VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

Breast Cancer Diagnosis using Geometrical Descriptors Obtained from Adaptive Convex Hulls of Suspicious Regions

İdil IŞIKLI ESENER[1]*, Şükriye KARA[2], Semih ERGİN[3]

[1]*Bilecik Seyh Edebali University, Faculty of Engineering, Department of Electrical and Electronics Engineering, Bilecik, TURKEY*
[2]*Eskiehir Osmangazi University, Faculty of Engineering and Architecture, Department of Electrical and Electronics Engineering, Eskisehir, TURKEY*

Makale Bilgisi

Başvuru:
24/07/2021
Kabul:
17/10/2021

Abstract

Breast cancer is the leading cause of cancer-related deaths among women and also the most frequently diagnosed cancer type. Advances in technology introduce Computer-Aided Diagnosis (CAD) systems for breast cancer which gain importance in reduced mortality rate by the increased sensitivity of early diagnosis. Although any imaging technique can be adapted, mammography, with known effectiveness in early diagnosis, are mostly used in CAD systems for breast cancer. Hence, this paper focuses on design of a CAD system analyzing mammography images for breast cancer diagnosis and proposes a new approach for geometrical feature extraction for this system. The proposed scheme is verified on a subset of the publicly available Mammographic Image Analysis Society digital mammogram database. In the detection phase of this system, initially, adaptive median filtering is applied for noise reduction; artifact suppression and background removal is realized via morphological operations, and pectoral muscle removal is executed using a region growing algorithm. Then, Chan-Vese active contour modeling is utilized for the ROI detection. Thereupon, the center of gravity (CoG) of each ROI is determined, and a convex image is created by specifying 92 points, called as edge points, on the boundary curves of the related ROI. In the feature extraction stage of the diagnosis phase, the angles between each pair of edge points and the CoG, the Euclidean distance between edge points and the CoG, and the Euclidean distance between each pair of edge points are computed. These geometrical descriptors are utilized in the classification stage via the Random Forest classifier using the five-fold cross-validation technique. As a result, breast cancer diagnosis is achieved by an accuracy of 70.13%. Analyzing the overall confusion matrix constructed in the classification stage, it is clearly seen that although healthy and benign diagnoses are mixed, malignancy is diagnosed well by the proposed geometrical descriptors.

Keywords: *Digital Mammography, Computer-Aided Diagnosis, Feature Extraction, Geometric Descriptor*

* İletişim e-posta: idil.isikli@bilecik.edu.tr

1 Introduction

Breast cancer is indicated as it leads the cancer diagnosis and cancer-related deaths among women worldwide by the world cancer statistics [1]. Early diagnosis of breast cancer is currently the most effective strategy to reduce the related mortality rate while it can applicably be achieved by mammography [2]. However, the evaluation of mammography is human-dependent and is getting difficult as the similarities of normal and abnormal tissues within a breast increase due to the breast tissue type. Computer-Aided Diagnosis (CAD) systems for breast cancer diagnosis gain importance from this point of view as they increase the sensitivity of diagnosis by giving radiologists the opportunity of re-evaluation [3].

A typical CAD system has two major phases named as detection and diagnosis. The detection phase involves the segmentation of suspicious regions (Region-of-Interest – ROI) for breast cancer after a pre-processing stage where noise reduction, artifact suppression, background removal, and pectoral muscle removal are performed. In the diagnosis phase, descriptive features are extracted from ROIs, and health status classification is realized by using these features.

The abnormalities present in a mammography can either be benign (non-cancerous) or malignant (cancerous). The Breast Imaging Reporting and Data System (BI-RADS) states that various descriptors like shape, size, and margins can be used to characterize these abnormalities [4]. This characterization may be formulated as textural, geometrical, or both in CAD systems. Especially, geometrical analysis of ROIs seems to be a powerful tool for discriminating the abnormalities due to different morphologies of benign and malignant tumors [5]. Several methodologies are examined for geometrical shape analysis in the literature whether examined in breast cancer diagnosis [5] or not [6-10].

Inspired by the Edge Step (ES) approach introduced in [10], an Adaptive Convex Hull (ACH) approach is proposed for geometrical feature extraction in this paper. The proposed approach is verified via a publicly available Mammographic Image Analysis Society (MIAS) digital mammogram database [11] on a pre-designed CAD system [12, 13] for breast cancer diagnosis. The extracted features are then utilized for ROI diagnosis using 5-fold cross-validation technique via Random Forest classifier. In conclusion, an average accuracy of 70.13% is achieved for overall diagnosis while malignancy is diagnosed by an average accuracy of 90.43% following the proposed scheme.

