Computer-aided Diagnosis of Pulmonary Nodules in Thoracic Computed Tomography

by

Ted Win Way

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biomedical Engineering)
in The University of Michigan
2008

Doctoral Committee:

Professor Heang-Ping Chan, Co-chair
Professor Jeffrey A. Fessler, Co-chair
Professor Charles R. Meyer
Professor Douglas C. Noll
Associate Professor Berkman Sahiner
Research Associate Professor Lubomir M. Hadjiyski

To my family

Acknowledgements

I am grateful for the privilege of working in the lab of my advisor, Dr. Heang-Ping Chan, along with co-advisors Dr. Berkman Sahiner and Dr. Lubomir Hadjiiski. Under their tutelage and personal attention, I have grown as a researcher, learning to read papers critically, design experiments and interpret results, and explore ways to advance the field of computer-aided diagnosis. In addition, they have provided valuable guidance on the development of computer vision techniques for CAD. It is difficult to list all the ways they have been there for me in my development in such a few short sentences. I am also thankful for my Engineering co-advisor, Dr. Jeff Fessler, as he has continued to provide advice and discussion. From that first academic advising meeting, to discussions about the research, to presentations at group meetings, he has been a source of encouragement and understanding. I enjoyed taking the medical imaging and image processing courses with Dr. Doug Noll, whose ultrasound assignment turned out to be an image of a smiley face. I also thank Dr. Chuck Meyer for his image processing expertise, enthusiasm, and support in serving on this dissertation committee and previous master's thesis and qualifying exam committees.

I thank the faculty and post-doctoral research fellows in the lab for their friendship and discussions: Drs. Yi-ta Wu, Jun Ge, Ziazheng Shi, Jing Cui, Chuan Zhou, Yiheng Zhang, and Jun Wei. I was glad to make this journey with the support of fellow graduate students Desmond Yeo, Yingying Zhang-O'Connor, Dan Ruan, Kim Khalsa, Somesh Srivastava, Rongping Zheng, and Wayne Fung.

I'm grateful for the fellowship of the Ann Arbor Chinese Christian Church and the Daida and Thomas families. Finally, I thank God for the love and encouragement of Alice and my mother, father, and sister Sharon.

This research is supported by USPHS grant CA 93517. Chapter 2 was published in *Medical Physics* [1] and Chapter 3 was published in *Physics in Medicine and Biology* [2].

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Motivation

Lung cancer is the leading cause of cancer death in the United States, causing an estimated 160,400 deaths in 2007. At the time of diagnosis, most patients already present with advanced disease. Despite advances in treatment and diagnosis, the five-year overall survival rate is only 15% [3]. As for earlier detection, the "serendipitous discovery of lung cancer in asymptomatic people is currently the principal way in which stage I lung cancer is detected" [4]. Thus, there is great interest in determining whether routine screening can improve early detection and reduce the mortality rate. Previous trials in the 1970's to screen for lung cancer with chest X-ray and sputum (mucus or phlegm) analysis did not show significant reduction in mortality [5].

Currently, researchers are actively investigating screening with computed tomography (CT), as CT has been shown to have higher sensitivity in detecting small lung nodules compared to chest X-ray [6-12]. This suggests that CT screening has a strong potential for improving the likelihood of detecting lung cancer at an earlier and potentially more curable stage [13]. If CT screening is recommended, however, it would also exacerbate already mounting challenges of using CT for detection and diagnosis, namely, an ever increasing number of slices for interpretation. As more nodules are detected, more nodules need to be followed-up and managed. Despite the increasing spatial resolution of CT, assessing the likelihood of malignancy of nodules by visual inspection is difficult. This may be a reason why as many as 50% of nodules resected at surgery are benign [12], emphasizing the need to provide radiologists with tools to characterize nodules accurately and to handle large data sets.

1.1.1 Computer-aided Diagnosis (CAD)

Computer-aided diagnosis (CAD) may serve as a second reader by analyzing nodules and providing a malignancy estimate using computer vision and machine learning techniques. CAD may address some of the issues in lung nodule characterization, such as the increasing demand on radiologist's time caused by the increasing data volume, radiologist fatigue or distraction, and differences in radiologists' experience. Computers are playing an increasingly large role in radiology. In conventional radiography, x-ray images were recorded on screen-film systems, while today's radiologists view digital radiographs on display monitors. Computers have been vital in the development of medical imaging technology -- without computerized reconstruction, CT and MRI imaging would not be possible. The next role for computers may be in the interpretation of images.

1.2 Problem Statement

The goal of this dissertation is the investigation of the various components and development of an automated CAD system to assist radiologists in the classification (characterization) of lung nodules on helical multi-slice CT scans. A CAD system comprises segmentation, feature extraction, feature selection, and classification components. We aim to develop an effective CAD system that will assist radiologists in assessing lung nodule malignancy.

1.3 Summary of Contributions

The main contributions of this dissertation are summarized as follows:

- An extension of the 2D active contour model [14, 15] to include 3D information by adding three new energy terms to the cost function (Chapter 2) [16].
- An approach to search for segmentation parameters by using classification accuracy as quantified by the area under the receiver operating characteristic curve, A_z (Chapter 2) [1, 17].
- A study on the dependence of the performance of the CAD system on the primary versus metastatic status of lung nodules [18].

- A set of performance metrics to evaluate segmentation effectiveness (Chapter 2, Appendix 2) [1].
- An investigation of the effects of CT scanning parameters on segmentation performance for synthetic lung nodules in a chest phantom (Chapter 3) [2, 19] that may aid in providing a context to interpreting extracted interval change volume features from serial CT scans [20, 21].
- A simulation study to investigate the biases and variances of CAD systems built with various feature selection and classification methods using data drawn from representative feature space distributions (Chapter 4).
- The design of profile and gradient field features to describe the nodule surface, in addition to the use of morphological features and texture features to characterize the nodule (Chapter 5) [22].
- A two-loop leave-one-case-out resampling scheme to reduce the optimistic test A_z bias in the one-loop scheme (Chapter 5) [22].
- A preliminary observer study with six fellowship-trained thoracic radiologists to demonstrate that CAD can provide statistically significant improvement on radiologists' diagnostic accuracy (Chapter 6) [23, 24].

1.3.1 Other Contributions

Other contributions not detailed in this dissertation utilized the 3D active contour for various tasks. One was to study the effect of segmentation on the accuracy of nodule detection [25]. The others used the 3DAC to automatically segment phantom nodules for analysis of CT number. These include:

- A study of the accuracy of CT number estimation for lung nodules in multi-detector CT scans [26, 27].
- An analysis of the effect of patient body size and lung size on CT number accuracy for lung nodules [28].
- An investigation of single- and dual-energy CT calibration lines for assessing the calcium content of nodules [29].

1.4 Dissertation Outline

This dissertation focuses on the investigation of the various components of the CAD system and its overall performance and effectiveness. We introduce a preliminary CAD system in Chapter 2 that concentrated on automatic segmentation and classification of lung nodules. We extended the 2DAC model to include 3D curvature and 3D gradient energy terms, in addition to adding a lung mask energy term. The leave-one-case-out test classification performance was used to guide the simplex search for the best 3DAC segmentation parameters. Morphological and run-length statistics (RLS) features from the rubber band straightening transform (RBST) image were extracted. The segmentation performance of the 3DAC model trained with our data set was evaluated with 23 nodules available from the Lung Image Database Consortium (LIDC).

In Chapter 3, we describe the effects of CT scanning and reconstruction parameters on automated segmentation and volumetric measurements of nodules in CT images. We used phantom nodules of known sizes so that segmentation accuracy could be quantified in comparison to ground-truth volumes. CT scans of the phantom were acquired with a 16-slice scanner at various tube currents, pitches, and fields-of-view, and then reconstructed to various slice thicknesses. This study provided insight to automated lesion segmentation and volumetric measurements. The results may provide a guide to analyze interval changes in nodule volumes using serial CT scans. This will be useful because radiologists use nodule growth as an indicator of malignancy.

One obstacle all CAD researchers face is the limitation of available data. Ground truth for lung nodule malignancy is established by two-year follow-up, positron emission tomography, or biopsy, a potentially risky procedure for patients. We explored the finite sample size effect on various feature selection and classification methods in Chapter 4. In this simulation study, we generated data from known Gaussian distributions and compared various feature selection and classifier combinations. The sample size effects on the bias and variance of classifier performance were investigated.

We improved the CAD system by adding newly designed features to describe the nodule surface in Chapter 5. In addition to morphological and texture features, we designed new gradient field features to quantify the lung nodule boundary. The effects of two demographic features, age and gender, were also investigated. A “two-loop” leave-one-out resampling scheme was developed to estimate the test performance of the CAD system. We also compared the performance between the linear discriminant analysis (LDA) and support vector machine (SVM) with various kernels and parameters for a range of orders of features selected by principal component analysis.

In Chapter 6, we describe the observer study that was conducted with receiver operating characteristic (ROC) methodology to evaluate the effect of CAD on radiologists' characterization of lung nodules. Six fellowship-trained thoracic radiologists served as readers. All 6 radiologists achieved higher performance with CAD; four reaching statistical significance ($p<0.05$). This shows that CAD has the potential to improve radiologists' accuracy in assessing the likelihood of malignancy of lung nodules on CT.

Chapter 2
Segmentation and Classification Using 3D Active Contours

2.1 Abstract

We developed a computer-aided diagnosis (CAD) system to classify malignant and benign lung nodules found on CT scans. A fully automated system was designed to segment the nodule from its surrounding structured background in a local volume of interest (VOI) and to extract image features for classification. Image segmentation was performed with a 3D active contour (AC) method. A data set of 96 lung nodules (44 malignant, 52 benign) from 58 patients was used in this study. The 3D AC model is based on 2D AC with the addition of three new energy components to take advantage of 3D information: 1) 3D gradient, which guides the active contour to seek the object surface, 2) 3D curvature, which imposes a smoothness constraint in the z-direction, and 3) mask energy, which penalizes contours that grow beyond the pleura or thoracic wall. The search for the best energy weights in the 3D AC model was guided by a simplex optimization method. Morphological and gray-level features were extracted from the segmented nodule. The rubber band straightening transform (RBST) was applied to the shell of voxels surrounding the nodule. Texture features based on run-length statistics (RLS) were extracted from the RBST image. A linear discriminant analysis classifier with stepwise feature selection was designed using a second simplex optimization to select the most effective features. Leave-one-case-out resampling was used to train and test the CAD system. The system achieved a test area under the receiver operating characteristic (ROC) curve (A_z) of 0.83±0.04. Our preliminary results indicate that use of the 3D AC model and the 3D texture features surrounding the nodule is a promising approach to the segmentation and classification of lung nodules with CAD. The segmentation performance of the 3D AC model trained with our data set was evaluated with 23 nodules available in the Lung Image Database Consortium (LIDC). The lung

nodule volumes segmented by the 3D AC model for best classification were generally larger than those outlined by the LIDC radiologists using visual judgment of nodule boundaries.

2.2 Introduction

Lung cancer is the leading cause of cancer death for both men and women in the United States, accounting for 28% of all cancer deaths, or an estimated 163,510 lives in 2005. More people die from lung cancer than from colon, breast, and prostate cancers combined. While the five-year survival rate for lung cancers is only 15%, if detected and treated at its earliest stage (stage I), the five-year survival rate increases to 47% [30]. Unfortunately, most patients present clinically with advanced stage disease. The lack of a generally accepted screening test to reduce lung cancer mortality contributes to the poor prognosis of lung cancer. Furthermore, existing diagnostic tests to evaluate lung nodules are insufficient, with many lung nodules classified as indeterminate for malignancy. For this reason, approximately half of the indeterminate lung nodules resected at surgery are benign [31]. Reducing the number of biopsies for benign nodules will reduce health care costs and patient morbidity.

The Early Lung Cancer Action Project (ELCAP) was initiated in 1992 to assess the usefulness of annual low dose computed-tomography (CT) screening for lung cancer in a high-risk population [13]. Initial findings from the baseline screening of 1000 patients indicated that low dose CT can detect four times the number of malignant lung nodules and six times more stage I malignant nodules than chest radiography. These results have been confirmed by several groups of investigators [6-12]. These data suggested a strong potential for improving the likelihood of detecting lung cancer at an earlier and potentially more curable stage with CT [13]. The on-going National Lung Screening Trial funded by the National Cancer Institute is the first multi-center, randomized controlled trial, to evaluate the effectiveness of helical CT versus chest radiography for lung cancer screening.

Although CT may be more sensitive than chest radiography for the detection of lung cancer, potential impediments to the use of helical CT for lung cancer screening exist. For example, the chance of false negative detection due to the large volume of images in each multidetector CT examination is not negligible, the management of the large number of benign nodules or false-positive results that are detected may limit the cost-effectiveness of screening CT, and the follow up of nodules found on CT with serial CT examinations increases radiation exposure to the population [12]. One solution to address some of these issues may be computer-aided diagnosis (CAD), which has been shown to increase the sensitivity of breast cancer detection on mammography screening in clinical practice [32]. Computer-aided detection may reduce false negative detections, while computer-aided diagnosis (characterization) may increase the discrimination between malignant and benign nodules.

CAD systems typically involve the steps of segmentation, feature extraction, and classification. Various methods used in medical image segmentation such as thresholding [33], region growing [34, 35], and level sets [36, 37] have been evaluated. Segmentation of organs or other structures where the general shape is known has been performed with atlas-based segmentation methods [38]. While these methods may be effective for specific types of lesions and images, pulmonary nodules present a challenging problem due to their variability in shape and anatomic connection to neighboring pulmonary structures, such as blood vessels and the pleural surface.

Previous CAD development for CT focused mainly on automated detection [33, 39-46]. Recently there has been more work on the classification of malignant and benign nodules. McNitt-Gray *et al.* obtained 90.3% correct classification accuracy between 14 malignant and 17 benign cases [47]. Shah *et al.* achieved A_z values between 0.68 and 0.92 with 48 malignant and 33 benign nodules, using four different types of classifiers in a leave-one-out method. Features were extracted from contours manually drawn on a single representative slice of each nodule [48]. Armato *et al.* used an automated detection scheme, then manually separated nodules from non-nodules for the classification step. They achieved an A_z value of 0.79 using features such as radius of sphere of equivalent volume, minimum and maximum compactness, gray-level threshold, effective diameter, and location along the *z*-axis [49]. Kawata *et al.* used surface curvatures and ridge lines as features for description of 62 cases including 47 malignant and 15 benign nodules, showing good evidence of separation between malignant and benign classes in feature maps; no A_z value was reported [50]. Li *et al.* reported an A_z of

0.937 for distinction between 61 malignant and 183 benign nodules in a leave-one-out testing method, and an A_z of 0.831 for a randomly selected subset consisting of 28 primary lung cancers and 28 benign nodules [51]. Features used included diameter, contrast of segmented nodule, and those extracted from gray-level histograms of pixels inside and outside the segmented nodule. Aoyama *et al.* reported an A_z of 0.846 for classifying 76 primary lung cancers and 413 benign nodules using multiple slices (10 mm collimation and 10 mm reconstruction interval), which was a statistically significant improvement over 0.828 when only using single slices [52]. Suzuki *et al*. obtained an A_z of 0.882 by use of a massive training artificial neural network (MTANN) on a data set of 76 malignant and 413 benign nodules [53].

We are developing an automated system for classification of malignant and benign nodules extracted from CT volumes. Nodules were segmented from the image background using a 3D active contour (AC) method. Malignant and benign nodules were differentiated using morphological and texture characteristics. The weights for the energy terms in the AC model were optimized using the classification accuracy as a figure-of-merit. Our initial experience in nodule classification is reported in this paper. For comparison, we also analyzed the classification performance using radiologists' subjective estimation of likelihood of malignancy, and a classifier designed with feature descriptors provided by radiologists. The segmentation performance of the 3D AC model trained with our method was evaluated with 23 nodules available from the Lung Image Database Consortium (LIDC).

2.3 Methods and Materials

2.3.1 Data Sets

2.3.1.1 Clinical Data Set

We analyzed 96 lung nodules (44 malignant and 52 benign) from 58 patients. All cases were collected with Institutional Review Board approval. Of the 44 malignant nodules, 25 were biopsy-proven to be malignant, and 19 nodules were determined to be malignant either through positive PET scans or known metastatic nodules from confirmed cancers in other body parts. Of the 44 malignant nodules, 15 were primary cancers and 29 were metastases. Of the 52 benign nodules, 10 were biopsy-proven and 42 were determined to be benign by two-year follow-up stability on CT. Of the 96 nodules, 20 (21%) were juxta-pleural and 12 (12.5%) were juxta-vascular as indicated by expert radiologists.

Each CT image was 512 x 512 pixels. The CT scans were acquired with either GE Lightspeed CT/I (single-slice helical), QX/I (4 slice), Ultra (8 slice), or LightSpeed Plus (8 slice) scanners, using imaging techniques of 120 kVp, 80-400 mAs, and reconstructed slice interval of 1.25-5 mm. Linear interpolation was performed in the *z*-direction to obtain isotropic voxels before initial contour generation and segmentation to facilitate the implementation of the 3D segmentation and feature extraction operations in the CAD system. The interpolation does not recover the reduced spatial resolution in the z-direction.

A user interface was developed for displaying the CT images and recording nodule locations and ratings provided by radiologists. Two radiologists were trained in using the software and giving ratings for the data set. Each case was read by one of these experienced thoracic radiologists who marked volumes of interest (VOI's) that contained lung nodules. For each nodule, a confidence rating of the likelihood of malignancy on a 5-point scale was provided, 5 being the most likely to be malignant. Electronic rulers were used to measure the longest diameter of each nodule as seen on axial slices. The radiologists also recorded various feature descriptors for each nodule, such as conspicuity, edge (smooth, lobulated, or spiculated/irregular), and the presence of calcification. Each radiologist read approximately half of the cases. The distribution of the longest diameters of the 96 nodules is shown in Fig. 2.1. The longest diameter ranged from 3.9 mm to 59.8 mm, with a median of 13.3 mm and mean of 17.3 mm. Fig. 2.2 shows the distribution of the malignancy ratings of the nodules by the radiologists. The malignancy ratings for benign and malignant nodules overlap substantially, confirming that visual characterization of the nodules on CT images is not a simple task.

2.3.1.2 LIDC Data Set

The 23 nodules available to-date from the data set provided by the LIDC [54] were used for testing our 3D AC model. The LIDC database is intended to be a common data set available to all researchers for development of CAD systems and for comparison of their performance. The data set includes "gold standard" segmentation of each nodule by six expert chest radiologists. Each radiologist performed one manual and two semi-automatic markings of each nodule, resulting in a total of 18 boundaries for each. The 18 boundaries were used to generate a probability map (pmap), which was scaled to a range of 0 to 1000. A boundary of the nodule at a pmap threshold of 500, for example, is a contour that encloses all the voxels with values greater than or equal to 500, which means that those voxels were considered to be part of the nodule by more than 50% of the 18 "gold standard" segmentations. More information about the database can be found on the LIDC website, where the images are also free for download: (http://imaging.cancer.gov/reportsandpublications/reportsandpresentations/firstdataset)

2.3.2 Initial Contour Determination

Our nodule segmentation method has two steps: estimation of an initial boundary by *k*-means clustering and refinement of the boundary with a 3D active contour model. The VOI determined by the radiologist may contain other pulmonary structures in addition to the nodule, such as blood vessels or voxels that are outside the lung region (chest wall or mediastinum). A lung region mask determined by our automated nodule detection system described in the literature [39] is first applied to the VOI to exclude the voxels belonging to the chest wall or the pleura from further processing. Then a 3D weighted *k*-means clustering method [55] based on CT values is used for initial segmentation of the nodule from the other structures in the VOI. The VOI is assumed to contain two classes: the lung nodule (including other tissue but excluding the chest wall) and the background. Clustering is performed iteratively until the cluster centers of the classes stabilize as described elsewhere [55]. The voxels grouped into the non-background class may or may not be connected. A 26-connectivity criterion is used to determine the various connected objects in the 3D space and the largest one closest to the center is chosen as the nodule. We can make this assumption because of the *a priori* knowledge that the VOI contains a nodule and that this study is focused on classification, not on detection (determining whether objects are true nodules).

The lung nodule segmented by clustering may be attached to blood vessels or other structures. Once this main object is identified in the VOI, 3D morphological opening with a spherical structuring element is applied to the object to trim off some connected vessels or structures. The structuring element is chosen to be spherical in this application because nodules tend to be spherical in shape, while non-nodule objects such as blood vessels tend to be cylindrical. For each slice intersecting the object in the VOI, the radius of an equivalent circle with the same area was found. The radius of the structuring element was chosen experimentally as the average of the radii subtracted by 1. Equivalent radii of cross sections are used because the partial volume effect makes some objects more cylindrical than they truly are, resulting in structuring elements that are too large if volumes are used in the calculation. After morphological opening, the boundary of the resulting object is used as the initial contour for the active contour segmentation.

2.3.3 3D Active Contour Segmentation

2.3.3.1 The Active Contour Model

Deformable contour models, particularly the AC model introduced in the seminal paper by Kass *et al.* [14], are well-known tools for image segmentation. Active contours are energy-minimizing splines guided by various forces, or energies. The internal energies impose constraints on the contour itself, while external energies push the contour towards salient image features such as lines and edges. The contour is represented as a vector **v**(s)=(x(s),y(s)), where s is the parameter arc-length. The energy functional is defined as:

$$E^*_{snake} = \int_0^1 E_{snake}((v(s))ds \qquad (2.1)$$

The E^*_{snake} energy contains the various energy components that will be discussed later along with the energies we contribute. Segmentation of the object using the AC is thus achieved by minimizing E^*_{snake}.

2.3.3.2 Parametric Implementation of Continuous Splines

Using variational calculus or dynamic programming to minimize the total energy of the parametric representation of a continuous contour can result in instability and a tendency for points to bunch up together [15]. Instead of a continuous contour representation, the AC optimization algorithm in this study represents the contour by a set of vertices and uses a greedy algorithm to find the solution. The neighborhood for vertex **v**(c), $c = \{1, 2, ..., N\}$, is examined at each iteration, where N is the total number of polygon vertices. The vertex is then moved to the pixel with the minimum contour energy $E_{min}(\mathbf{v}(c))$. The process repeats until the number of vertices that moves is below a threshold. The final contour is obtained by minimizing the cost function:

$$\begin{aligned} E_{total} = \min_{E(c)} \sum_{c=1}^{N} [& w_{\hom} E_{\hom}(c) + w_{cont} E_{cont}(c) + w_{curv} E_{curv}(c) + w_{3Dcurv} E_{3Dcurv}(c) \\ & + w_{grad} E_{grad}(c) + w_{3Dgrad} E_{3Dgrad}(c) + w_{bal} E_{bal}(c) + w_{mask} E_{mask}(c)] \end{aligned} \tag{2.2}$$

where *E*(c) is the energy at a pixel in the neighborhood of vertex $\mathbf{v}(c) = (x(c), y(c)), c \in \{1, 2, ..., N\}$. In this energy functional, the internal energies include homogeneity (*hom*), continuity (*cont*), curvature (*curv*), 3D curvature (*3Dcurv*), and the external energies include gradient (*grad*), 3D gradient (*3Dgrad*), balloon (*bal*), and mask (*mask*). The weight w_j is a parameter assigned to each energy *j*, where *j* represents one of the eight energies: *hom, cont, curv, 3Dcurv, grad, 3D grad, bal, and mask.*

2.3.3.3 2D Energies

In this preliminary study, the energy terms other than 3D curvature and 3D gradient are calculated on the x-y planes of the CT slices intersecting the nodule. The vertices on each slice move in the x-y plane during the iteration. The continuity of the segmented nodule area between different slices is constrained by the 3D curvature and the 3D gradient terms which provide the 3D information in the current model.

A brief description of the 2D energy components is given here. Details can be found elsewhere in the literature [15, 56, 57]. Homogeneity energy [56] is a measure of how similar the pixel intensities inside the contour are. The contour divides each region-of-interest (ROI) into two regions: the area enclosed by the contour and the background excluding the chest wall. We seek to minimize the intensity variation within each region while maximizing the difference of the mean intensities between the two regions. The homogeneity energy is therefore calculated as the ratio of the within-regions sum of squares to the between-region sum of squares of the gray levels in the two regions. The continuity energy maintains regular spacing between the vertices of the contour. If the points could move in the neighborhood without this constraint, then they might move towards one another, leading to the ultimate collapse of the contour. The continuity energy is calculated as the deviation of the length of line segment between two vertices from the average line segment length over all vertices. The curvature energy smoothes the contour by discouraging small angles at vertices. There are many ways of estimating curvature, as investigated by Williams and Shah [15]. In our implementation, the second-order derivative along the contour is approximated by finite differences. If $\mathbf{v}(c)$ is a point on the contour as depicted in Fig. 2.3, then the second-order derivative at $\mathbf{v}(c)$ is $|\mathbf{v}(c\text{-}1) - 2\mathbf{v}(c) + \mathbf{v}(c+1)|$, which is used as the curvature energy. If the angle where two segments meet at a vertex is small, then this term will be large; conversely, when the angle is large, a low energy value results.

