IIOAB JOURNAL

ARTICLE

A COMPARATIVE STUDY ON SEGMENTATION AND CLASSIFICATION IN BREAST MRI IMAGING

Ahmet Haşim Yurttakal[1*], Hasan Erbay[2], Türkan İkizceli[3], Seyhan Karaçavuş[4], Gökalp Çinarer[1]

[1]*Computer Technologies Department, Bozok University, Yozgat, TURKEY*

[2]*Computer Engineering Department, Kırıkkale University, 71450, Kırıkkale, TURKEY*

[3]*Department of Radiology, University of Health Sciences, Istanbul, TURKEY*

[4]*Department of Nuclear Medicine, University of Health Sciences, Kayseri, TURKEY*

ISSN 0976-3104

ABSTRACT

Background: *Breast cancer is the type of cancer that develops from cells in the breast tissue. The breast cancer is leading cancer in women. One in every eight to nine women has breast cancer at some point during their lifetime. Computer-Aided Diagnosis (CAD) Technology is getting more important to assist radiologists not only to detect breast cancer tumor but also to interpret lesioned regions. The CAD, as a second reader in the clinic, improves the classification of malignant and benign lesions. On the other hand, Magnetic Resonance Imaging (MRI) is a highly recommended test for detecting and monitoring breast cancer tumors and interpreting lesioned regions since it has an excellent capability for soft tissue imaging. In MRI image analysis, the segmentation images are important objective because accurate measurement of the delineation of the regions of interest (ROI) is critical for the breast cancer diagnosis and treatment. Herein, by using MRI scans, we propose a semi-automated CAD system prototype to assist radiologists in detecting breast cancer tumors and interpreting lesioned regions. The prototype, first, pre-processes the raw selected suspicious region to reduce the noises and to reveal the structure. Later, using Expectation Maximization (EM), the prototype segments the pre-processed region. After that, we use the Discrete Wavelet Transform (DWT) for providing efficient multi-resolution sub and decomposition of signals. Then Random Forest Algorithm is used for feature selection. Finally, Naive Bayes, Linear Discriminant Analysis and C4.5 Decision Tree Algorithms are used to classify the features of the ROI in the diagnosis analysis. We tested the prototype CAD on 105 patients, among them, 53 are benign and 52 malign. 80% of the images are allocated for training and 20% of images reserved for testing. The CAD classified 20 patients correctly in case of 5 fold cross-validation. Only one patient is misclassified. The computer-aided diagnosis system with the C4.5 has accuracy 95.24%. Furthermore, C4.5 classifies the breast cancer tumors better than Naive Bayes and Linear Discriminant Analysis. We tested the prototype CAD on 105 patients, among them, 53 are benign and 52 malign. The computer-aided diagnosis system with the C4.5 has accuracy 95.24%. Furthermore, C4.5 classifies the breast cancer tumors better than Naive Bayes and Linear Discriminant Analysis.*

KEY WORDS
Breast cancer, computer aided

INTRODUCTION

Breast cancer is the most common cancer among women, comprising 23% of all female cancers all over the world [1]. In Western countries, one in every eight to nine women has breast cancer at some point during their lifetime [2].

Early detection of a breast cancer tumor is crucial in the treatment process. Mammography is a valuable tool because it can identify breast cancer at an early stage, before physical symptoms develop, as a footnote, screening mammography is the only test to date proven to reduce deaths due to breast cancer [3, 4, 5, 6]. To reduce false-negative diagnosis in mammography, a biopsy is recommended for lesions with greater than a 2% chance of having suspected malignant tumors [7] and, among them, less than 30 percent are found to have malignancy [8, 9]. To reduce unnecessary biopsies, recently, Magnetic Resonance Imaging (MRI) has also been used for the diagnosis of breast cancer [10, 11] since it has an excellent capability for soft tissue imaging and the most sensitive technique in detecting breast diseases, besides, it does not contain potentially dangerous radiation [12]. But interpreting MRI images is both time-consuming and requires reader experience. In recent years, automated systems developed with computer-assisted programs have been improved to noninvasively detect abnormal lesions and determine tissue characterization in medical images. In this context, the extraction of additional features from various imaging modalities by tissue analysis has been investigated intensively. In addition, with the increasing use of medical imaging for diagnosis, treatment planning and clinical experiments, it has become almost imperative to use computers to assist radiologists in clinical diagnosis and treatment planning. Thus, Computer Aided Diagnosis (CAD) is becoming a compulsory tool in assisting radiologists not only in detecting breast cancer tumors but also in interpreting lessoned regions [13, 14]. The CAD, as a second reader in the clinic, improves classification of malignant and benign lesions [15], so it needs to be explored further. The CAD is composed of two main stages: (1) the analysis stage (2) the diagnosis stage. In the analysis stage, after the images are preprocessed, the anatomical structures of the images and their features are extracted. So, reliable methods are required for both the delineation of the regions of interest (ROI) and obtaining their features. The analysis stage consists of series of procedures such that pre-processing, segmentation and feature extraction. In the diagnosis stage, diagnosis rules are determined according to the features obtained in the analysis stage. This stage consists of classification procedure. As a result, in the CAD, there are four crucial modules: preprocessing, segmentation, feature extraction and selection, and classification.

In the literature, nevertheless, when compared to mammography and ultrasound, relatively few CAD systems have been developed for breast MRIs [16, 17]. If the CAD systems in the literature are examined,

Received: 22 June 2018
Accepted: 11 Nov 2018
Published: 20 Nov 2018

***Corresponding Author**
Email:
ahmet.yurttakal@bozok.edu.tr
Tel.: +90-354-2175064
Fax: +90-354-2171780

IIOAB JOURNAL
ISSN: 0976-3104

one can observe that K nearest neighbors (KNN) and support vector machines (SVM) are mostly used as classifiers. In addition to these, Yassin et al. in their extensive literature review study on breast cancer [18] state that there are 2 studies using Linear Discriminant Analysis (LDA) and a study using Naive Bayes, while there are no studies using decision trees as Classification Algorithm with MRI images. We detail existing researches and compare with our study in the Discussion section.

