

# USING DISTANCE MEASURE BASED CLASSIFICATION IN AUTOMATIC EXTRACTION OF LUNGS CANCER NODULES FOR COMPUTER AIDED DIAGNOSIS

Maan Ammar<sup>1</sup>, Muhammad Shamdeen<sup>2</sup>, MazenKasedeh<sup>2</sup>, Kinan Mansour<sup>3</sup>, and  
Waad Ammar<sup>3</sup>

<sup>1</sup>AL Andalus University for Medical Sciences, Faculty of Biomed. Eng.,  
Al Kadmous, Syria

<sup>2</sup>Damascus University, Department of Biomedical Engineering, Syria

<sup>3</sup>Al Andalus University Hospital, Al Kadmous, Syria

## **ABSTRACT**

*We introduce in this paper a reliable method for automatic extraction of lungs nodules from CT chest images and shed the light on the details of using the Weighted Euclidean Distance (WED) for classifying lungs connected components into nodule and not-nodule. We explain also using Connected Component Labeling (CCL) in an effective and flexible method for extraction of lungs area from chest CT images with a wide variety of shapes and sizes. This lungs extraction method makes use of, as well as CCL, some morphological operations. Our tests have shown that the performance of the introduce method is high. Finally, in order to check whether the method works correctly or not for healthy and patient CT images, we tested the method by some images of healthy persons and demonstrated that the overall performance of the method is satisfactory.*

## **KEYWORDS**

*Nodules classification, lungs cancer, morphological operators, weighted Euclidean distance, nodules extraction*

## **1. INTRODUCTION**

According to World Health Organization (WHO) statistics, lung cancer cases reached in 2012 one million and eight hundred thousand (13%) of all cancer cases appeared in the world, surpassing breast cancer which came second with 11.9%[1]. Moreover, according to WHO website, there were 1.37 million death cases due to lung cancer in 2008, which means 18% of all death cases due to all cancer types. These facts made lung cancer a major concern for both specialists and scientists. It is well known that early diagnosis can improve the effectiveness of treatment and increase the patient's chance of survival[2]. The previous facts motivated researchers to pay a great attention to researches that work on automated diagnosis of lung cancer in a wide field known as Computer Aided Diagnosis Systems for Lung Cancer. A reliable computer diagnosis of the disease will help screening a large number of images created every day enabling specialized doctors to work with only little amount of candidate images and raising their efficiency[3].

The great attention paid to research on this subject led to more than 300 published scientific papers during the past 3 decades[2]. It motivated also several groups of researchers to write review papers that evaluate the overall situation of research on this subject, identify the challenges, and propose what is needed to improve the performance of the CAD approaches for

lung cancer detection and diagnosis. Heang-Ping Chan et al[3], wrote in 2008 a review paper on Computer Aided diagnosis of Lung Cancer using 158 references spanning 3 decades and finished to the conclusion that the developments in this area are still in early stage. Firmino et al.[4], published in 2014 another review paper on " Computer-aided detection system for lung cancer in computed tomography scans" using 77 references selected from 420 papers found in the famous related databases, and concluded that many, if not all systems, described in their survey have the potential to be important in clinical practice, but further research is needed to improve existing systems and propose new solutions. Bhavanishankar K. et al. [5], wrote a survey paper in 2015 using 68 references spanning almost the previous 2 decades, on "techniques for detection of solitary pulmonary nodules in human lung and their classifications" in an attempt to summarize various methods that have been proposed by several authors over the years of their research on detection, classification and diagnosis of lung nodules. In 2010, M.V. Sprindzuk, et al.[6], published a survey on Lung cancer differential diagnosis of pulmonary nodules detection using 101 references published in the period 2004-2009 and discussed several issues including early diagnosis, and reached an optimistic view of the ability of using the developed systems for different modalities of images, and called for improving the techniques of the image analysis to increase the sensitivity of diagnostic strategies. Finally, and in the same context, Ayman El-Baz, et al.[2], published in 2013 an extensive review paper with 46 pages on "Computer-Aided Diagnosis Systems for Lung Cancer: Challenges and Methodologies" using 364 references spanning 33 years of research. This review discussed extensively the subject of CAD systems for lung cancer using these references, through the four main steps of processing in these systems: segmentation of the lung fields (regions), detection of nodules inside the lung fields, segmentation of the detected nodules, and diagnosis of the nodules as benign or malignant. It addressed several issues like methodologies, training, testing databases and validation of methods used, identified the challenges researchers faced, and the strength and drawbacks of the existing approaches.

Concerning systems for detection and diagnosis of lung cancer nodules, all the review papers we mentioned above called for further improvements in the performance of the available systems by improving the techniques they use. Specifically, reference [2] specified accurate segmentation of lung fields (lungs from chest image) and detecting lung nodules (segmenting and detecting nodules in lung images) as challenges for further investigation in this field.

In this paper, we segment lungs from chest image accurately using what we called "lung mask" produced using 2D Connected Components Labeling 2D-CCL followed by some morphological operations, and a new method to segment lung nodules from extracted lungs image efficiently using 2D-CCL to find regions that are probable to be nodules, then, by extracting shape, texture and density features, necessary training is made, and the nodules are detected and extracted from test data. Finally, a performance consistency-check is made by testing images of healthy persons to show that the program detects no nodules in these cases.

## 2. RELATED WORKS

Different techniques were used by researchers to extract nodules from 2D and 3D, CT images. Using two dimensional CT images, Kaur R., et al. [7] used PCA (Principal Component analysis) to extract nodules from lung cancer CT images, and Miwa T., et al.[8] used morphological N-Quoit Filter to automatically extract nodules based on shape and gray level information. Homma N., et al. [9] used Gabor filter and the difference of pixel values along the object axis to detect nodules, and Gomathi M., et al. [10] used FPCM and extreme learning machine for the same purpose. Those were some sample references from the period (2002-2013). Recently, S. Makaju et al. [11], used in 2018 watershed technique for segmentation and some shape and density features to detect nodules., S. Wang et. al. [12] used in (2020) Residual Neural Networks and N. Khehrahet. al [13] used in (2020) the histogram and some morphological operators to extract the

lung, and a threshold based technique to select candidate nodules. The works mentioned above on detecting lung nodules from 2D chest CT images for Computer Aided diagnosis are naturally not exhaustive but give a good idea about the diversity of techniques and methods used.

Ammar M. et al. [14], used the CCL technique to extract liver area from the complicated 2D abdominal CT image to be used for diagnosis of liver cancer. Based on this experience, the authors explored the possibility of using the same technique for the detection and extraction of lung cancer nodules from 2D chest CT images, and presented the encouraging results they obtained in this paper.

### 3. USED DATA

Images from CT scans of lungs of 11 persons were provided by Alsham Imaging Center and 102 others by Tishreen Hospital. The images in each scan are about 80, and the thickness of each slice is 2 mm. The specialist selected one image from each scan to be used in this study. We divided the 113 images into 2 groups: (1) lungs of 98 cancer patients containing nodules, and (2) lungs of the remaining 15 persons with no nodules (from healthy persons). Both groups were used for lung area extraction from the CT image. For nodule detection, we used the first group for developing and testing the algorithm, and used the second group to check the consistency of the algorithm performance, since the algorithm that detects nodules in the lungs of cancer patients must detect "no nodules" in the images of the lungs of healthy persons. We transformed all the images from DICOM format to JPG format for processing in MATLAB environment. Fig. 1 shows an example from the CT slices of a healthy person, and three other images from CT slices of cancer patients lungs with different number of cancer nodules in each slice (4,8,1, respectively, in raster order).

