∎ 1256

Lung diseases detection caused by smoking using support vector machine

Sri Widodo*[1], Ratnasari Nur Rohmah[2], Bana Handaga[3], Liss Dyah Dewi Arini[4]
[1,4]Citra Medika Surakarta Medical Records and Health Information Academy, Central Java, Indonesia
[2,3]Electrical Enginnering Department, Muhammadiyah Surakarta University, Central Java, Indonesia
*Corresponding author, e-mail: widodosri1972@gmail.com[1], rnr217@ums.ac.id[2],
bana.handaga@ums.ac.id[3], leeansz_fortune@yahoo.com[4]

Abstract

Type of lung disease is very much manifold, but type of lung disease caused by smoking there are only 4, namely Bronchitis, Pneumonia, Emphysema and Lung Cancer. Doctors usually diagnose lung disease from CT scans using the naked eye, then interpret data one by one.This procedure is not effective. The aim of this research is improvement accuracy of lung diseases detection caused by smoking using support vector machine on computed tomography scan (CT scan) images. This study includes 4 (four) main points. First is the development of software for segmentation of lung organ automatically using Active Shape Model (ASM) method. Second is the segmentation of candidates who are considered illness by using Morphology Mathematics. The third process of lung disease detection using Support Vector Machine (SVM). Fourth is visualization of disease or lung disorder using Volume Rendering.

Keywords: ASM, CT scan, lung disease, morphological math, SVM

1. Introduction

Medical imaging system is a system used to examine structural abnormalities and organ dysfunction in human body. Examples of medical imaging systems are X-rays, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Single Photon Emission Computed Tomography (SPET) and Positron Emission Tomograph (PET). The system is very useful to help diagnose a disease and useful in implementation of computer aided surgery [1, 2]. One of the organ structural abnormalities is lung disease. Many types of lung disease, but lung diseases caused by smoking amount to 4 types, namely Bronchitis, Pneumonia, Emphysema and Lung Cancer [3, 4]. CT scan image of lung normal and lung diseases caused by smoking shown in Figure 1.

Lung specialists and radiologists use naked eye vision to read and diagnose lung disease in CT 2-D images [5]. The reading process is done by placing CT or X-ray images printed out on the reading lamp [6].This method is certainly not effective. In addition, lung specialists may differ in diagnosing lung disorders, including determining the type, shape, size and location of abnormalities in the lung. Therefore, it is necessary application that can detect or diagnose lung disease from CT scan image automatically.

Several studies related to the detection of abnormalities in the lung has been done by many researchers [7-12]. From the research that has been done by some previous researchers can be summarized as follows: almost all CT scan data used using image of lung edge very clear. Data of lung abnormalities used is big shape, making it easy to segment. The CT scan data for the experiment was cropping manually. Our study, the segmentation process can segment for obscure lung boundaries. In segmentation suspected lung disease can segment lung disease image in a small size and attach to the artery. Both of these processes will greatly affect process of detecting lungs abnormalities.

The aim of this study is improvement accuracy of lung diseases detection caused by smoking using support vector machine on Computed Tomography Scan (CT scan) images. This study includes 4 (four) main points. First is the development of software for segmentation of lung organ automatically using Active Shape Model (ASM) method. Second is segmentation of candidates who are considered illness by using Morphology Mathematics. The third process of lung disease detection using Support Vector Machine (SVM). Fourth is visualization of lung

disease using Volume Rendering. The developed software can detect from the original CT scan image which has an obscure lung edge and small abnormalities. Research carried out can segment lung images that have unclear edges from the original CT-Thoracic. Lung segmentation determines the success of lung abnormalities detection. Features used for detection are combinations of first and second order statistical features. In this study, 3-D reconstruction of images suspected of abnormalities in the lungs was carried out, so users could see the shape and location of abnormalities from different angles. To facilitate the detection process all programs are integrated in user interface.
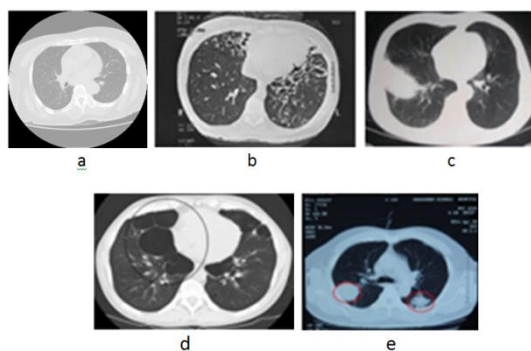

Figure 1. CT scan image of lung normal and lung diseases caused by smoking
(a) lung normal (b) bronchitis, (c) pneumonia, (d) emphysema and (e) lung cancer

2. Research Method

This research consists of 4 stages namely: lung organ segmentation automatically using Active Shape Model (ASM) method, segmentation of candidates who are considered illness by using Morphology Mathematics, lung disease detection using Support Vector Machine (SVM) and visualization of lung diseases using Volume Rendering. Stages of developed method, shown in Figure 2.