On the other hand, the segmentation module should include either breast segmentation or mass segmentation or both. Vast number of studies on mass segmentation exist in the literature, but studies on breast segmentation are limited. For more on improvements on CADs for breast cancer, see [15].

In this study, we present an alternative set of computational tools to segment and detect breast cancer tumor using MRI images. We employed the Expectation Maximization (EM) [19] to segment the ROI. Then, total of nine features are extracted, five of which are first order intensity-based statistical features [20] and four are GLCM based texture features introduced by Haralick et al. [21] in 1973. Later Random Forest Algorithm is used for feature selection and three different classification procedures (i.e. Naive Bayes, LDA, C4.5 Decision Tree Algorithm) are used. Numerical results obtained show that C4.5 Decision Tree Algorithm classifies the ROIs better than Naive Bayes, LDA.

The rest of the paper organized as follows. Sect. 2 presents the necessary materials and methods throughout the paper such that image acquisition, pre-processing, segmentation, pre-processing segmented image, feature extraction, feature selection and classification. In Sect. 3 we present some simulation results. Section 4 compares our study with the literature and presents conclusion

MATERIALS AND METHODS

The proposed CAD System is divided into five phases (i) image acquisition, (ii) ROI and segmentation of tumor region, (iii) image preprocessing, (iv) feature extraction and feature selection, (v) classification and evaluation. We tested the Proposed CAD on the dataset containing 105 patients MRI images with a dimension of 512x512 and all images are in DICOM format. Proposed CAD is implemented in the Python environment.

Image acquisition

The dataset is composed of breast MR images of patients. Breast MRI was performed at 1.5 Tesla (Achieva, Philips, The Netherlands) with the patients in a prone position using dedicated eight-channel breast coils. Non-fat suppressed T1 weighted scans in the axial plane (TR: 550 ms TE: 10 ms THK: 3 mm, FOV: 300 mm NSA: 2 T: 1.55 min), T2-weighted spoiled gradient echo (GRE) scans with fat-suppression were acquired in the axial plane (TR: 4000 ms TE: 125 ms FOV 300 mm NSA: 2 T: 1.40 min). Dynamic contrast-enhanced imaging was performed by means of a high-resolution T1-weighted gradient echo sequence with automated intravenous bolus application. Subtraction images were obtained for each contrast-enhanced series via subtraction of the non-enhanced series from the enhanced series. After performing of MRI, images were taken from MR device's workstation as DICOM images having slice thickness less than 2.0mm. As a result of image acquisition, a series of 512x512 unit16 images with 65536 different gray levels (0-65535) was obtained.

Dataset

The MRI images in the raw dataset were taken from Haseki Training and Research Hospital in. Istanbul, Turkey. The dataset consists of breast MRI images of 105 patients among them, 53 are benign and 52 malignant. Breast biopsy was performed in diagnosing tumorous regions. [Table 1] presents the characteristic of the dataset.

Table 1: Dataset characteristics

Cases	105
Benign	53
Malign	52
Image Resolution	512x512
Image Format	Dicom
Sequence	STIR
Slice Thickness	<3.00mm

Later, the dataset was randomly split into a training set and a test set. [Table 2] shows the splitting.

Table 2: Testing and training

Cases	Testing	Training
Benign	40	13
Malign	44	8

IIOAB JOURNAL
ISSN: 0976-3104

Pre-processing

Using the median filter, the noise at the intersection of the MRI image is reduced. After that, the filtered image is smoothed by Gauss filter. Then, the structural characters of the filtered image are revealed using top hat and bottom hat methods. Recall that the top hat transformation corresponds to the difference between the image and the state in which the opening operation is performed, and mathematically, defined as

$$\text{top}-\text{hat}(f,b) = f - (f \circ b) \qquad (1)$$

where $(f \circ b)$ is an opening process according to the structure b of the image f. top hat sharpens the peak values (bright areas in the image) in the image. On the other hand, bottom-hat is defined as the residual of a closing compared to the original image, mathematically,

$$\text{bottom}-\text{hat}(f,b) = (f \bullet b) - f \qquad (2)$$

Thanks to the transformation, the lower gray levels, that is, dark areas are detected. The pit areas become clear. Due to these features, top hat and bottom hat transformations help to separate the anomalies from other areas on the image [22].

Briefly, to improve contrast in the preprocessed morphotype process, the filtered image is added to the top hat filtered image and then subtracted from the bottom hat filtered image.

Segmentation of tumor using expectation maximization algorithm

Image segmentation is one of the essential steps in multi-dimensional signal processing. The purpose of the segmentation process is to cluster the intersection of the MR image pixels into salient image regions. Here, we use the Expectation Maximization (EM) method for segmenting the ROI in order to account for the spatial dependencies among pixels.

The EM algorithm [19] is an efficient iterative procedure to compute the maximum likelihood estimate in the presence of missing or unobserved latent data. The EM is perhaps most often used algorithm for unsupervised clustering.

Each iteration of the EM algorithm consists of two processes: the expectation (E-step) and the maximum likelihood (M-step). In the E-step, the missing data are estimated given the observed data and the current estimate of the model parameters. This is achieved using the conditional expectation, explaining the choice of terminology. In the M-step, the likelihood function is maximized under the assumption that the missing data are known. The estimate of the missing data from the E-step is used instead of the actual missing data. Convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration.