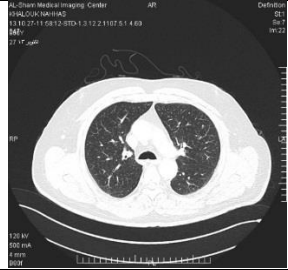

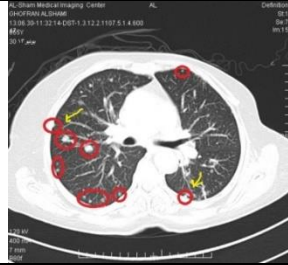
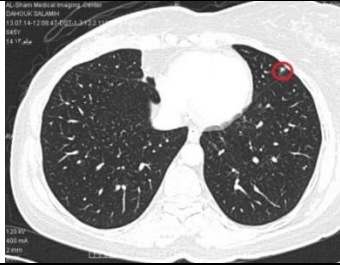
<p><b>A lung CT image of a healthy person</b></p> 	<p><b>A lung CT image of a cancer patient (4 nodules)</b></p> 
<p>534x512</p>	<p>703x512</p>
<p><b>A lung CT image of a cancer patient (8 nodules)</b></p> 	<p><b>A lung CT image of a cancer patient (1 nodule)</b></p> 
<p>548x512</p>	<p>657x512</p>

Fig. 1. A CT image of a healthy person (upper left), a CT image of a lung cancer patient with 4 nodules (upper right), a CT image with 8 nodules (lower left), and a CT image with 1 nodule (lower right). Nodules are marked by circles by specialized doctor.

#### 4. NODULES AUTOMATIC EXTRACTION METHOD

As can be seen in Fig. 1, the original lungs CT image is a complicated content one because it contains, as well as the lungs area, the name of the medical center, the name of the patient, the date, and several other types of information and shapes to help the doctor in diagnosis and archiving. Besides, the lung nodes and the nodules are rather similar in shape, and their gray levels are similar to those of the surrounding region. This situation makes direct extraction of the lungs with their pictorial content from the original image (by thresholding, for example) impossible. Therefore, the general method we use to extract the nodules consists of two main stages, as shown in Fig. 2. Of course, each stage consists of several steps. In the first stage, the lungs area is extracted from the original CT image, and in the second one, the nodules in the lungs area are detected and extracted. If no nodules found in the lungs image, it is marked as "healthy lungs". The output image contains the extracted nodules that can be used later in diagnosis research.

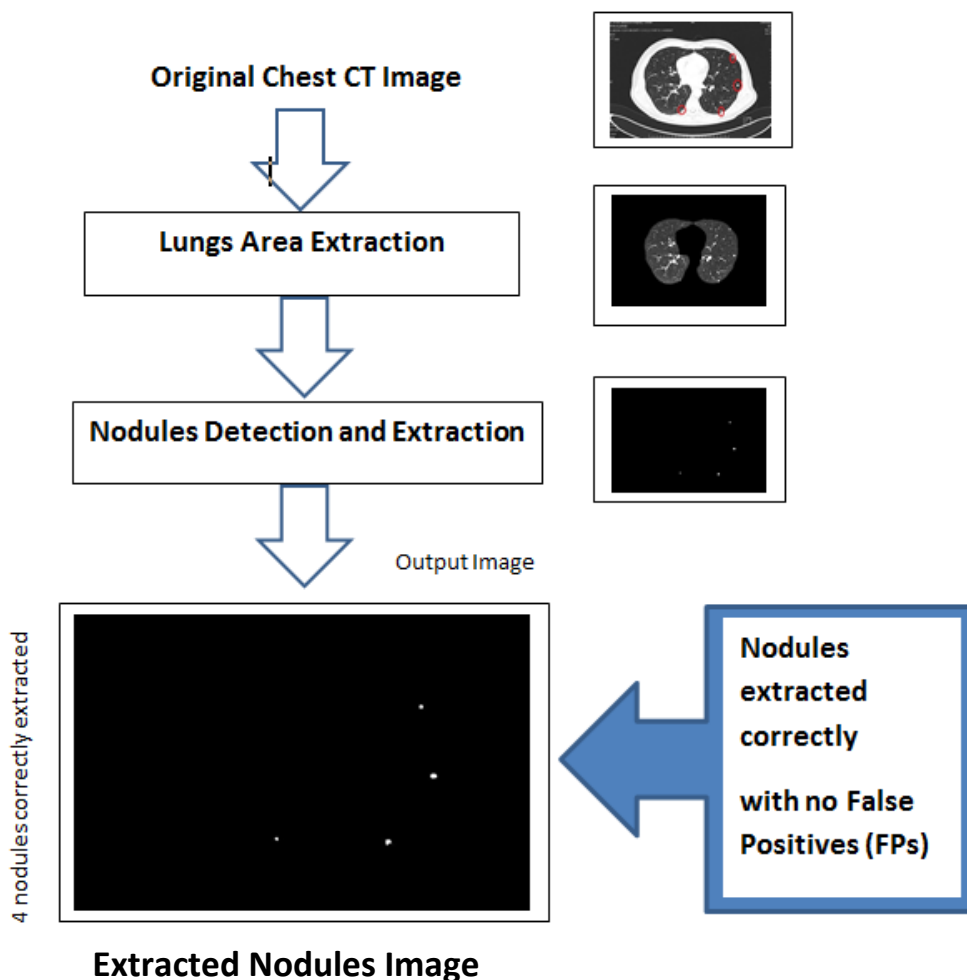


Fig. 2. The two main stages of the general nodules extraction method with results.

##### 4.1. Lungs Area Extraction Stage

The automatic extraction of lungs area from the chest CT image is a complicated process. It is rich of details that we should consider in order to extract the lungs accurately from different CT

shapes and sizes. We will explain this stage below in details due to its importance because accurate extraction of lungs area affects greatly the correct detection and extraction of nodules in the next stage. In this stage, we extract the lungs area from the original chest CT image through two essential processes. In the first one, the known "lungs mask" is extracted through several steps. Then the mask is multiplied by the original image to extract the lungs area from the chest image with the original gray levels preserved. The block diagram in Fig. 3 summarizes this stage, which will be explained below:

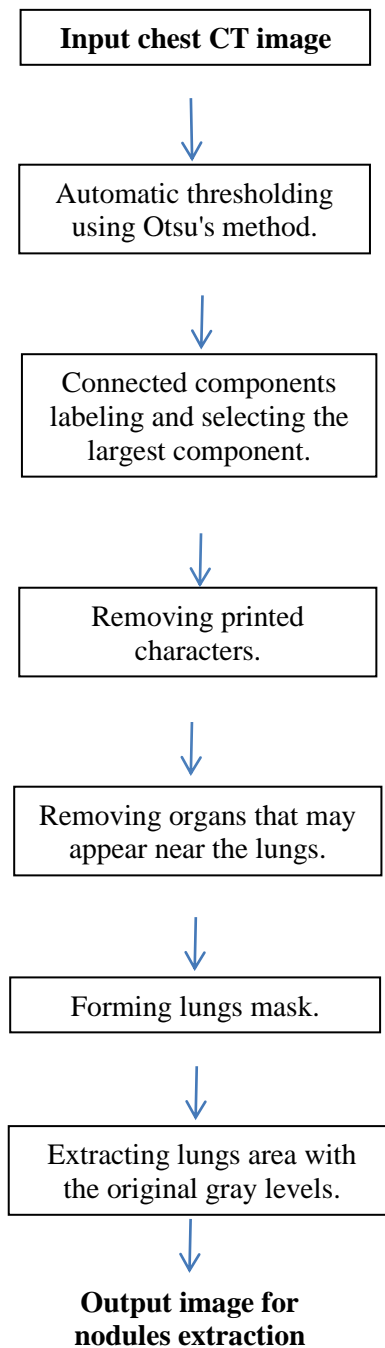


Fig. 3 A block diagram summarizes the lungs area extraction process.

