

# A New Feature Extraction Approach Using Contourlet Transform and T-Test Statistics for Mammogram Classification

N. GEDİK

**Abstract—** In this study, a CAD system is recommended for the classification of mammographic images. The images are classified as normal-abnormal and benign-malignant. The proposed system consists of three basic steps: the feature extraction, determination of the distinguishing capabilities of the features and selection, and classification. The distinguishing capabilities of the features mean determining the best or optimal features. Thanks to this determination, mammograms could be put into classes with high accuracy. The determination process is carried out using thresholding and t-test statistics. Classification is performed repeatedly for all threshold values using support vector machine. Among the obtained results of the classification, the optimal feature set, which has the best classification performance, is selected. Finally, to evaluate the optimal feature set, classification carries out applying 5-fold cross-validation.

**Index Terms—** Classification, CAD, Contourlet transform, Mammogram, SVM.

## I. INTRODUCTION

CANCER IS one of the most serious health problems globally. Of all the types of cancer, breast cancer is the most common type of cancer and threatens women (especially those over 40 years of age). It is stated that the number of new cases diagnosed as breast cancer is approximately 250,000 each year [1]. Struggling with cancer successfully, and reducing the mortality rates can only be achieved with early diagnosis and appropriate treatment [2, 3]. In order to diagnose breast cancer at an early stage, mammography is the widely used and accepted methodology by radiologists [4]. Screening mammograms is a challenging and busy task, and some cases can be overlooked and misinterpreted. Studies show that 10%-30% of cases of cancer have been overlooked (false negative), and 20%-30% of cases sent to biopsy do not have cancer (false positives) [5, 6]. Development of automated systems is a requirement to support experts as the second opinion in reducing false negative and positive rates. Computer-aided

diagnosis (CAD) systems provide appropriate vision and orientation for diagnosis [7, 8].

The development of CAD systems is a progressive area and one of the most important stages in the system is the selection and extraction of features. There are different approaches to feature extraction in CAD studies [9, 10]. Spatial data can be used to obtain features directly, or a transform can be used to obtain the features via a different domain by transferring the data [11,12]. Gedik [13] proposes a feature extraction approach combining fast finite shearlet transform (FFST) and t-test statistics. FFST is employed to decompose mammograms and to get shearlet coefficients for building the feature matrix. To eliminate the redundant features, a feature selection process is performed using distinguishing capability values of features counting by t-test statistics and thresholding. Jadoon et al. [14] present a mammogram classification system consisting of discrete curvelet transform and multilayer perceptron algorithm. First of all, the images are filtered, and then the curvelet transform is applied to the filtered images. The dense scale invariant feature transform (DSIFT) is applied to the curvelet transform sub-bands to obtain scale-invariant features. Classification is carried out by using multilayer perceptron classifier. Chen et al. [15] present a system that uses wavelet energy entropy and linear regression classifiers to classify mammographic images for breast cancer detection. In the method, the images are segmented, and the regions of interests (ROIs) are obtained. Wavelet energy entropy is calculated applying the wavelet transform to the ROIs. Wavelet energy entropy values are used as features, and the classification process is performed with linear regression classifier. Eltoukhy et al. [16] compare wavelet and curvelet transform to classify the mammograms. Wavelet and curvelet transforms are applied to the images individually, and the feature matrixes are formed by using their coefficients. The features are listed depending on their ability of discriminating the classes using the t-test statistics. Eltoukhy et al. [17], in another study, also use wavelet and curvelet transform for feature extraction comparatively. The features are also sorted according to their ability to distinguish the classes, and then thresholding is applied to

Nebi GEDİK, is with Institute of Health Sciences, University of Health Sciences, Istanbul, Turkey, (e-mail [nebi.gedik@sbu.edu.tr](mailto:nebi.gedik@sbu.edu.tr)).

 <https://orcid.org/0000-0002-1560-1058>

Manuscript received April 24, 2020; accepted January 25, 2020.

