Reduction of False Positives in Computerized Detection of Lung Nodules in Chest Radiographs Using Artificial Neural Networks, Discriminant Analysis, and a Rule-Based Scheme

Yuzheng C. Wu, Kunio Doi, Maryellen L. Giger, Charles E. Metz, and Wei Zhang

A computer-aided diagnosis (CAD) scheme is being developed to identify image regions considered suspicious for lung nodules in chest radiographs to assist radiologists in making correct diagnoses. Automated classifiers—an artificial neural network, discriminant analysis, and a rule-based scheme—are used to reduce the number of false-positive detections of the CAD scheme. The CAD scheme first detects nodule candidates from chest radiographs based on a difference image technique. Nine image features characterizing nodules are extracted automatically for each of the nodule candidates. The extracted image features are then used as input data to the classifiers for distinguishing actual nodules from the false-positive detections. The performances of the classifiers are evaluated by receiver-operating characteristic analysis. On the basis of the database of 30 normal and 30 abnormal chest images, the neural network achieves an A_Z value (area under the receiver-operating-characteristic curve) of 0.79 in detecting lung nodules, as tested by the round-robin method. The neural network, after being trained with a training database, is able to eliminate more than 83% of the false-positive detections reported by the CAD scheme. Moreover, the combination of the trained neural network and a rule-based scheme eliminates 96% of the false-positive detections of the CAD scheme.
Copyright © 1994 by W.B. Saunders Company

KEY WORDS: Artificial neural networks, discriminant analysis, rule-based scheme, lung nodules, receiver-operating characteristic (ROC) analysis.

$\mathbf{A}$RTIFICIAL NEURAL networks[1,2] have found widespread applications in medical imaging in recent years and have been shown to be a very powerful tool in pattern recognition and data classification.[3-16] In mammography, neural networks have been applied for the detection of microcalcifications,[5,6] and classification of breast lesions.[6,7] In chest radiography, neural networks have been used for classification of interstitial disease,[8] interpretation of neonatal images,[9] detection of lung nodules,[10-17] and estimation of scatter.[18] Neural networks have been applied also to other imaging modalities such as nuclear medicine for lesion detection in single-photon emission computed tomography images[19] and magnetic resonance imaging for image segmentation.[20]

Radiologists can fail to detect pulmonary nodules in up to 30% of cases with actually positive findings.[21] Diagnoses may be missed because of camouflaging effects of anatomic background, subjective or varying decision criteria, or distractions in clinical situations. Automated detection schemes that can detect potential nodules and alert radiologists of locations of possible nodule candidates may be very useful in improving diagnostic accuracy.

A computer-aided diagnosis (CAD) scheme is being developed to detect lung nodules and to help to reduce the number of false-negative diagnoses.[22-25] However, the number of false-positive detections obtained with the current CAD scheme is relatively high. Therefore, it may be useful to employ an automated classifier to further reduce the number of false-positive detections and thereby improve the overall performance of the CAD scheme.

We have attempted to apply artificial neural networks to distinguish lung nodules from false-positive detections of the CAD scheme based on image data directly.[10,11] In this attempt, 64- × 64-pixel regions of interest (ROI) that contained lung nodules and false-positive detections were selected from digitized chest radiographs that had been subjected to the CAD scheme. These ROIs were preprocessed by using a background-trend correction technique[26] to suppress nonuniform background. Each trend-corrected ROI was then sub-sampled to form a smaller 8 × 8 pixels ROI matrix. After being normalized, these ROIs were used as the

input to the neural network. The neural network performed well in detecting obvious lung nodules. However, the present version of the neural network was unable to detect very subtle nodules on the basis of image data.

As an alternative approach, we are applying automated classifiers to reduce the number of false-positive detections that are obtained when image features are extracted automatically by computer. The classifiers that are used include neural networks, discriminant analysis, and a rule-based scheme. The CAD scheme first identifies nodule candidates and then automatically extracts image features for each of the candidates identified. Using the extracted-image features, the classifiers are trained to distinguish nodules from non-nodule detections. Thus, the number of false-positive detections of the CAD scheme can be reduced when it incorporates the trained automated classifiers.

MATERIALS AND METHODS

In this study, 60 conventional, posterior-anterior chest radiographs obtained from 30 normal and 30 abnormal patients were selected for our database. The 30 abnormal cases had 32 pulmonary nodules with diameters ranging from 6 to 25 mm. The presence of pulmonary nodules was verified by computed tomography scans or radiographic follow-up. Calcified nodules and nodules with secondary features, such as atelectasis, were excluded. In the 30 normal cases, the absence of nodules was confirmed by the consensus of two chest radiologists.

The 60 chest radiographs were digitized by a Fuji optical drum scanner (Fuji Medical Systems, Tokyo, Japan)[27] using a 0.1-mm pixel size and 10-bit quantization. The images were subsequently averaged into a 512- × 512-pixel matrix with an effective pixel size of 0.6 mm.

The CAD scheme was based on a difference-image technique.[22-25] In this difference-image approach, two filtered images were created from an original chest image—one was signal enhanced, and the other signal suppressed. The nonuniform background in the original chest image can be substantially removed by subtracting the signal-suppressed image from the signal-enhanced image to obtain a difference image. Gray-level thresholding was performed on the difference image to locate initial nodule candidates. After being subjected to tests of shape, size, and contrast, the remaining nodule candidates were reported. A sensitivity of ~70% was achieved by this CAD scheme. A false-positive detection by the CAD scheme was defined as a nodule candidate, indicated by the CAD scheme, that was actually a normal anatomic structure such as ribs or pulmonary vessels. In this study, a "relaxed" threshold level was used so that 11 to 12 false-positive detections per image were produced by the CAD scheme. By applying the CAD scheme to our database of 60 radiographs (32 nodules), we

identified 727 nodule candidates that included 23 nodules and 704 false-positive detections.

