Sensors and Materials, Vol. 30, No. 8 (2018) 1859–1868
MYU Tokyo

S & M 1643

Lung Nodule Detection Using Ensemble Classifier in Computed Tomography Images

Chien-Cheng Lee,[1*] Shuo-Ting Tsai,[1] and Chin-Hua Yang[2]

[1]Department of Electrical Engineering, Yuan Ze University,
135 Yuan-Tung Road, Chung-Li Dist., Taoyuan City, 32003, Taiwan
[2]Department of Medical Imaging, National Taiwan University Hospital,
25, Lane 442, Sec.1, Jingguo Rd., Hsinchu City, 300, Taiwan

(Received April 5, 2017; accepted February 27, 2018)

Keywords: lung nodule, ensemble classifier, CT image

Lung cancer is the leading cause of cancer deaths. The main reason is that patients are mostly diagnosed with lung cancer in its third or final stage. Lung nodules are small growing tissues, which may become malignant tumors that cause early lung cancer lesions. Therefore, a computer-aided system of lung nodule detection would achieve early detection and facilitate early treatment. In this paper, we present a method of lung nodule detection in computed tomography (CT) images based on an ensemble classifier. The proposed nodule detection method includes lung parenchyma segmentation, nodule candidate detection, and nodule candidate classification. First, an adaptive thresholding algorithm is applied to segment the lung parenchyma. The lung region boundaries are also corrected by using a contour analysis algorithm. Second, the adaptive thresholding algorithm is employed to find the regions of interest. Meanwhile, lung nodule candidates are roughly detected by connected component analysis. To obtain a complete 3D structure, the method merges the rough detection results if they conform to the predefined merging conditions. Finally, a self-organizing map (SOM) algorithm is used to select the negative samples for the training data, and an ensemble classifier is applied to recognize the nodule regions. The experimental results show that the proposed method outperforms the previous methods.

1. Introduction

According to the World Health Organization, cancer is the second leading cause of death globally, and was responsible for 8.8 million deaths in 2015.(1) The most common cause of cancer death is cancer of the lung (1.69 million deaths). Cancer mortality can be reduced if cases are detected and treated early. When diagnosed at an early stage, lung cancer patients are more likely to respond to effective treatment, which can result in a greater probability of surviving and less expensive treatment. Screening is a popular early detection method for lung cancer. It identifies individuals with abnormalities suggestive of a specific cancer or pre-cancer who are asymptomatic and refer them promptly for diagnosis and treatment. Nowadays,

*Corresponding author: e-mail: cclee@saturn.yzu.edu.tw
https://dx.doi.org/10.18494/SAM.2018.1899

ISSN 0914-4935

discovering lung cancer in its initial stage is by finding a solitary lung nodule revealed by chest X-ray or computed tomography (CT). Thus, the development of a reliable computer aided diagnosis (CAD) system for lung cancer is one of the most vital research topics in medical image processing.

Many lung-nodule detection techniques have been proposed. Usually, these proposed methods consist of three stages. The first stage is to segment the lung parenchyma. The next stage is to detect the nodule candidates. Finally, a classification algorithm is used to classify the correct nodules. In the lung parenchyma segmentation, Retico *et al.* proposed a binary thresholding method to segment the lung parenchyma based on Hounsfild unit (HU) values.[2] Then, they used image morphological operations to correct regions of vessels and airway walls. Later, they also proposed a directional-gradient concentration (DGC) method and applied it to the pleura surface.[3] The DGC was combined with a morphological opening-based procedure to generate a list of nodule candidates. Ye *et al.* also utilized binary thresholding to segment the lung parenchyma. However, they used a chain code that defined eight directions to analyze and correct the region boundaries of lung parenchyma.[4] Gurcan *et al.* proposed a method of identifying lung regions by a k-means clustering technique.[5] Each lung slice is classified as belonging to the upper, middle, or lower part of the lung. Within each lung region, structures are segmented again using weighted k-means clustering.[5]