Processing the segmented tumor image based on 2D discrete wavelet transform

The input image represents the sum of the useful image and the noise image. Because these two random signals are not correlated, the correlation of the wavelet coefficients of the input image is the sum of the correlations of the wavelet coefficients of the useful image and of the noise image. Thus, wavelets let us analyze multi-resolution of images in detail, namely, it allows the signal to be described from the coarse level to the finest level [23, 24]. It is also useful for removing noise.

Wavelet transform plays an important role in image processing because wavelets allow both time and frequency analysis. The continuous wavelet transform of the signal $f(x)$ is defined as

$$W_{\psi}(s,\tau) = \int_{-\infty}^{\infty} f(x)\psi_{s,\tau}(x)dx \qquad (3)$$

with

$$\psi_{s,\tau}(x) = \frac{1}{\sqrt{s}}\psi\left(\frac{x-\tau}{s}\right) \qquad (4)$$

where ψ is mother wavelet, s is scaling, τ is translation. Note that Equation 3 transforms a continuous function of one variable into a continuous function of two variables.

On the other hand, in digital signal processing, a signal is represented by a discrete sequence. Thus, the discrete wavelet transform can be utilized to process it. Namely, a discrete signal $f(n), n = 0,1,2,\cdots,m$ can be represented as a weighted sum of wavelets $\psi(n)$ and coarse approximation $\vartheta(n)$, mathematically,

$$f(n) = \frac{1}{\sqrt{m}}\sum_{k} W_{\vartheta}(j_0,k)\vartheta_{j_0,k}(n) + \frac{1}{\sqrt{m}}\sum_{j=j_0}^{\infty}\sum_{k} w_{\psi}(j,k)\psi_{j,k}(n) \qquad (5)$$

IIOAB JOURNAL
ISSN: 0976-3104

where j_0 is an arbitrary starting scale. In Equation 5

$$W_\psi(j,k) = \frac{1}{\sqrt{m}} \sum_{n=0}^{m} f(n) \psi_{j,k}(n) \quad (6)$$

with

$$\psi_{j,k}(n) = 2^{j/2} \psi(2^j n - k) \quad \text{and} \quad W_g(j_0,k) = \frac{1}{\sqrt{m}} \sum_{n=0}^{m} f(n) \vartheta_{j_0,k}(n) \quad (7)$$

See [23] for more details and applications of wavelet transform.

Feature extraction

Feature selection is the final processing step where feature descriptors are used to quantify characteristics of the ROI. Here we use two sets of feature families: intensity-based statistical and texture matrix-based features.

Intensity-based statistical features: The intensity-based statistical features describe how grey levels within the ROI are distributed. [Table 3] presents some intensity-based statistical features mentioned in [20]. At the table, X represents the set of N_p voxels included in the ROI and $P_{(i)}$, the first order histogram with N_g discrete intensity levels. Moreover, $p_{(i)}$ represents the normalized first order histogram. The average grey level intensity within the ROI is represented by $\bar{X}$.

Table 3: Intensity based statistical features

Statistics	Formula
Entropy	$-\sum_{i=1}^{N_g} p_{(i)} \log_2(p_{(i)} + \zeta), \zeta \approx 2.2x10^{-16}$
Mean Deviation	$\frac{1}{N_p} \sum_{i=1}^{N_p} \lvert X_{(i)} - \bar{X} \rvert$
Standard Deviation	$\sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} (X_{(i)} - \bar{X})^2}$
Skewness	$\frac{\frac{1}{N_p} \sum_{i=1}^{N_p} (X_{(i)} - \bar{X})^3}{(\sqrt{\frac{1}{N_p} \sum_{i=1}^{N_p} (X_{(i)} - \bar{X})^2})^3}$
Kurtosis	$\frac{\frac{1}{N_p} \sum_{i=1}^{N_p} (X_{(i)} - \bar{X})^4}{(\frac{1}{N_p} \sum_{i=1}^{N_p} (X_{(i)} - \bar{X})^2)^2}$

Gray level co-occurrence matrix (GLCM) based features: A statistical approach of examining texture that considers the spatial relationship of pixels (i.e. regular repetition of an element, pattern on an image) is the gray-level co-occurrence matrix (GLCM). The GLCM functions characterize the texture of an image by calculating how often pairs of pixel with specific values and in a specified spatial relationship occur in an image, creating a GLCM, and then extracting statistical measures from this matrix. The process is known as GLCM based texture analysis and gives information on the disposition of the structures and their relations with the environment [20, 25].

For a GLCM matrix of size $N_g \times N_g$, let $P_{(i,j)}$ be the co-occurrence matrix. Also, let $p_{(i,j)}$ be the normalized co-occurrence matrix. [Table 4] shows some texture features and their explanations [20].

Table 4: Texture based statistical features

Statistics	Formula
Energy	$\sum_{i}^{N_g} \sum_{j}^{N_g} p_{(i,j)}^2$
Contrast	$\sum_{i}^{N_g} \sum_{j}^{N_g} (i-j)^2 p_{(i,j)}$
Homogeneity	$\sum_{i}^{N_g} \sum_{j}^{N_g} \frac{(i-\mu_i)(j-\mu_j) p_{(i,j)}}{\sigma_i \sigma_j}$
Correlation	$\sum_{i}^{N_g} \sum_{j}^{N_g} \frac{p_{(i,j)}}{1+\lvert i-j \rvert}$

IIOAB JOURNAL
ISSN: 0976-3104

Feature selection using the random forest algorithm

Random Forests (RFs) are an elegant method for feature ranking due to their relatively good accuracy, robustness and ease of use. RF consists of a number of decision trees. Every node in the decision trees is a condition on a single feature. Here, the Gini feature importance of the RF is employed that allows for an explicit feature elimination [26]. In other words, at each node within the decision trees of the random forest, the optimal split is sought using the Gini impurity measuring how well a potential split is separating the samples of the two classes in this particular node. With correlated features, strong features can end up with low scores and the method can be biased towards variables with many categories. For the mathematical formulation of the model and its analysis, we refer readers to see [27].