The balloon energy prevents the contour from collapsing onto itself, which is a well-known phenomenon for AC segmentation [58]. The normal direction $\mathbf{n}(c)$ to the contour is defined as the average of the normals to the two sides of the polygon that meet at vertex $\mathbf{v}(c)$. Let $\mathbf{v}'(c)$ be the new position where vertex $\mathbf{v}(c)$ moves to in the neighborhood. The balloon energy can then be calculated as the cosine of the angle between $\mathbf{n}(c)$ and $\mathbf{v}'(c)$- $\mathbf{v}(c)$. The weight w_{bal} determines whether the contour expands in the normal direction or the direction opposite to the normal. If the weight is negative, then a point moving farther along the normal direction will lower the energy, thus expanding the contour.

The gradient energy attracts the contour to object edges. To calculate the gradient magnitude, the image is first smoothed with a low-pass filter, chosen experimentally to be a Gaussian filter, $F(x,y) = e^{-(x^2+y^2)/\sigma^2}$, with σ=300μm. The partial derivatives are found in the vertical and horizontal direction, and the magnitude of the resulting vector is computed. The energy is defined as the negative of the gradient magnitude, so object edges with high gradient magnitudes will attract the contour. For image $I(x,y)$, the gradient energy is calculated as:

$$E_{3dgrad}(c) = -|\bar{\nabla}\,(I(x,y) ** F(x,y))|^2 \tag{2.3}$$

where ** denotes 2D convolution, and $\bar{\nabla}$ is the partial derivative gradient operator.

2.3.3.4 New Energies

2.3.3.4.1 3D Gradient

The 3D gradient energy is defined in a similar way to the 2D gradient energy. The 2D gradient magnitude image shows the edges of the object in the 2D image, but the 3D gradient magnitude image reveals the surface of the object, thus giving better shape information of the nodule. The 3D image containing the nodule is first smoothed with a 3D low-pass Gaussian filter:

$$F(x,y,z) = \frac{1}{(2\pi)^{3/2}\sigma}\exp\left[-\frac{1}{2}\frac{(x^2+y^2+z^2)}{\sigma^2}\right] \tag{2.4}$$

The energy is calculated in a similar way to the 2D method:

$$E_{3dgrad}(c) = -|\bar{\nabla}\,(I(x,y,z) ** F(x,y,z))|^2 \tag{2.5}$$

2.3.3.4.2 3D Curvature

We introduced the 3D curvature energy to take advantage of the information in the *z*-direction, which we found to improve segmentation results over 2D energies alone [16]. This energy is an extension of the curvature constraint idea in 2D, where the energy is calculated using the two nearest neighbor vertices. In 3D, the energy for each vertex is calculated with the nearest points on the contours above and below the current contour. With this energy, the 2D contour at a given slice will thus be constrained by the adjacent contours above and below. This prevents one contour from varying substantially from other contours and results in an overall smoothness in the z-direction.

To calculate this energy for vertex $\mathbf{v}_i(c)$ of the contour on the i^{th} slice, the closest points to $\mathbf{v}_i(c)$ on the contours in the slices above and below i are determined. Let $\mathbf{v}_{i+1}(c_{i+1})$ and $\mathbf{v}_{i-1}(c_{i-1})$, denote the closest points to vertex c on slices (i+1) and (i-1), respectively. Since these points are defined to be lying on the contours, they may be on the lines between vertices and are not necessarily the vertices that move during deformation of other slices. Note that the index of the point on the contour of the (i+1)th and (i-1)th slices may not be the same as c and, in fact, they may not be vertices of the contours. As described above, we used two new indices with subscripts, c_{i-1} and c_{i+1}, to denote that these are different indices on the respective contours. The 3D curvature energy is represented by an approximation to the second derivative of the contour in the z-direction,

$$E_{3dcurv}(c) = |\,\mathbf{v}_{i-1}(c_{i-1}) - 2\mathbf{v}_i(c) + \mathbf{v}_{i+1}(c_{i+1})\,| \qquad (2.6)$$

2.3.3.4.3 Mask Energy

Nodules attached to the pleura (juxtapleural nodules) present a challenge to segmentation. Both nodules and normal body tissues have a similar range of Hounsfield Units (HU). Region growing or threshholding methods will fail to segment nodules, because the pleura, chest wall, or mediastinum may be included in the contour. Gradient-based methods will not be able to detect the edge of the nodule either, as there is no well-defined boundary between nodule and normal tissues. Kostis [59] proposed connecting the two points of highest curvature (where the boundary of the pleura meets that of the nodule) and estimating the curvature of the wall boundary. That method may not be sensitive enough to local concavities, due to anatomic or pathological variations.

We have designed a mask energy to meet this challenge. The mask energy is a function of the distance from a vertex on the AC contour of the nodule to the lung boundary. This is calculated for each vertex that moves beyond the lung boundary (in the pleura or thoracic wall) during each iteration of the energy-minimizing procedure. An accurate lung boundary is therefore required for determining the mask energy. We will describe below our methods for finding the initial lung boundary and the subsequent local refinement used to produce an accurate boundary.

The first step is to determine the boundary of the pleura. Because of different CT scanning parameters, the k-means clustering technique with CT voxel value as the feature is used to segment the lung regions from the thorax in each CT slice instead of a simple threshold. The extracted lung regions are represented by polygons marking the lung boundaries. This process provides the initial lung boundaries in the entire slice. More details on this process may be found in the literature [39].

The initial lung boundaries are a general outline of the lungs, but they may not be sensitive enough to exactly delineate the boundary between a nodule and the pleural surface. If the estimated lung boundary is not close enough to the actual boundary, it may even trim off part of a juxtapleural nodule. To refine the boundary between a juxtapleural nodule and the pleural surface, we use *k*-means clustering [55] within each VOI, to determine the mean and standard deviation of the voxel values considered to be background (lung regions). Any voxels originally considered part of the pleura, chest wall, or mediastinum that fall within 3 standard deviations of this range will have their membership changed to be that of the lung region. The threshold of 3 standard deviations was chosen experimentally based on the separation between the distributions of voxel values of the nodules and the lung regions in the training samples. As depicted in Fig. 2.4, an indentation was created as the refined boundary included more of the area originally considered chest wall into the lung region.

An *indentation detection* [39] technique is used to fill in that indentation. This method detects an indentation by means of distance ratios. For every pair of points P_1 and P_2 along the lung boundary, three distances are calculated. Distances d_1 and d_2 are distances between P_1 and P_2 measured by traveling along the boundary in the counter-clockwise and clockwise directions, respectively. The third distance d_e is the Euclidean distance between P_1 and P_2. The ratio is calculated:

$$R_e = \frac{\min(d_1, d_2)}{d_e} \tag{2.7}$$

If the ratio is greater than a threshold, then an indentation is assumed, and it is filled by connecting the points P_1 and P_2 with a straight line. R_e was chosen to be 1.5 in our previous study [39]. Fig. 2.4 shows an example how the boundary improves as a result of this method.

This boundary marks where the lung region is. If a vertex $\mathbf{v}(c)$ moves to a position $\mathbf{v}'(c)$ that falls outside of the lung region into the chest wall, the mask energy is calculated as:

$$E_{mask} = |\mathbf{b}(c) - \mathbf{v}'(c)| \tag{2.8}$$

where $\mathbf{b}(c)$ is the point on the lung boundary closest to $\mathbf{v}(c)$. Instead of outright forbidding the nodule contour to grow into the chest wall, this energy allows for the fact that the lung boundary may not be completely accurate. The contour may grow into the chest wall, but it will be penalized the further away from the chest wall boundary it grows.

2.3.4 Feature Extraction

Gurney [60] has provided likelihood ratios for various characteristics that may be useful for discriminating malignant from benign nodules. Other features to discriminate malignant from benign nodules have been described by Erasmus [61]. We seek to quantify the characteristics of nodules by mathematical feature descriptors. The accuracy of the segmentation is important for extraction of some of the features. There is no single feature that can accurately determine whether a nodule is benign or malignant. For example, features such as the presence of calcification may be a strong indicator that a nodule is benign. However, it has been reported that 38% to 63% of benign nodules are non-calcified [61-63], and in the study by Swensen *et al.* [11], up to 96% benign nodules were non-calcified.

From the segmented nodule boundary, we extracted a number of morphological features including volume, surface area, perimeter, maximum diameter, and maximum and minimum CT value (HU) inside the nodule. We also extracted statistics from the gray-level intensities of voxels inside the nodule including the average, variance, skewness, and kurtosis of the gray-level histogram.

In addition to features that can be derived from the inside of the nodule, the tissue texture around the margin of the nodule is also important. The growth of malignant tumors tends to distort the surrounding tissue texture, while benign nodules tend to have smooth surfaces with more uniform texture around them. To derive these features from the texture around the nodule, the rubber band straightening transform (RBST) is first applied to planes of voxels surrounding the nodule. The *run-length statistics* (RLS) texture features are then extracted from the transformed images, as described below.

2.3.4.1 Rubber Band Straightening Transform (RBST)

The RBST was introduced by Sahiner *et al.* [64] for analysis of the texture around mammographic masses on 2D images. The RBST image is obtained by traveling along the boundary of the nodule, transforming the band of pixels surrounding the nodule into a rectangular image. In this way, spicules that grow out radially from an object may be transformed as approximately straight lines in the y-direction.

The RBST maps a closed path at an approximately constant distance from the nodule boundary of the original image to a row in the transformed image, as depicted in Fig. 2.5. The difference between the RBST and the transformation from Cartesian to polar coordinates is that the irregular or jagged lesion boundary will be transformed to a straight line in the horizontal direction, whereas in a Cartesian to polar transformation, only a circle of constant radius will be transformed to a horizontal straight line.

With 3D CT scan data, the texture around the whole nodule needs to be extracted. We apply the RBST to the original CT slices to extract texture in the axial planes. To adequately sample the texture in all directions, we slice the nodule with two additional sets of planes: by considering the nodule as a globe, one set contains the longitude lines that run through the north and south poles (*z*-direction in a CT scan), and the other through the east-west poles. In each set, four oblique planes (45^o apart) slice evenly along the lines of longitude on the nodule surface. The RBST is applied to a band of voxels surrounding the nodule on each of the oblique planes. Each RBST image is then enhanced by Sobel filtering in both the horizontal and vertical directions. Texture features based on run-length statistics are extracted from the Sobel-filtered RBST images.

RLS texture features were introduced by Galloway [65] to analyze the number of runs of a gray level in an image. A run-length matrix $p(i,j)$ stores information of the number of runs with pixels of gray-level i and run length j. In this study, the 4096 gray levels are binned into 128 levels before the run-length matrix is constructed to improve the statistics in the matrix. Galloway designed five RLS features extracted from $p(i,j)$ to describe the gray level patterns in the image: Short Run Emphasis (SRE), Long Run Emphasis (LRE), Gray-Level Nonuniformity (GLN), Run Length Nonuniformity (RLN), and Run Percentage (RP). Dasarathy and Holder proposed four more features [66] which are based on the idea of joint statistical measures of the gray levels and run length: Short Run

Low Gray-Level Emphasis (SRLGE), Short Run High Gray-Level Emphasis (SRHGE), Long Run Low Gray-Level Emphasis (LRLGE), and Long Run High Gray-Level Emphasis (LRHGE). Mathematical expressions for these RLS texture features are given in Appendix I. We extract these nine RLS texture features from each of the Sobel-filtered RBST images. Each feature is averaged over the slices in each of the three groups (axial x-y plane, north-south longitudinal planes, and east-west longitudinal planes), providing 3D texture information around the nodule.

2.3.5 Feature Selection and Classification

Many different features may be extracted from a nodule, but not all of them are effective in differentiating the malignant and benign nodules. To identify effective features to be used in the linear discriminant classifier, we employed stepwise feature selection using F-statistics [67]. The F-statistics is used to evaluate the significance of the change in a feature selection criterion, which is chosen to be the Wilks lambda (ratio of the within-class sum of squares to the total sum of squares of the two class distributions) in this study, when a feature is entered into or removed from the feature pool. Simplex optimization [68] is utilized to determine the best combination of thresholds (F_{in}, F_{out}, *tol*) that gives the highest figure-of-merit (FOM), the area under the ROC curve (A_z), where F_{in} is the F-to-enter, and F_{out} is the F-to-remove threshold. The *tol* threshold sets how correlated the features can be for selection.

2.3.6 Training and Testing

A leave-one-case-out resampling scheme was used for training the segmentation energy weights and feature selection. In a given cycle, one case that included all CT scans from the same patient was left out to be used as the test case while the other cases were used for training. The collection of the test results from all of the left-out cases after the leave-one-case-out cycles were completed was evaluated by ROC analysis [69]. Two simplex optimizations were embedded: one in the determination of segmentation weights, and the other in the selection of features. Simplex optimization was used to determine the set of weights that would result in the highest A_z from the feature selection and classification step. A schematic of the training and testing process is shown in Fig. 2.6 and the process is described below.

Step 1: Initialize with a set of weights for the 3D AC.

Step 2: Generate the boundaries based on the weights, and then extract features from the boundaries.

Step 3: Perform simplex optimization for feature selection using a leave-one-case-out resampling scheme for both feature selection and classifier weight determination. The simplex searches for the F_{in}, F_{out}, and *tol* thresholds that provide the highest test A_z from a linear discriminant classifier with the selected features as predictor variables to differentiate the malignant and the benign classes.

Step 4: Determine a new set of AC weights using the test A_z as the FOM for the simplex optimization of AC segmentation.

Step 5: Go back to Step 2 and the subsequent steps to determine A_z for the new weights. The iteration continues until simplex converges to the best A_z or a predetermined number of iterations is performed.

In the leave-one-case-out loop for feature selection (Step 3), we also used an alternative FOM, the partial area index $A_z^{0.9}$ (TPF above 0.9) for the feature selection process. The use of $A_z^{0.9}$ as the FOM would select features that maximize the specificity at the high sensitivity region [70], which is often more important than having a classifier with high average sensitivity over the entire specificity range. The classifier designed with the $A_z^{0.9}$ is compared with that designed with A_z.

2.3.7 Comparison with LIDC First Data Set

The performance of the trained 3D AC segmentation program was evaluated with the nodules in the independent LIDC data set. We used the set of 3D AC weights that provided the highest test A_z in the leave-one-case-out training and test process using our data set as described above. The 3D AC weights were then fixed and applied to the LIDC nodules.

To quantify performance, we propose to use an overlap measure in combination with a percentage volume error measure. Let A denote the object segmented using the 3D AC method and L denote the gold standard reference object. Let V_A be the volume of the object A, and V_L the volume of the object L, which is the volume of the LIDC object calculated at a specified pmap threshold in this study. The overlap measure is the ratio of the intersection of volumes relative to the volume of the gold standard reference object:

$$Overlap_1(A,L) = \frac{|V_A \cap V_L|}{|V_L|} \tag{2.9}$$

Alternatively, one may use an overlap measure that is defined as the ratio of the intersection of volumes relative to the union of the volumes of the segmented and gold standard reference objects:

$$Overlap_2(A,L) = \frac{|V_A \cap V_L|}{|V_A \cup V_L|} \tag{2.10}$$

where $|\cdot|$ denotes cardinality. These measures are extensions to the 3D volume from the 2D area overlap measures [57, 71]. In the expressions of the overlap measures, each of the volumes can be considered as the set of voxels comprising the volume.

Overlap$_1$(A,L) or *Overlap$_2$(A,L)* can provide one measure of the 3D AC performance relative to the "gold standard" object but neither of them gives a complete description. *Overlap$_1$(A,L)* represents the fraction of the gold standard object that is included in the segmented object, though there is no indication as to what fraction, if any, of the segmented object is outside the gold standard object. *Overlap$_2$(A,L)* represents the fraction of overlap relative to the union, but does not provide information on how large a fraction of the gold standard object is actually included in the segmented object and whether the non-overlap volume is contributed by the segmented object or by the gold standard.

To complement the information, we calculated the percentage volume error, V_{err}, defined in Eqn. (2.11) as the difference between the volumes of the segmented object V_A and the gold standard object V_L, relative to V_L:

$$V_{err} = \frac{V_A - V_L}{V_L} \times 100\% \tag{2.11}$$

From the two measures, *Overlap$_1$(A,L)* and V_{err}, one can derive a number of useful performance metrics, as detailed in Appendix II, that quantify the number of voxels correctly and incorrectly segmented as a part of the object, using the gold standard object as a reference.

2.3.8 Classification with Radiologist's Feature Descriptors and Malignancy Ratings

For comparison, we analyzed the accuracy of a classifier designed with features that were provided by the radiologists to describe the nodule characteristics. When the radiologists identified the nodule locations in each CT scan, they provided descriptors of the nodule characteristics including: (1) the longest diameter, (2) perpendicular diameter to the longest diameter, (3) conspicuity, (4) edge (smooth, lobulated, or spiculated/irregular), (5) presence or absence of calcification, (6) presence or absence of cavitation (7) presence or absence of fat, (8) attenuation (solid/mixed/ground glass opacity), (9) nodule location (the lobe of the lung), and (10) location (juxtavascular, juxtapleural). These descriptors were treated as input features to design a linear discriminant classifier. Again, leave-one-case-out resampling was used for stepwise feature selection and classifier weight determination. Simplex optimization [68] was employed to find the features that resulted in the highest test A_z.

The radiologists also provided a malignancy rating on a 5-point scale for each nodule based on subjective impression from the CT images (Fig. 2.2). We applied ROC analysis to the malignancy rating and estimated the A_z. This A_z value was also compared to the test A_z obtained by the computer classifier.

2.4 Results

2.4.1 Feature Selection and Classification Based on 3DAC

Table 2.1 shows the comparison of classification accuracy obtained with different methods. The test ROC curves for the various classifiers are shown in Fig. 2.8. When $A_z^{0.9}$ was used as the FOM in the leave-one-case out scheme described above, the training A_z was 0.88 ± 0.03 and the test A_z and $A_z^{0.9}$ were 0.78 ± 0.05 and 0.35, respectively. When A_z was used as the FOM, the training A_z was 0.87 ± 0.04 and the test A_z was 0.83 ± 0.04, with $A_z^{0.9}$ of 0.30. The difference between using $A_z^{0.9}$ and A_z as FOM was not significant (p=0.15), as estimated by the CLABROC program [72]. The distribution of classifier scores for A_z as FOM is shown in Fig. 2.7. An average of 4.1 features was selected. Four of the most frequently selected features along with the number of times selected out of 58 leave-one-case out cycles are:

- Long-range low gray-level emphasis on the axial planes (58)
- Run-length non-uniformity in the north-south oblique planes (56)
- Maximum CT number: (55)
- Long-range high gray-level emphasis on the axial planes (54)

This indicates that similar features were consistently selected over the different leave-one-case-out cycles, even though a different case was left out each time.

2.4.2 Comparison with Radiologist's Feature Descriptors and Malignancy Ratings

For classification using the radiologist-provided feature descriptors of the nodules, on average only one feature, the longest diameter, was consistently selected with A_z as the FOM. The test A_z was 0.80 ± 0.05 with $A_z^{0.9}$ of 0.24. The difference in A_z between the classifier based on radiologist-provided feature descriptors and the computer classifier did not achieve statistical significance ($p = 0.40$). When $A_z^{0.9}$ was used as the FOM, the test A_z and $A_z^{0.9}$ was 0.82 ± 0.04 and 0.32, respectively ($p = 0.48$). Using the radiologist's malignancy rating (Fig. 2.2) as input to the ROC analysis resulted in an A_z of 0.84 ± 0.04, with an $A_z^{0.9}$ of 0.33. The performance of the computer classifier was comparable to that from radiologists' assessments of the likelihood of malignancy ($p = 0.98$).

2.4.3 Segmentation Evaluation on LIDC Data Set

The 3D AC model with weights trained by the nodules in our data set, as described above, were tested on the 23 LIDC nodules. The mean and median overlap measures for the "gold standard" volumes defined at various thresholds (from 100 to 1000 in steps of 100) of the probability map (pmap) are shown in Fig. 2.9(a). For a given nodule, the number of voxels included within a pmap threshold, i.e., the common volume that radiologists agreed to be a part of the nodule, decreased as the pmap threshold increased. There were eight nodules with no voxels at pmap threshold of 1000 because there were no common voxels that all 18 segmentations agreed to be part of the nodule.

These nodules were excluded in the calculation of *Overlap$_1$(A,L)* and the percentage volume error at pmap threshold of 1000 because the values would be undefined. The average and median were calculated with the remaining nodules. As seen in Fig. 2.9(a), the mean of *Overlap$_1$(A,L)* increases from 0.62 to 0.95 as the pmap increases. The median of the *Overlap$_1$(A,L)* follows a similar trend as the mean, increasing from 0.64 to 1.0. The relatively high values of *Overlap$_1$(A,L)* and its increasing trend with increasing pmap value indicate that a substantial fraction of the voxels that all radiologists marked as a part of the nodule was consistently included in the AC segmented volume. The mean of *Overlap$_2$(A,L)* ranges from 0.07 to 0.63, with the maximum at a pmap threshold of 400. The median of *Overlap$_2$(A,L)* follows a similar trend as the mean with a range from 0.009 to 0.67, reaching a maximum at the pmap threshold of 300. The small values of *Overlap$_2$(A,L)* result from the overestimation of the volumes by AC segmentation, which is also shown by the percentage volume errors.

The percentage volume error relative to the radiologists' manually segmented nodule volumes was calculated using Eqn. (2.11) and plotted in Fig. 2.9(b). The average percentage volume error was lowest at a pmap threshold of 300, with a mean of 2% and a median of 10%. The volume error increased rapidly as the pmap threshold increased because the number of common voxels decreased. At pmap thresholds greater than 800, there were very few common voxels from the radiologists' outlines so that the percentage volume error exceeded 500%. The high value of *Overlap$_1$(A,L)* indicated that most of these common voxels were included in the computer-segmented volumes. The relationship between the percentage volume error and the nodule volume calculated at the pmap threshold of 500 is plotted in Fig. 2.10. The threshold of 500 was chosen since at least half of the contours provided by radiologists enclosed these voxels to be a part of the true nodule.

2.5 Discussion

There is no ground truth for lesion boundaries in medical images. The most commonly used gold standard is subjective manual segmentation by radiologists. The LIDC studied intra- and inter-observer variability in manual segmentation of lung nodules by experienced thoracic radiologists [54]. It found large variabilities among radiologists due to the difficulty in defining the boundaries of ill-defined nodules, a task that even experienced radiologists are not required to perform clinically. The LIDC has provided a data set of 23 nodules, each with a probability map ("pmap" image) derived from 18 boundaries manually outlined by six expert thoracic radiologists (each providing one manual and two semi-automatic segmentations). The probability map can be used as the "gold standard" boundaries for evaluation of segmentation by computer methods. For our data set of nodules, we did not attempt to obtain a gold standard because even experienced radiologists have no standardized method for defining nodule boundaries. To reduce inter- and intraobserver variation, it will be necessary to have multiple radiologists segment each nodule multiple times, as done by the LIDC. This approach will be impractical to perform within one institution, since even a small data set like the one used in this study contained over 950 CT slices that intersected the nodules.

It is difficult to analytically find a set of energy weights that would provide effective segmentation for all nodules. The difficulty can be attributed to (i) energy calculations required for the linear cost function in Eqn. (2.2) being highly non-linear, (ii) lung nodules growing in many different irregular shapes, and (iii) boundaries between nodule and lung regions varying from very distinct to very fuzzy. One empirical method of determining the weights could be manually segmenting the lung nodules and training the contour weights to fit these case samples, using an overlap measure or distance measure as an FOM in the optimization process. However, since there are large intra- and inter-observer variabilities even among experienced thoracic radiologists as to what constitutes accurate segmentation, our overall goal is not to conform the segmented objects to subjectively estimated boundaries. Rather, the features extracted from the generated boundaries should provide accurate classification between malignant and benign nodules. We therefore used the A_z or $A_z^{0.9}$of the feature selection step as the FOM to guide the search for the best weights in the 3D AC model. This approach not only takes into consideration classification accuracy during segmentation, it also has the advantage of eliminating the need for manually drawing the nodule boundaries by radiologists for all the training samples. Nevertheless, it would be interesting in a future study to examine how well the classifier performs if features are extracted from manually drawn contours provided by radiologists in comparison to classification by automated segmentation as described in this study.

We examined the segmentation of the nodules in the LIDC data set by our 3D AC model trained with A_z as an FOM. The average and median overlap measures at various thresholds and the percentage volume errors based on the LIDC pmap give an indication of segmentation performance. The average percentage volume errors at the pmap threshold below about 500 were in the range of -20.4% to 6.2%. The average percentage volume error at the pmap threshold of 500 was 85.7%.