This paper is organized as follows: In the following section, the database used in this paper is introduced and the methods used for each stage in both detection and diagnosis phases are explained. The experimental results are stated and discussed in Section 3 and the main conclusions are given in Section 4.

2 Materials and methods

2.1 Database

The proposed geometrical feature extraction scheme is substantiated on the MIAS database that has 322 Mediolateral Oblique (MLO) – view digital mammography images, with 330 diagnoses as 207 normal, 69 benign cancers and 54 malignant cancers, belong to 161 objects having fatty, fatty-glandular, and dense tissue types [11]. The images are at a size of 1024 × 1024 with a resolution of 8 bits/pixel in in ". pgm" imaging format, and ground truth information of each diagnose is also presented in the database. Sample mammography images of each breast tissue type of each health status is shown in Figure 1.

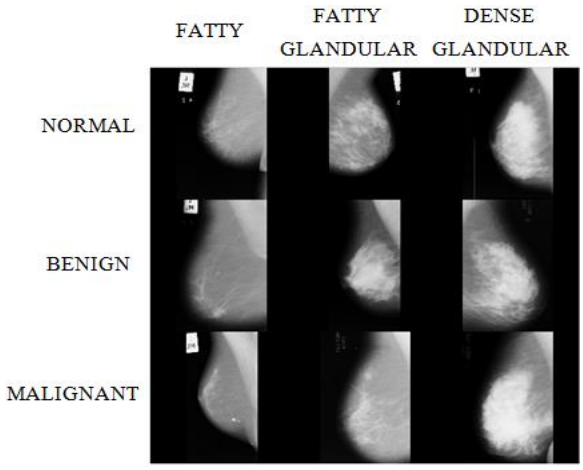


Figure 1. Sample mammography images from the MIAS database [12]

2.2 Detection phase of the CAD system

A typical CAD system has two major phases named as detection and diagnosis. The detection phase involves the segmentation of suspicious regions (Region-of-Interest – ROI) for breast cancer after a pre-processing stage where noise reduction, artifact suppression, background removal, and pectoral muscle removal are performed.

2.2.1 Pre-processing

All mammography images in the MIAS database are initially resized to 256×256 using bi-cubic interpolation.

A MLO-view mammography image consists of not only the breast parenchyma but also artifacts, pectoral muscle, and background as shown in Figure 2. Besides, as in all pattern recognition studies, any type of noise exists on digital images and should be reduced to increase the reliability of both segmentation and feature extraction stages of a CAD system. Accordingly, adaptive median filter is applied to the images for digitization noise reduction while preserving gross details such as pectoral muscle and ROI edges.

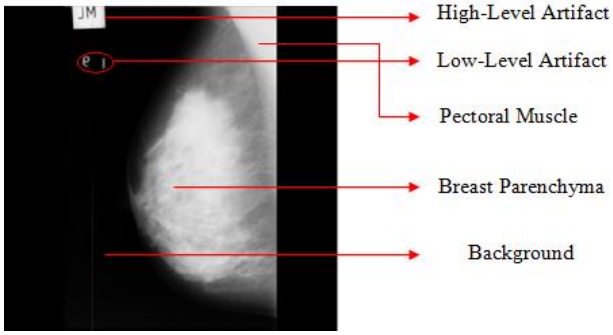


Figure 2. A sample MLO-view mammography image in the MIAS database [12]

The presence of artifacts and pectoral muscles in mammography images plays an obstructor role since they have similar characteristics on intensity with ROIs. That being the case, noise reduction is followed by some morphological operations for artifact suppression, background removal, and left-alignment of the images while a region growing algorithm is performed for pectoral muscle removal [12]. The phases realized for the pre-processing stage are visualized in Figure 3 on a sample mammography image.

2.2.2 Segmentation of suspicious regions

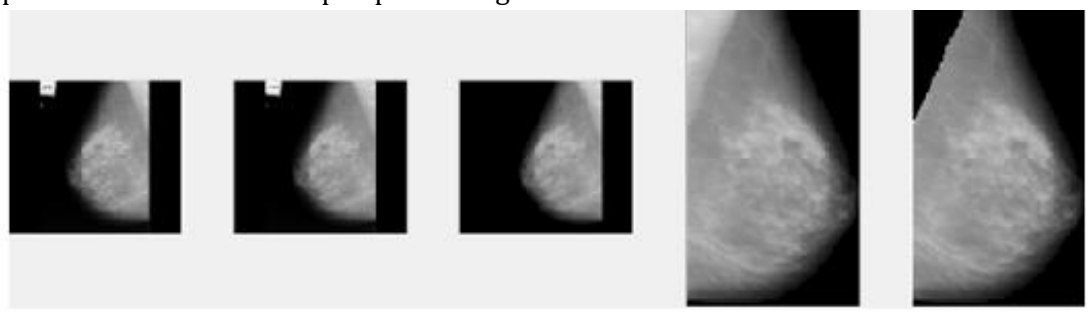
The Chan-Vese active contour modelling [14] is used iteratively by manual-initial-contour-selection for suspicious region, Region-of-Interest (ROI), detection on a mammography image [13]. The initial-contour is selected as a single point with high possibility to contain any abnormality. Then, the Chan-Vese active contour modelling is run at most two iterations according to a defined decision criterion, and connected component analysis followed by morphological operations are applied on the segmented regions for irrelevant pixel removal and enhancement, respectively [13]. Figure 4 visualizes the mentioned phases for ROI detection.

