35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

A New Method for Pulmonary Nodule Detection using Decision Trees*

A. Tartar, N. Kılıç, A. Akan, *Senior Member, IEEE*

Abstract— **A computer-aided detection (CAD) can help radiologists in diagnosing of lung diseases at an early level. In this study, a new CAD system for pulmonary nodule detection from CT imagery is presented by using morphological features and patient information properties. Decision trees are utilized for classification and overall detection performance is evaluated. Results are compared to similar techniques in the literature by using standard measures. Proposed CAD system with random forest classifier result in 90.5 % sensitivity and 87.6 % specificity of detection performance.**

I. INTRODUCTION

Today, lung cancer is still a primary cause of cancer death worldwide. Especially, lung cancer is one of the main public health issues in the developed industrial countries [1]. This makes the treatment of lung cancer a very important task. Early level detection of pulmonary nodules is of high importance for improving the patient's chance of survival. However, the detection of cancer takes quite a long time and the evaluation of each scan containing numerous sectional images is a tiresome process. Consequently the performance of radiologist's interpreter can be negatively affected.

It is extremely important task to develop Computer Aided Diagnosis (CAD) algorithms that potentially find true positive findings and can help radiologists in the detection of early level pulmonary nodules. Thus, CAD system is an compact tool providing radiologists with a second opinion to improve the sensitivity of their diagnosis decision-making process [2]. Today, CAD systems are frequently utilized to detect plenty abnormalities in routine clinical work.

In other studies reported in the literature, CAD systems were developed by using the features of nodule candidate patterns with image-processing techniques by classifying the shape of pulmonary nodule patterns [3] and by using morphological features [4]. To classify lung nodules, neural network approaches [5] and Fisher linear discriminant classifier [6] were proposed. Similarly, CAD systems were presented by using a genetic algorithm with the random subspace method [7], a support vector machine [8] and a random forest classifier [9].

In this study, geometric features based on the basic morphological shape information of 2D pulmonary nodule patterns and patient information properties were utilized for feature extraction. To evaluate the performance analysis of the proposed CAD system with decision trees, different feature sets are created for selecting the features that will be used in classification algorithms.

To perform a careful validation of the proposed system, entirely independent training and testing datasets are used. All nodules in the proposed system are first trained using a dataset provided by Istanbul University, Cerrahpasa School of Medicine.

The remainder of this study is organized as follows. The proposed CAD system and its algorithm are described in Section 2. This section includes the database information, morphological image processing for feature extraction and classifier algorithms of decision tree as well. Overall performance results are presented in Section 3. The results of the proposed CAD system contain performance comparisons with five other previously reported CAD systems. Conclusions are drawn in Section 4.

II. MATERIALS AND METHODS

II.1 Database and Imaging Protocol

In this section, a brief description of the dataset used is provided. Dataset containing 95 pulmonary nodule and 75 non-nodule patterns obtained from CT images of 63 patients was utilized. Images are collected from 39 male and 24 female patients whose ages range from 25 to 78 years. The mean age is 55.4±12.3 years. The number of pulmonary nodules detected in the right and left lung parenchyma is 67 (20 in the upper part, 20 on the bottom part, 27 in pleural case) and 28 (12 in the upper part, 8 on the bottom part, 8 in pleural case), respectively. The average nodule diameter is 6.42±3.00 mm. The diameter distribution of the nodules utilized in the database is shown in Fig.1. Nodule and non-nodule pattern samples used in dataset are also given in Fig.2. The 2D pulmonary nodule patterns are manually marked on CT image by radiologists.

The dataset was obtained from chest CT images of patients scanned by using "Sensation 16" CT scanner (Siemens Medical Systems) between 2010 and 2012 at Radiology Department, Cerrahpasa School of Medicine,

*Research supported by Scientific Research Projects Coordination Unit of Istanbul University.

A. Tartar is with the Department of Engineering Sciences, University of Istanbul, Avcilar Campus, 34320 Istanbul, Turkey (corresponding author to provide phone: +90(212)4737070; fax: +90(212)4737180; e-mail: atartar@istanbul.edu.tr).

N. Kılıç is with the Department of Electrical and Electronics Engineering, University of Istanbul, Avcilar Campus, 34320 Istanbul, Turkey (e-mail: niyazik@istanbul.edu.tr).