2.1. Active Shape Model (ASM)

Active Shape Model (ASM) is an object modeling technique that is very powerful and one of the best methods on computer vision and computer graphics, especially in image segmentation. This method includes a segmentation method that uses training data (supervised). The stages in ASM include: forming a model using the position of a landmark from the training data, training a model to find gray value information from each landmark using images on trainingset, image search, at this stage, an initial estimate of the form is applied manually to an invisible image. The initial form must be placed on the edge of the object from an invisible image and responded at the same time. Furthermore ASM uses edge profiles and covariance matrices from the normal mean derivatives to find the best movements [13].
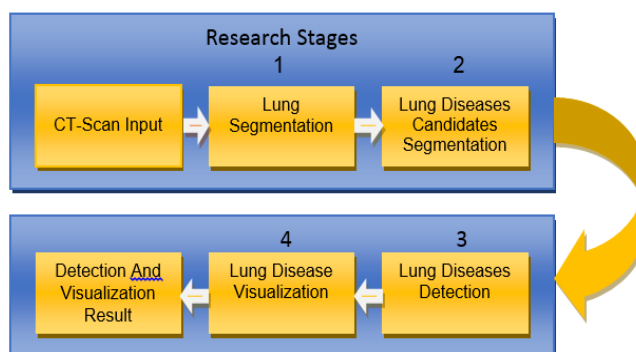

Figure 2. Stages of developed method

2.2. Morphological Operation.

Morphological operations are operations that are commonly applied to binary images (black and white) to change the structure of objects contained in images. The core of morphological operations involves two arrays of pixels. The first array is an image that will be subjected to morphological operations, while the second array is called the kernel or structuring element. Some morphological types include: dilatation (dilation) and Erosion. Dilation and erosion operations do add and subtract píksel on binary imagery. The dilatation function is to add pixels to the edges of each binary object, namely the area that has a value of 1. The erosion function is to remove 8 pixels from a binary object that is related to the edges of the object. Erosion is the removal of the object point (1) being part of the background (0) [14-16].

2.3. Support Vector Machine (SVM)

Support Vector Machine (SVM) developed by Boser, Guyon, and Vapnik, was first presented in 1992 at the Annual Workshop on Computational Learning Theory. The basic concept of SVM is actually a harmonious combination of computational theories that have existed decades before, such as hyperplane margins (Duda and Hart in 1973, Cover in 1965, Vapnik 1964, etc.), the kernel was introduced by Aronszajn in 1950, as well as other supporting concepts. The difference with a neural network is that a neural network tries to find a dividing hyperplane between classes, SVM tries to find the best hyperplane in input space. The basic principle of SVM is a linear classifier, and then developed so that it can work on non-linear problems by incorporating the kernel trick concept in high-dimensional workspaces. Figure 3 shows hyperplane that separates two classes (-1 and $+1$). The pattern incorporated in the class 1 is symbolized as red (box), while the pattern in class $+1$ is symbolized as yellow (circle). The classification problem can be translated by finding a line (hyperplane) that separates the two groups. Various alternative discrimination lines are shown in Figure 3 (a). The best hyperplane separator between two classes can be found by measuring the margin of hyperplane, and looking for the maximum point. Margin is distance between hyperplane and the closest pattern of each class. This closest pattern is called support vector. Figure 3 (b) shows the best hyperplane (solid line), which is line that is located right in the middle of two classes, while the red and yellow points in black circle are support vector. Finding a hyperplane location is core of learning process in SVM. SVM was originally designed as a binary classification, which is only to separate two classes. Therefore, SVM classification method by building and combining several more binary classifications was chosen to do the problem of multiclass. Some of the most popular SVM methods are One Against One (OAO), One Against All (OAA) and Half Against Half (HAH) [17, 18].
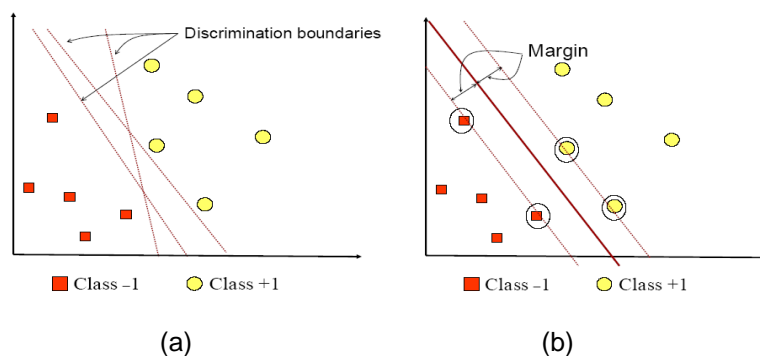

(a)   (b)

Figure 3. Hyperplane that separates two classes (-1 and $+1$)
(a) various of alternative discrimination lines, (b) the best hyperplane (solid line)

2.4. Volume Rendering

Volume rendering is one technique that is highly capable in image visualization and manipulation. In this technique there is no need for discretion from the surface, so that the integrity of the volume image data is well preserved. That is why this technique can produce high quality displays, although of course the implication is that computing time is quite long [19, 20].

