Available online at www.dergipark.org.tr/en

INTERNATIONAL ADVANCED RESEARCHES
and
ENGINEERING JOURNAL

Journal homepage: www.dergipark.org.tr/en/pub/iarej

e-ISSN: 2618-575X

International Open Access

Volume 05
Issue 02

August, 2021

Research Article

A new region-of-interest (ROI) detection method using the chan-vese algorithm for lung nodule classification

Ali Cinar [a,*], ***Bengisu Topuz*** [b] ***and Semih Ergin*** [c]

[a]*Dept. of Electrical and Electronics Engineering, Kastamonu University, Kastamonu, Turkey*
[b]*Bioengineering Division, Institute of Science, Hacettepe University, Ankara, Turkey*
[c]*Dept. of Electrical and Electronics Engineering, Eskisehir Osmangazi University, Eskisehir, Turkey*

ARTICLE INFO

Article history:
Received 12 January 2021
Revised 17 April 2021
Accepted 24 June 2021

Keywords:
Chest x-ray classification
Nodule classification
Nodule detection
Rib detection
Rib suppression
ROI detection

ABSTRACT

Suspicious regions in chest x-rays are detected automatically, and these regions are classified into three types, including “malignant nodule”, “benign nodule”, and “no nodule” in this study. Firstly, the areas except the lung tissues are removed in each chest x-ray using the thresholding method. Then, Poisson noise was removed from the images by applying the gradient filter. Ribs may overlap onto nodules. Since this circumstance will make the detection of a nodule difficult, it is necessary to distinguish and suppress the ribs. The location of the rib bones is determined by a template matching method, and then the corresponding bones are suppressed by applying the Gabor filter. After this stage, suspicious tissues in the chest x-rays are specified using the Chan-Vese active contour without edges. Then, some features are extracted from these suspicious regions. Six different features are extracted: Statistical, Histogram of Oriented Gradients (HOG)-based, Local Binary Pattern (LBP)-based, Geometrical, Gray Level Co-Occurrence Matrix (GLCM) Texture-based and Dense Scale Invariant Feature Transform (DSIFT)-based. Then, the classification stage is achieved using these features. The best classification result is obtained using statistical, LBP-based, and HOG-Based features. The classification results are evaluated with sensitivity, accuracy, and specificity analyses. K-Nearest Neighbour (KNN), Decision Tree (DT), Random Forest (RF), Logistic Linear Classifier (LLC), Support Vector Machines (SVM), Fisher’s Linear Discriminant Analysis (FLDA), and Naive Bayes (NB) methods are used for the classification purpose separately. The random forest classifier gives the best results with 57% sensitivity, 66% accuracy, 81% specificity values.

1. Introduction

Lung cancer is one of the most wide spread causes of death in the World [1]. Structurally, it is based on an uncontrolled proliferation of cells from normal lung tissue to form a mass within the lung. This nodule structure on the lung is often noticed during radiological imaging of the lung rather than as a symptom. The shape and structure of the nodule roughly indicate whether the nodule may be benign or malignant. However, for definitive diagnosis, extract of the nodule and pathological examinations are essential.

Today, different imaging methods (x-ray, tomography, MRI, etc.) are utilized depending on the structure of the nodules and their location in the tissue. However, as this nodule structure varies from patient to patient in general and is difficult to detect, new approaches are needed in imaging systems. As with other types of cancer, early diagnosis is critical to overcome lung cancer. The X-ray imaging system is one of the oldest and widely used diagnostic systems in the world. In the images obtained from this system, the bones appear black, and the soft tissue appears lighter. Thanks to this contrast difference, the nodule within the bones and soft tissue can be easily distinguished. In addition, the low cost and easy access to x-ray, less radiation, and allowing more frequent follow-up of nodule development make it more preferred than other imaging systems [2].

In the literature, there are many works on lung-based illnesses detection and classification. Abbas et al. developed

* Corresponding author. Tel.: +90-366-280-2932.
E-mail addresses: alicinar@outlook.com.tr *(A.Cinar)*, bengisutopuz@gmail.com *(B.Topuz)*, sergin@ogu.edu.tr *(S.Ergin)*
ORCID: 0000-0002-3947-7848 *(A.Cinar)*, 0000-0001-6698-5154 *(B.Topuz)*, 0000-0002-7470-8488 *(S.Ergin)*

DOI: 10.35860/iarej.857579

a computer-aided diagnosis (CAD) system for detecting lung nodule cancer earlier. They enhanced the contrast of images using contrast limited adaptive histogram equalization. The segmentation process was done using Otsu’s thresholding method. Then, some morphological filters were used to remove background and other geometrical objects. After that, denoised images were obtained using Discrete Wavelet Transform. GLCM was used for feature extraction, and some features were extracted, such as correlation, energy, etc. Principal component analysis was used for feature selection. SVM was used for the classification of benign or malignant tumors. Classification results were given in terms of accuracy, specificity, sensitivity, peak signal to noise ratio, and root mean square errors [3]. Parveen and Khan developed a CAD system for the detection and classification of pneumonia in chest x-ray images. Features are extracted using the HOG technique. SVM, DT and RF were used as classifiers. Classification results were given in terms of accuracy, recall, precision, and F1 score [4]. Gonzalez and Ponomaryov developed a CAD system for lung cancer detection. Firstly, the background tissue was eliminated using the thresholding technique and morphological operations. Then, suspicious regions were calculated using priori information and Hounsfield Units. After that, shape and textural features were extracted. Shape features include area, eccentricity, circularity, and fractal dimension. Texture features include mean, variance, energy, entropy, inverse difference moment, kurtosis, skewness, contrast, smoothness, and correlation. SVM was used as a classifier. Classification results were given in terms of sensitivity, specificity, and accuracy [5]. Tun and Khaing developed a CAD system for lung cancer detection and classification. Median filter was used for preprocessing. Otsu's thresholding method was used for segmentation. Physical dimension measures and the GLCM method were used for feature extraction. Physical dimension measures include area, perimeter, and eccentricity. GLCM-based features are entropy, contrast, correlation, energy, and homogeneity. The artificial neural network was used for classification [6].

The purpose of this study is to detect and classify the different tissues in chest x-rays. At the end of the study, chest x-rays are classified as benign, malignant, and with no nodules. In this context, the database prepared by the Japanese Radiological Technology Association was used [7]. This dataset was chosen because it is free and a public dataset. Moreover, this dataset was used in many works, and thus it gives the advantage to compare with the other works' results.

This study includes 7 sections. In section 2, preprocessing of chest x-ray images is expressed. Detection of suspicious regions is explained in section 3, and feature extraction methods are declared in section 4. In section 5, classification results are given. Conclusions are given in section 6, and discussion is given in section 7. The block diagram shows how the algorithm classifies chest x-ray images in Figure 1.

