ESKİŞEHİR TECHNICAL UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY
A- APPLIED SCIENCES AND ENGINEERING

2022, 23(1), pp. 70-86, DOI: 10.18038/estubtda.906920

A HYBRID TEXTURAL AND GEOMETRICAL FEATURE EXTRACTION TO REVEAL HIDDEN INFORMATION FROM SUSPICIOUS REGIONS ON MAMMOGRAMS

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

[1] Electrical and Electronics Engineering, Faculty of Engineering, Bilecik Seyh Edebali University, Bilecik, Turkey
[2] Electrical and Electronics Engineering, Faculty of Engineering and Architecture, Eskisehir Osmangazi University, Eskiehir, Turkey
[3] Radiology, Faculty of Medicine, Eskisehir Osmangazi University, Eskiehir, Turkey

ABSTRACT

A mammographic feature extraction scheme through textural and geometrical descriptors is examined to implement in a computer-aided diagnosis system for breast cancer diagnosis in this paper. This scheme is verified on a selected subset of suspicious regions (Region of Interest – ROI) detected on a publicly available mammogram image database constructed by the Mammographic Image Analysis Society. The ROI detection is succeeded using the Chan-Vese active contour modelling after some pre-processing operations which are median filtering, morphological operations, and a region-growing method performed for digitization noise reduction, artifact suppression and background removal, and pectoral muscle removal, respectively, applied on mammogram images. Then, a new adaptive convex hull approach is introduced for extracting geometrical descriptors of ROIs accompanied by the Haralick features extracted from the gray-level co-occurrence matrices for textural description. In addition to geometrical and textural features, a hybrid mammographic feature vector is constructed by concatenating these features. All three feature vectors are separately utilized to diagnose ROIs via Random Forest classifier using 5-fold cross-validation. Experimental studies show that the textural features diagnose benignity more specifically and malignancy more accurately; and they are more effective on discriminating healthy ROIs when concatenated with geometrical features. Hence, a feature combination of these three features is proposed for diagnosis. The proposed feature combination is determined to be more effective for more accurate diagnoses of benignity and malignancy.

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

1. INTRODUCTION

Early diagnosis of breast cancer gains more importance since breast cancer is the most common type of cancer in the female populace and the number of related demises is increasing day by day [1]. Hence, breast imaging techniques need to be evaluated in terms of their usability for early diagnosis. Today's technology allows breast imaging via mammography, breast Magnetic Resonance Imaging (MRI), sonography, Computer Tomography (CT), Positron Emission Tomography (PET), and thermography [2, 3]. Although breast MRI can be used as primer imaging for patients having high risk of breast cancer, it is usually performed after a patient is diagnosed with cancer to assess the extent of it [2, 3]. Similarly, CT and PET are used to evaluate metastasis or response to therapy [2]. Thermography has the risk to increase the false negative and false positive measurements due to asymmetric internal body temperature [3]. Sonography, on the other hand, has limitations of low-resolution images and low sensitivity [2, 3]. Mammography, despite holding bounded information because of not much different X-ray permeability of healthy and cancerous cells due to low dose radiation application and not being suitable for dense breast, is a well-known and mostly used radiological imaging technique for early diagnosis of breast cancer diagnosis [2-5]. Besides, re-analysing mammography images via Computer Aided Diagnosis (CAD) systems can give the chance of early diagnosis, despite the lack of information and reduce the breast cancer-caused mortality rate [6-8].

*Corresponding Author: idil.isikli@bilecik.edu.tr
Received: 31.03.2021 Published: 30.03.2022

CAD systems designed for breast cancer diagnosis are responsible for both detection and diagnosis of this disease. This responsibility requires four fundamental stages. Difficulties for the detection phase of the CAD systems, such as low contrast of the images, digitization noise, labelling artifacts, background, and pectoral muscles in whole breast mammograms, are eliminated in the pre-processing stage. The second stage is the detection of suspicious regions, called as ROIs, for breast cancer. The descriptive and discriminative features of the ROIs are then extracted in the third stage, and finally, diagnosis is achieved using these features in the classification stage.

The feature extraction stage of any CAD system directly affects the system performance by the way how data is represented. Besides utilizing statistical features [9-18] for breast cancer diagnosis, use of mammographic features [17, 19-22] like shape, spicule index, contour, size, density, and brightness would be more consistent through a radiologist's evaluation. Although pixel intensity is a sufficient measure for brightness, additional textural information [11, 16, 17, 23-25] is required for density determination as well as geometrical techniques [16, 26-30] should be used for shape and contour definition. In addition, subspace projections are also used for breast cancer diagnosis in the literature [31, 32].

In medical science, radiologists analyse the suspicious regions in a mammography image by their size, density, shape, contour traceability, and accompanying findings. It is known that radiolucent, regular-shaped, and bounded-contoured ROIs sign the benign cancer possibility while radiopacity, irregular-shape and unbounded-contour increase the risk of malignancy [33]. Besides, radiologists state that the breast tissue type directly affects the interpretation above symptoms. Therefore, a radiologist leaded especially for extracting some geometrical descriptors in the earlier phases of this study before some mathematical and geometrical expressions are developed. Under these circumstances, textural, and geometrical feature extraction techniques such as Centroid Distance (CD) approach [29] and Edge Step (ES) approach [30] are thought to be descriptive for health status. Accordingly, feature vectors defining ROIs' textural and geometrical structures are examined via a pre-designed CAD system [34, 35] for breast cancer diagnosis, in this paper. Noise reduction, background removal, artifact suppression, and pectoral muscle removal are carried out [34] and the ROIs are identified [22] in the pre-processing and detection stages of this CAD system, respectively. In the feature extraction stage, an Adaptive Convex Hull (ACH) approach is examined for geometrical feature extraction while textural formation of ROIs is formalized by Haralick features [36-38] computed from Gray-Level Co-Occurrence Matrix (GLCM) of mammogram ROIs. These features are utilized in the classification stage using a 5-fold cross-validation technique via Random Forest classifier.

This paper is organized as follows. The database used in this paper is introduced in the following section where the theoretical explanations on the applied methods for each of above-mentioned stages are also given. The experimental studies including feature vector construction, classification, and performance evaluation are described in Section 3. In Section 4, the achieved results are specified and extendedly discussed while the main conclusions are given in the last section.

2. MATERIALS AND METHODS

2.1. Database

The discriminative power of the geometrical and textural features is discussed over the health status classification results verified on the publicly available Mammographic Image Analysis Society (MIAS) digital mammogram database [39]. This database comprises two-breast Mediolateral Oblique (MLO) view mammography images of 161 objects having fatty, fatty-glandular, and dense tissue types with 330 diagnoses as 207 normal, 69 benign cancers and 54 malignant cancers, at a size of 1024×1024 and resolution of 8 bits/pixel in ". pgm" imaging format, with their ground truth information. Figure 1 shows sample images for each breast tissue type in columns and health status in rows.

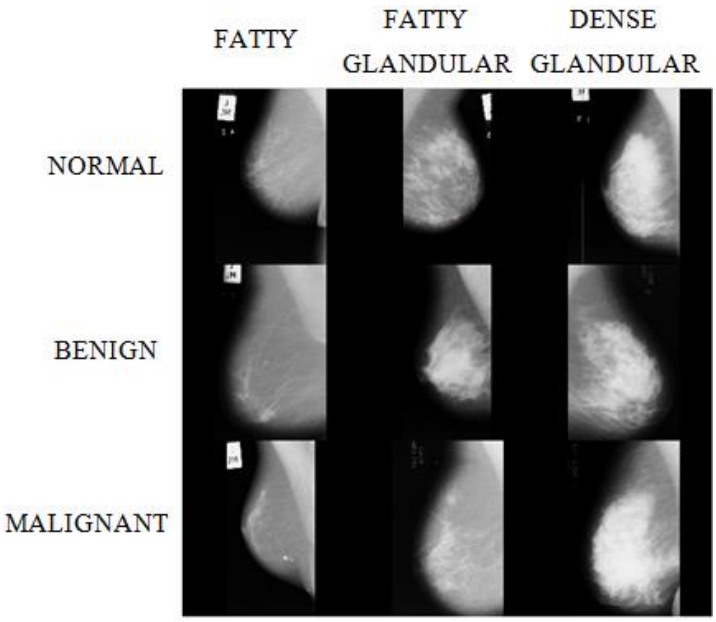


Figure 1. Sample mammography images from the MIAS database [34].