3. Results and Analysis

Data used in this study is CT images of 30 patients in axial slices. Image size of 505x427 pixels, and a thickness of 0.5-10 mm. CT scan image were taken from Ir. Soekarno Sukoharjo Regional General Hospital, Central Java Indonesia.

3.1. Lung Segmentation with Active Shape Model (ASM)

Lung segmentation is done to separate lung field with the surrounding tissue. Method used is Active Shape Model. In this paper we do not discuss about the segmentation of lung using ASM. This is caused by ASM has been discussed in another paper [5]. Lung segmentation algorithm using ASM can be explained as follows:
a. Start
b. Input CT scan data
c. Read CT scan data
d. Make training data by marking the edge of the lung
e. Form a model using a landmark position from the image on the training data.
f. Shape Alignment
g. Application of the Principle Component Analyzer (PCA) that transmits the dataset to the new coordinate system as the first largest variant in one-dimensional data.
h. Training Model to find gray value information from each landmark using the image on trainingset.
i. Image Search uses edge profiles and covariance matrices from the normal mean derivatives to find the best moves.
j. Calculate Landmark Movements
k. Input testing data
l. Putting a model from ASM on a lung CT scan image that will be segmented.
m. Summing the contours of the segmentation results with the CT scan image being tested
n. See the results of segmentation
o. Save the results of segmentation
p. Finish

Results of lung segmentation can be seen in Figure 4. Segmentation of the lung fields using ASM may segment the lung fields that have unclear lung borders, which is difficult to do by conventional segmentation methods. So the lungs that have a large (nodular) abnormality, and lies in the lung boundary that causes the lung edge is not clear, can be well segmented. Results of our study show that segmentation approach with an Active Shape Model has accuracy 96.8%, sensitivity 88.2% and specificity 99.1%.
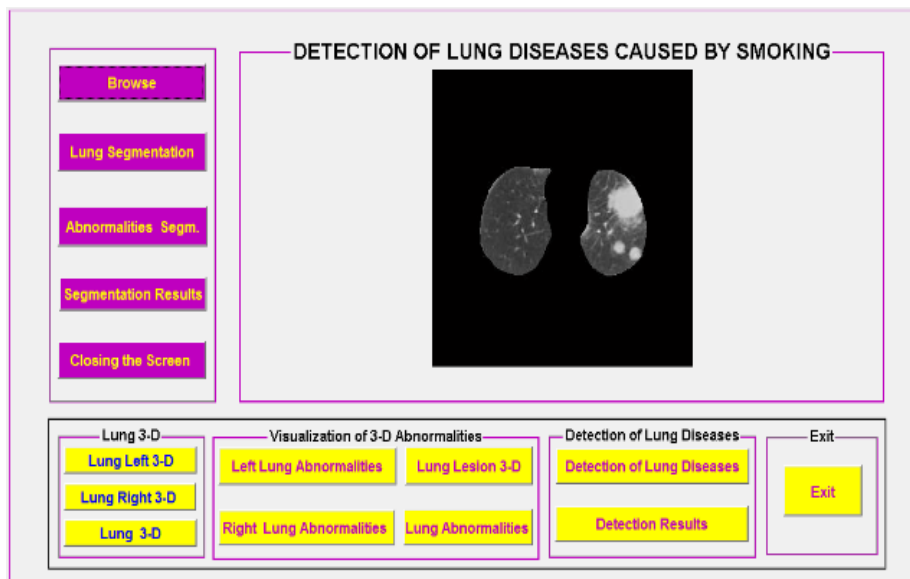

Figure 4. Results of lung segmentation

3.2. Lung Diseases Candidates Segmentation using Morphology Mathematics

Method used for segmentation of lung diseases candidate is morphology mathematics. Image segmentation algorithms of lung diseases candidate can be explained as follows:

a. Start
b. Read the CT scan Paru Image
c. Image of Subtraction Results.
d. Image of Threshold with Value 60.
e. Results of Edge Detection with Convolution Prewitt.
f. Dilation Process Results with Value: 1.5.
g. Image of Masking process Results.
h. Image of Erosion Process.
i. Showing the results of the Lung Diseases Candidate Image
j. Save Results
k. Finish