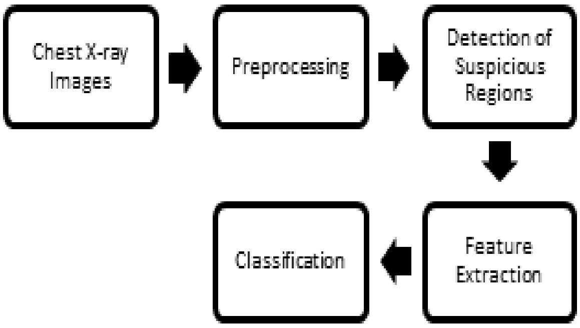


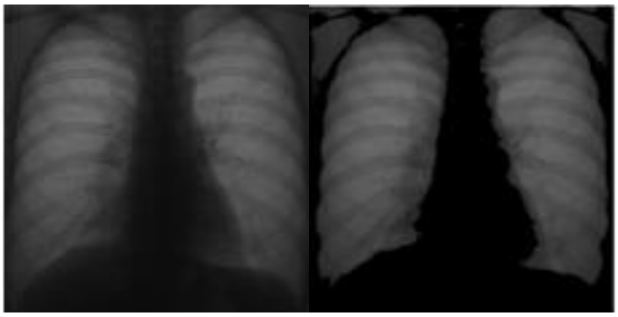
Figure 1. Block Diagram

Nodules appear spherical on chest x-rays. Roughly defined, generally benign nodules have a uniform shape, and malignant nodules have an irregular shape. The greater the diameter of the nodule, the more likely it is to be malignant. In X-rays, ribs may overlap with the nodular region. This is undesirable and may prevent accurate detection of the nodular region. For this reason, the image is preprocessed in order to suppress the ribs. Preprocessing plays a crucial role in image processing and improves the characteristics of the image, and ensures better classification results. After preprocessing, ribs' positions are detected and then suppressed. Subsequently, suspicious regions are specified in the lung and features are extracted from these suspected regions, and the classification stage starts. Different feature extraction methods are tried, and it is seen that which features express benign nodule, malignant nodule, or non-nodule region well according to classification results. Many different classifiers are employed in the classification stage, and it is seen which classifier makes the best classification. Thus, a new region-of-interest (ROI) detection method using the chan-vese algorithm is developed for lung nodule classification. Classification results are given with respect to sensitivity, accuracy and specificity as in the other studies in the literature using the database akin to this study. So as to compare the results obtained with other studies in the literature in a fair manner, only the studies using this database are considered.

2. Preprocessing of Chest X-ray Images

2.1 Segmentation and Denoising of Lung Area

A random chest x-ray is shown in Figure 2a. To segment only lung area, a pixel value 45 was chosen as a threshold value, and all pixel values under 45 changed with 0. Thus, mostly lung area was extracted. However, some non-lung tissues were still visible. A weighted gradient filter was applied to the segmented image to remove Poisson noise, and a gradient image was obtained [8]. Then, a Gaussian weighted filter was applied to the gradient image for smoothing [8]. The segmented denoised image is demonstrated in Figure 2b.



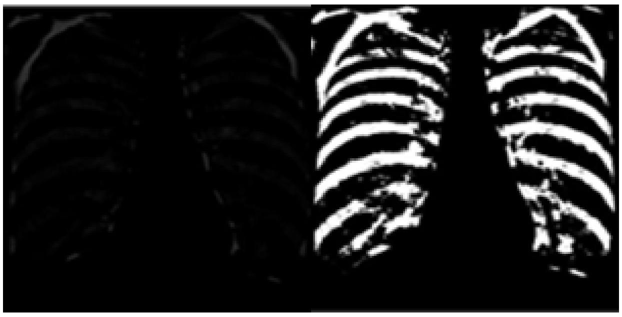
(a) (b)

Figure 2. a) The Original image, b) The Segmented denoised image

2.2 Rib Detection and Suppression

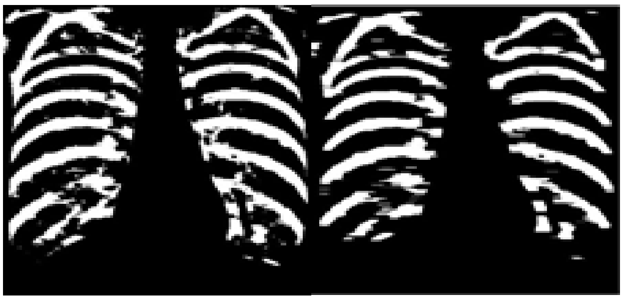
200x200 median filter [9] was used to smooth the segmented denoised image [10]. The difference between median filtered and segmented denoised images is shown in Figure 3a [10]. Then, binarization was applied [10]. A threshold value of 0.0005 was chosen for binarization. The pixels with a value of less than 0.0005 were transformed into 0, whereas the ones with a greater value than 0.0005 were changed into 1 as can be seen in Figure 3b. Thus, only the ribs area was extracted.

By completing the binarization operation, ribs were extracted, but there were tissues connected to the ribs. Then, some morphological operations(i.e., closing and opening) were applied to eliminate non-rib tissues. Firstly, a closing operation was applied. The line-shaped structure element with a radius of 20 pixels was created for this operation, and the gaps got filled as in Figure 4a. Afterward, an opening operation was performed.



(a) (b)

Figure 3. a) The difference between median filtered and the segmented image, b) The binarization applied image



(a) (b)

Figure 4. a) The image after the closing operation, b) The image after the opening operation

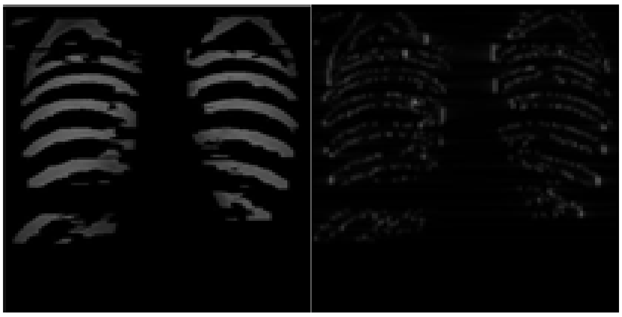
The line-shaped structure element with a radius of 40 pixels was formed for this operation, and the lines having a radius less than 40 pixels were removed, as can be seen in Figure 4b.