After the initial identification of nodule candidates, automated image feature-extraction techniques were used. Nine image features were calculated for each of the nodule candidates regarding the contrast, effective size, degree of circularity, degree of irregularity, slope of contrast, slope of circularity, slope of irregularity, edge gradient, and standard deviation of edge gradient. These image features were then used as the input to the automated classifiers—a neural network, discriminant analysis, and a rule-based scheme—to distinguish the actual nodules from false-positive detections (ie, non-nodules). The performance of the classifiers was evaluated by receiver-operating characteristic (ROC) analysis.[28-32]

Image Feature Analysis

For each of the nodule candidates initially identified by the CAD scheme, automatic gray-level region growing[33] is performed. In such growing, the size of the grown regions is analyzed as a function of the gray levels. The gray levels relative to the peak of the nodule candidate for region growing are kept the same for each of the nodule candidates. Fig 1, A and B show a nodule and the corresponding contour lines, respectively; contour lines represent the grown areas corresponding to the various gray levels. The area within each contour line will be referred to as a "grown region." Nine image features are calculated for the nodule candidates. Contrast is defined as the gray-level interval used in the region growing process, ie, the difference between the maximum and the minimum pixel values within a grown region. The effective size, the degree of circularity and the degree of irregularity are defined in terms of parameters that are demonstrated in Fig 2. A circle that has the same area as the grown region is centered at the centroid of the grown region. We denote the area of the grown region by A_g and the portion of A_g that is within the circle by A_c. Perimeters of the circle and the grown region are denoted by P_c and P_g, respectively. Effective size is defined as the diameter of the circle that has the same area as the grown region. The degree of circularity is defined as the ratio of the area of the grown region within the circle to the area of the grown region. If a grown region is a perfect circle, the degree of circularity will be equal to 1. On the other hand, if a grown region has a narrow linear shape, the degree of circularity will be close to 0. The degree of irregularity is defined as:

$$1 - (P_c/P_g)$$

A rounded grown region will have a low irregularity, whereas an irregularly shaped grown region will have a high degree of irregularity.

As shown in Fig 1C, the effective size, the degree of circularity, and the degree of irregularity are calculated as functions of the contrast of a grown region. Note that when the contrast is increased to a certain point, the effective size and the degree of irregularity increase abruptly, whereas the degree of circularity decreases abruptly. This transition indicates that the grown region has merged with the surrounding background adjacent to the nodule candidate. The grown region at the point just before this transition, as
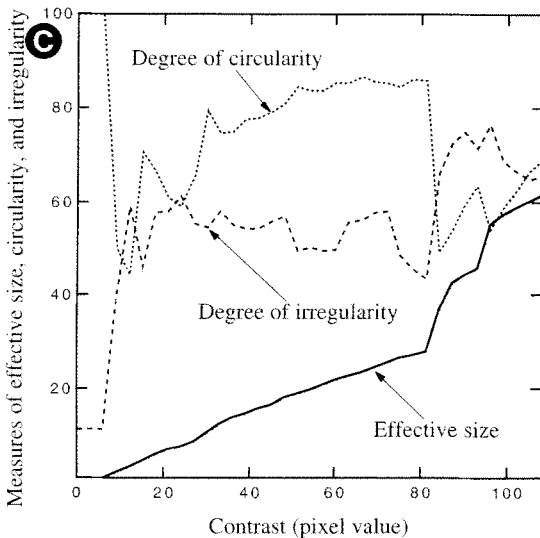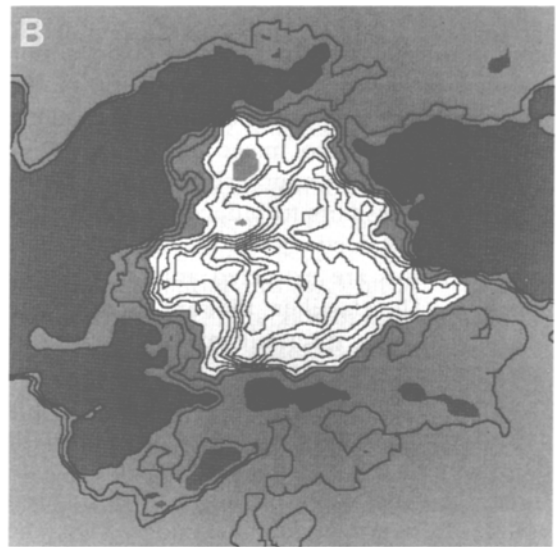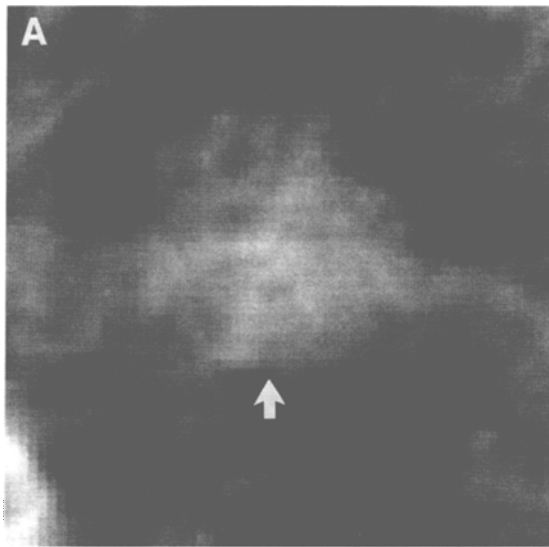
Fig 1. (A) Original image of an actual nodule (arrow indicates nodule); (B) the corresponding contour line image after region growing; (C) a diagram indicating relationships between the contrast of the grown region and degree of circularity (. . . .); degree of irregularity (----); and effective size (—).

illustrated by the light area in Fig 1B, appears visually to be similar to the actual shape and size of the nodule candidate in Fig 1A. Thus, we define this point as the transition point. The values of the image features at the transition point are used as input to the classifiers.

As the contrast increases from zero to its value at the transition point, the shape of the nodule candidate grows from a starting point to approximately the full size of the candidate. The average slopes of effective size, circularity, and irregularity within the contrast range from zero to the transition point define three additional features that are used for the input data to the classifiers.

Two additional features used in this study are derived from an edge gradient analysis.[34] For each nodule candidate, the magnitude and orientation of edge gradients are calculated for all pixels within a 50- × 50-pixel ROI, centered at the same pixel that was used previously for initializing the region-growing process. For an ROI that

contains a nodule and vessels, edge gradients are generally oriented in many different directions. Therefore, the distribution of the number of pixels in different directions is relatively uniform with a small standard deviation. However, for ROIs containing rib edges, the gradient-orientation shows a definite directionality, and the standard deviation of the number of pixels in different directions is usually large. Thus, we use the average edge gradient and the standard deviation of edge gradient also as features for the input to the classifiers.

