This is the published version:


Available from Deakin Research Online:

http://hdl.handle.net/10536/DRO/DU:30018296

Reproduced with the kind permission of the copyright owner.

Copyright: 2008, IEEE.
A Random Forest for Lung Nodule Identification

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 method is presented for identification of lung nodules. It includes three stages: image acquisition, background removal, and nodule detection. The first stage improves image quality. The second stage extracts long lobe regions. The third stage detects lung nodules. The method is based on the random forest learner. Training set contains nodule, non-nodule, and false-positive patterns. Test set contains randomly selected images. The developed method is compared against the support vector machine. True-positives of 100% and 85.9%, and false-positives of 1.27 and 1.33 per image were achieved by the developed method and the support vector machine, respectively.

I. INTRODUCTION

Lung cancer is formed from uncontrolled growths in lung cells. The uncontrolled growths found in lung are known as lung nodules which can be malignant or benign. Lung nodules refer to a range of focal abnormalities considered as small, round, opacity, roughly spherical, and restricted on abnormal tissue [1, 2].

Lung imaging techniques such as multi-detector row CT scanning is preferable for detecting lung nodule. Currently, it generates more than 300 image slices per subject from a single breath hold. With the growing concern of lung cancer, low-dose helical computed tomography (CT) protocol is used due to the high spatial and contrast resolution of the anatomical structures. This enables the expert radiologist to visualise the lung nodules.

Recent studies indicates an inter-reader variability in the detection of nodules amongst expert radiologists [3]. An automated diagnostic system can thus provide initial nodule detection which may help expert radiologists in their decision making.

Recently, a number of approaches have been formulated for detection of nodules in 2D CT lung images [4-9]. Some existing methods use classification approach for detection of lung nodules. One of the current trends is utilisation of ensemble learners which employs a large amount of weak classifiers with boosting. In [10], a method is described 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.

This paper employs the concept of pattern classification to form an automated lung nodule detection method. Two pattern classes are formed namely nodule and non-nodule.

A random forest [11] learner 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. 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.

Since an individual tree is unpruned, the terminal nodes can contain only a small number of observations. The training data is 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]. A summary of the random forest algorithm for classification is given below [12]:

• 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 among those variables.
• Predict new data by aggregating the predictions of the K trees, i.e., majority votes, 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. Using a random selection of features to split each node yields error rates that compare to AdaBoost [13]. An estimate of the error rate can be obtained by the following [12]:

• 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% [14] of the times. Calculate the error rate, and call it the “out-of-bag” estimate of error rate.
We employ the random forest algorithm to form the proposed method for detection of lung nodules in 2D CT images.

II. PROPOSED LUNG NODULE DETECTION METHOD

The proposed method consists of the following components (see Fig. 1): (i) image acquisition, (ii) background removal, and (iii) nodule detection.

A. Image Acquisition

Image acquisition is the process of acquiring 2D medical lung nodule image datasets. We used a lung nodule dataset from the Lung Imaging Database Consortium (LIDC), National Imaging Archive [15] by National Cancer Institute. A total of 32 scans containing 5721 images are selected. 411 images that contained expert-radiologists identified nodules. Fig. 2 shows sample lung images from this dataset.

B. Background Removal

Background removal is the process of identifying the lung lobe region from the background region. Figure 3 shows the process flow of the background removal algorithm. The image is converted to its binary representation. Region labelling algorithm is then applied on the binary image. An image mask is created to highlight the lung region of the image. After executing the initial region of interest (ROI) algorithm, the lung tissue region is extracted. As shown in Fig. 4, there exists non-lung lobe region. After applying final ROI algorithm, the non-lung lobe region has been discarded from the lung image.

C. Nodule Detection

Nodule detection is the process of identifying the presence and location of the lung nodules. This stage includes two sections: data preparation and nodule detection. For data preparation for the classifier (see Fig. 5), the first step is to obtain the dimension for each region in the background removed image. Sliding windows of 30 X 30, 56 X 56, and 82 X 82 were formed to scan through the image. The scanning was initiated on the top-left-corner of the lung lobe region. In each of the iteration in 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 of the next row (see Fig. 6). The region covered by the sliding window was extracted and the non-lung lobe regions were eliminated. The lung lobe regions were then passed on to the classifier for detection.