The lungs area extraction process explained briefly in the block diagram in Fig. 3 is detailed below:

1- Thresholding the input original image automatically using Otsu's method [15] to select the threshold, then all pixels values in the image below this threshold are set to zero "0", and the rest ones are set to "1", resulting in a binary image. Fig. 4(b) shows the result of thresholding the original image shown in Fig. 4(a).

2- Labeling the connected components in the thresholded image to give each distinct object (connected component) a unique label that enables us, in principle, to compute all possible kinds of features and use them appropriately to extract lungs area or to distinguish the nodules.

3- Finding the largest component which is the closed region surrounding the lungs, by selecting the component that has the maximum area measured by number of pixels in the thresholded image, Fig. 4(c) shows the largest component obtained from Fig. 4(b).

4- Applying a "closing" process to the complement of Fig. 4(c) using a circular "structuring element" with a diameter (D=8) to remove any remaining printed characters or tiny objects attached to the largest component and to close small holes. The result of this process is shown in Fig. 4 (d).

5 – Applying a hole-filling process to the complement of Fig. 4 (d) to get the complete inner area of the largest component, shown in Fig. 4 (e).

6- Removing any organs may remain near the lung: This is done by applying an "opening" process followed by a "closing" one using a circular "structuring element" with a diameter (D=10). This case does not appear in some CT scans. Fig.4 (g) shows an organ removed from Fig. 4(f) by this step. Note that we use here a different image to show this case, which does not appear in all CT images.

7- Multiplying Fig. 4(d) by Fig. 4(e) to get what we called Lung Mask shown in Fig. 4 (h).

8- Finally, multiplying the Lung Mask by the original image in Fig. 4(a) to extract the lungs area with original gray levels preserved, as shown in Fig. 4 (i).

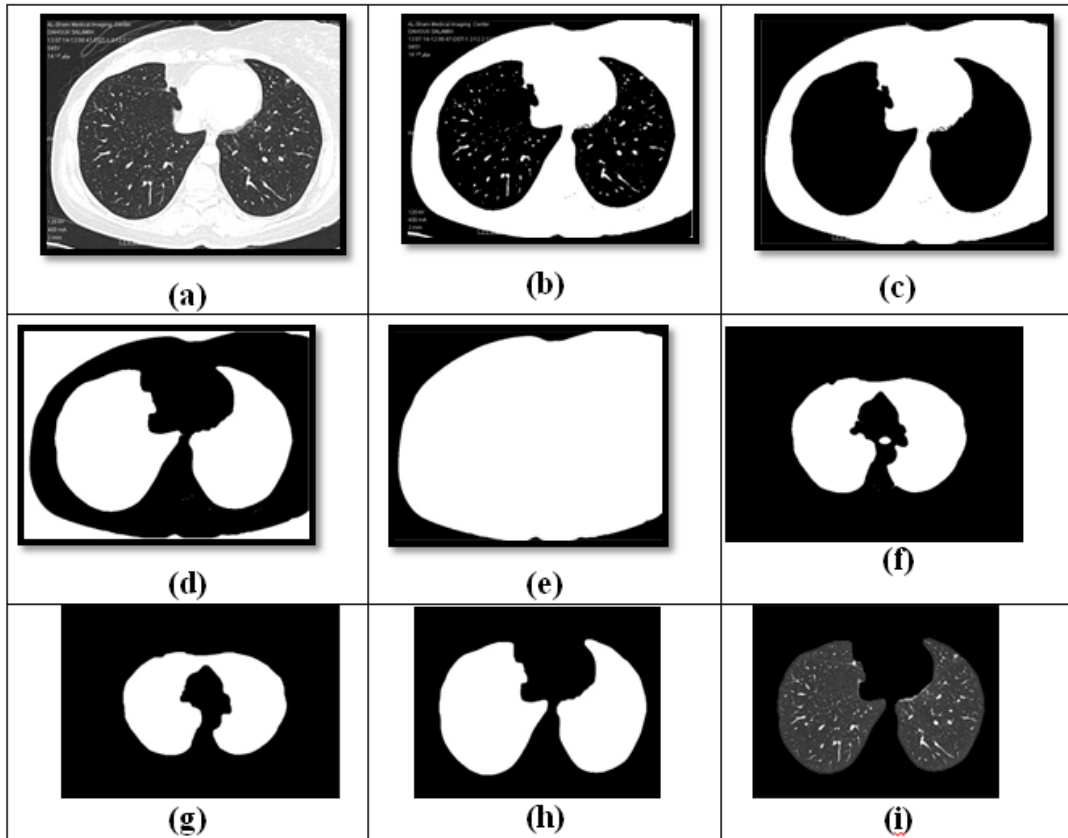
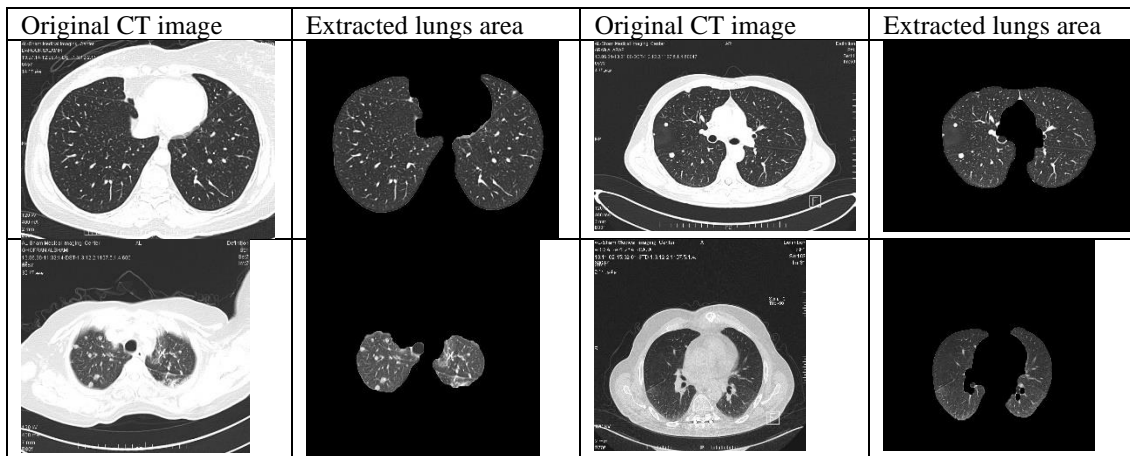


Fig. 4. Sample results of the eight steps lungs extraction process.

The method explained above extracted successfully lungs areas from CT images of different general shapes and complications, as shown in Fig. 5. These results give an idea of the flexibility and effectiveness of the proposed method.



