The Lung TIME—Annotated Lung Nodule Dataset and Nodule Detection Framework

Martin Dolejší, Jan Kybic
Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic

Michal Polovinčák, Stanislav Tůma
Department of Imaging Methods, Faculty Hospital Motol in Prague, Czech Republic

ABSTRACT

The Lung Test Images from Motol Environment (Lung TIME) is a new publicly available dataset of thoracic CT scans with manually annotated pulmonary nodules. It is larger than other publicly available datasets. Pulmonary nodules are lesions in the lungs, which may indicate lung cancer. Their early detection significantly improves survival rate of patients. Automatic nodule detecting systems using CT scans are being developed to reduce physicians' load and to improve detection quality. Besides presenting our own nodule detection system, in this article, we mainly address the problem of testing and comparison of automatic nodule detection methods. Our publicly available 157 CT scan dataset with 394 annotated nodules contains almost every nodule types (pleura attached, vessel attached, solitary, regular, irregular) with 2-10mm in diameter, except ground glass opacities (GGO). Annotation was done consensually by two experienced radiologists. The data are in DICOM format, annotations are provided in XML format compatible with the Lung Imaging Database Consortium (LIDC). Our computer aided diagnosis system (CAD) is based on mathematical morphology and filtration with a subsequent classification step. We use Asymmetric AdaBoost classifier. The system was tested using TIME, LIDC and ANODE09 databases. The performance was evaluated by cross-validation for Lung TIME and LIDC, and using the supplied evaluation procedure for ANODE09. The sensitivity at chosen working point was 94.27% with 7.57 false positives/slice for TIME and LIDC datasets combined, 94.03% with 5.46 FPs/slice for the Lung TIME, 89.62% sensitivity with 12.03 FPs/slice for LIDC, and 78.68% with 4,61 FPs/slice when applied on ANODE09.

Keywords: Lungs, nodule detection, CAD development, performance evaluation.

1. INTRODUCTION

Lung cancer is one of the leading causes of cancer related death (31% of males and 26% of females) in the USA[1] and other developed countries. CT is considered to be the most accurate imaging modality available for early detection and diagnosis of lung cancer.[2] It allows detecting pathological deposits as small as 1mm in diameter.[3] These deposits are called lung nodules. The best way how to improve patients' five-year survival rate is early detection—the survival rate is 49% for the lung cancer localized in early-stage (stage I).[4] Compare this with the survival rate for all stages and types of lung cancer combined, which is only 14%.

Automatic nodule detection systems can improve the nodule detection rate,[3] serving as a second reader, or identifying suspect regions for a human radiologist. A lot of CAD systems for nodule detection are being developed. These CAD systems usually have two stages, because this architecture improves efficiency. The first stage quickly detects suspicious regions (candidate detector), The second processing stage improves specificity by removing not-nodule structures detected in the first one.

There are two main groups of candidate detectors: density-based and model-based approaches. Density-based detection methods concentrate on voxel intensities and employ e.g. multiple thresholding,[5] region-growing,[6] and many others, to identify nodule candidates. Model-based detection approaches take advantage of the relatively compact shape of (small) lung nodules. Techniques such as "N-Quoit filter",[7] template-matching,[8] and object-based deformation have been proposed to identify small nodules in the lungs. A classifier is usually used for the

e-mail: dolejm1@fel.cvut.cz

second stage. It is e.g. a threshold classifier,[9] 3D Markov random field models,[10] or some other pattern recognition techniques. Large number of different features based on intensity, volume, shape and other measurement are used as inputs for the classifiers.

