Automated Identification of Lung Nodules

S. L. A. Lee, A. Z. Kouzani, and E. J. Hu
School of Engineering and IT, Deakin University
Waurn Ponds, VIC 3217, AUSTRALIA
slale, kouzani, erichu@deakin.edu.au

***Abstract*- A system that can automatically detect nodules within lung images may assist expert radiologists in interpreting the abnormal patterns as nodules in 2D CT lung images. A system is presented that can automatically identify nodules of various sizes within lung images. The pattern classification method is employed to develop the proposed system. A random forest ensemble classifier is formed consisting of many weak learners that can grow decision trees. The forest selects the decision that has the most votes. The developed system consists of two random forest classifiers connected in a series fashion. A subset of CT lung images from the LIDC database is employed. It consists of 5721 images to train and test the system. There are 411 images that contained expert- radiologists identified nodules. Training sets consisting of nodule, non-nodule, and false-detection patterns are constructed. A collection of test images are also built. The first classifier is developed to detect all nodules. The second classifier is developed to eliminate the false detections produced by the first classifier. According to the experimental results, a true positive rate of 100%, and false positive rate of 1.4 per lung image are achieved.**

I. INTRODUCTION

Lung nodule refers to a range of abnormalities considered as small, round opacity, roughly spherical, restricted on abnormal tissue [1]. Some lung abnormalities lead to lung cancer which is a top cancer killer in the world. Detection of lung nodules can be achieved through computed tomography (CT) imaging.

With the constant improvement in the CT imaging technology, the amount of data per subject increases continuously. Currently, an average of 300 image slices per subject can be acquired in a scan. Whilst the additional data benefits the accuracy of nodule visualisation, it also increases the complexity of inspection and interpretation, and may affect the evaluation judgment by expert radiologists.

An intelligent diagnostic system may serve as a preliminary interpreter to assist the expert radiologists. Recent studies show that there exist an inter-reader variability in the detection of nodules by the expert radiologists [2]. Therefore, automated approaches may help improve the precision of lung nodule detection.

In the past years, numerous methods have been developed by researchers for automated detection of nodules in lung images. Some of these methods are reviewed in the following section. However, since automated lung nodule detection is a very challenging problem, the achieved lung nodule detection rate can be further increased. There is still room for improvement in detection accuracy as well as speed of lung nodules.

The main contribution of this paper is the utilisation of the random forests to formulate a method for automated detection of lung nodules in CT images. A random forest [3] is an ensemble learning method that grows many classification trees. To classify an object from an input vector, the input vector is put down each of the trees in the forest. Each tree gives a classification. The forest selects the classification that has most votes. The random forest has demonstrated to be accurate and fast.

The developed method employs the concept of pattern classification for the detection of lung nodules. Two pattern classes are formed namely nodule and non-nodule. The random forest-based classifier is trained to classify the patterns belonging to the nodule and the non-nodule classes. A procedure for a two stage detection algorithm consisting of two random forest-based classifiers is proposed, and tested. The developed method minimises the false-detection rate of lung nodule lesion within multi-slice helical CT images and achieves full nodule detection.

II. EXISTING LUNG NODULE DETECTION METHODS

In the following, we have reviewed some existing approaches that can automatically detect lung nodules.

Klik et al [4] formed an algorithm to differentiate between benign and malignant nodules. For the benign nodule, the characteristics properties are flattened surface and direct attachment to plate-like structures neighbourhood in the fissures. Hessian matrix based on the eigenvalues was utilised to detect the fissure. Hough transforms was performed to the nodule boundary and the detected fissure voxels to enable an accurate partition of benign nodules. The system is trained using a *k*NN classifier with k = 10, Parzen classifier ($\sigma = 1.0$), a linear discriminant classifier and a quadratic discriminant classifier. A specificity of 95% with the sensitivity higher than 65% for quadratic discriminant classifier were achieved.

Clifford et al [5] described a technique using the wavelet and bi-orthogonal wavelet as the pre-processing module to form a precise and sharp CT image. A thresholding followed by a morphological transform were used to extract the features. A fuzzy inference system based on the extracted features was used to locate the severity of the lung nodules. The detection rate was not reported.