Segmentation process of lung diseases candidate using morphology mathematics is shown in Figures 5 and 6. Segmentation result of lung diseases candidate shown in Figure 7. Segmentation of lung disease candidates using morphological math can segment the shape of abnormality with a very small size and attached to the artery, which is difficult for some researchers to do. This is because the image suspected as a lung abnormality is not clearly visible. This segmentation will affect the accuracy of lung disease detection processThe results of segmentation of disease candidate showed that the segmentation approach with morphology mathematics had 94.1% accuracy, 93.5% sensitivity, and 94.2% specificity.
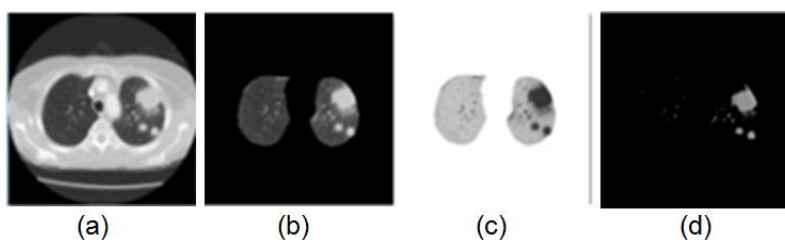
Figure 5. Subtraction process of image
(a) original ct images, (b) lung image, (c) negative image,
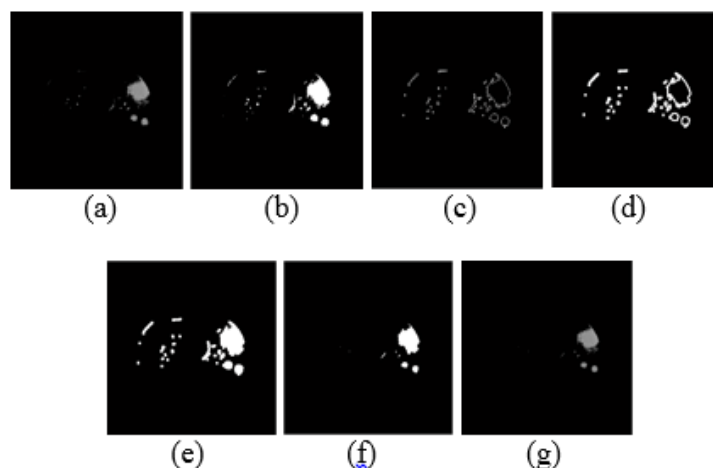(d) subtraction results between b and c image

Figure 6. Morphology process of lung diseases candidate
(a) image of subtraction results, (b) image of threshold with value 60,
(c) results of edge detection with convolution prewitt, (d) dilation process results with
value: 1.5, (e) image of masking process results, (f) image of erosion process,
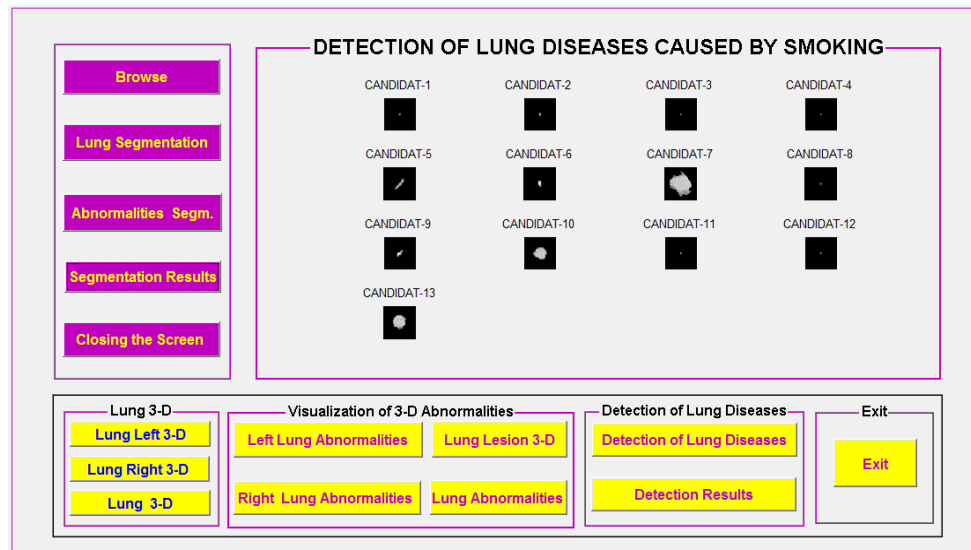(g) lung diseases candidate image

Figure 7. Segmentation result of lung diseases candidate

3.3. Detection of Lung Diseases using SVM

The method used to detect lung diseases caused by smoking is the Support Vector Machine (SVM) with Principle Component Analysis (PCA). In this paper we do not discuss about Principle Componen Analysis (PCA). This is because PCA has been discussed in another paper [6]. Algorithms for detecting lung disease caused by smoking can be explained as follows:
a. Start
b. CT scan Image input
c. Read the CT scan image entered
d. Creating file names Image of lung abnormalities.
e. Creating a sequence of training and trial data.
f. Creating an overall matrix of images of lung abnormalities in 1 array.
g. Creating a training data matrix
h. Create feature extraction using PCA
i. Creating training data tables
j. Detect pulmonary abnormalities using SVM.
k. the results of testing each data that was tested
l. Calculate the accuracy of the detection process
m. Finish