For developing of a nodule detection system a dataset with annotated nodules (ground truth) is needed. This training dataset defines nodules for the system, and should be as large as possible, and capture as many different nodule types as possible. To obtain a universally applicable detector, the data should be acquired with different CT machines under different settings. Otherwise the resulting detection system performance would suffer when the acquisition conditions change, or a different machine is used. The most important aspect of a good training dataset is a correct annotation. There are several issues why the annotation made by two radiologist can differ:

1. The true ground truth is not known, unless biological samples are examined, or until we wait a sufficiently long time to evaluate the changes of the suspected nodule in time. Radiologists can have different experience and hence interpret the same observations differently.
2. Radiologists can have different trade-offs between specificity and sensitivity.
3. Consensus search can make the annotations more robust yet less precise.
4. There is no standard definition of a nodule. Instead, nodules are often defined indirectly, using examples from annotated datasets.
5. Attempts to standardize a nodule definition such as proposed by the Lung Imaging Database Consortium (LIDC)[11] are still not universally accepted by radiologist from other groups.

1.1 Publicly available datasets

There are very few publicly available thoracic CT scan datasets with annotated nodules. The LIDC* database contains 84 scans with annotated nodules. Small nodules are marked by a single point, for large ones all boundary voxels are marked. Another public dataset is the part of the "Nelson study"[12] published by the ANODE09† project. It contains 5 example CT scans with annotations (single point per nodule), and 50 test scans without annotations. Results on the test set can be evaluated at the web pages of the project.

2. THE LUNG TIME DATASET

The Lung TIME dataset (http://cmp.felk.cvut.cz/projects/LungTIME/) which we have collected consists of two parts. The first part (TIME1) contains data of adolescent patients from the Faculty Hospital in Motol, Prague, Czech Republic with 148 scans, slice thickness 5 mm, slice spacing 1 mm and transversal resolution 0.58±0.06 mm. There are 355 nodules in this part. This dataset has large slice overlap and even the smallest nodules are visible in more than one slice. Patients were included in this study because of suspicion of soft tissue or bone tumors metastasis in lungs. Some patients had no findings, some patients underwent multiple scans.

The second part (TIME2) consists of 9 adult patient data also from Motol hospital. CT scans was acquired with slice thickness 5 mm, slice spacing 5 mm and transversal resolution 0.71±0.05 mm. There are 39 nodules in this part. Since the TIME2 dataset has much lower slice overlap, some small lesions are visible only in one slice. Adult patients were scanned during primary lung cancer treatment. The distribution of nodules over scans is shown in Figure 1. Example of nodule spatial distribution in one scan is shown in Figure 2.

Both parts of the Lung TIME dataset were acquired by Somatom AR Star CT scanner, a single detector machine by Siemens (1998). A low-dose spiral technique[13] (LDSCT) was used. We have used the following acquisition parameters: voltage of X-ray generator output 110kV, X-ray tube current 63mA, time of rotation 1.5s, pitch 2. All patients were imaged in the cranio-caudal direction, in maximal inspiration following three deep inspirations. The scans were preformed during two breath holds, resulting in a slight respiratory discontinuity between the upper and lower part of the 3D image. Volumetric data are saved slice by slice in DICOM 3.0 format, the scanning parameters are saved as metadata in each DICOM file.