Zhao et al [6] developed an algorithm based on the support vector machine and genetic algorithms for false positive nodule reduction in CT images. There were 15 features and

only 9 were incorporated in the system. By using leave-one-out cross validation, a sensitivity of 98.5% and specificity of 82.9% were obtained.

Gori et al [7] described a system compromised of a dot-enhancement filter to select the potential nodule candidate and a neural classifier to reduce the false-positive nodule detection. A sensitivity of 86.5% with false positives of 6 per scan was reported. Jia et al. [8] proposed an automated nodule detection method which could identify the pulmonary nodule. It contained segmentation of lung parenchyma, trachea and main airway bronchi elimination, filtering of nodule candidates, detection of nodule candidates, feature extraction, and classification. The classification approach used was not described in detail. A sensitivity of 95% was reported.

Takizawa et al. [9] presented a nodule discrimination method based on a statistical analysis of CT scans. They used a relationship between pulmonary nodules, false positives, and image features in CT scans. The method was applied to 218 actual thoracic CT scans with 386 actual pulmonary nodules. A receiver operating characteristic analysis was used and the result was 93.1%.

Some existing methods used classification for detection of lung nodules. One of the current trends is the appearance of ensemble learners which utilized a large amount of weak classifiers with boosting. Ochs et al. [10] described a method for voxel-by-voxel classification of airways, fissures, nodules, and vessels from CT images. Twenty-nine CT scans were used. The AdaBoost algorithm was used. The feature set consisted of voxel attenuation and a small number of features based on the eigenvalues of the Hessian matrix. The detection rate for the nodule was 94.5%.

Considering the lung nodule detection literature, it is clear that there is still room for improvement in the lung nodule detection accuracy as well as speed.

III. Random Forests

A random forest [3] predictor is an ensemble of individual classification tree predictors. For each observation, each individual tree votes for one class and the forest predicts the class that has the plurality of votes. The number of randomly selected variables $m_{\_try}$ to be searched through for the best split at each node has to be determined.

Whilst a node is split using the best split among all variables in standard trees, in a random forest the node is split using the best among a subset of predictors randomly chosen at that node. The largest tree possible is grown and is not pruned. The root node of each tree in the forest contains a bootstrap sample from the original data as the training set. The observations that are not in the training set, are referred to as "out-of-bag" observations.

Since an individual tree is unpruned, the terminal nodes can contain only a small number of observations. The training data are run down each tree. If observations *i* and *j* both end up in the same terminal node, the similarity between *i* and *j* is increased by one. At the end of the forest construction, the similarities are symmetrised and divided by the number of trees. The similarity between an observation and itself is set to one. The similarities between objects form a matrix which is symmetric, and each entry lies in the unit interval *[0, 1]*. Breiman defines the random forest as [3]:

> A random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(\mathbf{x},\Theta_k), k=1,\ldots\}$ where $\{\Theta_k\}$ are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\boldsymbol{x}$.

A summary of the random forest algorithm for classification is given below [11]:

- Draw *K* bootstrap samples from the training data.
- For each of the bootstrap samples, grow an unpruned classification tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample *m* of the predictors and choose the best split from among those variables.
- Predict new data by aggregating the predictions of the *K* trees, i.e., majority votes for classification, average for regression.

The random forest approach works well because of: (i) the variance reduction achieved through averaging over learners, and (ii) randomised stages decreasing correlation between distinctive learners in the ensemble.

The generalisation error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them. Using a random selection of features to split each node yields error rates that compare to AdaBoost [12]. An estimate of the error rate can be obtained, based on the training data, by the following [11]:

- At each bootstrap iteration, predict the data that is not in the bootstrap sample, called ``out-of-bag" data, using the tree which is grown with the bootstrap sample.
- Aggregate the out-of-bag predictions. On the average, each data point would be out-of-bag around 36.8% [13] of the times. Calculate the error rate, and call it the "out-of-bag" estimate of error rate.