The algorithm for detecting lung abnormalities caused by smoking using SVM can be explained as follows:
a. Start
b. Input: x, y, kernel functions, kernel parameters, C
c. Alpha0 = [ ]; % (empty set).
d. b0 = [ ]; % (empty set)
e. Determine Mmax as the maximum value of y; % of labels: 1, 2, 3 ... ... k
f. Determine Min as the minimum value of y; % of labels: 1, 2, 3 ... ... k
g. M=Mmax-Mmin + 1 as the number of classes
h. For i = 1 to M, do the following:
   1. Change the label value to +1 for classes I and -1 for other than i
   2. Call the computer program for SVM two classes to get alpha and b from the two class classification problems obtained
   3. Alpha0 = [alpha0 alpha]
   4. B0 = [b0 b]
   5. Finished
   6. Output: alpha0, b0
i. Detection results

Lung diseases detection caused by smoking using support vector machine... (Sri Widodo)

i. Finish

Stages in detection of lung disease caused by smoking using Support Vector Machine (SVM) [21] can be described in Figure 8. The first step in diseases detection process is feature extraction. Features used is analysis of texture. Texture analysis method used in this study is first order and second-order statistical feature extraction also called Gray Level Co-occurrence Matrix (GLCM) [22]. Grey-Level Co-occurrence Matrix texture measurements have been the workhorse of image texture since they were proposed by Haralick in the 1970s. The GLCM is a tabulation of how often different combinations of pixel brightness values (gray levels) occur in an image. First order texture measures are statistics calculated from original image values, like variance, and do not consider pixel neighbor relationships. Second order measures consider the relationship between groups of two (usually neighboring) pixels in the original image. Features of the first order include: mean, standard deviation, smoothness, third moment, uniformity, entropy. To calculate features are used the following formulas. Mean is a measure of mean intensity:

$$m = \sum_{i=0}^{L-1} Z_i \, P(Z_i) \tag{1}$$

standard deviation is a measure of contrast average:

$$\sigma = \sqrt{\sum_{i=0}^{L-1} (Z_i - m)^2 p(Z_i)} \tag{2}$$

smoothness measure relative smoothness of intensity in the region:

$$R = 1 - \frac{1}{1+\sigma^2} \tag{3}$$

third moment to measure skewness of histogram:

$$\mu_3 = \sum_{i=0}^{L-1} (Z_i - m)^3 \, p(Z_i) \tag{4}$$

uniformity:

$$U = \sum_{i=0}^{L-1} p^2(z_i) \tag{5}$$

entropy:

$$e = - \sum_{i-0}^{L-1} p(z_i) \, log_2 \, p(z_i) \tag{6}$$

Second-order statistics done by creating cooccurence matrix, ie a matrix that represents the adjacency relationship between pixels in the image at various orientation and spatial distance. One technique to obtain the statistical characteristics of second order is to calculate probability of adjacency relationship between two pixels at a certain distance and angular orientation [23-25]. This approach works by forming a cooccurence matrix of image data, followed by determining the characteristics as a function of the matrix between them. Occurrences means joint events, ie number of occurrences of level of neighboring pixel values with one another level pixel values within (d) and orientation angle (θ) specific. Occurrences matrix formation of an image with 45^0 in the angular orientation shown in Figure 9.

While GLCM features used in this study are 20 features that include first-order features are rotated on the axis 0^0, 45^0, 90^0 and 135^0. The second step is the process of classification using Support Vector Machine (SVM) [17, 18]. In the process of classification using SVM, performed two steps, the first step is the training, while the second phase is the testing step. The amount of training data is 20 per class, while the testing is done by using the image data that has not done the training is number 15 CT scan.

The training phase is used to obtain the coordinates of support vector, weight, and distance suppot vector, whereas the testing phase is to use data other than the training data to get the results of the classification, so as to know the level of accuracy. Some second-order statistical texture features (GLCM) has been selected as the input system classifier that is considered able to present the nature of the texture of the object. Results of lung diseases caused by smoking detection shown in Figure 10.
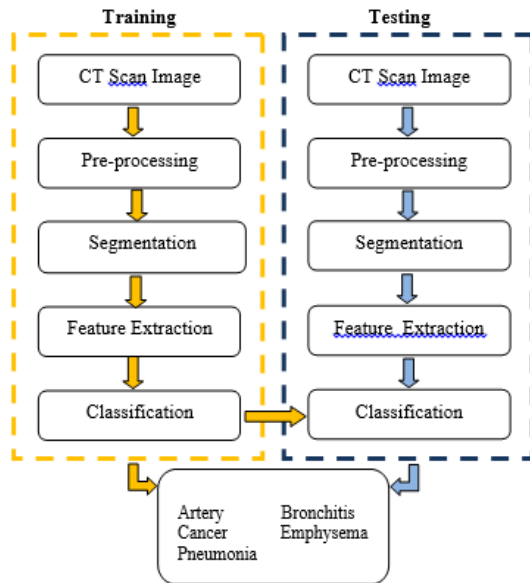
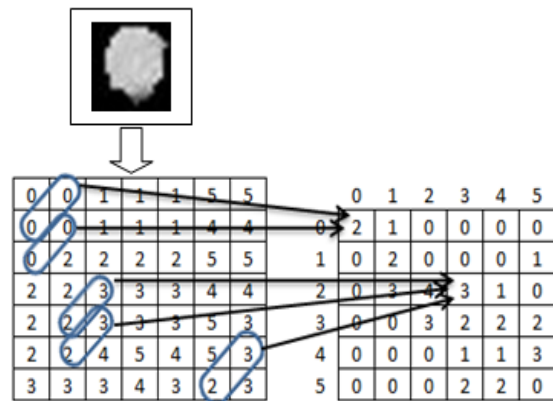Figure 9. Occurrences matrix with angle 45^0