*http://imaging.cancer.gov/programsandresources/InformationSystems/LIDC/

†http://anode09.isi.uu.nl/

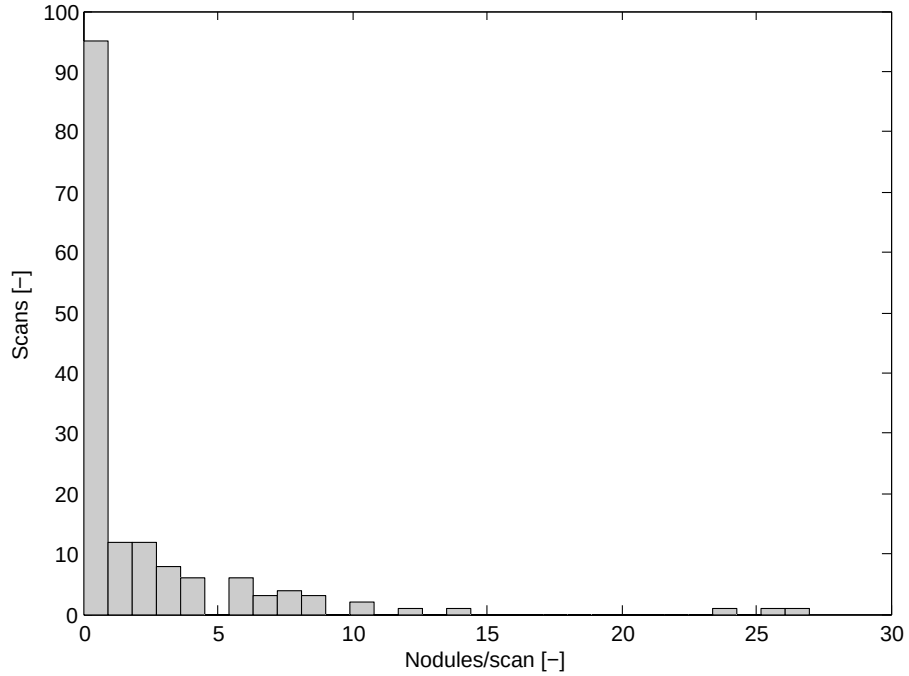

Figure 1. Distribution of nodules over scans. There is one scan with 72 nodules, not included in the histogram.

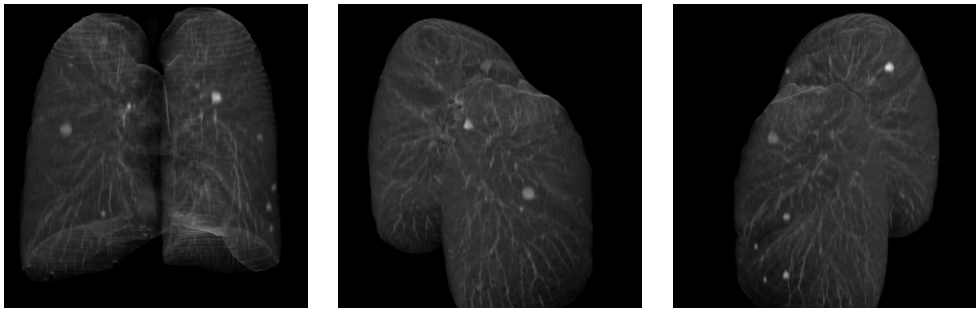

Figure 2. Example of spatial nodule distribution in one volume of the Lung TIME dataset.

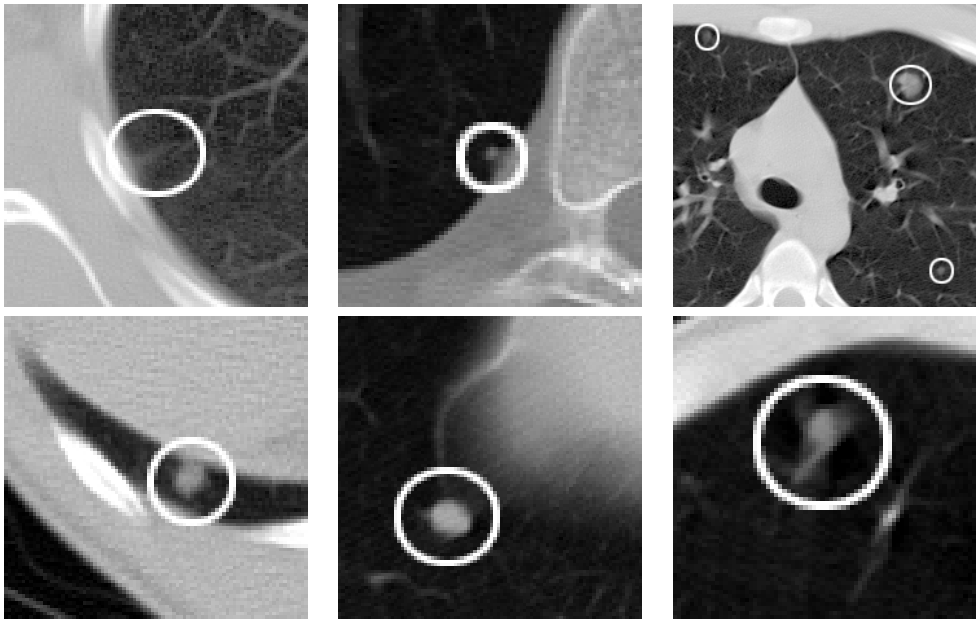

Figure 3. Examples of annotated nodules from the Lung TIME dataset.