With regard to the 36.8%, the random forest forms a set of tree-based learners. Each learner gets different training set of *n* instances extracted independently with replacement from the learning set. The bootstrap replication of training instances is not the only source of randomness. In each node of the tree the splitting attribute is selected from a randomly chosen sample of attributes. As the training sets of individual trees are formed by bootstrap replication, there exists on average $1/e \approx 36.8\%$ of instances not taking part in construction of the tree. The random forest performs well compared to some other popular classifiers. Also, it has only two parameters to adjust: (i) the number of variables in the

random subset at each node, and (ii) the number of trees in the forest. It learns fast.

We employ the random forest algorithm to form the proposed system for detection of lung nodules in 2D CT images. The developed system classifies the lung nodule patterns against the non-nodule patterns within the lung images.

IV. Lung Image Data

A large collection of CT lung images from the Lung Imaging Database Consortium (LIDC) database [14] was acquired. This collection contained 42 scans of different subjects. Each scan contained a varying number of image slices. The images were captured by different CT scanners including Siemens, Toshiba, and General Electric.

Out of the 42 original scans, we kept only 32 scans. The reason is that the images associated with 11 scans had gray-level ranges inconsistent with those of the all other scans. In addition, one scan contained a corrupt XML file so that the nodule information could not be retrieved. Therefore, we kept 32 scans whose gray-level ranges were similar varying between 0 and about 4100. These 32 scans contained a total of 5721 image files. There were 411 images out of 5721 images that contained nodule patterns. These scans were captured by Siemens and General Electric CT scanners with x-ray tube current exposure ranging from 75mA to 344mA. All images were of the size 512×512 in DICOM format. The pixel size varied from 0.488mm to 0.762mm. The slice thickness ranged from 2.0mm to 3.0mm.

The location of lung nodules within the images was marked by expert radiologists. A two-phase process was formulated to enable multiple radiologists at different centres to asynchronously review and annotate each CT image series. Each case was reviewed by four radiologists. In the first phase, named "blinded", each radiologist reviewed the CT series independently. In the second phase, named "unblinded", the results from all four blinded reviews were compiled and presented to each of the four radiologists for a second review. Each radiologist was able to review his/her own annotations as well as those of the other radiologists. The final unblinded review was created using the results from each radiologist's unblinded review. The nodule information was stored in a XML file for each scan. Fig. 1 shows two lung images marked by the expert radiologists.

We developed a converter program that used the information in the XML file for each scan and extracted out the nodule regions from the lung images. For nodule patterns that could fit within a 30×30 region, we extracted from the image the corresponding region surrounding the nodule pattern. There were a total of 386 such nodule patterns. On the other hand, for nodule patterns that could not fit within a 30×30 region, we extracted the entire nodule pattern first, and then resized the extracted pattern into a 30×30 region. There were a total of 817 such nodule patterns. Overall, we created 1203 30×30 nodule files.

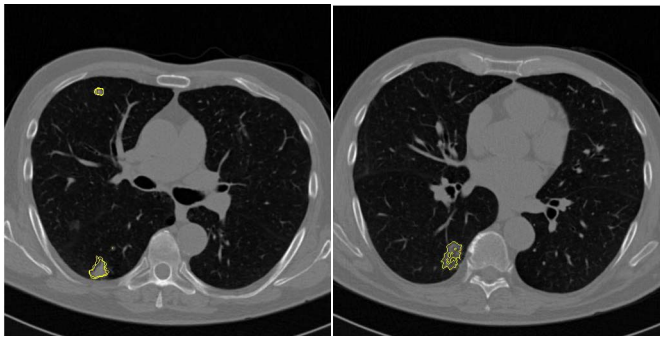

Fig. 1. Samples of lung images marked by expert radiologists from the LIDC database [14]: (left) image 17630, (right) image 17710.