2.2. Pre-Processing

The ROIs in a mammography image typically differs through brighter intensity levels from the breast parenchyma. However, as it is clearly seen in Figure 2, the presence of low- and/or high-level labelling artifacts and pectoral muscles in a mammogram also have brighter intensity levels than the breast parenchyma. Hence, a pre-processing stage is necessitated for a successful ROI detection process.

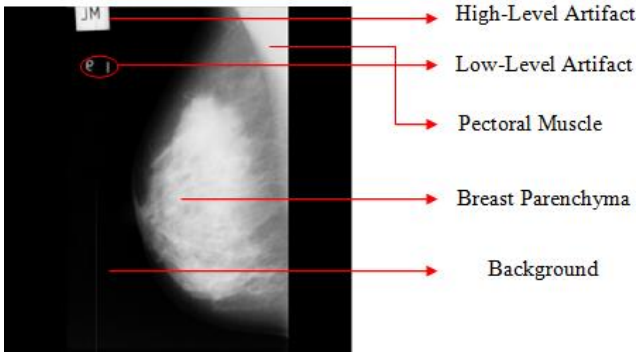


Figure 2. Sample MLO-view mammography image [34].

Initially, all mammography images are resized to a size of 256×256 using bi-cubic interpolation, for ease of operation, and adaptive median filter is applied on mammography images for digitization noise reduction while preserving gross details such that pectoral muscle and ROI edges [34]. Then, labelling artifacts are suppressed and the background is removed by defining the breast parenchyma as the largest area of the binarized noise-reduced mammograms [34]. As the last step of the pre-processing stage, pectoral muscle removal is realized using a region growing algorithm [34]. Figure 3 demonstrates the pre-processing stage on a sample image [35].

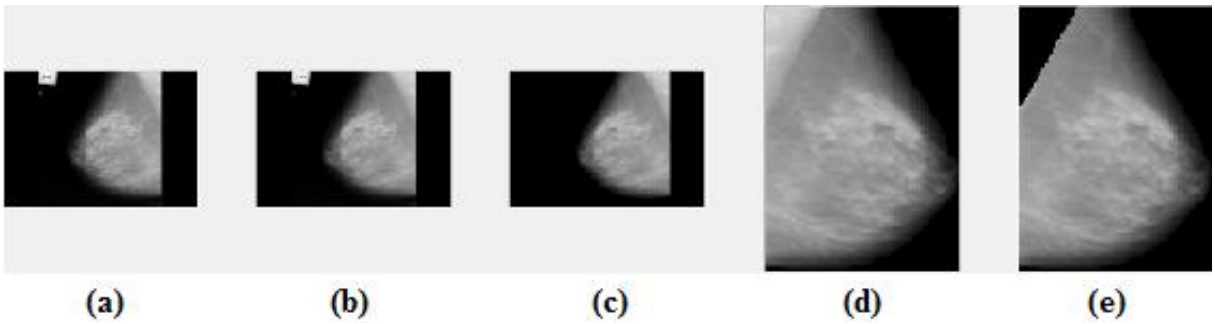


Figure 3. The phases for the pre-processing stage of the proposed system: (a) The original image; (b) The noise-reduced image; (c) The artifacts-suppressed image; (d) The background-removed and left-aligned image; (e) The pectoral muscle-removed image [35].

2.3. Region-of-Interest (ROI) Detection

The ROI detection is realized by using Chan-Vese active contour modelling [40] by manual initial seed point selection and morphological operations are applied for ROI enhancement [35]. This scheme is visualized on a sample mammography image in Figure 4.

2.4. Feature Extraction

2.4.1. Haralick feature extraction

The textural structure of ROIs is defined by Haralick features, introduced by Haralick et al. [36], Soh et al. [37], and Clausi [38], extracted from the GLCM of the relevant ROIs, in this paper.

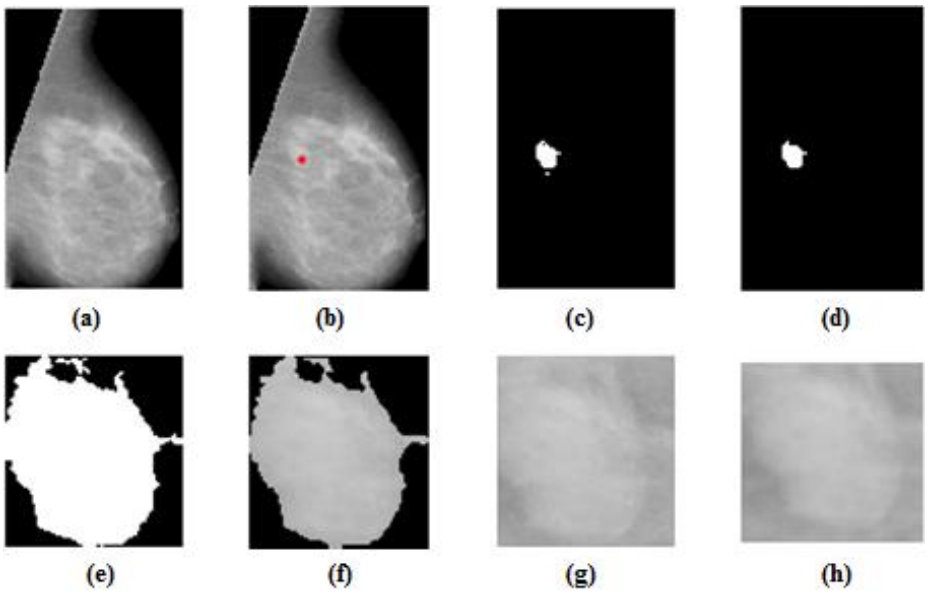


Figure 4. The phases for the adopted ROI detection system: (a) The pre-processed image; (b) The user-defined initial seed point (red) for the Chan-Vese algorithm; (c) The segmented ROI in the original binary image; (d) The enhanced ROI in the original binary image; (e) The binary ROI image; (f) The gray-scaled ROI image; (g) The segmented ROI region; (h) The reference ROI region [35].

The GLCM of any image I is constructed by computing the second order joint probability density function $P(i,j|d,\theta)$, and customizes the textural feature map of this image by pixel frequency calculation, in a defined spatial relationship, based on gray-level intensity values [36]. The spatial relationship of any pixel pair is defined by the distance d and the angle θ between these pixels. The function $P(i,j|d,\theta)$ for an N-gray-level image I is computed by (1), and is resulted in an $N \times N$ –sized matrix that comprises of the probabilities of the co-occurrence of intensities i and j between each reference pixel $I(x,y)$ and its neighbor $I(x+\Delta_x, y+\Delta_y)$ at a distance d through the direction of θ [36].

$$P(i,j|d,\theta) = \sum_{i=1}^{N}\sum_{j=1}^{N}\begin{cases}1 & , \quad I(x,y) = i \; and \; I(x+\Delta_x, y+\Delta_y) = j \\ 0 & , \quad otherwise\end{cases} \tag{1}$$

The frequently used directions θ, for GLCM construction, are shown in Figure 5 for unit distance, and the textural, Haralick, features extracted from the GLCMs with their mathematical representations are given in Table 1.