Dataset	Scans	Slices	Nodules	Slice thickness	Slice spacing
TIME1	148	246.2	355	5.0	1.0
TIME2	9	63.8	39	5.0	5.0
LIDC[11]	38	223.6	202	2.2	1.9
ANODE09[12]	50	451.5	unknown	1	0.7

Table 1. Important parameters of datasets (mean number of slices per scan, number of nodules in the dataset, and mean slice thickness and slice spacing in mm).

The nodule annotation is a result of a consensus of two experienced radiologists, after each of them has performed his/her independent reading. Each nodule was marked by a single point near its center of gravity in each slice containing even a small part of a nodule. The dataset contains a large number of nodules of different types (Figure 3). The annotations were made using a ScanView software by Dr. Jan Krásenský and converted to XML formatted files compatible with the LIDC dataset.

Our Lung TIME dataset is now the largest publicly available dataset. This dataset is also unique because it contains a huge number of CT examination of adolescent patients. The summary of publicly available datasets (Lung TIME, LIDC, and Anode09) parameters are in Table 1.

2.1 Automatic Nodule Detection

Our automatic nodule detection system is designed mainly to identify the suspect regions for a human radiologist. Because of this, we have preferred a highly sensitive system with a slightly higher number of false positives (FPs) detections, minimizing the number of false negatives (oversights).

The system works as follows. First, we segment the lungs by thresholding[14] and further binary image processing. As a parenchymal nodule candidate detector we use an algorithm based on multi-scale filtering by a Gaussian filter. The local maxima present at all scales are treated as nodule candidates. For juxta-pleural nodule candidates detection we use methods of mathematical morphology. The irregularities on the lung boundary are detected by subtracting of morphologically closed lung mask from original one.[14] The progress of juxta-pleural detector is illustrated in Figure 4. Finally each suspicious object is segmented by local thresholding and a set of 38 features derived from shape, intensity, covariance matrix, and 10 different Hessian matrices is computed. Volume, shape and intensity features[15] are measured on the segmented nodule candidate. For Hessian

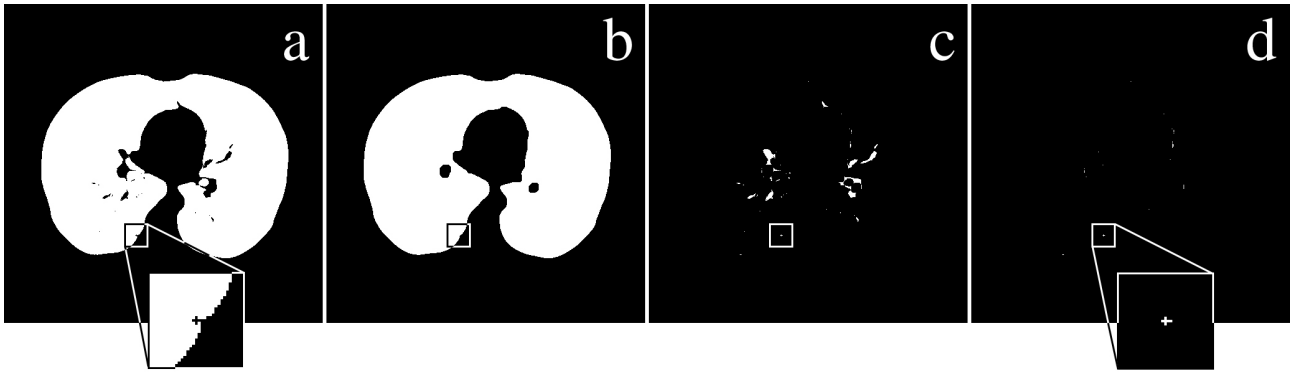

Figure 4. Juxta-pleural nodule detection: (a) lung mask, (b) closed lung mask, (c) the difference mask, (d) final mask after large object and non-edge point elimination. White pixels are the detected juxta-pleural nodule candidates, a real nodule is shown magnified in images (a), (d).