After the lung parenchyma segmentation, a nodule candidate detection algorithm is followed. These algorithms can be divided into two types. One is 2D-based, and the other is 3D-based. The 2D-based algorithms detect nodule candidates only from a single slice image. In contrast, the 3D-based algorithms consider several consecutive slices to find the nodule candidates. Sivakumar and Chandrasekar used a weighted fuzzy clustering to segment nodules for lung cancer images. Then, the lung nodules were classified as normal or abnormal by using the support vector machine (SVM).[6] Osman *et al.* combined the 3D CT regions of interest (ROIs) slices to form a 3D ROI image. Next, a 3D template was determined to find structures with properties similar to those of nodules.[7] Li *et al.* made use of the eigenvalues and eigenvectors derived from the Hessian matrix to calculate geometric features from 3D CT scans. The geometric features provide information of stick, plate, and ball-like objects. Three selective enhancement filters were developed for dot, line, and plane, which can simultaneously enhance objects of a specific shape (for example, dot-like nodules) and suppress objects of other shapes (for example, line-like vessels).[8] Teramoto and Fujita proposed a fast lung-nodule detection scheme in chest CT images using a cylindrical nodule-enhancement filter with the aim of improving the workflow for diagnosis in CT examinations.[9] Elizabeth *et al.* used a snake algorithm to segment the lung parenchyma from each slice. Then, ROIs were later extracted from the lung parenchyma using a region growing algorithm; the shape and texture features were extracted. Finally, a radial basis function neural network (RBFNN) was used for classification.[10]

To distinguish nodules from candidates, machine learning algorithms are widely used to classify the candidates as normal/abnormal. Sivakumar and Chandrasekar[11] and da Silva Sousa *et al.*[6] used SVM classifiers to classify the nodule regions. Böröczky *et al.* used a genetic algorithm as the classifier.[12] Antonelli *et al.* utilized five classifiers and also used different combinations of these classifiers to test the classification performance.[13] Lee *et al.* proposed a two-step classification architecture to distinguish nodule candidates that combined

genetic algorithm-linear discriminant analysis (GA-LDA) and random subspace method (RSM).[14]

In this paper, we propose a novel method for lung-nodule detection in CT images based on an ensemble classifier. The proposed nodule detection method includes lung parenchyma segmentation, nodule candidate detection, and nodule candidate classification. First, an adaptive thresholding algorithm is applied in the system to segment the lung parenchyma. Second, the adaptive thresholding algorithm is employed again to find the ROIs. Meanwhile, lung nodule candidates are roughly detected by the connected component analysis. Finally, a self-organizing map (SOM) algorithm is used to select the negative samples for the training data, and an ensemble classifier is applied to recognize the nodule regions.

The rest of this paper is structured as follows. In Sect. 2, we describe our proposed method. The experimental results are presented in Sect. 3. Finally, concluding remarks are made in Sect. 4.

2. Materials and Methods

The proposed method is summarized by the flowchart shown in Fig. 1. First, a series of CT slices is inputted. Next, a segmentation algorithm is applied to these slices to segment the lung regions. Third, a nodule candidate detection module is used to detect the nodule candidates. Finally, an ensemble classifier classifies the nodule candidates as nodules or non-nodules.

2.1 Lung region segmentation

To segment the lung regions, we first employed an adaptive binary thresholding algorithm.[15–17] The adaptive binary thresholding is based on the threshold value updating iteratively, as shown in the following steps.

Step 1. Set −500 HU as the initial threshold at $T^{(0)}$.

Step 2. Compute the average upper volume u_h.

$$u_h = Avg(I(x, y, z) \geq T^{(i)}), \tag{1}$$

where the 3D volume of a CT scan is denoted as $I(x, y, z)$, where the x and y indices represent the slice coordinates, and z indicates the slice number.

Step 3. Compute the average lower volume u_l.

$$u_l = Avg(I(x, y, z) < T^{(i)}), \tag{2}$$

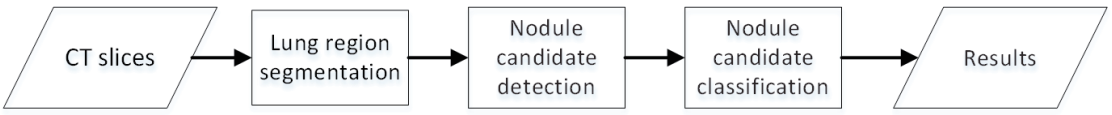


Fig. 1. Flowchart of the proposed method.