The database of nodule candidates used in this study includes 23 nodules and 704 false-positive detections. Figure 3 shows the average values of the input features for these nodules and false positives. In general, the differences between nodules and false positives are very small for most of the input features. However, the contrast, the effective size, the degree of circularity, and the standard deviation of edge gradient show larger differences, relative to their
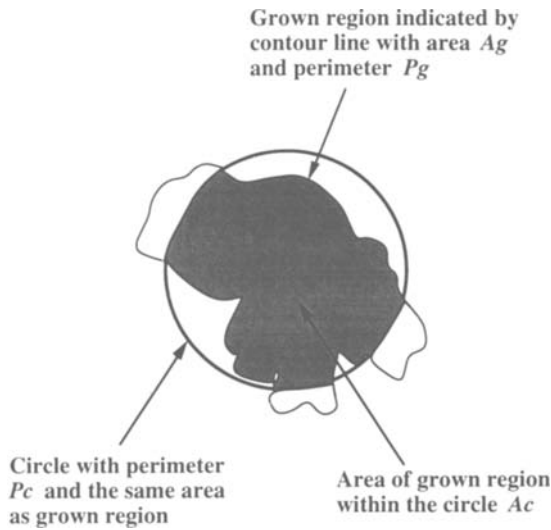
Fig 2. Illustration of definitions of parameters that were used to define three image features, ie, effective size, degree of circularity, and degree of irregularity for a nodule candidate.

standard errors, than the other features for the two classes of nodule candidates.

Neural Networks

A three-layer, feed-forward neural network is used in this study. Backpropagation and the generalized delta rule are used for the training algorithm.[35] A logistic function is used as the nonlinear activation function for each processing unit in the neural network, such that

$$o_{pj} = \frac{1}{1 + \exp\left\{-\left(\sum_i w_{ji} o_{pi} + \theta_j\right)\right\}}, \quad (2)$$

where o_{pj} is the jth element of the actual output pattern produced by the presentation of input pattern p, w_{ji} is the weight from the ith to jth units, and θ_j is the threshold of the jth unit. In the training process, the weights and thresholds are adjusted iteratively so that the difference between the output values and the desired results is minimized, according to the following rule:

$$\Delta w_{ji}(n + 1) = \eta(\delta_{pj} o_{pi}) + \alpha \Delta w_{ji}(n), \quad (3)$$

where n indicates the number of presentations of each input data pattern, which is also defined as iterations, η is the learning rate, δ_{pj} is the error signal, which is related to the difference between the output of the neural network and the target (ie, desired) output,[35] and α is a momentum term that determines the effect of past weight changes on the current direction of movement in weight space. Figure 4 shows the structure of the neural network that was used in this study. In the training process, when image features of a nodule or a false positive are provided to the input layer of the network, an output value of 1 or 0, respectively, is provided to the output unit as the desired value. The number of units in the hidden layer can be chosen according to the nature of input and output data. In this study, the

optimal performance, defined as having the greatest area under an ROC curve (A_z), is achieved by using eight hidden units and 1,000 training iterations (defined earlier).

Discriminant Analysis

On the basis of Bayesian statistics, signal detection theory, and an assumption of multidimensional normality of input data distributions, the optimal linear discriminant function[36] is given by (see Appendix):

$$D(\vec{x}) = (\vec{\mu}_s - \vec{\mu}_n)^t \sum\nolimits^{-1} \vec{x}, \quad (4)$$

where $\vec{x}$ is an input vector of feature values (nine, in this study), $\vec{\mu}$ is an average vector of a training database, and Σ is the 9×9 within-group variance matrix of the training database. Subscripts s and n denote two states of truth—signal and noise—that correspond to nodules and false-positive detections in this study, respectively. The discriminant function expressed by equation 4 is determined (trained) by calculating $\vec{\mu}$ and Σ from the training database.

Rule-Based Scheme

A rule-based scheme is developed by Matsumoto et al[37,38] to eliminate false-positive detections of the CAD scheme. This rule-based scheme is developed on the basis of the knowledge of experienced radiologists. The first step in this scheme is an analysis of the characteristics of the features of false-positive detections that occurred at rib crossings, rib and vessel crossings, end-on vessels, and aggregates of vessels. Threshold levels for the image features are chosen to remove as many false-positive detections as possible without loss of true nodules, by analyzing all cases in the database. Detailed description of this scheme can be found elsewhere.[34,35]
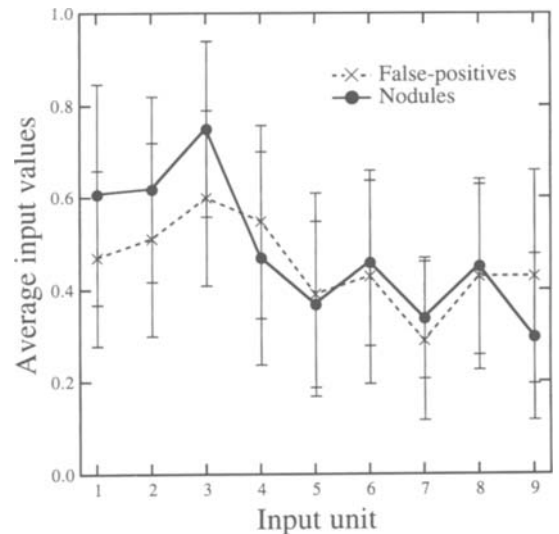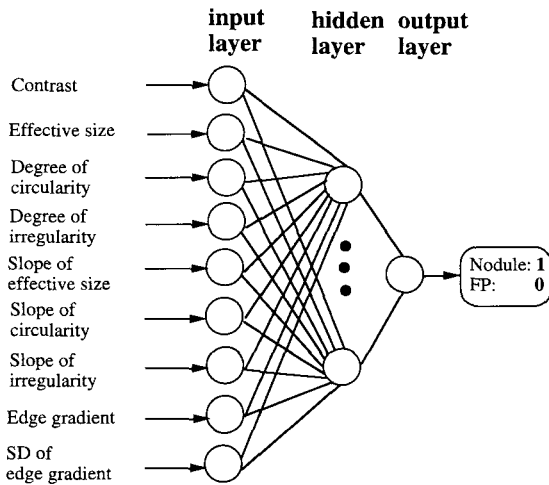

Fig 3. Average values of nine input features for cases used in this study, which include 23 nodules and 704 false positives. The nine input features are (from 1 to 9): contrast, effective size, degree of circularity, degree of irregularity, slope of effective size, slope of circularity, slope of irregularity, edge gradient, and standard deviation of edge gradient.