Classification algorithms

Image classification is based on the assumption that the image pixels depict one or more features and that each of these features belongs to one of the several distinct and exclusive classes.

We used the features described in [Tables 3 and 4]. Here we employed three different classification methods (Naive Bayes, linear discriminant analysis, C4.5 decision tree algorithm) to classify the ROI and compared their performance.

Bayesian classifiers are among the statistical classification techniques They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class.

Bayesian classifier is based on Bayes theorem, which says, for a set of classes $\Omega=\{\omega_1,\omega_2,\cdots,\omega_m\}$,

$$P(\omega_j \mid d)=\frac{P(d \mid \omega_j)P(\omega_j)}{P(d)} \tag{8}$$

where $P(\omega_j \mid d)$ is probability of instance d being in class ω_j , called posterior probability, $P(d \mid \omega_j)$ is probability of generating instance d given class ω_j , called conditional probability}, $P(\omega_j)$ is probability of occurrence of class ω_j , and $P(d)$ is probability of instance d occurring. In fact, Equation 8 is considered as Bayes classification when there is one attribute. We consider that there are more than one attributes, say n attributes. To simply task we assume that instances are independent and identically distributed. Then, given an n dimensional attribute vector $x=(x_1,x_2,\cdots,x_n)$, the class conditional probability is computed as:

$$P(x \mid \omega_j)=P(x_1 \mid \omega_j)P(x_2 \mid \omega_j)\cdots P(x_n \mid \omega_j) \tag{9}$$

and the generalization of Equation 8 is

$$P(\omega_j \mid x)=\frac{P(x \mid \omega_j)P(\omega_j)}{P(x)} \tag{10}$$

Naive bayes classification is the most likely or maximum posterior class that solves the optimization problem:

$$\arg_{\omega\in\Omega}\max P(\omega \mid \mathrm{x}) \tag{11}$$

Note that, in solving the optimization problem, we disregard the denominator $P(x)$ in Equation 10.

Naive Bayesian classifier assumes that the features in a dataset are mutually independent. In practice, the independence assumption is often violated, but naive Bayes classifiers still tend to perform very well under this unrealistic assumption [28].

Linear Discriminant Analysis (LDA), developed in 1936 by R. A. Fisher [29], searches for a linear combination of variables that best separates two or more classes. Suppose that the class set $\Omega=\{\omega_1,\omega_2\}$ has two classes and we need to discriminate between them. Then, we classify d d dimensional sample $\phi=\{x_1,x_2,\cdots,x_d\}$, let

$$Z=\beta_1 x_1+\beta_2 x_2+\cdots+\beta_d x_d \tag{12}$$

where $\beta_1,\beta_2,\cdots,\beta_d$ are unknowns and called predictors. The main objective of LDA is to find a linear combination of variables, as in Equation 12, that gives the maximum class separability. To reach this, Fisher defined the following score function

IIOAB JOURNAL
ISSN: 0976-3104

$$S(\beta) = \frac{\beta^T \mu_1 - \beta^T \mu_2}{\beta^T C \beta} \quad (13)$$

where $\beta = (\beta_1, \beta_2, \cdots, \beta_d)$ is the linear model coefficients vector, μ_1 and μ_2 are the mean vectors of the classes ω_1 and ω_2, respectively, and C is the pooled within group covariance matrix. In this method, we need to estimate the linear coefficients that maximize the score function to maximize the score function to maximize the discrimination between two considered classes. One can compute β and C as

$$\beta = C^{-1}(\mu_1 - \mu_2) \quad (14)$$

$$C = \frac{1}{d_1 + d_2}(d_1 \operatorname{cov}(\omega_1) + d_2 \operatorname{cov}(\omega_2)) \quad (15)$$

where $\operatorname{cov}(\mu_i)$ is covariance matrix for the class i and d_i is the number of observations in the class.

C4.5 is arguably the most popular algorithm used to generate DTs. C4.5 owes its popularity to solving various problems, such as continuous attributes and missing attribute values [30]. It is more robust in the presence of noise.

C4.5 uses gain ratio, evaluated by entropy, as an attribute selection measure to build a decision tree. Let S be the set of samples with n classes. Then the entropy of S is defines as

$$E(S) = -\sum_{i=1}^{n} p_i \log p_i \quad (16)$$

where p_i is the proportion of samples in S that belong to the i th class. When an attribute A splits the set S into subsets S_i, the average entropy (a.k.a information) is defined as

$$I(S, A) = \sum_i \frac{|S_i|}{|S|} E(S_i) \quad (17)$$

and information gain for the attribute A as

$$Gain(S, A) = E(S) - I(S, A) \quad (18)$$

Moreover, intrinsic information of the split A is

$$InstI(S, A) = \sum_i \frac{|S_i|}{|S|} \log\left(\frac{|S_i|}{|S|}\right) \quad (19)$$

and gain ratio

$$GA(S, A) = \frac{Gain(S, A)}{InstI(S, A)} \quad (20)$$

At each node in the DT, the attribute with the highest information gain ratio is chosen.

