Medical Image Analysis 13 (2009) 757–770

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification

K. Murphy [a,*], B. van Ginneken [a], A.M.R. Schilham [a], B.J. de Hoop [b], H.A. Gietema [b], M. Prokop [b]

[a] Image Sciences Institute, University Medical Center, Utrecht, The Netherlands
[b] Department of Radiology, University Medical Center, Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 1 July 2008
Received in revised form 26 June 2009
Accepted 2 July 2009
Available online 30 July 2009

Keywords:
Nodule detection
Lung nodule
CT
Automatic detection
kNN
Shape index

ABSTRACT

A scheme for the automatic detection of nodules in thoracic computed tomography scans is presented and extensively evaluated. The algorithm uses the local image features of shape index and curvedness in order to detect candidate structures in the lung volume and applies two successive k-nearest-neighbour classifiers in the reduction of false-positives.

The nodule detection system is trained and tested on three databases extracted from a large-scale experimental screening study. The databases are constructed in order to evaluate the algorithm on both randomly chosen screening data as well as data containing higher proportions of nodules requiring follow-up. The system results are extensively evaluated including performance measurements on specific nodule types and sizes within the databases and on lesions which later proved to be malignant. In a random selection of 813 scans from the screening study a sensitivity of 80% with an average 4.2 false-positives per scan is achieved. The detection results presented are a realistic measure of a CAD system performance in a low-dose screening study which includes a diverse array of nodules of many varying sizes, types and textures.

In 2008, the American Cancer Society estimates that 29% of all cancer deaths in the United States will be due to lung cancer. This makes lung cancer the leading cause of cancer deaths in the United States, killing more people than the next three most deadly cancers combined (colon cancer, breast cancer and pancreatic cancer) (American Cancer Society, 2008). One of the key issues in addressing this statistic is the fact that lung cancer is rarely diagnosed in the early, more treatable stage of the disease which can be relatively asymptomatic (American Cancer Society, 2008). Routine screening programmes using low-dose computed tomography (CT) scanning are currently being considered as a possible means to detect the first signs of lung cancer in apparently healthy but high-risk subjects (Gohagan et al., 2005; New York Early Lung Cancer Action Project Investigators, 2007; Novello et al., 2005; Swensen et al., 2005; Xu et al., 2006).

Early stage lung cancer manifests itself in the form of pulmonary nodules which are visible on thoracic CT scans. Pulmonary nodules are abnormal structures which are generally approximately spherical in shape, with those which are attached to the pleural surface (pleural nodules) being roughly hemispherical. These lesions can be detected on CT images even if they are only a few millimetres in diameter (Henschke et al., 2001). They appear as small bright spots surrounded by the darker lung parenchyma, with grey-values very similar to those of blood vessels in the lungs. Although the vast majority of pulmonary nodules are not malignant and do not require treatment, detection of such nodules is the first crucial step in identifying early stage lung cancer. Once the nodule has been detected, monitoring of its size, density, edge-smoothness and growth rate provide information about what treatment, if any, is appropriate.

Studies have shown that radiologists frequently fail to detect all visible nodules in CT scans (Gurney, 1996; Kakinuma et al., 1999; Swensen et al., 2002; White et al., 1996). The examination of a chest CT scan is a time-consuming and error-prone task while the radiologist is vulnerable to human-error and fatigue. These issues have motivated the development of computer-aided detection (CAD) solutions for lung nodule annotation on CT scans. Thus far it is generally accepted that the CAD solution is intended as an assistant to the radiologist which can quickly identify suspicious structures for his attention (Rubin et al., 2005).

The area of computer-aided nodule detection in CT is an active field of research with a wide range of approaches having been published in the literature of the last number of years (Arimura et al., 2004; Bae et al., 2005; Bellotti et al., 2007; Brown et al., 2003; Chang et al., 2004; Dehmeshki et al., 2007; Enquobahrie et al., 2007; Ge et al., 2005; Li et al., 2008; Matsumoto et al., 2006;

[*] Corresponding author. Address: Image Sciences Institute, University Medical Center, Heidelberglaan 100, Room Q0S.459, 3584 CX Utrecht, The Netherlands.
E-mail address: keelin@isi.uu.nl (K. Murphy).

1361-8415/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.media.2009.07.001

McCulloch et al., 2004; Mendonça et al., 2007; Paik et al., 2004; Retico et al., 2008; Wiemker et al., 2005; Ye et al., 2007; Zhang et al., 2005; Zhao et al., 2003, 2006). It is virtually impossible to compare these systems in any meaningful way due to the enormous variation in dataset size, scan properties, data selection criteria and reference standard generation. In the selection of literature referenced above the authors tested variously on simulated nodules, on nodules above or below particular size limits only, or on solid or non-solid nodules only. The number of scans used for testing varied between 5 and 500 with a median number of 29.5 and many of the studies included multiple scans from individual patients meaning a reduction in the diversity of available nodules. Furthermore the quality of the data used is highly variant, particularly with respect to slice thickness and radiation dose.

Although many studies have reported promising results based on their own data and experiments, a fully comprehensive system which is evaluated and proven on a large unfiltered dataset, truly representative of modern CT screening data has not been reported. In this work an algorithm for automatic nodule detection is presented and its performance is thoroughly evaluated on three large databases of trial screening data, the first with 1588 scans (2894 nodules), and the second and third with 1158 scans each (3451 and 1528 nodules respectively). The main components of the algorithm have been presented previously (Murphy et al., 2007) and showed promising results on a test set of 142 scans. Due to the increased size of the datasets used in this work the system performance can be reliably determined on particular nodule types and sizes allowing us to pinpoint its strong and weak points. In the long term the system can thus be refined or assisted by additional modular components to improve performance on specific nodule types. The system performance we measure is based on tests over thousands of individual nodules of differing shapes, textures, sizes and locations and found in both healthy subjects and subjects exhibiting pathology. The type of in-depth analysis and large-scale evaluation that is employed in this study is a crucial step for any algorithm under consideration for use in a clinical situation.

1. Materials

1.1. The Nelson trial data

The Nelson Trial is an experimental lung cancer screening programme currently taking place in the Netherlands and Belgium. The programme involves ongoing CT screening of members of the general population who are considered high-risk for the development of lung cancer due to being (former) heavy smokers. As a participant in this programme the University Medical Center Utrecht has access to a large database of low-dose chest CT scans. In all cases considered in this study, CT scanning was performed on a 16 detector-row scanner (Mx8000 IDT or Brilliance 16P, Philips Medical Systems, Cleveland, OH, USA). The scans were realized within 12 s, in spiral mode with 16 × 0.75 mm collimation, without contrast-injection and in inspiration. Exposure settings were low-dose: 30 mAs at 120 kVp for subjects weighing below 80 kg or 30 mAs at 140 kVp for those weighing over 80 kg. A soft reconstruction filter (Philips ‘B’) was used. All scans have a per-slice size resolution of 512 × 512, with the number of slices varying between 306 and 860 (on average 459 slices) for the data used in this work. Slice thickness is 1 mm with slice-spacing of 0.7 mm. Pixel spacing in the X and Y directions varies from 0.6 mm to 0.9 mm, as the field of view was set for every scan to include the outer rib margins at the widest dimension of the thorax.

For the purposes of the Nelson Trial observed pulmonary nodules are identified in the University Medical Center Utrecht (on-site) by an experienced observer and checked by a second observer in an independent medical facility (core lab). Observers scored any nodular structure not explained by atelectases, scars or infection. During the first phase of the trial (more than 1000 scans) extensive consultation between on-site readers and core lab ensured that this definition induced as little ambiguity as possible. Observers in the Nelson Trial are not required to mark findings whose automatically calculated volume (via Siemens LungCARE workstation software) is below 15 mm^3, (diameter d of corresponding volume sphere ≈3 mm) although they may do so if they choose. Nodules with volumes larger than 15 mm^3 must be documented and appropriate follow-up procedures are scheduled for patients with nodules larger than 50 mm^3 ($d \approx 4.5$ mm). Since there is no lower bound on the volume of structures to be reported the database includes extremely small or trivial detections as well as larger and more significant structures. Further details relating to the Nelson Trial protocols can be found in (Xu et al., 2006).

The nodule-markings (hereafter referred to as ‘annotations’) used as ground-truth in this work come primarily from a single on-site observer (Observer1). For 813 scans the findings of the second unblinded observer (Observer2) are available in addition. These scans are therefore used as test data in Database A (see Section 1.1.1), since the use of data with multiple available readings is always preferable. Observer1 is a CT technician with special training in evaluating and reporting cancer screening CTs (>3000 scans in 2 years) while Observer2 is a radiologist with 6 years of experience.

For the purpose of evaluating our nodule detection algorithm we have created three databases from the Nelson Trial data as described in the remainder of this section. Where a patient had more than one scan eligible for inclusion in a database the earliest scan was chosen and the remainder excluded. The system reported a detection as a true-positive if it was within seven voxels (≈1.4 mm) of an annotation. A summary of the number of scans and annotations in each database is provided in Table 1.

In each database we exclude a number of scans where our automatic lung segmentation procedure failed, since this paper focuses solely on the nodule detection aspect of a computer-aided diagnosis (CAD) system. Due to the amount of data involved it was not possible to check the lung segmentation and nodule detection results manually for every scan. Therefore only those scans where the segmented regions had an unusual volume or left/right ratio were checked for lung segmentation errors. A number of scans with minor lung segmentation errors are erroneously retained and may affect the nodule detection process. In particular several examples were observed where a large pleural nodule was excluded from the lung volume resulting in a false-negative detection.

1.1.1. Database A: all nodules from random scans

This database consists of an initial selection of 1588 non-consecutive scans made in 2005 as part of the Nelson Trial that year. Automatic lung segmentation was unsuccessful for 53 of these scans thus these were excluded leaving 1535 scans suitable for nodule detection. Included are 813 scans for which annotations

Table 1
Statistics on the number of scans in the three databases.