Fig 4. Structure of a three-layer, feed-forward neural network that was used for the detection of lung nodules based on nine input features.

Implementation of Classifiers

The CAD scheme, the neural network, and discriminant analysis are implemented on an IBM RISC 6000 POWER-station 560 (IBM, Purchase, NY). The rule-based scheme is implemented on a DEC VAX 3500 workstation.

Evaluation of Performance of Classifiers

In general, an evaluation of decision-making performance in terms of correct detections or classifications depends on the threshold value used. In this study, ROC analysis is used to objectively assess the overall performance of the classifiers by varying the threshold value. In ROC analysis, a curve that relates true-positive fraction (TPF) to false-positive fraction (FPF) is obtained instead of the single combination of TPF and FPF obtained from yes/no decisions based on a particular threshold value, which corresponds to only one point on the ROC curve. The area under an ROC curve (A_z) is used to summarize the performance of each classifier.

To estimate the ability of the neural-network and discriminant-analysis classifiers to generalize from the cases on which they are trained and make decisions concerning cases that have not been included in their training, we use both a round-robin method—also known as a leave-one-out method—and a jackknife method to evaluate the performance of the classifiers. In the round-robin method, all but one case in a data set is used to train a classifier. The single case that has been left out is then used to test the classifier. This procedure is repeated so that each case in the data set is used once as a testing case. In the jackknife method, one half of the cases in a data set are randomly selected as a training set and the other half are used as a testing set. Various training/testing sets can be generated by use of a random-number generator. The LABROC4 algorithm developed by Metz et al[29] is used to fit ROC curves to the continuous data that are obtained from the output unit of

the neural network and from the discriminant function (equation 4).

RESULTS

Performance of the Classifiers Based on Selected Data Sets

To evaluate the performance of classifiers in distinguishing nodules from false-positive detections, we used the round-robin and jackknife methods, as discussed earlier. For the round-robin method, we created three training data sets from the database. Each of the three data sets included all of the 23 nodules and 115 randomly selected false-positive detections. This round-robin test was used to evaluate the neural-network classifier and the discriminant-analysis classifier for each of the three training data sets. Note that the number of false-positive cases in each training data set was five times larger than the number of nodule cases. Therefore, we modified the training process of the neural network such that the number of training iterations for the nodules was five times greater than that for the false-positive candidates. This modification of the training process resulted in equal numbers of data entries for nodules and for false positives. ROC analysis was used to independently analyze the results obtained from the three training data sets. Averaged ROC curves for the two classifiers are shown in Fig 5. The neural network, which achieved an A_z value of 0.79, performed somewhat better than discriminant analysis ($A_z = 0.74$), although the difference was not statistically significant (two-tailed P value = .182 and 95% confidence interval[29] was [−0.110, 0.216]). ·

Alternatively, the jackknife method was used to evaluate the performance of the neural network. In the jackknife test, 10 different data sets were selected for the evaluation. Each of the data sets included all of the 23 nodule cases and a number of randomly selected false-positive cases. ROC analysis was performed separately on each data set, and the A_z values obtained from the 10 different data sets were averaged.

We used three different ratios of the number of false-positives to the number of nodules in these evaluations, namely, 1, 5, and 20. The
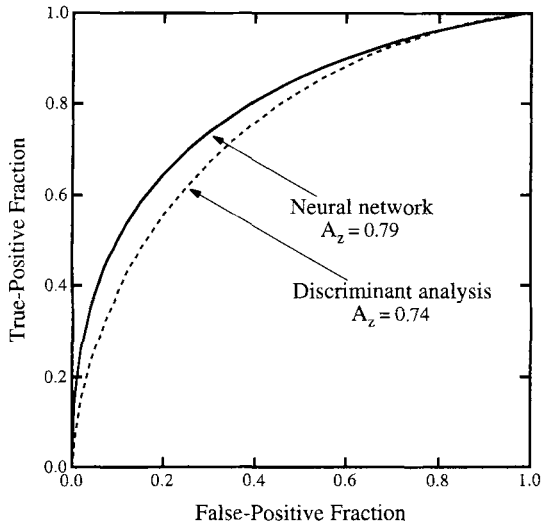
Fig 5. ROC analysis of the performance of the neural network, and discriminant analysis for distinguishing lung nodules from false positives based on results of three data sets using the round-robin method.

number of the nodules was kept the same at 23 for all three ratios used, whereas the number of false positives was varied accordingly. The average A_z values for the neural network and their standard deviations obtained with the three different ratios in the training data set are shown in Fig 6. The ratio of the number of
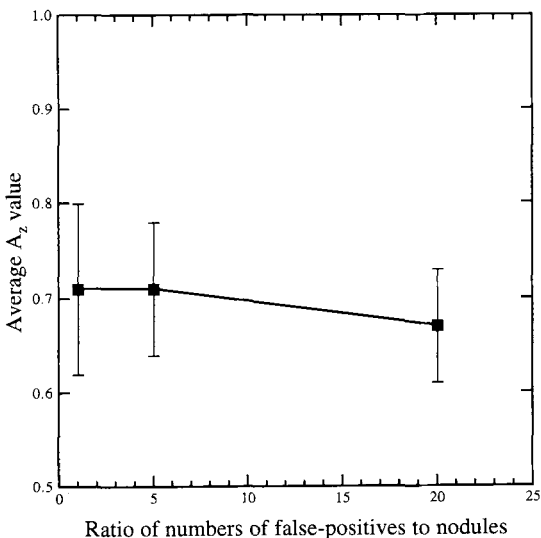

Fig 6. Performance of the neural network based on the jackknife method. Effect of the ratio of the number of false positives to the number of nodules in the training and testing data sets; A_z values for each of the ratios were averaged over 10 randomly selected sets.

false-positive cases to the number of nodule cases did not affect the performance of the neural network significantly (two tailed P value = .32 for the difference of A_z values between ratios 1 and 20), although the A_z value appeared to decrease slightly when the ratio increases. The A_z values obtained by the jack-knife method were slightly smaller than those obtained with the round-robin method, shown in Fig 5 because, in general, the round-robin method used a given number of cases more efficiently than the jackknife method in training a classifier. The variance of the A_z value decreased as the ratio increased because of the increase in the total number of cases involved.

