A CAD System for the Early Detection of Lung Nodules Using Computed Tomography Scan Images

https://doi.org/10.3991/ijoe.v15i04.9837

Hanan M. Amer (✉)
Mansoura University, Mansoura, Egypt
hanan.amer@yahoo.com.

Fatma E. Z. Abou-Chadi
The British University in Egypt, Cairo, Egypt

Sherif S. Kishk, Marwa I. Obayya
Mansoura University, Mansoura, Egypt

Abstract—In this paper, a computer-aided detection system is developed to detect lung nodules at an early stage using Computed Tomography (CT) scan images where lung nodules are one of the most important indicators to predict lung cancer. The developed system consists of four stages. First, the raw Computed Tomography lung images were pre-processed to enhance the image contrast and eliminate noise. Second, an automatic segmentation procedure for human's lung and pulmonary nodule candidates (nodules, blood vessels) using a two-level thresholding technique and morphological operations. Third, a feature fusion technique that fuses four feature extraction techniques: the statistical features of first and second order, value histogram features, histogram of oriented gradients features, and texture features of gray level co-occurrence matrix based on wavelet coefficients was utilised to extract the main features. The fourth stage is the classifier. Three classifiers were used and their performance was compared in order to obtain the highest classification accuracy. These are; multi-layer feed-forward neural network, radial basis function neural network and support vector machine. The performance of the proposed system was assessed using three quantitative parameters. These are: the classification accuracy rate, the sensitivity and the specificity. Forty standard computed tomography images containing 320 regions of interest obtained from an early lung cancer action project association were used to test and evaluate the developed system. The images consists of 40 computed tomography scan images. The results have shown that the fused features vector resulting from genetic algorithm as a feature selection technique and the support vector machine classifier give the highest classification accuracy rate, sensitivity and specificity values of 99.6%, 100% and 99.2%, respectively.

Keywords—Image processing, histogram thresholding, lung area segmentation, histogram of oriented gradients, contrast enhancement, nodule candidates identification, radial basis function neural network, discrete wavelet transform, genetic algorithm, support vector machine.

1 Introduction

Lung cancer has become one of the most important diseases that pose a great threat to humanity because of the high rates of air pollution, the spread of smoking in recent years and the difficulty of treatment. Developing early detection of this disease has become the concern of scientists in medical fields [1].

Early detection of lung cancer increases the chance of survival of the patient for a period of up to 5 years by up to a percentage of 70%, as well as it increases the chance of success of treatment whenever diagnosed in the early stages, this led to the increasing importance of work on the development of early detection systems [1].

One of the most accurate techniques used in the diagnosis of lung cancer is Computerized Tomography (CT) of the patient's chest, because it allows lung imaging on many sections, which results a large number of images, enabling radiologists and physicians to examine all parts of the lung [1]. But this large number of images resulting from the CT examination in addition to the use of low radiation doses to protect the patient from the risk of exposure to large amounts of radiation, made the examination of these images by a radiologist difficult and onerous task [1].This motivated scientists to develop computerized systems that process and analyze these images and allow automatic determination of the presence of pulmonary nodules. These systems are known as Computer-Aided Detection (CAD) systems [2].

In general, any CAD system for the detect the presence of pulmonary nodules automatically is composed of the following four stages: a preprocessing stage for image contrast enhancement and noise reduction, the automatic segmentation stage that aims to extract the human's lung area and nodules followed by a feature extraction procedure of the pulmonary nodule candidates and the final stage is the classification [2]. Figure 1 illustrates the main stages of processes in CAD system.

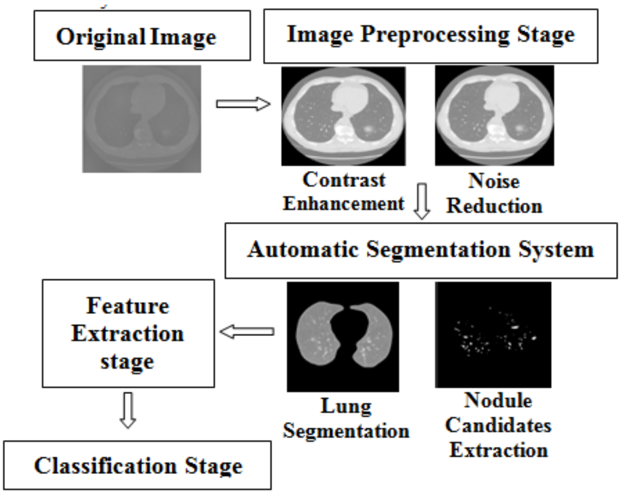


Fig. 1. The basic stages of CT-lung CAD system.