2.3 Diagnosis phase of the CAD system

In the diagnosis phase, descriptive features are extracted from ROIs, and health status classification is realized by using these features.

2.3.1 Feature extraction

Inspired by the Edge Step (ES) approach introduced by Türkoğlu and Hanbay [10], this paper proposes an Adaptive Convex Hull (ACH) approach for geometrical feature extraction. The ES approach defines the contour of a ROI based on selecting n points $(k_1, k_2, \cdots, k_n)$ with a fixed-step-size of pixel lengths on its boundary curve. Then, by using these n points and the center-of-gravity (CoG) of ROI, three parameters are computed which are the angle between each pair of points and the CoG $(\alpha_1, \alpha_2, \cdots, \alpha_n)$, edge-distance between each pair of points $(p_1, p_2, \cdots, p_n)$, and the edge-CoG distance of each point $(d_1, d_2, \cdots, d_n)$.



(a) (b) (c) (d) (e)

Figure 3. The phases for the pre-processing stage of the proposed system: (a) The original image; (b) The noise-reduced image; (c) The artifacts-suppressed image; (d) The background-removed and left-aligned image; (e) The pectoral muscle-removed image [13]

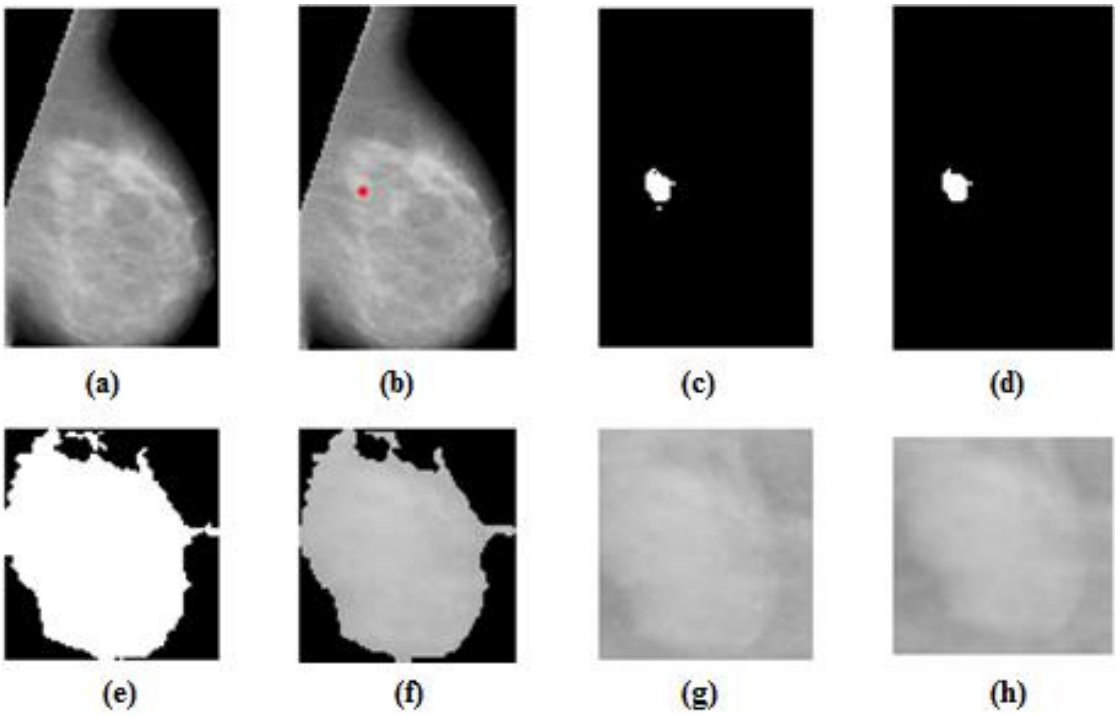


Figure 4. The phases for the adopted ROI detection system: (a) The pre-processed image; (b) The user-defined initial seed point (red) for the Chan-Vese algorithm; (c) The segmented ROI in the original binary image; (d) The enhanced ROI in the original binary image; (e) The binary ROI image; (f) The gray-scaled ROI image; (g) The ultimate segmented ROI region; (h) The reference ROI region [13]

These parameters are visualized in Figure 5. Türkoğlu and Hanbay, then computes some statistical formulas on these parameters for feature vector construction [10].