DOI: [10.17694/bajece.557693](https://doi.org/10.17694/bajece.557693)



#### IV. PROPOSED METHOD

Initially, mammograms are decomposed using CTsFL transform and the feature matrix (MxN) is generated from the CTsFL coefficients. Where M (lines) corresponds to every image and N (columns) correspond to CTsFL coefficients (features).

The proposed method aims to identify those which have the best classification performance among the whole feature set. The flow chart of the proposed method is indicated in the Figure 2 step by step. In order to accomplish the objective, t-test statistics by using Eq. (7) [23] is applied to the feature set, and coherence coefficients corresponding to each feature (each column in feature matrix (N)) are obtained. Finally, the coherence coefficients (cc) vector is created.

$$cc = \frac{|\mu_1 - \mu_2|}{\sqrt{\frac{(\delta_1)^2}{N_1} + \frac{(\delta_2)^2}{N_2}}} \quad (7)$$

where cc is the coherence coefficient,  $\mu_{1,2}$  and  $\delta_{1,2}$  are the mean and the standard deviations of the classes (subscript 1,2 donates normal-abnormal or benign-malignant), and N is the number of images in each class.

By applying dynamic thresholding over cc values, the feature matrix is reconstructed by new dimension (MxD), and the classification is performed using it to find the most effective feature set, iteratively. The generated new feature set is formed from the values of columns of the initial feature matrix (MxN) correspond to cc values greater than the applied threshold. The operations are repeated for all threshold values, and then the optimum feature set, which has the minimum number of feature with maximum accuracy, is determined. To evaluate the optimum feature set, classification is carried out again using 5-fold cross-validation with those features. All classification processes are performed using support vector machine (SVM).

#### V. EXPERIMENTAL RESULTS

In the evaluation of the proposed method, mammograms obtained from MIAS database [24] are used. Totally 322 images, which consist of 207 normal, 64 benign and 51 malignant, obtained from 161 patients are examined and labeled by expert radiologists. The images are originally 1024x1024 pixels size and about 50% of an image is background and noise. In this paper, those unwanted parts of the images are eliminated by using manual cropping and the ROI set is created. ROIs are 128x128 pixels size, and in case of cropping of abnormal images, the center of ROIs are the center of abnormality. For normal images, the process is randomly made. The ROI set is composed of 228 images including 114 abnormal and 114 normal. In the experiment process, the ROI set is divided into two sets; training (70%) and testing (30%). The training set is used to train the classifier (SVM) and the system performance is evaluated by using the test set. These are illustrated in Figure 2. The image distribution of ROIs is shown in Table 1.

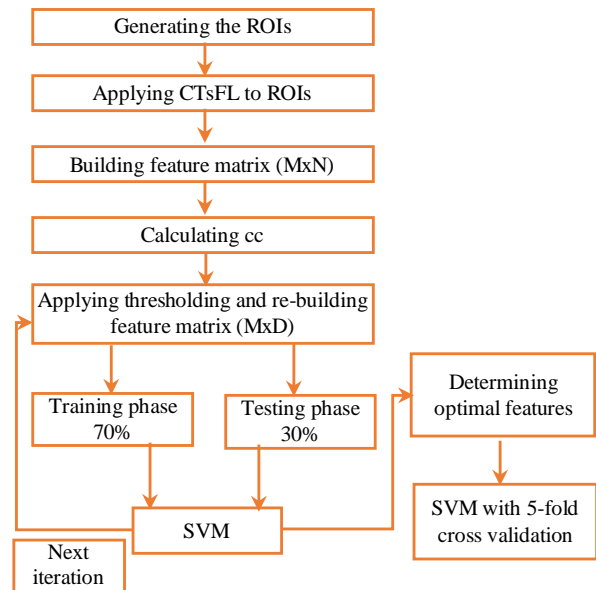


Fig. 2. The flow chart of the proposed method.