	A	B	C
#Scans before checks	1588	1158	1158
#Scans with lung segmentation failures	53	37	37
#Scans after removing failures	1535	1121	1121
#Scans in training set	722	580	580
#Scans in test set	813	541	541
#Nodules in final training set	1369	1763	760
#Nodules in final test set	1525	1688	768

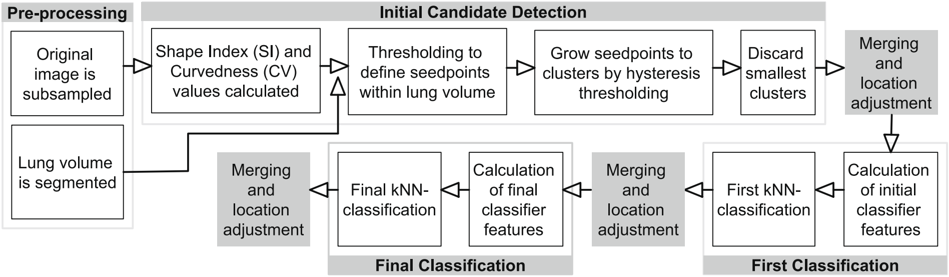

Pre-processing
Original image is subsampled
Lung volume is segmented
Initial Candidate Detection
Shape Index (SI) and Curvedness (CV) values calculated
Thresholding to define seedpoints within lung volume
Grow seedpoints to clusters by hysteresis thresholding
Discard smallest clusters
Merging and location adjustment
Calculation of initial classifier features
First kNN-classification
First Classification
Merging and location adjustment
Calculation of final classifier features
Final kNN-classification
Final Classification
Merging and location adjustment

Fig. 1. Overview of nodule detection scheme.

from both Observer1 and Observer2 are available, thus these were chosen as the test data. The remaining 722 scans which are used as training data are selected randomly from the Nelson database for 2005. The training data contains a total of 1369 nodules annotated by Observer1 only. The test data contains 1525 nodules including all observations from either Observer1 or Observer2.

The results of Observer1 and Observer2 were similar enough to make it acceptable that the training data has ground-truth from Observer1 only. In fact Observer2 annotated 1518 nodules in the test set, while Observer1 annotated 1525. Counting only nodules annotated by both radiologists in the same location, we find that 1518 nodules were common to both.

1.1.2. Database B: all nodules from suspect scans

For this database we selected all scans which contained at least one nodule with volume above 50 mm^3 or with edge-smoothness described as ‘‘Lobulated” or ‘‘Spiculated” or with category given as ‘‘Non-solid” or ‘‘Partially-solid”. These are nodule types which it is crucial to detect successfully, since they require follow-up or have an appearance more suggestive of malignancy than the more common smooth solid nodule (Jeong et al., 2007). The database contained a total of 1158 scans selected from the Nelson data available from the period 2004–2007. (Of these, 343 scans were also contained in database A). After removal of scans where the lungs could not be automatically segmented 1121 scans remained.

Within database B there are 34 patients who have now had the malignancy of a lesion proved conclusively by means of a biopsy. To ensure the system received some training on any special aspects of these lesions 14 of these (chosen at random) were included in the training data while the remaining 20 patients were included in the test set. The remaining data was divided at random between the training and test sets such that the final training set contained 580 scans while the test set contained 541.

All annotations in this database are provided by Observer1 since the data for Observer2 was not always available. There were a total of 1763 annotations in the training data and 1688 in the test set.

1.1.3. Database C: large suspect nodules

This database is a subset of Database B described above whereby all scans are examined but nodules with volume below 50 mm^3 are excluded for the purposes of training and testing our algorithm. In this way we measure the performance of our algorithm specifically in the detection of those nodules which require follow up and examine whether performance is improved by training only with such nodules. The training set in this case contained 760 nodules while the test set contained 768. As with Database B all annotations used for ground truth are made by Observer1.

2. Methods

In this section details of the nodule detection scheme will be provided. Fig. 1 provides an overview of the procedures which are described in detail in Sections 2.1–2.3.3. Note that the candidate detection procedure involves the use of a number of threshold values which were empirically determined during the system development on a small fully independent set of test data.

2.1. Pre-processing

Before beginning with nodule detection some initial processing is carried out on the original image data as described below.

2.1.1. Sub-sampling of image data

The first step is to down-sample the data to improve the speed of the algorithm. Use of the full-size images was extremely computationally expensive and gave little or no improvement to the results. The down-sampling is by means of block-averaging such that the matrix size of 512 × 512 in the original Nelson Trial images is reduced to 256 × 256, with the number of slices reduced to form isotropically sampled data. Linear interpolation is used to determine grey-values between voxel locations. The number of slices in the down-sampled scans varied between 149 and 428 with an average of 223 slices per scan.

2.1.2. Segmentation of lung volume

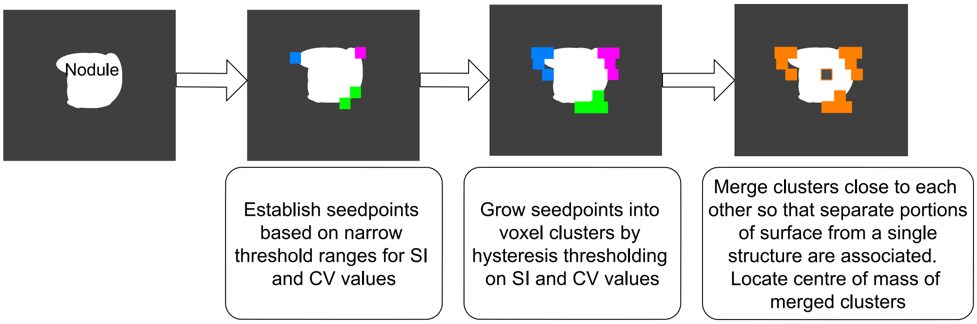
The second pre-processing step involves the segmentation of the lung volume from the surrounding tissues in the sub-sampled image. The mask obtained from this segmentation is used to ensure that nodule detection is performed within the lung volume only. This process has the two-fold advantage of reducing computation time and preventing the possible detection of false positive structures in regions of the image outside the lungs. Lung segmentation was carried out using an algorithm by Sluimer et al. (2005) based on that of Hu et al. (2001).

2.2. Initial candidate detection

The process of initial candidate detection is described in detail in the remainder of this section and depicted graphically in Fig. 2.

2.2.1. Shape index and curvedness

Our scheme for nodule detection utilises the shape index (SI) and curvedness (CV) features (Koenderink, 1990) to detect initial nodule candidates. These are 3D local image features which are calculated per voxel based on local attenuation values and give insight into the surface topology at every point in the image volume. The SI and CV are derived from the principal curvatures k_1 and k_2 (Koenderink, 1990) but have the advantage of decoupling topological shape and magnitude of curvature. In the context of nodule detection we are generally interested in voxels which show a roughly spherical shape and where the size of the supposed sphere is within a range of reasonable values. The shape index and



Establish seedpoints based on narrow threshold ranges for SI and CV values

Grow seedpoints into voxel clusters by hysteresis thresholding on SI and CV values

Merge clusters close to each other so that separate portions of surface from a single structure are associated. Locate centre of mass of merged clusters

Fig. 2. The scheme for initial nodule candidate detection.

curvedness values at a voxel are calculated using the principal curvatures k_1 and k_2 at that point as follows[1]

$$\mathrm{SI} = \frac{2}{\pi}\arctan\left(\frac{k_1 + k_2}{k_1 - k_2}\right) \tag{1}$$

$$\mathrm{CV} = \sqrt{k_1^2 + k_2^2} \tag{2}$$

The principal curvatures k_1 and k_2 are calculated for all voxels within the lung volume using first and second order derivatives of the image blurred with a Gaussian filter with scale $\sigma = 1$ voxel. This value for σ was empirically determined to reduce noise without removing important structural detail.

2.2.2. Seed point detection

Once the SI and CV values for the image are known, a set of seed points is established by the thresholding of these values according to empirically decided limits shown in Table 2. Voxels which have both SI and CV within the thresholds are selected as seeds. These seed points represent voxels which may lie on a nodule surface and whose locality deserves further exploration. For locations within five voxels of the pleural surface a slightly lower threshold for SI is used in order to increase the number of seed detections near the lung boundaries. This is necessary since pleural nodules do not present as much surface area for examination, and in addition the SI and CV values in their region may be affected by the topology of the adjacent pleural surface.

2.2.3. Cluster formation

The seedpoints are now expanded to form clusters of voxels of interest. The expansion is based on hysteresis thresholding (Canny, 1986) using broader thresholds as shown in Table 3. The final cluster therefore contains only voxels whose SI and CV values fall within the broader threshold range and which can be connected (using six-connectivity) to a seedpoint by a chain of other such voxels.

It should be noted that for a perfectly spherical structure, the voxels in the final cluster lie in the region of the blurred surface of the sphere. The centre of mass of the cluster is taken to be the point of interest at this stage.

A cluster whose original seedpoint lies within five voxels of the pleural surface is considered to be a pleural candidate. At this stage clusters whose volume is below a pre-determined threshold t_{vol} are discarded as their inclusion in the remaining processing steps was found to be extremely costly and more likely to introduce false-positives. During the development phase t_{vol} was empirically set at four voxels for candidates in the pleural region and at 15 voxels for the remaining candidates.

2.2.4. Cluster merging

At this stage a large number of clusters have been detected. Each one represents a region of surface in the image and it is reasonable to suppose that a true structure such as a nodule may have more than one cluster representing it. Except in the case where the nodule is extremely large and unusually shaped these clusters will lie in close proximity to each other. Clusters with locations within three voxels of each other are therefore recursively merged until no more merges can be performed, and the procedure is then repeated for clusters within seven voxels of each other. Fig. 3 shows examples of candidate structures being merged. It will be seen that this merging procedure is repeated at several later points in the processing pipeline (see Fig. 1). Although in later stages only a very small number of structures (if any) will require merging, the procedure serves two main purposes. Firstly, it ensures that a single nodule is represented by a single detection rather than by two detections alongside each other. Secondly, where the candidates being merged are false-positives, it frequently results in an oddly shaped structure which can be more easily eliminated in later classification steps.

2.2.5. Candidate location adjustments

At this point the candidate locations are checked and adjusted to ensure that they sit at the brightest spot locally. This is important since the nodule location defined by the centre of mass of the voxel cluster is not always accurate in locating the nodule centre, particularly for pleural nodules or where the voxel cluster found is concentrated on one side of the nodule surface. The location adjustment procedure examines all local points with a maximum distance of three voxels from the original candidate location. At each local point the average grey-value over the point and its six connected neighbours is calculated. The location with the highest average local grey-value is chosen as the new candidate location. Local averaging is necessary to exclude the possibility that a bright voxel representing data noise is selected.

Table 2
Initial seed thresholds.

Value	Upper threshold	Lower threshold
SI	1	0.8 (near pleural surface) 0.9 (elsewhere)
CV	1	0.3

Table 3
Hysteresis thresholds.