A. Akan is with the Department of Electrical and Electronics Engineering, University of Istanbul, Avcilar Campus, 34320 Istanbul, Turkey (e-mail: akan@istanbul.edu.tr).

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

Istanbul University. CT scans were acquired at a tube potential voltage of 120 kVp. All CT images are in size of 512x512 pixels and stored as DICOM (Digital Imaging and Communications in Medicine) format files, directly from the CT modality.

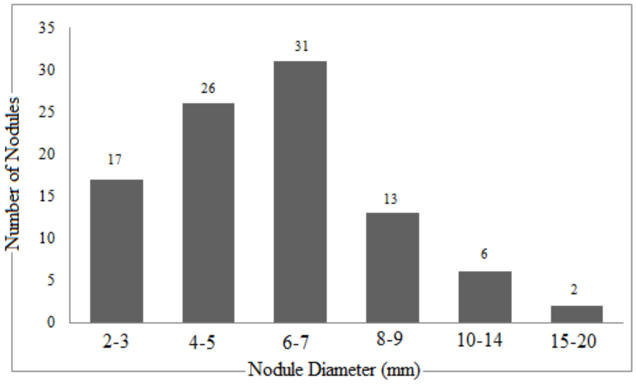

Fig. 1: Histogram representing the distribution of the diameter for the 67 nodules of the database.

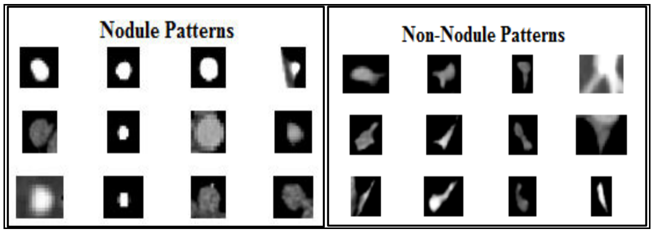

Fig. 2: Nodule and Non-Nodule pattern samples used in dataset.

II.2 Morphological Image Processing

Morphology is the main tool of the mathematical sets underlying the development of techniques that extract the conceptual features from an image [10]. In our approach, geometric features of pulmonary nodules were obtained by using the regional descriptors of the 2D patterns based on the basic morphological shape information. In this work, the geometric features consist of the area, perimeter, diameter, solidity, eccentricity, aspect ratio, compactness, roundness, circularity, ellipticity of the patterns. Besides, patient information properties, gender and age, are performed as features of the risk factors in detection of pulmonary nodules [21].

The complete listing of the set of features in the dataset is given by its definitions in Table 1. A total of 12 features are evaluated for extracting features of the patterns. In Table 1, *Solidity* denotes the proportion of the pixels in the convex hull that are also in the region. *Eccentricity* is the eccentricity of the ellipse that has the same second moments as the region. Also it is the ratio of the distance between the foci of the ellipse and its major axis length. The value of eccentricity is between 0 and 1. Measurements of compactness, roundness, circularity and ellipticity are computed by the definitions given in Table 1 [11].

Table 1
Geometric features and patient information properties used for pulmonary nodule detection

Measure	Definition
Area	A
Perimeter	P
Diameter	D
Solidity	S
Eccentricity	E
Aspect Ratio	$\frac{Min.Diameter\ (M)}{Max.Diameter\ (L)}$
Compactness	$\frac{P^2}{4\pi A}$
Roundness	$\frac{4A}{\pi L^2}$
Circularity	$\frac{4\pi A}{P^2}$
Ellipticity	$\frac{\pi L^2}{2A}$
Patient Information Properties	
Gender	-
Age	-