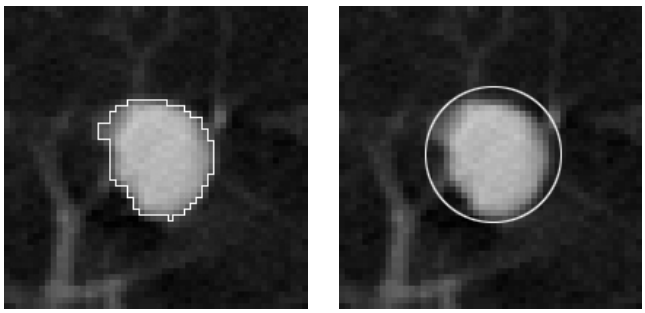

Figure 5. Example of nodule segmentation using thresholding (left) and outer ellipsoid obtained from covariance matrix eigenvalues.

features computations a $20 \times 20 \times 20$mm window is used. Example of threshold segmentation, and visualization of the covariance matrix eigenvalues is in Figure 5.

Then a classification step is applied. We use Asymmetric AdaBoost,[16] a greedy, cost sensitive algorithm to construct a complex—strong classifier as a linear combination of simple—weak ones. The only parameter of cost function is asymmetry[15] k. The cost of a wrong classification for the whole "nodule" class is set to be k-times greater than cost of a wrong classification for the whole "non-nodule" class:

$$kN_t c_t = N_f c_f$$

where n_t and n_f stand for number of nodules and non-nodules respectively, C is cost of one nodule or not-nodule in accordance to the lower index.

2.2 Experiments

The purpose of the experiments was twofold: first to test the performance of our detection system on the available datasets (TIME, LIDC, ANODE) and second, to evaluate the effect of the differences between datasets on the classification results. We evaluate the performance differently for Lung TIME+LIDC and ANODE09. For TIME+LIDC we consider a nodule correctly detected, when the ground truth point is no further than 5mm from any point detected by our automatic system. ANODE evaluation consider a finding as true nodule, when its distance to any ground truth nodule in the scan is less than 1.5 times the radius of that nodule.

The nodule annotations for TIME and LIDC are available, thus we could use a cross-validation evaluation scheme on both datasets together and also on each dataset separately, resulting in three experiments and three ROC curves. In each experiment, all detected candidates are used to construct a feature vector set. The information about the patient, scan, or dataset is not used. We divide the feature vector set into 10 subsets. All 10