Value	Upper threshold	Lower threshold
SI	1	0.7 (near pleural surface)
CV	1.3	0.2

[1] The definition of CV used here excludes the scaling constant of $\sqrt{2}$ used by Koenderink (1990) which serves only to enforce unit curvedness on the unit sphere.

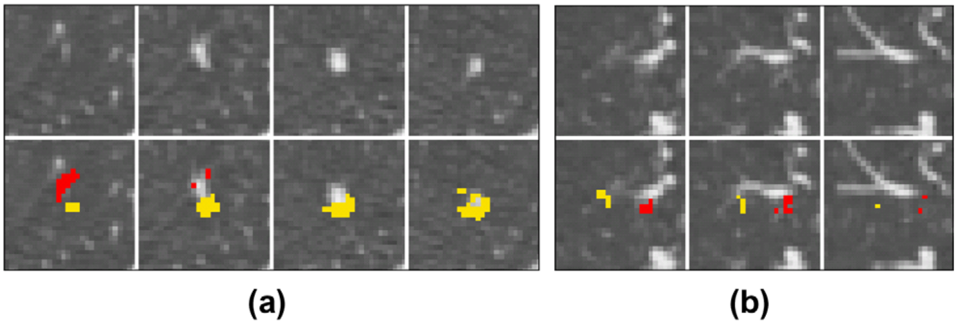


Fig. 3. Two examples of merging structures. Top rows show consecutive axial slices. Bottom rows show the same slices with structures to be merged in contrasting colours. (a) A true-positive structure is initially detected as two individual objects which will be merged. (b) Two false-positive detections to be merged into a single object. This object was subsequently rejected as a false-positive.

2.3. False positive reduction

The false-positive reduction consists of two consecutive classification steps using k-Nearest-Neighbour (kNN) classifiers (Cover and Hart, 1967). During pilot experiments other supervised classification methods (Duda et al., 2001) were tested however kNN was found to achieve the best results. The nature of the data does not lend itself well to linear classification and testing with a Support Vector Machine classifier obtained slightly worse results than those from the kNN classifier. In all cases *k* was set at the (odd) square-root of the number of samples in the training set. Experimentation with the value of *k* did not yield any improvement in experiments during the development phase.

Prior to classification, it is necessary to generate training data to train the kNN classifiers as well as possible. This procedure is described in Section 2.3.1 below and illustrated in Fig. 4.

For the initial classification a small number of features which are efficient to calculate are used in order to substantially reduce the number of candidates without too much overhead. The number of candidates remaining after the initial classification is then small enough to allow efficient calculation of a larger set of more complex features for the final classification step. In both classification steps feature selection was carried out by 'Sequential Forward Floating Selection' (SFFS) (Pudil et al., 1994) to establish the most discriminative subset of all calculated features. The SFFS procedure used leave-one-out training and testing on the training dataset only, with the area under the Receiver Operating Curve (ROC curve Metz, 1986) as the criterion to be optimized. The maximum number of features to be selected is set at 15 for the first classifier and 50 for the second.

Leave-one-out training and testing on the training dataset is also used to determine a threshold value for the posterior probability given by the soft classification from the first kNN classifier. This threshold is used to establish which items from the first classification will be sent as candidates to the second classification step. The threshold is chosen such that 90% of true nodules are correctly identified by the classifier in the training set.

Similarly a posterior probability threshold is determined via training and testing on the training dataset such that 90% of true nodules are retained after the final kNN-classification step. After nodule detection is completed these results are stored, and by reducing this threshold the system performance at lower sensitivity rates can be examined. In the remainder of this work this will be referred to as varying the "operating point" of the system.

The individual classification steps are described in Sections 2.3.2 and 2.3.3 below.

2.3.1. Training data generation

The key requirement in generating training data is that the circumstances under which the data is gathered should resemble as

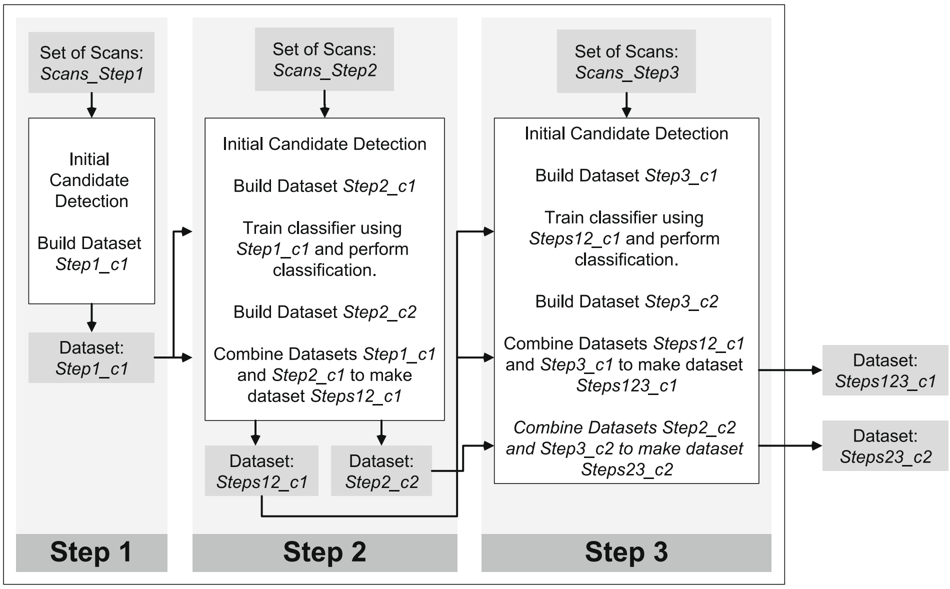

Set of Scans: Scans_Step1
Initial Candidate Detection
Build Dataset Step1_c1
Dataset: Step1_c1
Set of Scans: Scans_Step2
Initial Candidate Detection
Build Dataset Step2_c1
Train classifier using Step1_c1 and perform classification.
Build Dataset Step2_c2
Combine Datasets Step1_c1 and Step2_c1 to make dataset Steps12_c1
Dataset: Steps12_c1
Dataset: Step2_c2
Set of Scans: Scans_Step3
Initial Candidate Detection
Build Dataset Step3_c1
Train classifier using Steps12_c1 and perform classification.
Build Dataset Step3_c2
Combine Datasets Steps12_c1 and Step3_c1 to make dataset Steps123_c1
Combine Datasets Step2_c2 and Step3_c2 to make dataset Steps23_c2
Dataset: Steps123_c1
Dataset: Steps23_c2
Step 1
Step 2
Step 3

Fig. 4. Generation of training sets.

closely as possible those which will occur during the testing phase. For example, the training data for the second classification step should consist of candidates marked positive by a well-trained classifier in the initial kNN-classification. For this reason we use a three-step training set generation procedure, which is illustrated in Fig. 4, resulting in final training sets, Steps123_c1 for the first classifier and Steps23_c2 for the second. The description given here is generic–specific details for individual databases are provided in Section 3.

The scans available for training are divided randomly into three similarly sized groups, Scans_Step1, Scans_Step2 and Scans_Step3. In the first step we do initial candidate detection on the images in Scans_Step1 and use the generated candidates along with the ground-truth information to create a training set with all the features necessary for use in a first-classification step. Since we have no prior training data available no kNN classification is carried out on the Scans_Step1 data.

In step 2 we use the images in Scans_Step2, and as in step 1 we detect initial candidates and build a training set suitable for use in the first-classification. Next we make use of the training data that was created in step 1 to train a kNN-classifier and perform an initial classification to reduce the number of false-positives. The remaining candidates emerging from this first classification are now used to build a training dataset containing all the features used in the second-classification step.

The third step in the procedure uses the images from Scans_Step3 and repeats the procedure outlined in step2, except that in the first classification training data is a combination of that obtained in step 1 and in step 2. The reason for the inclusion of step 3 as a separate process (rather than simply combining step2 and step3 into a single step) is that the classifier trained with this combined data in step 3 gives even more accurate output than the first-step classifier in step 2. Its output is therefore closer to what we expect to occur in the testing situation and more valuable as training data.

Training sets produced in this manner will contain an extremely high number of false examples compared to true examples. For this reason we reduce the size of the false classes at the end of each step in the training procedure, such that the ratio of false objects to true objects in each dataset is roughly 3:1. This ratio was found to give optimal results during the development phase. The removal of false items is done by the data condensation method of Mitra et al. (2002) in order not to alter the distribution of the samples. This method produces a small representative subset from a larger dataset by the selection of points in a multi-scale fashion. The accuracy of representation by the subset is measured in terms of the error in density estimates of the original and reduced sets. The value of the parameter k determines the scales at which the data is viewed. In this work k is set initially at 15 and the set of false items is repeatedly condensed until further condensation would reduce it below the desired size ($3 \times$ number of true examples). The parameter k is then decremented and condensation is attempted again. This loop is repeated until $k \leqslant 2$ or the dataset reaches precisely the desired size.

2.3.2. Initial kNN classification

For the first kNN classifier a total of 18 features were calculated prior to feature selection. The entire list of calculated features is presented in Table 4 along with reference IDs which are used in this text. Details of which features were selected by SFFS in various experiments are presented in Table 6.

The features calculated at this stage relate to the geometric properties of the clusters of voxels found by the initial candidate detection step (a1–a9) as well as the grey-values in the region around the candidate locations (a10–a18). The shape of the voxel cluster gives an idea whether the structure in question is elongated (i.e. vessel-like) or roughly spherical, while its size can be a valuable feature in differentiating between true and false candidates. The grey-value features are examined over spherical kernels of voxels around the candidate location to try to eliminate structures which do not lie in a sufficiently bright region. Note that the descriptions in the table make reference to the ‘radius’, r, of the voxel cluster. This value is calculated by summing the maximum diameter of the cluster in each of the directions X, Y and Z, dividing by 3 to get an average diameter, and by 2 to get an average radius. The ‘kernel halfsizes’ mentioned refer to the radii (in voxels) of the spherical kernels in question.

When the first classification step is complete, the candidates are merged again where necessary, as described in Section 2.2.4. Although the original merging step was exhaustive, the subsequent location adjustment step (see Section 2.2.5) means that further merges may now be possible. It is particularly necessary to check for merge possibilities in case the location adjustments have placed two candidates in the same location. Merging is not carried out until after the first classification has been completed because at this stage the number of false-positives has been reduced and the possibility of invalid merges between true and false items is greatly reduced. After the exhaustive merging the nodule locations are once again adjusted to the brightest local points as described in Section 2.2.5.