The ACH approach selects a fixed number of n points for all ROIs, resulting in different step-sizes for each ROI, contrarily to the ES approach. The same parameters which are the angles (α_n), the edge-to-CoG distances (d_n), and the sequential edge distances (p_n), are then computed. The geometrical feature vectors are constructed by concatenating these parameters themselves and $n \times 3$–dimensional features are achieved.

In this paper, 92 points are selected on each ROI and a 276×1–dimensional feature vector is obtained for each ROI.

2.3.2 Classification

In this paper, a selected subset of 110 ROIs, detected on the mammography images in the MIAS database, having 52 normal, 39 benign, and 19 malignant diagnoses are used for the experimental studies. The geometrical features of these images are utilized via Random Forest classifier with 5-fold cross-validation technique. It means that 95% of each class (49 normal, 37 benign cancers, and 18 malignant cancers) are used for training while the remaining 5% (3 normal, 2 benign cancers, and 1 malignant cancers) are treated as the test parts.

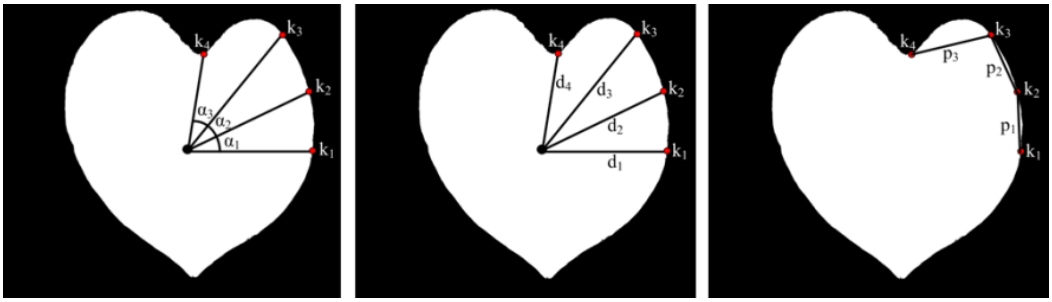


Figure 5. The ES approaches on a sample ROI: (a) angles (α_n); (b) edge-to-CoG distances (d_n); (c) sequential edge distances (p_n) [10]

2.3.3 Performance Evaluation

The sensitivity (SNS), specificity (SPC), positive predictive value (PPV), negative predictive value (NPV), false-positive rate (FPR), false negative rate (FNR), false discovery rate (FDR), false omission rate (FOR), and accuracy (ACC) metrics are used for the performance evaluation of the proposed feature extraction scheme. These metrics are computed as given in Table 1 for each fold in the cross-validation technique, and averages of each are evaluated.

Table 1. Performance evaluation metrics and their mathematical representations.

Performance Evaluation Metrics	Mathematical Representations
TP: True Positive TN: True Negative FP: False Positive FN: False Negative	
SNS	$\%\ SNS = \frac{TP}{TP + FN} \cdot 100$
SPC	$\%\ SPC = \frac{TN}{TN + FP} \cdot 100$
PPV	$\%\ PPV = \frac{TP}{TP + FP} \cdot 100$
NPV	$\%\ NPV = \frac{TN}{TN + FN} \cdot 100$
FPR	$\%\ FPR = \frac{FP}{FP + TN} \cdot 100$
FNR	$\%\ FNR = \frac{FN}{TP + FN} \cdot 100$
FDR	$\%\ FDR = \frac{FP}{TP + FP} \cdot 100$
FOR	$\%\ FOR = \frac{FN}{TN + FN} \cdot 100$
ACC	$\%\ ACC = \frac{TP + TN}{TP + TN + FP + FN} \cdot 100$

3 Results and discussion

In this paper, diagnosis of breast cancer is realized based on the geometrical structures of ROIs defined on their convex hulls. The average performance evaluation metrics succeeded for overall diagnosis using Random Forest classifier are shown in Figure 6.