The accurate extraction of the lungs from the CT chest images is an essential step in the CAD systems. Techniques previously reported in lung segmentation are based on intensity variation (thresholding methods), image region (merge region, split region, and the region growing techniques), and others that are based on the object texture, motion tracking, and edge detection [2]. In the present work, a Novel Image Size Dependent Normalization Technique (ISDNT) was adopted.

According to clinical opinions of physicians, the blood vessels and pulmonary nodules are presented in the CT scan image as having lower contrast values and higher gray values [2].Several attempts were reported for nodule extraction. The thresholding techniques [2] where the extraction of the nodule candidates is based on the intensity variation between the lung parenchyma and the nodule candidates were utilized.

For feature extraction, four types of feature extraction techniques [3, 4] were utilized: the Histogram of Oriented Gradients (HOG) features the statistical features, the texture features of Gray Level Co-Occurrence Matrix (GLCM) based on wavelet coefficients, and the Value Histogram (VH) features.

Classification approaches have been proposed such as Artificial Neural Networks (ANN), linear discriminate analysis classifier, rule-based, Bayesian classifier, support vector machine (SVM), and k-NN [5]. In the present work ANN, SVM, and Radial Basis Function Neural Network (RBF-NN) were used.

To increase the accuracy of the classification, fusion technique was used. It can be classified into three different levels, namely, data fusion at the level of data, feature fusion at the level of features, and decision fusion at the decision level [6]. A fusion step at the feature level has been adopted.

The performance of the developed system is compared with that of previous reported classifiers: ANN classifier, RBF-NN classifier and SVM classifier [5]. The paper is organized as follows: Section 2 is a description of the dataset used. Section 3 describes the different stages the proposed system. Section 4 discusses the results and Section 5 is the final conclusion.

2 The Dataset

Forty CT scans containing 320 regions of interest (ROI) were made available from the Early Lung Cancer Action Project (ELCAP) association [7]. The images in this database are available in format of Digital Images and Communication in Medicine (DICOM) and have a resolution of 0.76×0.76×1.25. The size of pulmonary nodules that were considered in this work varies from 3 mm to 30 mm. Figure 2 shows a typical example of the chest CT images.

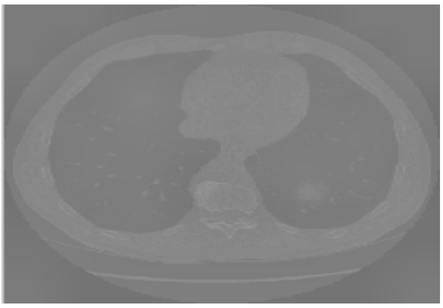


Fig. 2. A typical example of CT lung image from ELCAP database image.

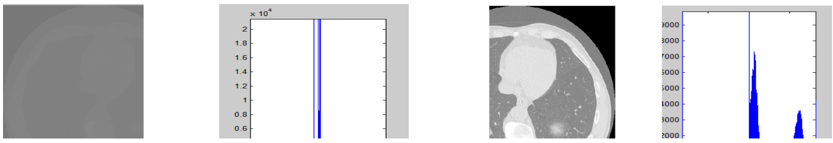
3 Methodology

The developed CAD system consists of the following four stages:

3.1 Image pre-processing

Physicians use a low radiation doses during the CT scan to protect the patient from the risk of exposure to large amounts of radiation but this leads to low-resolution images. On the other hand, processing the CT scan itself is accompanied by the exposure of images to noise from different sources which reduces the image quality. Preprocessing was accomplished in two steps: enhancing the image contrast and denoising of the CT chest image.

Image contrast enhancement: Enhancement of the image contrast of CT scan images increases the accuracy of nodule detection. Hence, a comparative study of three image contrast enhancement techniques; histogram equalization, adaptive Histogram Equalization, and a novel Image Size De-pendent Normalization Technique (ISDNT) [8] were utilized. The visual comparison of the contrast enhanced images showed that the ISDNT technique gives the best results. Figure 3 shows the CT image before and after contrast enhancement using the ISDNT.