2.3.3. Final kNN classification

The full set of 135 features calculated for the final kNN classification are detailed in Table 5 with assigned IDs to which we refer in this text. Table 6 lists which features were selected in various experiments. Note that the descriptions in Table 5 make reference to the ‘radius’, r, of the voxel cluster (explained in Section 2.3.2) and to the radius r_{seg}, which is calculated in the same way but

Table 4
The features calculated for the first kNN classifier. See text in Section 2.3.2.

ID	Description	Notes
Features of the voxel cluster		
a1	Cluster size (number of voxels)	
a2	Compactness1, $\frac{ClusterSize}{(dim_x)(dim_y)(dim_z)}$	dim_i = width in dim. i
a3	Compactness2, $\frac{ClusterSize}{max\_dim^3}$	$max\_dim = max_i(dim_i)$
a4	Ratio max_dim:min_dim	$min\_dim = min_i(dim_i)$
a5	Ratio max_dim:med_dim	$med\_dim = median_i(dim_i)$
a6	Ratio A_{med}:A_{max} where A_{max}, A_{med} and A_{min} are the eigenvalues for the eigenvectors of the cluster data by principal component analysis	
a7	Ratio A_{min}:A_{max}	as for a6 above
a8	Sphericity, $\frac{num\_cluster\_voxels\_in\_sphere\_S}{vol\_sphere\_S}$ where $sphere\_S$ is a sphere at the candidate location with radius r	
a9	Ratio Sphericity:r	
Features of voxels in spherical kernels at the candidate location		
a10-a18	On grey-values over spherical kernels K: Average, Median, Standard-Deviation	Halfsizes of K: 1 (a10-a12), 3 (a13-a15), r (a16-a18)

Table 5
The features calculated for the final kNN classifier. See text in Section 2.3.3

ID	Description	Notes
Features of the voxel cluster		
b1-b9	Features a1-a9 as described in Table 4	
b19	$min\_dim = min_i(dim_i)$	dim_i = width in dim. i
b20	$max\_dim = max_i(dim_i)$	dim_i = width in dim. i
Features of voxels in spherical kernels at the candidate location		
b10-b18	Features a10-a18 as described in Table 4	
b21-b26	On grey-values over spherical kernels K: Min, Max	Halfsizes of K: 1 (b21-b22), 3 (b23-b24), r (b25-b26)
b27-b36	On SI over spherical kernels K: Average, Median, Std-Dev, Min, Max	Halfsizes of K: 3 (b27-b31), r (b32-b36)
b37-b46	On CV over spherical kernels K: Average, Median, Std-Dev, Min, Max	Halfsizes of K: 3 (b37-b41), r (b42-b46)
Features calculated on randomly chosen points on a spherical surface around the candidate location.		
b47-b76	Features of Gradient orientation values: Average(Avg), Median, Max, Min, Std-Dev, Coefficient of Variation, Ratio Max:Min, Ratio Std-Dev:Median, Ratio Median:Avg, Ratio Median:Max	30 points on sphere rad = 3 (b47-b56), 50 points on sphere rad = r (b57-b66), 50 points on sphere rad = r_{seg} (b67-b76)
b77-b106	Features of Gradient magnitude values: Average(Avg), Median, Max, Min, Std-Dev, Coefficient of Variation, Ratio Max:Min, Ratio Std-Dev:Median, Ratio Median:Avg, Ratio Median:Max	30 points on sphere rad = 3 (b77-b86), 50 points on sphere rad=r (b87-b96), 50 points on sphere rad = r_{seg} (b97-b106)
Features of voxels in the candidate segmentation		
b107-b115	Features a1-a9 as described in Table 4 but calculated this time over the segmented voxels NOT the cluster voxels	
b116	$min\_dim = min_i(dim_i)$	
b117	$max\_dim = max_i(dim_i)$	
b118-b122	On grey-values over segmented voxels: Average, Median, Std-Dev, Min, Max	
b123-b127	On SI of segmented voxels: Average, Median, Std-Dev, Min, Max	
b128-b132	On CV of segmented voxels: Average, Median, Std-Dev, Min, Max	
b133	Ratio Num segmented voxels: Num ROI voxels	
b134	Ratio {Distance from candidate location to the farthest point in the segmentation}: {Number of voxels in the segmentation}	
Other features		
b135	Posterior probability of being a true nodule from the first classication step	

Table 6
Features selected during experiments on the 3 databases A, B and C

Classifier 1				Classifier 2											
	A	B	C		A	B	C		A	B	C		A	B	C
a1	1	1	1	b1	1	1	1	b46	0	0	1	b92	0	0	1
a2	0	1	0	b5	0	0	1	b49	0	0	1	b93	0	0	1
a3	1	0	1	b7	0	0	1	b52	1	0	0	b94	0	0	1
a4	0	0	1	b8	0	0	1	b54	0	0	1	b103	1	0	1
a7	0	0	1	b12	0	1	0	b55	1	0	0	b107	0	0	1
a8	1	1	0	b13	0	0	1	b56	1	0	0	b113	0	1	0
a9	0	1	0	b21	0	0	1	b57	0	0	1	b115	0	0	1
a10	1	1	1	b22	0	1	0	b58	0	1	1	b116	0	0	1
a11	1	1	1	b24	0	0	1	b62	0	1	1	b120	1	1	0
a12	1	0	1	b25	1	0	0	b64	1	0	1	b122	0	1	0
a13	1	0	0	b26	0	0	1	b65	1	1	1	b123	1	1	1
a14	1	0	1	b27	0	0	1	b66	0	1	1	b124	1	0	1
a15	1	0	1	b28	0	1	1	b67	0	1	0	b125	1	0	1
a16	1	0	1	b29	1	1	0	b68	1	0	0	b126	1	0	1
a17	0	1	0	b36	0	0	1	b70	1	0	0	b129	0	0	1
a18	0	1	0	b39	0	0	1	b72	0	0	1	b130	0	0	1
				b40	0	0	1	b75	1	0	1	b131	0	1	0
				b41	1	0	1	b79	0	1	0	b134	1	1	1
				b44	0	0	1	b83	0	1	1	b135	1	1	1
				b45	0	0	1	b90	0	0	1				
Total	10	8	10									Total	20	19	44

using the voxels contained in the nodule segmentation. The ‘kernel halfsizes’ mentioned refer to the radii (in voxels) of the spherical kernels in question.

All of the features calculated in the first classification step are re-used (b1–b18), and the posterior probability of a structure being a true nodule, which was calculated by the first classifier, is used as a feature at this stage also (b135).

Features of the SI and CV values in spherical kernels placed at the candidate location are added at this time (b27–b46), in addition to features of gradient orientation and magnitude on hypothetical spherical surfaces around the centre of the structure (b47–b106). Most true nodules have a shape that is roughly spherical and a gradient field that is roughly radially symmetric. For true-positives we therefore expect the gradient magnitude to be similar at all points on a generated spherical surface, while gradient orientation should always be roughly normal to the surface of the sphere. These features are calculated on randomly chosen points from spherical surfaces of various sizes. The number of points to be sampled on a sphere of a particular size was empirically determined. The gradient orientation is the component of gradient in the radial direction, and is defined at a point p as $(r \cdot g)/|r||g|$ where g is the gradient vector at p and r is the radial vector from the sphere centre to p.

We segment the candidate structures using an algorithm described by Kostis et al. (2003) in order to calculate features of the segmented objects. In order to improve the accuracy of the segmentations the process was carried out on a region of interest centred at the candidate location but extracted from the full resolution image rather than the sub-sampled version. This was done since the segmentation on the sub-sampled image was more prone to failure, however for convenience the coordinates of the segmented voxels were converted back to their equivalent values in the sub-sampled image for further processing. The features of the segmented voxels (b107–b134) which are calculated are broadly similar to those which were calculated relating to the initially detected voxel clusters. With these features we aim to determine whether the size and shape of the segmented object are commensurate with those of a true nodule and we expect that a good segmentation will give more accurate feature information than that obtained from the initial voxel cluster.

Finally the candidate merging procedure is performed one more time (as previously, new merges may be possible following the location adjustment step after the first kNN classification), and the locations are adjusted to the brightest local point for a final time. This step aims primarily to ensure that we find only one detection per annotation (and thus do not count surplus correct detections as false-positives).

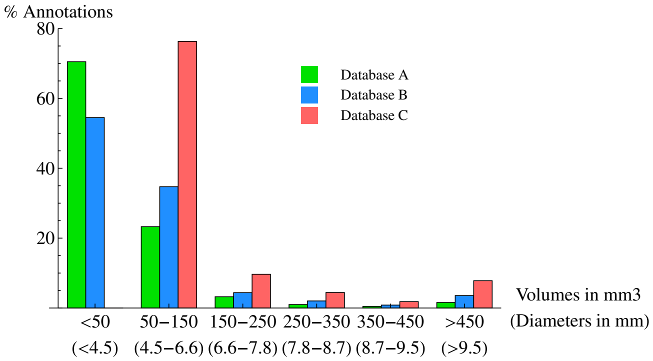


Fig. 5. The proportions of nodules of various sizes in the three databases. Diameter measurements refer to diameters of spheres with volumes equivalent to those of the nodules.

3. Results

In this section we present the experiments performed and the results for nodule detection on each of the databases created and analyse the algorithm performance. Statistics relating to the numbers of scans, and the numbers and sizes of nodules in each of the databases are illustrated in Figs. 5 and 6.

3.1. Database A

The training sets for Database A were formulated as described in Section 2.3.1. The initial 750 training scans were divided randomly into three groups of 250 scans each, to be used in the three steps of the training set generation procedure. After the removal of scans where lung segmentation failed, the three groups contained 242, 243 and 237 scans respectively.

The training set for the initial classifier contained 5776 samples, 1351 of which were true nodules, while the training set for the final classifier contained 3436 samples with 819 true-positives (see Fig. 6b).

The testing phase began with the SFFS feature selection procedure in which features were chosen for each classification step as shown in Table 6. The first classification proceeded with 10 selected features while the final classifier used a total of 20 features. The number of scans processed in the test data is 813 after removal of scans where the lung segmentation failed.

The results of the nodule detection for Database A are shown in Table 7, while the FROC curve achieved by varying the operating point in the final classification is shown in Fig. 8a.