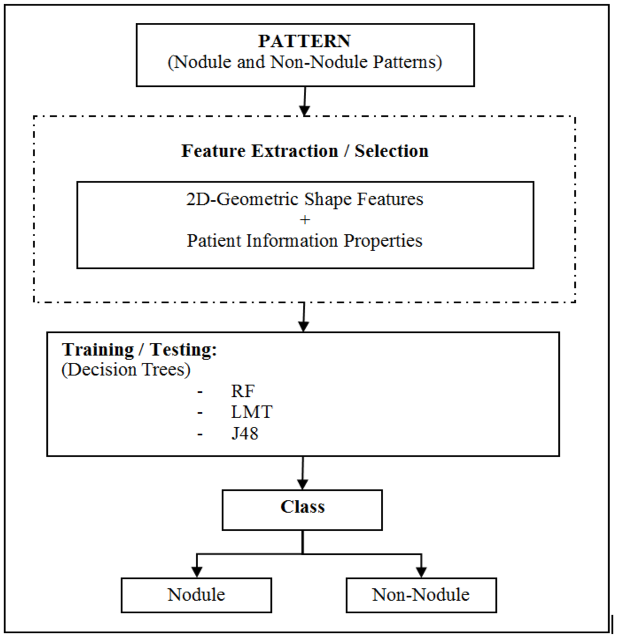
PATTERN
(Nodule and Non-Nodule Patterns)
Feature Extraction / Selection
2D-Geometric Shape Features
+
Patient Information Properties
Training / Testing:
(Decision Trees)
- RF
- LMT
- J48
Class
Nodule
Non-Nodule

Fig. 3: The computer-aided detection algorithm scheme for pulmonary nodule detection.

The computer-aided diagnosis algorithm scheme designed in this work is shown in Fig.3.

In this study, different feature sets are introduced for selecting the best features for classification algorithms. These feature sets are tabulated in Table 2. Feature selection is one of the most critical tasks in building of detection models. Feature selection can reduce both data size and computational complexity. Trial and error methods were used to create the feature sets in Table 2.

Table 2
Feature sets used in classification algorithms.

The number of features	Features' name
12	All features.
10	Gender, area, ellipticity, age, solidity, eccentricity, perimeter, compactness, aspect ratio, diameter.
8	Gender, area, ellipticity, age, solidity, eccentricity, perimeter, compactness.
6	Gender, area, ellipticity, age, solidity, eccentricity.
4	Gender, area, ellipticity, age.
2	Gender, area.

II.3 Nodule Classification

In order to classify the pulmonary nodules, Random Forest (RF), Logistic Model Tree (LMT) and J48 decision tree classifiers are utilized.

Random Forest was proposed by Leo Breimans in 1999 [12]. It is a new development in tree based classifiers and fast proven to be one of the most important algorithms in the machine learning systems. It is defined as a combination of tree predictors depend on the values of a random vector sampled independently and with the same distribution for all trees in the forest. Random Forest has given robust and improved results of classifications on standard data sets. It is providing very good competition to neural networks and ensemble techniques on different classification problems. Random Forest is related to be special type of ensembles using bagging and random splitting methods to grow multiple trees [12, 13]. There are several advantages on the Random Forest method. Especially, Random Forest can predict what features are important in the classification. It can process efficiently on large data sets. Also it can be utilized as an effective method to estimate missing data. The number of trees is 10 in the training / testing process.

Logistic Model Trees (LMT) is defined as a combination of tree induction and logistic regression. It is one of the popular techniques for the nominal prediction classes and numeric of supervised learning tasks. The LMT algorithm can deal with binary and multi-class target variables, numeric/nominal attributes and missing values [14]. The size of the tree is taken as 1 in the training/testing process.

J48 is an implementation of the C4.5 decision tree learner. A decision tree is a flowchart like tree where each internal node, branch and leaf nodes denote a test on an attribute, the outcome of the test and classes, respectively. To classify an unknown pattern, the attribute values of the pattern are tested against the decision tree. The classification process is continued till no further split is possible [15]. The size of the tree is taken as 3 in the training / testing process. The sizes of the tree for all the decision tree classifiers are selected the values providing the best performance in classification.

For the 5-fold cross validation, classifier probabilities are determined for candidates in 20% of the training cases, based on a classifier trained on the remaining 80% of cases in the training set. Classification processes were provided by using data mining software called the Weka tool version 3.7.7 which is available from http://www.cs.waikato.ac.nz/ ml/weka. Tests are done on a PC with Intel Core i7, 1.90 GHz CPU and 4.00 GB RAM.

III. RESULTS

Various classification methods have been utilized for feature extraction in medical pattern recognition. In our study, geometric features based on the basic morphological image processing of 2D patterns and patient information properties were utilized for feature extraction. Our proposed CAD system is performed with the number of 12, 10, 8, 6, 4 and 2 features by using RF, LMT and J48 decision tree classifiers. Six features providing the best performance in sensitivity of CAD system are gender, area, ellipticity, age, solidity and eccentricity for RF and J48 classifiers.