There were still some tissues connected to the ribs. To remove these tissues and record ribs' positions, template matching was applied. Totally two templates were utilized. One of the rib bones from the left side of the lung and one of the rib bones from the right side of the lung were used as a template. The rib bone template, which was selected from the left side of the lung, was used to extract rib bones from the left side of the lung, and the rib bone template which was chosen from the right side of the lung was employed to extract the rib bones from right the side of the lung. For the left side of the lung, the template is shown to the image pixel by pixel. A threshold value was determined to detect rib bones. In each iteration, the amount of overlap between the template and the image was examined. If this amount was larger than the threshold value, the rib bones were detected, and their position was recorded. Thus, more smooth rib structures with reduced tissue were obtained. The same operation was applied to the right side of the lung. Later, the recorded rib bones' pixel values were changed with their original value. After template matching, the detected ribs with their original values are shown in Figure 5a.

Detected ribs should be suppressed. Gabor filter is used for that purpose [11]. The Gabor Filter with a 2 wavelength in pixels/cycle and 0° orientation was applied to the image (see Figure 5a) in order to suppress ribs and the magnitude response of the filter was obtained as seen in Figure 5b [12].

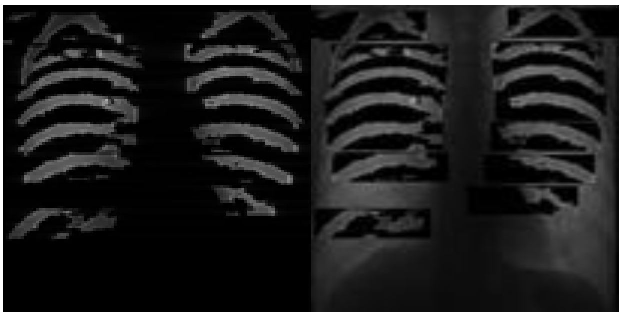
Then, the images in Figure 5a and Figure 5b were summed in the form of a matrix summation. The summation of images is shown in Figure 6a. The new ribs in Figure 6a were replaced with the ribs in the original image which is Figure 2a and Figure 6b illustrates the placement.

The black pixels (non-rib) in Figure 6b were replaced with their original values in Figure 2b, and the new image was obtained in Figure 7a. 3x3 median filter was applied to the image in Figure 7a, and the median filtered image is shown in Figure 7b. Afterwards, an opening operation was applied to the image in Figure 7b. The line-shaped structure element with a radius of 10 pixels is created for this operation.



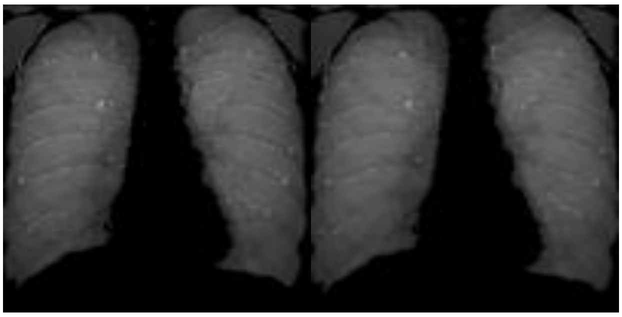
(a) (b)

Figure 5. a) The detected ribs, b) The magnitude response of the filter



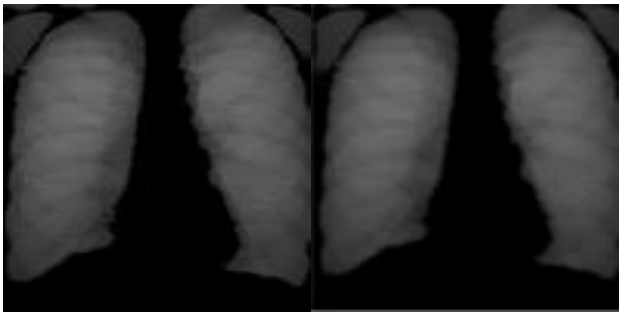
(a) (b)

Figure 6. a) The summation of image matrices, b) The placement of the new ribs to the original image



(a) (b)

Figure 7. a) The replacement of black pixels with their original values, b) The image after median filtered



(a) (b)

Figure 8. a) The image after the opening operation, b) The ribs suppressed image

The lines having a radius less than 10 pixels were eliminated removed as seen in Figure 8a. Finally, Gaussian Filter [13] with a standard deviation of 2 was applied to the image in Figure 13, thus suppressed ribs of the image are shown in Figure 8b [12].

3. Detection of Suspicious Regions in Chest X-ray Images

3.1 The Chan-Vese Algorithm-Based Approach

In order to obtain more accurate results in the classification of nodules, the suspicious areas on the lung were specified, and feature extraction was made in these suspicious areas. In this study, Chan-Vese active contour without edges method was made use of. The Chan-Vese active contour method without edges is based on the segmentation problem formulated by Mumford and Shah [14].

Equation (1) gives the function that minimizes energy compared to c_1, c_2, and C [14]. u_0 represents the image, c_1 can be interpreted as the average value of everything inside the C contour, and c_2 can be interpreted as the average value of everything outside the C contour. Ω_1 refers to the area within the contour, and Ω_2 refers to the area outside the contour.

$$F(c_1,c_2,C) = \int_{\Omega_1=w}(u_o(x,y)-c_1)^2\,dxdy + \int_{\Omega_2=\Omega-w}(u_o(x,y)-c_2)^2dxdy + \mathrm{v}|\mathrm{C}|, C=\partial w, w\subset\Omega \quad (1)$$

Chan and Vese interpreted the first two terms in Equation (2) as two forces in the article [14]. The first term forces the contour to reduce, and the second term forces the contour to expand. These two forces are compensated when the contour reaches the boundary of the object of interest. The logic of the algorithm is shown in Figure 9 in four cases.

$$F(c_1,c_2,C) = \int_{inside(C)}|u_0-c_1|^2\,dx + \int_{outside(C)}|u_0-c_2|^2\,dx = F_1(C)+F_2(C) \quad (2)$$

In the images in Figure 9, the black parts are indicated by -1, and the gray parts are indicated by 1. When the case (a) is examined in Figure 9, the first contour covers the entire object (-1) and some gray regions (+1). Therefore, c_1 is approximately 0, and c_2 is 1. Subtracting c_2 from the image outside the contour yields 0. Thus, the term F_2 is zero. Since c_1 approaches zero, a large positive number is reached when subtracting c_1 from the image remaining in the contour and finding the sum of the squares as shown by the formula. Therefore, F_1> 0. For F_1> 0 and F_2 = 0, the contour will narrow in the next step. The other cases in Figure 9 can be interpreted in this way.