Step 4. Compute the new threshold value.

$$T^{(i+1)} = \frac{u_h + u_l}{2} \tag{3}$$

Step 5. Continue steps 2 through 4 until $T^{(i)} = T^{(i+1)}$.
Step 6. Set the optimal threshold $T_{opt} = T^{(i+1)}$.

After the optimal threshold selection, thresholding operation is performed to roughly segment the bone and muscle from the image as defined below.

$$I_{Bin}(x, y, z) = \begin{cases} 1 \text{ if } I(x, y, z) \geq T_{opt} \\ 0 \text{ if } I(x, y, z) < T_{opt} \end{cases} \tag{4}$$

Figure 2(a) shows an example of the CT slice, and Fig. 2(b) is the result of the thresholding. Next, the region with the extreme outer contour is regarded as a body mask I_{BM}, as shown in Fig. 2(c). The body mask is utilized to segment the body region I_{Body} using intersection operation, as shown in Fig. 2(d). Then, the thresholding operation defined in Eq. (5) is performed to segment the lung parenchyma I_{Lung}, as shown in Fig. 2(e). Finally, a morphological closing operation is used to fill holes in the lung parenchyma to obtain a lung mask I_{LM}, as shown in Fig. 2(f).

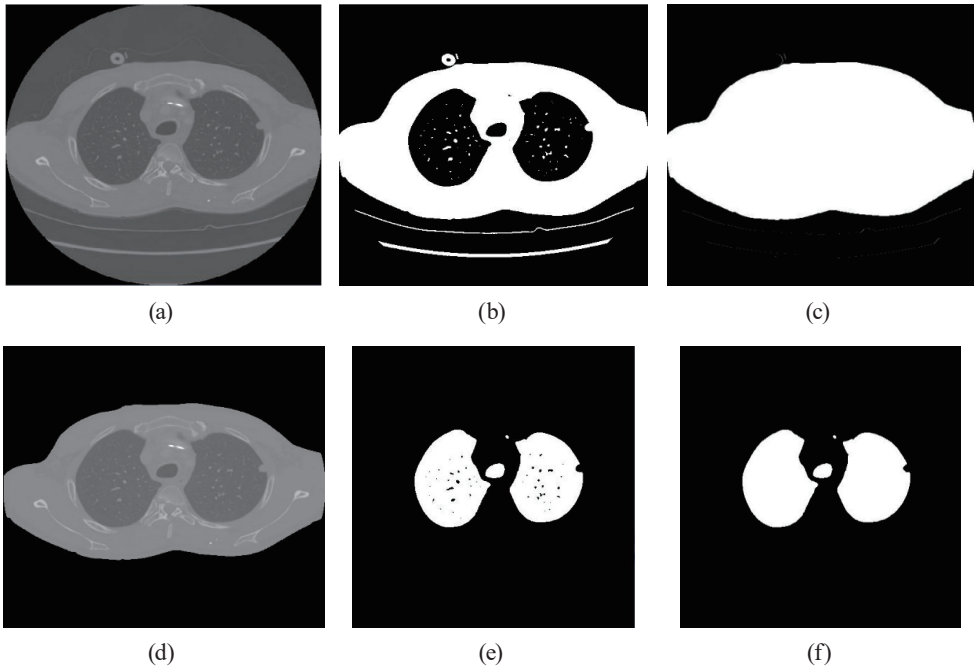


Fig. 2. Process of lung parenchyma. (a) Original CT image, (b) thresholding using T_{opt}, (c) body mask, (d) body region, (e) lung parenchyma, and (f) lung mask.