The sudden increase in the values of $Overlap_1(A,L)$ and the percentage volume error at pmap threshold of 500 was caused by the way that the boundary voxels were marked in the LIDC data set. These boundary voxels were assigned a value of 32767 in the pmap without the orignal voxel values given. In our calculation of nodule volume, we included the boundary voxels to be part of the volume for pmap < 500, i.e., treating the voxel values of the boundary voxels as 499. For nodule volumes at pmap threshold of greater than or equal to 500, the boundary voxels would be outside the nodule volume. This resulted in a large transition in the nodule volume at pmap threshold of 500, especially for small nodules, as shown in Fig. 2.10.

For 17 of the 23 (74%) nodules, the percentage volume error was below 100%. Three nodules had large errors. One had ground-glass opacity texture, while the other two had low contrast between the nodule and lung. Two of those were small (with longest diameters of 4.32 mm and 4.98 mm based on the gold standard boundaries), while the images are very noisy for the larger one. Representative slices through the center of the three lung nodules are shown in Fig. 2.11. These nodules contributed most to the high volume percentage error, due to incorrect segmentation of the attached blood vessels or due to incorrect expansion of the active contour beyond the faint edges.

The 3D AC segmentation energy weights that provided the best features and A_z for nodules in our clinical data set therefore agrees to a certain extent with the boundaries perceived by radiologists in the LIDC data set. If the purpose of the segmentation is to simulate radiologists' manual segmentation at a chosen pmap threshold, the 3D AC model should be trained with a set of nodules with gold standard boundaries at the same threshold. The 3D AC weights optimized in this manner will likely provide segmented boundaries for test nodules in better agreement with the manual boundaries than the current training. As discussed above, whether the boundaries that are in agreement with experienced radiologists' manual segmentation will provide higher classification accuracy than our current segmentation method remains to be investigated. This study can be pursued when the LIDC data set is large enough to provide both training and testing samples for malignant and benign nodules.

It is generally defined and accepted that solitary pulmonary nodules are less than 3 cm in longest diameter [73, 74], but the data set used in this study included 14 masses greater than 3 cm, two of which were benign. Although one motivation for CAD tools is to assist radiologists with less-obvious (smaller) indeterminate nodules, we intend to train a CAD system that can analyze a reasonably broad range of different types of nodules and masses. We therefore included all types of nodules that we collected in the data set. We extracted morphological and gray level features in addition to texture features to be used in the input feature pool for design of our classification system. However, the stepwise feature selection with simplex optimization selected mainly texture features. This indicates that features such as the size or shape of nodules may not be as discriminatory, likely because benign objects, such as those caused by inflammatory processes, also result in nodules of varying sizes and shapes. On the other hand, the texture around benign nodules may not be the same as that caused by a malignant growth. In these cases, texture information would be more discriminatory than shape descriptors. Another indication of this is that the longest diameter feature was the one selected most consistently out of the radiologist-extracted feature space, but the same feature was not selected in the combined morphological and texture features extracted by the computer from the 3D AC boundaries. Combinations of texture features seemed to provide better discrimination, even though the longest diameter is a relatively discriminatory feature as evidenced by the A_z of 0.75 using this feature alone. Thus, we believe that the inclusion of nodules greater than 3 cm in longest axis would provide the texture information important for training, not necessarily for size or shape, and that the trained CAD system may be used for analysis of nodules or masses over a reasonably broad ranges of sizes because its performance does not depend on the size of the nodule or masses.

There were nodules for which the classifier did not perform well. One example is shown in Fig. 2.12. This nodule was malignant, but the classifier gave a score indicating a low likelihood of malignancy. This nodule was embedded and located between branching blood vessels near the lung hilum, which resulted in poor segmentation ($Overlap_1(A,L)$ measure of 0.67 and 78% volume error). Furthermore, the texture features extracted from this nodule would not be a good indicator of spiculation or malignant growth, because the blood vessels occupied much of the surrounding tissue volume. Even though our segmentation method is fairly robust with juxtavascular nodules, an embedded nodule among large pulmonary vessels presents a difficult case. Improved segmentation methods utilizing information such as vessel tracking to remove vessels and new features will have to be investigated in future studies.

We are improving our current method by expanding our data set and analyzing the classification performance of different types of nodules. For example, nodules from primary lung cancer may have different characteristics than metastatic nodules, although both types would be considered malignant. Our long term goal is to aid the differentation of malignant and benign nodules detected in screening, making identification of primary lung cancer of prime importance. Thus, training classifiers that are specific to the features of primary lung cancer may improve the performance of the classifier in the screening population.

Another aspect of our current system that needs improvement is the method to determine which object is the nodule in the VOI. Currently we choose the largest object close to the center, since the VOI's were marked by radiologists. However, if a combined detection and classification system is to be developed in the future, the VOI may not center at the automatically detected object, and the object may not be the largest one in the VOI. More intelligent methods for differentiating nodules from other normal lung structures in the VOI segmented by clustering will have to be investigated.

Although the use of leave-one-case-out validation results is a commonly accepted approach in CAD literature because of the difficulty of collecting a large enough database for training, validation, and independent testing, it is prudent to keep in mind that the performance of the CAD system may not be considered generalizable to the patient population until its performance is verified with a truly independent test set that is not seen by the CAD system or the trainer during the developmental process. Our test results show comparable performance between CAD and radiologists' assessment of the likelihood of malignancy of the nodules. In a clinical situation, radiologists may be able to utilize other information such as patient history and clinical data, in addition to image data, to assess the likelihood of malignancy. An advanced CAD system may also merge all available information into a diagnostic recommendation. At the current stage, we focus on optimizing the use of image data to extract diagnostic information to avoid the masking of the image information by other dominant risk factors such as smoking history or age. Further, CAD is not intended to be used as a stand-alone diagnostic tool. After an effective classifier is designed, it is necessary to determine whether radiologists would improve their classification of lung nodules with CAD.

2.6 Conclusion

Our results demonstrate that 3D AC can segment lung nodules automatically. Automated feature extraction from the segmented boundary and classification can achieve an accuracy comparable to that of an experienced radiologist. The computer classifier thus has the potential to provide a second opinion to radiologists for assessing the likelihood of malignancy of a lung nodule. When the 3D AC trained with our clinical data set was applied to the LIDC data set, the segmented volumes by the computer algorithm were in general larger than those manually segmented by the radiologists. It remains to be investigated whether the 3D AC model trained using gold standard boundaries at a given threshold such as those provided by the LIDC database can achieve higher classification accuracy than that achieved with our current approach. Comparison of the two approaches will be pursued when a large data set is available in the LIDC database.

2.7 Tables

Table 2.1: Comparison of classification performance of the classifiers, in terms of A_z and $A_z^{0.9}$ obtained from leave-one-case-out testing. The classifiers were designed with different feature sets or different FOMs during simplex optimization.

Methods	A_z as FOM		$A_z^{0.9}$ as FOM	
	A_z	$A_z^{0.9}$	A_z	$A_z^{0.9}$
Computer Classifier	0.83 ± 0.04	0.30	0.78 ± 0.05	0.35
Feature Descriptors by Radiologist	0.80 ± 0.05	0.24	0.82 ± 0.04	0.32
Likelihood of Malignancy by Radiologist	0.84 ± 0.04	0.33		

2.8 Figures

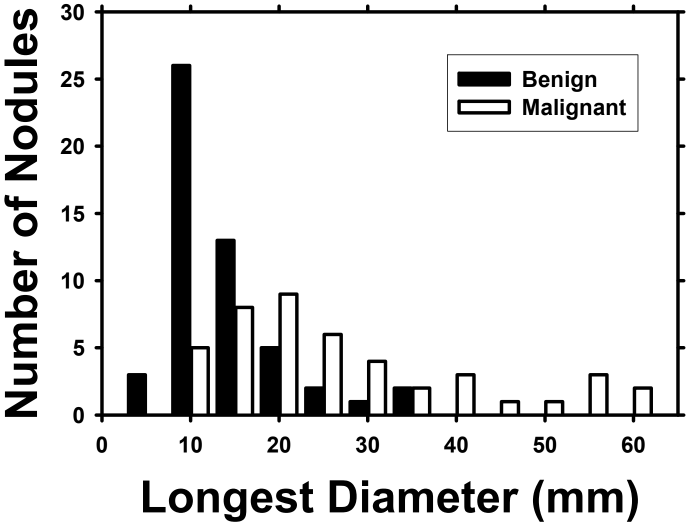


Figure 2.1: Distribution of the longest diameters of the lung nodules in the data set, as measured by experienced thoracic radiologists on the axial slices of the CT examinations.

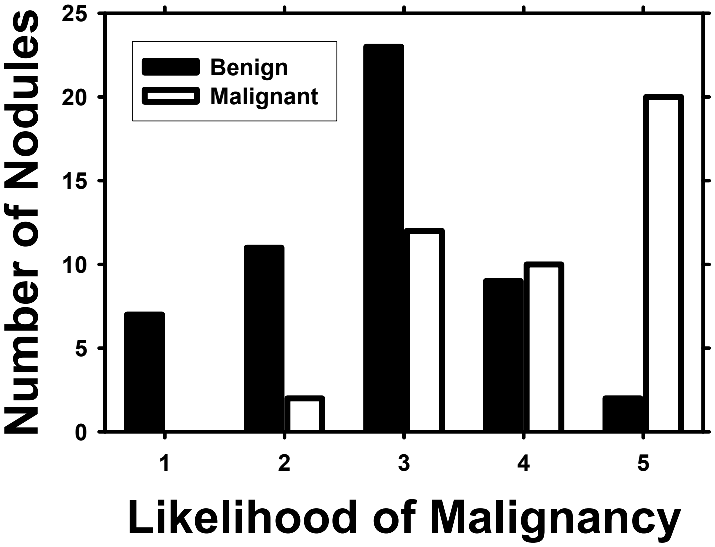


Figure 2.2: Confidence ratings for the likelihood of a nodule being malignant (1=most likely benign, 5=most likely malignant) by experienced thoracic radiologists.

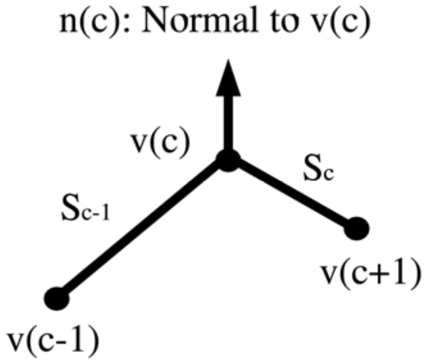


Figure 2.3: The vertices of the polygon and positions used in the active contour model.

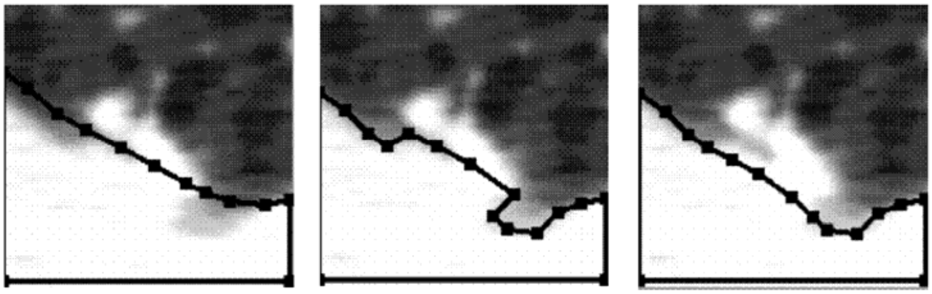


Figure 2.4: An example demonstrating the correction of lung segmentation from the pleural surface. From left to right: the initial pleural boundary, indentation created along the lung boundary after local refinement, and corrected lung boundary after indentation is filled.

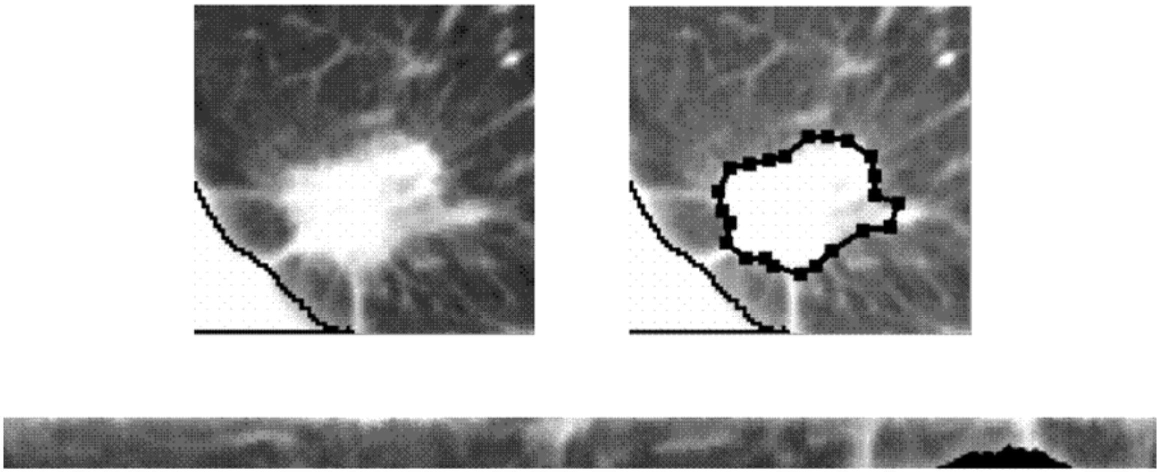


Figure 2.5: The Rubber Band Straightening Transform (RBST). Top Left: An ROI containing a nodule. Top Right: The active contour boundary from which the RBST image is extracted. Bottom: The RBST image that will be Sobel-filtered, from which run-length statistics may be extracted. The black area of the RBST image corresponds to the pixels where the chest wall is masked out.

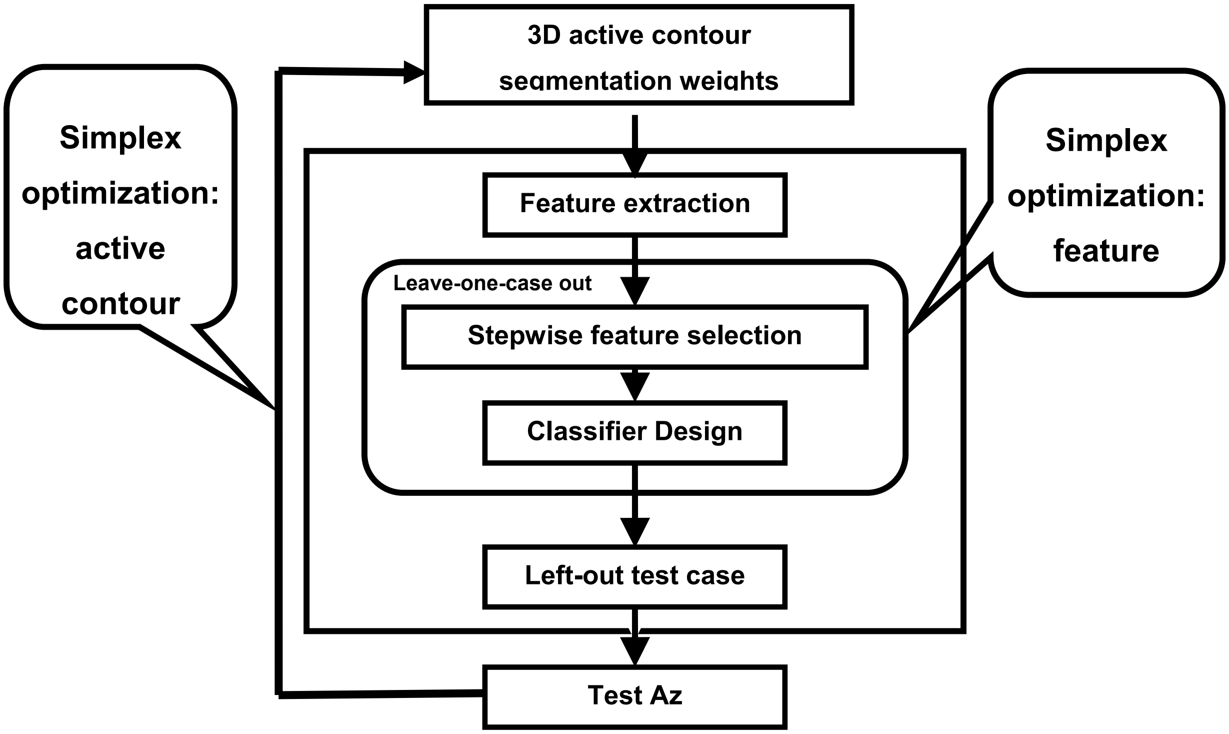

3D active contour segmentation weights
Simplex optimization: active contour
Simplex optimization: feature
Feature extraction
Leave-one-case out
Stepwise feature selection
Classifier Design
Left-out test case
Test Az

Figure 2.6: Flow chart showing the simplex optimization process for selection of weights in the 3D AC model and classifier design.

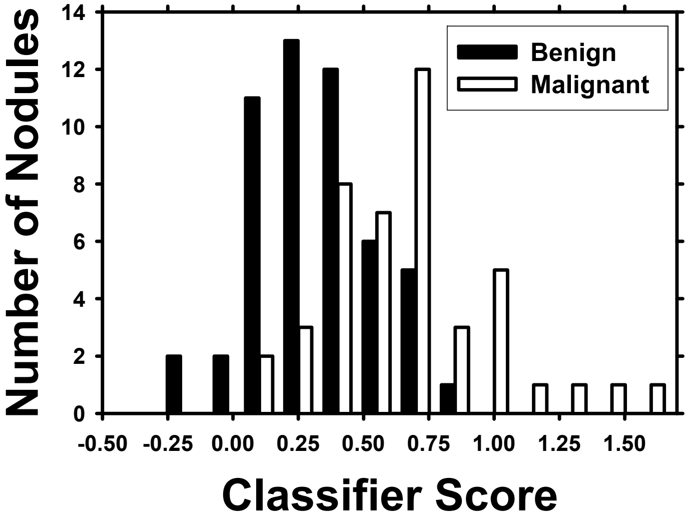


Figure 2.7: Test discriminant scores of lung nodules from the leave-one-case-out segmentation training and testing method.

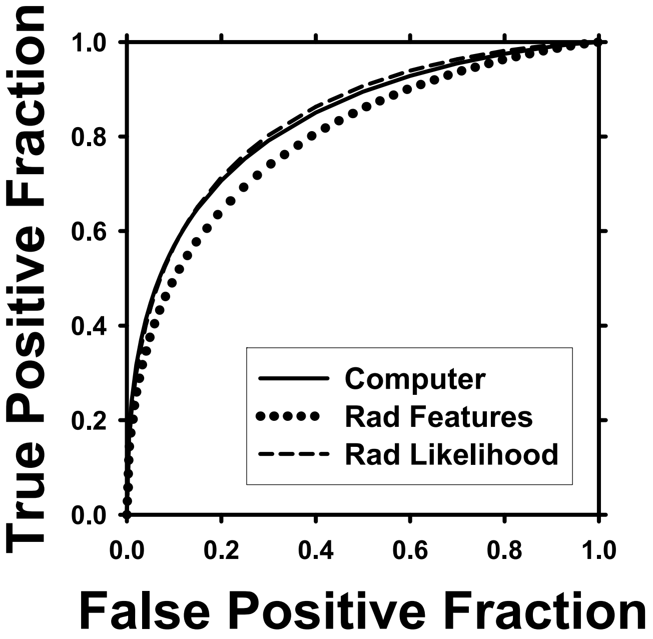
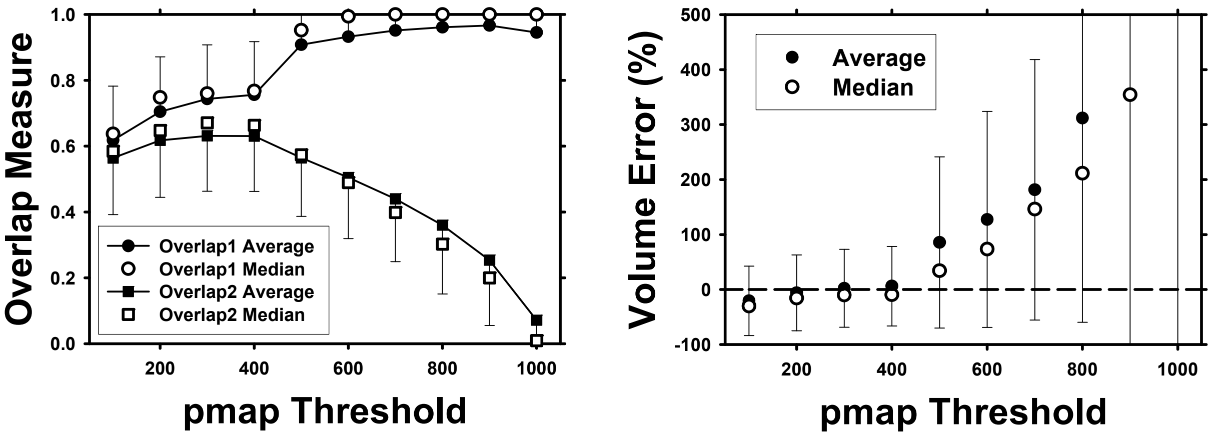


Figure 2.8: ROC Curves comparing the different results for optimization using A_z as FOM. Computer A_z=0.83, Rad Features A_z =0.80, Rad Likelihood A_z =0.84.



(a)

(b)

Figure 2.9: Overlap measures (a) and volume percentage errors (b) at different pmap thresholds for testing of the 3D AC segmentation using the 23 LIDC nodules. The error bars indicate one standard deviation from the average (only one side shown for clarity). Two overlap measures are shown: *$Overlap_1(A,L)$* relative to the gold standard volume and *$Overlap_2(A,L)$* relative to the union of the segmented volume and the gold standard volume. Note the increasing volume error as the pmap threshold increases because the LIDC-defined nodule volume decreases with increasing pmap threshold values. A pmap value of 1000 means the intersection of all 18 LIDC manually and semi-automatically drawn contours by radiologists. Eight of the nodules contain no voxels in the intersection at pmap of 1000 so that the average and the median were calculated from the remaining 15 nodules.

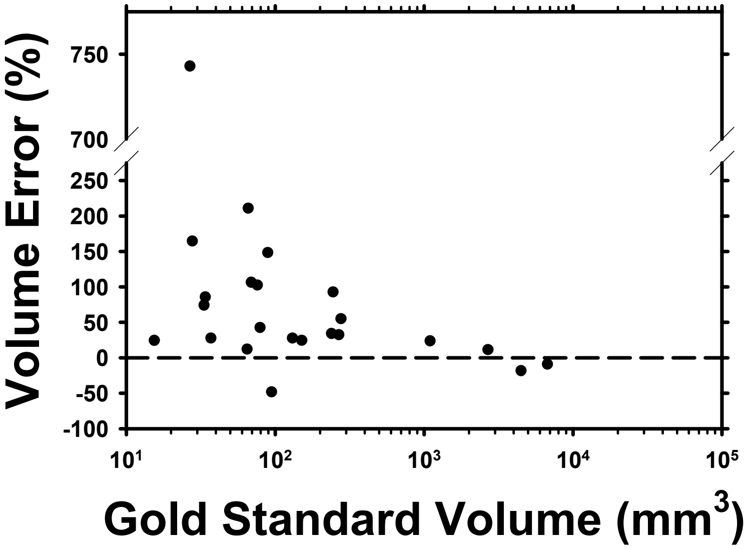


Figure 2.10: Percentage of volume error relative to the volume (in log scale) enclosed within the contour defined by a pmap threshold of 500 for each of the 23 LIDC test nodules. One small juxta-vascular nodule had a volume error of 743% because the blood vessel was erraneously segmented.