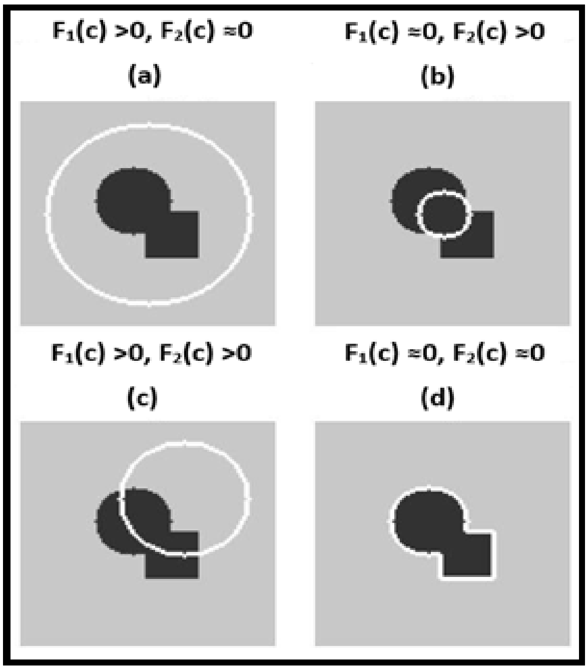


Figure 9. The procedure of the chan-vese algorithm [14]

In the Chan-Vese active contour without edge algorithm, the level set function shows the contour values ϕ (x, y). The mathematical expression of the contour curve C is given in Equation (3).

$$C = \{(x, y): \phi(x, y) = 0\}, \forall(\mathrm{x,y}) \in u_0 \qquad (3)$$

The change of the contour over time according to the level set function is given in Equation (4).

$$\frac{\partial C}{\partial t} = \frac{\partial \phi(x,y)}{\partial t} \qquad (4)$$

4. Feature Extraction Methods

4.1 Statistical Features

Statistical features' success has been explained in many papers. They are extracted from suspicious regions These features are given in Table 1. The feature vector has a dimension that is 12x1 [15]. In Table 1, $\mathrm{x_i}$ represents its sample, and N represents the total sample number.

Table 1. The statistical features and their mathematical representations

Statistical Features	Mathematical Representations
Energy	$\sum_{i=1}^{N} x_i^2$
Mean	$\mu = \frac{1}{N}\sum_{j=1}^{N} x_i$
Variance	$\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2$
Maximum	$max\{x_i, 1 \leq i \leq N\}$
Minimum	$min\{x_i, 1 \leq i \leq N\}$
Standard Deviation	$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^2}$
Skewness	$\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\sigma^3}$
Kurtosis	$\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\sigma^4}$
Area Descriptor [16]	σ / μ
Mean Energy	$\mu_{Energy} = \frac{1}{N}\sum_{i=1}^{N} x_i^2$
Energy Variance	$\frac{1}{N}\sum_{i=1}^{N}(X_i^2 - \mu_{Energy})^2$
Entropy	$\sum_{i=1}^{N} p(X_i) log_2\, p(X_i)$

4.2 Histogram of Oriented Gradients (HOG)-Based Features

HOG-based features yield good results in biomedical pattern recognition problems [17]. HOG-based features are extracted from suspicious regions.

The image gradient vector is defined as a metric for each pixel containing pixel intensity changes on both the x-axis and the y-axis. The definition is given by the gradient of a continuous multivariate function, a vector of partial derivatives of all variables. Suppose that position F (x, y) records the intensity value of the pixel at (x, y), the gradient vector of the pixel (x, y) is defined as in Equation (5).

$$\nabla f(x, y) = \begin{pmatrix} g_x \\ g_y \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \begin{pmatrix} f(x+1,y) - f x-1,y) \\ f(x,y+1) - f(x,y-1) \end{pmatrix} \qquad (5)$$

The expression ∂f / ∂x is a partial derivative in the + x direction, calculated as the difference in intensity between adjacent pixels on the left and right of the target pixel. Similarly, ∂f / ∂y is a partial derivative in the + y direction, calculated as the difference in intensity between adjacent pixels above and below the target pixel. The magnitude is the L2 norm of the vector and is calculated as in Equation (6). The orientation is the arctangent of the ratio of partial derivatives to each other in two directions and is calculated as in Equation (7).

$$\nabla f(x, y) = \sqrt{{g_x}^2 + {g_y}^2} \qquad (6)$$

$$\theta = \arctan\left(\frac{g_y}{g_x}\right) \qquad (7)$$

4.3 Local Binary Pattern (LBP)-Based Features

LBP-based features are extracted from suspicious regions. LBP is plain yet a useful tissue operator that marks the pixels of an image by thresholding the neighbourhood of each pixel and sees the result as a binary number [18]. Due to its discriminatory strength and computational simplicity, the LBP tissue operator has become a popular approach in a variety of applications. The most significant aspect of the LBP operator in actual applications is its robustness to monotonic grayscale changes.

The LBP operator replaces the value of pixels of an image with decimal numbers named LBP codes encoding the local structure around each pixel. Each center pixel is compared with its eight neighbours. The neighbours smaller than the value of the central pixel have bit 0, and the other neighbours greater than or equal to the value of the central pixel have bit 1. For each given central pixel, a binary number is generated by combining all of these binary bits clockwise, starting from one of the upper left neighbours. The resulting decimal value of the generated binary number replaces the central pixel value.

The LBP representation of an image is calculated as in Equation (8). In the equation, R is the radius of the circle, P is the number of pixels in the neighbourhood, u is the unit step function, g_i represents the intensity value of the ith neighbouring pixel and g_c stands for the intensity value of the central pixel in the neighbourhood.

$$LBP(P,R) = \sum_{i=0}^{P-1} u(g_i - g_c) \cdot 2^i \qquad (8)$$

If the binary pattern consists of no more than two 0-1 or 1-0 transitions, the local binary pattern is called uniform. The rotation-independent patterns are reached by rotating each bit circularly to the minimum value. The rotation-independent patterns are calculated as in Equation (9). The expression U in the equation expresses the uniform criterion.

$$LBP^{riu2}(P,R) = \begin{cases} \sum_{i=0}^{P-1} u(g_i - g_c) & U\big(LBP(P,R)\big) \leq 2 \\ P+1 & U\big(LBP(P,R)\big) \leq 2 \end{cases} \qquad (9)$$

4.4 Geometrical Features

Geometrical features are derived from the suspicious regions in the images. The largest interconnected tissue is specified in the suspicious region, and features; orientation, major axis length, minor axis length, eccentricity, solidity, fullness ratio, diameter, convex area, area, roundness, ovality, and perimeter are extracted.