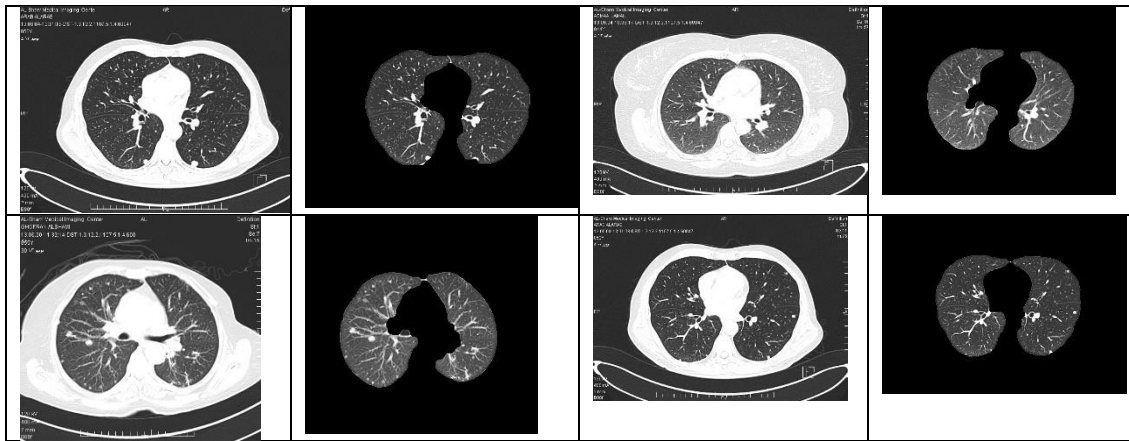


Fig. 5. Sample results of extracting lungs areas from images of different general shapes, sizes, and complication, with very good accuracy.

## 4.2. Extraction of Nodules from Lungs Image

After having the lungs area extracted, we extract nodules through three main stages: (1) Preparing the lungs area image for feature extraction, (2) Extracting features used for classification, and (3) Classifying the regions remaining in the prepared image into "nodules", and "not nodules". Regions classified as "nodules" remain in the resultant image and the others are removed, as shown in Fig. 2. We explain the three stages in the following in necessary details.

### 4.2.1. Preparing the Lung Area for Feature Extraction

Preparing the image of lungs area for feature extraction is done in three steps, as shown in Fig. 6. These steps are:

- (1) Binarizing the image using Otsu's method to select the threshold in the same way used in step 1 of lungs area extraction process explained in section 4.1.
- (2) Labeling connected components in the lungs area binary image.
- (3) Removing small components with areas less than 15 pixels, since as our investigation of all images showed that the regions of such areas are not nodules. Removing these small components will save some of the computation time needed for feature extraction and classification. Fig. 7 shows a sample result with the remaining CCs.



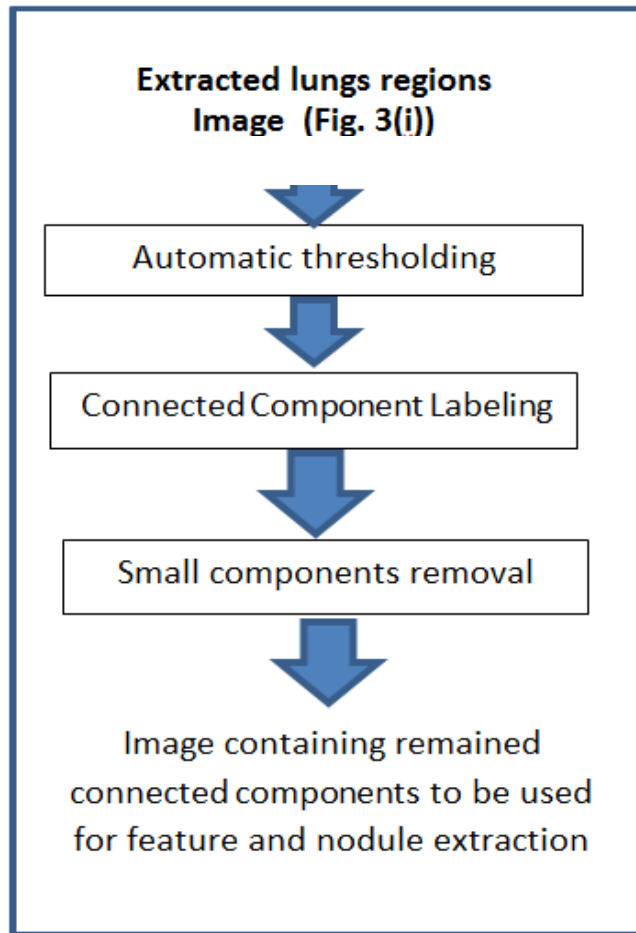


Fig. 6. Preparing the lungs image used for feature extraction.

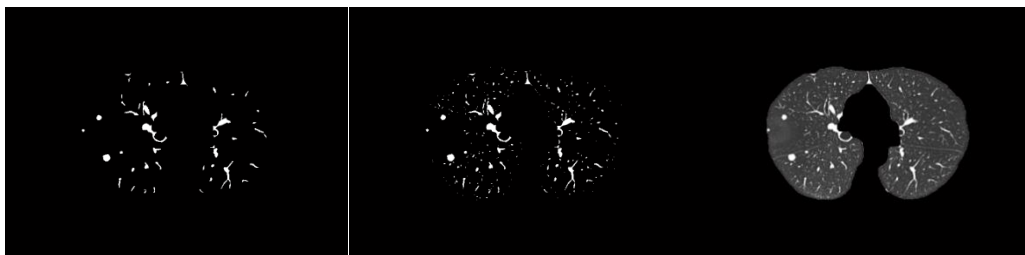


Fig. 7. An extracted lungs image (left), binarized image (middle), remaining CCs after small components removing (right).

#### 4.2.2. Feature Extraction

Specialized doctors diagnosed the nodules used in this study. We found by visual interactive investigation of our data using MATLAB programming facilities that, in general, the nodules have a rather circular shape with almost specific distribution of gray levels, Fig. 8 shows two examples of the interactive examination we made. Therefore, we extracted two groups of features: shape features group and density features group. We will explain the two groups in the following two subsections.

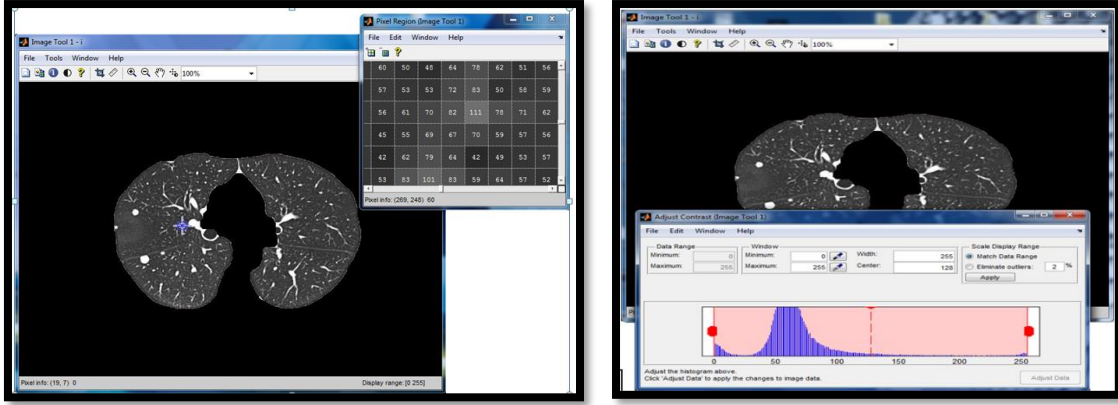


Fig. 8. Two examples (gray level measuring (left) and threshold selection(right)) of the response of the visual interactive image examination using MATLAB facilities we used.