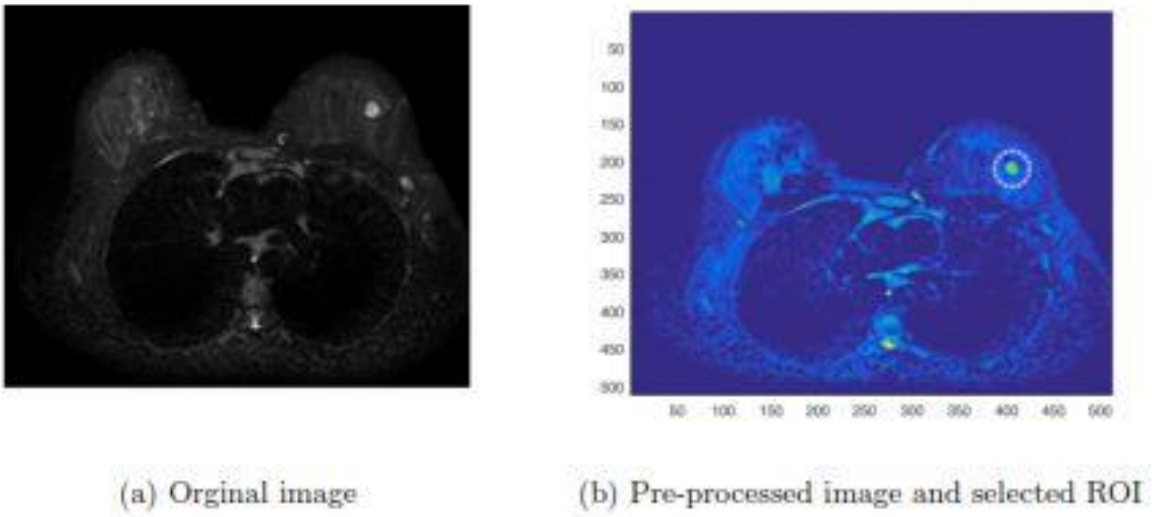
RESULTS AND DISCUSSION

The CAD is implemented in the Python environment. We tested the CAD on the dataset containing 105 patients MRI images, among them, 53 are benign and 52 malign, These MRI images are obtained from local hospital in Turkey.

Data exploring

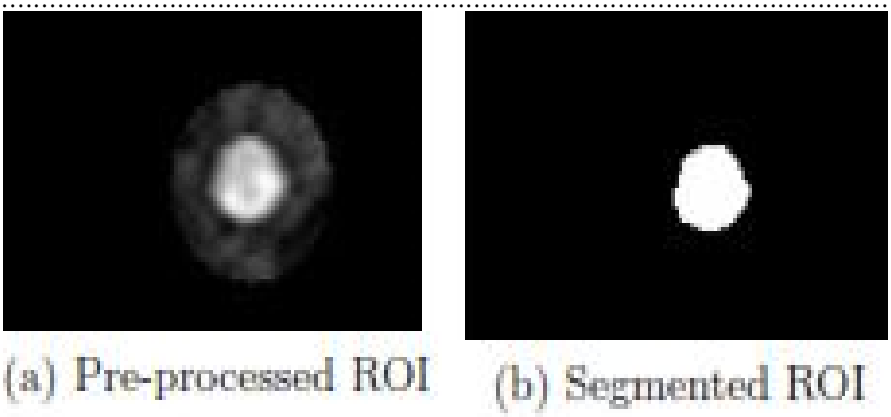
Intersection of MRI image at [Fig. 1(a)] is pre-processed to obtain [Fig. 1(b)]. During the pre-processing procedure, the original image is applied to median filter for reducing the noise, and then gaussian filter for smoothing the image. After that, top hat and bottom hat operations are performed to the resulting image. Then, ROI, ellipse region in [Fig. 1(b)], is selected. Later, the ROI is pre-processed to obtain [Fig. 2(a)] then, segmented via EM clustering technique to obtain [Fig. 2(b)]. The segmented images, then, is applied to 2D-discrete wavelet transform before texture analysis. [Fig. 3] presents the texture analysis results.

IIOAB JOURNAL
ISSN: 0976-3104



(a) Orginal image (b) Pre-processed image and selected ROI

Fig. 1: Original image (a) is processed to obtain (b) and the ROI is selected.



(a) Pre-processed ROI (b) Segmented ROI

Fig. 2: Pre-processed ROI (a) is segmented to obtain (b).

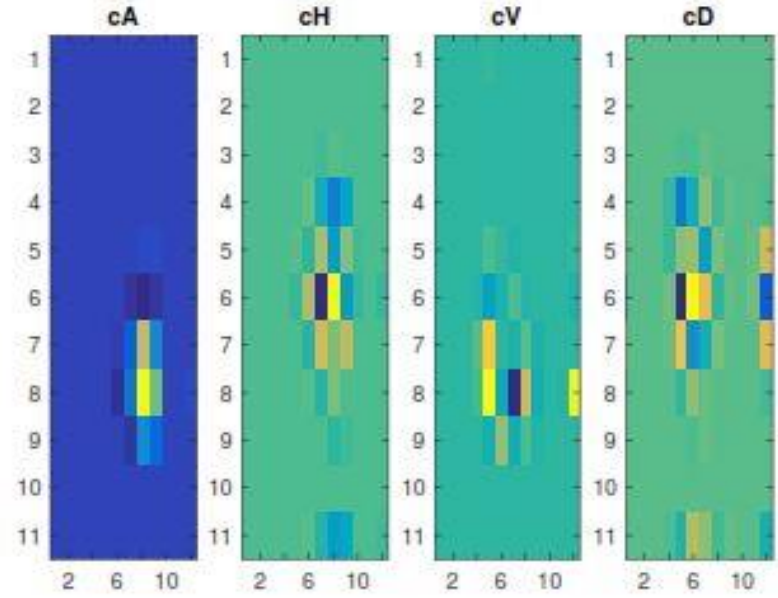


Fig. 3: Wavelet texture of segmented ROI by EM.

[Table 5] presents quantitative values of the features described in [Tables 3 and 4] in the dataset.