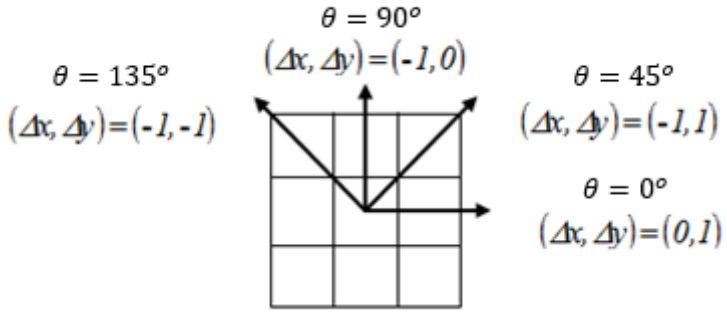


Figure 5. The frequently used directions for GLCM construction ($d = 1$).

2.4.2. Adaptive convex hull approach

The ES approach is proposed by Türkoğlu and Hanbay [30] for geometrical shape examinations of the leaves inspired by the CD method introduced by Mahdikhanlou and Ebrahimnezhad [29]. The CD method uses the distances of a fixed-value of ß-degree-incremented n $\left(n = \frac{360}{ß}\right)$ points $(p1, p2, \cdots, pn)$ on the boundary curve to the center-of-gravity (CoG) for a geometrical description of the leaves. The ES approach determines the n $\left(n = \frac{\text{Perimeter of the ROI}}{\text{step}-\text{size}}\right)$ points $(p1, p2, \cdots, pn)$ on the boundary curve by a fixed-size of pixel lengths (step-size) resulting in different ß-degrees unlike the CD approach. Then, $n \times 3$ – dimensional feature matrix for each ROI is constructed by computing the angle between each pair of points and the CoG $(ß1, ß2, \cdots, ßn)$, edge-distance between each pair of points $(k1, k2, \cdots, kn)$, and the edge-CoG distance of each point $(d1, d2, \cdots, dn)$. These features are visualized in Figure 6.

In this paper, an ACH approach is examined where the number of the points (n) on the boundary of the ROIs are fixed to a certain number for each ROI, resulting in the same value of angles $\left(ß = \frac{Perimeter\ of\ the\ ROI}{n}\right)$ between each boundary point and the CoG of the ROI; but different in each ROI. Thus, the ACH approach differs from the ES approach by the same angle values within ROIs and from CD by different angle values between ROIs. Then, the angles $(ß_n)$, the edge-CoG distances (d_n), and (d) the edge distances (k_n) are computed and $n \times 3$ – dimensional features are achieved as in the ES approach.

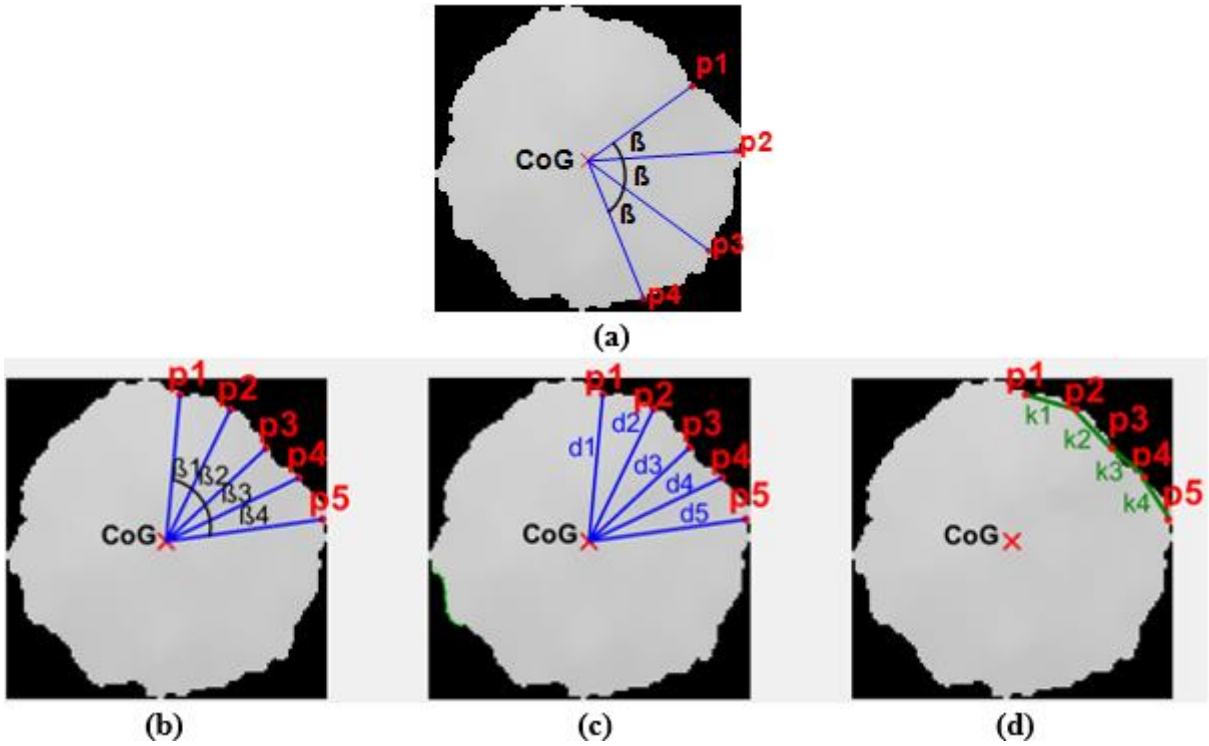


Figure 6. The CD and ES approaches on a sample ROI: (a) The CD approach; (b) The ES approach – angles ($ß_n$); (c) The ES approach – edge-to-CoG distances (d_n); (d) The ES approach – sequential edge distances (k_n).

Table 1. Textural features and their mathematical representations [24].

Textural Features		Mathematical Representations
$P(i,j)$: $GLCM = \begin{bmatrix} P(1,1) & \cdots & P(1,N) \\ \vdots & \ddots & \vdots \\ P(N,1) & \cdots & P(N,N) \end{bmatrix}$		
$P_x(i) = \sum_{j=1}^{N} P(i,j)$, $P_y(j) = \sum_{i=1}^{N} P(i,j)$		
$P_{x+y}(k) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ i+j=k}}^{N} P(i,j)$, $k = 2,3,\cdots,2N$ $P_{x-y}(k) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ \lvert i-j\rvert=k}}^{N} P(i,j)$, $k = 0,1,\cdots,N-1$		
$\mu_x = \sum_i \sum_j i \cdot P(i,j)$, $\mu_y = \sum_i \sum_j j \cdot P(i,j)$		
$\sigma_x = \sum_i \sum_j (i-\mu_x)^2 \cdot P(i,j)$, $\sigma_y = \sum_i \sum_j (j-\mu_y)^2 \cdot P(i,j)$		
***f*1**	Autocorrelation [37]	$\sum_i \sum_j (i \cdot j) \cdot P(i,j)$
***f*2**	Contrast [36, 37]	$\sum_{n=0}^{N-1} n^2 \left\{ \sum_{i=1}^{N} \sum_{\substack{j=1 \\ \lvert i-j\rvert=n}}^{N} P(i,j) \right\}$