3.2. Database B

This database consisted of scans which contained at least one nodule of clinical interest due to its size or physical properties. Most of these scans also contain other smaller or less significant nodules, meaning that the database contains a wide range of nodule sizes (see Fig. 5).

The training sets were created as described in Section 2.3.1. The 600 training scans were divided into three groups of 200 scans each, to be used in the three steps of the training set generation procedure. This division was random with the exception of the 14 patients known to have a malignancy which were placed in the third set, Scans_Step3. After removal of the scans where lung segmentation failed the three groups had 191, 196 and 193 scans respectively.

The training set for the initial classifier contained 7868 samples, 1715 of which were true nodules, while the training set for the final classifier contained 4490 samples with 1090 true-positives (see Fig. 6b).

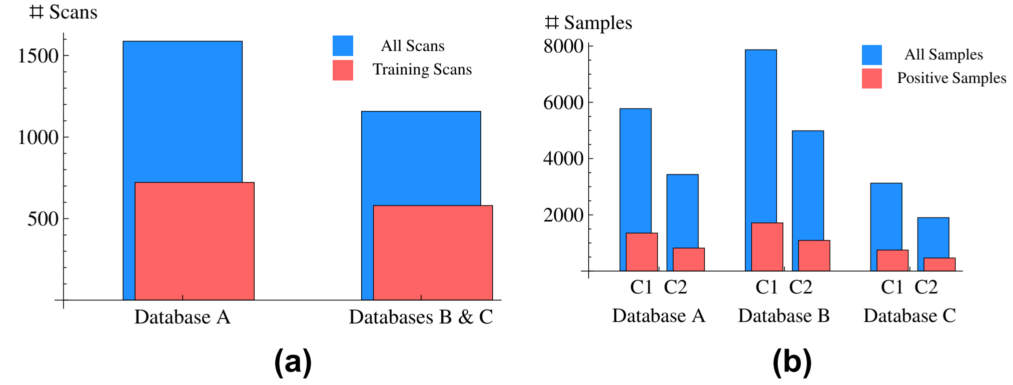


Fig. 6. (a) Number of scans per database. (b) Number of samples in training data sets. Classifier 1 training data is indicated by *C1* and classifier 2 training data by *C2*.

Table 7
Results for experiments on database A.

Number of Scans	813	
Number of annotations	1525	
	Sensitivity	FP per scan
After initial candidate detection	97.2%	649.0
After first classification	92.3%	77.3
After final classification		
– At around 4 FP per scan	80.0%	4.2

The testing phase began with the SFFS feature selection procedure in which features were chosen for each classification step as shown in Table 6. A total of eight features were selected for the initial classification, with 19 features being chosen for the final classifier. The number of scans processed in the test data is 541 after removal of scans where the lung segmentation failed.

The results of the nodule detection for Database B are shown in Table 8 and illustrated in an FROC curve in Fig. 8a. A selection of nodules from Database B are shown in Fig. 7. These nodules were detected by the system with varying degrees of certainty (posterior probabilities). It is difficult to make generalisations based on a small number of samples such as is shown in Fig. 7, in particular since the 3D characteristics of these structures cannot be gauged from these images. However it can be seen that the nodules in the top row, which were detected with posterior probabilities (*pp*) higher than 0.9, generally have a reasonably spherical appearance and are frequently clear of any other structures in the locality. Nodules in the second row ($0.45 < pp < 0.9$) have similar characteristics and are also successfully detected at an overall rate of four FP per scan. In the third row the nodules shown have $0.35 < pp < 0.45$ which is around the borderline point for detection at a rate of four FP per scan. These nodules tend to have less spherical surfaces and/or are close to other structures or surfaces which make their detection more difficult. Finally the nodules presented in the bottom row are among those which are not detected at a rate of four FP per scan ($pp < 0.35$). As with the row above, these nodules have awkward shapes, or are located alongside other confounding surfaces and structures, or in regions of pathology.

3.3. Database C

This database contained precisely the same images as Database B, but only those nodules with volumes above 50 mm^3 are considered in both the training and testing steps. The procedure to divide the data into groups and build the training sets was exactly as described in Section 3.2 except that only nodules with volumes above 50 mm^3 were included. The training set for the initial classifier contained 3127 samples, 748 of which were true nodules, while the training set for the final classifier contained 1900 samples with 465 true-positives (see Fig. 6b).

The testing phase began with the SFFS feature selection procedure in which features were chosen for each classification step as shown in Table 6. The initial classifier chose 10 features for use in classification while the final classifier chose a total of 44 features. The number of scans processed in the test data is 541 after removal of scans where the lung segmentation failed.

The results of the nodule detection for Database C are shown in Table 9 and illustrated in an FROC curve in Fig. 8a.

4. Analysis

It is clear from the FROC curves shown in Fig. 8a that the best performance was achieved on Database A, with the algorithm performing slightly better on Database C than on Database B. In this section we analyse the algorithm performance in detail in order to determine what factors contribute to a successful or poor performance. In Sections 4.1, 4.2, and 4.3 we examine the algorithm performance on Database B in relation to nodule size, nodule location and nodule type. The performance trends which we see for Database B in relation to these sub-groups of nodules are found to be similar for Databases A and C, therefore we do not repeat the same

Table 8
Results for experiments on database B.

Number of scans	541	
Number of annotations	1688	
	Sensitivity	FP per scan
After initial candidate detection	97.7%	750.5
After first classification	91.9%	69.0
After final classification		
– At around 4 FP per scan	72.4%	4.0

Table 9
Results for experiments on database C.

Number of scans	541	
Number of annotations	768	
	Sensitivity	FP per scan
After initial candidate detection	98.2%	752.1
After first classification	92.2%	51.2
After final classification		
– At around 4 FP per scan	77.7%	4.2

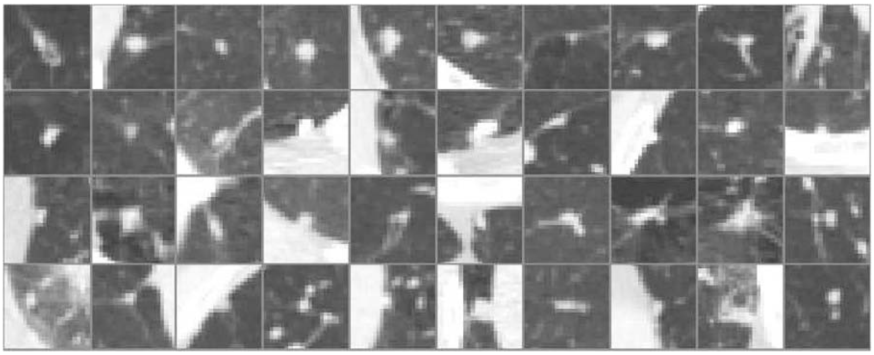


Fig. 7. Randomly selected sample nodules from database B with varying posterior probabilities (*pp*) assigned by the detection system. All images are in axial view. The cutoff *pp* for detection at 4 FP per scan was 0.41. Top Row: Nodules detected with high pp ($pp > 0.9$). Second Row: Nodules detected with lower pp ($0.45 < pp < 0.9$). Third Row: Nodules with borderline pp values ($0.35 < pp < 0.45$). Bottom Row: Nodules not detected at 4FP per scan ($pp < 0.35$).

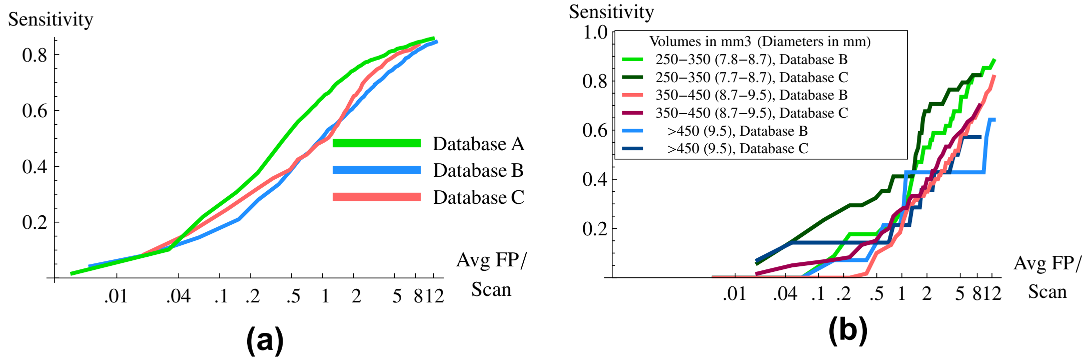


Fig. 8. (a) Overall system FROC curves for all three databases. (b) Comparison of performance on databases B and C for larger nodule sizes. The X-axes are on a logarithmic scale to show performance at lower FP rates more clearly. Diameter measurements refer to diameters of spheres with volumes equivalent to those of the nodules.

analysis on those databases. In the remainder of this section performance on larger nodule sizes and proven malignancies is examined, false-positive detections are categorised and possible reasons for sporadic high false-positive rates are investigated.

4.1. Analysis by nodule size

The distribution of nodule volumes in the testing data of Database B is shown in Fig. 9a while Fig. 9b shows the performance of the algorithm on each of these volume groups. It is clear that the performance is substantially worse on the groups containing the largest nodules, particularly those above 350 mm^3 ($d \approx 8.7$ mm) in volume.

4.1.1. Performance on larger nodules

Database C was constructed to test the performance of the algorithm when training and testing was performed only on nodules above 50 mm^3 in volume. It is clear from Fig. 8a that its overall performance is slightly better than that of Database B which contains the same nodules but also a large quantity of smaller nodules (see Fig. 5). In Fig. 8b we compare the performance of the algorithm on database B and C on the three groups of nodules with largest volumes, and see that in each case the performance on Database C is slightly better than that on Database B.

4.2. Analysis by nodule location

In the Nelson Trial the radiologist annotating nodule locations was required to give an indication where in the lung the nodule was located. The locations were categorized using the distance d between the costal pleural surface and the hilum as a guideline as follows: Pleural (Nodule is attached to the costal pleural surface), Peripheral (Distance to the costal pleural surface is less than one third of d), Central (Distance to the costal pleural surface is more than two-thirds of d), Other (Distance to the costal pleural surface is between one and two-thirds of d).

In Fig. 10 we examine the distribution of nodule locations in the test data of Database B and the performance of our algorithm in terms of these nodule locations. The system performs best on peripheral nodules which are near the pleural surface but not attached to it. Nodules in the category ‘other’, which do not lie in the region of the mediastinum and are not attached to the pleural surface are also relatively easy to detect, while central nodules and pleural nodules appear to present a much greater challenge.