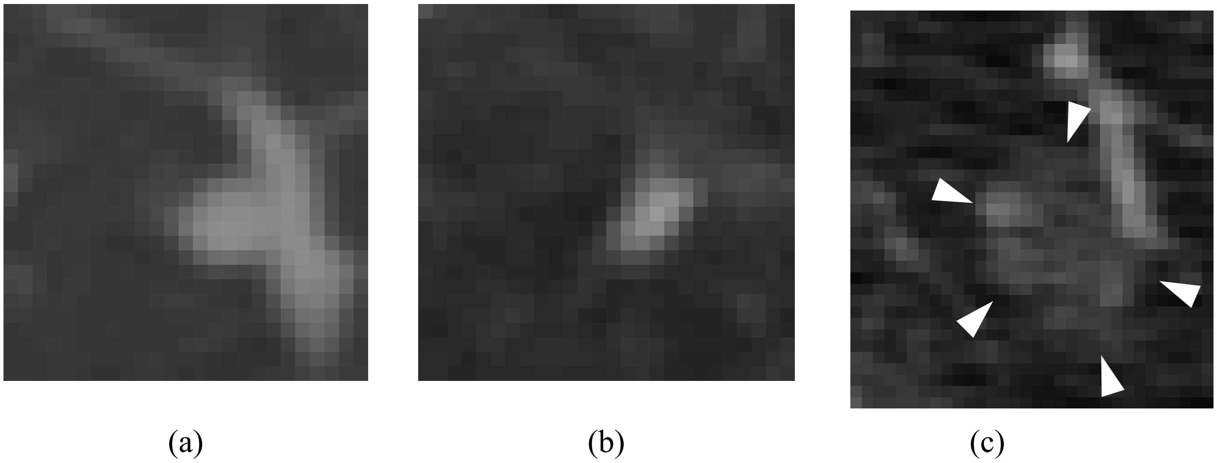
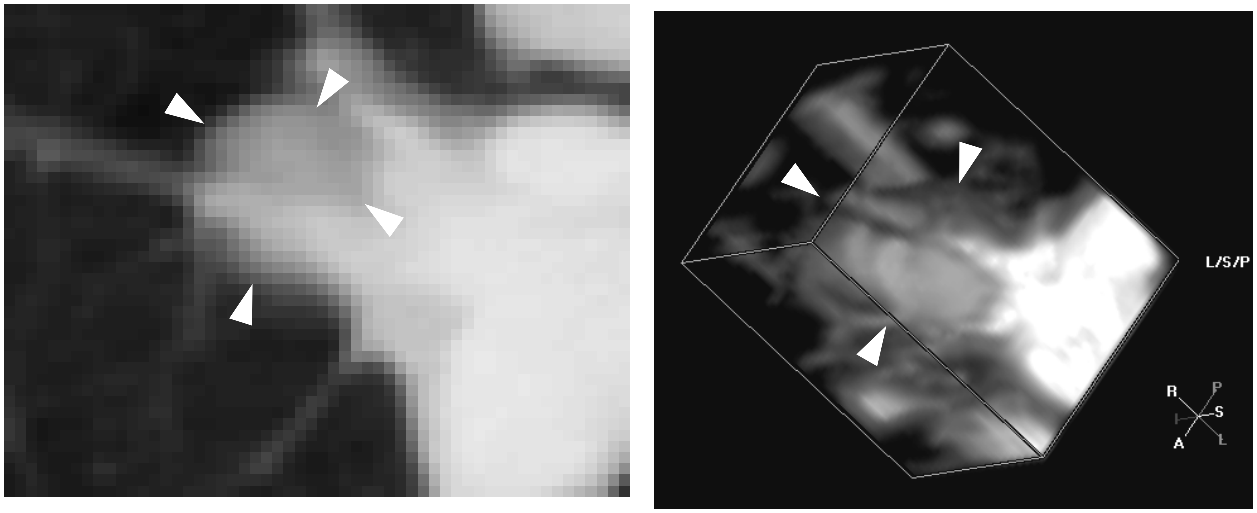


Figure 2.11: Representative slices (not to scale) from difficult-to-segment LIDC nodules: (a) small, faint juxtavascular nodule (longest diameter 4.32 mm), (b) small nodule (longest diameter 4.98 mm), and (c) juxtavascular (longest diameter 11.92 mm) low contrast nodule in noisy image.



(a) (b)

Figure 2.12: An example of a nodule which was difficult to segment because it was embedded in thick blood vessels, leading to inaccurate classification. (a) axial slice through the nodule, and (b) 3D volume containing the nodule.

2.9 Appendices

2.9.1 Appendix 1: RLS Features

The Gallaway run-length features are described below, where $p(i,j)$ is the run-length matrix that stores information on the number of runs with pixels of gray-level i and run length j. M is the number of gray levels, N is the number of runs, n_r is the total number of runs, and n_p is the number of pixels in the image:

- Short Run Emphasis (SRE)

$$SRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{j^2} \qquad (2.A1)$$

- Long Run Emphasis (LRE)

$$LRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j) \cdot j^2 \qquad (2.A2)$$

- Gray-Level Nonuniformity (GLN)

$$GLN = \frac{1}{n_r} \sum_{i=1}^{M} \left(\sum_{j=1}^{N} p(i,j) \right)^2 \qquad (2.A3)$$

- Run Length Nonuniformity (RLN)

$$RLN = \frac{1}{n_r} \sum_{j=1}^{N} \left(\sum_{i=1}^{M} p(i,j) \right)^2 \qquad (2.A4)$$

- Run Percentage (RP)

$$RP = \frac{n_r}{n_p} \qquad (2.A5)$$

Dasarathy and Holder presented four more features [66]. These are based on the idea of joint statistical measures of the gray levels and run length:

- Short Run Low Gray-Level Emphasis (SRLGE)

$$SRLGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{i^2 \cdot j^2} \qquad (2.A6)$$

- Short Run High Gray-Level Emphasis (SRHGE)

$$SRHGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j) \cdot i^2}{j^2} \tag{2.A7}$$

- Long Run Low Gray-Level Emphasis (LRLGE)

$$LRLGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j) \cdot j^2}{i^2} \tag{2.A8}$$

- Long Run High Gray-Level Emphasis (LRHGE)

$$LRHGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j) \cdot i^2 \cdot j^2 \tag{2.A9}$$

2.9.2 Appendix 2: Segmentation Performance Metrics

By combining the overlap measure *$Overlap_1(A,L)$* in Eqn. (2.9) and the percentage volume error (V_{err}) in Eqn. (2.11), one can define a number of performance metrics that quantify the number of voxels correctly and incorrectly segmented as a part of the object, using the gold standard object as a reference:

- **True positive fraction (*TPF*)**: the fraction of the voxels that are in the gold standard object and are included in the segmented object:

$$TPF = Overlap_1(A,L) \tag{2.A10}$$

- **False positive ratio (*FPR*):** the ratio of the voxels that are in the segmented object but not in the gold standard object, relative to the gold standard object

$$FPR = V_{err} + [1\text{-}\ Overlap_1(A,L)] \tag{2.A11}$$

- **False negative fraction (*FNF*)**: the fraction of the voxels that are in the gold standard object but not included in the segmented object:

$$FNF = 1\text{-}\ Overlap_1(A,L) \tag{2.A12}$$

- **Non-overlapping volume ratio (*NOVR*):** the ratio of voxels in the gold standard object and in the segmented object that do not overlap, relative to the gold standard object:

$$\begin{aligned} NOVR &= V_{err} + 2^*[1\text{-}\ Overlap_1(A,L)] \\ &= FPR+FNF \end{aligned} \tag{2.A13}$$

The equations above use *$Overlap_1(A, L)$*, but the conventional overlap measure in Eqn. (2.10) can also be expressed in terms of *$Overlap_1(A, L)$* and *NOVR*:

$$Overlap_2(A,L) = \frac{Overlap_1(A,L)}{NOVR + Overlap_1(A,L)} \tag{2.A14}$$

Therefore, the two measures *$Overlap_1(A, L)$* and V_{err} provide a complete description of the segmentation performance. As an example, we have plotted the metrics described above in Fig. 2.A1 as derived from the average *$Overlap_1(A, L)$* and V_{err} values shown in Fig. 2.9.

From Eqn. (A13), it is seen that *Overlap$_1$(A, L)* can be expressed in terms of *Overlap$_2$(A, L):*

$$TPF = Overlap_1(A,L) = \frac{Overlap_2(A,L) \cdot \left(V_{err} + 2\right)}{1 + Overlap_2(A,L)} \qquad (2.A15)$$

and thus *Overlap$_2$(A, L)* in conjunction with V_{err} can also provide similar performance metrics defined above, although the relationships are more involved. These analyses indicate that it is important to include the percentage volume error as a complement to either of the overlap measures. These metrics were calculated for the LIDC data set as shown in Fig. 2.A1.

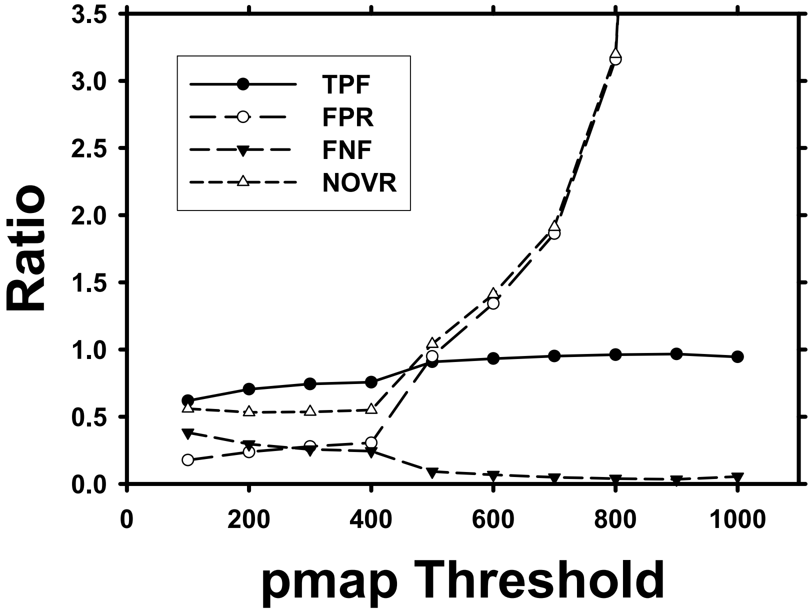


Figure 2.A1: The performance metrics *TPF, FPR, FNF,* and *NOVR* derived from the average *Overlap$_1$(A,L)* and V_{err} measures shown in Fig. 2.9.

Chapter 3
Effect of CT Scanning Parameters on Volumetric Measurements of Pulmonary Nodules by 3D Active Contour Segmentation: A Phantom Study

3.1 Abstract

The purpose of this study is to investigate the effects of CT scanning and reconstruction parameters on automated segmentation and volumetric measurements of nodules in CT images. Phantom nodules of known sizes were used so that segmentation accuracy could be quantified in comparison to ground-truth volumes. Spherical nodules having 4.8, 9.5, and 16 mm diameters and 50 and 100 mg/cc calcium contents were embedded in lung-tissue-simulating foam which was inserted in the thoracic cavity of a chest section phantom. CT scans of the phantom were acquired with a 16-slice scanner at various tube currents, pitches, fields-of-view, and slice thicknesses. Scans were also taken using identical techniques either within the same day or five months apart for study of reproducibility. The phantom nodules were segmented with a 3-dimensional active contour (3DAC) model that we previously developed for use on patient nodules. The percentage volume errors relative to the ground-truth volumes were estimated under the various imaging conditions. There was no statistically significant difference in volume error for repeated CT scans or scans taken with techniques where only pitch, field-of-view, or tube current (mA) was changed. However, slice thickness significantly ($p<0.05$) affected volume error. Therefore, to evaluate nodule growth, consistent imaging conditions and high resolution should be used for acquisition of the serial CT scans, especially for smaller nodules. Understanding the effects of scanning and reconstruction parameters on volume measurements by 3DAC allows better interpretation of data and assessment of growth. Tracking nodule growth with computerized segmentation methods would reduce inter- and intraobserver variabilities.

3.2 Introduction

Computed tomography (CT) is more sensitive than chest radiography in detecting small and subtle lung nodules [6-13]. With the advent of multi-detector row helical CT scanners that offer ever thinner slices, even more nodules are detected [75]. As it is recommended to follow up on most [76], if not all [73, 77] detected nodules, the management of these nodules may overwhelm radiologists, especially if lung cancer screening with CT becomes recommended practice.

Computer-aided diagnosis (CAD) tools have the potential to assist radiologists as a second reader in detecting, classifying (characterizing), and managing lung nodules. Currently, a CAD system can analyze a single scan, segmenting and then extracting features from the nodule to estimate its likelihood of malignancy. Features such as morphology (volume, size, and shape) [1, 49, 50], histograms of grey-levels, the contrast of the nodule [51], enhancement after contrast injection [78], and texture [1] have been used. There has also been increasing attention given to volume doubling rate as an indicator of malignancy [79, 80].

To track volume growth, an accurate boundary of the nodule is needed, and it is challenging for many reasons. First, it is difficult to determine nodule boundaries definitively, as suggested by the considerable inter- and intra-observer variabilities among radiologists in assessing lung nodule size [81-83]. Second, the method used to measure volume may vary. Nodule size can only be measured in two dimensions on a CT slice, typically based on its longest and perpendicular axes. It is difficult to assess doubling time, especially for smaller nodules, for which a small change in diameter corresponds to a large change in the volume. For example, a sphere with a 3 mm diameter doubles in volume when the diameter increases by only 26% to 3.78 mm. This 0.78 mm increase is about the size of 1.1 pixels in a typical scan with a field-of-view (FOV) of 36 cm.

Different window width and level settings can also affect the measurement of diameters [84]. These factors contribute to inaccuracies in the measured volume [85], resulting in inconsistency and uncertainty in detecting volume change in serial CT scans. The large inter- and intra-observer variations in radiologists' manual segmentation demonstrated in the 23 nodules provided by the Lung Image Database Consortium (LIDC) highlight the difficulty in judgment of nodule boundaries on CT scans [83]. Yankelevitz *et al.* found that computerized methods are accurate on phantom nodules and applied the methods to clinical patient scans over time. They concluded that computerized methods using CT volumetric information that are not affected by windowing settings or intra- and inter- observer errors will be useful for nodule volume estimation and for assessing growth [86, 87].

Even though computerized methods may be immune to some factors that may influence radiologists, such as variation in windowing, there are still challenges due to image acquisition parameters [88, 89]. Various groups have investigated the effects of image acquisition parameters using nodule-mimicking spheres with known volumes. Ko *et al*. reported computer calculated volumes from regions of interest (ROI's) marked by radiologists. Using a threshold method for segmentation, they found that tube current-time (20 and 120 mAs), reconstruction algorithm, quantitative volume calculation method, nodule attenuation (solid at 50 HU and ground-glass at -360 HU), and nodule size significantly affected volume error [90]. Goo *et al.* measured volumes using a thresholding method and reported that section thickness and threshold significantly affected absolute volume error [91].

The major differences between this investigation and previous studies [84, 85, 90, 91] include the segmentation algorithm, the greater ranges of scanning and reconstruction parameters and nodule sizes, and the effect of calcium concentration of the nodules. In addition, we investigated the reproducibility of repeated scans performed either consecutively or separated by a few months (five in this study) as in clinical follow-up studies. An automated nodule segmentation method was used so the volume errors were assessed free of inter- and intra-observer variabilities.

Although phantom nodules are different from real pulmonary nodules in many ways, phantom studies will allow comparison with the ground-truth volumes and systematic analysis of the relative trends of the volume error dependence on CT imaging conditions. This would be virtually impossible to perform with nodules in patient scans. Besides dose concerns, motion artefacts can be a major variable, making it difficult to determine what actually contributes to the measured volume errors. The fixed size of chest phantoms also precludes nodule volume variations due to lung volume change when breathing [92]. Despite the fact that the absolute volume errors estimated with phantoms will not be applicable to real nodules, the results will provide useful information for the selection of CT imaging protocols for clinical follow-up and for the interpretation of nodule volume changes.

3.3 Methods and Materials

3.3.1 Chest Phantom and Spherical Nodules

The chest phantom consisted of a CIRS Model 003 tissue equivalent transaxial thorax section phantom (CIRS, Norfolk, VA) sandwiched between two large water-equivalent bolus sections [27]. The body shape and lung cavities of the water bolus sections match those of the thorax section. The lung cavities in the thorax section were filled with a foam that has similar CT number Hounsfield Units (HU) as lung tissue in CT scans [27]. Nodule-mimicking spheres of diameters 4.8 mm, 9.5 mm, and 16 mm were scanned to evaluate the dependence on nodule size. These spheres were made of an epoxy resin with added fillers that would produce x-ray linear coefficients nearly identical to that as water in the CT energy range. The amount of calcium carbonate ($CaCO_3$) added to the spheres determined their calcium density. These sizes were chosen by taking into consideration the nodule sizes that are of clinical significance (3 to 30 mm) and the expected larger errors in volume estimation for smaller nodules in CT scans. They were inserted in slits cut in the foam of the lung section. The interface between sections was filled with Vaseline to minimize air gaps.

In Experiment I, five 4.8-mm-diameter 50 mg/cc $CaCO_3$ spheres were placed in the left lung cavity, and six 4.8-mm-diameter 100 mg/cc $CaCO_3$ spheres were placed in the right lung cavity. In Experiment II, scans taken five months later, 10 nodules of the same size were placed in the thorax section phantom. Five $CaCO_3$ spheres of 50 mg/cc were placed in the right lung cavity, and five 100 mg/cc spheres were placed in the left lung cavity. After scanning was completed for one size (e.g. 4.8 mm), all 10 spheres were replaced with those of another size (e.g. 9.5 mm). Examples of the CT scans acquired are shown in Figure 3.1.

3.3.2 Reproducibility of Lung Nodule Volume Measurement

CT scans were acquired with a GE LightSpeed VCT 64 slice scanner operating in 20 mm collimation mode (GE Healthcare, Waukesha, WI). Images were acquired in two sessions (Experiment I and Experiment II) five months apart. Within each experiment, three scans (Scans 1, 2, and 3) using identical parameters were taken for each imaging condition. The same imaging conditions were used in Experiments I and II to simulate serial exams where a follow-up scan would be taken months later. Thus, variability between the two experiments and within each experiment can be assessed.

We focused on protocols used with GE 16-slice scanners in two clinical trials. These conditions were employed because neither trial had published protocols for the new 64-slice scanners at the time we initiated our study, and the 64-slice scanner could be operated in a 20-mm collimation instead of 40-mm collimation mode, which is similar to that of the 16-slice scanners. The first source of parameters was the National Lung Screening Trial (NLST), sponsored by the National Cancer Institute (NCI), which enrolled over 50,000 high-risk subjects to compare the effectiveness of CT and chest radiography in lung cancer detection. Scans on GE 16-slice scanners were acquired with the 16 x 1.25 mm detector configuration, 120 kVp, 80 mA, 0.5 sec rotation time, and 1.375:1 pitch. The images were reconstructed to 1.25 mm slice thickness and 1.25 mm slice interval for a thin slice data set, in addition to the 2.5 mm slice thickness and 2.0 mm slice interval data set. The second source was the Lung Tissue Resource Consortium (LTRC), sponsored by the National Heart, Lung and Blood Institute, the goal of which was to create a lung tissue database to understand the pathogenetic mechanisms of lung diseases. CT scans were collected as part of the process. Scans on GE 16-slice scanners were acquired at 140 kVp, 300 mA, and 0.5 sec rotation time, 1.375:1 pitch. The data images were reconstructed at 0.625 mm slice thickness and 0.625 mm slice interval, and a retrospective reconstruction was also done at 1.25 mm thickness and 0.625 mm interval. Since both are large scale studies, and participants in the NLST underwent multiple scans over time, accuracy and reproducibility under these CT protocols are of interest. An additional high resolution imaging condition was chosen in anticipation of thin-slice scans that may become more common with multi-detector row CT scanners. All scans were reconstructed with the Standard kernel in the GE scanners. The imaging conditions for the reproducibility study are listed in Table 3.1.

3.3.3 Effects of CT Parameters on Volume Errors

For the study of the effects of CT scanning and reconstruction parameters on nodule volume estimation, each parameter was varied in a range that may be used in clinical examinations, including the pitch (0.531:1 to 1.375:1), slice thickness (0.625 mm to 2.5 mm), tube current (80 mA to 400 mA), and field-of-view (25 cm to 36 cm).

3.3.4 Nodule Segmentation

3.3.4.1 Image Pre-processing

A CT scan input to the segmentation program is first pre-processed to obtain isotropic voxels. Linear interpolation in the *z*-axis is performed if the reconstructed slice interval of the CT scan is greater than the pixel size on the axial plane. Bilinear interpolation in the axial plane is used instead if the slice interval is smaller. The interpolation does not improve the spatial resolution of the CT data, but it is performed to facilitate the implementation of the 3-dimensional (3D) segmentation operations.

Using a graphical user interface, each nodule location is manually identified by setting a box that contains the nodule on one scan based on visual inspection. For other scans where the nodules are in the same positions, an automatic nodule detection program extracts volumes-of-interest (VOIs) from the scan [39, 93]. The automatic detection is not a part of this study but is used to reduce the time of otherwise manually marking VOIs. For nodules that are close to the pleura, the VOI may include voxels from the pleura and the chest wall. For these nodules, a lung region masking program previously developed in our laboratory [39] is applied to the VOI to exclude these voxels from further processing.

3.3.4.2 Parametric Active Contour (AC) Segmentation

Deformable contour models, particularly the AC model [14], are well-known tools for image segmentation. ACs are energy-minimizing splines guided by various forces, or energies. We implemented a parametric model, representing the contour as vertices of a polygon, and minimized it based on a greedy algorithm described by

Williams and Shah [15]. We implemented two new energies to take advantage of 3D volumetric data and one energy to penalize the total energy if the contour grows into the chest wall and mediastinum [1]. The energy function is shown in Equation (1).

$$E_{total} = \min_{E(c)} \sum_{c=1}^{N} [w_{\text{hom}} E_{\text{hom}}(c) + w_{cont} E_{cont}(c) + w_{curv} E_{curv}(c) + w_{3Dcurv} E_{3Dcurv}(c) + \\ w_{grad} E_{grad}(c) + w_{3Dgrad} E_{3Dgrad}(c) + w_{bal} E_{bal}(c) + w_{mask} E_{mask}(c)] \qquad \textbf{\textit{(3.1)}}$$

where E(c) is the energy at a pixel in the search neighbourhood of vertex $v(c)$. In this energy function, the internal energies include homogeneity (*hom*), continuity (*cont*), curvature (*curv*), 3D curvature (*3Dcurv*), and the external energies include gradient (*grad*), 3D gradient (*3Dgrad*), balloon (*bal*), and mask (*mask*). The weight w_j is a parameter assigned to each energy *j*, where *j* represents one of the eight energies: *hom, cont, curv, 3Dcurv, grad, 3D grad, bal, and mask.* Details of how the weights were trained have been described elsewhere [1].

In this study, we applied the 3DAC model trained on patient nodules [1] to the segmentation of phantom nodules on CT scans. An initial contour for the nodule in the VOI is generated using a *k*-means clustering and object identification method and is used to initialize the 3DAC as explained in detail in [1].

3.3.5 Manual Segmentation by a Radiologist

To provide a reference for comparison with the 3DAC contours, a fellowship-trained thoracic radiologist manually segmented some 4.8-mm-diameter nodules. We selected three scanning conditions, which were identical except for slice thickness values of 0.625, 1.25, and 2.5 mm. The other parameters were 1.375:1 pitch, 0.625 mm slice interval, 120 kVp, 160 mA, 36 cm FOV, and 0.5 sec rotation time. The radiologist segmented the five 4.8-mm-diameter nodules with 100 mg/cc density in the right lung of the chest phantom. There were a total of 15 manually segmented nodules from the five nodules and three different slice thickness conditions.

The radiologist was only informed that the phantom nodules were spherical, without being given any size information. A graphical-user interface (GUI) program developed in our laboratory displayed a zoomed in ROI containing the nodule to be segmented. The radiologist was free to adjust the brightness and contrast of the image.

He was instructed to outline the nodule by a polygon following where he judged to be the nodule boundary in the image.

3.3.6 Data Analysis

The volume of a segmented phantom nodule (V_{3DAC}) was calculated by multiplying the number of voxels contained in the contours with the volume of each voxel. The segmented volume was compared with the ground-truth volume of the phantom nodule (V_{true}) and the difference was reported as the percentage volume error given in Equation (2):

$$\%\, Vol\, err = \frac{V_{3DAC} - V_{true}}{V_{true}} \times 100\% \qquad \textbf{\textit{(3.2)}}$$

where V_{true} is calculated by the volume formula of a sphere, Vol=(4/3)πr^3. The volume error percentage is a signed value: positive for overestimation and negative for underestimation of the true volume.

The volume errors were calculated for all nodule sizes imaged under various CT imaging conditions. We estimated statistical significance based on Student's *t*-test for paired data and one-way ANOVA for groups of data. Because multiple comparisons were made for pairs of conditions and their statistical significance was estimated, the Bonferroni correction [94] procedure was used, when appropriate, to adjust the threshold for the p value for statistical significance, which is usually set at 0.05 without this correction.

3.4 Results

3.4.1 Lung Nodule Volume Reproducibility

All nodule-mimicking spheres were successfully segmented as judged by visual inspection. Tables 3.2 and 3.3 list the average volume errors calculated from the 3DAC boundary using Eqn. (2) for the 4.8-mm-diameter spheres from Experiments I and II, respectively. The same imaging conditions were used in the two experiments conducted five months apart. The volume errors in these tables were analyzed separately for each density (either 50 or 100 mg/cc of $CaCO_3$) and each scan. Six averages for each imaging condition are listed for each experiment, since there were three scans taken for each condition and two densities for each.

3.4.1.1 Analysis by Density

The two-tailed paired *t*-test was performed where each pair consisted of the average volume error of the 50 mg/cc and the 100 mg/cc densities in one scan. There were five different imaging conditions in both Experiments I and II, each of which had three identical scans, for a total of 30 pairs for statistical analysis. The volume error difference between the densities was not statistically significant ($p = 0.845$).