***f*3**	Correlation (MATLAB R2021a)	$\frac{\sum_i \sum_j (i-\mu_x)\cdot(j-\mu_y)\cdot P(i,j)}{\sigma_x \cdot \sigma_y}$
***f*4**	Correlation [36, 37]	$\frac{\sum_i \sum_j (i,j)\cdot P(i,j) - \mu_x \cdot \mu_y}{\sigma_x \cdot \sigma_y}$
***f*5**	Cluster Prominence [37]	$\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\{i+j-\mu_x-\mu_y\}^4 \cdot P(i,j)$
***f*6**	Cluster Shade [37]	$\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\{i+j-\mu_x-\mu_y\}^3 \cdot P(i,j)$
***f*7**	Dissimilarity [37]	$\sum_i \sum_j \lvert i-j\rvert \cdot P(i,j)$
***f*8**	Energy [36, 37]	$\sum_i \sum_j \{P(i,j)\}^2$
***f*9**	Entropy [37]	$-\sum_{i=0}^{N-1}\sum_{j=0}^{N-1} P(i,j)\cdot log\{P(i,j)\}$
***f*10**	Homogeneity (MATLAB R2021a)	$\sum_i \sum_j \frac{1}{1+\lvert i-j\rvert}\cdot P(i,j)$
***f*11**	Homogeneity [37]	$\sum_i \sum_j \frac{1}{1+(i-j)^2}\cdot P(i,j)$
***f*12**	Maximum Probability [37]	$\max_{i,j} P(i,j)$
***f*13**	Sum of Squares: Variance [36]	$\sum_i \sum_j (i-\mu)^2 \cdot P(i,j)$
***f*14**	Sum Average [36]	$\sum_{i=2}^{2N} i \cdot P_{x+y}(i)$
***f*15**	Sum Variance [36]	$\sum_{i=2}^{2N} (i-f_{14})^2 \cdot P_{x+y}(i)$
***f*16**	Sum Entropy [36]	$-\sum_{i=2}^{2N} P_{x+y}(i)\cdot log\{P_{x+y}(i)\}$
***f*17**	Difference Variance [36]	variance of P_{x-y}
***f*18**	Difference Entropy [36]	$-\sum_{i=2}^{2N} P_{x-y}(i)\cdot log\{P_{x-y}(i)\}$
***f*19**	Information Measure of Correlation 1 [36]	$\frac{HXY-HXY1}{max\{HX;HY\}}$ $HXY = -\sum_i \sum_j P(i,j)\cdot log(P(i,j))$ $HXY1 = -\sum_i \sum_j P(i,j)\cdot log\left(p_x(i)\cdot p_y(i)\right)$ Hx ve HY, p_x ve p_y'nin entropi değerleridir.
***f*20**	Information Measure of Correlation 2 [36]	$(1-exp[-2\cdot(HXY2-HXY)])^{1/2}$ $HXY2 = -\sum_i \sum_j P_x(i)\cdot P_y(i)\cdot log\{P_x(i)\cdot P_y(i)\}$
***f*21**	Inverse Difference Normalized [38]	Normalized Homogeneity (MATLAB R2021a)
***f*22**	Inverse Difference Moment Normalized [38]	Normalized Homogeneity [37]

3. EXPERIMENTAL STUDY

In the experimental studies, 110 ROIs, confirmed by a radiologist as accurately detected on mammography images in the MIAS database, are used. This selected dataset consisted of diagnoses as 52 normal, 39 benign cancers and 19 malignant cancers.

3.1. Feature Vector Construction

The textural description of the ROIs is provided by the Haralick features in this paper. Accordingly, the features given in Table 1 are extracted from the GLCMs composed by $\theta = 90^o$ and $d = \{1,\ 2,\ 3,\ 4\}$, and four 22×1–dimensional feature vectors are obtained. The resultant 22×1–dimensional textural feature vector is the mean of these four vectors and named as "Haralick Feature Vector (HaFV)" in this paper.

Geometric identification is realized by the ACH approach by selecting the number of points (n) on the boundary of the ROIs as 92 for each ROI. The angles (ß), the edge-to-CoG distances (d_n), and (d) the sequential edge distances (k_n) are computed and concatenated for each ROI to construct its 276×1–dimensional "Adaptive Convex Hull Feature Vector (AChFV)".

Besides, 298×1–dimensional Hybrid Feature Vector (HyFV) of a ROI is constructed by concatenating the AChFV and HaFV. The feature vector construction process is summarized in Table 2.

Table 2. Feature vector construction process.

Haralick Feature Vector (HaFV)	**Adapted Edge Step Feature Vector (AChFV)**	**Hybrid Feature Vector (HyFV)**
22×1	276×1	298×1
Haralick Features in Table 1 extracted from GLCM composed by $\theta = 90^o$ and $d = \{1,\ 2,\ 3,\ 4\}$.	$\begin{bmatrix} (ß_1 & \mathrm{d}_1 & \mathrm{k}_1)' \\ (ß_2 & \mathrm{d}_2 & \mathrm{k}_2)' \\ & \vdots & \\ (ß_{92} & \mathrm{d}_{92} & \mathrm{k}_{92})' \end{bmatrix}$	$\begin{bmatrix} \mathrm{ACH\_FV} \\ \mathrm{Ha\_FV} \end{bmatrix}$

3.2. Classification and Performance Evaluation

The 22-dimensional Haralick features, the 276-dimensional AChFV features, and the 298-dimensional hybrid features are utilized in the diagnosis stage via Random Forest (RF) classifier, linear Support Vector Machines (SVM) classifier, Logistic Linear Classifier (LOGLC), and Artificial Neural Networks (ANN) using 5-fold cross-validation. A feed-forward backpropagation network consisting one hidden layer with 20 neurons is trained and simulated for the ANN classifier. The discriminative power of each feature vector is evaluated using the metrics of 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). The classification process is repeated for each fold in the cross-validation technique, and the PEMs of each health status class are computed as given in Table 3. The PEMs of any diagnosis process are then evaluated in terms of weighted average of the per class metrics, for each fold. The average PEMS of the diagnosis process is the arithmetic mean of the weighted metrics computed in each fold.

Table 3. Performance evaluation metrics and their mathematical representations.