#### 4.2.2.1. Shape features used in the study

After labeling the remaining components in the lung image, we extracted the following shape features for every component:

$$F_{s1}: \text{form factor} = \frac{4\pi \cdot \text{Area}}{\text{Perimeter}} \quad ; F_{s2}: \text{roundness} = \frac{4 \cdot \text{Area}}{\pi \cdot \text{MaxDiameter}^2}$$

$$F_{s3}: \text{solidity} = \frac{\text{Area}}{\text{ConvexArea}} \quad ; F_{s4}: \text{extent} = \frac{\text{TotalArea}}{\text{Area Bounding Rectangle}}$$

$$F_{s5}: \text{compactness} = \frac{\sqrt{(4 \cdot \text{Area})/\pi}}{\text{MaxDiameter}} \quad ; F_{s6}: \text{aspect ratio} = \frac{\text{MaxDiameter}}{\text{MinDiameter}}$$

Where  $F_s$  stands for: shape feature.

#### 4.2.2.2. Density features

Density features are those related to gray levels of image components. Several density features can be extracted from a gray image [16]. We extracted the following ones:

$$F_{d1}: 3^{\text{rd}} \text{ normalized moment: } \mu_2 = 255 * \sum_{i=0}^{255} (z_i - m)^2 p(z_i)$$

$$F_{d2}: 4^{\text{th}} \text{ normalized moment: } \mu_3 = 255 * \sum_{i=0}^{255} (z_i - m)^3 p(z_i)$$

$$F_{d3}: \text{standard deviation: } \sigma = \sqrt{\mu_2}$$

$$F_{d4}: \text{Smoothness measure: } R = 1 - \frac{1}{1 - \sigma^2}$$

$$F_{d5}: \text{similarity measure: } U = \sum_{i=0}^{255} p^2(z_i)$$

$$F_{d6}: \text{entropy: } e = - \sum_{i=0}^{255} p(z_i) * \log_2\{p(z_i)\}$$

Where:  $m = \sum_{i=0}^{255} z_i * p(z_i)$ ,  $F_{d}$ : density feature,  $m$ : average of gray levels;  $z_i$ : value of the pixel at gray level (i);  $p(z_i)$ : value of the histogram at gray level (i).

#### 4.2.3. Detection of nodules from extracted lungs with remaining CCs

The image of the extracted lungs area with the remaining connected components CCs, like that in Fig. 7, contains nodules and not-nodules CCs. Therefore, detecting nodules in this situation

belongs to the Pattern Recognition standard two-classes problem, in which a test sample must be classified whether to belong to the first class (nodule, here), or to the second class (not-nodule). This situation is exactly similar to signature verification problem in which the test signature must be judged whether to be genuine or a forgery. Ammar M. developed a reliable system for signature verification that gives its decision based on a threshold on a Weighted Euclidian Distance Measure computed from suitable features [17]. This system is US patented [18], and still working commercially in hundreds of US banks since about two decades examining about 3 million checks every day. Based on that, this approach was expected to work well in nodule detection. We detected nodules-components by computing the Weighted Euclidean Distance (WED) of each CC in the image using shape and density features explained in the previous two subsections, then using a global threshold obtained from the training group, we classified each CC with a WED less than this global threshold as a "nodule", because the low WED means that the CC resembles the nodules more than the arbitrary CCs. CCs that have a WED equal or larger than the threshold is considered as not-nodule.

In summary, the classification process is done as follows:

- (1) Labeling the components (regions) remained in each image after small components removal.
- (2) Extracting all features mentioned above (shape and density ones) for every component labeled in the image.
- (3) Computing the WED measure for every component which is the Weighted Euclidean Distance to the known cluster learned from group (1).
- (4) Using a global threshold applied to the distance measures of all labeled components, if the distance is less than the threshold, it is judged to be a "nodule", otherwise, it is judged to be "not- nodule", and removed from the image.
- (5) At the end of this process, the image will contain only the components judged as "nodule", like those appearing in Fig. 2.

#### **4.2.3.1. Types of decisions in this classification process and their impact on the results**

There are three types of decision in the classification process mentioned above:

- 1 – True Positive (TP): a component judged to be a "nodule", and at the same time, it is diagnosed by the doctor as "nodule".
- 2 – False Positive (FP): a component judged to be a "nodule", while it is not (not diagnosed by the doctor).
- 3 – False Negative (FN): a component judged to be "not a nodule" and removed from the image, while it is a nodule and diagnosed by the doctor.

**TP is the desired decision because it is the correct one.**

**FP** is not a correct one but it is not dangerous because it gives an alarm that there is a "nodule", but in fact there is no nodule. The negative effect of this decision is that the doctor has to examine the slice image for the availability of a nodule.

**FN** is a wrong decision because it overlooks an actual nodule available in the image. This can be dangerous *if the image contains only the nodule overlooked and removed from the image*, and this nodule is malignant . If there are more than one nodule and only one of them is detected, then this case will not be dangerous because the specialist will examine this slice.

Keeping the above considerations about decisions and their impact on the results in mind, we will discuss some representative experimental results we obtained very closely (at the individual CC level). In this way, we can give the reader an honest and informative view of our results and their analysis.

For view convenience, we will present detailed results using shape features only. Table 1 shows feature values for a test image containing 4 remained CC including one diagnosed as nodule, as well as the distance measure and the doctor diagnosis. "T" in the last column means that the connected component number 3 is diagnosed as "nodule" by the doctor. The WED measure of this CC is appreciably low compared with the other 3 ones which means that it can be easily separated from them by a threshold, and consequently making a TP decision.

Table 1: Shape feature values, WED measure, and doctor diagnosis of connected components of an image in the test data with 4 remaining components after small ones removal.

CC #	Form Factor	Roundness	Solidity	Extent	Compactness	Aspect ratio	Distance	Doc. Diag.
1	12.0391719	1001.60565	0.833333	0.535714	31.64815402	2.188309697	1.539006	
2	16.5141083	432.901445	1	0.9375	20.80628379	1.136570857	1.435795	
3	31.7389353	23175.2454	0.886957	0.653846	152.2341793	1.281090021	0.201863	T
4	17.431828	7495.76987	0.82	0.414141	86.57811425	2.48330078	1.306872	

#### 4.2.3.2. A Discussion inside the Classification Process

We will now proceed to more complicated case in which there are 40 remaining CCs and 5 diagnosed nodules (number: 12, 19, 34, 39, 40 in Table 2). Intuitively, the separation process between diagnosed nodules and the other remaining CCs will be more difficult here. Table 2 shows the same information as in Table 1, just explained above, but for another image, naturally.

If we consider this table, and consequently its related image alone, we find that a threshold value  $TH=1.39$  on the WED measure will give the following result:

TP=100%, FP=6, FN=0.

This is a very excellent result since all diagnosed nodules are correctly detected, with 6 FPs and no FNs. False positives in this case do not lay any new burden on the specialist because he must examine this slice any way.