IIOAB JOURNAL
ISSN: 0976-3104

Table 5: Statistic of features in the dataset

	Contrast	Correlation	Energy	Homogeneity	Mean	Std_Dev	Entropy	Kurtosis	Skewness
mean	1.090130	0.111163	0.540387	0.844589	0.014990	0.156886	3.732237	6.300306	0.856687
std	0.304420	0.076212	0.086720	0.033901	0.007165	0.012822	0.341759	1.938972	0.299210
min	0.579670	-0.052528	0.342828	0.756250	-0.006134	0.132644	2.986560	3.483746	0.255409
max	1.772321	0.269580	0.683801	0.892634	0.031909	0.176936	4.237545	12.198865	1.605046

[Fig. 4] shows violin plot that is used to visualize the distribution of the data and the probability density. According to the figure, the median of contrast, homogeneity, energy and standard deviation might give good results in the classification. However, the median value of the correlation and mean properties may not give good results in classification.

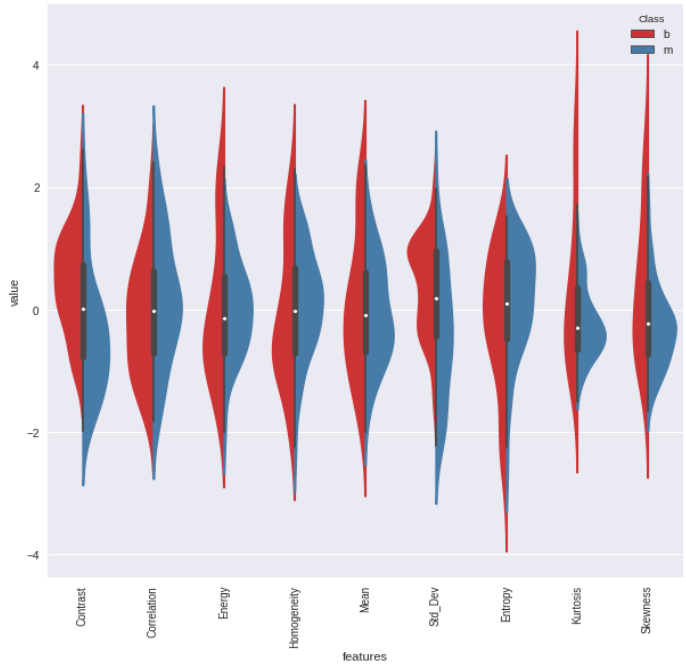


Fig. 4: Violin Plot for data visualization.

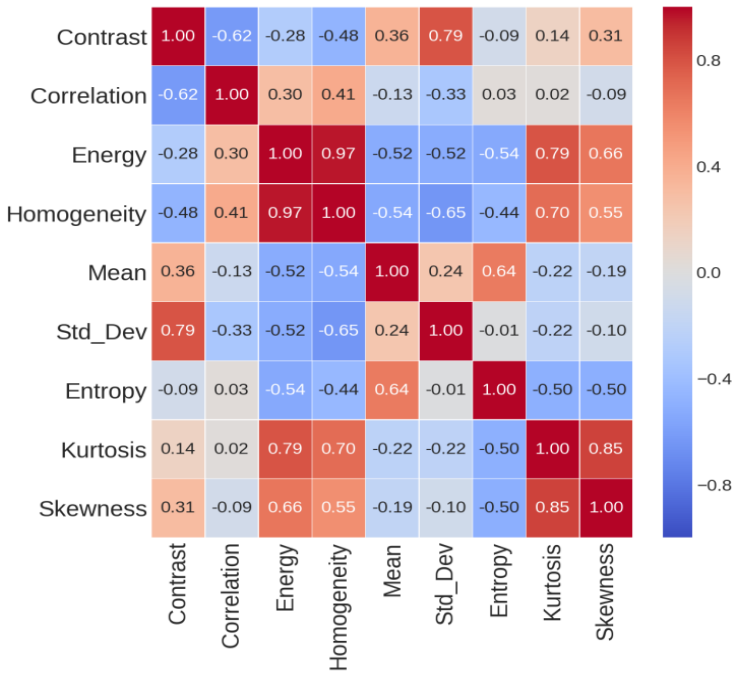


Fig. 5: Correlation matrix.

IIOAB JOURNAL
ISSN: 0976-3104

[Fig. 5] shows heat map for correlation. Correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and -1. A value of ±1 indicates a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. The direction of the relationship is indicated by the sign of the coefficient; a + sign indicates a positive relationship and a – sign indicates a negative relationship. According to the figure, Contrast, Standard Deviation, Energy, Homogeneity, Kurtosis, Skewness data are correlated. These features may be useful in classification

In random forest classification method there is a feature importance's attributes. [Fig. 6] shows feature importance's. The green bars are the feature importance's of the forest, along with their inter-trees variability. The plot shows that 4 features are more informative and the rest are less informative.

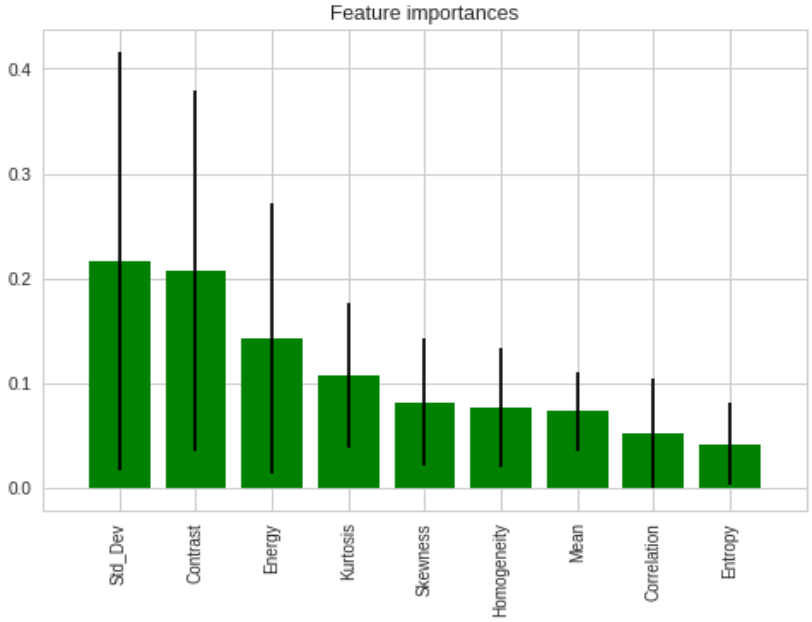


Fig. 6: Feature Importance's.