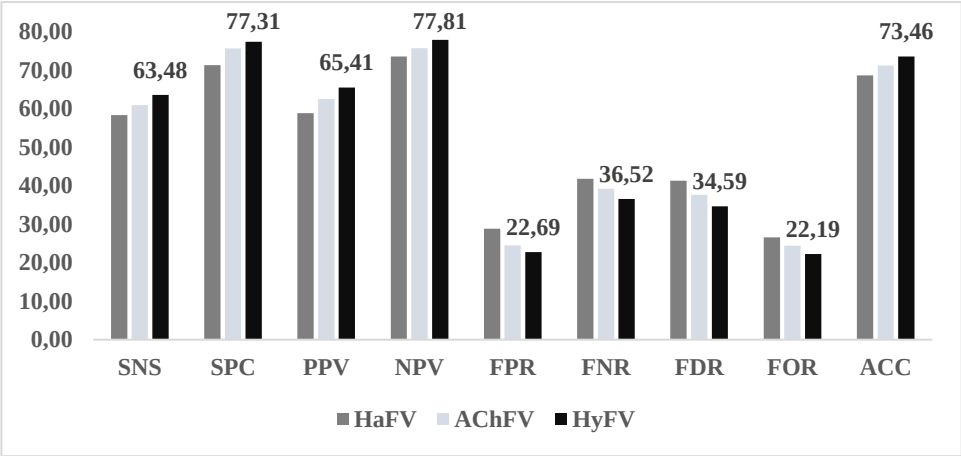
Performance Evaluation Metrics (PEMs)	**Mathematical Representations**			
C: Number of health status classes ($C = 3$)	i: Class index ($i = 1, 2, 3$)		N_i: number of samples belong to i^{th} class	
TP_i: number of True Positives belong to i^{th} class	TN_i: number of True Negatives belong to i^{th} class	FP_i: number of False Positives belong to i^{th} class	FN_i: number of False Negatives belong to i^{th} class	
	Per class metrics		**Weighted metrics**	
SNS	$\% SNS_i = \frac{TP_i}{TP_i + FN_i} \cdot 100$		$\% SNS = \frac{\sum_{i=1}^{C} SNS_i \cdot N_i}{\sum_{i=1}^{C} N_i}$	
SPC	$\% SPC_i = \frac{TN_i}{TN_i + FP_i} \cdot 100$		$\% SPC = \frac{\sum_{i=1}^{C} SPC_i \cdot N_i}{\sum_{i=1}^{C} N_i}$	
PPV	$\% PPV_i = \frac{TP_i}{TP_i + FP_i} \cdot 100$		$\% PPV = \frac{\sum_{i=1}^{C} PPV_i \cdot N_i}{\sum_{i=1}^{C} N_i}$	
NPV	$\% NPV_i = \frac{TN_i}{TN_i + FN_i} \cdot 100$		$\% NPV = \frac{\sum_{i=1}^{C} NPV_i \cdot N_i}{\sum_{i=1}^{C} N_i}$	
FPR	$\% FPR_i = \frac{FP_i}{FP_i + TN_i} \cdot 100$		$\% FPR = \frac{\sum_{i=1}^{C} FPR_i \cdot N_i}{\sum_{i=1}^{C} N_i}$	
FNR	$\% FNR_i = \frac{FN_i}{TP_i + FN_i} \cdot 100$		$\% FNR = \frac{\sum_{i=1}^{C} FNR_i \cdot N_i}{\sum_{i=1}^{C} N_i}$	
FDR	$\% FDR_i = \frac{FP_i}{TP_i + FP_i} \cdot 100$		$\% FDR = \frac{\sum_{i=1}^{C} FDR_i \cdot N_i}{\sum_{i=1}^{C} N_i}$	
FOR	$\% FOR_i = \frac{FN_i}{TN_i + FN_i} \cdot 100$		$\% FOR = \frac{\sum_{i=1}^{C} FOR_i \cdot N_i}{\sum_{i=1}^{C} N_i}$	
ACC	$\% ACC_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \cdot 100$		$\% ACC = \frac{\sum_{i=1}^{C} ACC_i \cdot N_i}{\sum_{i=1}^{C} N_i}$	

4. RESULTS AND DISCUSSION

In this paper, textural and geometrical features are utilized on a pre-designed CAD system [34, 35] in the light of a radiologist-eye, for breast cancer diagnosis. The experiments are performed on a computer with I5− 7200U at 2.5 GHz and 8-Gb memory and have been carried out using MATLAB R2021a.

In the first phase of these experiments, breast cancer diagnosis is realized in three processes consist of textural, geometrical, and hybrid (an ensemble of textural and geometrical) classification using the HaFVs, AChFVs, and HyFVs, respectively. The average PEMs achieved in these processes are shown in Tables 4-7. Analysing the PEMs in Table 4 for all three cases, it is seen that better diagnosis is succeeded when HyFVs are used via all classifiers except the SVM. The RF classifier is found to diagnose the ROIs more accurately. Hence, the average PEMs of normal, benign, and malignant diagnoses via RF classifier are individually analysed. Although the HyFVs show the best performance for the overall diagnosis, it is seen, from the Figures 7-9, that the HaFVs, the AChFVs, and the HyFVs have different impacts on different health status. Figure 8 indicates that the ROIs without any abnormalities are more accurately diagnosed when the HyFVs are used, and the HaFVs play a key role for malignancy diagnosis as seen in Figure 8. In case of benignity, Figure 9 shows that the HyFVs provide more sensitive and accurate diagnosis while the HaFVs diagnose more specifically reducing the false positives.

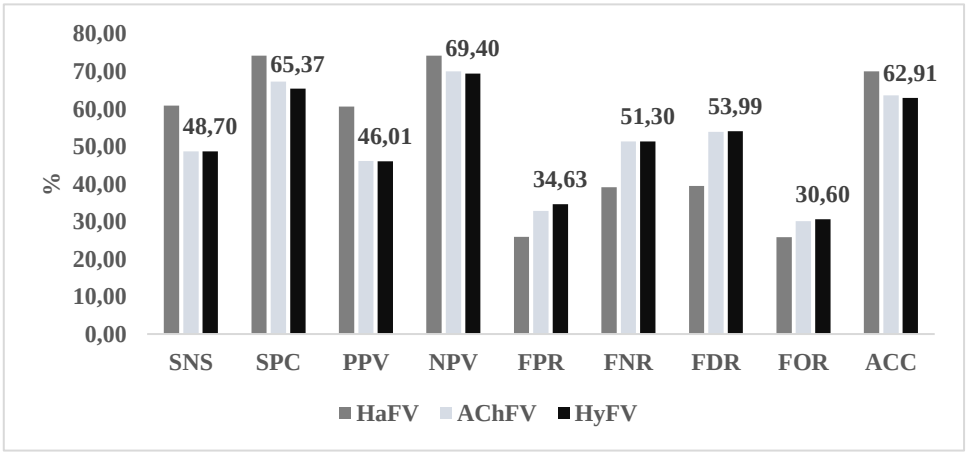
Table 4. Findings of overall diagnosis processes achieved via RF classifier.

Classifier	Average PEMs
RF	

Total Confusion Matrices

HaFV		Classified to			AChFV		Classified to			HyFV		Classified to		
		Normal	Benign	Malignant			Normal	Benign	Malignant			Normal	Benign	Malignant
Ground Truth	Normal	38	15	2	Ground Truth	Normal	36	18	1	Ground Truth	Normal	38	16	1
	Benign	19	19	2		Benign	12	24	4		Benign	11	26	3
	Malignant	4	6	10		Malignant	3	7	10		Malignant	2	9	9

Table 5. Findings of overall diagnosis processes achieved via SVM classifier.

Classifier	Average PEMs
SVM	

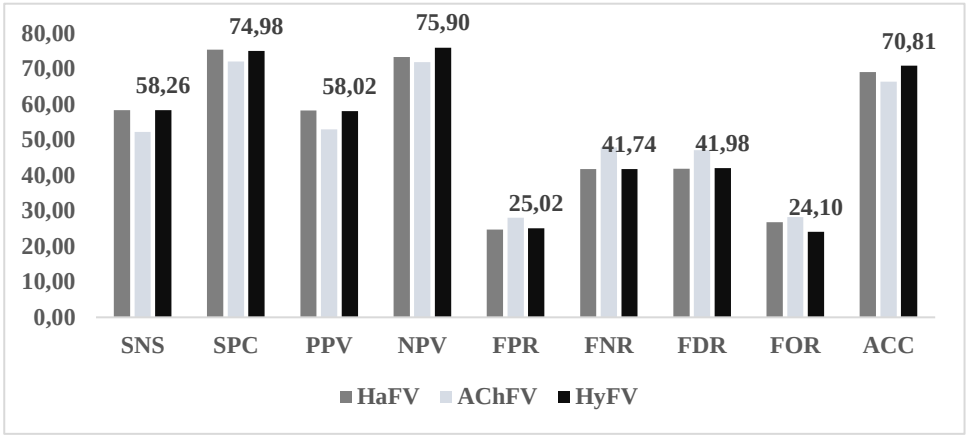
Total Confusion Matrices

HaFV		Classified to			ChFV		Classified to			HyFV		Classified to		
		Normal	Benign	Malignant			Normal	Benign	Malignant			Normal	Benign	Malignant
Ground Truth	Normal	36	18	1	Ground Truth	Normal	37	15	3	Ground Truth	Normal	38	15	2
	Benign	16	19	5		Benign	18	17	5		Benign	21	16	3
	Malignant	4	1	15		Malignant	5	13	2		Malignant	5	13	2