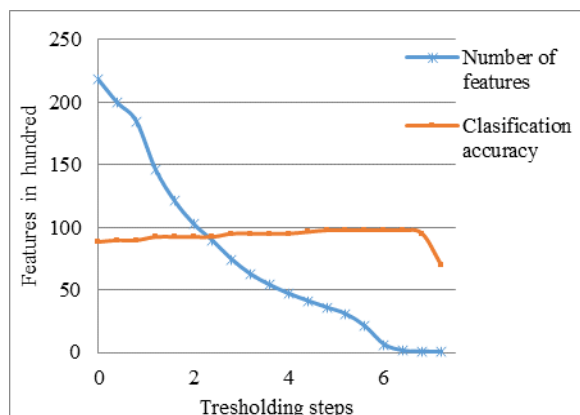
TABLE I  
NUMERICAL DISTRIBUTION OF ROIs.

	Benign	Malignant	Total
Abnormal	64	50	114
Normal	—	—	114
Total			228

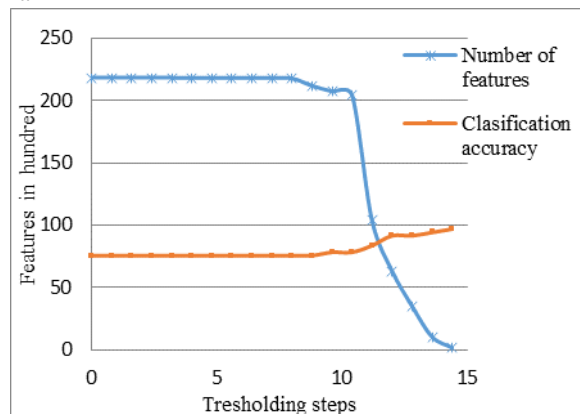
Classification is performed in two-phases; normal-abnormal separation and benign-malignant separation. Figure 3a and 3b illustrate results of the normal-abnormal and benign-malignant classification according to the number of features and the thresholding steps. From the classification results made for all threshold values, maximum accuracies are 97.36% and 97.29% for normal-abnormal and benign-malignant classification respectively. For the maximum accuracies points, while the number of features in the normal-abnormal classification reduces from 21842 to 129, it reduces from 21842 to 175 for benign-malignant classification. Because the goal is to get the maximum accuracy result with the minimum number of features, those points are selected as the optimum features point. To evaluate the result according to the optimum points, classification is performed using 5-fold cross-validation. Figure 4 presents the average value of normal-abnormal and benign-malignant classification results via 5-fold cross validation. Standard deviation values of different classification results obtained from different folds are shown with an error bar above the average classification accuracy bar.

Comparison of the proposed method with previous studies is shown in Table 2. The table contains closely related studies and contains the following comparisons: classification accuracy, number of feature data ensuring the best classification accuracy, feature extraction method, database and classifier. With a general evaluation on the accuracy values in the table, it is observed that the results are close to each other except for [17]. If the normal-abnormal and benign-malignant classification is

evaluated separately, the presented study has the highest accuracy value for the normal-abnormal classification. For benign-malignant classification, the results are almost the same value, except for [17] and for [13] with 98.8% accuracy value. If an evaluation is made in terms of the size of the feature data that used, because all the studies aim to obtain the most successful and the smallest size of feature data from the large feature data, the presented study shows the most successful result.



a



b

Fig. 3. Classifications results a) normal-abnormal separation b) benign-malignant separation.

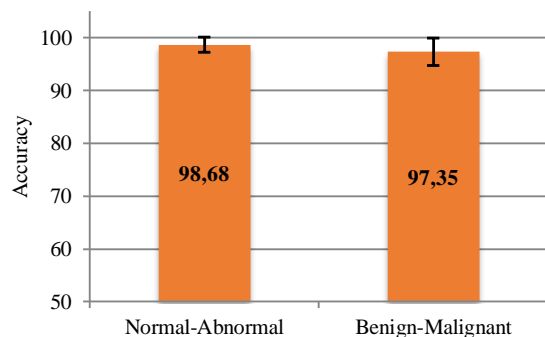


Fig. 4. Average classification results obtained from 5-fold cross validation.