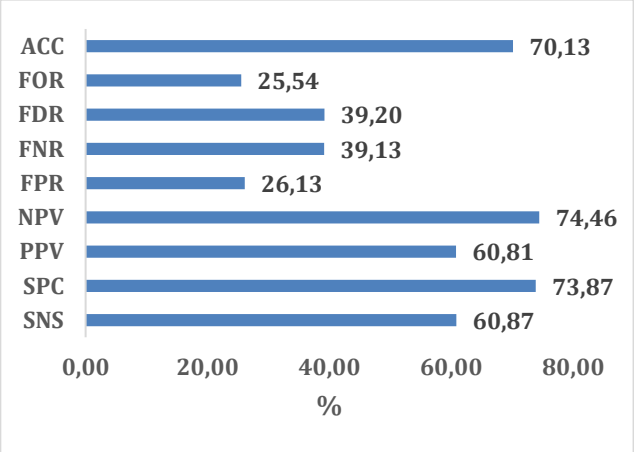


Figure 6. The average performance evaluation metrics for overall diagnosis

Although an overall accuracy of 70.13% is achieved, this study is concluded by a relatively lower sensitivity rate. Therefore, the performance evaluation metrics are analyzed for diagnosis of each class individually, in Figures 7-9.

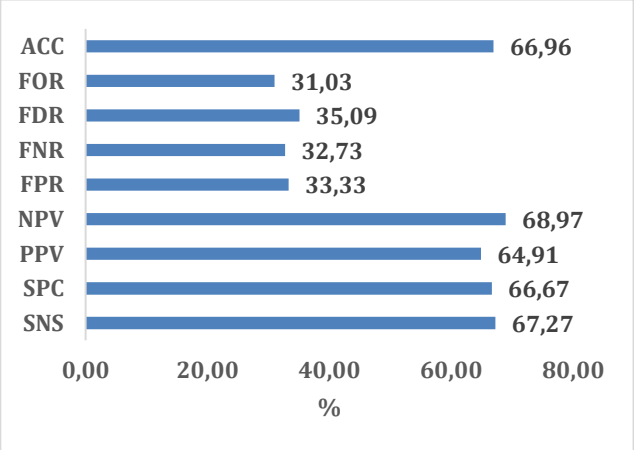


Figure 7. The average performance evaluation metrics for normal (healthy) diagnosis

The high FNRs in Figures 7 and 8 show the excessive number of healthy ROIs diagnosed as having abnormality, and ROIs having benignity although they are diagnosed as non-benign, respectively. This consequence is concluded with large FPRs and thus low ACC in Figures 7 and 8.

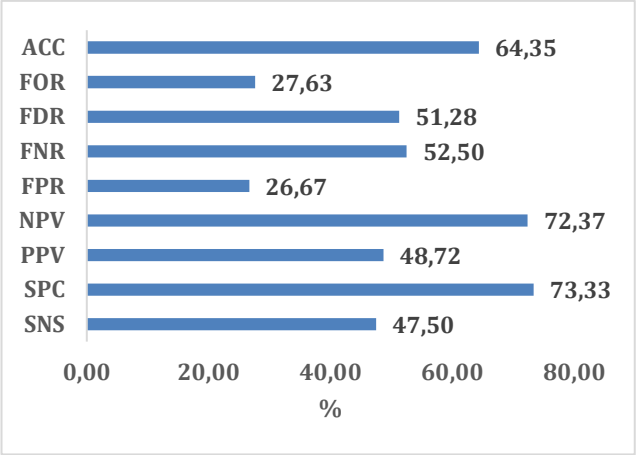


Figure 8. The average performance evaluation metrics for benign diagnosis

In case of diagnosing malignancy, the superiority of ACC, FPR, NPV, and SPC in Figure 9, shows the success of ACH vectors as a malignancy marker.

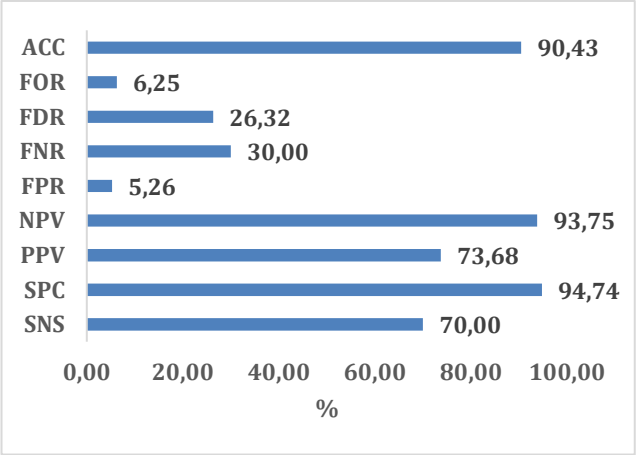


Figure 9. The average performance evaluation metrics for malignant diagnosis

Analyzing the overall confusion matrix given in Table 2, it is clearly seen that although healthy and benign diagnoses are mixed, malignancy is diagnosed well by the proposed geometrical descriptors.