Area refers to the total number of pixels in the suspicious tissue. The convex area is determined by plotting the smallest convex to contain the suspicious tissue and finding the total number of pixels within that convex. The solidity is found by dividing the total number of pixels (area) within the suspicious tissue into the convex area. The fullness ratio is determined by plotting the smallest rectangle to contain the suspicious tissue and dividing the total number of pixels in the suspicious tissue by the total number of pixels in the rectangle. While the major and minor axis lengths are found, an ellipse is drawn to cover the suspicious tissue. The major axis refers to the length of the ellipse's x-axis in pixels. The minor axis refers to the length of the ellipse's y-axis in pixels. The orientation is the angle between the x-axis and the x-axis of the ellipse. Eccentricity is the ratio of the distance between the foci of the ellipse and the major axis length. The perimeter gives the total number of pixels around the suspicious tissue. The diameter is calculated in Equation (10) as the diameter of a circle that has the same area with the suspected tissue.

$$D = \sqrt{\frac{4 * Area}{\Pi}} \qquad (10)$$

Ovality indicates how close the suspicious tissue is to the oval shape and is calculated as in Equation (11) [19].

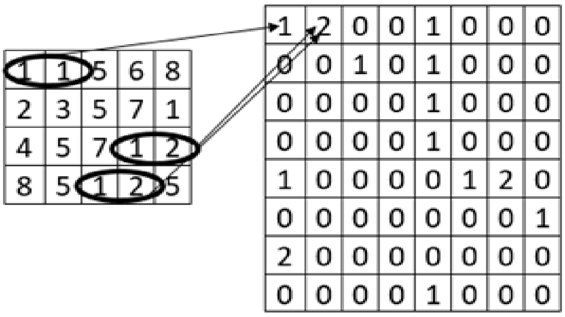
$$O = 2 * \frac{major\ axis\ length - minor\ axis\ length}{major\ axis\ length + minor\ axis\ length} \qquad (11)$$

Roundness is calculated as in Equation (12). r is the major axis length divided by the minor axis length [19].

$$Y = \frac{Area}{\Pi r^2} \qquad (12)$$

4.5 Gray Level Co-Occurrence Matrices (GLCM)-Based Texture Features

GLCM is used for tissue analysis [20]. GLCM, a square matrix, yields certain properties about the spatial distribution of gray levels in the tissue. Each element (i, j) of the GLCM matrix is the number of repetitions of the i and j pixel values at a distance d in Ɵ direction to each other. An example of how the GLCM matrix is generated is given in Figure 10.



1	1	5	6	8
2	3	5	7	1
4	5	7	1	2
8	5	1	2	5

1	2	0	0	1	0	0	0
0	0	1	0	1	0	0	0
0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0
1	0	0	0	0	1	2	0
0	0	0	0	0	0	0	1
2	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0

Figure 10. Creating a GLCM matrix

Figure 10 illustrates how often i and j pixel values repeat for d = 1 and Ɵ = 0 °. When the i and j pixel values are 1, 1 is written to the first column of the first row in the GLCM matrix since there are no elements in the matrix where the other i and j pixels are equal to 1. Likewise, since there are 2 pairs of pixels in the matrix with a pixel value of i = 1 and j = 2, 2 is written to the second column of the first row in the GLCM matrix.

Within the scope of the study [20-22], the features extracted from the GLCM matrices were utilized. Equations used for feature extraction are given in Table 2. Features and their mathematical expressions are given in Table 3. A total of 22 different features are extracted from the GLCM matrix [23]. The features given in the table are calculated using the equations given at the beginning of the table. As seen in Table 2; P(i,j) represents elements of the GLCM matrix, and Ng represents the row or column number of the GLCM matrix.

4.6. Dense Scale Invariant Feature Transform (DSIFT)-Based Features

DSIFT descriptors is a well known method to extract features from lesions [24]. DSIFT performs scale invariant feature transform on non-overlapping image blocks with the given radius and returns a 128-dimensional feature vector for each [24].

Table 2. The required equations

$P(i,j): GLCM = \begin{bmatrix} P(1,1) & \cdots & P(1,N_g) \\ \vdots & \ddots & \vdots \\ P(N_g,1) & \cdots & P(N_g,N_g) \end{bmatrix}$
$P_x(i) = \sum_{j=1}^{N_g} P(i,j) \quad , \quad P_y(i) = \sum_{i=1}^{N_g} P(i,j)$
$P_{x+y}(k) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P_{i+j=k}(i,j)\ , k = 2,3,\dots,2N_g$ $P_{x-y}(k) = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} P_{\lvert i-j\rvert=k}(i,j)\ , k = 0,1,\dots,N_g - 1$
$\mu_x = \sum_i\sum_j i\ .\ P(i,j) \quad , \quad \mu_y = \sum_i\sum_j j\ .\ P(i,j)$
$\sigma_x = \sum_i\sum_j (i-\mu_x)^2\ .\ P(i,j)$ $\sigma_y = \sum_i\sum_j (j-\mu_y)^2\ .\ P(i,j)$

Table 3. The GLCM texture features and their mathematical representations

GLCM Texture Features	Mathematical Representations
Autocorrelation [21]	$\sum_i\sum_j (i\ .\ j)\,.\,P(i,j)$
Contrast [20, 21]	$\sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{\substack{j=1 \\ \lvert i-j\rvert=n}}^{N_g} P(i,j) \right\}$
Correlation (MATLAB SUITE)	$\dfrac{\sum_i\sum_j (i-\mu_x).(j-\mu_y).(P(i,j))}{\sigma_x\ .\sigma_y}$
Korelasyon [20, 21]	$\dfrac{\sum_i\sum_j (i\ .\ j).P(i,j) - \mu_x\ .\ \mu_y}{\sigma_x\ .\sigma_y}$
Cluster Prominence [21]	$\sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} \{i+j-\mu_x-\mu_y\}^4.\ P(i,j)$
Cluster Shade [21]	$\sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} \{i+j-\mu_x-\mu_y\}^3.\ P(i,j)$
Dissimilarity [21]	$\sum_i\sum_j \lvert i-j\rvert\,.\,P(i,j)$
Energy [20, 21]	$\sum_i\sum_j \{P(i,j)\}^2$
Entropy [21]	$-\sum_{i=0}^{N_g-1}\sum_{j=0}^{N_g-1} P(i,j).\log\{P(i,j)\}$
Homogeneity (MATLAB SUITE)	$\sum_i\sum_j \dfrac{1}{1+\lvert i-j\rvert^2}.P(i,j)$
Homogeneity [21]	$\sum_i\sum_j \dfrac{1}{1+(i-j)^2}.P(i,j)$

Table 3. The GLCM texture features and their mathematical Representations (continue)