If we have to set one threshold for both images represented in Table 1 and Table2, we find that this threshold ( $TH=1.39$ ) works well but we will have one more FP from Table 1. The result in this case becomes:

TP=100%, FP=7, FN=0.

Now if we want to minimize FPs, we may make  $TH=1.27$ . This will make the results using the two slices (two persons) containing 6 diagnosed nodules as follows: TP=83%, FP=5, FN=1, that is because nodule number 4 in Table 1 and number (25, 31, 32) in Table 2, will not be FP. On the other hand, we have in this case 1 FN (number 12 in Table 2). However, this FN is not actually a

problem because 4 nodules are correctly detected and consequently the doctor must examine this slice any way.

Moreover, if we wish to consider more slices (persons) using the same threshold (TH) for general decision, we recognize that we must make a compromise between TP, FP, and FN, according to some considerations we set.

For the final result we reported below in Table 3, we selected (automatically) the threshold on the WED measure computed using shape features, and both of shape and density features, for all connected components remained in images after removing small ones, so that we got the best result.

Table 2: Shape feature values, distance measure and doctor diagnosis of connected components of another image in the test data with 40 remaining components after small ones removal.

CC #	form factor	roundness	Solidity	Extent	Compactness	Aspect ratio	Distance	Doc. Diag.
1	10.4249357	3446.79225	0.7	0.291667	58.70938805	3.259510526	2.030108	
2	17.4637494	119186.896	0.635714	0.317857	345.2345521	7.007055113	6.696163	
3	12.9455689	68606.4649	0.521368	0.260684	261.9283583	7.197314177	5.289446	
4	13.8022678	977.054729	0.9375	0.625	31.25787467	2.468279653	1.617706	
5	10.2907557	9158.48268	0.6	0.259615	95.69996176	4.350863356	2.631402	
6	12.5218755	21887.498	0.625	0.350877	147.9442394	5.307419623	3.217701	
7	19.9771266	80344.6825	0.807339	0.285714	283.4513759	5.893114633	4.7493	
8	18.4375056	6146.43619	0.904762	0.575758	78.39921041	2.420066516	1.261981	
9	31.6384011	17473888.8	0.251653	0.147641	4180.178087	2.387995592	834.8104	
10	22.0959009	12503.9979	0.929825	0.552083	111.8212766	2.56996834	1.124257	
11	14.4832858	1526.04864	0.904762	0.542857	39.06467248	2.372396484	1.534869	
12	13.8022678	631.237785	0.9375	0.6	25.12444596	1.444622811	1.382321	T
13	14.259656	5131.94222	0.878788	0.358025	71.63757549	3.355926203	1.955388	
14	24.5138915	1021218.48	0.501157	0.32003	1010.553552	1.417501282	48.28374	
15	15.3966423	709.030683	1	0.833333	26.62763008	1.869131001	1.442422	
16	11.8582659	3916.56401	0.846154	0.366667	62.58245769	4.424257992	2.719478	
17	9.74569055	7612.43347	0.621622	0.547619	87.24926058	5.988500393	3.786939	
18	10.2644282	14561.2714	0.535714	0.375	120.6700933	5.413293088	3.32917	

19	18.1783462	1143.61606	0.958333	0.638889	33.81739285	1.242003801	1.267455	T
20	12.6698853	2963.42367	0.875	0.7	54.43733705	3.592471641	2.14353	
21	19.0638325	4343.9594	0.897436	0.648148	65.90872018	1.912235477	1.114146	
22	14.2623434	7875.69015	0.815789	0.430556	88.74508522	4.470520541	2.639714	
23	9.77406321	8358.06802	0.520833	0.357143	91.42246999	4.193451023	2.521288	
24	37.7743416	2065085.55	0.460245	0.328423	1437.040553	1.747152727	98.21622	
25	18.3437668	1030.94052	0.956522	0.733333	32.10826243	1.249302461	1.280673	
26	20.7676771	145079.296	0.582938	0.297101	380.892762	3.19479427	6.443388	
27	16.2469754	3922.51568	0.875	0.7	62.6299903	2.888039732	1.636473	
28	15.6814289	1901.00829	1	0.875	43.60055382	2.498670381	1.576498	
29	13.4273865	14015.7852	0.698113	0.256944	118.3882813	4.452181394	2.616838	
30	22.6416841	61562.2386	0.679245	0.406015	248.1173887	2.282543301	2.328933	
31	17.4495279	610.07542	1	0.85	24.69970486	1.263164345	1.365554	
32	23.3841595	12049.1814	0.916667	0.555556	109.7687634	2.367511703	0.988834	
33	17.3240175	26716.5399	0.707865	0.375	163.4519499	3.268363446	1.702242	
34	25.4311313	7373.25296	0.835616	0.61	85.86764791	1.115704069	0.770938	T
35	11.7170336	180720.806	0.57971	0.4	425.1126974	11.34107117	11.16144	
36	26.8874838	16670.4559	0.985294	0.736264	129.1141199	2.252758863	0.791035	
37	20.657524	4675.95165	0.972973	0.72	68.38093042	2.164376577	1.18298	
38	17.3879834	13401.8703	0.862745	0.366667	115.7664473	3.851118966	2.110713	
39	23.1133404	2157.21731	1	0.809524	46.44585348	1.127255908	1.161356	T
40	30.6282678	9547.1584	0.985507	0.85	97.70956144	1.256972084	0.668926	T

#### 4.2.3.3. Training

The training to obtain the global threshold is done on 30 images among the 98 ones containing nodules diagnosed by the specialized doctor. If we refer to Table 1 again, it is obvious that the CC number 3 is easily separable by a threshold from the other CCs due to the considerable difference in the WED value, and can be detected as a nodule. This is an example, but the complete training was done using all shape features, and both of shape and density features in 30 images from the 98 images containing nodules used in this study. It is worth noting that the

number of remaining CCs in an image reaches 41 in some images. We have already shown the results of 40 remaining CCs in Table 2.

## 5. RESULTS AND DISCUSSION

After training on the 30 images using shape features once, and both shape and density features once again to obtain the global threshold that separates the nodules from the other CCs in each case, we obtained the results shown in Table 3, where: TP is "True Positive", "FP" is "False Positive", "FN" is False Negative, and the "Sensitivity" is computed by this equation:  $Sensitivity = TP / (TP + FN)$ .

Table 3. Results of the nodules detection using shape and density features on test cases of 68 images.

Feature kind	Sensitivity%	FPs (per case)	No. CCs	No. of Nodules Diagnosed by doctor
Shape features	95.1	2.1	1836	173
Shape +Density features	97.2	1.98	1836	173

We can conclude from Table 3 that: the average diagnosed nodules is about 2.54 per image, and the average number of remaining CCs in an image before detection and extraction is 27. Using the density features with shape ones improves the sensitivity by about 2%, and the improvement in FPs reaches about 6%.

### 5.1. Performance consistency check

As a consistency check of the performance of the nodule detection and extraction algorithm, we tested the lung images of the 15 healthy persons as a consistency-test-group. The result was excellent where 9 false positives (FPs) appeared in 8 images (4% of CCs, as an average). Fig. 9 shows the result for a test image containing seven nodules where all detected with one FP, and the second is for a healthy person where no nodules are detected, and consequently will be classified as "healthy lung".