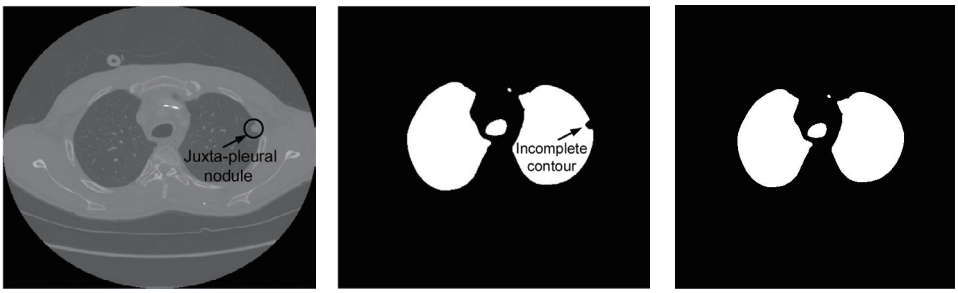
$$I_{Lung}(x,y,z) = \begin{cases} 1 & \text{if } I_{Body}(x,y,z) < T_{opt} \text{ and } I_{BM}(x,y,z) = 1 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

One kind of lung nodule, the juxta-pleural nodule, appearing on the lung boundary will cause an incomplete contour of the lung mask, as shown in Fig. 3. To overcome this problem, we trace the lung contour to find an arch-like curve. If an arch-like curve is found, the endpoints of the curve are connected to correct the lung contour. The result is shown in Fig. 3(c).

2.2 Nodule candidate detection

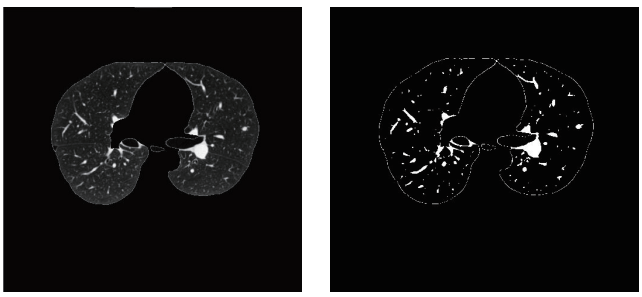
The adaptive binary thresholding described in Sect. 2.1 is applied to binarize the lung regions for nodule candidate detection. The intersection of the original CT slice and the lung mask is calculated first, then inputted to the adaptive thresholding algorithm as shown in Fig. 4(a). As HU values of nodules are usually higher than those of the other tissues in the lung, the initial threshold $T^{(0)}$ of the adaptive thresholding is set to −100. The result is shown in Fig. 4(b).

After the binarization of lung regions, connected components in the lung region are acquired. Meanwhile, the centroids of all connected components are also calculated. Nodule candidates are labeled as follows.



(a) (b) (c)

Fig. 3. Correction of lung mask contour. (a) Juxta-pleural nodule appearing on the lung boundary, (b) incomplete contour of the lung mask, and (c) contour of the lung mask after correction.



(a) (b)

Fig. 4. Binarization of the lung region. (a) Intersection of the CT slice and the lung mask and (b) binarization result.

Step 1. Start from slice $z = 2$.
Step 2. Calculate centroids of all connected components on slices $z - 1$, z, and $z + 1$.
Step 3. Calculate distance D_p between all connected components on slices $z - 1$ and z.
Step 4. Calculate distance D_n between all connected components on slices $z + 1$ and z.
Step 5. If $D_p < 6$ mm and $D_n < 6$ mm, these three connected components on slices $z - 1$, z, and $z + 1$ are labeled as nodule candidates, as shown in Fig. 5(a).
Step 6. Continue steps 2 through 5 until termination.

Figure 5(b) shows the labeled nodule candidates. However, some V-shaped or Λ-shaped structures may be regarded as different nodule candidates. Thus, if the distance between any two labeled candidates is less than 15 mm and are connected in 3D space, these two candidates are merged together.

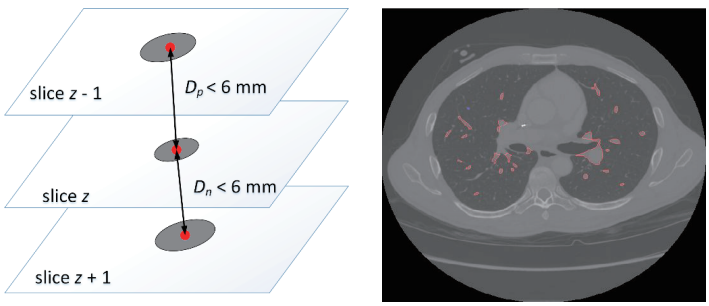
2.3 Nodule candidate classification

In the nodule candidate detection, most of the cylinder-like or ellipsoid-like structures could be detected. In other words, these candidates not only included the nodules but also vessels. Thus, in this study, we developed an ensemble classification algorithm to classify the candidates as nodules or non-nodules. To distinguish between them, six features were extracted including three geometrical features and three HU features.[12] These features are described as follows:

A. Volume: volume of a candidate structure.
B. Compactness mean: mean value of the compactness measure of each slice of a candidate structure.
C. Sphere density: ratio of volume of a candidate structure and the volume of the minimal bounding sphere, defined as

$$sd = \frac{volume}{\frac{4}{3}\pi r_{min}^3}, \tag{6}$$

where r_{min} is the radius of the minimal bounding sphere.



(a) (b)

Fig. 5. (Color online) Nodule candidate labeling. (a) Labeling process and (b) labeled nodule candidates.

D. HU mode: the mode of all voxel HU values in a candidate structure.
E. Mean of HU min: mean value of the minimum voxel HU among the corresponding location on the previous and next slices, defined as

$$HU_{min} = Avg(I_{min}(x, y, z) \in Candidate), \tag{7}$$

where $I_{min}(x, y, z) = min\{I(x, y, z-1), I(x, y, z), I(x, y, z+1)\}$, and the x, y, and z indices represent the voxel coordinate in a candidate structure.

F. HU skewness: the skewness of the voxels of a candidate structure, defined as

$$sk = \frac{E[(I(x, y, z) - u)^3]}{E[(I(x, y, z) - u)^2]^{3/2}}, \tag{8}$$

where the x, y, and z indices represent the voxel coordinate in the candidate structure, and u is the mean value of the structure voxels.

Training samples should be provided before the classification of nodules. However, the data usually exhibit a large imbalance in the distribution of the target classes. Specifically, there were more negative samples than positive samples. It was reasonable because the number of nodules was far less than that of the detected candidates. In such cases, maintaining the same percentage for each target class was important in the training process of a classifier. Otherwise, the classification model would trend to the negative target class.

To prevent this, we used a SOM network to select the negative samples.[18] The goal of the SOM network is to make different parts of the network respond similarly to certain input patterns. Thus, the SOM network can form an abstract representation from the input space. In order words, the SOM neurons corresponding to the smaller selected samples can represent a mass of input samples. In this study, the input space is constructed from the negative samples. The number of neurons in the SOM network indicated the number of selected negative samples. After the training of the SOM network, these neurons could be used to represent the negative samples.

An ensemble classifier is used in this study to classify the candidate nodules as nodules or non-nodules.[19] The proposed ensemble classifier consists of multilayer perceptron (MLP), SVM, and AdaBoost, as shown in Fig. 6.[20,21] The features extracted from the candidate nodules are presented to the ensemble classifier. Three SOM-based selectors with different initial weights selected the proper number of negative samples. These negative samples are combined with the positive samples to form the training samples for the classifiers. Finally, a weighted voting method is utilized to determine whether the candidate was a nodule or not.

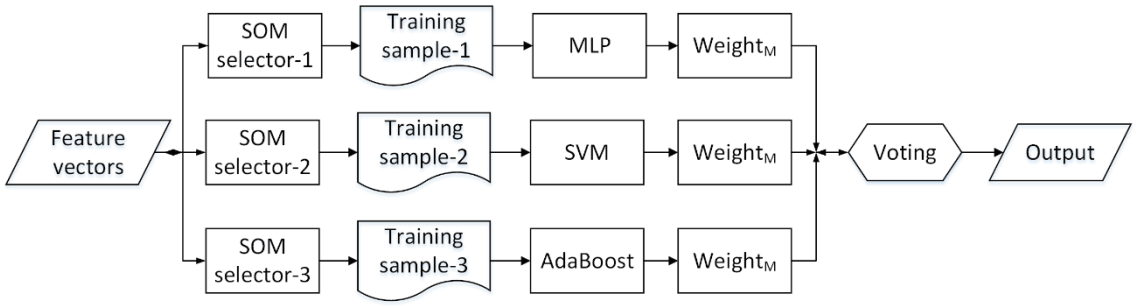


Fig. 6. Ensemble classifier for candidate nodule classification.