Maximum Probability [21]	$\max_{i,j} P(i,j)$
Variance [20]	$\sum_i\sum_j (i-\mu)^2.P(i,j)$
Sum Average [20]	$\sum_{i=2}^{2N_g} i.P_{x+y}(i)$
Sum Variance [20]	$\sum_{i=2}^{2N_g} (i - Sum\ Average)^2.P_{x+y}(i)$
Sum Entropy [20]	$-\sum_{i=2}^{2N_g} P_{x+y}(i)\,.\log\{P_{x+y}(i)\}$
Difference Entropy [20]	$-\sum_{i=2}^{2N_g} P_{x-y}(i)\,.\log\{P_{x-y}(i)\}$
Information Measure of Correlation 1 [20]	$\dfrac{HXY - HXY1}{max\{HX;HY\}}$ $HXY = -\sum_i\sum_j P(i,j).\log(P(i,j))$ $HXY1 = -\sum_i\sum_j P(i,j).\log\left(p_x(i).p_y(j)\right)$ HX and HY are entropies of p_x and p_y
Information Measure of Correlation 2 [20]	$(1-\exp[-2.(HXY2-HXY)])^{1/2}$ $HXY2 = -\sum_i\sum_j P_x(i).P_y(j).\log\{P_x(i).P_y(j)$
Inverse Difference Normalized [22]	$\sum_i^{N_g}\sum_j^{N_g} \dfrac{1}{1+\lvert i-j\rvert^2/{N_g}^2}.P(i,j)$
Inverse Difference Moment Normalized [22]	$\sum_i^{N_g}\sum_j^{N_g} \dfrac{1}{1+(i-j)^2/{N_g}^2}.P(i,j)$

5. Experimental Study

5.1 Database

The database [10] of Japanese Society of Radiological Technology (JSRT) was used for this study. It is a digital public database with and without lung nodules. It contains 247 chest x-rays, among which 154 x-rays are abnormal, and 93 x-rays are normal. Out of 156 abnormal x-rays, 100 of them contain malignant nodules, and 54 of them contain benign nodules. Nodule diameters vary from 1mm to 60 mm. All x-ray images have a size of 2048×2048 pixels and a gray-scale color depth of 8 bits [10].

5.2. Revealing of Suspicious Tissue Regions

In all of the images, the Chan-Vese active contour without edge algorithm made the same number of iterations and found suspicious regions. For each suppressed rib image like in Figure 8b, the suspected regions of the left lung and the right lung were specified separately, and the starting contour was applied in the same manner.

The initial contour was taken as in Figure 11. The pixel value of the black pixels (pixel value=0) in each suppressed rib image was changed to 118 before suspicious regions were specified. For the left lung, the replacement of black pixels in a suppressed rib image with the value of 118 pixels is given in Figure 12a. The Chan-Vese algorithm had 1500 iterations, and suspicious regions were segmented. The white areas in Figure 12b show the locations that the algorithm found suspicious. The same procedures were performed in the right lung. Some examples that how the algorithm found suspicious regions were given in Figure 13a and Figure 13b.

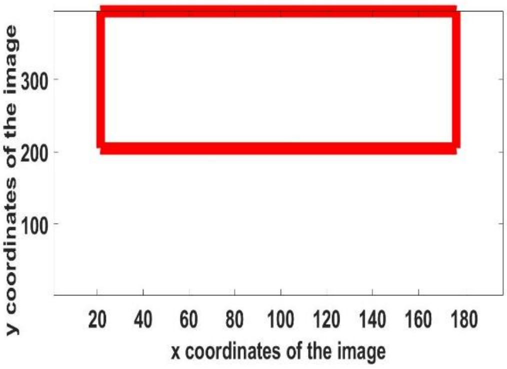
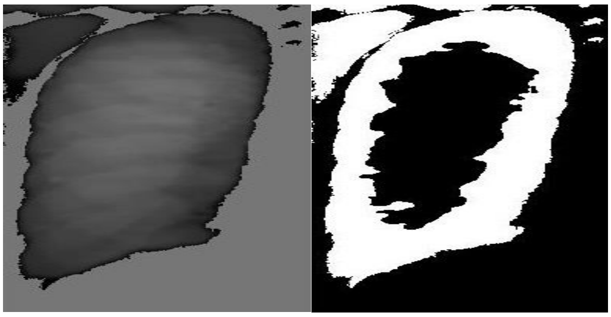
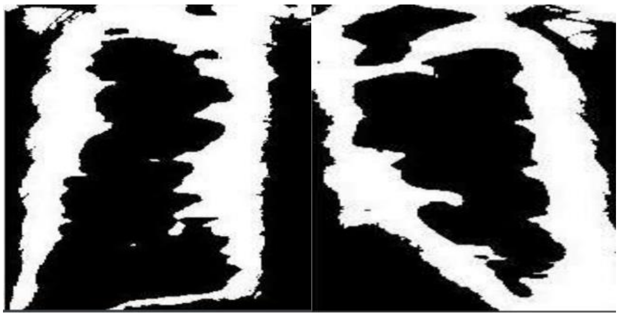


Figure 11. The frame (red colored rectangle) for the initial Contour



(a) (b)

Figure 12. a) The image after the replacement of black pixels with the value of 118, b) The segmentation of the suspicious region



(a) (b)

Figure 13. a) The example of how the algorithm found suspicious regions on left lung, b) The example of how the algorithm found suspicious regions on right lung

5.3. Feature Extraction and Classification of Detected Suspicious Tissues

The regions with a size of 32x32 were selected randomly from the suspicious regions in each image, and feature vectors were created from these regions. For each image, first, all of the feature extraction methods described in subsection 4 were extracted separately, and the classification success of each feature vector was examined. Then, the feature vector was created by using the features with high success, and the classification successes were explored again.

The method of 10-fold cross-validation was utilized in the classification. 90% of the images in each class were used for the training and 10% for the test purposes. The classification was completed in 10 stages. At each stage, 10% of each class entered the test phase, and the classification success was calculated on the basis of sensitivity, accuracy, and specificity. Then, these values were summed and averaged, so the system's sensitivity, accuracy, and specificity were found. The formulas for performance evaluation metrics (sensitivity, accuracy, and specificity) are given in the Table 4. Classifiers used; KNN (k = 5) [25], RF(number of trees = 100) [26], DT [27], NB [28], LLC [29], SVM [30], and FLDA [31].

Specificity percentage shows that how the algorithm is well enough to identify lungs without disease. Sensitivity percentage shows that how the algorithm is well enough to identify lungs with disease. Accuracy percentage shows that how the algorithm differentiates disease and without disease conditions correctly.

Firstly, all feature extraction methods described in subsection 4 was used separately to see which method gave the best result. The region with a size of 32x32 was converted to 1024x1 size of vector, and the statistical features were generated from 1024x1 vector. The feature vector obtained using only statistical properties has a size of 12x1. All the other features were extracted from the region with a size of 32x32 without converting to a vector. The GLCM matrix was formed by taking d = 2 and ϴ = 0°. The GLCM-based texture features were extracted from this GLCM matrix.