4.3. Analysis by nodule type

The radiologist annotating nodules for the Nelson Trial was also required to note the type of nodule which had been found, according to the following list of categories: (A) Solid nodule, (B) Partially solid nodule, (C) Non-solid nodule, (D) Benign fat, (E) Benign calcification, (F) Complete calcification, (G) Central calcification, (H) Plate-like atelectasis. For the purposes of this analysis categories B and C are considered together as ‘non/part solid nodules’, and categories D, E, G, H are grouped under the title ‘Other’ due to the very low occurrence rate of these nodule types. Fig. 11a shows the distribution of nodule types in the test data of database B and Fig. 11b shows the performance of the algorithm in detecting the various types.

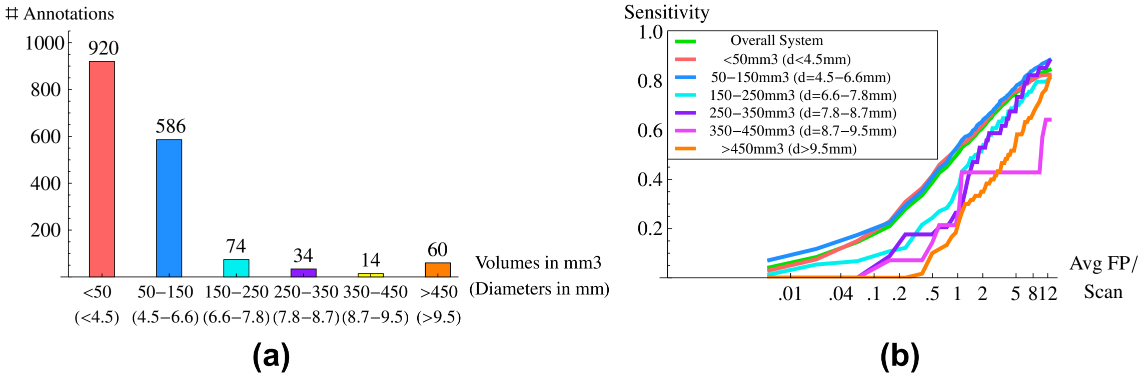


Fig. 9. (a) Distribution of nodule sizes in database B. (b) FROC analysis of system performance per nodule-size group for database B. The X-axis is on a logarithmic scale to show performance at lower FP rates more clearly. Diameter measurements refer to diameters of spheres with volumes equivalent to those of the nodules.

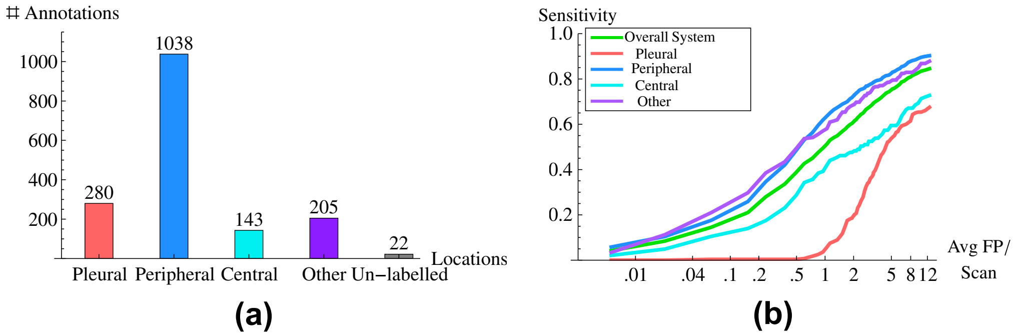


Fig. 10. (a) Distribution of nodule locations in database B. (b) FROC analysis of system performance per nodule-location for database B. The X-axis is on a logarithmic scale to show performance at lower FP rates more clearly.

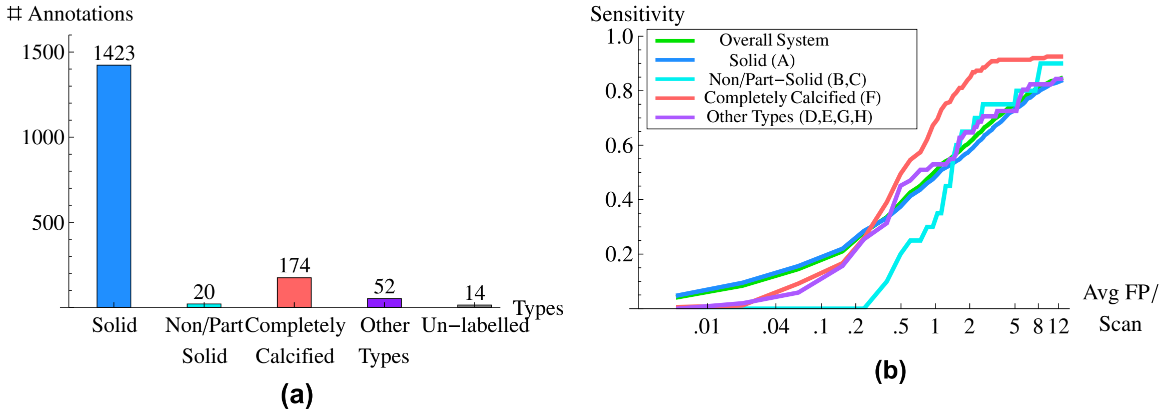


Fig. 11. (a) Distribution of nodule types in database B. (b) FROC analysis of system performance per nodule type for database B. The X-axis is on a logarithmic scale to show performance at lower FP rates more clearly.

4.4. Detection rate of proven malignancies

Within the test set for database C there are 20 patients who have now had a lesion proved malignant by means of a biopsy. To determine the ability of our system to detect proven cancer at an early stage, and in line with our normal protocol (see Section 1.1) we took the earliest available scan from each of these patients for inclusion in our database.

Of the 20 malignant lesions in the test set, one was excluded since the nodule was not yet visible in the earliest scan of the patient which was processed by our algorithm.

Fourteen of the remaining 19 malignancies (73.7%) were successfully detected. Two of the five malignant nodules undetected by the system were erroneously excluded from the lung segmentation and thus could not be identified. These results are based on the same operating point as that which gave an average of four false-positives per scan over the entire test set in Database C. Over these 19 images there was an overall sensitivity of 89.2% with an average of 5.7 false-positives per scan.

The five missed malignant nodules are depicted in Fig. 12.

4.5. Analysis of false-positive items

In order to analyse the types of FP arising from the system 1013 FP items in 340 scans were examined by a radiologist in training [2]. In the selection of FP items to be examined, scans with more than 15 FP items were excluded since these potentially contained pathology responsible for many of the FP items (see Section 4.6). The scan selection was otherwise random. The FP items in question were detected at an operating point giving an average 4FP per scan over the full database B.

The observer was asked to categorise each FP item into one of 13 categories as follows: Nodule, pathology, vessel, vessel junction, fissure, scar tissue, hilum, mediastinum, protruding rib, pleural wall, fluid/phlegm, spinal column, other. The results of the observer analysis are illustrated in Fig. 13 which also shows the changing trend in FP sources at different operating points of the CAD system. It should be noted that these results, as with all cases of single-reading, represent the opinion of a single observer and may be open to interpretation in some cases. They are nonetheless indicative of the general trend in the processed data.

The observer judged that 159 of the 1013 FP detections were in fact nodules (15.7%). Many of these were too small to have been annotated for the Nelson Trial but nonetheless it is acceptable for a system such as this to identify them.

Vessels, vessel junctions and parts of fissures made up 464 (45.8%) of the FP items examined. Also with relatively high proportions were parts of the pleural surface (7.9%), parts of the spinal column (7.9%) and protruding ribs (7.3%). A further 4.8% of FP items were caused by fluid/phlegm in the bronchi. Additional lesser difficulties are presented by the central regions of the hilum (2.4%) and the mediastinum (0.5%).

The remaining 15.8% of FP items are classified in the category ‘other’ which included regions of the bronchi (4), artefacts (4), pacemaker leads (3), atelectasis (4) and calcium in the coronary artery (1).

Fig. 14 shows examples of false positive structures which were assigned to various categories. Samples of the false-positives which were later identified as nodules are presented on the top left of this

[2] A physician with two years of experience in radiology, particularly in CT lung cancer evaluation.

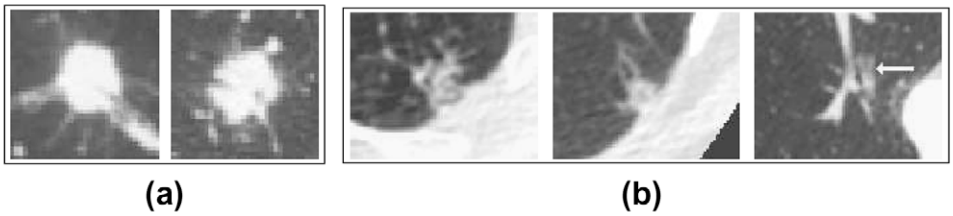


Fig. 12. (a) Two large malignant nodules which could not be detected by the system as they were erroneously excluded from the lung segmentations. (b) Three malignant nodules which were not detected by the CAD system at an average 4FP per scan. Note that the first example was also not detected by the radiologist until a follow-up scan. At an increased false-positive rate of 5FP per scan all three nodules are successfully identified by the system.

image. Many of these were very small structures which were overlooked during the original annotation procedure. The leftmost structure is attached to a vessel junction making it more difficult for an observer to notice. On the top right of the image, vessels and vessel junctions which were erroneously detected as nodules are shown. Again, these are mainly very small structures at turning- or branch-points of vessels. Examples of mediastinal and pleural structures which were detected as nodules are shown on the bottom left of Fig. 14. It is clear that surfaces with properties similar to those of nodules occur frequently in these regions. Pathological structures and an example of atelectasis (rightmost image) are shown on the bottom right. Such structures are very similar in appearance to true nodules.

4.6. Scans with a high proportion of false-positive items

It was noted that a number of scans had an unusually high rate of false-positives compared to the overall system average. In database B, at an overall average rate of four FPs per scan, 19 scans contained 15 or more FPs. The number of FPs per item was maximum in a scan with 103 FP items and the average over the 19 scans was 32.4 FPs per scan These 19 scans were examined by a trainee radiologist to determine whether a clinical reason for the high number of FPs could be identified.