Elimination of False-Positive Detections

To evaluate the ability of the classifiers to eliminate false-positive detections, the neural network and discriminant analysis were first trained by using a training data set and then tested using the entire database. This approach was used for reasons discussed later. Each of the training data sets contained all of the 23 nodule cases and 115 false-positive cases selected randomly from the database. Figure 7, A and B, show histograms of the output obtained from the neural network and from the discriminant function, respectively, after training on one of the training data sets. With the neural network, ~80% of false-positive detections could be eliminated without loss of true nodules if a threshold level of 0.9 was selected. However, discriminant analysis could eliminate only about 30% of false-positive detections without a loss of true nodules (threshold value = 0.4). Table 1 summarizes the performance of the two classifiers in eliminating false-positive detections when they were trained with each of the three training data sets.

ROC analysis was performed on the outputs obtained from the neural network and the discriminant function for the three different training data sets. The average ROC curves obtained in this way are shown in Fig 8. Note that when the false-positive fraction was 0.2, the averaged true-positive fraction of the neural network was almost 1.0. Therefore, about 80% of false-positive detections could be eliminated by the neural network without substantial loss of true-positive detections.
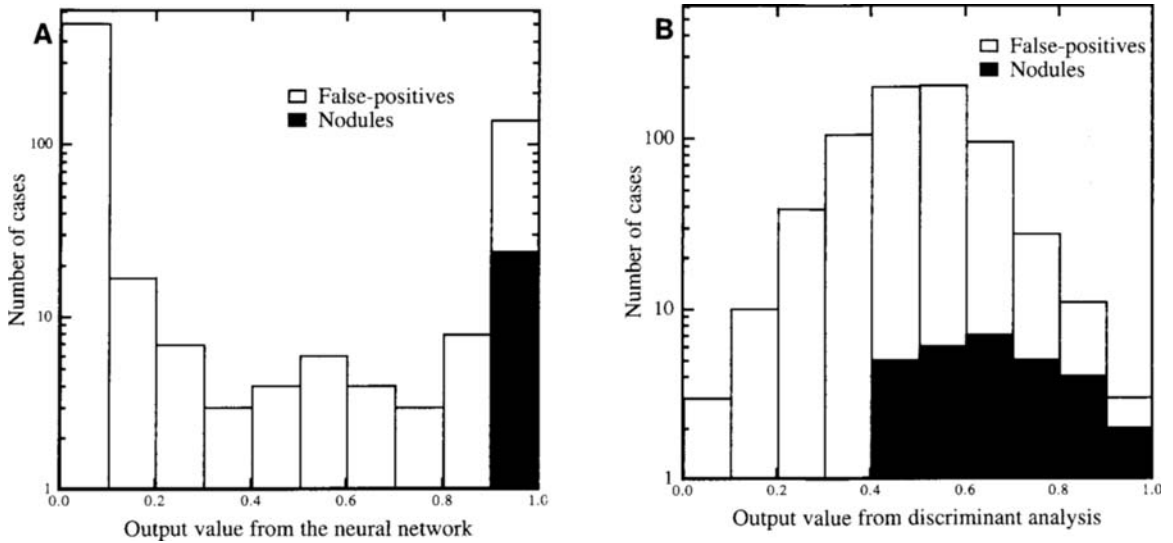
Fig 7.   Histograms of the output values from the classifiers, trained by using data set A in Table 1, for elimination of false-positive detections; (A) the neural network; (B) the discriminant analysis.

The rule-based scheme[37,38] was developed to eliminate false-positive detections based on the analysis of various types of false positives. Based on the database of 30 normal and 30 abnormal images that included 23 nodules and 704 false positives, the rule-based scheme was able to reduce the number of false-positive detections to 277 without loss of true nodules. Therefore, ~60% of the false-positive detections were eliminated by the rule-based scheme. Table 1 shows that the neural network was able to perform better than the rule-based scheme (discussed earlier) in eliminating false-positive detections, whereas the performance of discriminant analysis was poorer than that of the rule-based scheme. However, in comparing the rule-based scheme with the neural network and discriminant analysis, it is important to notice that the rule-based scheme was established on the basis of all of the cases in the database, whereas the neural network and discriminant

analysis were trained based on a training data set and then tested against a larger case set that included the training set. Hence, the obtained results were partially biased in favor of the rule-based scheme.

Performance of the Neural Network with Two Groups of False Positives

The neural network and the rule-based scheme were two essentially independent classifiers that may distinguish nodules from false positives in different ways. Therefore, the types of false positives eliminated by the two classifiers could be different. The neural network was applied to distinguish both nodules from false positives that were identified as such by the rule-based scheme, and also nodules from false positives that were not identified as such by the rule-based scheme. The performance of the neural network was evaluated by using the round-robin method in connection with two

Table 1. Comparison of Performances Among Classifiers in Elimination of False-Positive Detections

Classifier	Discriminant Analysis			Neural Network		
	A	B	C	A	B	C
Remaining nodules after test	23	23	23	23	23	23
Remaining FP after test	495	517	504	123	67	101
Nodules lost	0	0	0	0	0	0
% of FP eliminated	30	27	28	83	91	86

Total number of nodules and false positives in the database are 23 and 704, respectively. A, B, and C indicate three different data sets used in the training of the classifiers.

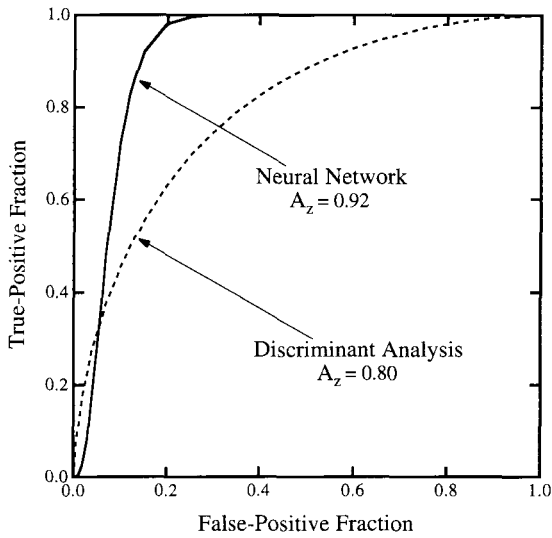Abbreviation: FP, false-positive detections.