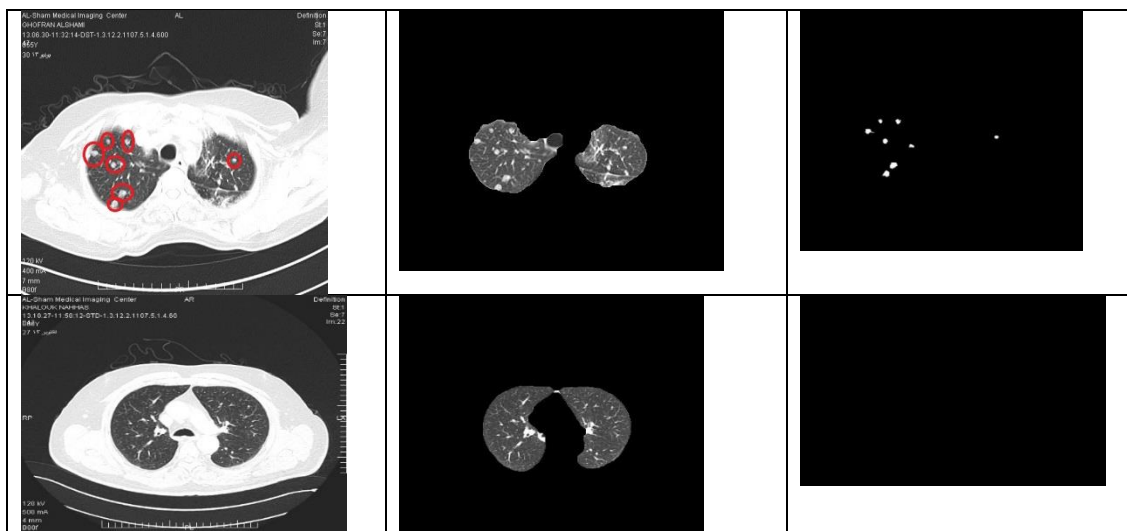


Fig. 9. The first column is the input image, the second column is the extracted lungs image, and the third column is the result of detection and extraction of nodules. In the first row: TPs=7, and FPs=1. In the 2<sup>nd</sup> row: TPs=0, FPs=0, healthy person (perfect result).

### 5.2. Comparison with some other works

We compared our method with 9 works conducted in the period 2010-2020 as shown in Table 4 below. We can see in this table that the introduced method gave a general performance comparable with the best results reported, with little bit higher sensitivity, and the FPs rate is among the low ones.

Table 4. Performance comparison between other works and the introduced method.

Authors	year	Sensitivity%	FPs per case	No. nodules
Liu [19]	2010	97	4.3	32
Cascio [20]	2012	97	6.1	148
Orozco [21]	2012	96.15	2	50
Teramoto [22]	2013	80	4.2	103
Shao [23]	2012	89.47	11.9	44
Bergtholdt [24]	2016	85.90	-	
Wu [25]	2017	79.23	-	
Saien [26]	2018	83.98	0.02	
Khehrah [12]	2020	93.75	0.13	
Monif(our work)	2021	97.2	2.1	173

### 6. FURTHER INVESTIGATION AND FUTURE WORK

In an attempt to find some way to improve the performance of the method further, we examined the images containing nodules and compared them with the numerical results. We found that a considerable part of wrong decisions is related to what we can call Nodules Attached to the Shield (NAS), where NAS is explained as follows: In Fig. 10 below, we can see in the upper right image two nodules diagnosed by the specialized doctor (surrounded by circles). Those two nodules are attached to the shield (white area surrounding the lungs) and called "NAS".

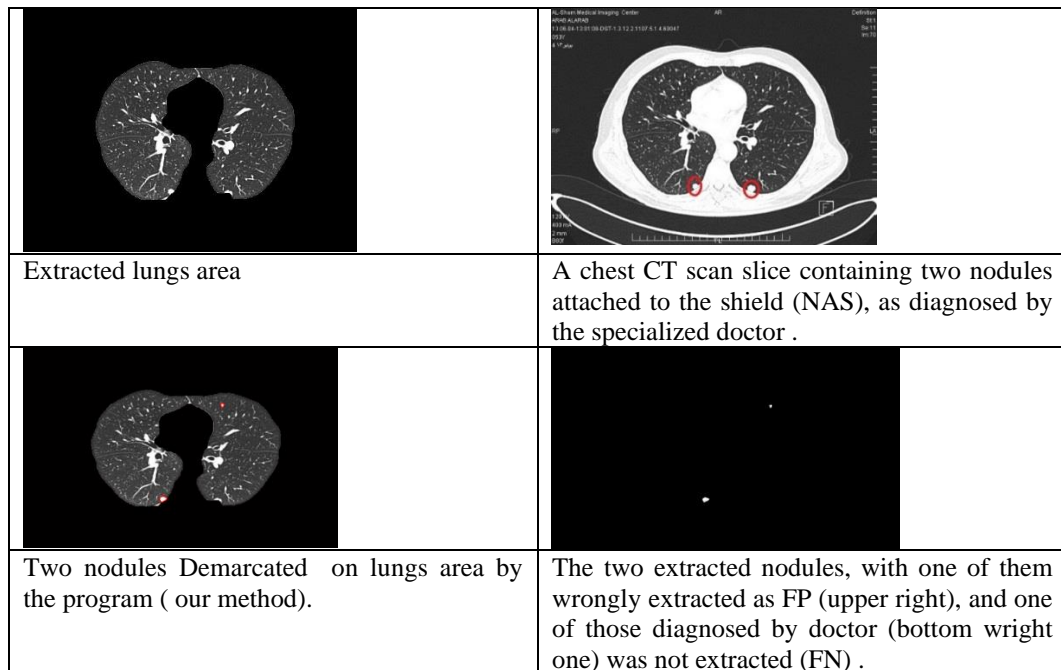


Fig. 10. Two nodules extracted by the method in the right bottom image, one of them (upper one) is FP. And one of the two nodules diagnosed by the doctor was not detected (FN).



We plan to work on improving the performance in two directions:

- 1 – Improving the segmentation of the NAS,
- 2 – Improving the correct classification rate ( sensitivity ) by using the feature selection technique developed by Ammar M. [17].

## 7. CONCLUSIONS

We have introduced a method for the automatic detection and extraction of lungs cancer nodules from CT images using connected component labeling (CCL) and weighted Euclidian distance measure based classification. The obtained results have shown clearly the high performance of the method in both extraction of lungs area from the CT image and in the correct detection of the cancer nodules. This high performance was also supported by the consistency check we made which reveals the ability of the method to detect the healthy image at the same time.

The experiments conducted and their results analysis presented in this paper enable us to conclude that using CCL technique with appropriate sequence and parameters of some morphological operations may lead to a high performance approach for extracting lungs areas from complex chest CT image with a wide variety in shapes and sizes. We can conclude also that using CCL technique with the high flexibility it offers in manipulating CCs in the extracted lungs areas, with the usage of WEDM and a threshold based classification, may provide an efficient method for detecting and extracting nodules to be used later for diagnosis. We found also that using density features with shape ones improves to some extent the performance. Our further investigation have also shown that improving the segmentation of NAS may improve the final results. We have set this point as a future work.

We hope that we have introduced a positive effort in the general direction of the research for building an actual automatic lung cancer detection and diagnosis systems.

## ACKNOWLEDGMENTS

The authors wish to thank Alsham Medical Imaging Center and Tishreen Hospital for providing the lung images data. We appreciate also the support provided by Al Andalus University for Medical Sciences and its hospital.