We performed a study on the locations of the 1203 nodule within the entire 5721 lung images and produced a map indicating the likelihood of nodule occurrence within lung images. In the map, which is shown in Fig. 2-left, the pixel brightness relates to the likelihood of nodule occurrence. Using the map, we were able to work out the image regions in which the likelihood of a nodule presence was zero. This information was used to speed up the detection phase by excluding such regions from examination later during our experiments. Fig. 2-right displays the subset of lung images within which nodules can appear.

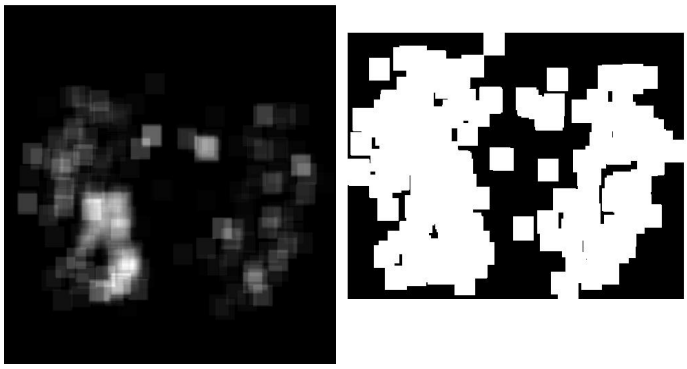

Fig. 2. (left) Map indicating the likelihood of nodule occurrence locations; (right) Subset of lung images within which nodules occur.

V. System Development Procedure

The developed system consists of two random forest classifiers connected in a series fashion used as the detector of nodule patterns within 2D lung images. The first classifier is named nodule-detector, and the second classifier is called false-positive-reducer.

A. Nodule-Detector

The nodule-detector was developed to detect as many nodule patterns as possible. To satisfy this goal, the decision making had to be done is such a way that the number of false-positive would not be minimised.

In order to develop the nodule-detector, a training set consisting of two groups of images was formed: nodule and non-nodule. The nodule group consisted of two collections of

30×30 nodule patterns. The first collection contained 386 nodule patterns that could fit within a 30×30 region (see Fig. 3-top). The second collection contained 817 nodule patterns that could not fit within a 30×30 region but resized to a 30×30 region (see Fig. 3-bottom). Overall, the nodule group contained 1203 30×30 nodule patterns.

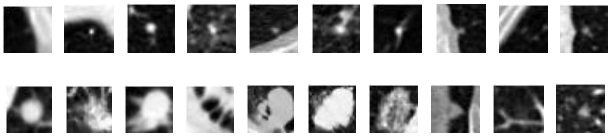

Fig. 3. Sample of 30×30 nodule group: (top) first collection, and (bottom) second collection.

The non-nodule group, on the other hand, also contained two collections of 30×30 non-nodule patterns. The first collection contained 1156 expert-marked non-nodule patterns from the 32 used scans from the LIDC database (see Fig. 4-top). The second collection contained 1203 randomly captured regions of sizes 30×30, 47×47, 64×64, and 82×82 that did not contain any nodule patterns from within the lung lobe areas (see Fig. 4-bottom). These 1203 regions were all resized to 30×30.

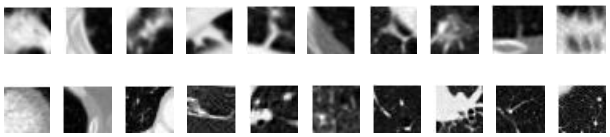

Fig. 4. Sample of 30×30 non-nodule group: (top) first collection, and (bottom) second collection.

We varied the two random forest parameters, no-of-trees-grown and no-of-variables-at-each-split, as follows. The first parameter, no-of-trees-grown, was varied from 1 to 100 with an increment of 1. For each tree grown, the second parameter, no-of-variables-at-each-split, was varied from 1 to 50 with an increment of 1. For each classifier that was made of a specific number of trees and variables, the classification error was calculated. The random forest with 50 trees and 33 variables was selected because it produced the lowest classification error amongst all tried forests. The procedure for the nodule-detector to identify nodules was as follows.