Classification Performance

A model that would just repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet. To avoid it, 84 patients' images are allocated for training and 21 patients' images reserved for testing. Three different classification algorithms (Naive Bayes, the LDA and C4.5) are fed separately with these features. [Fig. 7] shows confusion matrix. According to the figure, C4.5 Decision Tree Algorithm classified 20 patients correctly with a success rate of 95.24%. Only one patient is misclassified. Naive Bayes Algorithm classified 19 patients correctly with a success rate of 90.48%. LDA Algorithm classified 18 patients correctly with a success rate of 85.71%.

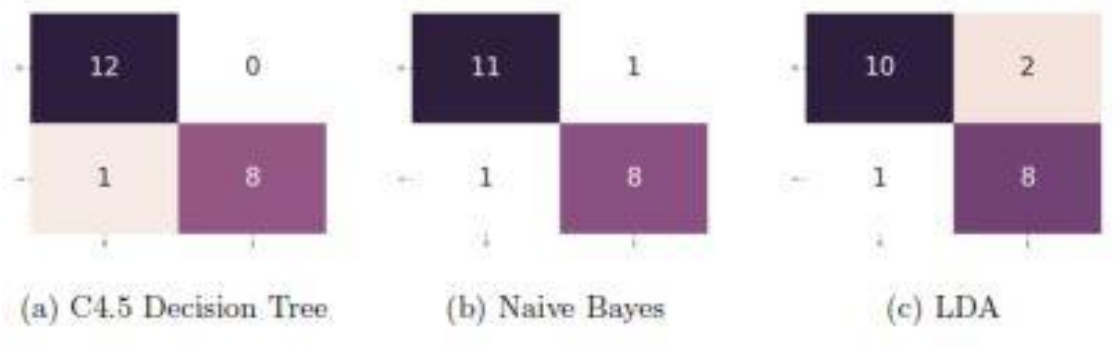


Fig. 7: Confusion Matrix.

The features selected in the random forest feature selection are used to train the C4.5 Decision Tree, Naive Bayes and LDA classifiers by 5-fold cross validation method. Using some metrics such that precision, recall and F-beta score, we compare the performance of the three algorithms. These metrics of the

IIOAB JOURNAL
ISSN: 0976-3104

classifiers are tabulated in [Table 6]. Observe that C4.5 Decision Tree Algorithm has better classification accuracy than the others.

Table 6: Classification Report

	C 4.5 Decision Tree	Naive Bayes	LDA
Precision	1.000	0.889	0.800
Recall	0.889	0.889	0.889
F1 Score	0.941	0.829	0.842
Mean Auc	0.800	0.790	0.600
Accuracy	0.9524	0.9048	0.8571

As we mentioned above, in the CAD, there are four crucial modules: preprocessing, segmentation, feature extraction and selection, and classification. The preprocessing module prepares the images for the subsequent stages and responsible for cleaning the medical image and removing noise from it through a set of image preprocessing operations. The segmentation module, in fact, is a procedure of dividing the input image into several regions according to the visual characteristics. In the features extraction and selection module, features are extracted from the cleaned images, then the most discriminative features are selected. The selected features are capable of differentiating between normal and cancerous regions in order to minimize the classification error. Despite the large effort, there is still no agreement on the features that are most suitable for the target task. Many kinds of features such as dynamic features, textural features, and morphological features have been traditionally used in tumor classification. Along with the segmentation, the classification module is regarded as the heart of the CAD.

Karaçavuş et al. [31] used otsu segmentation and the KNN classifier in the TNM staging system they developed and achieved a result of 84% accuracy. In their study, they extracted ten textual features and among them they used five features in classification of tumor stage. Waugh et al. [32] achieved 74.7% accuracy using KNN classification. On the other side, Cai et al. [33] analyzed the effects of 28 features in four different classes to the dataset collected under different imaging protocols, in result, they obtained that instead of using all features only five features with the highest diagnostic affect rates have the highest classification accuracy of 82.8% using SVM. They incorporated fuzzy c-means clustering and a gradient vector flow snake algorithm to segment the ROI.

The first time in the literature in breast MRI, we used the EM method to segment ROI which is the crutial step in the CAD. Later, carefully selected total nine features, among them five from intensity-based statistical and four from texture matrix-based features, fed to a classifier, in particular, C4.5 decision tree classification.

CONCLUSION

In this paper, we presented a novel user-independent time-saving CAD to diagnose breast cancer tumor using MRI images. The CAD is composed of two main stages: (1) the analysis stage (2) the diagnosis stage. On the other side, the analysis stage consists of series of procedures such that pre-processing, segmentation and feature extraction. The CAD uses the EM for segmenting the ROI. Later, some intensity-based statistical and texture matrix-based features are obtained. Finally, using three different methods such that C4.5 Decision Tree, Naive Bayes and LDA, the ROI is classified.