## VI. CONCLUSION

In the present study, a CAD system is represented via a new feature extraction approach to classify mammogram images, and its implementation is carried out. The proposed new feature extraction approach includes the contourlet transform and the t-test statistic. The main purpose is to purify the features that make the most contribution to a successful classification by eliminating redundant ones. Success criterion is to achieve maximum classification success by using the minimum number of features among the all obtained features. According to the experimental results, maximum classification rate is 98.68% for normal-abnormal classification using 129 features among 21842 features, while maximum classification rate is 97.35% for benign-malignant classification using 175 features among 21842 features. Considering the results, the method has an acceptable success to classify mammogram images.

## REFERENCES

- [1] M.D. Chin, K.K. Evans, J.M. Wolfe, J. Bowen, J.W. Tanaka, "Inversion effects in the expert classification of mammograms and faces", *Cognitive Research: Principles and Implications*, vol. 3, 2018, pp. 31.
- [2] Y. Wang, H. Shi, S. M., "A new approach to the detection of lesions in mammography using fuzzy clustering", *J. Int. Med. Res.* vol. 39, no. 6, 2011, pp. 2256–2263.
- [3] N.J. Massat, A. Dibden, D. Parmar, J. Cuzick, P.D. Sasieni, S.W. Duffy, "Impact of screening on breast cancer mortality: the UK program 20 years on", *Cancer Epidemiology and Prevention Biomarkers*, vol. 25, no. 3, 2016, pp. 455-62.
- [4] T. Onega, L.E. Goldman, R.L. Walker, D.L. Miglioretti, D.S. Buist, S. Taplin, B.M. Geller, D.A. Hill, R. Smith-Bindman, "Facility mammography volume in relation to breast cancer screening outcomes", *J. Med. Screen.* vol. 23, 2016, pp. 31.
- [5] M.M. Pawar, S.N. Talbar, "Genetic fuzzy system (GFS) based wavelet co-occurrence feature selection in mammogram classification for breast cancer diagnosis" *Perspectives in Science*, vol.8, 2016, pp. 247–250.
- [6] L. Berlin, "Radiologic errors, past, present and future", *Diagnosis*, vol. 1, no. 1, 2014, pp. 79–84.
- [7] Y. Li, H. Chen, Y. Yang, L. Cheng, L. Cao, "A bilateral analysis scheme for false positive reduction in mammogram mass detection", *Computers in Biology and Medicine*, vol. 57, 2015, pp. 84–95.
- [8] N. Gedik, A. Atasoy, "Performance evaluation of the wave atom algorithm to classify mammographic images", *Turk. J. Elec. Eng. & Comp. Sci.*, vol.22, 2014, pp. 957–969.
- [9] V. Chaurasia, S. Pal, "A novel approach for breast cancer detection using data mining techniques", *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, no. 1, 2014, pp. 1-17.
- [10] L. Dora, S. Agrawal, R. Panda, A. Abraham, "Optimal breast cancer classification using Gauss–Newton representation based algorithm", *Expert Systems with Applications*, vol. 85, 2017, pp. 134-145.
- [11] N. Gedik, "Breast cancer diagnosis system via contourlet transform with sharp frequency localization and LS-SVM", *Journal of medical imaging and health informatics*, vol. 5, 2015, pp. 1–9.
- [12] W. Yang, L. Tianhui, "A Robust Feature Vector Based on Waveatom Transform for Mammographic Mass Detection," *ICVR 2018 Proceedings of the 4th International Conference on Virtual Reality*, Hong Kong, pp.133-139, 24-26 February 2018.
- [13] N. Gedik, "A new feature extraction method based on multi-resolution representations of mammograms", *Applied Soft Computing*, vol. 44, no. 1, 2016, pp. 128-133.
- [14] M.M. Jadoon, Q. Zhang, I.U. Haq, A. Jadoon, A. Basit, S. Butt, "Classification of mammograms for breast cancer detection based on curvelet transform and multi-layer perceptron", *Biomedical Research*, vol. 28, no. 10, 2017, pp. 4311-4315.