(a)

b)

Fig. 3. (a st image and (b) the results obtained

Denoising of the CT lung images: Artifacts decrease of the quality of CT images. A previous study [13] compared the performance of six image denoising techniques: Gaussian filter, average filter, weighted average filter, Wiener filter, median filter, and wavelet filter and concluded that the Weiner filter gives the best results [9]. Figure 4 shows an example of a denoised image using Wiener filter.

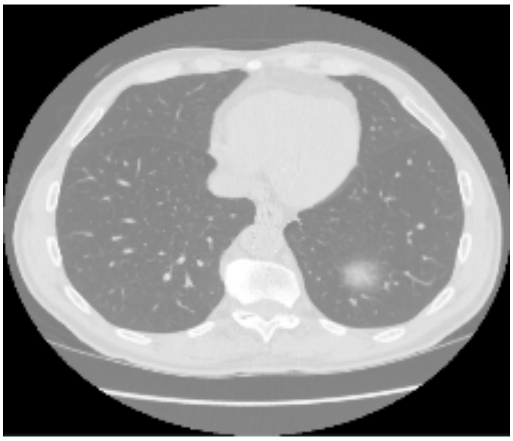


Fig. 4. A denoised CT lung image

3.2 Lung segmentation

Having preprocessing CT chest images, the next step is to extract the human lungs area from CT chest images. The proposed algorithm of human lungs segmentation [10] consists of three main steps; calculation of an optimal threshold, then segment the thorax from the background, and finally segment the lungs.

To calculate the optimal gray-level threshold, a diagonal gray-level histogram was constructed using the diagonal pixels intensity of all CT chest images of a complete scan. The resulted histogram was found to have three clear peaks (Figure 5). This is a common feature in all CT chest scan images. These peaks represent the following; peak P1is formed from black background pixels intensity, peak P2is formed from the pixels of low intensity representing the external region that surrounds the thorax area and the internal parenchyma of lung, and peak P3is formed from the pixels of high intensity which represent blood vessels, bones of the rib cage, heart, and pulmonary nodules. Accordingly, the choice of a gray-level point that divides the distance between the second and third peaks equally as an optimal gray-level threshold was used in the present automatic segmentation work. The optimal gray-level threshold is calculated according to the following equation;

$$L = (P_2 + P_3)/2. \quad (1)$$

Segment the thorax from background: The thorax extraction includes the removal of all image components external to the chest area. First, the bi-level thresholding technique was applied to obtain a binary image. Then a morphological operation and median filter of size 15*15 were applied to obtain the thorax binary mask which will be multiplied with the preprocessed image to get the segmented thorax area. Figure 6 shows the steps in detail.

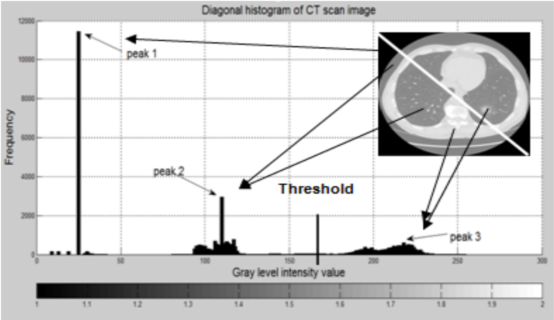


Fig. 5. The diagonal gray-level histogram of CT scan images.

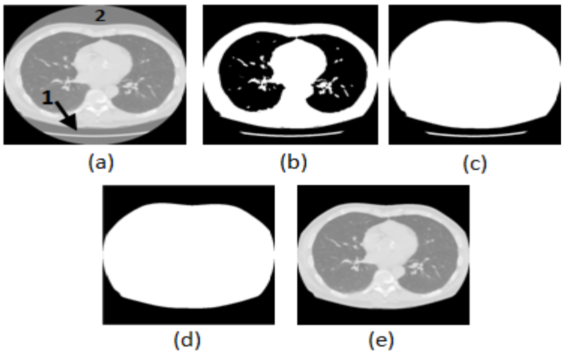


Fig. 6. The steps performed to segment the thorax (a) The preprocessed image, (b) The binary image, (c) The filled image, (d) The filtered image, (e) The segmented thorax.