Experimental results indicate that the EM segmentation and C4.5 classifications, having accuracy 95.24%, successfully distinguish the tumor's region from the other regions.

CONFLICT OF INTEREST
There is no conflict of interest.

ACKNOWLEDGEMENTS
None

FINANCIAL DISCLOSURE
None

REFERENCES

[1] Ozmen V. [2008] Breast cancer in the world and Turkey. J Breast Health. 4(2):6-12.
[2] Aydıntuğ S. [2004] Meme kanserinde erken tanı. Sted. 13(6):226-229.
[3] Cady B, Michaelson JS. [2001] The life-sparing potential of mammographic screening. Cancer: Interdisciplinary International Journal of the American Cancer Society. 91(9):1699-1703.
[4] Smith RA, Cokkinides V, von Eschenbach AC, et al. [2002] American Cancer Society guidelines for the early detection of cancer. CA: a cancer journal for clinicians. 52(1):8-22.
[5] Feig SA, D'Orsi CJ, Hendrick RE, et al. [1998] American College of Radiology guidelines for breast cancer screening. AJR. American journal of roentgenology. 171(1):29-33.
[6] Berg WA. [2010] Benefits of screening mammography. Jama. 303(2):168-169.
[7] Sickles EA. [1991] Screening for breast cancer with mammography. Clinical imaging. 15(4):253-260.

IIOAB JOURNAL
ISSN: 0976-3104

[8] Adler DD, Helvie, MA. [1992] Mammographic biopsy recommendations. Current Opinion in Radiology. 4(5):123-129.
[9] Sickles EA. [1991] Periodic mammographic follow-up of probably benign lesions: results in 3,184 consecutive cases. Radiology. 179(2):463-468.
[10] Prince JL, Links JM. [2006] Medical imaging signals and systems. Upper Saddle River, NJ: Pearson Prentice Hall.
[11] Wollins DS, Somerfield MR. [2008] Q and a: magnetic resonance imaging in the detection and evaluation of breast cancer. Journal of oncology practice. 4(1):18-23.
[12] Kuhl C. [2007] The current status of breast MR imaging part I. Choice of technique, image interpretation, diagnostic accuracy, and transfer to clinical practice. Radiology. 244(2):356-378.
[13] Deurloo EE, Muller SH, Peterse JL, Besnard AP, Gilhuijs KG. [2005] Clinically and mammo graphically occult breast lesions on MR images: potential effect of computerized assessment on clinical reading. Radiology. 234(3):693-701.
[14] Nattkemper TW, Arnrich B, Lichte O, et al. [2005] Evaluation of radiological features for breast tumour classification in clinical screening with machine learning methods. Artificial intelligence in medicine. 34(2):129-139.
[15] Hadjiiski L, Sahiner B, Chan HP. [2006] Advances in CAD for diagnosis of breast cancer. Current opinion in obstetrics & gynecology. 18(1):64.
[16] Levman JE, Warner E, Causer P, Martel AL. [2014] A vector machine formulation with application to the computer-aided diagnosis of breast cancer from DCE-MRI screening examinations. Journal of digital imaging. 27(1):145-151.
[17] Shi J, Sahiner B, Chan HP, et al. [2009] Treatment response assessment of breast masses on dynamic contrast-enhanced magnetic resonance scans using fuzzy c-means clustering and level set segmentation. Medical physics. 36(11):5052-5063.
[18] Yassin NI, Omran S, El Houby EM, Allam H. [2017] Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: a systematic review. Computer methods and programs in biomedicine.
[19] Dempster AP, Laird NM, Rubin DB. [1977] Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological). 1-38.
[20] Zwanenburg A, Leger S, Vallières M, Löck S. [2016] Image biomarker standardization initiative. arXiv preprint arXiv:1612.07003.
[21] Haralick RM, Shanmugam K. [1973] Textural features for image classification. IEEE Transactions on systems, man, and cybernetics. (6):610-621.
[22] Burçin KURT, NABİYEV VV. [2010] Dijital mamografi görüntülerinin kontrast sınırlı adapt if histogram eşitleme ile iyileştirilmesi.
[23] Debnath L, Shah FA. [2002] Wavelet transforms and their applications. Boston: Birkhäuser. 2-14.
[24] Stollnitz EJ, DeRose AD, Salesin DH. [1995] Wavelets for computer graphics: a primer. 1. IEEE Computer Graphics and Applications. 15(3):76-84.
[25] Pathak B, Barooah D. [2013] Texture analysis based on the gray-level co-occurrence matrix considering possible orientations. International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering. 2(9):4206-4212.
[26] Menze BH, Kelm BM, Masuch R, et al. [2009] A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC bioinformatics. 10(1):213.
[27] Biau G. [2012] Analysis of a random forests model. Journal of Machine Learning Research. 13:1063-1095.
[28] Rish I. [2001] An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence. New York: IBM. 3(22):41-46.
[29] Fisher RA. [1936] The use of multiple measurements in taxonomic problems. Annals of eugenics. 7(2):179-188.
[30] Quinlan JR. [2014] C4. 5: programs for machine learning. Elsevier.
[31] Karacavus S, Yılmaz B, Tasdemir A, et al. [2018] Can Laws Be a Potential PET Image Texture Analysis Approach for Evaluation of Tumor Heterogeneity and Histopathological Characteristics in NSCLC? Journal of digital imaging. 31(2):210-223.
[32] Waugh SA, Purdie CA, Jordan LB, et al. [2016] Magnetic resonance imaging texture analysis classification of primary breast cancer. European radiology. 26(2):322-330.
[33] Cai H, Liu L, Peng Y, Wu Y, Li L. [2014] Diagnostic assessment by dynamic contrast-enhanced and diffusion-weighted magnetic resonance in differentiation of breast lesions. BMC Cancer.14:366.