Figure 8. Stages in detection of lung disease
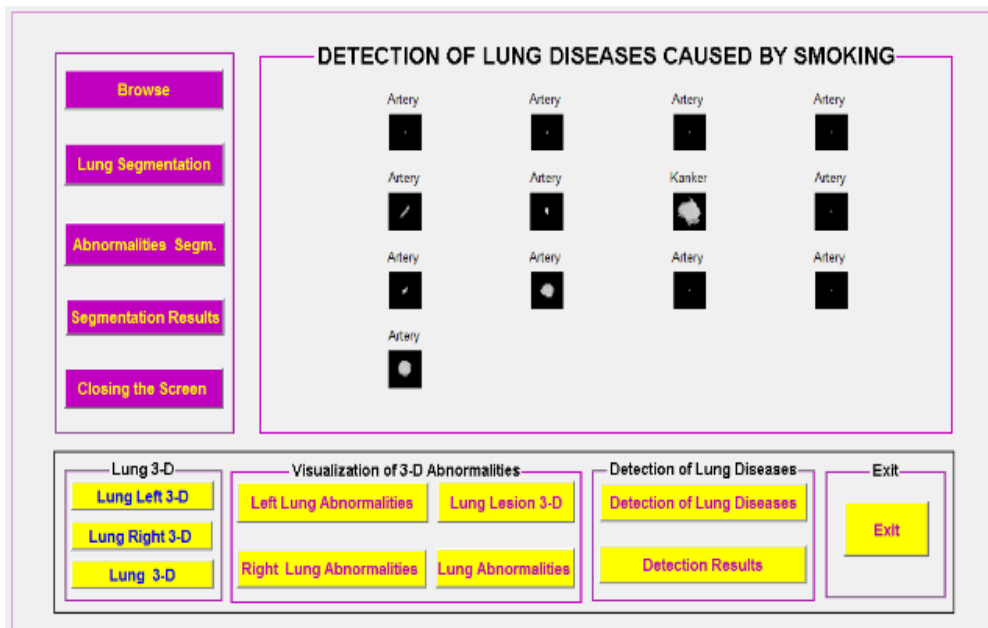caused by smoking

Figure 10. Results of lung disease detection caused by smoking

Classification results will be made a comparison with groundtruth, so that would be obtained four values, each of which is a true positive, false negative, false positive and true negative. True positive (TP) shows the image of lung disease caused by smoking identified precisely in its class. False positive (FP) is the image of lung disease caused by smoking identified another class. True negative (TN) is an image not lung disease caused by smoking, identified the right not lung disease caused by smoking. False negative (FN) shows the image that should not lung disease caused by smoking, identified as lung disease caused by smoking. Based on the four values obtained true positive rate (TPR), which is known as sensitivity. Sensitivity formula is (7):

$$TPR = \frac{TP}{TP+FN} \tag{7}$$

the false positive rate (FPR) or specificity is a value that indicates the level of error in identification obtained by the (8):

$$FPR = \frac{FP}{FP+TN} \tag{8}$$

while the value that indicates the accuracy of identification is obtained from the (9):

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{9}$$

the greater the degree of accuracy, sensitivity and specificity declare the results more accurate prediction system. Results of Lung Disease Detection Caused by Smoking shown in Figure 10.

Testing process is done after CT scan image of lung disease caused by smoking is a segmentation process to separate lung disease image with surrounding tissue, and also to determine Region of Interest (ROI) of lesion object. Furthermore image is normalized to 60x60 pixels with gray degree 255. Testing process is done by dividing sample data into two. The first group is training data, second group as testing data. The amount of sample data used is 150. Which consists of 75 training data and 75 data testing. The number of classes there are 5, artery class, cancer class, pneumonia class, bronchitis class, and emphysema class. Each class has 30 training data. If feature extraction uses 75 training data, then training data used is data 1 to 75, while data 76 to 150 is used as test data. Accuracy of testing data by using 75 data test can be seen in Figure 11.
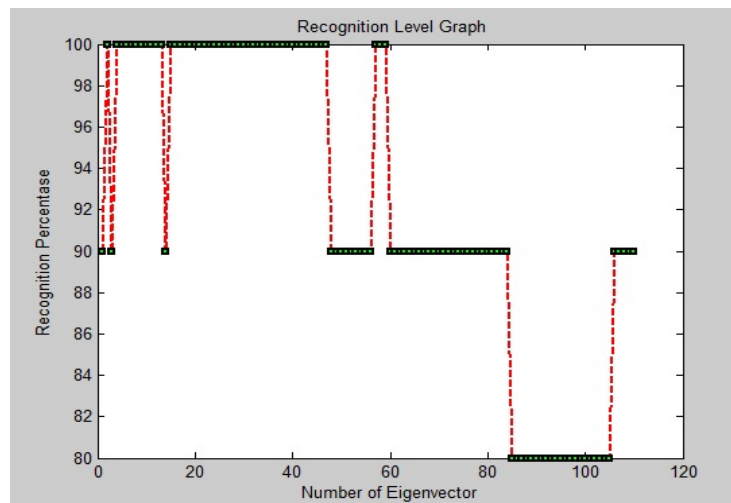