Segment the lungs within the thorax: The goal of this step is to separate human lungs area from the thoracic area. To extract the lung area, the bi-level thresholding technique was applied to obtain a binary image. Then the morphological operations and a median filter were used to obtain the lung binary mask. This was multiplied with the thorax image to obtain the segmented lung area. Figure 7 shows the resulted images.

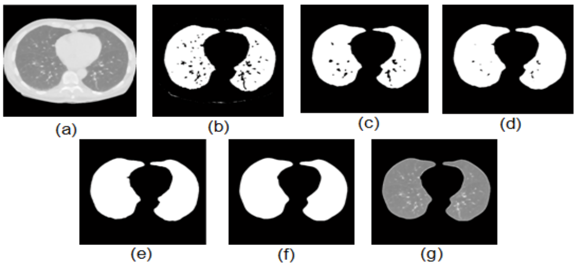


Fig. 7. The lungs segmentation technique (a) The segmented thorax image, (b) The binary image, (c) The filtered image, (d) The dilated image, (e) The filled image, (f) The closed image, (g) The segment lungs image.

3.3 Extraction of nodule candidates

The objective of this step is to extract the regions of interest (ROIs) that composed of the nodule candidates using the bi-level thresholding technique and a median filter of size 5*5. The resultant image was multiplied with the gray level lung image to obtain the ROIs in the CT chest images. The performance of the proposed framework was evaluated by comparing the resulted area with those obtained by three other different techniques. These techniques are Otsu thresholding, local entropy-based transition region extraction and thresholding, and the basic global thresholding [11]. The region non-uniformity criteria [11] was used to compare the performance of the four thresholding methods.

Region non-uniformity: Region non-uniformity is defined as:

$$NU = \frac{|F_T|}{|F_T + B_T|}\frac{\sigma_f^2}{\sigma^2} \tag{2}$$

where σ^2 is the whole image variance, and σ_f^2 represents the foreground variance, B_T and F_T are the background and foreground area pixels in the segmented image [11]. According to a non-uniformity (NU) measure the segmented image of smallest NU measure is the best histogram thresholding technique. The calculated NU measures of each segmented image for all applied histogram thresholding techniques are shown in Table 1.

Table 1. The NU measure calculated for the segmented images using the four histogram thresholding techniques

The Histogram Thresholding Technique	NU
Otsu thresholding technique	0.0225
Local entropy-based transition region extraction thresholding technique	0.1533
basic global thresholding technique	2.2029
proposed framework	0.0188

By visual comparison of the images in Figure 8 and comparing the results tabulated in Table 1 shows that the developed system gives the highest accuracy to detect the pulmonary nodule candidates.

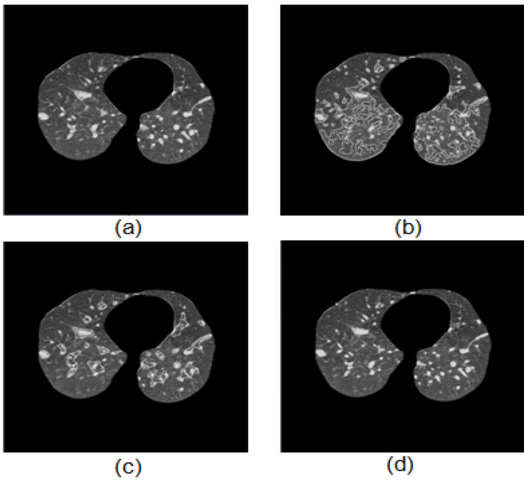


Fig. 8. (a) Otsu's method, (b) Basic Global Thresholding, (c) Local Entropy-Based Transition Region, and (d) the proposed framework.

3.4 Feature extraction

Feature extraction process aims to extracting a set of features which represent the information that is used in analysis and classification process. The goal of feature extraction is to achieve significant data reduction and to determine informative measures. In the present work, four different techniques of feature extraction were used; the first and second order of the statistical features [12], the Histogram of Oriented Gradients (HOG) features [13], the Value Histogram (VH) feature [12], and the texture features of Gray Level Co-Occurrence Matrix (GLCM) based on wavelet coefficients [13].

3.5 Feature fusion

In the process of features fusion, a new set of features was created from different sets of features obtained from different domains after removing the insignificant and redundant features. Therefore, the four different feature vectors were fused in a new hybrid feature vector using a simple concatenation procedure.