3.4.1.2 Reproducibility Analysis

To determine whether there were differences in the volume errors among the three identical scans acquired on the same day, analysis of variance (ANOVA) was performed on the three identical scans. Each scan contained 11 (Experiment I) or 10 (Experiment II) nodules for each of the five different imaging conditions in Experiments I and II. For a given imaging condition, the difference in average volume errors was not statistically significant ($p > 0.05$).

To determine reproducibility for scans separated by a few months, ANOVA was used to analyze the volume errors from the six scans (three from each experiment) taken under identical imaging conditions, each of which contained 11 or 10 nodules. For each of the five imaging conditions in Experiments I and II, the difference in the volume measurements ($p > 0.05$) between the scans acquired five months apart did not achieve statistical significance.

3.4.1.3 Volume Error Variability

Table 3.4 shows a comparison of the volume errors averaged over all three scans for a given density and imaging condition. Higher variability of volume errors occurred with larger slice thicknesses and slice intervals. The greatest volume errors occurred for the scan using 2.5 mm slice thickness and 2.0 mm slice interval, which were the lowest resolution scans in this reproducibility study.

3.4.2 Effects of CT Parameters on Volume Errors

3.4.2.1 Dependence on Slice Thickness

The average volume errors for varying slice thicknesses of 0.625 mm, 1.25 mm, and 2.5 mm were compared for tube currents ranging from 80 to 320 mA. The other parameters were fixed at 1.375:1 pitch, 0.625 mm slice interval, 120 kVp tube voltage, 0.5 sec rotation time, and 36 cm FOV. The dependence of the average volume errors on nodule size is shown in Figure 3.2 for a tube current of 160 mA as an example. The trends are similar within the tube current range of 80 to 320 mA studied. The average volume errors of the 4.8, 9.5, and 16 mm spheres for all slice thicknesses and tube currents ranged from 19.4% to 45%, 9.9% to 23.5%, and -0.8% to 1.7%, respectively.

For the 160 mA case, there were three different slice thickness values and three nodule sizes, resulting in nine paired t-tests. A p-value of less than (0.05/9 = 0.0056) was considered statistically significant after application of the Bonferroni correction. There was no statistically significant difference in the average volume error ($p > 0.0056$) when the slice thickness changed from 0.625 to 1.25 mm for all nodule sizes. However, a change in slice thickness from 1.25 mm to 2.5 mm resulted in a statistically significant ($p < 0.0056$) difference in the average volume error for the 4.8 mm and 16 mm nodules, but not for the 9.5 mm nodules ($p = 0.0073$). A change in slice thickness from 0.625 mm to 2.5 mm affected average volume error significantly ($p < 0.0056$) for all nodule sizes. The results for the various nodule sizes and all three tube currents are summarized in Table 3.5.

3.4.2.2 Dependence on Tube Current (mA)

For each slice thickness, three different tube currents of 100, 200, and 400 mA were used to scan the three different nodule sizes. Table 3.6 lists the standard deviations of the HU values in several regions-of-interest of a slice intersecting approximately the center of the nodules in each scan. The standard deviations represented the relative noise levels in the CT scans due to the tube current changes. As expected, the noise decreased as the slice thickness increased and the tube current decreased. It was found that changing the noise level within the range studied did not significantly ($p > 0.05$) affect the average volume error of the nodules for all nodule sizes.

For the 4.8 mm spheres, the volume errors for varying slice thickness and tube current are shown in Figure 3.3. The parameters fixed were 0.531:1 pitch, 0.625 mm slice interval, 120 kVp, 0.8 sec rotation time, and 36 cm FOV. Note that the pitch and rotation time for this set of parameters differ from those used for the data shown in Figure 3.2 and Table 3.5, although the trends are similar. With a slice thickness of 0.625 mm, the average volume errors were 21.6%, 20.8%, and 21.7% for 100, 200, and 400 mA scans, respectively. When the slice thickness was increased to 2.5 mm, the corresponding errors increased to 42.3%, 43.3%, and 40.7% for the three tube currents, respectively.

There was no statistically significant change (all $p > 0.6$) in average volume errors when slice thickness changed from 0.625 to 1.25 mm for the 100, 200, and 400 mA scans. However, further increase in slice thickness from 0.625 mm or 1.25 mm to 2.5 mm resulted in a statistically significant ($p < 0.0056$) difference in average volume error for all tube currents.

3.4.2.3 Dependence on Pitch

The effects of varying pitch from 0.531:1, 0.969:1, and 1.375:1 are shown in Figure 3.4. The CT parameters were fixed at 0.625 mm slice thickness and slice interval, 120 kVp tube voltage, 400 mA tube current, 0.8 sec rotation time, and 36 cm FOV. The volume error decreased as the nodule size increased, similar to the trends observed in Figure 3.2. The average volume errors of the 4.8, 9.5, and 16 mm spheres ranged from 20.6% to 21.7%, 10.4% to 10.6%, and -0.4% to -0.6%, respectively. The variation in pitch did not significantly affect the volume error of segmentation ($p > 0.05$) within the range studied for all nodule sizes.

3.4.2.4 Dependence on Field-of-View

The effects of varying FOV values from 25 cm to 36 cm on volume error are presented in Figure 3.5. Note that increasing the FOV increases the pixel size on the axial plane and also may affect the interpolated voxel size used for the segmentation. The other parameters were fixed at 0.531:1 pitch, 0.625 mm slice thickness and interval, 120 kVp tube voltage, 400 mA tube current, and 0.8 sec rotation time. The CT scan containing 4.8 mm spheres with 25 cm FOV were not used because the beginning and end slice acquisition location were set erroneously, resulting in spheres that were not completely scanned. The average volume errors of the 4.8, 9.5, and 16 mm spheres ranged from 20.3% to 22.4%, 10.3 to 12.4%, and -1.0% to -0.1% respectively. There was no statistically significant change ($p > 0.006$, Bonferroni correction) in the average volume errors within the range of FOV studied.

3.4.3 Partial Volume Effects on Volume Errors

The overestimation in nodule volume is mainly caused by partial volume effects. Typical segmented contours for the 4.8 mm nodule at three slice thicknesses and the same 0.625 mm slice interval are shown in Figure 3.6, rows (1)-(3). The segmented boundaries are visually reasonable although the average volume errors of 20% to 45% (see Figure 3.3) seem excessive. The segmented volume increases as slice thickness increases although the true nodule size is the same. To further demonstrate the effect, a spherical object of radius 4.8 mm was digitally generated and shown in Figure 3.6, row (4). Comparison of both the digitally generated slices and the actual CT slices showed that partial volume effects generally cause the volume to appear larger than the true volume. This is because the nodule boundary is blurred and more pixels that are outside the true nodule boundary become brighter and are considered part of the nodule. There are also additional slices that appear to contain part of the nodule while the true nodule may not have intersected those slices. These effects become stronger when the slice thickness increases.

3.4.4 Manual Segmentation by a Radiologist

Figure 3.7 shows a comparison of a computer generated boundary of a discretized 4.8 mm sphere, the 3DAC result, and the radiologist's hand-drawn boundaries. The average volume errors of the manually segmented volumes for the five 4.8-mm-diameter nodules in each of the 0.625, 1.25, and 2.5 mm slice thicknesses were 239.3 ± 36.5%, 214.8 ± 34.4%, and 275.6 ± 40.3%, respectively. The corresponding 3DAC average volume errors for the same nodules and imaging conditions were 18.6 ± 5.5%, 23.6 ± 7.6%, and 45.2 ± 5.4%, respectively.

3.5 Discussion

In this paper, our focus is to evaluate the effects of CT scanning and reconstruction parameters on the accuracy and reproducibility of automated nodule volume estimation. Although the absolute magnitude of the volume errors estimated using phantom nodules may be different from those of real nodules in patient scans, this study reveals important trends for the dependence of the volume errors on CT imaging conditions and the variability of these measurements. This information may serve as a guide when automated or manual methods are used to assess interval volume change of a nodule from serial CT scans.

In one part of this study, we asked a fellowship-trained thoracic radiologist to segment nodules imaged with varying slice thicknesses. The radiologist's hand-drawn boundaries appear to have included the partial volume voxels. Quantitatively, the volume errors of the manual segmentation ranged from 150 to 350%, compared to the 9% to 52% for the 3DAC. Radiologists are not required to outline lesions in their clinical practice. They are trained to estimate whether a nodule changes size in serial CT scans. The judgment of where the lesion boundaries are on an image is subjective, as indicated by the large inter- and intra-observer variabilities among experienced chest radiologists in the LIDC study [83]. Since the percentage volume change for a given diameter change depends strongly on the nodule size, an overestimation of the nodule diameter can cause large underestimation in nodule growth rate for small nodules. This underscores the difficulty partial volume effects will impose on assessment of nodule growth.

The advantages of automated volume segmentation over manual segmentation include immunity from variability due to changing window and level settings. Radiologists may change the settings to better view the nodule, resulting in differing perceptions of the boundary and thus different volumes [84]. Computerized analysis uses the CT numbers of the voxels themselves in a deterministic algorithm and it will produce exactly the same result for identical input. However, even for CT scans under identical imaging conditions, the acquired images contain statistical variations in the x-ray photons recorded at the detector and other uncertainties of the CT scanner. For example, the starting scan position of the CT scanner is not perfectly reproducible. The slice locations relative to the anatomical structures are therefore not identical in repeated scans even if the phantom is not repositioned.

The reproducibility of volume measurements from scans taken on the same day or five months apart were evaluated in this study. The comparison of segmented volume errors for CT scans with identical imaging conditions between the two experiments shows that, on average, the volume errors agree to within a few percent in the absence of motion or other physiological changes associated with a real patient. This consistency was also observed by Wormanns [95], who reported high precision in lung nodule volume measurement based on an automated method applied to two scans taken 10 minutes apart using the same imaging conditions.

Our study indicated that the average volume error relative to the ground truth were consistently overestimated, except for the 16 mm nodules at thin slice thickness that occasionally showed slight underestimation within experimental uncertainties. The overestimation is mainly caused by partial volume averaging rather than the AC segmentation settings as demonstrated in our analysis of the partial volume effect. Nodule size and slice thickness had the largest effects on the measured volume accuracy. For a given CT condition, the smaller the nodules, the higher the percentage volume errors.

We chose the nodule sizes used in this study because there is general consensus that pulmonary nodules of clinical significance in CT evaluation have longest diameters in the range of 3-30 mm. Nodules less than 10 mm should be followed up with CT [13, 77], and we chose the 4.8 mm and 9.5 mm to represent the low and high end of the range. In addition, 15 to 22 mm nodules are at intermediate risk for cancer, so we chose 16 mm as the representative size. Our results indicate that, for nodules greater than 16 mm, the volume errors due to the factors considered in the current study are negligible so that the size range of the nodules of interest has been covered.

In this study, typical CT scans have 0.703 mm resolution in the axial plane. The voxels are isotropic after interpolation to a volume of $(0.703)^3$ mm^3. At this voxel size, a 4.8 mm diameter sphere contains approximately 168 voxels if the effect of discretization is ignored. Seven slices (4.8 mm / 0.703 mm) is the smallest number of slices that can represent this nodule. A 20% over-estimation in volume is the result of 33 extra voxels, which corresponds to less than five pixels per slice. In other words, if five more voxels in each slice are considered part of the nodule instead of the background for each slice, it will result in a 20% volume error. Thus, the volume error for small nodules is especially sensitive to the uncertainties in the segmented boundary, as a slight deviation due to partial volume effect and reconstruction artefacts such as a blurry or irregular edge would result in substantial percentage error. When the slice thickness is large, the blurred boundary due to partial volume averaging contributes to extra slices for the nodule. The additional slices and additional voxels in each slice together would have caused the high volume errors (>40% for 2.5 mm slice thicknesses) for the 4.8 mm nodules.

This is contrasted to a study by Yankelevitz *et al.* that reported volume errors of ±3% for 3.2 and 3.96 mm diameter phantom nodules [87]. The difference may be attributed to two main factors. First, Yankelevitz *et al.* scanned the phantom at a high resolution of 1 mm beam collimation and 9.6 cm FOV, resulting in pixel size of 0.188 x 0.188 mm in the axial plane, with reconstructed slice interval of 0.5 mm and trilinear interpolation to obtain isotropic voxels. In our study, the smallest slice interval of 0.625 mm would have resulted in interpolated isotropic voxels with volumes of 0.244 mm^3, which is 36.7 times larger than their isotropic voxels with volumes of 0.00665 mm^3 (=0.188^3). Second, our AC parameters were trained with patient nodules [1] rather than the current set of phantom nodules. We can expect that if the AC parameters were trained to fit the edge characteristics of phantom nodules, the AC segmented volumes could be adjusted to be closer to their ground-truth volumes. However, the resulting AC algorithm may not perform well for patient nodules. Because segmentation of patient nodules is the goal for developing the automated AC method, we chose to apply the trained AC to phantom nodules without re-training.

The CT scanning pitch did not significantly affect volume error ($p > 0.05$) for phantom nodules of all sizes when the other parameters were fixed as chosen in this study. However, this experiment was only performed for thin slices (0.625 mm thickness and interval). Future experiments are needed to determine whether this result would still hold with larger slice thicknesses and intervals. There was also no statistically significant effect of FOV on the average volume error, although the average volume error for the 36 cm FOV was consistently slightly lower than that for the 25 cm or 30 cm FOV. One may expect larger errors for larger FOVs because of the increase in the voxel size and thus the partial volume blurring. However, larger voxel size could result in a smaller segmented volume due to the poorer approximation to a sphere by the large discrete voxels. For simplicity, consider an example in 2D. Because the segmented nodule contour was represented by a polygon, smaller voxels would allow more vertices to form the polygon. Consider a polygon inscribed in a circle by placing five vertices on the border of the circle. The area of this polygon would be smaller than that if 10 vertices were placed on the border of the circle. Since the segmented nodule volume was generally over-estimated, the reduced volume due to increased voxel size could reduce the volume error if the reduction was greater than the increase due to the increased partial volume effect.

There was no statistically significant dependence of volume errors on tube current for pitch settings of 0.531:1 and 1.375:1 and for slice thicknesses from 0.625 mm to 2.5 mm. The volume errors therefore were relatively independent of dose for the phantom nodules. For assessment of nodule growth, it may be sufficient to use high resolution but low dose serial scans.

We have demonstrated that there is a large difference in the estimated volume when the CT scan is acquired with different imaging conditions. To minimize the error, thin slice CT scans should be used, especially for small nodules. Furthermore, to estimate the growth of a nodule, it is important to use the same scanning and reconstruction parameters in follow-up scans. Currently, follow-up scanning is performed for nodules after certain time intervals, but there are no specifications on the parameters to be used. A baseline scan and follow-up scans using different slice thicknesses, for example, may result in similar volume measurements when in fact the nodule size has changed. The error in volume change estimation may lead to misdiagnosis and delay in treatment.

The volume errors also raise questions regarding protocols used to screen for lung cancer. Protocols such as the NLST trial that used 2.5 mm slice thickness and 2.0 mm slice intervals may be too thick to accurately assess volume and growth. A thin-slice protocol should be used or a work-up scan at a higher resolution should be performed after a nodule is detected. However, this needs to be balanced with other considerations for screening, such as the larger number of images to be interpreted and archived associated with thin-slice CT screening, patient dose, and the costs of work-up.

In clinical CT scans, variabilities such as patient motion and the change in nodule size and boundary characteristics over time will further degrade the reproducibility between serial exams. Furthermore, clinical nodules in general have less sharp and more irregular boundaries than spherical phantom nodules. These characteristics may cause additional volume errors compared with those estimated in this study, and cause large inter- and intra-observer variabilities even for experienced chest radiologists [83]. These errors will be impossible to estimate because there is no ground truth volume for clinical nodules and the cause and magnitude of the errors may change from case to case. With computer segmentation, although the percentage volume errors for small nodules are large, the reproducibility of volume estimation from repeated scans is within a few percent. It may be expected that if a consistent CT protocol and computerized segmentation algorithm are used for serial CT scans to assess nodule growth, the error in assessing the volume change could be less than the absolute volume error.

There are limitations in this study. First, it will be of strong interest to estimate the smallest possible volume change that can be estimated with confidence using automated segmentation. Second, it is not known whether the same trends observed in this phantom study would be seen for real lung nodules. Third, it is also not known whether the dependence of volume errors on imaging and reconstruction parameters is consistent for CT scans acquired with scanners from different manufacturers. Fourth, for the evaluation of CT parameters on volume errors, we fixed the slice interval at 0.625 mm to reduce the number of variables. The effects of this parameter and its interaction with other parameters on volume error are therefore still unknown. Finally, we did not employ targeted reconstructions with smaller FOVs such as 9.6 cm that would reduce volume averaging in the axial plane. These and other issues will be investigated in future studies.

In summary, we have found that scanning and reconstruction parameters of CT scans affect automatic volume measurement by the 3DAC method. This investigation has important clinical implications because comparing nodule volumes measured from two different scans to determine whether there is growth is a commonly used method in initial diagnosis of lung cancers. In the larger context of a computerized image analysis system, this shows that not only is the segmentation algorithm important, but the method of image acquisition for the CT scans used in the segmentation also affects the outcome. Thus, to accurately follow up on a nodule to detect interval change in volume, the scanning and reconstruction parameters should be properly chosen and kept constant between the initial and follow-up scans to minimize the variability in the volume change evaluation.

3.6 Tables

Table 3.1: Table of scanning and reconstruction parameters for reproducibility study, which included the techniques used in the NLST and LTRC protocols and a high resolution technique. In both protocols, an initial reconstruction and a retrospective reconstruction are performed from the original projection data.

Protocol	Pitch	Slice Thickness (mm)	Slice Interval (mm)	kVp	mA	time (sec)
Hi-resolution	0.531:1	0.625	0.625	120	400	0.8
NLST initial	1.375:1	1.25	1.25	120	80	0.5
NLST retrospective	1.375:1	2.50	2.00	120	80	0.5
LTRC initial	1.375:1	0.625	0.625	140	300	0.5
LTRC retrospective	1.375:1	1.250	0.625	140	300	0.5

Table 3.2: Experiment I: the means and standard deviation of volume errors of 4.8 mm phantom nodules in each of the three repeated scans in the reproducibility study for various scanning and reconstruction parameters. For each set of parameters, there were five spheres with 50 mg/cc $CaCO_3$ and six spheres with 100 mg/cc $CaCO_3$ in each scan.

Pitch	Thickness (mm)	Interval (mm)	kVp	mA	time (sec)	Density	Scan 1		Scan 2		Scan 3	
						mg/cc	Avg.	Std. dev.	Avg.	Std. dev.	Avg.	Std. dev.
0.531:1	0.625	0.625	120	400	0.8	50	24.9%	5.9%	24.9%	3.4%	22.2%	4.0%
						100	19.2%	5.0%	20.4%	6.7%	15.7%	6.0%
1.375:1	1.25	1.25	120	80	0.5	50	24.4%	7.3%	29.0%	7.0%	25.5%	4.5%
						100	20.0%	7.9%	23.7%	6.5%	21.6%	8.9%
1.375:1	2.50	2.00	120	80	0.5	50	35.4%	22.9%	45.4%	13.3%	51.4%	13.5%
						100	43.5%	9.1%	48.1%	4.3%	44.1%	7.8%
1.375:1	0.625	0.625	140	300	0.5	50	21.3%	4.7%	21.8%	4.9%	20.2%	6.7%
						100	19.2%	7.5%	16.4%	6.8%	18.5%	5.8%
1.375:1	1.250	0.625	140	300	0.5	50	18.2%	2.5%	15.7%	3.5%	19.0%	4.2%
						100	17.0%	5.9%	20.0%	5.0%	16.0%	5.7%

Table 3.3: Experiment II (five months after Experiment I): the means and standard deviation of volume errors for 4.8 mm phantom nodules in each of the three repeated scans in the reproducibility study for various scanning and reconstruction parameters. For each set of parameters, there were five spheres with 50 mg/cc $CaCO_3$ and five spheres with 100 mg/cc $CaCO_3$ in each scan.

Pitch	Thickness (mm)	Interval (mm)	kVp	mA	time (sec)	Density	Scan 1		Scan 2		Scan 3	
						mg/cc	Avg.	Std. dev.	Avg.	Std. dev.	Avg.	Std. dev.
0.531:1	0.625	0.625	120	400	0.8	50	21.3%	6.0%	21.2%	4.7%	21.5%	4.6%
						100	21.7%	4.0%	22.2%	3.1%	22.6%	5.7%
1.375:1	1.25	1.25	120	80	0.5	50	25.3%	3.3%	24.9%	7.7%	28.4%	6.8%
						100	22.8%	5.9%	21.7%	5.5%	25.0%	7.5%
1.375:1	2.50	2.00	120	80	0.5	50	50.4%	10.2%	48.7%	11.8%	50.2%	9.8%
						100	48.7%	5.5%	43.5%	9.6%	42.4%	6.7%
1.375:1	0.625	0.625	140	300	0.5	50	23.0%	3.9%	20.0%	2.8%	24.0%	4.2%
						100	21.5%	6.3%	21.3%	2.6%	19.2%	6.7%
1.375:1	1.250	0.625	140	300	0.5	50	21.4%	6.2%	20.8%	4.2%	17.6%	3.4%
						100	14.8%	3.1%	15.5%	4.8%	14.9%	1.9%

Table 3.4: Comparison of the volume errors for 4.8 mm diameter phantom nodules for various parameters taken during Experiments I and II. The volume errors for the nodules with 50 mg/cc $CaCO_3$ and 100 mg/cc $CaCO_3$ are separately analyzed. For each set of parameters, the averages and standard deviations are calculated for volume errors from three identical scans.

Pitch	Slice Thickness (mm)	Slice Interval (mm)	kVp	mA	time (sec)	Density (mg/cc)	Experiment I		Experiment II	
							Avg	Std. dev.	Avg	Std. dev.
0.531:1	0.625	0.625	120	400	0.8	50	18.5%	5.9%	21.3%	4.8%
						100	24.0%	4.4%	22.2%	4.8%
1.375:1	1.25	1.25	120	80	0.5	50	21.8%	7.5%	26.2%	6.0%
						100	26.3%	6.2%	23.2%	6.0%
1.375:1	2.50	2.00	120	80	0.5	50	45.2%	7.2%	49.8%	9.9%
						100	44.1%	17.3%	44.9%	7.5%
1.375:1	0.625	0.625	140	300	0.5	50	18.0%	6.5%	22.3%	3.8%
						100	21.1%	5.2%	20.7%	5.2%
1.375:1	1.25	0.625	140	300	0.5	50	17.7%	5.5%	19.9%	4.7%
						100	17.6%	3.6%	15.1%	3.2%

Table 3.5: Analysis of statistical significance in changes in average volume errors for varying nodule size, tube current, and slice thickness values. Significant (with Bonferroni correction) and insignificant changes are marked with “S” and “NS”, respectively. Other parameters were fixed at 1.375:1 pitch, 0.625 mm slice interval, 120 kVp tube voltage, 0.5 sec rotation time, and 36 cm FOV. The scan of 4.8 mm spheres at 80 mA and 1.25 mm slice thickness was not performed due to an oversight.

Nodule size (mm diameter)	**Tube current (mA)**	**Slice thickness change (mm)**		
		0.625 to 1.25	**1.25 to 2.5**	**0.625 to 2.5**
4.8	**80**	N/A	N/A	S
	160	NS	S	S
	320	NS	S	S
9.5	**80**	NS	S	S
	160	NS	NS	S
	320	NS	S	S
16	**80**	NS	NS	S
	160	NS	S	S
	320	NS	S	S

Table 3.6: Standard deviations of HU values of regions-of-interest (ROI's) on CT slices. The ROI's were placed in the heart muscle, lung parenchyma, and muscle near the vertebra. Since the chest phantom was not moved when the scanning parameters were changed, the slice and position of the ROI's were the same for all entries of this table. The parameters kept constant were 0.531:1 pitch, 0.625 mm slice interval, 120 kVp, 0.8 sec rotation time, and 36 cm FOV.