Figure 11. Accuracy of testing data

3.4. 3-D Visualization of Lung Disease Caused by Smoking

To assist user in detecting of lung disease caused by smoking, the developed software comes with modules for the visualization of the lung and lung disease caused by smoking. This is because the medical image is generally a collection of 2D image slices are difficult to interpret, to make it more resembles its original form, subsequent 2D medical images is transformed into a 3-D image. Visualization is done by using volume rendering. Algorithm for 3-D visualizing of lung disease can be explained as follows:

a. Start
b. Input 2-D CT scan Image.
c. Reading a collection of 2-D medical images to be set based on the position spatially.
d. Applying the technique of image rendering using volume rendering (VR).

e. Volume rendering is implemented using modules from central mathwork, with vol3d.m name.
f. Smoothing process of lung image.
g. Render texture field of x, y, z simultaneously. The aspect ratio of the data [1 1 0.6], which means that one unit is one unit in y and 0.6 units in z.
h. Create a linear alphamap by increasing opacity.
i. The result of lung volume rendering is displayed using the command view (3), the purpose of which is to display 3D shape with az=-37.5 el=30.
j. Results of 3-D reconstruction
k. Rotation Process
l. Save results
m. Finish

Results of 3-D visualization of lung disease caused by smoking and its location in the lung area are shown in Figure 12.
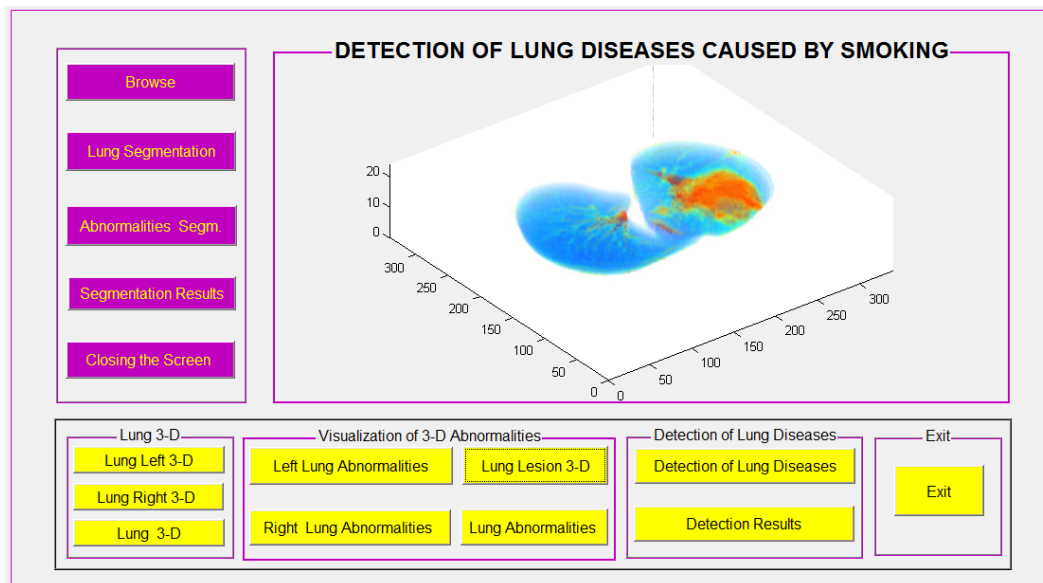

Figure 12. Visualization of lung disease caused by smoking

4. Conclusion

Segmentation with ASM can successfully segment the lung fields that have a large abnormality, resulting in an unclear edge of the lung, this is due to the low contrast value. Segmentation of lung disease candidate with Morphology Mathematics successfully segmented the abnormality well, even abnormalities attached to the artery, which is difficult for some researchers to do. After testing with 75 testing data, by using Support Vector Machine Methode using 75 training data, showed average accuracy of 90%. Based on the tests, it can be concluded that software detection of lung diseases by using Support Vector Machine (SVM) proved capable of being used as a model for the detection of lung disease caused by smoking.

References

[1] JD Bronzino. The Biomedical Engineering. Handbook 2nd Edition. Boca Raton: CRC Press. 2000: 61.