- Load forest, plus the information associated with the expert identified nodules for the images. This information includes the three-dimensional centre-of-mass location for nodules of less than 3mm, or edge-maps providing the complete outline around the nodules that were greater than 3mm.
- Each test image was loaded at a time. We randomly selected 10 images containing one or more nodules, and 10 images containing no nodules.
- A 30×30 sliding window was formed to scan through the image. The scanning was carried out on the rectangular region whose top-left-corner was on row 98 and column 22 and bottom-right-corner was on row 448 and column 509. The reason was that the likelihood of nodule outside the described rectangular region was zero (see Fig. 2-right). In each iteration of the scanning process, the sliding window was shifted to the right by one pixel. Once the window reached the last pixel of the row, it was then moved to the beginning pixel of the next row (see Fig. 5). The region covered by the sliding window was extracted, and then passed on to the classifier for detection. The problem with this approach was multiple detection regions around the nodule or multiple false detection regions around the non-nodule (see Figure 4). To address this problem, we formulated the following algorithm:

i. Two 512×512 detection and false-detection masks were created with pixel values of all '0's. If the classifier returned 'detection', the associated 30×30 region of the detection mask was filled with '1's (see Fig. 6-left).
ii. For each expert-identified nodule, we formed a window whose size and location were determined by the nodule's size and location. We set the window's pixel values to '1's. This window was placed on the detection mask according to the expert-identified location of the nodule. Next, the degree-of-match between the window pixels and the detection mask pixels under the window was calculated. If the degree-of-match was greater than a threshold, the expert-identified nodule was ticked as detected. Using this approach, if another detection region occurs close to an existing detection, it would not be considered as a new nodule. This is because for the second detection, the same detected nodule will be ticked again as detected. The threshold was set by trial and error to reduce the occurrence of multiple detection of the same nodule.
iii. On the other hand, if the degree-of-match was below the threshold, then we do as follows. We formed a window whose size and location were determined by the size and location of the sliding window. We set the window's pixel values to '1's. This window was placed on the false-detection mask according to the actual location of the sliding window. Next, the degree-of-match between the window pixels and the false-detection mask pixels under the window was calculated. If the degree-of-match was less than a threshold, the region under the sliding window was considered as a false-detection and filled with '1's (see Fig. 6-right). Otherwise, the region under the sliding window is ignored. This is because the region had been previously marked as false-detection. Therefore, if two neighbouring sliding window regions happen to be false-detection, only one false-detection is recorded. The false-detection regions are stored into files for further use. This

process was repeated until the sliding window covered the entire lung image under examination.

- In addition, the abovementioned lung image scanning process is repeated for the sliding window size of 47×47, 64×64, and 82×82 to be able to detect nodules of size greater than 30×30. In these iterations, the region under the sliding window is extracted and resized to 30×30 for nodule and non-nodule detection. Table I shows a summary of the test results for the nodule-detector.

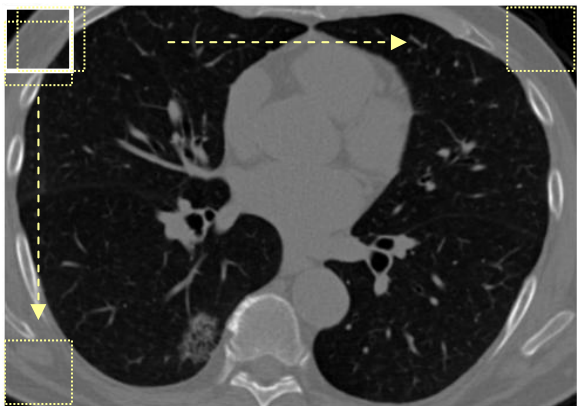

Fig. 5. Window sliding approach that is used to scan lung images.