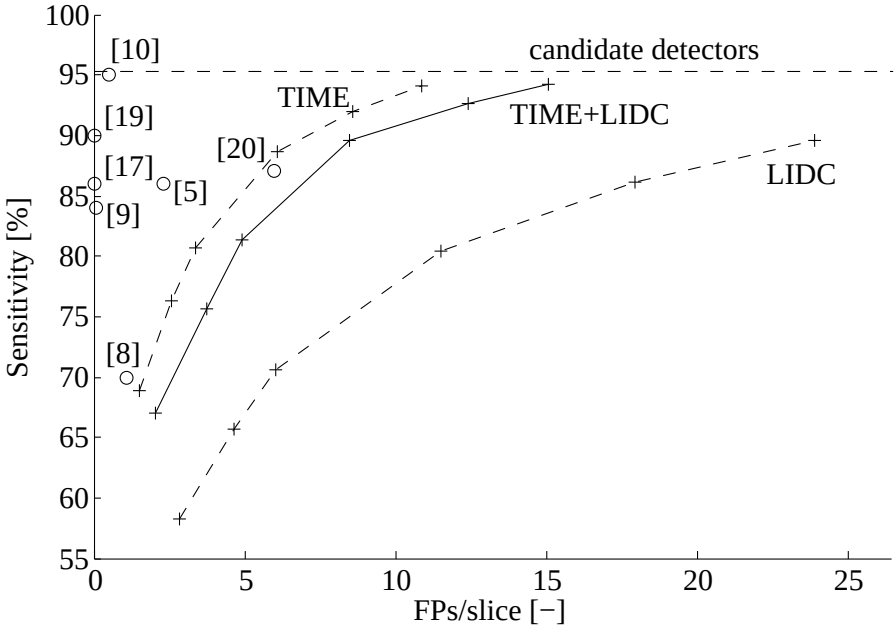

Figure 6. ROCs of our system with Asymmetric AdaBoost applied on TIME and LIDC datasets show, that results of algorithm are dataset dependent. The parameter changing along the ROC curves is the asymmetry k of the Asymmetric AdaBoost. (Results with respect to FPs/scan look very similar.)

FPs/scan	**1/8**	**1/4**	**1/2**	**1**	**2**	**4**	**8**	**average**
small nodules	0.009	0.017	0.035	0.069	0.094	0.113	0.151	0.070
large nodules	0.010	0.020	0.040	0.081	0.110	0.132	0.165	0.080
isolated nodules	0.011	0.023	0.046	0.092	0.125	0.150	0.150	0.085
vascular nodules	0.005	0.010	0.019	0.038	0.052	0.091	0.117	0.047
pleural nodules	0.009	0.018	0.035	0.071	0.096	0.096	0.135	0.066
peri-fissural nodules	0.021	0.042	0.084	0.168	0.229	0.229	0.314	0.155
all nodules	0.009	0.019	0.037	0.075	0.102	0.122	0.157	0.074

Table 2. Standard evaluation table of ANODE09 project. It shows the sensitivity of the system on several levels of FP/scan.

combinations of 9 subsets are used for learning, the remaining one for testing. Then we compute the average sensitivity and number of FPs/slice over 10 testing sets. We run such a cross-validation multiple times with different cost functions $k = 1, 2, 3, 10, 30, 100$ to obtain ROC curve. Results of cross-validation for TIME and LIDC datasets are in Figure 6.

Annotations for ANODE09 dataset are not public, so we trained the Asymmetric AdaBoost ($k = 0.7, 1, 2, 3, 10, 30, 100$) on TIME and LIDC sets and applied those classifiers onto ANODE09. Finally we interpolate the results between classifier's working points using a classifier output before binarization. Results of our system obtained from the evaluation algorithm at the project's web pages are in Figure 7. The evaluation was divided into two parts, because the ANODE09 evaluation algorithm has the limit of 5000 finds. We decided to show this ROC separately, because the performance of detection step is unknown (the upper limit in Figure 6. could be confusing). The standard ANODE09 table for first part of the ROC is in Table 2.

Note, that the results are database dependent. There are two reasons for that:

1. The acquisition parameters of each dataset are different. See Table 1. for details.
2. The annotation in datasets is different. Some pathologic structures annotated in LUNG Time as "nodule" are almost identical to "not-nodules" in LIDC, e.g. top left one in Figure 3.