[2]   Clinton Brian Fookes. Medical Image Registration and Stereo Vision Using Mutual Information. Queensland University of Techonology, Brisbane, Australia, 2003.

[3]   IGK Sajinadiyasa, IM Bagiada, IB Ngurah Rai. Prevalence and Smoking Risk of Lung Disease at the Lung Polyclinic at the Sanglah Central General Hospital, Denpasar. *Jurnal Penyakit Dalam.* 2010; 11(2).

[4]   IfeanyiChukwu O Onor, Daniel L Stirling, Shandrika R Williams, Daniel Bediako, Amne Borghol, Martha B Harris, Tiernisha B Darensburg, Sharde D Clay, Samuel C Okpechi, Daniel F Sarpong. Clinical Effects of Cigarette Smoking: Epidemiologic Impact and Review of Pharmacotherapy Options. *International Journal Environment Researc and Public Health.* 2017; 14: 1147.

[5]   Widodo S, Wijiyanto. Software Development For Three Dimensional Visualization Of Lung on Computed Tomography Scans Using Active Shape Model And Volume Rendering. *Journal of Theoretical and Applied Information Technology (JATIT).* 2014; 65(1): 154-160.

[6]   Widodo S, Ratnasari Nur Rohmah, Bana Handaga. *Classification of Lung Nodules and Arteries in Computed Tomography Scan Image Using Principle Component Analysis.* 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE). Yogyakarta. 2017: 152-157.

[7]   Golosio B, Masala GL, Piccioli A, Oliva P, Carpinelli M. A novel multithreshold method for nodule detection in lung CT. *Med Phys.* 2009; 36(8): 3607–3618.

[8]   Messay T, Hardie R, Rogers S. A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. *Med Image Anal.* 2010; 14(3): 390–406.

[9]   Riccardi A, Petkov TS, Ferri G, Masotti M, Campanini R. Computer-aided detection of lung nodules via 3D fast radial transform, scale space representation, and Zernike MIP classification. *Med Phys.* 2011; 38(4): 1962–1971.

[10] Camarlinghi N, Gori I, Retico A, Bellotti R et al. Combination of computer-aided detection algorithms for automatic lung nodule identification. *Int J CARS.* 2011; 7(3): 455–464.

[11] Tan M, Deklerck R, Jansen B et al. A novel computer-aided lung nodule detection system for CT images. *Med Phys.* 2011; 38(10): 5630–5645.

[12] Hardie RC, Rogers SK, Wilson T, Rogers A. Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs. *Med Image Anal.* 2008; 12(3): 240–258.

[13] TF Cootes, A Hill, CJ. Taylor, J Haslam. The use of active shape models for locating structures in medical images. *Image Vis. Computing.* 1994; 12(6): 355–366.

[14] Yu-qian, Zhao, etc. *Medical Images Edge Detection Based on Mathematical Morphology.* Proceedings: IEEE Engineering in Medicine and Biology 27th Annual Conference Shanghai, China. 2005: 1-4.

[15] Angenent, S., Eric Pichon, and Allen Tannenbaum. Mathematical Methods in Medical Image Processing. Buletin of the American mathematical society. 2000.

[16] Charles RG, Edward RD. Morphological methods in image and signal processing. *New Jersey: Prentice Hall.* 1988.

[17] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press. 2000.

[18] Burges CJC. Tutorial on support vector machines for pattern recognition. *Data Min Knowl Disc.* 1998; 2(2): 121–167.

[19] Drebin, R, L Carpenter, et al. Volume Rendering. SIGGRAPH.1988: 665-674.

[20] Wang Hongjian. Three-Dimensional Medical CT Image Reconstruction, Chongqing Engeering Technology Research Center for Information Management in Development. Chongqing Technology and Business University, Chongqing, China. 2009; 400067.

[21] Widodo S, Wijiyanto. Texture Analysis to Detect Malaria Tropica in Blood Smears Image Using Support Vector Machine. *International Journal of Innovative Research in Advanced Engineering (IJIRAE).* 2014; 1(8): 301-306.

[22] Mark S. Nixon A and Alberto S. Aguado. Feature Extraction And Image Processing. Second Edition. AcademicPress is an imprint of Elsevier. 2008.

[23] Widodo S, Pipin Widyaningsih. Software Development for Detecting Malaria Tropika on Blood Smears Image Using Support Vector Machine. *International Journal of Engineering Sciences & Research Technology (IJESRT).* 2015; 4(1): 39-44.

[24] Kustanto, Widodo S, Sri Tomo. Software Development to Detect Lung Nodules in Computed Tomography Scan Image Using Support Vector Machine. *International Research Journal of Engineering and Technology (IRJET).* 2015; 2(6): 354-360.

[25] Widodo S, Ratnasari Nur Rohmah, Bana Handaga. *Software Development to Detect Lung Cancer in Computed Tomography Scan Image.* 5th National and International Conference-Research to Serve Society-Huachiew Chalermprakiet University. Thailand: 35-45.