The classifiers were compared for each feature number, and overall performance results of the proposed CAD system were given in Table 3. Furthermore, confusion matrixes of the classifiers having the best performance of the sensitivity in the proposed CAD system were shown in Table 4. The performance measurements are given by,

$$Sensitivity = \frac{TP}{TP+FN} \tag{1}$$

$$Specificity\ (SPC) = \frac{TN}{TN+FP} \tag{2}$$

$$TCA = \frac{TP+TN}{Total\ Number\ of\ Dataset} \tag{3}$$

$$RMSE = \sqrt{\frac{\sum(y^{'}-y)^2}{n}} \tag{4}$$

where TP and FN are the number of nodules classified as true positive and false negative, respectively. TCA, the abbreviation of Total Classification Accuracy, represents the probability of correctly classified patterns. The RMSE denotes the root of mean squared error. For RMSE, y, $y^{'}$ and n denote actual value, predicted value and number of data patterns, respectively. Sensitivity is the number of correctly predicted positives divided by the total number of positive cases. Specificity is the number of correctly predicted negatives divided by the total number of negative cases. AUROC represents the area under the receiver operating characteristic curve. Kappa statistics is a chance-corrected measure of agreement between the classifications and the true classes. If Kappa is equal to 1, it indicates perfect agreement. If Kappa is equal to 0, it represents chance agreement.

In table 4, A, B and C denote the confusion matrixes of RF classifier, LMT classifier and J48 classifier, respectively. In confusion matrix, “a” depicts positive class (Nodule) and “b” is negative class (Non-nodule).

The area under ROC curves as a function of number of features is illustrated in Fig. 4. As seen in the figure, AUROC has the best value with 12 features for LMT and with 6 features for RF and J48 classifiers. In addition, Kappa statistics (0.748) as well as sensitivity and specificity are higher for RF classifier using 6 features.

Table 3
Overall performance results of the proposed CAD system.

	Feature Number	Sensitivity (%)	SPC (%)	TCA (%)	AUROC	Kappa	RMSE
RF	12	86.3	76.8	73.5	0.838	0.448	0.423
LMT	**12**	**87.4**	**84.2**	**86.5**	**0.923**	**0.726**	**0.336**
J48	12	64.2	68.5	79.4	0.802	0.601	0.379
RF	10	93.7	85.0	72.4	0.827	0.410	0.431
LMT	10	85.3	80.0	80.6	0.848	0.604	0.405
J48	10	63.2	65.0	73.5	0.800	0.481	0.426
RF	8	95.8	91.1	77.6	0.854	0.527	0.425
LMT	8	88.4	82.8	80.6	0.851	0.600	0.397
J48	8	63.2	65.0	73.5	0.800	0.481	0.426
RF	**6**	**90.5**	**87.5**	**87.6**	**0.923**	**0.748**	**0.324**
LMT	6	84.2	79.2	80.6	0.857	0.605	0.398
J48	**6**	**89.5**	**85.9**	**85.9**	**0.858**	**0.712**	**0.363**
RF	4	86.3	81.7	82.4	0.885	0.640	0.371
LMT	4	87.4	81.8	80.6	0.799	0.601	0.417
J48	4	86.3	80.6	80.0	0.810	0.590	0.414
RF	2	73.7	64.8	68.2	0.766	0.352	0.497
LMT	2	80.0	70.3	71.2	0.781	0.406	0.438
J48	2	88.4	77.1	71.2	0.698	0.392	0.448

Table 4
Confusion matrixes classified by the proposed methods.

A

	Predict	
Actual	a	b
A	86	9
B	12	63

B

	Predict	
Actual	a	b
A	83	12
B	11	64

C

	Predict	
Actual	a	b
a	85	10
b	14	61

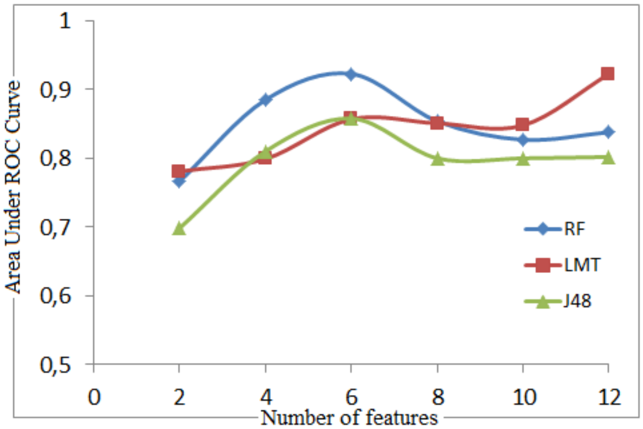

Fig. 4: Area under ROC curves as a function of number of features.