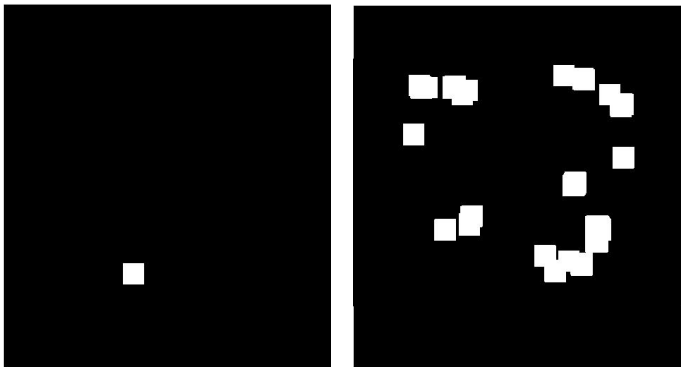

Fig. 6. Image 17710's (left) detection mask showing its first true-detection, and (right) false-detection mask showing multiple false-detections.

B. False-Positive-Reducer

The first random forest classifier was able to detect all expert-identified nodules within the 20 scanned images. However, it produced a large number of false-detections as well. In order to reduce the number of false-detections, we developed a second random forest classifier as follows.

- Similarly, a training set consisting of two groups of images was formed: false-detection and true-detection. The false-detection group consisted of 400 out of 611 false-detection patterns from the 20 images (see Fig. 6).
- The true-detection group contained the same 1203 nodule patterns used in the training of the first random forest classifier.
- The second random forest classifier was applied to the same 20 sets of images used in the testing of the first random forest classifiers, and whilst all nodules were detected, no false detection was also recorded.

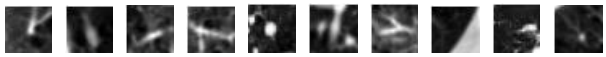

Fig. 7. Sample of 30×30 false detection-patterns collected from nodule-detector testing.

TABLE I.
SUMMARY OF THE TESTS RESULTS FOR NODULE-DETECTOR.

	Image Name	Number of expert-identified nodule	True Positive	False Positive
NODULE	6078	8	8	13
	6082	7	7	13
	6090	6	6	24
	6094	8	8	37
	6098	8	8	61
	6102	8	8	32
	17036	7	7	47
	17630	10	10	71
	17710	7	7	99
	22250	9	9	3
NONNODULE	4450	0	0	15
	14918	0	0	9
	24308	0	0	32
	24356	0	0	11
	29566	0	0	23
	29710	0	0	39
	30982	0	0	47
	31146	0	0	21
	31246	0	0	6
	31290	0	0	8

VI. DISCUSSIONS

The proposed lung detection system was formed from the series connection of both the nodule-detector and the false-positive-reducer classifiers. To test the system, a collection of test images were constructed. It included 15 lung images randomly selected from the 411 images containing nodules and 5 images randomly selected from the images without any nodules. These images were not used in the training phase of both the nodule-detector and the false-positive-reducer classifiers. The system could identify all expert-identified nodules present within the 20 tested images. In addition, it produced only 28 false detections for the 20 images. Table II shows a summary of the test results for the proposed system consisting of both classifiers.

The results demonstrate that the proposed random forest-based system performs well not only in detection of nodule patterns of different sizes, but also in producing a very few false detections. The nodule detection rate of 100% and false detection rate of 1.4 per image outperform most reported exiting lung nodule detection systems.

The implemented systems were trained and tested on an Intel Xeon CPU 5130 @2.00 GHz on-board of a Dell Desktop. The codes were written and executed in Matlab. The image scanning processes were quite slow taking several hours for each image. We will optimise our codes for speed,

and also compile the codes to form executable binaries. These should improve the execution time of our system.

TABLE II.
SUMMARY OF THE TEST RESULTS FOR THE PROPOSED SYSTEM CONSISTING OF BOTH CLASSIFIERS IN SERIES.