Table 4. The evaluation criteria and their mathematical expressions

Evaluation Criteria	Mathematical Expressions
TP: True Positive, TN: True Negative FP: False Positive, FN: False Negative S: Class Number, Vj = data number who belongs to jth class	
Sensitivity (SNS)	$SNS\% = \frac{TP}{TP+FN} \cdot 100$
Specificity (SPC)	$SPC\% = \frac{TN}{TN+FP} \cdot 100$
Accuracy (ACC)	$ACC\% = \frac{\sum_{J=1}^{S} SNS_J \cdot V_j}{\sum_{j=1}^{S} V_J}$

The feature vector created using GLCM-based texture features only has a size of 22x1. The rotation invariant LBP-based features were extracted. The radius of the circle is 1, and the number of pixels in the neighbourhood is 8. The feature vector obtained using LBP-based features only has a size of 10x1. The geometrical features were extracted. The feature vector obtained using only geometrical features has a size of 12x1. 16x16 cell size HOG was applied to the region with a size of 32x32, and the features were obtained. The feature vector obtained using HOG only has a size of 36x1. DSIFT-based features were extracted. The feature vector created using only DSIFT-based features has a size of 64x1. All feature vectors' sensitivity, accuracy, and specificity values for each classifier were calculated and presented in Table 5.

As seen in table 5, statistical features gave the best results. Then, statistical features were combined with the other features, and it was found which combination gave the best result. Statistical and HOG-based features were used together, and a 48x1 size vector was obtained.

Table 5. The sensitivity, accuracy, and specificity values for each classifier and feature vectors

	Criteria	KNN	FLDA	RF	DT	SVM	NB	LLC
Statistical Features	$ACC\%$	53%	59%	60%	56%	33%	59%	43%
	$SNS\%$	50%	49%	51%	50%	26%	51%	33%
	$SPC\%$	76%	78%	79%	78%	63%	79%	67%
GLCM-based texture features	$ACC\%$	39%	43%	44%	38%	48%	45%	49%
	$SNS\%$	36%	37%	39%	34%	39%	43%	40%
	$SPC\%$	70%	71%	71%	68%	71%	73%	71%
LBP-based features	$ACC\%$	42%	50%	54%	46%	52%	45%	52%
	$SNS\%$	41%	41%	47%	42%	42%	40%	43%
	$SPC\%$	71%	72%	75%	72%	73%	71%	73%
Geometrical Features	$ACC\%$	40%	48%	46%	44%	48%	47%	47%
	$SNS\%$	38%	40%	39%	41%	39%	40%	39%
	$SPC\%$	70%	71%	71%	72%	71%	72%	70%
HOG-based features	$ACC\%$	43%	49%	53%	43%	55%	45%	55%
	$SNS\%$	38%	43%	44%	40%	46%	39%	45%
	$SPC\%$	72%	73%	74%	71%	75%	72%	75%
DSIFT-based features	$ACC\%$	50%	45%	58%	45%	47%	51%	51%
	$SNS\%$	45%	41%	48%	41%	37%	43%	41%
	$SPC\%$	75%	72%	77%	72%	69%	74%	72%

Table 6. The sensitivity, accuracy, and specificity values for each classifier and combination of feature vectors

	Criteria	KNN	FLDA	RF	DT	SVM	NB	LLC
Statistical and HOG-based Features	$ACC\%$	52%	57%	64%	54%	32%	52%	43%
	$SNS\%$	50%	51%	55%	49%	25%	47%	33%
	$SPC\%$	76%	78%	81%	77%	62%	76%	67%
Statistical and LBP-based Features	$ACC\%$	52%	51%	60%	54%	32%	53%	41%
	$SNS\%$	50%	43%	51%	50%	25%	48%	33%
	$SPC\%$	76%	74%	79%	77%	62%	77%	66%
Statistical and DSIFT-based Features	$ACC\%$	51%	45%	59%	50%	32%	51%	43%
	$SNS\%$	48%	39%	49%	46%	25%	43%	34%
	$SPC\%$	76%	72%	78%	75%	62%	75%	67%
Statistical, HOG- and DSIFT-based Features	$ACC\%$	51%	43%	60%	51%	32%	53%	42%
	$SNS\%$	49%	39%	49%	46%	26%	45%	34%
	$SPC\%$	76%	74%	79%	77%	62%	77%	66%
Statistical, HOG- and LBP-based Features	$ACC\%$	54%	55%	66%	50%	32%	53%	43%
	$SNS\%$	51%	49%	57%	46%	27%	50%	33%
	$SPC\%$	78%	78%	81%	75%	64%	77%	67%

Statistical and LBP-based features were used together, and a 22x1 size vector was obtained. Statistical and DSIFT-based features were used together, and a 76x1 size vector was obtained. It was seen that the combination of statistical and HOG-based features gave the best results. Afterwards, DSIFT-based features and LBP-based features were added to the combination of statistical and Hog-based features separately to see which gave the best results. Statistical, HOG-based, and DSIFT-based features were used together, and a 112x1 size vector was obtained. Statistical, HOG-based, and LBP-based features are used together, and 58x1 size vector is obtained. Sensitivity, accuracy, and specificity values for each classifier and combination of feature vectors were calculated and presented in Table 6.

6. Discussion

The findings are in line with the literature. If the classification results (accuracy=66%, sensitivity=57%, and specificity=81%) are compared to the studies' results which use the same JSRT database in the literature, it gives the top eighth result among the other up to date works [10]. It is important to compare the proposed study with other studies that only used the JSRT database in terms of being fair because the choice of the database directly affects the results. Moreover, all the manuscripts that compared use image processing techniques. It is also essential in terms of making a fair comparison because it is known that many deep learning algorithms are used in this area.

7. Conclusions

In this study, a new region of interest (ROI) detection method using the Chan-Vese algorithm for lung nodule classification was proposed. Firstly, chest x-rays were preprocessed to suppress ribs. Then, suspicious tissue regions were determined in each x-ray by applying the Chan-Vese algorithm. The suspicious regions were successfully specified with a sensitivity score of 73%. The regions with a size of 32x32 were selected randomly from suspicious regions for each image, and then, different feature vectors were extracted from these selected regions. Finally, the chest x-rays were classified into three groups, including no nodule, benign nodule, and malignant nodule.

As seen in Table 6, the whole aggregation of the HOG-, LBP-based and statistical features yielded the best classification results (accuracy=66%, sensitivity=57%, and specificity=81%) with the random forest classifier when tree number was 100. Thus, starting from the best individual performance among all the types of the features, the dual combinations of the features in which the statistical features remain constant in every possible case were tested. Then the best triplet combination was constructed using the best dual combinations of the features according to accuracy, sensitivity, and specificity performances. This manner of increase in the number of features is an elaborative search in order to identify the most suitable feature-classifier combination. The random forest classifier mostly generated better classification results compared to the other classifiers.

Declaration

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The authors also declared that this article is original, was prepared in accordance with international publication and research ethics, and ethical committee permission or any special permission is not required.