A ROC curve is usually utilized as a technique to visualize the performance of classifiers and is extremely useful to compare the performance of different classifiers in medical decision-making. The graph denotes the tradeoff between the true positive and false positive rates. The Area under ROC (*AUROC*) used here is largely adopted to represent the expected performance of a classifier. The AUROC of a classifier is equivalent to the probability that the classifier ranks a randomly chosen positive instance higher than a randomly chosen negative instance [16]. ROC curves showing the performance of our proposed CAD system are plotted in Fig. 5.

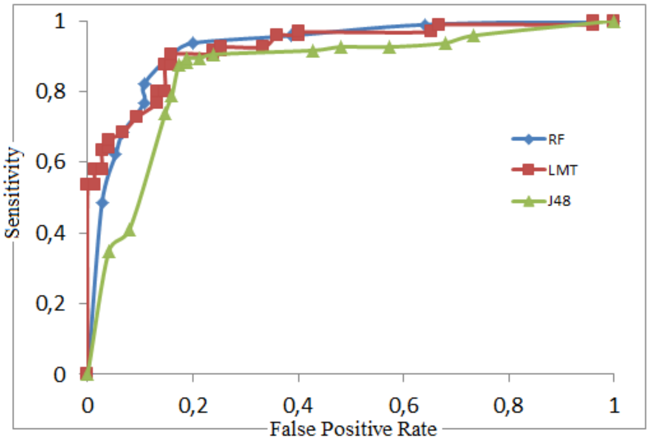

Fig. 5: ROC curves showing CAD performance with decision the methods proposed.

III.1 Performance Evaluations

To evaluate the performance of the proposed CAD system, the results of this study were compared with previously reported CAD systems. Opfer and Wiemeker utilized the dataset comprised of 93 cases (2-3 mm slice thickness) with 127 nodules [17]. Gori et al. utilized a total of 19 CT scans containing 45 internal nodules [18]. Messay et al. used a total of 84 CT scans with a total of 143 nodules in the range of 3-30 mm in nodule size [19]. Suzuki et al. utilized low-dose CT images scanned from 71 different patients with a total of 121 nodules (8-20 mm nodule size interval), totaling 101 CT scans (10 mm slice thickness and 0.586-0.684 pixel interval) [20].

A comparison of the performances of reported CAD systems is shown in Table 5. As seen in the table, the proposed study achieves a sensitivity of 90.5 % in the range of 2-20 mm nodule size for RF classifier by using only six features. All other CAD systems also have reasonable sensitivity values in pulmonary nodule detection. It is extremely important to consider the small nodule size in a CAD system. This increases the probability of early detection of nodule. By these results, it can be seen that the proposed CAD system represents a relatively high sensitivity.

Table 5
Comparison of the performances of reported CAD systems.

CAD System	Nodule Size (mm)	Reported Sensitivity (%)
Opfer and Wiemeker [17]	≥ 4	74.0
Gori et al. [18]	≥ 5	74.7
Hardie et al. [6]	NA	78.1
Suzuki et al. [20]	8-20	80.3
Messay et al. [19]	3-30	82.66
Proposed study	**2-20**	**90.5**

IV. CONCLUSIONS

In this paper, a new computer-aided system for pulmonary nodule detection from CT imagery is presented by using various decision tree algorithms.

Various classification algorithms for CAD systems have been extensively studied in the literature. In order to reduce the complexity of the algorithm and the computational load, the use of fewer features is extremely important, while maintaining an acceptable detection performance. For example, the CAD system in Messay et al. [19] uses 40 features selected from a set of 245 features with sensitivity of 82.66 %, Hardie et al. [6] uses a subset of 46 features selected from a set of 114 features by sensitivity of 78.1 %, respectively. In this study, proposed CAD system utilizes 6 features by RF classifier with sensitivity of 90.5 %. In addition, as shown in Table 4, false positive (FP) rate is shown to decline in RF classifier. In addition, the sensitivity of J48 with 2 features is very close to the sensitivity of RF and J48 with 6 features. However, the values of the specificity (SPC), TCA, Auroc and Kappa are small for J48 with 2 features.

