyze the system are precision, recall, F-score and specificity. High classification accuracy is achieved while using the fusion of features. "Table 1" shows the aforementioned metrics for the same using 64 numbers of Gaussian in GMM.

## 6. CONCLUSION

In this study, dyadic wavelet transform is investigated for the classification of microcalcification in

digital mammograms. For the classification task, the robust GMM classifier is used. In feature extraction, the multi resolution characteristics of a given digital mammogram are explored by the use of dyadic wavelet transform. In order to analyze the performance of the proposed system, the parameters such as decomposition level, number of Gaussian used in GMM are varied. The decomposition level is varied from 2 to 6 and the number of Gaussian used is 4, 8, 16, 32 and 64. Results show that higher classification accuracy can be obtained by fusing the energy and entropy features extracted from dyadic wavelet transform and classifying with GMM classifier.

## 7. REFERENCES

- El-Naqa, I., Y. Yang, M.N. Wernick, N.P. Galatsanos and R.M. Nishikawa, 2002. A support vector machine approach for detection of microcalcifications. *IEEE Trans. Med. Imag.*, 12: 1552-1563. DOI: 10.1109/ICIP.2002.1040110
- Faye, I. and B.B. Samir, 2009. Digital mammograms classification using a wavelet based feature extraction method. *Proceedings of the IEEE Conference on Computer and Electrical Engineering*, Dec. 28-30, IEEE Xplore Press, Dubai, pp: 318-322. DOI: 10.1109/ICCEE.2009.39
- Gopinath, R.A., J.E. Odegard and C.S. Burrus, 1994. Optimal wavelet representation of signals and the wavelet sampling theorem. *IEEE Trans. Circ. Syst. II: Analog Digital Signal Process.*, 41: 262-277. DOI: 10.1109/82.285705
- Heath, M., K.W. Bowyer, D. Kopans, P. Kegelmeyer Jr and R. Moore *et al.*, 1998. Current Status of the Digital Database for Screening Mammography. *Digital Mammography*, 13: 457-460. DOI: 10.1007/978-94-011-5318-8\_75
- Karahaliou, A.N., I.S. Boniatis, S.G. Skiadopoulos, F.N. Sakellaropoulos and N.S. Arikidis *et al.*, 2008. Breast cancer diagnosis: Analyzing texture of tissue surrounding microcalcifications. *IEEE Trans. Inform. Technol. Biomed.*, 6: 731-738. DOI: 10.1109/TITB.2008.920634
- Mallat, S.G., 1989. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Patt. Anal. Mach. Intell.*, 11: 674-693. DOI: 10.1109/34.192463
- Nakayama, R., Y. Uchiyama, K. Yamamoto, R. Watanabe and K. Namba, 2006. Computer-aided diagnosis scheme using a filter bank for detection of microcalcification clusters in mammograms. *IEEE Trans. Biomed. Eng.*, 53: 273-283. DOI: 10.1109/TBME.2005.862536
- Peng, R., H. Chen and P.K. Varshney, 2009. Noise-enhanced detection of micro-calcifications in digital mammograms. *IEEE J. Selected Top. Signal Process.*, 1: 62-67. DOI: 10.1109/JSTSP.2008.2011162
- Reynolds, D.A. and R.C. Rose, 1995. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech Audio Process.*, 3: 72-83. DOI: 10.1109/89.365379
- Sameti, M., R.K. Ward, J. Morgan-Parkes and B. Palcic, 2009. Image feature extraction in the last screening mammograms prior to detection of breast cancer. *IEEE J. Selected Topics Signal Process.*, 1: 46-52. DOI: 10.1109/JSTSP.2008.2011163
- Samulski, M. and N. Karssemeijer, 2011. Optimizing case-based detection performance in a multiview CAD system for mammography. *IEEE Trans. Med. Imag.*, 4: 1001-1009. DOI: 10.1109/TMI.2011.2105886
- Steffen, P., P.N. Heller, R.A. Gopinath and C.S. Burrus, 1993. Theory of regular M-band wavelet bases. *IEEE Trans. Signal Process.*, 41: 3497-3511. DOI: 10.1109/78.258088
- Tirtajaya, A. and D.D. Santika, 2010. Classification of microcalcification using dual-tree complex wavelet transform and support vector machine. *Proceedings of the IEEE International Conference on Advances in Computing, Control and Telecommunication Technologies*, Dec. 2-3, IEEE Xplore Press, Jakarta, pp: 164-166. DOI: 10.1109/ACT.2010.52
- WCRF, 2013. World Cancer Research Fund International.
- Wei, L., Y. Yang, R.M. Nishikawa and Y. Jiang, 2005b. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. *IEEE Trans. Med. Imag.*, 24: 371-380. DOI: 10.1109/TMI.2004.842457
- Wei, L., Y. Yang, R.M. Nishikawa, M.N. Wernick and A. Edwards, 2005a. Relevance vector machine for automatic detection of clustered microcalcifications. *IEEE Trans. Med. Imag.*, 24: 1278-1285. DOI: 10.1109/TMI.2005.855435