Fig 8. ROC analysis for eliminating false-positive detections using the neural network and the discriminant analysis that were trained by using three data sets; ROC curves were average results based on the three training data sets.

data sets: one containing all of the nodules and a number of false positives identified by the rule-based scheme, and another containing all of the nodules and a number of false positives that were not ·identified by the rule-based scheme. ROC analysis of the results of the two round-robin tests yielded A_z values of 0.74 and 0.72, respectively. Thus, it appeared that there was no substantial difference between the performance of the neural network in identifying these two groups of false positives, although the neural network was able to recognize the false positives identified by the rule-based scheme slightly better than the false positives that were not identified by the rule-based scheme. This result suggests that the neural-network and rule-based classifiers distinguish between nodules and false positives in substantially different ways. Therefore, some of the false positives that were not eliminated by one of the classifiers may be eliminated by the other.

Performance of Combining the Neural Network, Discriminant Analysis, and the Rule-Based Scheme

As we discussed in the previous section, the false-positive detections eliminated by the rule-based scheme were largely distinct from those eliminated by the neural network. Therefore, one can expect that the number of false-positive

detections might be reduced even further by combining two or three of the classifiers.

The results obtained from two pairings of the classifiers are summarized in Table 2. The neural network and the discriminant function were first trained individually by using selected training data sets. To combine discriminant analysis and the neural network, we applied one classifier after another. When the neural network was trained on the training data set A, the number of false-positive detections could be reduced to 112, as shown in Table 2, by combining the neural network with discriminant analysis. Note that the neural network and discriminant analysis could reduce the number of false-positive detections only to 123 and 495, respectively, if they were applied individually. We also combined the neural network with the rule-based scheme to reduce the number of false-positive detections. We first applied the rule-based scheme so that the number of false positives was reduced to 277. The neural network, which had been trained by training data set A, was then applied to the remaining 277 false positives. As a result, all of the true nodules were retained, but 668 (96%) of the original 704 false positives were eliminated. ROC analysis yielded an A_z value of 0.93 for this combination of classifiers. The ROC curves of the two combined classifiers, averaged over three different training data sets, are shown in Fig 9.

We found that the number of false-positive detections could not be reduced further by combining all three classifiers because the false-positive detections eliminated by discriminant analysis were eliminated also by the rule-based scheme. This was in contrast to the false-positive detections that were eliminated by the rule-based scheme and by the neural network, which were largely distinct.

One should note that the results in Tables 1 and 2 and in Figs 8 and 9 were for elimination of false-positive detections by automated classifiers that had been trained by a training data set. Each of the training data sets contained all of the nodules and a portion of the false-positives that were used for testing. Therefore, our results do not provide an unbiased assessment of the ability of the neural network and discriminant analysis to recognize nodules (all of which

Table 2. Performance in Eliminating False-Positive Detections: Combinations of Classifiers

Combination	Rule-Based Scheme/ Neural Network			Discriminant Analysis/ Neural Network		
	A	B	C	A	B	C
Remaining nodules after test	23	23	23	23	23	23
Remaining FP after test	36	24	46	112	66	98
Nodules lost	0	0	0	0	0	0
% of FP eliminated	96	97	94	84	91	86

Total number of nodules and false-positives in the database are 23 and 704, respectively. A, B, and C indicate three different data sets used in the training of the classifiers.

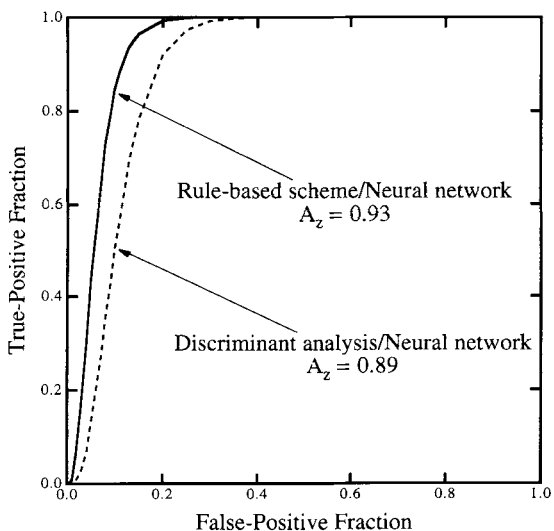Abbreviation: FP, false-positive detections.

were in both the training sets and the testing set), but rather indicate the ability of the network to identify false-positive detections.

DISCUSSION

We found that an artificial neural network can be trained to distinguish between nodules and false-positive detections, and to serve as a useful tool in eliminating false-positive detections in the CAD scheme for detection of lung nodules in chest images. Our neural network achieved an average A_z value of 0.79 when it was tested by the round-robin method on three selected data sets. Moreover, after being trained by selected training sets, the neural network was able to eliminate more than 83% of the false-positive detections reported by the CAD scheme,

Fig 9. ROC curves for eliminating false-positive detections combining the neural network with the discriminant analysis and the rule-based scheme. The neural network and the discriminant analysis were trained by using three data sets. ROC curves were average results based on the three training data sets.

although this level of performance should be considered as the upper limit, as will be discussed later. The performance of the neural networks, a nonlinear and nonalgorithmic approach, appeared to surpass that of linear discriminant analysis in distinguishing nodules from false-positive detections.

The prevalence of cases in the two classes—nodules and false-positive detections—was unbalanced in our database. We modified the iteration scheme of the training process for the two classes of input data to compensate for this unbalanced prevalence so that the total number of training iterations for nodules and false positives were made equal. An alternative way to deal with unbalanced prevalence is to modify the generalized delta rule in the training algorithm so that the error term is not weighted by the number of cases in each class. The effect of adjusting the error term on the performance of the neural network may be similar to the approach that we used in the study reported here, but that possibility requires further investigation. It is important to note that increasing the number of iterations or using a weighted error term for nodule cases may help the neural network to learn the patterns of nodules and false positives in a more balanced way when the prevalence of the two classes in the database is unbalanced. However, increasing the number of iterations is not equivalent to increasing the actual number of cases in the training database, and it will not increase the ability of the neural network to generalize to the population of cases at large.