3. Experimental Results

To evaluate the performance of our proposed method, we used datasets from the National Institute of Health's Lung Imaging Database Consortium (LIDC).[22,23] The main benefit of this database is that the ground truth is provided by medical specialists; thus, we could verify the detection results efficiently. The LIDC datasets are stored in Digital Imaging and Communications in Medicine (DICOM) format. The size of the images was 512 × 512, with a slice gap of 1 mm. In this study, we used 31 CT scans including 7699 slices with 66 nodules. Nodules were found in each series.

In the effort to find the best parameters, the distribution of radii in all the nodules was investigated as shown in Fig. 7. From the figure, the average radius is about 4.6 mm, and most of the radii are less than 15 mm. Therefore, the maximum radius of a nodule is set to 15 mm in this study. These parameters were used in the lung contour correction and nodule candidate labeling. To verify the performance of the nodule candidate detection method, all 31 CT scans were presented to the algorithm. The results show that 59 nodule candidates were detected among 66 nodules. The seven failures were caused by the very low HU values of the nodules. The detection rate of the nodule candidates was about 89.39%.

To evaluate the performance of the proposed ensemble classifier, two hidden layers of 15 and 10 neurons were adopted in MLP neural network. The radial basis function kernel was used in the SVM classifier and the AdaBoost was composed of 100 weak classifiers. Sixteen CT scans including 40 nodules were used in training and 15 CT scans including 26 nodules were used in the test. In this study, we selected five sets of training data. The five negative sample sets were selected by the SOM network selectors with the number of neurons equal to 8 × 8, 9 × 9, 10 × 10, 10 × 10, and 10 × 10. The experimental results demonstrated a sensitivity rate of 100% and a specificity rate of 86.07%. The comparison results of the proposed method with other methods are listed in Table 1. From the table, it is clearly seen that our proposed method outperformed the other methods overall. The sensitivity describes the fraction of diseased patients who were correctly classified by radiologists, while the specificity describes the fraction of nondiseased patients who were correctly classified, defined as

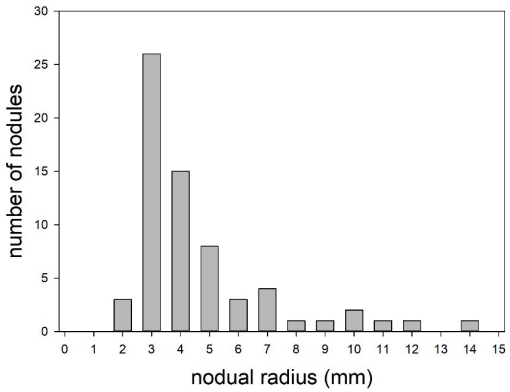


Fig. 7. Distribution of the radii of nodules.

Table 1
Comparison results of our method with other methods.

Method	Sensitivity (%)	Specificity (%)
Our method	100	86.07
Böröczky *et al.*[12]	100	56.4
Yeh *et al.*[24]	94.4	74.4
Antonelli *et al.*[13]	92.5	83.5
Lee *et al.*[14]	87	81
da Silva Sousaa *et al.*[11]	84.84	96.15
da Silva *et al.*[25]	70	100

$$sensitivity = (true\ positive) / (true\ positive + false\ negative), \tag{9}$$

$$specificity = (true\ negative) / (true\ negative + false\ positive). \tag{10}$$

4. Conclusions

In this paper, we presented a method for lung-nodule detection in CT images based on an ensemble classifier. The proposed method includes lung parenchyma segmentation, nodule candidate detection, and nodule candidate classification. In this study, we used an adaptive thresholding methodology to segment the lung region and an ensemble classifier combining MLP, SVM, and AdaBoost to classify candidate nodules as nodules or non-nodules. The method was applied on datasets from the LIDC database to evaluate the performance. It was demonstrated that the overall performance of the proposed method is better than those of the other methods.

Acknowledgments

This work was supported by the Ministry of Science and Technology (Grant number: MOST 105-2622-E-155-015-CC3).