Having formed the hybrid feature vector, the next step is to remove any redundant and correlated information which is known as “feature selection”. In the present work, the GA algorithm [20] was applied to the hybrid feature vector as a feature selection technique. The performance of each feature vector and the new hybrid feature vector was then compared.

3.6 Nodules detection

The final stage aims to classify the extracted nodule candidates into nodules and non-nodules (blood vessels). Three classifiers were utilized and their performance was compared. These are: Support Vector Machine (SVM)[15], Multi-Layer Feed-Forward Neural Network (ANN) [15], and Radial Basis Function Neural Network (RBF-NN) [14]. The classifiers were trained and their performance was compared using the classification accuracy rate (CAR), sensitivity (S), and specificity (SP) measures [20]. For the training and testing steps of each classifiers, 25% of the available data set size was used for the training phase and they were tested using 75% of the available dataset size.

4 Experimental Results

The classification accuracy rate (CAR), sensitivity (S) and specificity (SP) were calculated for each classifier using the four types of features and hybrid feature vector. Tables 2-4 show the CAR, S, and SP obtained from the 3 different classifiers.

Comparing the results of classification shown in the three tables, it is clear that the use of the feature fusion technique led to the highest classification results for the three classifiers. The CAR reached 96.3%, 97% and 95%, for the three classifiers ANN, RBF-NN and SVM, respectively. The specificity (S) reached 99.1%, 100% and 100% for the three classifiers ANN, RBF-NN and SVM, respectively. SP reached to 94%, 91% and 95%, False Positive (FP) of values 0.06, 0.1, and 0.058 and the False Negative (FN) values of 0.008, 0.0 and 0.0 for the three classifiers Multi-Layer Feed-Forward Neural Network (ANN), Radial Basis Function Neural Network (RBF-NN) and Support Vector Machine (SVM), respectively. Table 5 depicts the number of features before and after selection using the GA algorithm and the CAR, S and SP corresponding to each classifier.

As clear from Table 5, the application of GA feature selection technique has increased the CAR, S and SP in addition to reducing the feature vector size. This has led also to a reduction in the computational time. The number of features resulted from using the RBF-NN classifier decreased significantly but the CAR and SP are relatively lower than those of the other two classifiers. While the results show that both the ANN and SVM have equal values of CAR but the number of features in the case of the SVM is less than that of ANN.

Table 2. The classification accuracy rate (CAR) for the three classifiers.

Classifiers Features	ANN	RBF-NN	SVM
Wavelet Features	85.8%	87.5%	94%
VH Features	89.6%	95.4%	87%
HOG Features	63%	65.8%	71.6%
Statistical Features	83.7%	78.3%	95.8%
Hybrid Features	96.3%	97%	95%

Table 3. The sensitivity (S) of three classifiers

Classifiers Features	ANN	RBF-NN	SVM
Wavelet Features	83.6%	95.9%	100%
VH Features	88.6%	96.6%	81.6%
HOG Features	62.8%	69%	67.8%
Statistical Features	98.8%	96%	99.1%
Hybrid Features	99.1%	100%	100%

Table 4. The sensitivity (SP) of three classifiers

Classifiers Features	ANN	RBF-NN	SVM
Wavelet Features	88.4%	81.7%	89.6%
VH Features	90.6%	94.3%	95%
HOG Features	63%	63.6%	77.7%
Statistical Features	75.8%	70.5%	93%
Hybrid Features	94%	91%	95%

Table 5. The classification of accuracy rate (CAR), the sensitivity (S) and the specifivity (SP) of the three classifiers and the number of hybrid features before and after using the generic algorithm (GA) technique.

Classifier	Hybrid feature size	Selected feature size	CAR	S	SP
ANN	182	169	99.6%	100%	99.2%
RBF-NN	182	115	99.2%	100%	98.4%
SVM	182	153	99.6%	100%	99.2%

5 Conclusion

In the present work, a Computer-Aided Detection system (CAD) for early detection of lung nodules in CT scans images has been developed. The proposed system consists of four main stages. These are; image preprocessing stage to enhance the image contrast of the CT images, an automatic segmentation stage to automatically extract the human's lung and the nodule candidates, a feature extraction and selection stage and a classification stage to classify the detected nodules.