Table 2. Overall confusion matrix constructed in diagnosis stage.

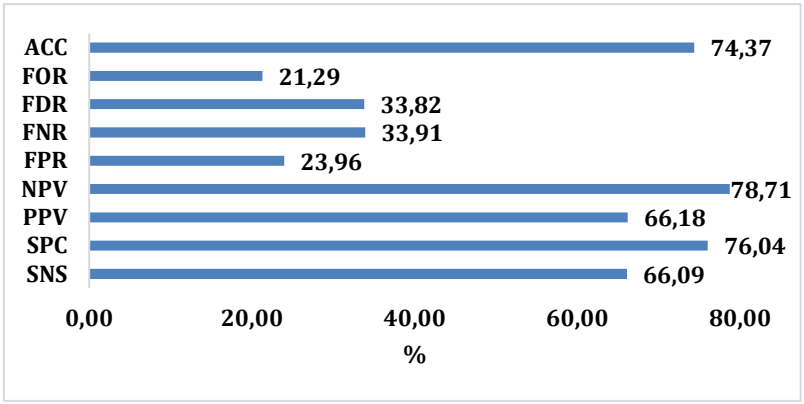
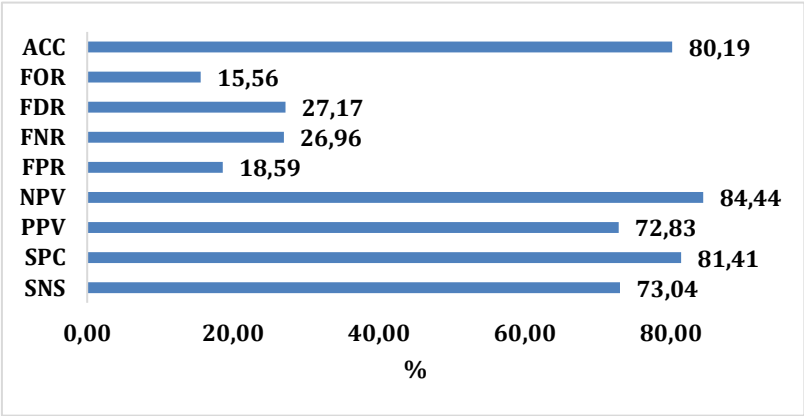
		Classified to		
		Normal	**Benign**	**Malignant**
Ground Truth	**Normal**	37	17	1
	Benign	17	19	4
	Malignant	3	3	14

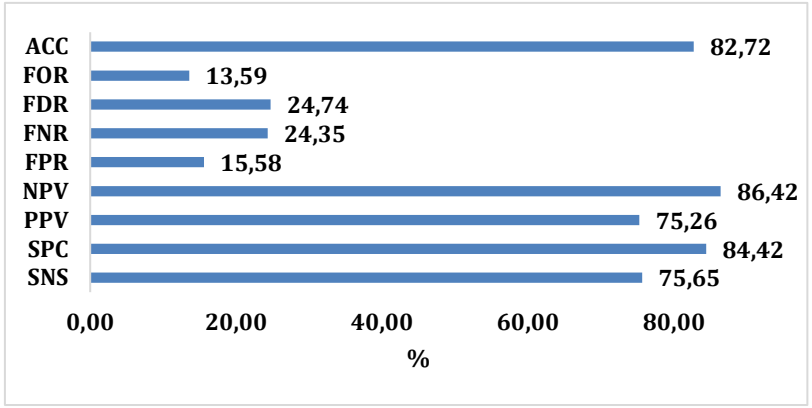
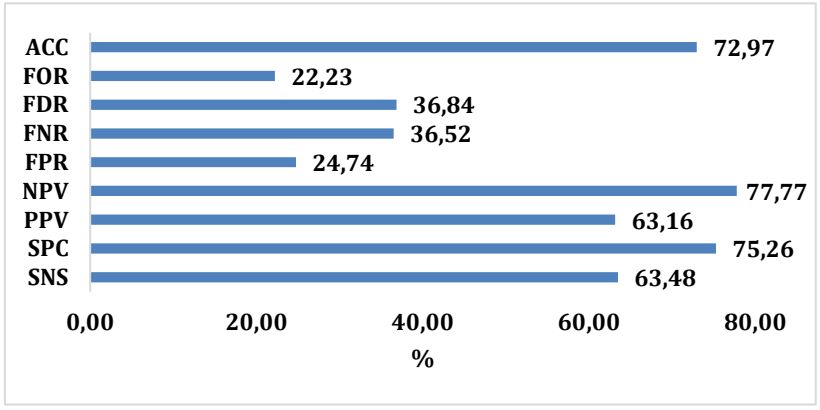
Besides, a benchmark study is executed on the same mammography images, by running the CAD system in this paper, in order to evaluate the performance of the proposed ACH approach. This study consists of four diagnosis experiments using four different feature sets constructed using state-of-the-art techniques in the related literature. The first feature set is the Local Configuration Pattern (LCP) vectors obtained by applying LCP algorithm on wavelet-decomposed ROIs [3], and named as LCP features in this paper. The second feature set is the LCP-based features [3] which are constructed by concatenating the LCP features, some statistical measurements of these features, and the energies of four-level wavelet-decomposed sub-bands of each ROI. The third feature set, namely statistical features, is calculated by the same statistical measurements [3] computed directly from ROIs rather than LCP features. The fourth set is the Haralick features extracted on gray-level co-occurrence matrices of each ROI [15].