References

1 World Health Organization: Cancer Key Facts, http://www.who.int/mediacentre/factsheets/fs297/en/ (accessed February 2018).
2 A. Retico, P. Delogu, M. E. Fantacci, I. Gori, and A. Preite Martinez: Comput. Biol. Med. **38** (2008) 525.
3 A. Retico, M. E. Fantacci, I. Gori, P. Kasae, B. Golosio, A. Piccioli, P. Cerello, G. De Nunzio, and S. Tangaro: Comput. Biol. Med. **39** (2009) 1137.
4 X. Ye, X. Lin, J. Dehmeshki, G. Slabaugh, and G. Beddoe: IEEE Trans. Biomed. Eng. **56** (2009) 1810.
5 M. N. Gurcan, B. Sahiner, N. Petrick, H. P. Chan, E. A. Kazerooni, P. N. Cascade, and L. Hadjiiski: Med. Phys. **29** (2002) 2552.
6 S. Sivakumar and C. Chandrasekar: Int. J. Eng. Technol. **5** (2013) 179.
7 O. Osman, S. Ozekes, and O. N. Ucan: Comput. Biol. Med. **37** (2007) 1167.
8 Q. Li, S. Sone, and K. Doi: Med. Phys. **30** (2003) 2040.
9 A. Teramoto and H. Fujita: Int. J. Comput. Assisted Radiol. Surg. **8** (2013) 193.
10 D. S. Elizabeth, H. K. Nehemiah, C. S. Retmin Raj, and A. Kannan: IET Image Process. **6** (2012) 697.
11 J. R. F. da Silva Sousa, A. C. Silva, A. C. de Paiva, and R. A. Nunes: Comput. Methods Programs Biomed. **98** (2010) 1.
12 L. Böröczky, L. Zhao, and K. P. Lee: IEEE Trans. Inf. Technol. Biomed. **10** (2006) 504.
13 M. Antonelli, M. Cococcioni, B. Lazzerini, and F. Marcelloni: Pattern Anal. Appl. **14** (2011) 295.
14 M. C. Lee, L. Boroczky, K. Sungur-Stasik, A. D. Cann, A. C. Borczuk, S. M. Kawut, and C. A. Powell: Artif. Intell. Med. **50** (2010) 43.
15 W. J. Choi and T. S. Choi: Comput. Methods Programs Biomed. **113** (2014) 37.
16 J. Dehmeshki, X. Ye, X. Lin, M. Valdivieso, and H. Amin: Comput. Med. Imaging Graphics **31** (2007) 408.
17 S. Hu, E. A. Hoffman, and J. M. Reinhardt: IEEE Trans. Med. Imaging **20** (2001) 490.
18 T. Kohonen: Biol. Cybern. **43** (1982) 59.
19 L. Rokach: Artif. Intell. Rev. **33** (2010) 1.
20 Y. Freund and R. E. Schapire: J. Comput. Syst. Sci. **55** (1997) 119.
21 C. Cortes and V. Vapnik: Mach. Learn. **20** (1995) 273.
22 M. F. McNitt-Gray, S. G. Armato Iii, C. R. Meyer, A. P. Reeves, G. McLennan, R. C. Pais, J. Freymann, M. S. Brown, R. M. Engelmann, P. H. Bland, G. E. Laderach, C. Piker, J. Guo, Z. Towfic, D. P. Y. Qing, D. F. Yankelevitz, D. R. Aberle, E. J. R. van Beek, H. MacMahon, E. A. Kazerooni, B. Y. Croft, and L. P. Clarke: Acad. Radiol. **14** (2007) 1464.
23 S. G. Armato Iii, G. McLennan, M. F. McNitt-Gray, C. R. Meyer, D. Yankelevitz, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, A. P. Reeves, B. Y. Croft, and L. P. Clarke: Radiology **232** (2004) 739.
24 C. Yeh, J. F. Wang, M. T. Wu, C. W. Yen, M. L. Nagurka, and C. L. Lin: Comput. Med. Imaging Graphics **32** (2008) 270.
25 E. C. da Silva, A. C. Silva, A. C. de Paiva, R. A. Nunes, and M. Gattass: Comput. Methods Programs Biomed. **90** (2008) 230.