Table 6. Findings of overall diagnosis processes achieved via LOGLC classifier

Classifier: LOGLC

Average PEMs



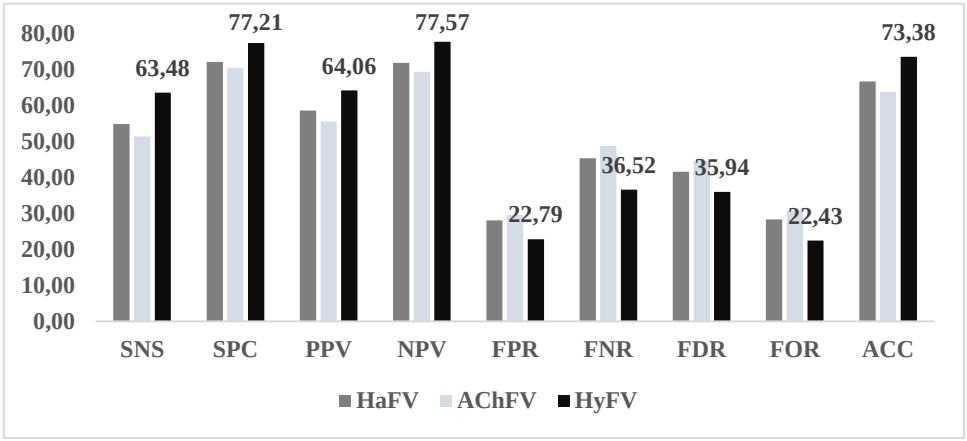
Total Confusion Matrices

HaFV		Classified to			AChFV		Classified to			HyFV		Classified to		
		Normal	Benign	Malignant			Normal	Benign	Malignant			Normal	Benign	Malignant
Ground Truth	Normal	34	14	7	Ground Truth	Normal	34	18	3	Ground Truth	Normal	38	14	3
	Benign	17	16	7		Benign	12	21	7		Benign	11	24	5
	Malignant	2	1	17		Malignant	4	11	5		Malignant	4	11	5

Table 7. Findings of overall diagnosis processes achieved via ANN classifier

Classifier: ANN

Average PEMs



Total Confusion Matrices

HaFV		Classified to			AChFV		Classified to			HyFV		Classified to		
		Normal	Benign	Malignant			Normal	Benign	Malignant			Normal	Benign	Malignant
Ground Truth	Normal	31	23	1	Ground Truth	Normal	27	28	0	Ground Truth	Normal	38	14	3
	Benign	13	24	3		Benign	12	23	5		Benign	12	25	3
	Malignant	2	10	8		Malignant	3	8	9		Malignant	3	7	10

The PEM scores and the confusion matrices in Table 6 show that the healthy and benignity diagnoses power of the HyFVs are increased when the LOGLC classifier is used however the malignancy discriminability is decreased as it is also by using the SVM classifier as seen in Table 5. There is no significant difference found on diagnosis either the Random Forest or the ANN classifier is used (see Tables 4 and 7). The PEM scores and the confusion matrices achieved by all classifiers except SVM, show that deficiency of textural features in benign diagnosis is relatively compensated when used in combination with geometric features.

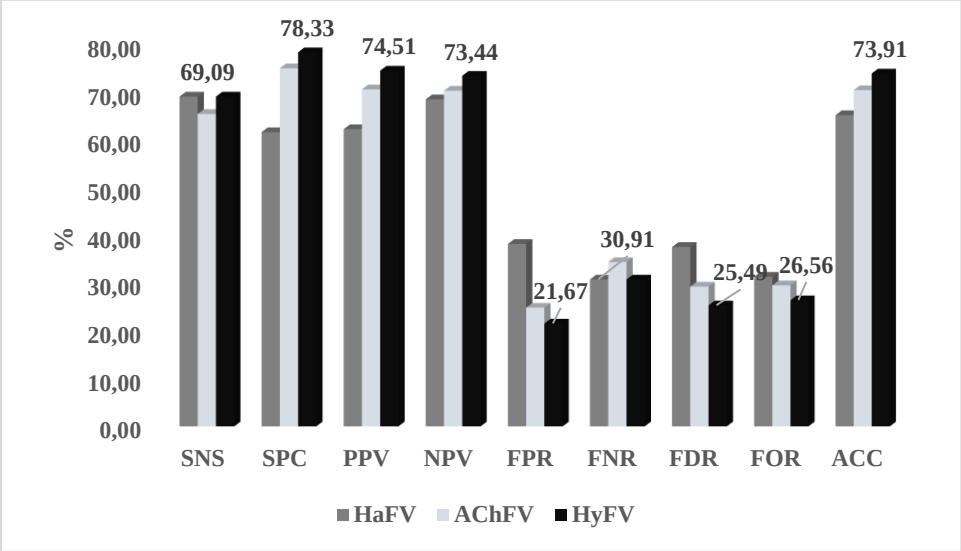


Figure 7. Average PEMs of normal diagnoses, achieved when RF classifier is used.

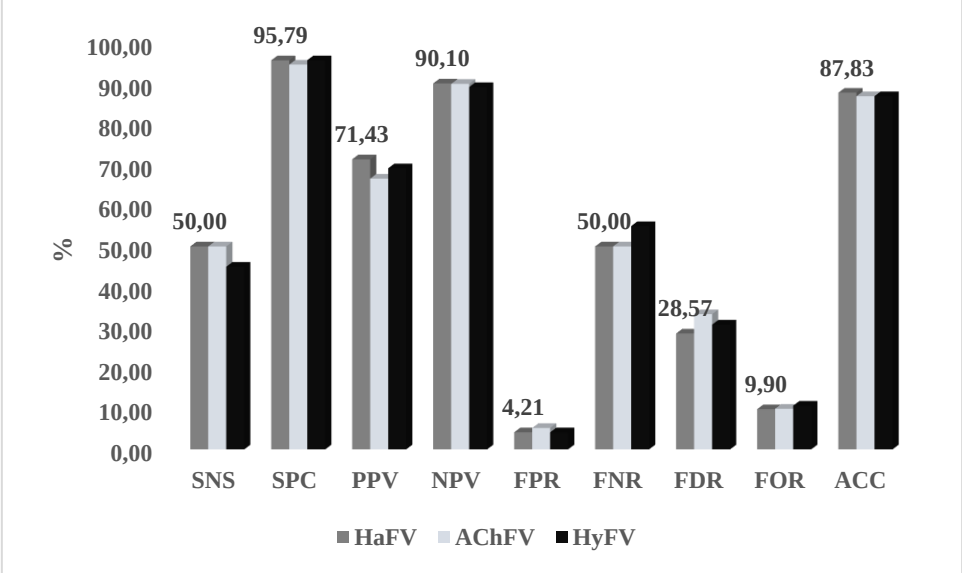


Figure 8. Average PEMs of malignant diagnoses, achieved when RF classifier is used.