Table 3 presents the comparative results of diagnosis using these four feature sets with the proposed geometrical descriptors. Although the exiguousness with respect to performance evaluation metrics is obtained using the proposed descriptors, an analysis on the confusion matrices shows the mastery of the ACH approach on malignancy diagnosis, especially.

The weakness of the proposed descriptors is due to the confusion of normal and benign ROIs. The confusion matrices in Table 3 clearly indicate that there is no different tendency even if the other feature extraction techniques are preferred.

Table 3. The comparison between the proposed and previously extracted features according to different success metrics.

<table>
<tr><th>No</th><th>Feature</th><th>Performance Evaluation Metrics for Overall Diagnosis</th><th>Confusion Matrix</th></tr>
<tr><td>1</td><td>LCP Features [3]</td><td> 
</td><td>
<table>
<tr><th></th><th></th><th colspan="3">Classified to</th></tr>
<tr><th></th><th></th><th>Normal</th><th>Benign</th><th>Malignant</th></tr>
<tr><td rowspan="3">Ground Truth</td><td>Normal</td><td>42</td><td>12</td><td>1</td></tr>
<tr><td>Benign</td><td>15</td><td>22</td><td>3</td></tr>
<tr><td>Malignant</td><td>6</td><td>2</td><td>12</td></tr>
</table>
</td></tr>
<tr><td>2</td><td>LCP-based Features [3]</td><td> 
</td><td>
<table>
<tr><th></th><th></th><th colspan="3">Classified to</th></tr>
<tr><th></th><th></th><th>Normal</th><th>Benign</th><th>Malignant</th></tr>
<tr><td rowspan="3">Ground Truth</td><td>Normal</td><td>47</td><td>7</td><td>1</td></tr>
<tr><td>Benign</td><td>13</td><td>24</td><td>3</td></tr>
<tr><td>Malignant</td><td>3</td><td>4</td><td>13</td></tr>
</table>
</td></tr>
</table>

3	**Statistical Features [3]**		
4	**Haralick Features [15]**		

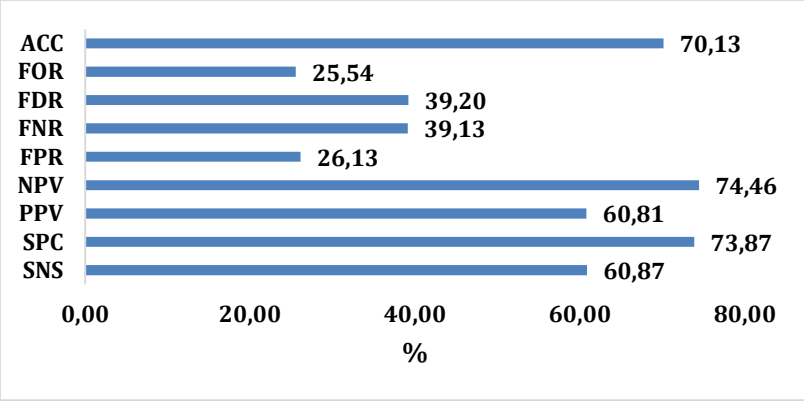
Classified to

		Normal	**Benign**	**Malignant**
Ground Truth	**Normal**	48	6	1
	Benign	10	26	4
	Malignant	2	5	13

Classified to

		Normal	**Benign**	**Malignant**
Ground Truth	**Normal**	42	12	1
	Benign	16	20	4
	Malignant	4	5	11