Of the 19 scans checked, nine were deemed to have moderate or severe pathology (mainly interstitial lung disease and fibrosis) which caused the high occurrence of false positive detections. A further three scans contained an extremely high number of nodules, many of which were not annotated due to the sheer volume of nodules already marked and the fact that most of them were clearly benign calcifications. The nodules marked by the CAD system were therefore valid detections and should have been considered true-positives. The remaining seven scans displayed no clear reason for the high false-positive rate apart from poor inhalation in one case and density patterns typical in the lungs of heavy smokers in the others.

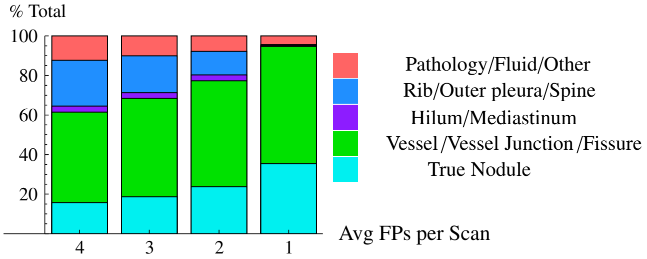


Fig. 13. An illustration of the distribution of false-positive types at various operating points in the CAD system. (This analysis is based on 1013 FP items from database B.)

The exclusion of the 12 scans with pathology or widespread unmarked calcifications increased the system sensitivity from 72.4% to 75.2% at four FP per scan, demonstrating that the presence of a small number of such scans has a noticeable effect on the overall system performance.

5. Discussion

In this work a CAD system for the automatic detection of pulmonary nodules has been presented and extensively evaluated, particularly in the context of lung cancer screening data. Although many such systems have been presented in previous literature, to our knowledge no study of this scale has been carried out before. This type of analysis is vital in the development and improvement of a CAD system in preparation for clinical use. The results we have presented illustrate both the strong and weak points of the system as well as highlighting the general difficulty of detecting all nodule types simultaneously without the introduction of large numbers of false-positive detections.

Our analysis of the system performance categorized by nodule size showed that larger nodules, which generally pose a greater threat to the patient, remain the most difficult to detect (Fig. 9). Although the performance was slightly improved by eliminating the high number of smaller nodules from the training data (see Fig. 8b) the results were not substantially better. In a supervised learning system such as this, the poor performance can be partially attributed to the smaller numbers of training samples available for larger nodules. However it also seems likely that the more irregular shapes and density patterns exhibited by larger nodules make them difficult to identify conclusively by means of local image features. This theory is backed up by the outcome of feature selection in database C compared to databases A and B (Table 6). For

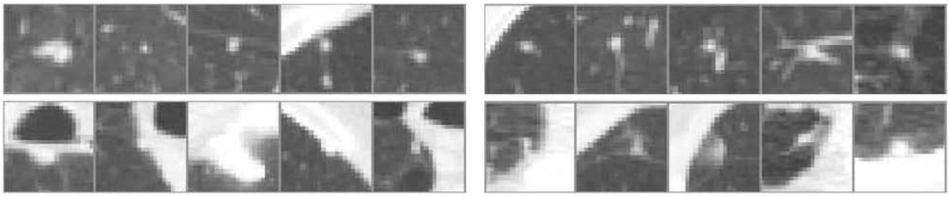


Fig. 14. False-positive items examined and categorised by an expert. All images in axial view. Top Left: 5 items categorised as true nodules. Top Right: 5 items categorised as vessel or vessel junctions. Bottom Left: 5 items categorised as mediastinum/rib/pleural surface. Bottom Right: 5 items categorised as pathology/atelectasis.

databases A and B, which contain predominantly nodules with volumes below 50 mm^3, 20 and 19 features respectively are selected for the second stage classifiers. Database C contains exactly the same set of nodules as Database B excluding those below 50 mm^3, and the feature selection procedure identifies a total of 44 features to be retained for the best results. This increase in the number of features required may be indicative of the much greater diversity in the appearance of larger nodules.

In considering the proficiency of the CAD system in the context of nodule locations (Fig. 10), it can be seen that peripheral nodules present the fewest difficulties. This is to be expected since they lie in an area where their surfaces are unlikely to be compounded by any other structures such as vessels or the mediastinal wall. In addition the training data contains more peripheral nodules than any other type. The system performance is also extremely good on nodules remote from both the mediastinum and the costal-pleural surface (Category 'Other'), again since they are likely to be isolated or have only minor vessel connections.

In contrast however, the performance on pleural nodules and central nodules is relatively poor. Central nodules are located around the area of the mediastinum which has a complex topology with many partially spherical surfaces and high intensity values. Nodules in this region are therefore much more difficult to detect and isolate correctly and have a surrounding area dissimilar to that of other nodule types. After the first classification step we retain 90.9% of the central nodules. This figure is comparable with that for nodules in other locations, implying that it is the second classification step which is mainly at fault in the incorrect elimination of these candidates. The lower performance rate in the detection of pleural nodules is likely to be caused by similar factors. Since they are attached to the pleural surface they are usually roughly hemispherical in shape and therefore significantly different in appearance to all other nodule types. The initial candidate detection step correctly identifies 96.4% of the pleural nodules annotated, however after the first classification only 81.8% are correctly retained which is considerably lower than the general system sensitivity at this point. It is clear that nodules in contact with the pleural surface require special treatment and that a separate procedure should be developed for the detection of these lesions in our CAD system.

The system performance categorised by nodule-type (Fig. 11) shows that the results are clearly substantially better for completely calcified nodules, particularly as the rate of FPs increases. This can be explained by their dense bright appearance and clearly defined boundaries. There is little difference in the system performance among the remaining types at false-positive rates above two FP per scan. Below two FP per scan the performance for the detection of non or part-solid nodules is substantially worse than for other types. This is to be expected due to the limited number of training examples and the atypical appearance of these nodules. However the initial candidate detection step identifies 100% of the non/part solid lesions, and after the first classification 95% of these are retained, indicating that it is the second classification step which fails to assign a high probability of being a nodule to these structures. The development of features which distinguish efficiently between non-solid lesions and false positive items will substantially improve the CAD system performance in this respect.

In the detection of malignant nodules our system performed reasonably well with 73.7% of malignancies being detected at four FP per scan. Two of the five malignant lesions which were not detected were incorrectly excluded from the lung segmentation (Fig. 12a). This highlights the importance of an accurate lung segmentation in a complete CAD system. A further three malignant lesions were not detected (Fig. 12b) at this system operating point, but at an increased false-positive rate of five FP per scan these three nodules are successfully identified. Notably, the first of these three nodules, which was located close to the apex of the lung, was missed by the radiologist in the scan which was processed by the system, and was not detected until a follow-up scan 3 months later. This demonstrates the potential of a CAD system in pinpointing lesions which might otherwise be overlooked by a radiologist.

The analysis of false-positive items (Fig. 13) illustrates that the majority of false detections relate to vessel structure or fissures. This is to be expected since these are the most predominant bright structures in the lung volume. At four FP per scan 23.1% of false positive detections are caused by rib, spine or pleural surface, supporting the aforementioned suggestion that extra measures are required for accurate pleural nodule detection. At an operating point with one FP per scan however, all of these false-positives (with a single exception) are eradicated and vessels or vessel junctions dominate the false-positive group making up 86% of those items which were not categorised as true nodules. Interestingly the true nodule category makes up 35% of the total at this operating point compared to 16% at four FP per scan. This implies that the false-positive items which are actually true nodules tend to have a higher posterior probability of being a nodule than other genuinely false-positive structures. The system therefore shows promise that with some refinement these structures could be efficiently distinguished.

It is expected that the algorithm utilised would be equally successful on other types of chest CT data assuming that appropriate training data is available and nodule visibility is equivalent. Where nodule visibility is reduced (in data with thicker slices for example) the system results may deteriorate, however radiologist readings are also likely to be less accurate on such datasets. Some parameter tuning is likely to be required when using different data, particularly for the threshold values for SI and CV and the limit on cluster size t_{vol}. If the data used to train the system has different properties than the data to be analysed then system performance is likely to be compromised and will depend largely on the magnitude of these data discrepancies.

The focus of this work has been on the detection of pulmonary nodules which is clearly an isolated step in a computer-aided diagnosis processing pipeline. An important addition to this pipeline would be a nodule classification algorithm to determine whether a nodule is likely to be malignant or not. This characterization task has been undertaken by other authors, for example Kawata et al. (2004) and Reeves et al. (2006). Among the most important indicators for nodule malignancy are size, shape, opacity and growth rate (Aoki et al., 2000), all of which require an accurate nodule segmentation in order to be evaluated. Nodule correspondence between temporal scans can be determined (Betke et al., 2003), enabling the growth rate to be calculated fully automatically.

This work has demonstrated the challenges of creating a nodule detection system which works effectively on low-dose data with a diverse range of nodule types and in a variety of circumstances such as the presence of pathology. The presented system is largely successful on the majority of encountered nodules and achieves an accuracy comparable to that of other systems which are evaluated in much more limited conditions. Future work will focus on the improvement of demonstrated weak points of the system in order to develop an effective broad-spectrum nodule detection algorithm.

References

American Cancer Society, 2008. Cancer Facts and Figures.

Aoki, T., Nakata, H., Watanabe, H., Nakamura, K., Kasai, T., Hashimoto, H., Yasumoto, K., Kido, M., 2000. Evolution of peripheral lung adenocarcinomas: CT findings correlated with histology and tumor doubling time. American Journal of Roentgenology 174 (3), 763–768.

Arimura, H., Katsuragawa, S., Suzuki, K., Li, F., Shiraishi, J., Sone, S., Doi, K., 2004. Computerized scheme for automated detection of lung nodules in low-dose computed tomography images for lung cancer screening. Academic Radiology 11 (6), 617–629.

Bae, K.T., Kim, J.S., Na, Y.H., Kim, K.G., Kim, J.H., 2005. Pulmonary nodules: automated detection on CT images with morphologic matching algorithm – preliminary results. Radiology 236, 286–294.

Bellotti, R., De Carlo, F., Gargano, G., Tangaro, S., Cascio, D., Catanzariti, E., Cerello, P., Cheran, S.C., Delogu, P., De Mitri, I., Fulcheri, C., Grosso, D., Retico, A., Squarcia, S., Tommasi, E., Golosio, Bruno, 2007. A CAD system for nodule detection in low-dose lung CTs based on region growing and a new active contour model. Medical Physics 34 (12), 4901–4910.