Slice Thickness (mm)	**Tube Current (mA)**	**Heart**	**Parenchyma**	**Muscle**
0.625	100	11.78	9.29	16.84
	200	8.60	6.66	11.16
	400	6.38	5.07	10.98
1.25	100	10.02	8.11	15.03
	200	7.73	6.04	11.12
	400	5.92	4.86	9.51
2.5	100	7.39	5.78	9.25
	200	5.48	4.61	9.07
	400	4.50	3.58	8.57

3.7 Figures

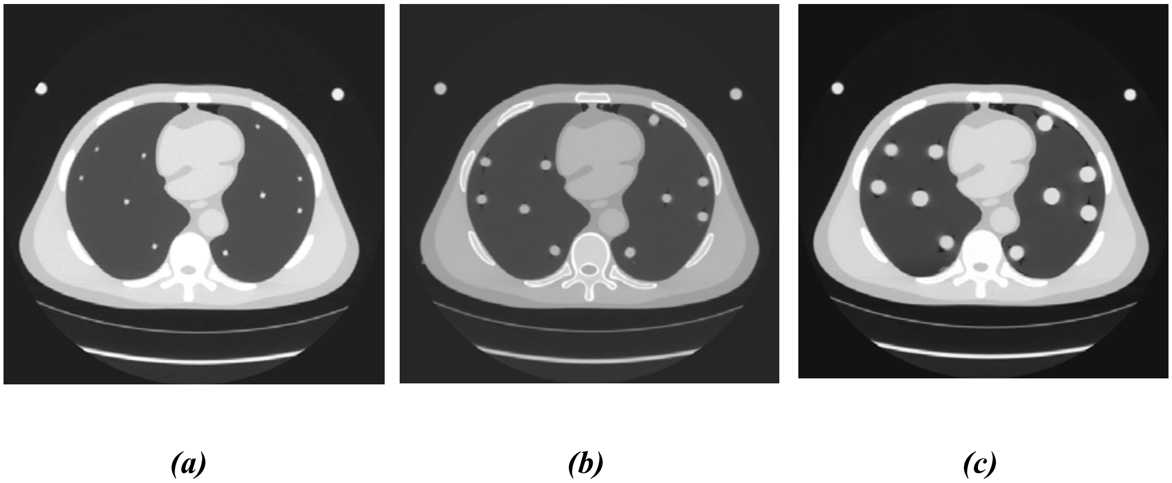


Figure 3.1. Examples of 2-dimensional (2D) axial slices for the three phantom nodule configurations used in Experiment II: The nodules have diameters of (a) 4.8 mm, (b) 9.6 mm, and (c) 16 mm.

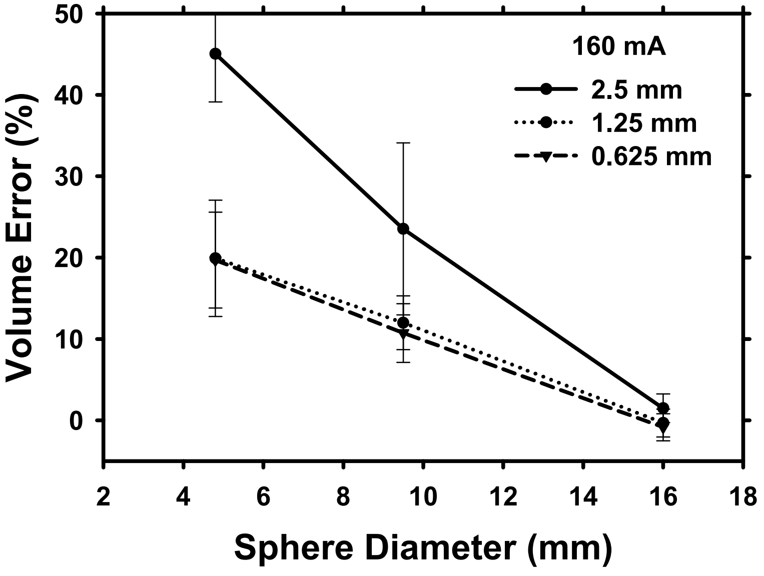


Figure 3.2: The effect of slice thickness on volume error at 120 kVp tube voltage, 160 mA tube current, 0.5 sec rotation time, 0.625 mm slice interval, 1.375:1 pitch, and 36 cm FOV. The error bars indicate ± 1 SD of a measurement.

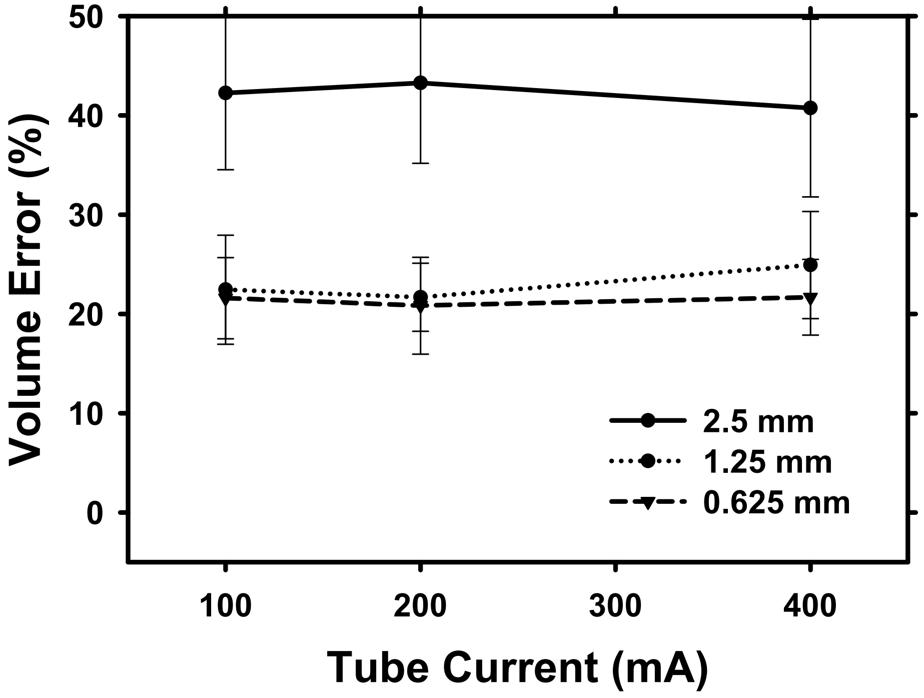


Figure 3.3: Effect of slice thickness and tube current on volume error for 4.8 mm spheres. The parameters fixed were 0.531:1 pitch, 0.625 mm slice interval, 120 kVp, 0.8 sec rotation time, and 36 cm FOV. The error bars indicate ± 1 SD of a measurement.

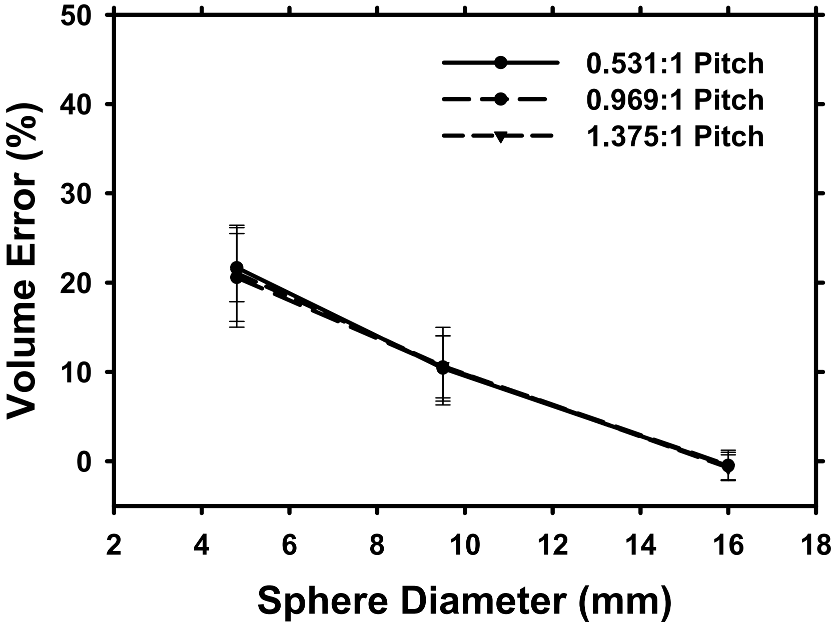


Figure 3.4: Effect of pitch on volume error. The scanning parameters were fixed at 0.625 mm slice thickness and slice interval, 120 kVp tube voltage, 400 mA tube current, 0.8 sec rotation time, and 36 cm FOV. The error bars indicate ± 1 SD of a measurement.

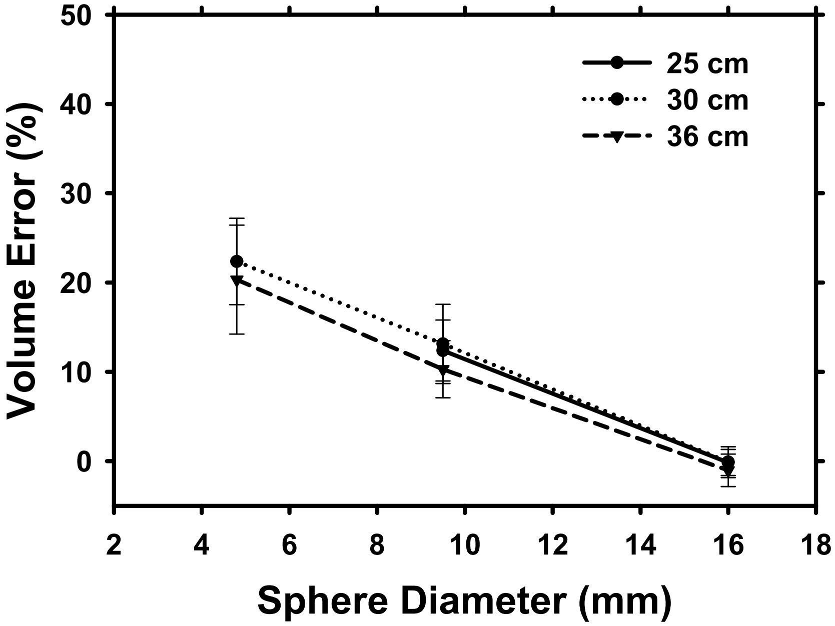


Figure 3.5: Effect of the scanning field of view (FOV) on volume error. The other parameters were fixed at 0.531:1 pitch, 0.625 mm slice thickness and interval, 120 kVp tube voltage, 400 mA tube current, and 0.8 sec rotation time. The error bars indicate ± 1 SD of a measurement.

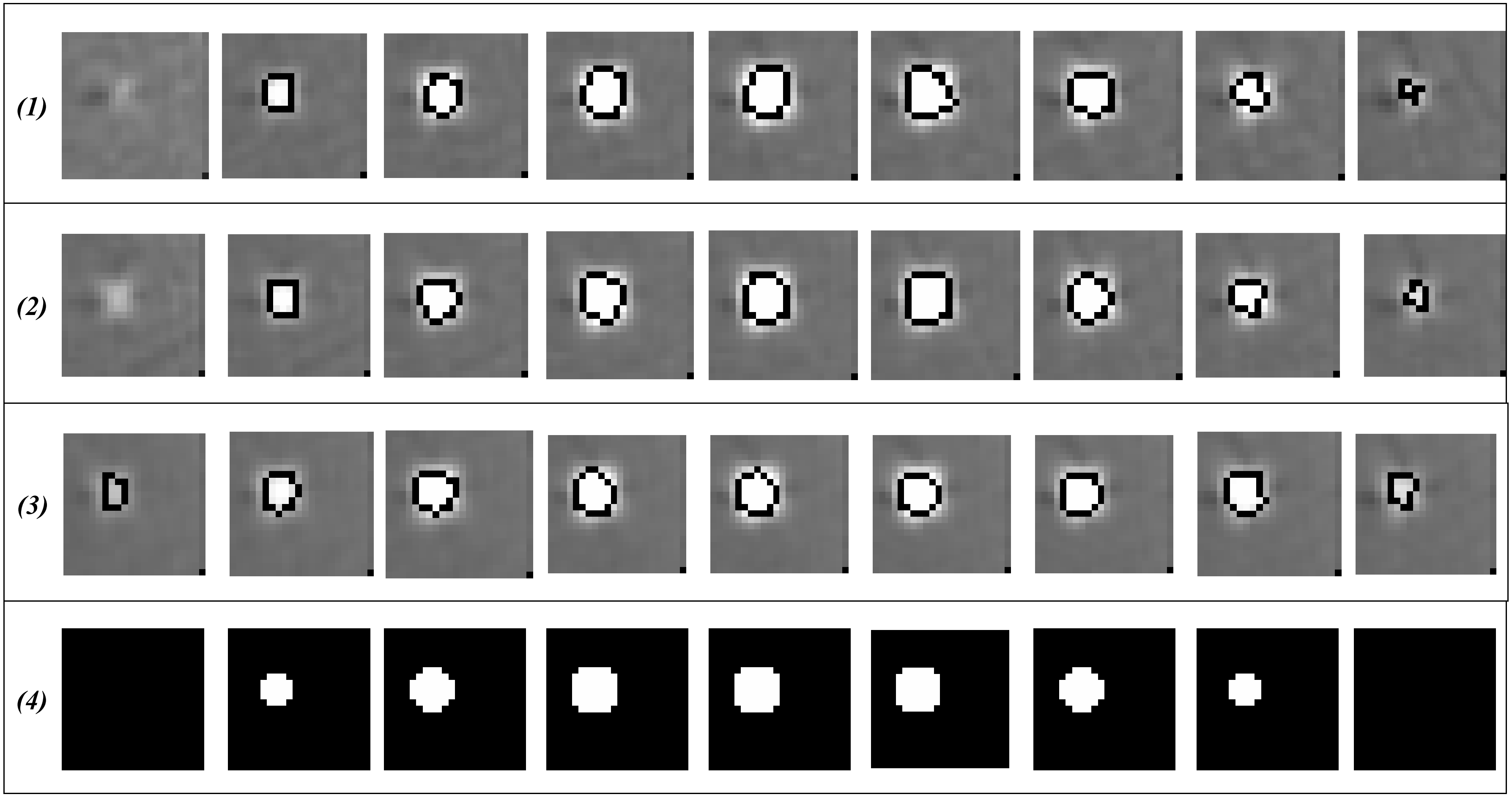


Figure 3.6: Images of all slices with 3DAC segmented contours of a 4.8 mm phantom nodule and various slice thicknesses. Row (1): 0.625 mm, 17% error. Row (2): 1.25 mm, 22% error. Row (3): 2.5 mm, 24% error. Other parameters were kept constant: 0.625 mm slice interval, 120 kVp, 320 mA, 0.5 sec rotation time, 36 cm FOV. Note the additional slice at 2.5 mm slice thickness that was segmented due to the high voxel intensity caused by partial volume averaging. Row (4) contains a computer-generated sphere of 4.8 mm diameter, symmetrically aligned with the pixel array, at 0.625 mm slice interval for comparison.

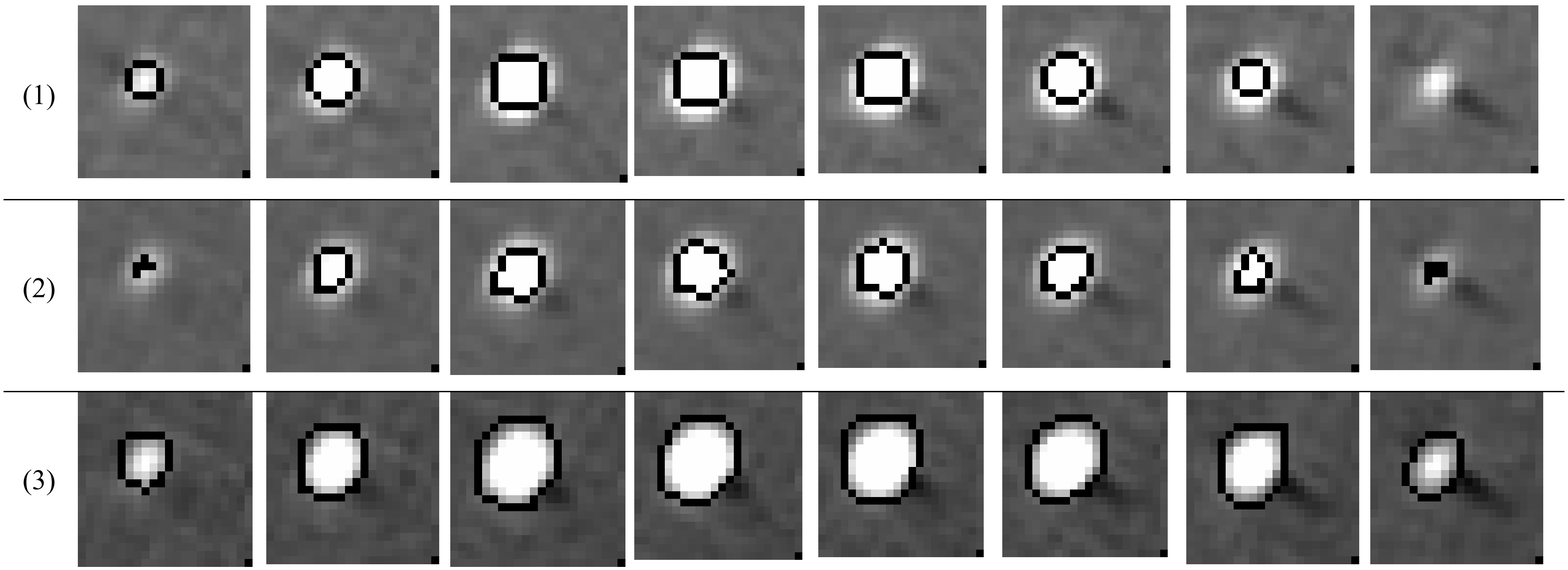


Figure 3.7: Comparisons of contours for a 4.8 mm diameter nodule. Imaging conditions were 1.375:1 pitch, 0.625 mm slice thickness and interval, 120 kVp, 160 mA, 0.5 sec rotation, and 36 cm FOV. Row (1): Contours of a computer-generated 4.8-mm-diameter discretized sphere; row (2): 3DAC output; row (3): radiologist segmentation.

Chapter 4
Effect of Finite Sample Size on Feature Selection and Classification: A Simulation Study

4.1 Abstract

The small number of samples available for training and testing is often the limiting factor in finding the most effective features and designing an optimal computer-aided diagnosis (CAD) system. Training on a limited set of samples introduces bias and variance in the CAD system relative to a CAD system trained with an infinite sample size. We conducted a simulation study to investigate the dependence of the classification performance on design sample size for combinations of feature selection techniques and classifiers. Two feature selection techniques, the stepwise feature selection and sequential floating forward search (SFFS) and two commonly used classifiers, Fisher's linear discriminant analysis (LDA) and support vector machine (SVM), were investigated. Samples were drawn from multi-dimensional feature spaces of multivariate Gaussian distributions with equal or unequal covariance matrices and unequal means. Classifier performance was quantified by the area under the receiver operating characteristic (ROC) curve, A_z. The mean A_z values obtained by resubstitution and hold-out methods were evaluated for training sample sizes ranging from 15 to 100 per class. The number of simulated features available for selection was chosen to be 50, 100, and 200. It was found that the LDA and SVM with radial kernel performed similarly for most of the conditions tested in this study, although the SVM classifier showed a slightly higher hold-out performance for some conditions. The SFFS method was comparable to the SFS method. The understanding of the performance of various feature selection technique and classifier combinations when limited samples are available is expected to facilitate development of effective CAD systems.

4.2 Introduction

Advances in computer processing power, memory capacity, imaging technologies, and image processing algorithms have greatly improved the diagnostic information available to radiologists. Image processing and analysis tools allow computers to aid radiologists in previously time consuming tasks. Computer-aided detection and diagnosis (CAD) software further facilitates image interpretation. Computer-aided detection systems mark suspicious areas on images that radiologists may have overlooked to prompt them to examine that area more carefully. Computer-aided diagnosis software for cancer provides a malignancy estimate of suspicious tissue.

For these computer aids to be effective, however, a CAD system would need to extract salient features from the images, choose only the features that can discriminate between classes, and accurately classify previously unseen samples. Ideally, there would be a large number of training samples to design the CAD system, but it is expensive and time-consuming to collect case samples with ground truth. The development of CAD systems for automatic detection and diagnosis of lung nodules on CT can serve as an example. For detection studies, the gold standard for determining whether abnormal-appearing tissue is considered a nodule is often determined by consensus among radiologists, yet considerable inter-observer variability makes the truth uncertain. For diagnosis studies, a nodule is considered benign when it shows no change for at least two years. Determination of malignancy often requires biopsy. If multiple nodules are present in the lungs, biopsy may not be performed for every nodule because of the risks and expenses. These factors limit the number of samples available to train, validate, and test CAD systems.

An important issue in CAD system development is whether the performance on training data is generalizable to the population at large. It is therefore useful to estimate the bias and variance of the classifier performance on previously unseen samples. This would allow users to predict the performance when the CAD system is applied to unknown cases in clinical practice. Classifiers for the differentiation of true and false lesions or for the differentiation of malignant and benign lesions are some of the main components in a CAD system. Studies of sample size effects on classifier design exemplify similar problems in development of CAD systems. Previous simulation studies have focused on the effect of finite sample size on classifier performance when the samples were drawn from multivariate Gaussian distributions of various dimensionality. Resubstitution and hold-out methods were used for estimating classifier performance. In the resubstitution method, the classifier performance is measured by applying the classifier to the training samples that have been used to design it. In the hold-out method, the samples are partitioned into training and test samples. The classifier is designed with only the training samples and then evaluated on the independent test samples. Chan *et al.* [96] compared the sample size effects on the design of the linear discriminant, the quadratic discriminant, and the backpropagation artificial neural networks (ANN). For feature spaces of Gaussian distributions with unequal covariance matrices and 3 to 15 dimensions, the linear discriminant analysis (LDA) classifier was inferior to the quadratic discriminant or the ANN when there were a large number of training samples. However, with a small number of training samples available, a simpler classifier such as the LDA or ANN with few nodes may be preferred. A small sample size becomes even more limiting when one has to select the most effective features from a large pool of available features using the same small sample set. Sahiner *et al.* [97] investigated the effect of sample size, number of available features, and the parameters for stepwise feature selection (SFS) on LDA performance. They found that the resubstitution estimate was always optimistically biased, except when there were too few features. The hold-out estimate was always pessimistically biased when the classifier was trained on only the training samples.

In recent studies, Sahiner *et al.* investigated the bias and variance of various resampling methods in predicting the performance of a classifier for unknown samples when the classifier is trained with a finite sample size. Two classifiers, Fisher's LDA [98] and backpropagation ANN [99] were evaluated. Under their study conditions, they found that the prediction accuracy depends strongly on the resampling method, especially for large feature dimensionality and small sample sizes. Li and Doi [100] performed a simulation study and proposed an automated threshold selection method to minimize the overtraining effect in rule-based classifier design. Li and Doi also performed another study [101] to compare evaluation methods for CAD systems such as the bias of the estimated performance, the generalization performance, and the uniqueness of the CAD scheme. Beiden *et al.* [102] focused on the variance of competing classifiers. They concluded that, in comparing various classifiers, the variance contributed by the finite training sample is the dominant component. This is opposed to the conventional wisdom that the finite training sample size affects the bias of measures of performance, while the variance can be attributed mainly to the finite number of test samples.

The interaction between the feature selection method and the classifier used will also influence the training of a classifier. Some combinations of methods may perform better than others given a small sample size, some may generalize better to unknown samples, and others may result in lower variance. Jain and Zongker [103] compared various feature selection algorithms and concluded that the sequential forward floating search (SFFS) [104] method performed better than other methods. Kudo and Sklansky [105] also concluded that SFFS was effective for small- and medium-scale problems while genetic algorithms would be better suited for large-scale problems.

The goal of this study is to investigate combinations of feature selection techniques and classifiers and to compare their performance on two classes of data drawn from multivariate Gaussian distributions with unequal means and either equal or unequal covariance matrices. The effects of the covariance matrices, finite sample size, and the dimensionality of the feature spaces on the bias of the classifier performance relative to that of infinite training sample were studied. Although both feature selection methods and classifiers have been investigated extensively in the literature, there are only limited studies on combinations of these two important processes for various feature selection techniques and classifiers. This study is expected to provide CAD system designers further understanding on the sample size effects and their interaction with feature selection and classification methods. This information may serve as a guide in future CAD system development and prompt further investigations on these important issues for classifier design.

4.3 Methods and Materials

A typical CAD system is shown in Fig. 4.1. In the training phase, a large number of features can be extracted from the image samples. Many of these features may not be useful for the classification task at hand and a feature selection method is used to choose the most effective features. A classifier is then built using the selected features as input predictor variables. Both the feature selection and the classifier parameters should be trained on the training samples only. The performance of the trained classifier on unknown cases is then estimated on the independent samples that have been held out for testing.

4.3.1 Random Sample Generation

In this simulation study, the training and test samples for the two classes were drawn randomly from two multivariate Gaussian distributions of three different types: (1) equal covariance matrices with unequal means, (2) unequal covariance matrices with unequal means, and (3) equal covariance matrices estimated from clinical data with unequal means. Although clinical data may not follow a Gaussian distribution, this idealized distribution keeps the number of parameters to investigate manageable, and allows us to gain insight into the effect of finite sample size and relative performance among the various feature selection and classifier methods.