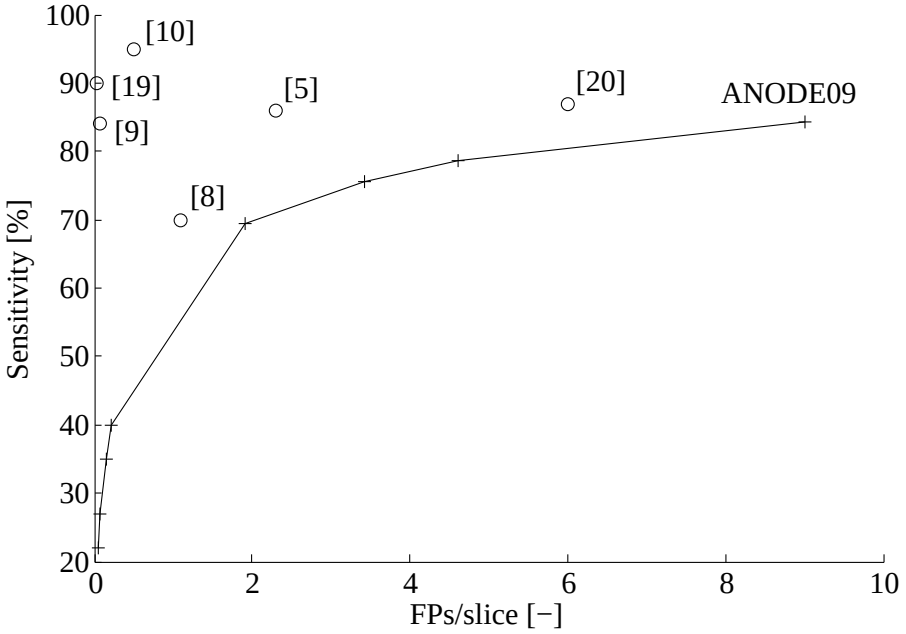

Figure 7. ROC of our system with Asymmetric AdaBoost applied on ANODE09 dataset proves our claim from TIME and LIDC comparison, that the results of algorithm are dataset dependent. The parameter changing along the ROC curve is the asymmetry k of the Asymmetric AdaBoost. (Results with respect to FPs/scan look very similar.)

3. RESULTS AND CONCLUSIONS

The new publicly available database of lung nodules was created as a collaboration between the Czech Technical University and Faculty Hospital in Motol, Prague. The database contains 394 small ($<$ 10mm) nodules of different types (juxta-pleural, parenchymal, vessel attached, regular, irregular) in 157 CT-scans, marked by two experienced radiologists. Our Lung TIME is the largest publicly available lung nodule CT database.

We have observed, that a CAD system can have a significantly different performance when tested on different datasets. It shows that it is therefore incorrect to compare detection results using independent datasets. Future algorithms for automatic nodule detection will have to work with outputs from different CT scanners under different settings. It follows from our results that creating such universal algorithms is very challenging. We conclude that to develop and train such algorithms, a concerted effort will be needed to create very large datasets of annotated CT scans acquired in a variety of conditions. With more publicly available data, the development, testing and comparison become more precise and meaningful.

We have developed a new automatic two-step CAD nodule detection algorithm based on Asymmetric AdaBoost classification.

Acknowledgments

This work was supported by the Czech Ministry of Education, project MSM6840770012; the Czech Science Foundation, project 102/07/1317; and the grant agency of Czech Ministery of Health under project NR8314-3/2005.

REFERENCES

[1] American Cancer Society, "Cancer facts & figures 2007," (December 2007).

[2] Kaneko, M., Eguchi, K., Ohmatsu, H., Kakinuma, R., Naruke, T., Suemasu, K., and Moriyama, N., "Peripheral lung cancer: screening and detection with low-dose spiral CT versus radiography," *Radiology* **201**(3), 798–802 (1996).

[3] Das, M., Mühlenbruch, G., Mahnken, A. H., Flohr, T. G., Gündel, L., Stanzel, S., Kraus, T., Günther, R. W., and Wildberger, J. E., "Small pulmonary nodules: Effect of two computer-aided detection systems on radiologist performance," *Radiology* **241**(2), 564–571 (2006).

[4] Ries, L. and et al., "Seer cancer statistics review 19731996," tech. rep., National Cancer Institution, Bethesda, MD (1999).

[5] Ko, J. P. and Betke, M., "Chest CT: Automated nodule detection and assessment of change over time—preliminary experience," *Radiology* **218**, 267–273 (2001).

[6] Fiebich, M., Wietholt, C., Renger, B. C., Armato, S., Hoffmann, K., Wormanns, D., and Diederich, S., "Automatic detection of pulmonary nodules in low-dose screening thoracic CT examinations," in [*Proceedings of SPIE*], 1434–1439, SPIE, Bellingham, Washington, USA (1999).

[7] Okumura, T., Miwa, T., Kako, J., Yamamoto, S., Matsumoto, M., Tateno, Y., Linuma, T., and Tohru, M., "Automatic detection of lung cancers in chest CT images by variable N-Quoit filter," in [*14th International Conference on Pattern Recognition*], **2**, 1671–1673, IEEE (1998).