Forty CT scans with 320 regions of interest (ROI) were made available from the early lung cancer action project (ELCAP) association to train and test the classifiers. The size of pulmonary nodules that were considered varies from 3 mm to 30 mm.

An Image Size Dependent Normalization Technique (ISDNT) was utilized for enhancing the CT image contrast and a Wiener filter was used to ameliorate the CT image quality in the image preprocessing stage.

For the automatic segmentation stage, the bi-level thresholding technique was applied to the preprocessed CT images and median filter and mathematical morphological operations were utilized to suppress any unwanted pixels.

In the third stage, four feature extraction techniques were utilized. These are: are the statistical features of first and second order, the Value Histogram (VH) feature, the Histogram of Oriented Gradients (HOG) features, and the texture features of Gray Level Co-Occurrence Matrix (GLCM) based on wavelet coefficients. A feature fusion step was employed on the four different sets of extracted features to produce the hybrid features vector. The five feature vectors were then used as the input to three types of classifiers and their performance was evaluated. The classifiers are: Artificial Neural Network (ANN), Radial Basis Function Neural Network (RBF-NN), and Support Vector Machine (SVM). Each classifier was trained using 25% of the dataset and tested using the remained 75% of available data input.

The Classification Accuracy Rate (CAR), the Sensitivity (S), and the Specificity (SP) were calculated for each classifier using each of the five feature vectors.

Comparing the CAR, S, and SP resulted from each classifier has showed that the hybrid features gave the highest CAR, S, and SP. This leads to conclude that the feature fusion technique increased the detection accuracy of pulmonary nodules and improves the system performance.

An attempt was made to increase the classification accuracy, enhance the system performance and to reduce the computational time using the Genetic Algorithm (GA) as a feature selection algorithm on the hybrid features vector. The CAR, S and SP results of the three learned classifiers; ANN, RBF-NN, and SVM showed an increase in the values of CAR, S and SP of the three classifiers. The CAR reached 99.6%, 99.2% and 99.6% for the three classifiers respectively. Based on these results, it can be concluded that applying the (GA) as a feature selection technique to the hybrid feature vector increases the classification performance of the system significantly.

In conclusion, the SVM classifier gives the highest CAR, S, and SP values of 99.6%, 100% and 99.2%, respectively. Table 6 shows a comparison of the performance of the suggested system and five systems reported in previously published researches. The comparison shows that the suggested system achieves the best classification rate and the lowest false positives**.**

Still much work is needed for discriminating benign and malignant tumors of the lung nodules. This is the aim of the next stage of the work.

Table 6. The accuracy and false positive of the proposed system and other work in the pulmonary nodules detection

Published Work	Accuracy	False Positives
Choi et al. , 2013 [15]	97.61%	2.27
Kuruvilla et al. , 2.014 [16]	93.30%	2.00
Demir et al. , 2015 [17]	90.12%	2.45
Manikandan et al. , 2016 [18]	94.00%	0.38
Sweetlin et al. , 2017 [19]	94.00%	
Proposed system	99.60%	0.008

6 References

[1] Donhauser J., "Early detection of lung cancer using low-dose CT screening", Siemens Healthcare GmbH, 2016.

[2] Narayanan B., Hardie R., Kebede T., "Analysis of Various Classification Techniques for Computer Aided Detection System of Pulmonary Nodules in CT", Aerospace and Electronics Conference (NAECON) and Ohio Innovation Summit (OIS), 2016 IEEE National, 2016. https://doi.org/10.1109/NAECON.2016.7856779

[3] Unnikrishnan S., Shamya C., Neenu P.A., "An Overview of CAD Systems for Lung Cancer Detection", International Journal of Engineering Research and General Science, vol. 4, no. 2, March-April, 2016.

[4] Firmino M., Angelo G., Morais H., Dantas M., Valentim R., "Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy", Bio Medical Engineering OnLine, 2016. https://doi.org/10.1186/s12938-015-0120-7

[5] Panpaliya N., Tadas N., Bobade S., Aglawe R., Gudadhe A., " A Survey on Early Detection and Prediction of Lung Cancer ", International Journal of Computer Science and Mobile Computing IJCSMC, Vol. 4, Issue. 1, January 2015.