	Image Name	Number of expert-identified nodules	True Positive	False Positive
NODULE	1391	4	4	0
	1411	2	2	1
	4282	6	6	1
	5572	3	3	5
	6680	3	3	1
	7574	4	4	4
	8228	4	4	2
	12851	2	2	1
	13263	4	4	0
	14470	1	1	2
	14574	3	3	5
	19178	1	1	1
	31844	3	3	0
	33500	2	2	2
NON NODULE	8700	0	0	0
	9197	0	0	2
	12832	0	0	1
	20374	0	0	0
	33680	0	0	0

VII. CONCLUSION

An automated lung nodule identification system consisting of two random forest-based classifiers connected in a series fashion was developed. Training sets consisting of nodule, non-nodule, and false-detection patterns are constructed. A collection of test images are also built which includes 15 lung images randomly selected from the 411 images containing nodules, plus 5 images randomly selected from the images with no nodules. These images were not used in the training phase of the classifiers. The developed system could identify all expert-identified nodules present within the 20 tested images. In addition, it produced only 28 false detections for the 20 images. The results demonstrate that the proposed random forest-based system performs well not only in detection of nodule patterns of different sizes, but also in producing a very few false detections. The nodule detection rate of 100% and false detection of 1.4 per image outperform most reported exiting lung nodule detection systems. The developed random forest-based system proved to perform well for the lung nodule identification problem considered in this work.

ACKNOWLEDGEMENT

The support of the Victorian Partnership for Advanced Computing (VPAC) under an e-Research Program Grants Scheme is gratefully acknowledged.

REFERENCES

[1] J. H. Austin, N. L. Mueller, and P. J. Friedman, "Glossary of terms for CT of the lungs: recommendations of the nomenclature," *Committee of the Fleischner Society. Radiology*, vol. 331, pp. 200:327, 1996.

[2] S. G. Armato III, G. McLennan, M. F. McNitt-Gray, C. R. Meyer, D. Yankelevitz, and e. al., "Lung Image Database Consortium Developing a Resource for the Medical Imaging Research Community," *Radiology*, vol. 232, pp. 739-748, 2004.

[3] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.

[4] M. A. J. Klik, E. M. v. Rikxoort, J. F. Peters, H. A. Gietema, M. Prokop, and B. v. Ginneken, "Improved Classification of Pulmonary Nodules by Automated Detection of Benign Subpleural Lymph Nodes." in *ISBI*, 2006.

[5] S. C. Clifford, V. Saravanan, and D. M. R. Vimala, "Lung Nodule Diagnosis From CT Images Using Fuzzy Logic." in *International Conference on Computational Intelligence and Multimedia Applications*, 2007.

[6] L. Zhao, L. Boroczky, and K. P. Lee, "False positive reduction for lung nodule CAD using support vector machines and genetic algorithms," *International Congress Series* vol. 1281, pp. 1109-1114, 2005.

[7] I. Gori, R. Bellotti, P. Cerello, C. Cheran, G. D. Nunzio, M. E. Fantacci, P. Kasae, L. Masala, A. P. Martinez, and A. Retico, "Lung Nodule Detection in Screening Computed Tomography." in *IEEE Nuclear Science Symposium Conference Record*: IEEE, 2006.

[8] T. Jia, D.-Z. Zhao, J.-Z. Yang, and X. Wang, "Automated Detection of Pulmonary Nodules in HRCT Images." in *1st International Conference on Bioinformatics and Biomedical Engineering, 2007 (ICBBE)*: IEEE, 2007, pp. 833-836.

[9] H. Takizawa, S. Yamamoto, and T. Shiina, "Accuracy Improvement of Pulmonary Nodule Detection Based on Spatial Statistical Analysis of Thoracic CT Scans," *IEICE TRANS. INF. & SYST.*, vol. 90-D, pp. 1168-1174, 2007.

[10] R. A. Ochs, J. G. Goldin, A. Fereidoun, H. J. Kim, K. Brown, P. Batra, D. Roback, M. F. McNitt-Graya, and M. S. Brown, "Automated classification of lung bronchovascular anatomy in CT using AdaBoost," *Medical Image Analysis* vol. 11, pp. 315-324, 2007.

[11] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News*, vol. 2, pp. 18-20, 2002.

[12] Y. Freund and R. E. Schapire, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 14, pp. 771--780, 1999.

[13] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, 1996.

[14] "Lung Imaging Database Consortium (LIDC)." [Online]. Available: http://imaging.cancer.gov/programsandresources/InformationSystems/LIDC.