[8] Lee, Y., Hara, T., Fujita, H., Itoh, S., and Ishigaki, T., "Automated detection of pulmonary nodules in helical CT images based on an improved template-matching technique," *IEEE Transactions on Medical Imaging* **20**, 595–604 (July 2001).

[9] Zhao, B., Gamsu, G., Ginsberg, M. S., Jiang, L., and Schwartz, L. H., "Automatic detection of small lung nodules on CT utilizing a local density maximum algorithm," *Journal of Applied Clinical Medical Physics* **4**, 248–260 (Summer 2003).

[10] Takizawa, H., Yamamoto, S., Matsumoto, T., Tateno, Y., Iinuma, T., and Matsumoto, M., "Recognition of lung nodules from X-ray CT images using 3D Markov random field models," in [*Proceedings of the 16 th International Conference on Pattern Recognition*], (2002).

[11] Armato III, S. G., McLennan, G., McNitt-Gray, M. F., Meyer, C. R., Yankelevitz, D., Aberle, D. R., Henschke, C. I., Hoffman, E. A., Kazerooni, E. A., MacMahon, H., Reeves, A. P., Croft, B. Y., and Clarke, L. P., "Lung image database consortium: Developing a resource for the medical imaging research community," *Radiology* **232**(3), 739–748 (2004).

[12] Xu, D. M., Gietema, H., de Koning, H., Vernhout, R., Nackaerts, K., Prokop, M., Weenink, C., Lammers, J.-W., Groen, H., Oudkerk, M., and van Klaveren, R., "Nodule management protocol of the NELSON randomised lung cancer screening trial," *Lung Cancer* **54**(2), 177–184 (2006).

[13] Tůma, S., Neuwirth, J., Čumlivská, E., Mališ, J., Kybic, J., Šanda, J., Fricová-Poulová, M., Adla, T., and Polovinčák, M., "Nízkodávková technika spirální CT plic v diagnostice metastatických ložiskových nález ˚u u dětí a v dorostovém věku.," *Česko-Slovenská Pediatrie* **61**(4), 179–185 (2006).

[14] Dolejší, M. and Kybic, J., "Automatic two-step detection of pulmonary nodules," in [*Proceedings of SPIE*], *Medical Imaging 2007: Computer-Aided Diagnosis* **6514**, 3j–1–3j–12, SPIE (February 2007).

[15] Dolejší, M., Kybic, J., Polovinčák, M., and Tůma, S., "Reducing false positive responses in lung nodule detector system by Asymmetric AdaBoost," in [*Proceedings of 2008 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*], 656–659, IEEE (May 2008).

[16] Viola, P. and Jones, M., "Fast and robust classification using Asymmetric AdaBoost and a detector cascade," in [*Advances in Neural Information Processing Systems 14*], Dietterich, T. G., Becker, S., and Ghahramani, Z., eds., MIT Press, Cambridge, MA (2002).

[17] Zhang, X., McLennan, G., Hoffman, E. A., and Šonka, M., "Automated detection of small-size pulmonary nodules based on helical CT images," in [*Proceedings of 19th International Conference, Information Processing in Medical Imaging*], 664–676, Springer-Verlag, Berlin, Germany (July 2005).

[18] Armato III, S. G., Altman, M. B., Wilkie, J., Sone, S., Li, F., Doi, K., and Roy, A. S., "Automated lung nodule classification following automated nodule detection on CT: A serial approach," *Medical Physics* **30**, 1188–1197 (June 2003).

[19] Mendonça, P. R. S., Bhotika, R., Zhao, F., and Miller, J. V., "Lung nodule detection via Bayesian voxel labeling," in [*Information Processing in Medical Imaging, 20th International Conference*], *Lecture Notes in Computer Science*, 134–146, Springer, Berlin, Germany (July 2007).

[20] Gori, I., Bellotti, R., Cerello, P., Cheran, S. C., Nunzio, G. D., Fantacci, M. E., Kasae, P., Masala, G. L., Martinez, A. P., and Retico, A., "Lung nodule detection in screening computed tomography," in [*Proceedings of IEEE Nuclear Science Symposium Conference*], **4322**, IEEE (2006).