A set of N_s samples was generated from each class distribution using a random number generator. The detail of the two classes is described below. This set was then randomly partitioned into N_{train} training and N_{test} test samples per class. We varied N_{train} and fixed N_{test} to be 100 for a given feature space to study the effect of training sample size on classifier performance. For a given number of training and testing samples, 1000 experiments were performed with a new set of samples generated for each experiment. Keeping N_{test} fixed for different experiments allowed us to directly investigate the dependence of the variance of the performance measure on the number of training samples, without the confounding effects of the variation of the number of test samples. The resubstitution and hold-out test performances of the classifier were quantified by the area under the receiver operating characteristic curve, A_z. The mean and the variance of the resubstitution and hold-out test A_z for the given sample size were estimated from the 1000 experiments.

4.3.1.1 Equal Covariance Matrices and Unequal Means

The first condition simulated two classes with multivariate Gaussian distributions and equal covariance matrices. Without loss of generality, we used two identity matrices because a common arbitrary covariance matrix for both classes can be simultaneously diagonalized and the variances of the individual feature components normalized to unity [96]. The mean feature vector of the first class was zero, $\boldsymbol{\mu}_1 = \mathbf{0}$, and the difference in the class means, $\Delta\mu(i)$, between the two classes for feature i was given by [97]:

$$\Delta\mu(i) = \mu_2(i) - \mu_1(i) = \alpha\beta^i, i = 1,...,M \; and \; \beta < 1 \tag{4.1}$$

where M is the dimensionality of the available feature space from which a number of features may be selected. The squared Mahalanobis distance between the two classes Δ [97] was computed as:

$$\Delta = \frac{\alpha^2\beta^2}{1-\beta^2}\left(1-\beta^{2\mathbf{M}}\right) \tag{4.2}$$

since all the diagonal values of the covariance matrix were 1. The parameter β was set to be 0.9 and α was chosen such that Δ = 3.0. Feature i therefore has decreasing ability to separate the two classes as i increases. The specific form of the features and the values of these parameters were not critical for the purpose of this simulation study; they were designed to generate a set of features that have varying discriminatory power to distinguish the two classes. For the equal covariance matrix condition, the Mahalanobis distance can be used to determine the ideal A_z value of the optimal classifier trained and tested with the true (infinite-sized) population, denoted as $A_z(\infty)$ [96]. In this study, the Mahalanobis distance was selected such that $A_z(\infty)$=0.89, which is representative of the range of A_z values achieved in CAD literature. The classification accuracy for M = 50, 100, and 200 was investigated.

4.3.1.2 Unequal Covariance Matrices and Unequal Means

This condition simulates two classes that are distinctly different from each other. The covariance matrix of the first class was diagonalized and scaled as the identity matrix, $\Sigma_1 = I$, with $\boldsymbol{\mu}_1 = \mathbf{0}$. The covariance matrix of the second class, Σ_2, was simultaneously diagonalized such that it had eigenvalues $v_i, i = 1,\ldots,M$, where M is the feature space dimensionality. The values of v_i were generated by

$$v_i = 1 + \varepsilon(\gamma^{M-i} - 1),\; i = 1,\ldots,M\;,\; \varepsilon = \frac{v_{\max} - 1}{-1 + \gamma^{M-1}} \tag{4.3}$$

where γ = 1.5, $v_{\max} = v_1$ = 3, and the smallest eigenvalue $v_{\min} = v_M$ was set to 1. The eigenvalues of the covariance matrix for the second class therefore decreased exponentially from $v_{\max}$ to 1 as the feature number changed from 1 to M. The values of the mean vector of the second class, $\boldsymbol{\mu}_2$, were calculated according to Eqn. (4.1), where β = 0.9. For the unequal covariance matrix condition, there is no closed-form solution that relates the mean and covariance matrices of the class distributions to $A_z(\infty)$. However, a close approximation for $A_z(\infty)$ in terms of the Bhattacharyya distance [106, 107] has been derived [Barrett, Abbey and Clarkson, 1999]. In this study, the value of α in Eq. (4.1) was chosen such that the Bhattacharyya distance between the two classes was 3/8, which corresponded to $A_z(\infty) \approx 0.89$. With the selected values of α and β, the squared Mahalanobis distance was 1.66, which was lower than that in the equal covariance matrix condition. The non-identity covariance matrix was designed such that the greatest separation in the mean value corresponded with the greatest eigenvalue in the covariance matrix. Since our goal was to compare the performance of various features selection methods and classifiers, the specific values of $v_{\max}$ and $v_{\min}$ were not critical.

4.1.3.3 Equal Covariance Matrices Based on Clinical Data and Unequal Means

To simulate features from clinical data that may be encountered by a CAD system, we first extracted features from volumes-of-interest containing lung nodules from computed tomography (CT) scans. These features were extracted with the goal of classifying the lung nodules as malignant or benign [1]. They included morphological features such as volume and perimeter, in addition to gray-level statistics, texture features from run-length statistics [65, 66], gradient field, and radii features. The means and covariance matrices of each class were estimated from a database of 124 malignant and 132 benign nodules. These estimated means and covariance matrices were assumed to be the true underlying multivariate Gaussian distributions of the population for this study. We assumed that the two classes had the same multivariate Gaussian distribution with covariance matrix $\Sigma = (\Sigma_1 + \Sigma_2)/2$, where Σ_1 and Σ_2 were estimated from the malignant and benign classes, respectively, of the clinical data.

4.3.2 Feature Selection Methods

Typical feature selection strategies include the “top-down” and “bottom-up” methods. Marill and Green introduced the “top-down” method [108], which is initialized with the entire feature space. Features are removed after certain criteria have been met to obtain the set of remaining features to be used. Its counterpart is the “bottom-up” method, which is initialized with the empty set, and features are added until certain criteria have been met [109]. The disadvantage of these methods is the “nesting effect,” in which features removed are no longer considered or features added cannot be removed. The stepwise feature selection (SFS) and sequential floating forward selection (SFFS) methods were designed to overcome the nesting effect.

4.3.2.1 Stepwise Feature Selection (SFS)

SFS uses a selection criterion based on F-statistics [110, 111] and addresses the nesting issue by allowing features to be added to and removed from the set of selected features. Initially, all features are tested to find the one that provides the smallest value of a selection criterion, which was Wilks’ lambda in this study. Wilks’ lambda is defined as the ratio of the within-group sum-of-squares to the total sum of squares:

$$\lambda_k = \frac{\sum_{i \in class1}\left(h^{(k)}(X_i) - m_1^{(k)}\right)^2 + \sum_{i \in class2}\left(h^{(k)}(X_i) - m_2^{(k)}\right)^2}{\sum_{i=1}^{N}\left(h^{(k)}(X_i) - m^{(k)}\right)^2} \tag{4.4}$$

where *d* is the dimensionality of the selected feature subspace, *h(X)* is the discriminant score for the input vector X_i consisting of the selected features for case *i*, $h^{(d)}(X_i) = b^T X_i + b_0$, with $b^T = [b_1, b_2, \ldots, b_d]$ and b_0 being the LDA coefficients, $m_1^{(d)}$ and $m_2^{(d)}$ are the means of the discriminant scores for classes 1 and 2 respectively, $m^{(d)}$ is the mean of the discriminant scores for both classes, and *N* is the number of available training samples [97]. The smaller the value of Wilks’ lambda, the smaller the spread within each class relative to the spread of the entire sample; indicating that the separation of the two classes is larger and better classification can be achieved.

To determine whether to include a feature when *d* features have already been selected, the *F-to-enter* value is calculated [112] for each feature that has not been selected:

$$F = (N - d - 2)\left(\frac{\lambda_d}{\lambda_{d+1}} - 1 \right), \tag{4.5}$$

where λ_d and λ_{d+1} are the Wilks' lambda values before and after entering the feature to the pool of selected features. The feature with the largest *F-to-enter* value is added to the selected features if its value is higher than a threshold F_{in}. A lower F_{in} threshold means that it is easier to add more features, resulting in a larger set of selected features. After a feature is entered, each feature in the selected pool is tested for removal by calculating the *F-to-remove* value, which is defined similarly to *F-to-enter*. The feature with the smallest *F-to-remove* value that is also lower than a threshold F_{out} is removed. A lower F_{out} makes it more difficult to remove features, which will lead to a larger set of selected features. This process of entering and removing a feature is repeated until no more features satisfy the criteria for entry or removal. Another threshold is the tolerance term, which prevents a feature from being entered when it is highly correlated with the already selected features, even if the feature satisfies the F_{in} threshold. Because the thresholds are not known *a priori*, and it is not practical to search through all combinations, we set $F_{out} = F_{in} - 1$, where F_{in} was varied from 2 to 7 to cover a reasonable range of values, and the tolerance threshold was fixed at 0.001. These thresholds result in a wide range of the number of features selected, allowing us to demonstrate the effect of finite sample size on feature selection and classifier performance.

4.3.2.2 Sequential Floating Forward Search (SFFS)

A disadvantage of SFS is that it only allows one feature to be added or discarded at a time. The Plus-*l*-minus-*r* method [113] allows the addition of *l* or removal of *r* features at a time, but there is no theoretical way to predict the best *l* and *r* values. Pudil *et al.* [104] introduced the floating search method, where the number of features added or removed at each step changes dynamically, and a pre-defined number of desired features controls the stopping criterion. The SFFS method is initialized with the best performing combination of two features. The procedure terminates when the number of selected features reaches the pre-determined number of desired features plus delta, where delta was set to 5 in this study. This allows the SFFS algorithm to search for combinations of features of cardinality beyond the desired number of features. The best feature combination corresponding to the desired cardinality can then be chosen.

SFFS is a suboptimal search method that assesses the performance of combinations of features. A table stores the best performing feature combinations of cardinalities of 1 through a number beyond the total number of desired features plus delta. As features are added and removed, the performance of features is assessed. If a better performing combination of the same cardinality is found, then that combination is updated in the table. The procedure terminates when the best feature set of the desired cardinality is found. The SFFS process is illustrated in an example below. We chose to examine the performance of 5, 8, 11, 14, 17, and 20 desired features because that encompassed the range of features selected by the SFS method for all but a few extreme cases under our simulation conditions.

Suppose we have a set of features, F_j, where j = 1 to M. Assume that F_3 and F_9 are selected to initialize the SFFS method. The method starts with the feature addition stage. All unselected features are each tested for their discriminatory ability in combination with the already-selected features. We used the Mahalanobis distance between the two classes as the figure-of-merit when the feature is added to the already selected features. Assume that among all the unselected features, feature F_{12} results in the greatest Mahalanobis distance between the two classes when combined with (F_3, F_9). To determine whether to add F_{12}, each feature is removed in turn, with the discriminatory performance of the remaining features calculated. If (F_3, F_9) has the best performance compared to (F_3, F_{12}) and (F_9, F_{12}), then F_{12} is kept. After deciding that F_{12} will be kept, the next feature among the unselected features that provides the greatest Mahalanobis distance with the already selected features, F_5 for example, is evaluated. Each feature is removed in turn and the remaining features are tested, e.g., all three-feature combinations from among (F_3, F_9, F_{12}, F_5) are tested. If (F_3, F_9, F_{12}) results in the greatest Mahalanobis distance between the two classes, F_5 is added to the selected feature set, and the procedure of adding features is repeated until the testing by removal of the most recently added feature does not result in the highest performance combination. If (F_3, F_9, F_{12}) does not result in the greatest Mahalanobis distance between the two classes, F_5 is still added to the selected feature set, but the feature addition stage is stopped and the algorithm proceeds to the feature removal stage. From the selected feature set, (F_3, F_9, F_{12}, F_5) in our example, each feature is removed and all the other three-feature combinations are evaluated. For example, if (F_3, F_{12}, F_5) results in the highest Mahalanobis distance, then this combination is compared to the best three-feature combination stored in the table. If (F_3, F_{12}, F_5) has better performance, then this is updated in the table as the best three-feature combination so far. The feature removal stage continues with the next smallest cardinality (though in this example, the next smallest cardinality is two, and the best combination of two features has already been found If this combination performs better than what is stored in the table another feature is removed. If no combination of the remaining features is better than what has already been found, then one returns to the feature addition stage. When the number of selected features reaches the number of desired features plus delta, the algorithm terminates. It is then possible to look up the desired cardinality in the table and find the corresponding best performing combination of features for that cardinality that has been searched.

4.3.3 Classification Methods

A large number of linear and non-linear classifiers have been developed in the literature for various pattern recognition and machine learning problems. We selected two commonly used classifiers, Fisher's LDA and the support vector machine (SVM) with two different kernels, as examples of linear and non-linear classifiers to compare their performance in combination with the SFS and SFFS methods.

4.3.3.1 Linear Discriminant Classifier

The linear discriminant classifier uses the means and covariance matrices of the two class distributions to calculate a linear decision boundary separating the two classes. The classifier is described as [106, 114]:

$$h_l(X) = (\mu_2 - \mu_1)^T \overline{\Sigma}^{-1} X + \frac{1}{2}\left(\mu_1^T \overline{\Sigma}^{-1} \mu_1 - \mu_2^T \overline{\Sigma}^{-1} \mu_2\right) \tag{4.6}$$

where $\overline{\Sigma} = (\Sigma_1 + \Sigma_2)/2$ and X is the feature vector. The means and covariance matrices have to be estimated from the available training samples. A nonlinear transformation of the sample means and covariance matrices results in the LDA coefficients. The LDA coefficients are then linearly combined with the test data to obtain the discriminant scores, which are transformed nonlinearly into a performance measure. The variances due to the estimated parameters propagate to the mean classifier performance, resulting in a bias through the second derivative of the transformation function [96].

It is known that the LDA classifier is optimal for multivariate normal distributions with equal covariance matrices. The classifier performance in the limit of large training samples can be calculated by the Mahalanobis distance:

$$A_z = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\sqrt{\Delta/2}} e^{-u^2/2} du \tag{4.7}$$

In this study, we set Δ=3 for the equal covariance matrix condition, and thus the maximum achievable $A_z(\infty)$ by the optimal linear discriminant is 0.89 in the limit of a large number of training samples. For the unequal covariance matrix condition, Δ=1.66 for the chosen feature space, which corresponds to A_z= 0.82 using Eqn. (4.7) for this second simulation condition. Based on the estimated means and covariance matrices of the clinical data, we had Δ=4.91 for the third simulation condition, and the corresponding A_z value from Eqn. (4.7) was 0.941. Since the means and covariance matrices for this third condition were estimated from the small available clinical sample, the estimated Mahalanobis distance between the two classes may be optimistically biased compared to the true Mahalanobis distance between the malignant and benign populations.

4.3.3.2 Support Vector Machine (SVM)

The SVM works similarly to the LDA by constructing a decision hyperplane to separate classes using training data. A brief overview of the SVM is given here, with more details in the literature [115]. Geometrically, the SVM maps the original data to a higher dimension space via a kernel *K*. A decision hyperplane is constructed in this higher dimension such that the distance between the training samples of both classes and the hyperplane is maximized. This distance between a training sample and the hyperplane is called the margin, and the SVM calculates the hyperplane with the largest margin.

Suppose we have labeled training samples $\{\mathbf{x}_i, y_i\}$, $i = 1 \ldots N$, $y_i \in \{-1,1\}$, $\mathbf{x}_i \in \mathbf{R}^d$, where *N* is the number of samples and *d* is the dimensionality of the selected feature space (number of selected features). In the SVM formulation, the data appears in the form of dot products, $\mathbf{x}_i \cdot \mathbf{x}_j$. First, the SVM algorithm uses a mapping, Φ, to transform the data to some other Euclidean space *H*, $\Phi : R^d \mapsto H$. The transformation depends only on the dot products in *H* of the form $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. There exist kernel functions *K* such that $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$, and only *K* is needed in the training algorithm. No explicit knowledge of Φ is necessary. Various kernels have been investigated in the literature, and we chose two commonly used ones, the radial and polynomial kernels [116-118]. In the following, the SVM with the radial and polynomial kernels are referred to as SVM(rad) and SVM(poly), respectively. We implemented the SVM with the freely available *mySVM* [119] software.

4.3.4 Simulation Study

The number of training samples per class randomly drawn from the class distributions was 15, 20, 30, 40, 50, 60, 80, 90, and 95. The number of test samples per class was fixed at 100 so that the variances in the hold-out classification performance due to the test set size are kept relatively constant. The dimensionality of the input feature spaces, M, were chosen to be 50, 100, and 200.

Combinations of the two feature selection methods (SFS and SFFS) and three classifiers (LDA, SVM(rad), and SVM(poly)) were trained and tested on the available samples. For each combination, there were three different types of feature space distributions, as discussed above. The resulting resubstitution and hold-out A_z values, in addition to the variances were compared.

4.4 Results

The results of the simulation study for various combinations of feature selection and classification methods are described below. For a given number of training samples, the mean A_z obtained by resubstitution or the hold-out performance is estimated by averaging the results of 1000 experiments. For simplicity, mean A_z will be referred to as A_z in the following discussion.

4.4.1 Equal Covariance Matrices with Unequal Means

In Figs. 4.2, 4.3, and 4.4, the two feature selection methods are compared for the LDA, SVM(rad), and SVM(poly) classifiers, respectively. The A_z values for the resubstitution and the hold-out methods are plotted as a function of $1/N_{train}$ for $M = 50$, 100, and 200. Figs. 4.5 and 4.6 compare the performance of the classifiers for the SFS and SFFS methods, respectively. These two last figures contain the same data as Figs. 4.2 - 4.4, but are organized in a different way to facilitate the comparison between different classifiers. Examples of the standard deviation values when the two feature selection methods are used with the LDA classifier for M = 100 are shown in the first row of Fig. 4.13. The standard deviations of the SVM classifiers (not shown) had similar magnitudes.

4.4.2 Unequal Covariance Matrices

Comparisons of the feature selection techniques for the LDA, SVM(rad), and SVM(poly) classifiers are shown in Figs. 4.7, 4.8, and 4.9, respectively. The results of comparing the classifier performances with input features selected by SFS and SFFS are shown in Figs. 4.10 and 4.11, respectively. Similar to the equal covariance case, Figs. 4.10 and 4.11, contain the same data as Figs. 4.7 - 4.9. The A_z values of the resubstitution and hold-out estimates of different methods are compared and plotted as a function $1/N_{train}$. Examples of the standard deviation values when the two feature selection methods are used with the LDA classifier for M = 100 are shown in the second row of Fig. 4.13. The standard deviations of the SVM classifiers had similar magnitudes.

4.4.3 Equal Covariance Matrices (Clinical)

In these experiments, the two classes had unequal means but the same covariance matrix derived from the features extracted from lung nodules on CT scans. There were $M = 61$ features available for selection. The classifier A_z values from the LDA, SVM(rad), and SVM(poly) classifiers with the SFS and SFFS feature selection techniques are compared in Fig. 4.12. Examples of the standard deviation values when the two feature selection methods are used with the LDA classifier are shown in the third row of Fig. 4.13.

4.5 Discussion

Numerous feature selection and classifier methods have been investigated in the literature. It is difficult to determine which combination of feature selection and classifier methods would be the most effective for a given classification task. When a specific combination is selected based on a limited number of samples available, as often in the case of CAD system development, the potential for overtraining is high, and the performance of the resulting CAD system as predicted by the training data may not be generalizable to the population at large. To investigate the performance of various combinations, we conducted a simulation study with sample data randomly drawn from multivariate Gaussian distributions, which allowed us to generate arbitrarily large number of samples. Although the results may not be directly applicable to features extracted from clinical data since the class distributions may not be Gaussian, the relative performance of various combinations may serve as a guide to understanding the characteristics of the combinations.

In Fig. 4.2, the effect of increasing the feature space dimensionality is demonstrated by the graphs in each column. Since the ability of feature i to separate the two classes decreased with increasing i, the feature index, in Eqn. (4.2), the contribution of additional features beyond $i = 25$ was close to zero. An effective feature selection algorithm should be impervious to these additional features, which were essentially noise since the difference in the means of a given feature between the two classes from which samples were drawn was close to zero. However, the hold-out A_z for any number of training samples decreased as the number of features available, M, increased. This trend could also be seen in Figs. 4.3 and 4.4 for the SVM(rad) and SVM(poly) classifiers, respectively. Nevertheless, the hold-out performance for LDA was similar whether features were selected by SFS or SFFS.

Fig. 4.3 compares the SVM(rad) classifier performance with the SFS and SFFS methods. The classifier hold-out performance for the two methods was similar. Although the resubstitution bias for both SFS and SFFS increased with an increased number of features selected, the SFS results showed a slightly higher resubstitution bias because more features were selected based on the F_{in} and F_{out} thresholds. It is interesting to note that at $M = 200$, the number of features selected had minimal effect on hold-out performance.

For the comparison of feature selection methods with the SVM(poly) classifier shown in Fig. 4.4, SFFS resulted in a higher resubstitution bias than SFS when few features were selected. When the number of training samples was large, the classifier hold-out A_z with SFFS still achieved better performance than with SFS, since the resubstitution bias with SFFS was slightly lower. The hold-out A_z of the classifier with SFS was low at $M = 200$ when the number of training samples was large, but that was not seen with SFFS. This can be attributed to the fact that the number of features selected by SFS could be larger than 60 for small F_{in}, which was greater than the range of the number of features set to be chosen for the SFFS. The excessively large number of selected features may have resulted in the greater bias of SVM(poly) with SFS. Fig. 4.5 shows a comparison of the LDA, SVM(rad), and SVM(poly) classifiers with SFS. The resubstitution A_z, especially when there were few training samples, increased with increasing M whereas the hold-out A_z decreased, indicating overtraining. The number of selected features also increased as M increased (not shown in the graphs). The same set of selected features from SFS was used for each classifier. The resubstitution performance of the LDA and SVM(rad) classifiers were essentially the same, while the SVM(poly) showed more optimistic bias, especially when the number of selected features was large (low F_{in} and F_{out} thresholds) or when the number of training samples was large. The hold-out performance of the SVM(poly was in general worse than the performance of the LDA and SVM(rad) classifiers.

It is interesting to note that given the same selected features from the SFS method, the SVM(rad) had slightly better hold-out performance than the LDA, especially when the number of available features M and the number of selected features were large. Since the data were drawn from multivariate normal distributions with identical covariance matrices, it is expected that LDA would theoretically provide the optimal performance. However, LDA estimated the means and covariance matrices from the available training samples of both classes, and the limited sample size could result in poor estimates. For small training sample sizes, the hold-out performance of SVM(rad) was less pessimistically biased compared to LDA. The difference in the performance between LDA and SVM(rad) decreased as the training sample size increased. When the number of training samples was the highest for the experimental conditions studied, the LDA hold-out performance was similar to the SVM(rad) performance when the number of selected features was small. For the SFS method, for a given training sample size, the hold-out A_z values for SVM(rad) was less dependent on the F_{in} and F_{out} thresholds than that of LDA when the original feature space dimensionality was high (M=100 or 200) and vice versa when the dimensionality was low (M=50). Given conditions similar to this experiment, SVM(rad) would have a slight advantage over LDA when the available training sample size is small.

It was also observed that SVM(poly) had the highest resubstitution and hold-out biases, especially for small training sample sizes and with the SFFS method as shown in Fig. 4.6. The resubstitution bias of SVM(rad) was slightly less than that of LDA, which may have resulted in better hold-out performance of SVM(rad). The resubstitution A_z of the classifiers with SFFS remained fairly constant regardless of M.

For the equal covariance matrix case, for both SFS and SFFS methods, SVM(rad) obtained a slightly higher hold-out performance, and it may hold a slight advantage when there are few training samples available. Although LDA is theoretically the optimal classifier, the lack of training samples to accurately estimate the covariance matrices and means may have contributed to its poorer performance. However, when a large training sample is available, LDA performed similarly to SVM(rad). The two feature selection methods are comparable when they are combined with the three classifiers studied.

Fig. 4.7 shows the effect of the feature selection methods on LDA when the two classes had unequal covariance matrices and unequal means. Under this condition, LDA would not be the optimal classifier, but its performance in combination with various feature selection methods would be of interest because LDA is a commonly used classifier. The hold-out performance of LDA was similar whether features were selected by SFS or SFFS. The resubstitution A_z of LDA with SFFS had much larger optimistic bias than the LDA with SFS for small training sample sizes. The trend of the resubstitution curve depends on feature selection method. With SFFS, the resubstitution A_z decreased monotonically as N_{train} increased, whereas with SFS, the resubstitution A_z decreased and then increased as N_{train} increased when the number of selected features and M were small. Although the number of features selected for both methods were similar at M = 50, the number of features from SFS had greater influence on hold-out performance.

A similar comparison was made for the SVM(rad) in Fig. 4.8. There was a greater change in resubstitution bias for SVM(rad) with SFFS compared to with SFS as the training sample size varied. Whether features were selected by SFS or SFFS, the SVM(rad) hold-out performance was virtually the same. For M = 200, the number of features selected had little influence on classifier performance. Given the experimental design, the additional features may not provide much discriminatory power, and the SVM(rad) classifier may have effectively disregarded them.