Betke, M., Hong, H., Ko, J.P., 2003. Landmark detection in the chest and registration of lung surfaces with an application to nodule registration. Medical Image Analysis 7, 265–281.

Brown, M.S., Goldin, J.G., Suh, R.D., McNitt-Gray, M.F., Sayre, J.W., Aberle, D.R., 2003. Lung micronodules: automated method for detection at thin-section CT – initial experience. Radiology 226, 256–262.

Canny, J., 1986. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 8 (6), 679–698.

Chang, S., Emoto, H., Metaxas, D.N., Axel, L., 2004. Pulmonary micronodule detection from 3D chest CT. In: Medical Image Computing and Computer-Assisted Intervention, vol. 3217, pp. 821–828.

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13 (1), 21–27.

Dehmeshki, J., Ye, X., Lin, X., Valdivieso, M., Amin, H., 2007. Automated detection of lung nodules in CT images using shape-based genetic algorithm. Computerized Medical Imaging and Graphics 31 (6), 408–417.

Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, second ed. John Wiley and Sons, New York.

Enquobahrie, A.A., Reeves, A.P., Yankelevitz, D.F., Henschke, C.I., 2007. Automated detection of small pulmonary nodules in whole lung CT scans. Academic Radiology 14, 579–593.

Ge, Z., Sahiner, B., Chan, H., Hadjiiski, L.M., Cascade, P.N., Bogot, N., Kazerooni, E.A., Wei, J., Zhou, C., 2005. Computer-aided detection of lung nodules: false positive reduction using a 3D gradient field method and 3D ellipsoid fitting. Medical Physics 32 (8), 2443–2454.

Gohagan, J.K., Marcus, P.M., Fagerstrom, R.M., Pinsky, P.F., Kramer, B.S., Prorok, P.C., Ascher, S., Bailey, W., Brewer, B., Church, T., Engelhard, D., Ford, M., Fouad, M., Freedman, M., Gelmann, E., Gierada, D., Hocking, W., Inampudi, S., Irons, B., Johnson, C.C., Jones, A., Kucera, G., Kvale, P., Lappe, K., Manor, W., Moore, A., Nath, H., Neff, S., Oken, M., Moore, A., Plunkett, M., Price, H., Reding, D., Riley, T., Schwartz, M., Spizarny, D., Yoffie, R., Zylak, C.and the Lung Screening Study Research Group, 2005. Final results of the Lung Screening study, a randomized feasibility study of spiral CT versus chest X-ray screening for lung cancer. Lung Cancer 47 (1), 9–15.

Gurney, J.W., 1996. Missed lung cancer at CT: imaging findings in nine patients. Radiology 199, 122–177.

Henschke, C.I., McCauley, D.I., Yankelevitz, D.F., Naidich, D.P., McGuinness, G., Miettinen, O.S., Libby, D., Pasmantier, M., Koizumi, J., Altorki, N., Smith, J.P., 2001. Early lung cancer action project: a summary of the findings on baseline screening. The Oncologist 6, 147–152.

Hu, S., Hoffman, E.A., Reinhardt, J.M., 2001. Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. IEEE Transactions on Medical Imaging 20, 490–498.

New York Early Lung Cancer Action Project Investigators, 2007. CT Screening for lung cancer: diagnoses resulting from the New York Early Lung Cancer Action Project. Radiology 243(1), 239–249.

Jeong, Y.J., Yi, C.A., Lee, K.S., 2007. Solitary pulmonary nodules: detection, characterization, and guidance for further diagnostic workup and treatment. American Journal of Roentgenology 188, 57–68.

Kakinuma, R., Ohmatsu, H., Kaneko, M., Eguchi, K., Naruke, T., Nagai, K., Nishiwaki, Y., Suzuki, A., Moriyama, N., 1999. Detection failures in spiral CT screening for lung cancer: analysis of CT findings. Radiology 212, 61–66.

Kawata, Y., Niki, N., Ohmatsu, H., Kusumoto, M., Kakinuma, R., Yamada, K., Mori, K., Nishiyama, H., Eguchi, K., Kaneko, M., Moriyama, N. (2004). Pulmonary nodule classification based on nodule retrieval from 3-D thoracic CT image database. In Medical image computing and computer-assisted intervention, vol. 3217, pp. 838–846.

Koenderink, J.J., 1990. Solid Shape. MIT Press, Cambridge, MA.

Kostis, W.J., Reeves, A.P., Yankelevitz, D.F., Henschke, C.I., 2003. Three-dimensional segmentation and growth rate estimation of small pulmonary nodules in helical CT images. IEEE Transactions on Medical Imaging 22 (10), 1259–1274.

Li, Q., Li, F., Doi, K., 2008. Computerized detection of lung nodules in thin-section CT images by use of selective enhancement filters and an automated rule-based classifier. Academic Radiology 15 (2), 165–175.

Matsumoto, S., Kundel, H.L., Gee, J.C., Gefter, W.B., Hatabu, H., 2006. Pulmonary nodule detection in CT images with quantized convergence index filter. Medical Image Analysis 10 (3), 343–352.

McCulloch, C.C., Kaucic, R.A., Mendonça, P.R.S., Walter, D.J., Avila, R.S., 2004. Model-based detection of lung nodules in computed tomography exams. Academic Radiology 11 (3), 258–266.

Mendonça, P.R.S., Bhotika, R., Zhao, F., Miller, J.V., 2007. Lung nodule detection via Bayesian voxel labeling. In Information Processing in Medical Imaging, vol. 4584, pp. 134–146.

Metz, C.E., 1986. ROC methodology in radiologic imaging. Investigative Radiology 21 (9), 720–733.

Mitra, P., Murthy, C.A., Pal, S.K., 2002. Density based multiscale data condensation. IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (6), 734–747.

Murphy, K., Schilham, A.M.R., Gietema, H., Prokop, M., van Ginneken, B., 2007. Automated detection of pulmonary nodules from low-dose computed tomography scans using a two-stage classification system based on local image features. In Proceedings of the SPIE, vol. 6514, pp. 651410-1–651410-12.

Novello, S., Fava, C., Borasio, P., Dogliotti, L., Cortese, G., Crida, B., Selvaggi, G., Lausi, P., Brizzi, M.P., Sperone, P., Cardinale, L., Ferraris, F., Perotto, F., Priola, A., Scagliotti, G.V., 2005. Three-year findings of an early lung cancer detection feasibility study with low-dose spiral computed tomography in heavy smokers. Annals of Oncology 16 (10), 1662–1666.

Paik, D.S., Beaulieu, C.F., Rubin, G.D., Acar, B., Jeffrey Jr., R.B., Yee, J., Dey, J., Napel, S., 2004. Surface normal overlap: a computer-aided detection algorithm with application to colonic polyps and lung nodules in helical CT. IEEE Transactions on Medical Imaging 23 (6), 661–675.

Pudil, P., Novovicova, J., Kittler, J., 1994. Floating search methods in feature selection. Pattern Recognition Letters 15 (11), 1119–1125.

Reeves, A.P., Chan, A.B., Yankelevitz, D.F., Henschke, C.I., Kressler, B., Kostis, W.J., 2006. On measuring the change in size of pulmonary nodules. IEEE Transactions on Medical Imaging 25 (4), 435–450.

Retico, A., Delogu, P., Fantacci, M.E., Gori, I., Preite Martinez, A., 2008. Lung nodule detection in low-dose and thin-slice computed tomography. Computers in Biology and Medicine 38 (4), 525–534.

Rubin, G.D., Lyo, J.K., Paik, D.S., Sherbondy, A.J., Chow, L.C., Leung, A.N., Mindelzun, R., Schraedley-Desmond, P.K., Zinck, S.E., Naidich, D.P., Napel, S., 2005. Pulmonary nodules on multidetector row CT scans: performance comparison of radiologists and computer-aided detection. Radiology 234, 274–283.

Sluimer, I.C., Prokop, M., van Ginneken, B., 2005. Towards automated segmentation of the pathological lung in CT. IEEE Transactions on Medical Imaging 24 (8), 1025–1038.

Swensen, S.J., Jett, J.R., Hartman, T.E., Midthun, D.E., Mandrekar, S.J., Hillman, S.L., Sykes, A., Aughenbaugh, G.L., Bungum, A.O., Allen, K.L., 2005. CT screening for lung cancer: five-year prospective experience. Radiology 235 (1), 259–265.

Swensen, S.J., Jett, J.R., Sloan, J.A., Midthun, D.E., Hartman, E.E., Sykes, A., Aughenbaugh, G.L., Zink, F.E., Hillman, S.L., Noetzel, G.R., Marks, R.S., Clayton, A.C., Pairolero, P.C., 2002. Screening for lung cancer with low-dose spiral computed tomography. American Journal of Respiratory and Critical Care 165, 508–513.

White, C.S., Romney, B.M., Mason, A.C., Austin, J.H., Miller, B.H., Protopapas, Z., 1996. Primary carcinoma of the lung overlooked at CT: analysis of findings in 14 patients. Radiology 199 (1), 109–115.

Wiemker, R., Rogalla, P., Blaffert, T., Sifri, D., Hay, O., Shah, E., Truyen, R., Fleiter, T., 2005. Aspects of computer-aided detection (CAD) and volumetry of pulmonary nodules using multislice CT. British Journal of Radiology 78, S46–S56. Spec No. 1.

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

Ye, X. , Lin, X. , Beddoe, G., Dehmeshki, J., 2007. Efficient computer-aided detection of ground-glass opacity nodules in thoracic CT images. In: Conference Proceedings on IEEE Engineering in Medicine & Biology Society, vol. 1, pp. 4449–4452.

Zhang, X., McLennan, G., Hoffman, E.A., Sonka, M., 2005. Automated detection of small-size pulmonary nodules based on helical CT images. In: Information Processing in Medical Imaging, no. 3565, pp. 664–676.

Zhao, B., Gamsu, G., Ginsberg, M.S., Jiang, L., Schwartz, L.H., 2003. Automatic detection of small lung nodules on CT utilizing a local density maximum algorithm. Journal of Applied Clinical Medical Physics 4 (3), 248–260.

Zhou, J., Chang, S., Metaxas, D.N., Zhao, B., Schwartz, L.H., Ginsberg, M.S., 2006. Automatic detection and segmentation of ground glass opacity nodules. Medical Image Computing and Computer-Assisted Intervention 9(Pt 1), 784–791.