In this study, we first used round-robin and jackknife methods to evaluate the performance of the neural network and discriminant analysis in distinguishing lung nodules from false positives. However, in the second phase of this

study, the neural network and discriminant analysis were trained to eliminate false-positive detections of the CAD scheme by using a training data set that was part of the testing set. It is important to note that the round-robin and jackknife methods provide an unbiased test of the classifiers' ability to learn from training cases and generalize the learned knowledge to a different set of cases in the testing process. Therefore, results from these tests more reliably indicate the level of performance that the neural network or discriminant analysis can achieve when it is applied in practical situations. Evaluation of the classifiers in terms of their ability to eliminate false positives from the CAD scheme, on the other hand, does not require an unbiased test of the neural network's ability to detect nodules. Ideally, we would train the network with a larger data set and apply the trained network to an entirely different data set to evaluate its performance. However, we were not able to split our database into completely separate training and testing data sets because the number of confirmed nodules in our database currently is very small. Nevertheless, our results indicate an upper limit on the number of false positives that the neural network would be able to eliminate if trained adequately to recognize all or a majority of the nodules.

This upper limit on the number of false positives that the neural network would be able to eliminate can be explained by the example shown in Fig 7A. The neural network was trained by all of the nodules and a portion of the false positives in the database and then tested on all of the cases in the database. All of the nodules were recognized by the neural network, and its output values were larger than 0.9 for all nodule cases. With its threshold level set at 0.9, the neural network was able to eliminate 83% of the false positives without losing any nodules. If the neural network had not been trained to recognize all of the nodules, the output values of some of the nodule cases would be less than 0.9. As a result, the threshold value would have to be lowered to avoid losing true nodules, and the neural network would not be able to eliminate as many false positives.

The majority of false-positive detections eliminated by the neural networks were not included in the training data set. Therefore, our results

suggest that the neural network may be a useful tool for reducing the number of false-positive detections from our CAD scheme. The false-positive detections that were eliminated by the neural network were, in general, different from those eliminated by the rule-based scheme and discriminant analysis. We found that a combination of the neural network and the rule-based scheme eliminated 96% of the false-positive detections reported by the CAD scheme. This result suggests that after incorporation of a properly trained neural network and the rule-based classifier, our current CAD scheme would potentially achieve a sensitivity of 70% for detecting lung nodules in chest radiographs with only ~ 0.5 false positives per image.

Nine image features were extracted from chest images and used as input data to the classifiers in this study. These nine features may not have included all of the useful information that can be used in detecting nodules. Other features such as the location of a nodule candidate in a chest image and the density of the candidate may be useful. Expanding the input feature set and its effect on the performance of the neural network remains to be investigated in the future.

APPENDIX: DERIVATION OF THE DISCRIMINANT FUNCTION

Let $\vec{x}$ be a vector of random variables. A decision rule for determining whether $\vec{x}$ belongs to the class s (signal) or the class n (noise) can be expressed as:

$$\text{If} \begin{cases} \dfrac{P(\vec{x}|s)}{P(\vec{x}|n)} > c', & \text{assign } \vec{x} \text{ to s.} \\[2ex] \dfrac{P(\vec{x}|s)}{P(\vec{x}|n)} < c', & \text{assign } \vec{x} \text{ to n.} \end{cases}$$

By taking logarithm of both sides of the above inequalities, we can define a discriminant function

$$D(\vec{x}) = \ln\left(\frac{P(\vec{x}|s)}{P(\vec{x}|n)}\right) + c, \qquad (A1)$$

where c is a constant such that the decision rule

becomes

$$\text{If} \begin{cases} D(\vec{x}) > C_0, & \text{assign } \vec{x} \text{ to s,} \\ D(\vec{x}) < C_0, & \text{assign } \vec{x} \text{ to n,} \end{cases}$$

where C_o is a critical value.

Assuming N-dimensional normal distribution for $\vec{x}$, one can write

$$P(\vec{x}|s) = \left[(2\pi)^{N/2} \left| \sum_s \right|^{1/2} \right]^{-1}$$

$$\cdot \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu}_s)^t \sum_s{}^{-1}(\vec{x} - \vec{\mu}_s) \right]. \quad \text{(A2)}$$

$$P(\vec{x}|n) = \left[(2\pi)^{N/2} \left| \sum_n \right|^{1/2} \right]^{-1}$$

$$\cdot \exp\left[-\frac{1}{2}(\vec{x} - \vec{\mu}_n)^t \sum_n{}^{-1}(\vec{x} - \vec{\mu}_n) \right]. \quad \text{(A3)}$$

Where $\vec{\mu}$ is an average vector and $\sum$ is a variance matrix. If we assume that the distributions of signal and noise have the same variances and covariances, then

$$\sum_s = \sum_n = \sum. \quad \text{(A4)}$$

Substituting equations A2, A3, and A4 into A1,

we have

$$D(\vec{x}) = \ln\left(\frac{P(\vec{x}|s)}{P(\vec{x}|n)} \right) + c$$

$$= \frac{1}{2}\left[(\vec{x} - \vec{\mu}_n)^t \sum{}^{-1} (\vec{x} - \vec{\mu}_n) \right.$$

$$\left. - (\vec{x} - \vec{\mu}_s)^t \sum{}^{-1} (\vec{x} - \vec{\mu}_s) \right] + c$$

$$= (\vec{\mu}_s - \vec{\mu}_n)^t \sum{}^{-1} \vec{x}. \quad \text{(A5)}$$

It can be shown that this discriminant function is equivalent to an ideal statistical decision maker who uses likelihood ratio as a decision variable for input data that have the same variances and covariances for positive and negative Gaussian distributions.[4] If $\Sigma_s \neq \Sigma_n$, the linear classifier that maximizes the area under the ROC curve (A_z), is given by[4]

$$D(\vec{x}) = (\vec{\mu}_s - \vec{\mu}_n)^t \left(\sum_s + \sum_n \right)^{-1} \vec{x}. \quad \text{(A6)}$$

ACKNOWLEDGMENT

REFERENCES

1. Rumelhart DE, McClellend JL (eds): Parallel Distributed Processing. Cambridge, MA, MIT, 1986

2. Hecht-Nielsen R: Neurocomputing. Reading, MA, Addison-Wesley, 1990

3. Boone JM, Sigilitto VG, Shaber GS: Neural networks in radiology: An introduction and evaluation in a signal detection task. Med Phys 17:234-241, 1990

4. Wu Y, Doi K, Metz CE, et al: Simulation studies of data classification by artificial neural networks: Potential applications in medical imaging and decision making. J Digit Imaging 6:117-125, 1993

5. Wu Y, Doi K, Giger ML, et al: Computerized detection of clustered microcalcifications in digital mammograms: Applications of artificial neural networks. Med Phys 19:555-560, 1992