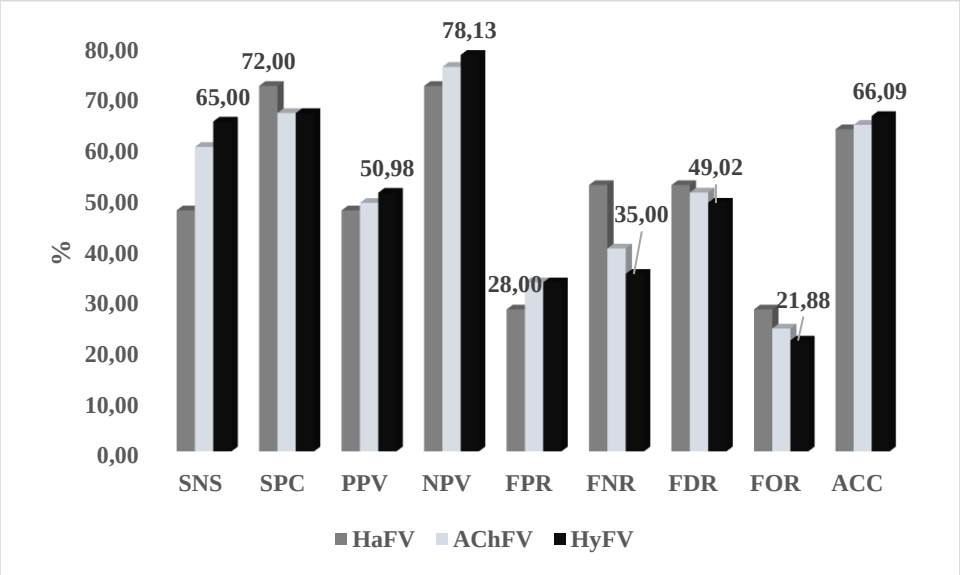


Figure 9. Average PEMs of benign diagnoses, achieved when RF classifier is used.

In line with the findings obtained in the first phase of the experiments, a combination of the HaFV, the AChFV, and the HyFV via majority voting is employed for diagnosis in the second phase. The average PEMs achieved for each health status in this phase are shown in Figure 10. The accuracy values for benign and malignant diagnoses increased, and the number of false positives is decreased by this feature combination.

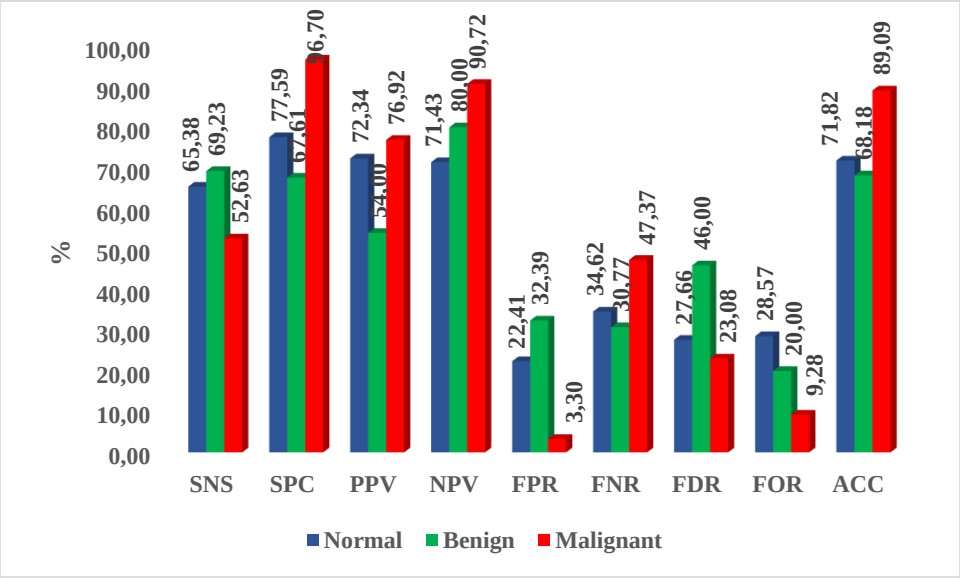


Figure 10. Average PEMs of diagnosis processes via feature combination, achieved when RF classifier is used.

Although there is a slight deterioration in the PEM values of healthy diagnoses and an insignificant increase in overall PEMs (see Figure 11), more accurate diagnoses of benignity and especially malignancy show that the usage of feature combination is more effective.

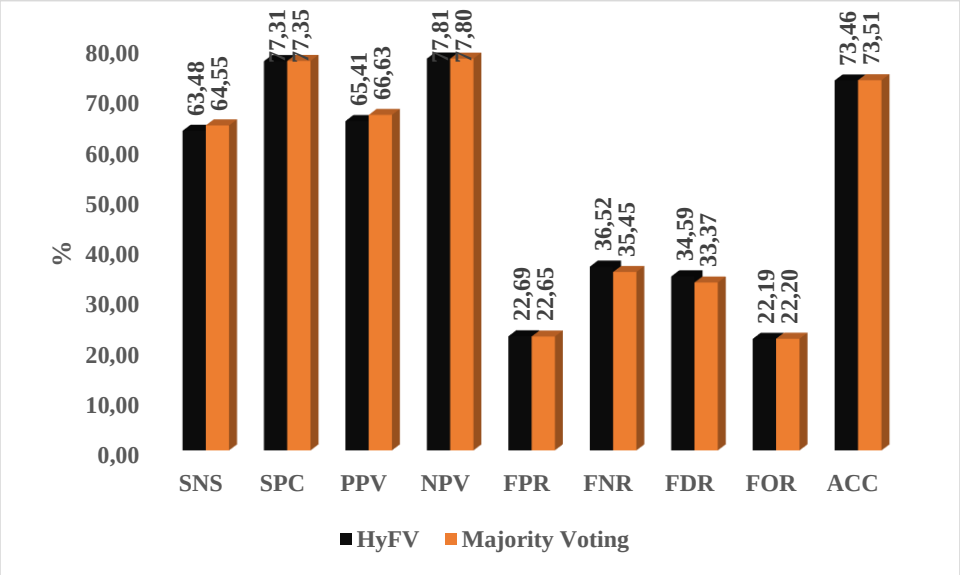


Figure 11. Average PEMs of diagnosis processes: HyFV vs. feature combination, achieved when RF classifier is used.

5. CONCLUSION

The chance of decrement in breast cancer-caused mortality rate due to the increment in early diagnosis is concluded in a great importance of CAD systems designed for breast cancer diagnosis. The diagnostic performance of a CAD system is known to be most directly related to the feature extraction stage. Although several feature extraction strategies are evaluated for breast cancer diagnosis in the literature, it is more feasible and consistent to use ROI-specific features. Radiologists state that any ROI is evaluated in terms of its brightness, density, shape, and contour structure during the examination. The density-related information may be revealed through textural definitions while shape and contour structure of a ROI may be defined geometrically. Accordingly, Haralick features extracted from GLCMs are utilized for textural feature vector construction. A new adaptive convex hull approach is introduced and applied for geometrical feature vector creation. Besides, the geometrical and textural feature vectors are concatenated, so that a hybrid feature vector is constituted to analyse the combined effect of these features as well as the individual feature performances on diagnosis. The RF, SVM, LOGLC, and ANN classifiers are accomplished using a 5-fold cross-validation technique for ROI diagnosis using textural, geometrical, and hybrid features. The experimental studies indicate that textural features are a bit more successful for diagnosing ROIs including malignancy. One can easily infer from the PEM scores that the proposed hybrid features, which are the concatenation of textural and geometrical features, are more fruitful on classifying both healthy and benign ROIs. Additionally, a feature combination of these three types of features via majority voting combination technique is proposed for diagnosis. The benign and malignant ROI cases are more successfully identified with this proposed feature combination and a little bit increment is achieved on the average PEMs of whole diagnosis process. A combination of geometrical features and notional textural features for breast tissue type may be planned to examine for future studies for more accurate breast cancer diagnosis.

CONFLICT OF INTEREST

The authors stated that there are no conflicts of interest regarding the publication of this article.

REFERENCES

[1] World Health Organization, available at https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ (accessed January 2020).

[2] Wang, L. Early diagnosis of breast cancer. Sensors 2017; 17 (7): 1572-1591.