Author Contributions

S. Ergin supervised and guided the study. B. Topuz did some preprocessing operations on chest x-ray images, interpreted the JSRT database, and adjusted parameters in the chan-vese algorithm. A. Cinar did some preprocessing, feature extraction, classification, and detected suspicious regions.

References

1. Stewart, B., C.P. Wild, *World Cancer Report 2014*. 2015, WHO Press.
2. Garfinkel, L., G. Murphy, W.J. Lawrence, R.J. Lenhard, *American Cancer Society Textbook of Clinical Oncology.* 1995, The Society Press.
3. Abbas, W., K.B. Khan, M. Aqeel, M.A. Azam, M.H. Ghouri, F.H. Jaskani, *Lungs Nodule Cancer Detection Using Statistical Techniques.* IEEE 23rd International Multitopic Conference, 2020, Pakistan. p. 1-6.
4. Parveen, S., K.B. Khan, *Detection and classification of pneumonia in chest X-ray images by supervised learning.* IEEE 23rd International Multitopic Conference, 2020, Pakistan, p. 1-5.
5. Gonzalez, E.R., V. Ponomaryov, *Automatic Lung nodule segmentation and classification in CT images based on SVM*. 9th International Kharkiv Symposium on Physics and Engineering of Microwaves, Millimeter and Submillimeter Waves, 2016, Ukraine. p. 1-4.
6. Tun, K.M.M, A.S. Khaing, *Feature Extraction and Classification of Lung Cancer Nodule using Image Processing Techniques.* International Journal of Engineering Research & Technology, 2014. **3**(3): p. 2204-2210.
7. Shiraishi, J., S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi , K. Komatsu, M. Matsui, H. Fujita, Y. Kodera, K. Doi, *Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules*. AJR, 2000. **174**(1): p. 71-74.
8. Khan, K.B., A.A. Khaliq, M. Shadid, J.A. Shah, *A new approach of weighted gradient filter for denoising of medical images in the presence of Poisson noise*. Tehnički vjesnik, 2016. **23** (6): p. 1755-62.
9. Gallagher, N., G. Wise, *A theoretical analysis of the properties of median filters*. IEEE Transaction on Acoustic Speech Signal Processing, 1981. **29** (6): p. 1135–1141.
10. Wang, C., A. Elazab, J. Wu, Q. Hua, *Lung nodule classification using deep feature fusion in chest radiography*. Computerized Medical Imaging and Graphics, 2017. **57**: p. 10-18.
11. Gabor, D., *Theory of communication*. Journal of the Institution of Electrical Engineers- Part III: Radio and Communication Engineering, 1946. **93**(26): p. 429–457.
12. Soleymanpour, E., H.R. Pourreza, E. Ansariour, M. Sadooghi, *Fully Automatic Lung Segmentation and Rib Suppression Methods to Improve Nodule Detection in Chest Radiographs*. Journal of Medical Signals and Sensors, 2011. **1**(3): p. 191-199.
13. Haddad, R.A., A.N. Akansu, *A, Class of Fast Gaussian Binomial Filters for Speech and Image Processing*, IEEE Transactions on Acoustics, Speech and Signal Processing, 1991. **39**(3): p. 723-727.
14. Chan, T.F., L.A. Vese, *Active contours without edges*. IEEE Transactions on Image Processing, 2001. **10**(2): p. 266-277.
15. Esener, I.I., S. Ergin, T. Yuksel, *A New Feature Ensemble with a Multistage Classification Scheme for Breast Cancer Diagnosis*. Journal of Healthcare Engineering, 2017. **2017**: p. 1-15.
16. Ahonen, T., A. Hadid, M. Pietkäinen, *Face description with local binary patterns: Application to face recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence,

2006. **28**(12): p. 2037-2041.

17. Song, L., X. Liu, L. Ma, C. Zhou, X. Zhao, Y. Zhao, *Using HOG-LBP features and MMP learning to recognize imaging signs of lung lesions*. 25.International Symposium On Computer-Based Medical Systems, 2012, Italy. p. 1-4.
18. Wang, L., D.C. He,*Texture Classification Using Texture Spectrum*, Pattern Recognition, 1990. **23**(8): p. 905-910.
19. Esener, I.I., *The Identification of Suspicious Regions on Mammography Images for Breast Cancer and the Classification of Breast Cancer Type*. PhD Thesis, Eskisehir Osmangazi University, 2017.
20. Haralick, R.M., K. Shanmugam, I. Dinstein, *Textural features of image classification.* IEEE Transactions on Systems, Man, and Cybernetics, 1973. **SMC-3**(6): p. 610-621.
21. Soh. L., C. Tsatsoulis, *Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices*. IEEE Transactions on Geoscience and Remote Sensing, 1999. **37**(2): p. 780-795.
22. Clausi, D.A., *An analysis of co-occurrence texture statistics as a function of grey level quantization*. Canadian Journal of Remore Sensing, 2002. **28**(1): p. 45-62.
23. Esener. I.I, S. Ergin, T. Yuksel, *A Genuine GLCM-based Feature Extraction for Breast Tissue Classification on Mammograms*. International Journal of Intelligent Systems and Applications in Engineering, 2016. **4**(Special Issue): p. 124-129.
24. Ergin. S., O. Kilinc, *Using DSIFT and LCP features for detecting breast lesions*. ISCSE, 2013. International Symposium on Computing in Science & Engineering. Proceedings: Izmir. p. 216-220.
25. Kim. J., B.S. Kim, S. Savarese, *Comparing image classification methods:K-nearest-neighbor and support-vector-machines.* 6. WSEAS International Conference on Computer Engineering and Applications, 2012, World Scientific and Engineering Academy and Society: USA. p. 133-138.
26. Akar. O, O. Gungor, *Classification of multispectral images using Random Forest Algorithm*. Journal of Geodesy and Geoinformation, 2012. **1**(2): p. 105-112.
27. Safavian, S.R., D. Landgrebe, *A survey of decision tree classifier methodology*. IEEE Transactions on Systems, Man, and Cybernetics, 1991. **21**(3): p. 660-674.
28. Rish, I., *An empirical study of the naive Bayes classifier*. IJCAI Workshop on Empirical Methods in artificial intelligence, 2001, IBM New York: USA. p. 41-46.
29. Webb, A.R., *Linear discriminant analysis in Statistical Pattern Recognition*. 2002, John Wiley & Sons.
30. Ozkan, K., S. Ergin, S. Isik, I. Isikli, *A new classification scheme of plastic wastes based upon recycling labels*, Waste Management, 2015. **35**: p. 29-35.
31. Fisher, R.A., *The use of multiple measurements in taxonomic problems*. Annals of Eugenics, 1936. **7**(2): p. 179-188.