5 **Geometrical Descriptors (proposed)**



		Classified to		
		Normal	**Benign**	**Malignant**
Ground Truth	**Normal**	37	17	1
	Benign	17	19	4
	Malignant	3	3	14

4 Conclusions

A feature extraction scheme based on the geometrical structure of ROIs is proposed for a CAD system for early diagnosis of breast cancer, in this paper. The geometrical structure of ROIs are identified by the ACH approach and three features, the angle between each pair of points and the CoG $(\alpha_1, \alpha_2, \cdots, \alpha_n)$, edge-distance between each pair of points $(p_1, p_2, \cdots, p_n)$, and the edge-CoG distance of each point $(d_1, d_2, \cdots, d_n)$, for each of 92-defined edge points on the convex hull of each ROI are computed. The 276×1–dimensional ACH feature vector of each ROI is then constructed by concatenating the above three features of each ROI. These feature vectors are verified on a selected subset of 110 ROIs having 52 normal, 39 benign, and 19 malignant diagnoses via Random Forest classifier with 5-fold cross-validation technique. As a result, breast cancer diagnosis is achieved by an accuracy of 70.13%. Analyzing the overall confusion matrix constructed in the classification stage, it is clearly seen that although healthy and benign diagnoses are mixed, malignancy is diagnosed well by the proposed geometrical descriptors.

References

[1] International Agency for Research on Cancer. "Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018". World Health Organizatio5n, Lyon, France, 263, 2018.

[2] Ergin S, Kılınç O. "A new feature extraction framework based on wavelets for breast cancer diagnosis". *Computers in Biology and Medicine*, 51, 171-182, 2014.

[3] Işıklı Esener İ, Ergin S, Yüksel T. "A new feature ensemble with a multistage classification scheme for breast cancer diagnosis". *Journal of Healthcare Engineering*, 2017, 1-15, 2017.

[4] D'Orsi CJ, Sickles EA, Mendelson EB, Morris EA, et al. ACR BIRADS® Atlas, Breast Imaging Reporting and Data System. American College of Radiology [internet]. 2013 [cited 2018 Dec 06].

[5] Souza JC, Silva TF, Rocha SV, Paiva AC, Braz G, Almeida JD, Silva AC. "Classification of malignant and benign tissues in mammography using dental shape descriptors and shape distribution". *2017 7th Latin American Conference on Networked and Electronic Media (LACNEM 2017)*, Valparaiso, Chile, 6-7 November 2017.

[6] Osada R, Funkhouser T, Chazelle B, Dobkin D. "Shape distributions". *ACM Trans Graph*, 21(4), 807-832, 2002.

[7] Yu M, Atmosukarto I, Leow WK, Huang Z, Xu R. "3D model retrieval with morphing based geometric and topologic topological feature maps". *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison*, WI, USA, 18-20 June 2003.

[8] Wu SG, Bao FS, Xu EY, Wang YX, Chang YF, Xiang QL. "A leaf recognition algorithm for plant classification using probabilistic neural network". *2007 IEEE International Symposium on Signal Processing and Information Technology*, Giza, Egypt, 15-18 December 2007.

[9] Mahdikhanlou K, Ebrahimnezhad H. "Plant leaf classification using centroid distance and axis of least inertia method". *2014 22nd Iranian Conference on Electrical Engineering (ICEE),* Tehran, Iran, 20-22 May 2014.

[10] Türkoğlu M, Hanbay D. "Plant recognition system based on extreme learning machine by using shearlet transform and new geometric features". *Journal of The Faculty of Engineering and Architecture of Gazi University*, 34(4), 2097-2112, 2019.

[11] Suckling J et al. "The Mammographic Image Analysis Society Digital Mammogram Database". *Exerpta Medica Int Congr Ser*, 1069, 375-378, 1994.

[12] Işıklı Esener İ, Ergin S, Yüksel T. "A novel multistage system for the detection and removal of pectoral muscles in mammograms". *Turkish Journal of Electrical Engineering & Computer Sciences*, 26(1), 35-49, 2018.

[13] Işıklı Esener İ, Ergin S, Yüksel T. "A practical Region-of-Interest (ROI) detection approach for suspicious region identification in breast cancer diagnosis". *2017 International Conference on Engineering Technologies (ICENTE17)*, Konya, Turkey, 7-9 December 2017.

[14] Chan TF, Vese LA. "Active contours without edges". *IEEE Transactions on Image Processing*, 10, 266-277, 2001.

[15] Işıklı Esener İ, Ergin S, Yüksel T. "A genuine GLCM-based feature extraction for breast tissue classification on mammograms". *International Journal of Intelligent Systems and Applications in Engineering*, 4, 124-129, 2016.