TABLE II  
THE COMPARISON OF THE PROPOSED METHOD WITH PREVIOUS STUDIES

	Database	ROIs	Method	Number of features	Classification	Accuracy	Classifier
Eltoukhy et al. [16]	MIAS	322 images (207 normal, 115 abnormal, 128x128 pixel)	Curvelet	5663	N-A	95,98%	SVM
				333	B-M	97,30%	
			Wavelet	1238	N-A	95,84%	
				150	B-M	96,56%	
Gedik [13]	MIAS	228 images (114 normal, 114 abnormal, 128x128 pixel)	FFST	8776	N-A	98,29%	SVM
				66	B-M	97,39%	
	DDSM	228 images (114 normal, 114 abnormal, 128x128 pixel)	FFST	35857	N-A	96,89%	
				272	B-M	98,08%	
Eltoukhy et al. [17]	MIAS	322 images (207 normal, 115 abnormal, 128x128 pixel)	Curvelet	not given	N-A	91,19%	SVM
				not given	B-M	91,18%	
			Wavelet	not given	N-A	89,51%	
				not given	B-M	89,58%	
Proposed method	MIAS	228 images (114 normal, 114 abnormal, 128x128 pixel)	CTsFL	129	N-A	98,68%	SVM
				175	B-M	97,35%	

N-A: Normal-Abnormal, B-M: Benign-Malignant

- [15] Y. Chen, Y. Zhang, H.M. Lu, X.Q. Chen, J.W. Li, S.H. Wang, "Wavelet energy entropy and linear regression classifier for detecting abnormal breasts", *Multimed Tools Appl.*, vol. 77, 2018, pp. 3813–3832.
- [16] M.M. Eltoukhy, I. Faye, B.B. Samir, "A statistical based feature extraction method for breast cancer diagnosis in digital mammogram using multiresolution representation", *Computers in biology and medicine*, vol. 42, no. 1, 2012, pp. 123–128.
- [17] M.M. Eltoukhy, I. Faye, "An optimized feature selection method for breast cancer diagnosis in digital mammogram using multiresolution representation", *Applied Mathematics and Information Sciences*, vol. 8, no. 6, 2014, pp. 2921-2928.
- [18] D. Sehrawat, A. Sehrawat, D. Jaiswal, A. Sen, "Detection and classification of tumor in mammograms using discrete wavelet transform and support vector machine", *International Research Journal of Engineering and Technology (IRJET)*, vol. 4, no. 5, 2017, pp. 1328-1334.
- [19] Y. Lu, M.N. Do, "A new contourlet transform with sharp frequency localization", *IEEE 2006 International Conference on Image Processing*, Atlanta, Georgia, U.S.A., pp.1629-1632, 8-11 October 2006.
- [20] V. Vapnik, "The Nature of Statistical Learning Theory", Springer-Verlag New York, 1995.
- [21] V. Vapnik, "Statistical Learning Theory", New York: Wiley, 1998.
- [22] V. Vapnik, E. Levin, Y.L. Cun, "Measuring the VC-dimension of a Learning Machine", *Neural Computation*, 6(5), MIT Press, Cambridge, 1994.
- [23] H. Liu, J. Li, L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns", *Genome Inf.*, vol. 13, no. 1, 2002, pp. 51–60.
- [24] <http://peipa.essex.ac.uk/info/mias.html> (20.11.2018)

## BIOGRAPHIES



Nebi Gedik received his B.S. degree in Electrical and Electronics Engineering from Firat University in 2001, his PhD degrees in Electrical and Electronics Engineering from Karadeniz Technical University in 2013, and his MSc degree in 2005 from Atatürk University. He is now an Assistant Professor at University of Health Science. His research interest includes medical image and signal processing, pattern recognition and machine learning.