Four steps were used to detect the nodules (see Fig. 7). The first step was to develop the XML converter program which contains expert identified nodule and non-nodule information in the XML file for each scan. The second step was to extract out the nodule and non-nodule regions from the corresponding lung images. For nodule patterns that could fit within a 30 X 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 X 30 region, we extracted the entire nodule pattern first, and then resized the extracted pattern into a 30 X 30 region. There were a total of 817 such nodule patterns. Overall, we created 1203 30 X 30 nodule files (see Fig. 8). For the non-nodule patterns, there were 1156 expert identified non-nodule patterns. Also, we used another 1770 randomly captured regions of sizes 30 X 30, 56 X 56, and 82 X 82 and then resize to 30 X 30 which did not contain any nodule patterns from within the lung lobe areas.
A total of 2926 30×30 non-nodule files were created (see Fig. 8).

The third step was to train the random forest classifier with the nodule and non-nodule patterns. The two random forest parameters, no-of-trees-grown and no-of-variables-at-each-split were varied, 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 80 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 the tried forests.

The last step was to test the random forest classifier and record the false positive rates. A total of 40 images from the LIDC database were used to test the random forest classifier and 100% true detection and 1203 false detection were recorded.
The random forest is then retrained, by adding the false detection images into the non-nodule regions. The two random forest parameters, no-of-trees-grown and no-of-variables-at-each-split were varied as described above and the random forest with the 79 trees and 18 variables produced the lowest classification error.

The support vector machine classifier with the RBF kernel was trained with the nodule and non-nodule regions including the false detection regions. The support vector machine-based classifier’s kernel parameter was varied. The kernel parameter of 0.01 produced the lowest classification error.

The random forest classifier and support vector machine classifier were tested with 15 images chosen from the LIDC database which was not previously used. 10 of the images contained nodule and the remainder 5 images did not contain any nodules. The results are illustrated in Table I.

<table>
<thead>
<tr>
<th>Image Name (dcm)</th>
<th>No. of expert-identified nodule</th>
<th>True Positive</th>
<th>False Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>12871</td>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>12904</td>
<td>7</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>12923</td>
<td>6</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>13280</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1399</td>
<td>8</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>1408</td>
<td>6</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>17634</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>17678</td>
<td>6</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>19996</td>
<td>0</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>29310</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>29530</td>
<td>6</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>33328</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>5762</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6106</td>
<td>6</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>6608</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

TABLE I. SUMMARY OF THE DETECTION RESULTS

III. DISCUSSIONS

In the automated nodule identification method, there were three stages involved in this proposed system namely (i) image acquisition, (ii) background removal, and (iii) nodule detection. LIDC database was utilised due to the reliable and extensive expert identified nodule information. Background removal was implemented in the system to reduce the computational time of nodule detection. It eliminated the non-lung lobe region from the scanning process in the nodule detection phase. Random forest based classifier was developed to detect the lung nodule presence in the scan.

The system tested on 10 images randomly selected from the 411 images containing nodules and 5 images randomly selected from the images without any nodules. Also, these images were not presented in the first testing phase. The support vector machine classifier based system was trained using the same two pattern classes as the abovementioned classifier.

Both the classifier were trained and tested on an Intel Xeon CPU 5130 @2.00 GHz on-board of a Dell Desktop.

The codes for training and testing both the classifier based system were written and executed in Matlab.

Table I indicates that the random forest classifier based method performs better than support vector machine classifier based system. The random forest classifier based system recorded 100% true positive and 1.27 false positive per scan. On the other hand, support vector machine classifier based system only recorded 85.9% true positive and 1.33 false positive per scan.

IV. CONCLUSIONS

The developed automatic detection of lung nodule system consists of three stages that are image acquisition, background removal, and nodule detection. 5721 images selected from the LIDC lung databases. In the training process, 1203 nodules patterns and 4129 non-nodule patterns including the false detection patterns from the first training phase were included. The system was tested on 10 images containing nodules and 5 images containing no nodules which were randomly selected from the 5721 images. The images were not involved in the training process of the system. The proposed random forest based classifier performs well to detect all the nodules in the images and recorded a low false detection rate. It results 100% sensitivity and 1.27 FP/scan.

ACKNOWLEDGMENT

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

REFERENCES