[6] Faria F., Santos J., Rocha A., Torres R., “A framework for selection and fusion of pattern classifiers in multimedia recognition ", Pattern Recognition Letters, Volume 39, April, 2014.

[7] Early Lung Cancer Action Program (ELCAP), available from: http://www.via.cornell.edu/lungdb.html. [Last cited on 2011 Dec 05].

[8] Al-Ameen Z., Sulong G., Gapar M., Johar M., "Enhancing the Contrast of CT Medical Images by Employing a Novel Image Size Dependent Normalization Technique", Int. Journal of Bio-Science and Bio-Technology, vol. 4, no. 3, September, 2012.

[9] Abou-Chadi F.E.Z., Obayya, M.I., Amer H.M., “A Computer-Aided System for Classifying Computed Tomographic (CT) Lung Images Using Artificial Neural Network and Data Fusion", Int. Journal of Computer Science and Network Security, vol.11 no.10, Oct. 2011.

[10] Kumari S., “Threshold Based Enhanced Segmentation Technique for Early Detection and Prediction of Lung Cancer", International Journal of Computer Science and Information Technologies, vol. 8, no. 3, pp. 335-337, 2017.

[11] Zuoyong L., Zhang D., Xuc Y., Liu C., "Modified local entropy-based transition region extraction and thresholding", Applied Soft Computing, vol. 11, pp. 5630–5638, 2011. https://doi.org/10.1016/j.asoc.2011.04.001

[12] Lingayat N.S., Tarambale M.R., "A Computer Based Feature Extraction of Lung Nodule in Chest X-Ray Image", International Journal of Bioscience, Biochemistry and Bioinformatics, Vol. 3, No. 6, November 2013.

[13] Orozco, H.M, Villegas, O.V, Sánchez, V.G., Domínguez, H.O., " Automated system for lung nodules classification based on wavelet feature descriptor and support vector machine", Madero Orozco et.al. Bio-Medical Engineering on Line, vol. 14, no. 9, 2015.

[14] Babu M.S., Murty, N.V., "A Novel Hybrid Application of Radial Basis Function Neural Network and Fuzzy Logic for Detection and Diagnosis of Lung Cancer ", International Journal of Engineering Science and Computing, vol. 6, no. 9, 2016.

[15] Choi, W.J., Choi, T.S., " Cancer Detection Using Artificial Neural Network and Support Vector Machine: A Comparative Study ", Entropy 2013, vol. 15, pp. 507-523, 2013.

[16] Kuruvilla, J., Gunavathi, K., "Lung cancer classification using neural networks for CT images", computer methods and programs in biomedicine, vol. 113, pp. 202–209, 2014.

[17] Demira, O., Çamurcub, A.Y., "Computer-aided detection of lung nodules using outer surface features”, Bio-Medical Materials and Engineering, vol. 26, pp. S1213–S1222, 2015. https://doi.org/10.3233/BME-151418

[18] Manikandan, T., Bharathi, N.," Lung Cancer Detection Using Fuzzy Auto-Seed Cluster Means Morphological Segmentation and SVM Classifier", J Med Syst., vol. 40, no. 181, 2016. https://doi.org/10.1007/s10916-016-0539-9

[19] Sweetlin, J.D., Nehemiah, H. Kh., Kannan, A., "Computer aided diagnosis of pulmonary hamartoma from CT scan images using ant colony optimization based feature selection", Alexandria University, Alexandria Engineering Journal, ISSN:111.-0168, 2017.

[20] Sokolova M., Lapalme G., " A systematic analysis of performance measures for classification tasks", Information Processing & Management, Volume 45, Issue 4, July 2009.

7 Authors

Hanan M. Amer currently works in the Department of Electronics and Communications Engineering, Faculty of Engineering at Mansoura University, Mansoura, Egypt.

Sherif S. Kishk works in the Department of Electronics and Communications Engineering, Faculty of Engineering at Mansoura University, Mansoura, Egypt.

Marwa I. Obayya works in the Department of Electronics and Communications Engineering, Faculty of Engineering at Mansoura University, Mansoura, Egypt.

Fatma E. Z. Abou-Chadi works in the Department of Electrical Engineering, Faculty of Engineering, The British University in Egypt, Cairo, Egypt.

Article submitted 14 October 2018. Resubmitted 29 November 2018. Final acceptance 06 December 2018. Final version published as submitted by the authors.