[3] Meenalochini G, Ramkumar S. Survey of machine learning algorithms for breast cancer detection using mammogram images. Materials Today: Proceedings, 2021; 37 (2): 2738-2743.

[4] Ergin S, Kılınç O. A new feature extraction framework based on wavelets for breast cancer diagnosis. Comput Biol Med, 2014; 51: 171-182.

[5] Heywang-Köbrunner SH, Hacker A, Sedlacek S. Advantages and disadvantages of mammography screening. Breast Care, 2011; 6 (3):199-207.

[6] Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin, 2011; 61 (2): 69-90.

[7] Üncü YA, Özdoğan H. Mamografi sistemlerinde ilgi alanı, türev ve ince gruplama seçimlerinin modülasyon transfer fonksiyonunun üzerine etkileri. Süleyman Demirel Üniversitesi Fen Edebiyat Fakültesi Fen Dergisi, 2020; 15 (1): 23-35.

[8] Üncü YA, Sevim G, Mercan T, Vural V, Durmaz E, Canpolat M. Differentiation of tumoral and non-tumoral breast lesions using back reflection diffuse optical tomography: A pilot clinical study. Int J Imaging Syst Technol, 2021; 1-9.

[9] Radovic M, Djokovic M, Peulic A, Filipovic N. Application of data mining algorithms for mammogram classification. In: 2013 IEEE 13th International Conference on Bioinformatics and Bioengineering (BIBE); 10-13November 2013; Chania, Greece, 1–4.

[10] Ganesan K, Acharya UR, Chua CK, Min LC, Matthew B, Thomas AK. Decision support system for breast cancer detection using mammograms. Proc Inst Mech Eng H, 2013; 227 (7): 721–732.

[11] Li JB, Wang YH, Chu SC, Roddick JF. Kernel self-optimization learning for kernel-based feature extraction and recognition. Inf Sci, 2014; 257: 70-80.

[12] Ramos RP, Nascimento MZ, Pereira DC. Texture extraction: an evaluation of ridgelet, wavelet and co-occurrence based techniques applied to mammograms. Expert Syst Appl, 2012; 39 (12): 11036-11047.

[13] Shradhananda B, Banshidhar M, Ratnakar D. Mammogram classification using two dimensional discrete wavelet transform and gray-level co-occurrence matrix for detection of breast cancer. Neurocomputing, 2015; 154: 1–14.

[14] Vallez N et al. Breast density classification to reduce false positives in CADe systems. Comput Biol Med, 2013;113 (2): 569–584.

[15] Imran S, Lodhi BA, Alzahrani A. Unsupervised method to localize masses in mammograms," in IEEE Access, 2021; 9: 99327-99338.

[16] Heidari M et al. Applying a random projection algorithm to optimize machine learning model for breast lesion classification. IEEE Trans Biomed Eng, 2021; 68 (9): 2764-2775.

[17] Loizidou K, Skouroumouni G, Nikolaou C, Pitris C. An Automated breast micro-calcification detection and classification technique using temporal subtraction of mammograms. IEEE Access, 2020; 8: 52785-52795.

[18] Heidari M, Mirniaharikandehei S, Liu W, Hollingsworth AB, Liu H, Zheng B. Development and assessment of a new global mammographic image feature analysis scheme to predict likelihood of malignant cases. IEEE Trans Med Imaging, 2020; 39 (4): 1235-1244.

[19] Sampaio WB, Diniz EM, Silva AC, Paiva AC, Gattass M. Detection of masses in mammogram images using CNN, geostatistic functions and SVM. Comput Biol Med, 2011; 41 (8): 653–664.

[20] Keleş A, Keleş, A, Yavuz U. Expert system based on neuro-fuzzy rules for diagnosis breast cancer. Expert Syst Appl, 2011; 38 (5): 5719–5726.

[21] Krishnan MMR, Banerjee S, Chakraborty C, Chakraborty C, Ray AK. Statistical analysis of mammographic features and its classification using support vector machine. Expert Syst Appl, 2010; 37 (1): 470–478.

[22] Verma B, McLeod P, Klevansky A. A novel soft cluster neural network for the classification of suspicious areas in digital mammograms. Pattern Recognit, 2009; 42 (9): 1845–1852.

[23] Papadopoulos A, Fotiadis DI, Costaridou L. Improvement of microcalcification cluster detection in mammography utilizing image enhancement techniques. Comput Biol Med, 2008; 38 (10): 1045–1055.

[24] Işıklı Esener İ, Ergin S, Yüksel T. A genuine GLCM-based feature extraction for breast tissue classification on mammograms. Int J Intell Syst Appl Eng, 2016; 4 (Special Issue): 124-129.

[25] Song R, Li T, Wang Y. Mammographic classification based on xgboost and dcnn with multi features. in IEEE Access, 2020; 8: 75011-75021.

[26] 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. In: 2017 7th Latin American Conference on Networked and Electronic Media (LACNEM 2017); 6-7 Nov. 2017; Valparaiso, Chile, 22-27.

[27] Osada R, Funkhouser T, Chazelle B, Dobkin D. Shape distributions. ACM Trans Graph, 2002; 21(4): 807–832.

[28] Yu M, Atmosukarto I, Leow WK, Huang Z, Xu R. 3D model retrieval with morphing based geometric and topologic topological feature maps. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 18-20 June 2003; Madison, WI, USA, II-656.

[29] Mahdikhanlou K, Ebrahimnezhad H. Plant leaf classification using centroid distance and axis of least inertia method. In: 2014 22nd Iranian Conference on Electrical Engineering (ICEE); 20-22 May 2014; Tehran, Iran, 1690-1694.

[30] Türkoğlu M, Hanbay D. Plant recognition system based on extreme learning machine by using shearlet transform and new geometric features (article in Turkish with an abstract in English). J Fac Eng Archit Gaz, 2019; 34 (4): 2097-2112.

[31] Azlan NAN, Lu CK, Elamvazuthi I, Tang TB. Automatic detection of masses from mammographic images via artificial intelligence techniques. IEEE Sens J, 2020 (21): 13094-13102.

[32] Mohanty F, Rup S, Dash B, Majhi B, Swamy MNS. An improved scheme for digital mammogram classification using weighted chaotic salp swarm algorithm-based kernel extreme learning machine, Appl Soft Comput, 2020; 91: 1568-4946.

[33] Işıklı Esener İ, Ergin S, Yüksel T. A coping with breast cancer diagnosis using a normalized texture feature set. In: 2017 International Conference on Engineering Technologies (ICENTE17); 7-9 December 2017; Konya, Turkey, 38-43.

[34] Işıklı Esener İ, Ergin S, Yüksel T. A novel multistage system for the detection and removal of pectoral muscles in mammograms. Turk J Elec Eng & Comp Sci, 2018; 26 (1): 35-49.

[35] Işıklı Esener İ, Ergin S, Yüksel T. A practical Region-of-Interest (ROI) detection approach for suspicious region identification in breast cancer diagnosis. In: 2017 International Conference on Engineering Technologies (ICENTE17); 7-9 December 2017; Konya, Turkey, 44-47.

[36] Haralick RM, Shanmugam K, Dinstein I. Textural features of image classification. IEEE Trans Syst, Man, Cybern Syst, 1973; SMC-3 (6): 610-621.

[37] Soh L, Tsatsaulis C. Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. IEEE Trans Geosci Remote Sens, 1999; 37 (2): 780 – 795.

[38] Clausi DA. An analysis of co-occurrence texture statistics as a function of grey level quantization. Can J Remote Sens, 2002; 28 (1): 45-62.

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

[40] Chan TF, Vese LA. Active contours without edges. IEEE Trans Image Process 2001; 10: 266-277.