## REFERENCES

- [1] WHO, Latest world cancer statistics, (2013), The International Agency for Research on Cancer Publications, Geneva: World Health Organization.
- [2] Ayman El-Baz, et al.,(2013) Computer-Aided Diagnosis Systems for Lung Cancer: Challenges and Methodologies, International Journal of Biomedical Imaging, Volume 2013, Article ID 942353.
- [3] Heang-Ping Chan et al, (2008) Computer-Aided Diagnosis of Lung Cancer and Pulmonary Embolism in Computed Tomography — A Review, Acad. Radiology, 15(5): 535–555.
- [4] Sprindzuk M.V., et al., (2010) Lung cancer differential diagnosis based on the computer assisted radiology: The state of the art, Pol J Radiology, 75(1): 67–80.
- [5] Firmino et al., (2014) Computer-aided detection system for lung cancer in computed tomography scans: Review and future prospects, BioMedical Engineering, 13:41.
- [6] Bhavanishankar K. et al., (2015) techniques for detection of solitary pulmonary nodules in human lung. and their classifications-A survey, International Journal on Cybernetics & Informatics (IJCI) Vol. 4, No. 1.

- [7] Kaur R., and Ada S. (2013), "Feature Extraction and Principal Component Analysis for Lung Cancer Detection in CT scan Images", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3, Issue 3, pp.187-190.
- [8] Miwa T., Kako J., Yamamoto S., Matsumoto M., Tateno Y., Inuma T., and Matsumoto T. (2002), "Automatic Detection of Lung Cancers in Chest CT Images by the Variable N-Quoit Filter", *Systems and Computers in Japan*, Vol. 33, No. 1, pp.53-63.
- [9] Homma N., Takei K. and Ishibashi T. (2008), "Combinatorial Effect of Various Features Extraction on Computer Aided Detection of Pulmonary Nodules in X-ray CT Images", *INFORMATION SCIENCE & APPLICATIONS*, Issue 7, Vol. 5, pp.1127-1136.
- [10] Gomathi M., and Thangaraj P. (2010), "A Computer Aided Diagnosis System For Detection Of Lung Cancer Nodules Using Extreme Learning Machine", *International Journal of Engineering Science and Technology*, Vol. 2, pp. 5770-5779.
- [11] Wang S., Dong L., Wang X., Xin Wang, *Open Med.* (2020) Classification of pathological types of lung cancer from CT images by deep residual neural networks with transfer learning strategy.
- [12] Khehrah N., Farid M. S., Bilal S. and Khan M. H., (2020) Lung Nodule Detection in CT Images Using Statistical and Shape-Based Features, *J. Imaging*, 6, 6 (14 pages).
- [13] Makaju S., Prasad P.W.C., Alsadoon A., Singh A. K. and Elchouemi A., (2018) Lung Cancer Detection using CT Scan Images, *Procedia Computer Science* 125- 107–114.
- [14] Ammar M., et al., (2011) Using Image Processing Techniques for Automatic Extraction of Liver Suspicious Regions from X-Ray Computed Tomography Images, *Tishreen University Journal for Research and Scientific Studies - Engineering Sciences Series*, Vol. (33) No. (3), pp.(217-235).
- [15] Nobuyuki Otsu (1979). "A threshold selection method from gray-level histograms". *IEEE Trans. Sys., Man., Cyber.* 9(1): 62–66.
- [16] Gonzalez, R., Wood, E. and Eddins, S. (2009) "Digital Image Processing Using MATLAB", Second Edition, Gatesmark Publishing.
- [17] Ammar M., Raising the Performance of Automatic Signature Verification Over that Obtainable by Using the Best Feature Set, (2011) *International Journal of Pattern Recognition and Artificial Intelligence*. 25-2, PP 183-206.
- [18] Ammar M., Method and apparatus for verification of signatures, United States Patent: No. 6424728 , 07/23/2002, U.S.A.
- [19] Liu Y, Yang J, Zhao D, Liu J (2010) A method of pulmonary nodule detection utilizing multiple support vector machines. In *Computer Application and System Modeling (ICCASM)*, International Conference On, vol. 10. Taiyuan; 10–11810121.
- [20] Cascio D, Magro R, Fauci F, Iacomi M, Raso G (2012) Automatic detection of lung nodules in CT datasets based on stable 3d mass-spring models. *ComputBiol Med*, 42(11):1098–1109.
- [21] Orozco HM, Osiris Vergara Villegas O, Maynez LO, Sanchez VGC, de Jesus Ochoa Dominguez H (2012) Lung nodule classification in frequency domain using support vector machines. In *Information Science, Signal Processing and Their Applications (ISSPA)*, 11th International Conference On. Montreal, QC; 870–875.
- [22] Teramoto A, Fujita H (2013) Fast lung nodule detection in chest CT images using cylindrical nodule-enhancement filter. *Int J Comput Assist RadiolSurg*, 8(2):193–205.
- [23] Shao H, Cao L, Liu Y (2012) A detection approach for solitary pulmonary nodules based on CT images. In *Computer Science and Network Technology (ICCSNT)*, 2nd International Conference On. Changchun; 1253–1257.
- [24] Bergtholdt, M.; Wiemker, R.; Klinder, T. (2016) Pulmonary nodule detection using a cascaded SVM classifier. In *Proceedings of the Medical Imaging: Computer-Aided Diagnosis*, San Diego, CA, USA, 27 February–3March 2016; Volume 9785, pp. 268–278.
- [25] Wu, P.; Xia, K.; Yu, H. (2016) Correlation coefficient based supervised locally linear embedding for pulmonary nodule recognition. *Comput. Methods Programs Biomed.* 136, 97–106.
- [26] Saien, S.; Moghaddam, H.A.; Fathian, M. (2018) A unified methodology based on sparse field level sets and boosting algorithms for false positives reduction in lung nodules detection. *Int. J. Comput. Assist. Radiol. Surg.*, 13, 397–409.

## AUTHORS

**Maan Ammar** Ph. D. in Information Engineering, Nagoya University, Japan, 1989, Professor at Al Andalus University for medical sciences, Biomedical engineering since 2014, Full professor at Applied Sciences University, Amman Jordan 2003, US patent of a commercial system serving hundreds of US banks since 2002 "Method and apparatus for verification of signatures", United States Patent: No. 6424728 , 07/23/2002, U.S.A. Published many papers in image processing and pattern recognition fields. Served as Head of biomedical engineering department–Damascus University for 8 years.



**Muhammad Shamdeen** Received B. SC in Biomedical Engineering in 2014 from Damascus University. Now he is a Master student in the biomedical department at Damascus University. Working in the field of medical image processing.



**MazenKasedeh** Received B. Sc in Biomedical Engineering in 2014 from Damascus University. Now he is a Master student in the biomedical department at FMEE. He is Interested in medical image processing.



**Kenan Mansour** MD Obstetrics and Gynaecology Specialist, Tishreen Hospital, 2011. Medical Manager of Al Andalus University Hospital, 2019 to present. Medical Education Master student, Syrian Virtual University, 2020. Interested in medical image analysis and diagnosis.



**Waad Ammar** MD General Surgery Specialist, Tishreen Hospital, 2020. At present, working at Al Andalus University Hospital. Medical Education Master student, Syrian Virtual University, 2020. Interested in medical image analysis and diagnosis.