6. Wu Y, Doi K, Giger ML, et al: Application of neural networks in mammography for the diagnosis of breast cancer. Proc SPIE Imaging Technol Appl 1778:19-27, 1992

7. Wu Y, Giger ML, Doi K, et al: Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer. Radiology 187:81-87, 1993

8. Asada N, Doi K, MacMahon H, et al: Potential usefulness of an artificial neural network for differential diagnosis of interstitial lung diseases: Pilot study. Radiology 177:857-860, 1990

9. Gross GW, Boone JM, Greco-Hunt V, et al: Neural networks in radiologic diagnosis: II. Interpretation of neonatal chest radiographs. Invest Radiol 25:1017-1023, 1990

10. Wu Y, Doi K, Giger ML, et al: Detection of lung nodules in digital chest radiographs: Artificial neural networks and discriminant analysis. Presented at the 34th annual meeting of AAPM, Calgary, Canada, August 23-27, 1992

11. Wu Y, Doi K, Giger ML, et al: Detection of lung nodules in digital chest radiographs: Comparison of artificial neural networks and discriminant analysis. Presented at the 78th scientific assembly and annual meeting of RSNA, Radiology 185 (P):156, 1992 (suppl, abstr)

12. Lo SCB, Freedman MT, Lin JS, et al: Profile-matching techniques and neural classifier for automatic nodule detections in pulmonary radiographs. Presented at the 78th scientific assembly and annual meeting of RSNA, Radiology 185 (P):196, 1992 (suppl, abstr)

13. Garg S, Floyd CE: Neural network localization of pulmonary nodules on digital chest radiographs. Presented at the 78th scientific assembly and annual meeting of RSNA, Radiology 185 (P):157, 1992 (suppl, abstr)

14. Kim JH, Min BG, Han MC, et al: Computer-assisted detection of lung nodules by using artificial neural net. Presented at the 78th scientific assembly and annual meeting of RSNA, Radiology 185 (P):156, 1992 (suppl, abstr)

15. Lo SCB, Freedman MT, Lin JS, et al: Automatic lung

nodule detection using profile matching and back-propagation neural network techniques. J Digit Imaging 6:48-54, 1993

16. Chiou YS, Lin JS, Fleming Lure YM, et al: Shape feature analysis using artificial neural networks for improvements of hybrid lung nodule detection (HLND) system. Image Processing, Medical Imaging VII, SPIE Medical Imaging 1993 Symposium, Newport Beach, CA, February 13-19, 1993

17. Chiou YS, Lin JS, Fleming Lure YM, et al: Neural network based hybrid lung nodule detection (HLND) system. 1993 IEEE International Conference on Neural Network, San Francisco, CA, March 28-April 1993.

18. Lo JY, Floyd CE, Bowsher JE, et al: Spatially varying scatter estimation in portable chest radiography with an artificial neural network. Presented at the 78th scientific assembly and annual meeting of RSNA, Radiology 185 (P):300, 1992

19. Floyd CE, Tourassi GD: An artificial neural network for lesion detection on single-photon emission computed tomographic images. Invest Radiol 28:667-672, 1992

20. Chan KK, Hayrapetian AS, Lau CC, et al: Neural network segmentation of double-echo MR images. Presented at the 78th scientific assembly and annual meeting of RSNA, Radiology 185 (P):157, 1992

21. Forrest JV, Friedman PJ: Radiologic errors in patients with lung cancer. West J Med 134:485-490, 1981

22. Giger ML, Doi K, MacMahon H: Image feature analysis and computer-aided diagnosis in digital radiography, III: Automated detection of nodules in peripheral lung fields. Med Phys 15:158-166, 1988

23. Giger ML, Doi K, MacMahon H, et al: Pulmonary nodules: Computer-aided detection in digital chest images. RadioGraphics 10:41-51, 1990

24. Giger ML, Ahn K, Doi K, et al: Computerized detection of pulmonary nodules in digital chest images: Use of morphological filters in reducing false-positive detections. Med Phys 17:861-865, 1990

25. Yoshimura H, Giger ML, Doi K, et al: Computerized scheme for the detection of pulmonary nodules A nonlinear filtering technique. Invest Radiol 27:124-129, 1992

26. Katsuragawa S, Doi K, MacMahon H: Image feature analysis and computer-aided diagnosis in digital radiography: Detection and characterization of interstitial disease in digital chest radiography. Med Phys 15:311-319, 1988

27. Ishida M, Kato H, Doi K, et al: Development of a new digital radiographic image processing. Proc SPIE 347:42-48, 1982

28. Metz CE: ROC methodology in radiologic imaging. Invest Radiol 21:720-733, 1986

29. Metz CE, Shen JH, Herman BA: New methods for estimating a binormal ROC curve from continuously-distributed test results. Presented at the 1990 Joint Meetings of the American Statistical Society and the Biometric Society, Anaheim, CA, August 1990

30. Metz CE: Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 24:234-245, 1989

31. Metz CE, Wang PL, Kronman HB: A new approach for testing the significance of differences between ROC curves measured from correlated data, in Deconinck F (eds): Information Processing in Medical Imaging. Nijhoff, The Hague, The Netherlands, 1989, pp 432-445

32. Metz CE: Quantification of failure to demonstrate statistical significance: The usefulness of confidence intervals. Invest Radiol 28:59-63, 1993

33. Pratt WK: Digital Image Processing. New York, NY, Wiley, 1978

34. Sanada S, Doi K, MacMahon H: Image feature analysis and computer-aided diagnosis in digital chest radiography: Automated delineation of posterior ribs in chest images. Med Phys 18:964-971, 1991

35. Rumelhart DE, Hinton GE, Williams RJ: Learning internal representations by error propagation, in Rumelhart DE, McClellend JL (eds): Parallel Distributed Processing. Cambridge, MA, MIT, 1986, pp 318-362

36. Lachenbruch PA, Discriminant Analysis. New York, NY, Hafner, 1975

37. Matsumoto T, Yoshimura H, Giger ML, et al: Potential usefulness of computerized nodule detection on screening programs for lung cancer. Invest Radiol 27:471-475, 1992

38. Matsumoto T, Yoshimura H, Doi K, et al: Image feature analysis of false-positive diagnoses produced by automated detection of lung nodules. Invest Radiol 27:587-597, 1992