An important feature of CAD systems desired by radiologists is that it is able to detect small nodule patterns. The dataset in our study is composed of nodules with relatively smaller diameters (> 2mm), as shown in Fig. 1, and Table 5. In conclusion, our proposed CAD system using RF classifier from decision trees is a promising method for the detection of pulmonary nodules.

ACKNOWLEDGMENT

This work was supported by Scientific Research Projects Coordination Unit of Istanbul University, Project Numbers: 24014 and 14381.

REFERENCES

[1] Cancer Facts and Figs, *The American Cancer Society*, 2009.

[2] Doi K., "Computer-aided diagnosis in medical imaging: historical review, current status and future potential", *Computerized Medical Imaging and Graphics*, 31(4–5):198–211, 2007.

[3] Iwano S., Nakamura T., Kamioka Y., Ikeda M., Ishigaki T., "Computer-aided differentiation of malignant from benign solitary pulmonary nodules imaged by high-resolution CT", *Computerized Medical Imaging and Graphics*, 32:416-422, 2008.

[4] Chen H., Zhang J., Xu Y., Chen B., Zhang K., "Performance comparison of artificial neural network and logistic regression model for differentiating lung nodules on CT scans", *Expert Systems with Applications*, 39, 11503-11509, 2012.

[5] Retico, A., Delogu, P., Fantacci, M.E., Gori, I., Martinez, A.P., "Lung nodule detection in low-dose and thin-slice computed tomography", *Computers in Biology and Medicine*, 38, 525-534, 2008.

[6] Hardie R.C., Rogers S.K., Wilson T., Rogers A., "Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiograhps", *Medical Image Analysis*, 12, 240-258, 2008.

[7] Lee M.C., Boroczky l., Sungur S-K., Cann A.D., Borczuk A.C., Kawut S.M., Powell C.A., "Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction", *Artificial Intelligence in Medicine*, 50:43-53, 2010.

[8] Wang Q., Kang W., Wu C. Wang B., "Computer-aided detection of lung nodules by SVM based on 3D matrix patterns", *Clinical Imaging*, Doi:10.1016, 2012.

[9] Lee S.L.A., Kouzani A.Z., Hu E.J., "Random forest based lung nodule classification aided by clustering", *Computerized Medical Imaging and Graphics*, 34, 535-542, 2010.

[10] Gonzales, R., Woods, R., "Digital Image Processing", *Prentice Hall*, 2007.

[11] Solomon C., Breckon T., "Fundamentals of Digital Image Processing: A practical approach with examples in Matlab", *Wiley-Blackwell*, 2011.

[12] Breiman, L., "Random forests", *Technical report*, Statistics department, University of California, Berkeley, 1999.

[13] Breiman, L., "Random forests", *Machine Learning*, 45, 5-32, 2001.

[14] Landwehr, N., Hall, M., Frank, E., "Logistic Model Trees", *Machine Learning*, 95(1-2), 161-205, 2005.

[15] Quinlan, R., "C4.5: Programs for machine learning", *Morgan Kaufmann Publishers*, San Mateo, CA, 1993.

[16] Fawett, T., "ROC graphs: notes and practical considerations for data mining researches", *Technical report*, HPL-2003-4 HP Labs, 2003.

[17] Opfer, R., Wiemker, R., "Performance analysis for computer-aided lung nodule detection on LIDC data", *Proceedings of SPIE Medical Imaging*, 6515, 65151C, 2007.

[18] Gori, I., Fantacci, M., Martinez, A.P., Retico, A., "An automated system for lung nodule detection in low dose computed tomography", *Proceedings of the SPIE on Medical Imaging, Computer Aided Diagnosis*, 6514, 65143R, 2007.

[19] Messay T., Hardie R.C., Rogers S.K., "A new computationally efficient CAD system for pulmonary nodule detection in CT imagery", *Medical Image Analysis*, 14:390-406, 2010.

[20] Suzuki, K., Armato III, S.A., Li, F., Sone, S., Doi, K., "Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography", *Medical Physics*, 30, 1602-1617, 2003.

[21] Hanamiya M., Aoki T., Yamashita Y., Kawanami S., Korogi Y., "Frequency and significance of pulmonary nodules on thin-section CT in patients with extrapulmonary malignant neoplasms", *European Journal of Radiology*, 2010.