A trend that is most evident at M = 50 was the decrease in hold-out bias when the number of features selected by SFS increased. A similar trend was observed for the equal covariance matrix condition although the dependence on the number of selected features was weaker. This was also true for LDA with SFS under the unequal covariance matrix condition at M=50. However, for the LDA and SVM(poly) classifiers under the equal covariance matrix condition, we observed the opposite trend of a larger hold-out bias when more features were selected. All of the trends above were observed for the range of SFS parameter investigated in our study. For the feature spaces in our study, if an exhaustive search were conducted in which the number of selected features spanned the closed interval [1,M], one would expect the hold-out bias to reach a minimum within the open interval (1,M).

The SVM(poly) classifier performance is shown in Fig. 4.9. Although the high resubstitution A_z was evidence that SVM(poly) with SFFS was overtrained for small training sample size, there was not a corresponding drop in hold-out performance, compared to the hold-out performance with SFS, where the resubstitution bias was not as great.

A comparison of classifier performance with SFS is shown in Fig. 4.10. When there were few training samples available, the SVM(poly) resubstitution bias was similar to those of the LDA and SVM(rad) classifiers. However, when the number of available training samples increased, the SVM(poly) resubstitution bias increased compared to those of the LDA and SVM(rad) classifiers. The SVM(rad) hold-out performance was comparable or slightly higher than the LDA performance, depending on the number of selected features. With a large number of training samples available, the SVM(rad) hold-out performance was better than that of the LDA, especially when M = 200 and the number of selected features was large, although the resubstitution values for LDA and SVM(rad) were similar. Note that, within the feature selection parameters investigated in this study, an increasing number of features selected for SVM(rad) resulted in decreasing hold-out bias, but the opposite trend was observed for SVM(poly) with both SFS and SFFS.

Comparing the classifiers with SFFS in Fig. 4.11, the SVM(poly) classifier had the highest resubstitution bias for small training sample sizes. Although the hold-out performance for SVM(poly) was similar to those of LDA when the training sample sizes were small, SVM(poly) had slightly better performance when large training samples were available. SVM(rad) consistently had a slightly higher hold-out performance than LDA, especially for $M = 50$ with small training sample sizes. This may be attributed to the fact that for the unequal covariance matrix, the LDA classifier is not optimal, and the SVM is a non-linear classifier.

The performance of the LDA, SVM(rad), and SVM(poly) classifiers with various feature selection methods using samples drawn from class distributions having the covariance matrix estimated from clinical data is shown in Fig. 4.12. SVM(rad) had less optimistic resubstitution bias and less pessimistic hold-out bias compared to LDA under the conditions of small training sample size. However, when the training sample size approached about 100 samples per class, LDA with SFS and SFFS provided slightly higher hold-out A_z than SVM(rad).

4.6 Conclusion

The LDA classifier has been used for many classification tasks in CAD applications because of the limited number of samples available for training and testing. A linear classifier would less likely overfit the training data because of the relatively few parameters to be trained. Recently, there has been increased interest in the SVM. Under our simulation conditions, we found that the SVM with the radial kernel performed slightly better than the LDA when the training sample size was small. However, the many variables that need to be selected for the SVM, such as the kernel function and the parameter values, may depend on the specific classification task. A different choice of kernel, such as the polynomial function in this simulation study, may result in lower performance than the LDA under some of the conditions. The limited conditions that we examined in the current simulation study demonstrated that the relative performances of the different combinations of classifier and feature selection methods depend on the feature space distributions, the dimensionality, and the available training sample sizes. Further investigations will be needed to determine if there can be simple rules of thumb to guide the choice among different classifiers, or among the kernel functions for SVM. For the evaluation of feature selection methods, we found that the SFS and the SFFS methods are comparable. It will be of strong interest to evaluate whether other feature selection methods, such as the principal component analysis that selects features independent of the classifier, may have different characteristics than the methods studied.

Choosing effective feature selection and classification methods is a vital part in the design of a CAD system. Although the conditions that we investigated are limited, our study has revealed some interesting properties of these methods and contributed to the knowledge that may facilitate the design of an effective CAD system under the constraint of limited available samples.

4.7 Figures

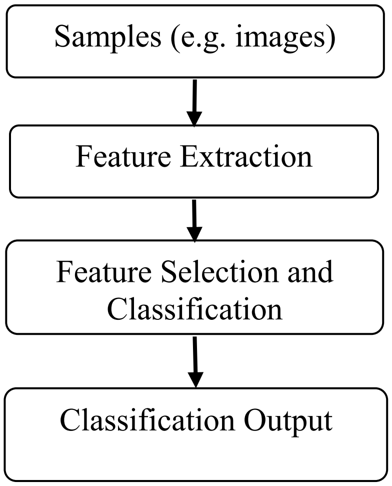


Figure 4.1: Flowchart of a typical computerized classification system.

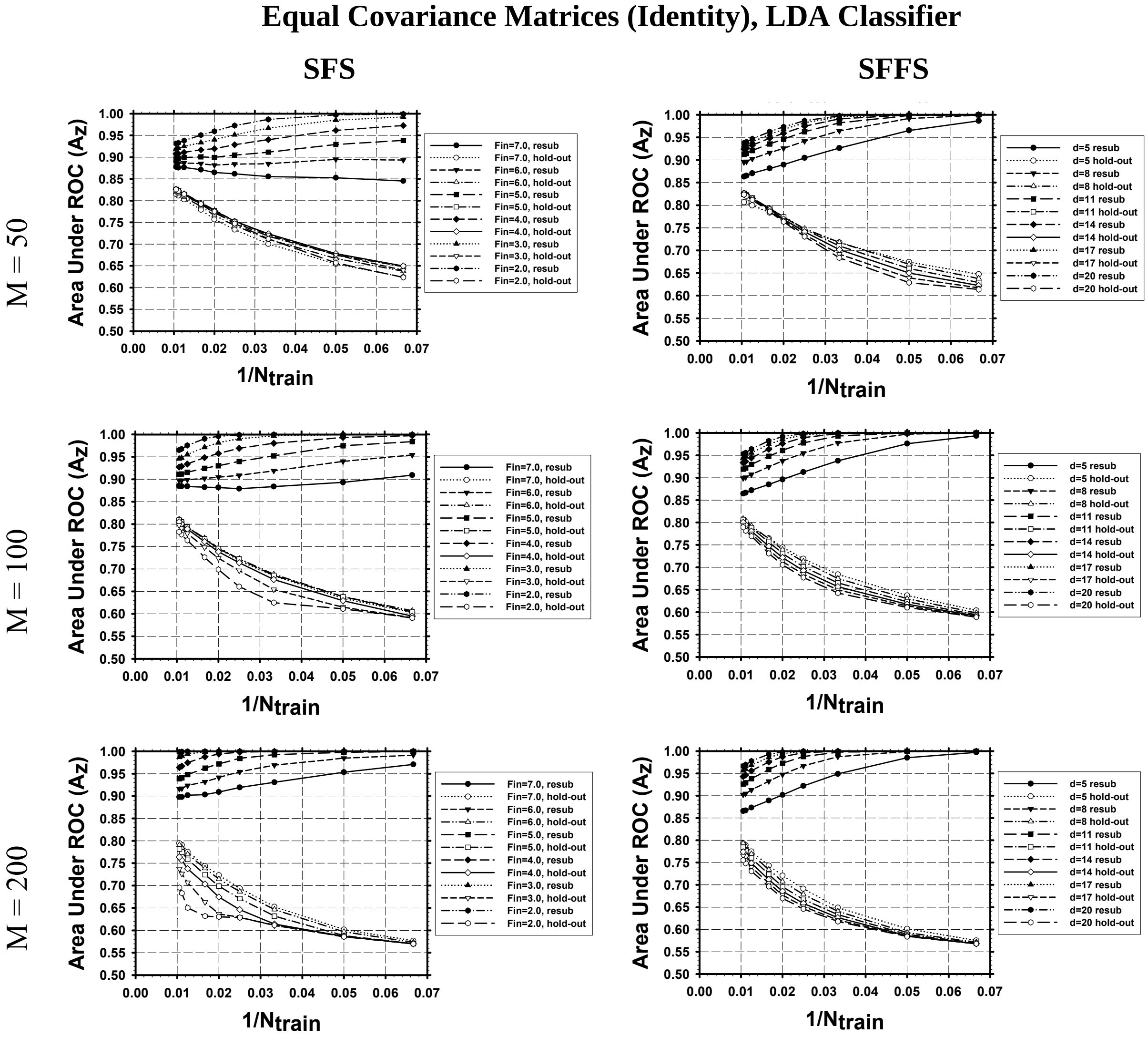


Figure 4.2: Dependence of the LDA classifier performance, A_z, on training sample size. The two class distributions were multivariate normal with equal covariance matrices and unequal means. The effect of increasing dimensionality of the feature space available for selection (*M*) is shown in each column. The comparison of the SFS and SFFS methods is shown in each row.

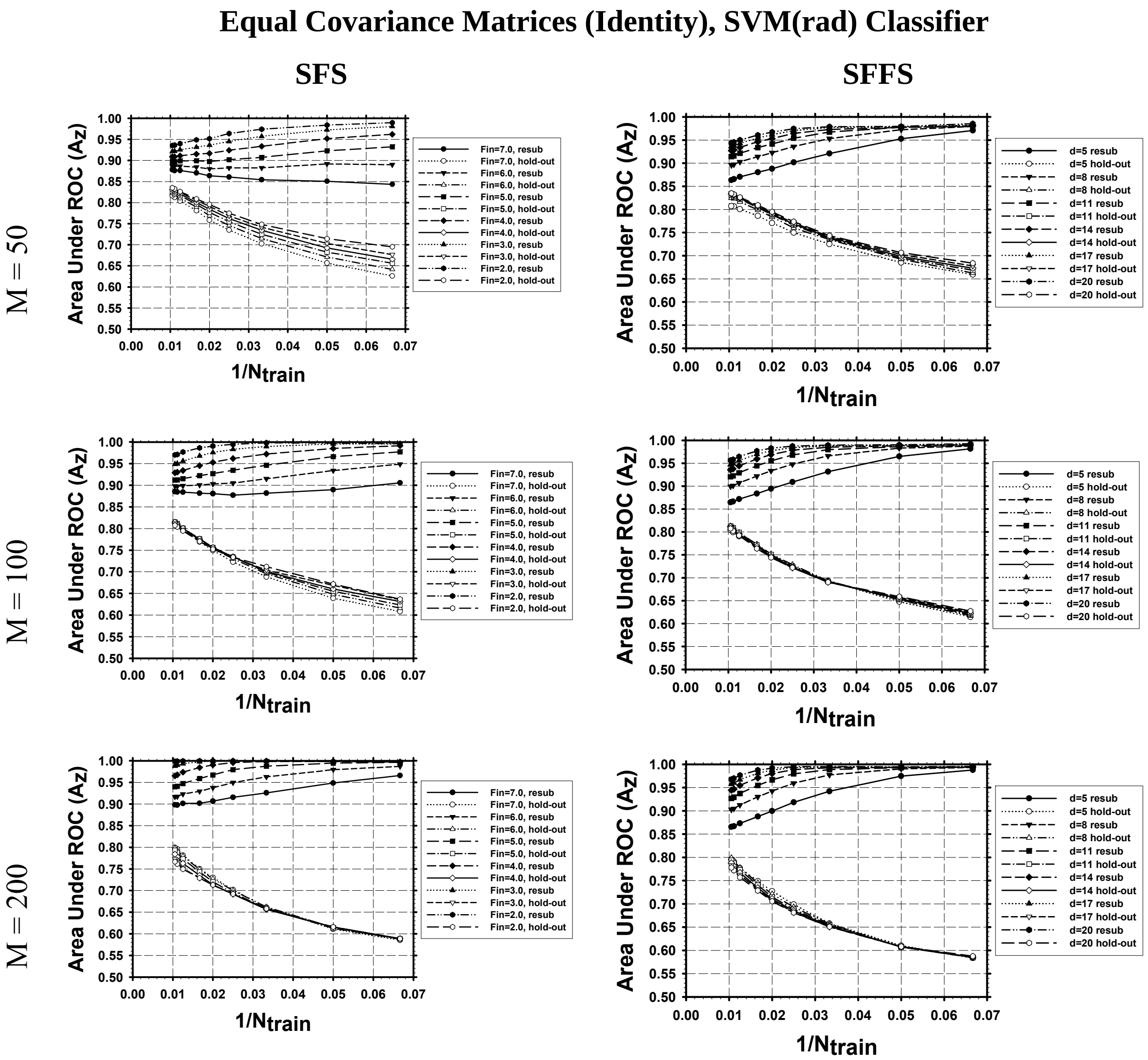


Figure 4.3: Dependence of the performance, A_z, of the SVM classifier with radial kernel on training sample size. The two class distributions were multivariate normal with equal covariance matrices and unequal means. The effect of increasing dimensionality of the feature space available for selection (M) is shown in each column. The comparison of the SFS and SFFS methods is shown in each row.

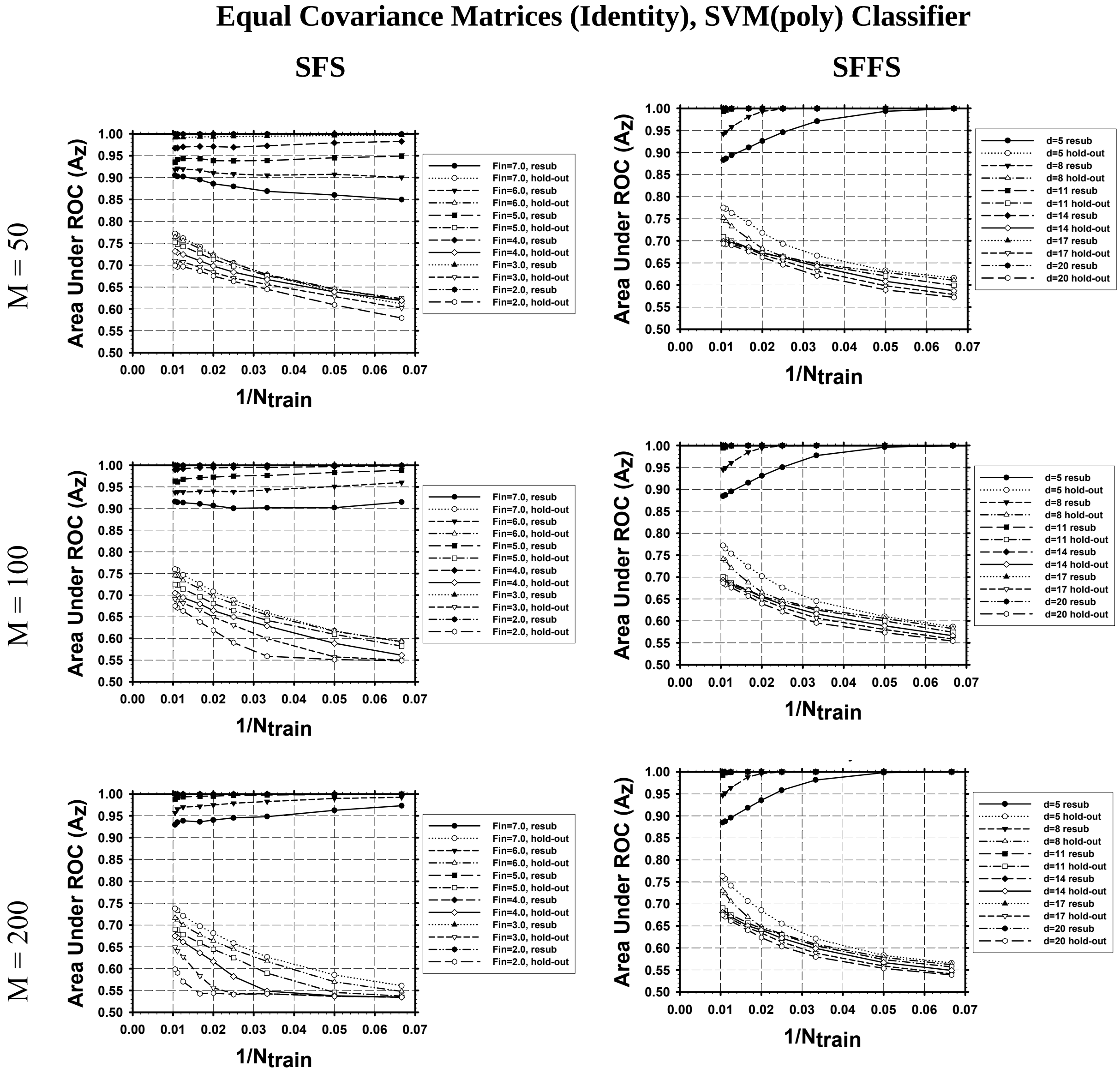
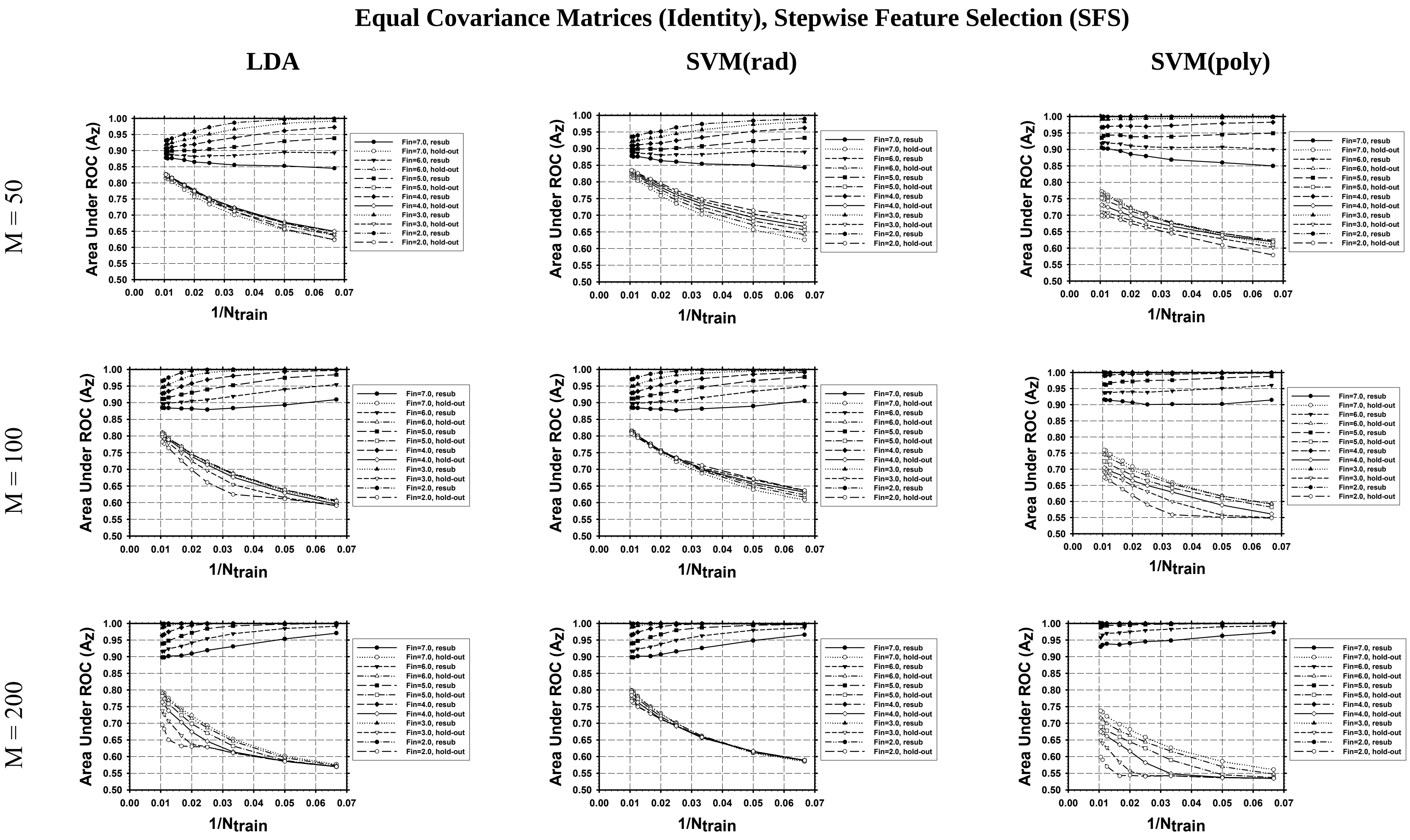


Figure 4.4: Dependence of the performance, A_z, of the SVM classifier with polynomial kernel on training sample size. The two class distributions were multivariate normal with equal covariance matrices and unequal means. The effect of increasing dimensionality of the feature space available for selection (M) is shown in each column. The comparison of the SFS and SFFS methods is shown in each row.



Equal Covariance Matrices (Identity), Stepwise Feature Selection (SFS)

LDA

SVM(rad)

SVM(poly)

M = 50

M = 100

M = 200

Figure 4.5: Comparison of the LDA, SVM(rad), and SVM(poly) classifiers with the same input features obtained from stepwise feature selection. The two class distributions were multivariate normal with equal covariance matrices and unequal means.

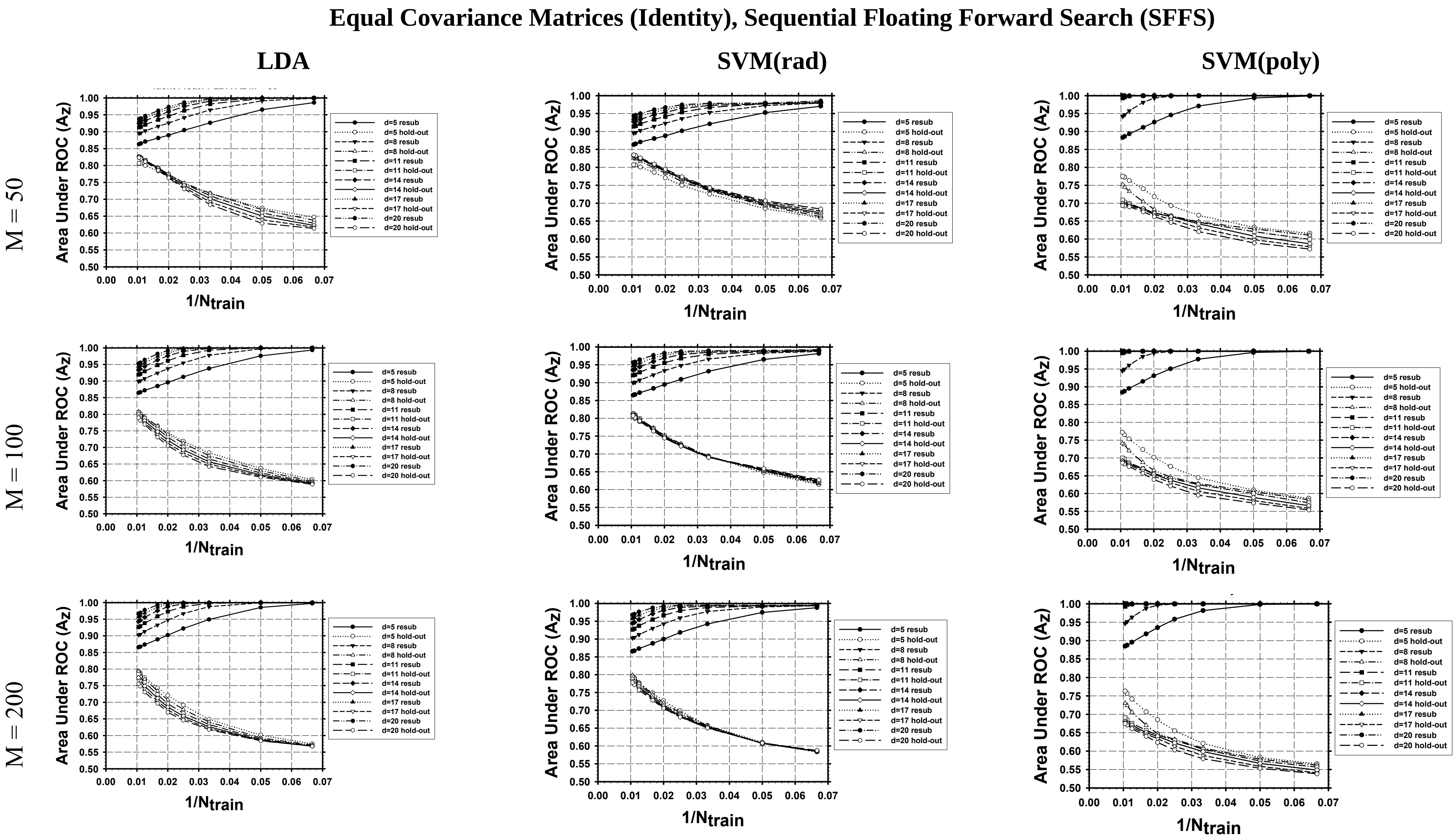


Figure 4.6: Comparison of the LDA, SVM(rad), and SVM(poly) classifiers with the same input features obtained from sequential floating forward search (SFFS). The two class distributions were multivariate normal with equal covariance matrices and unequal means.

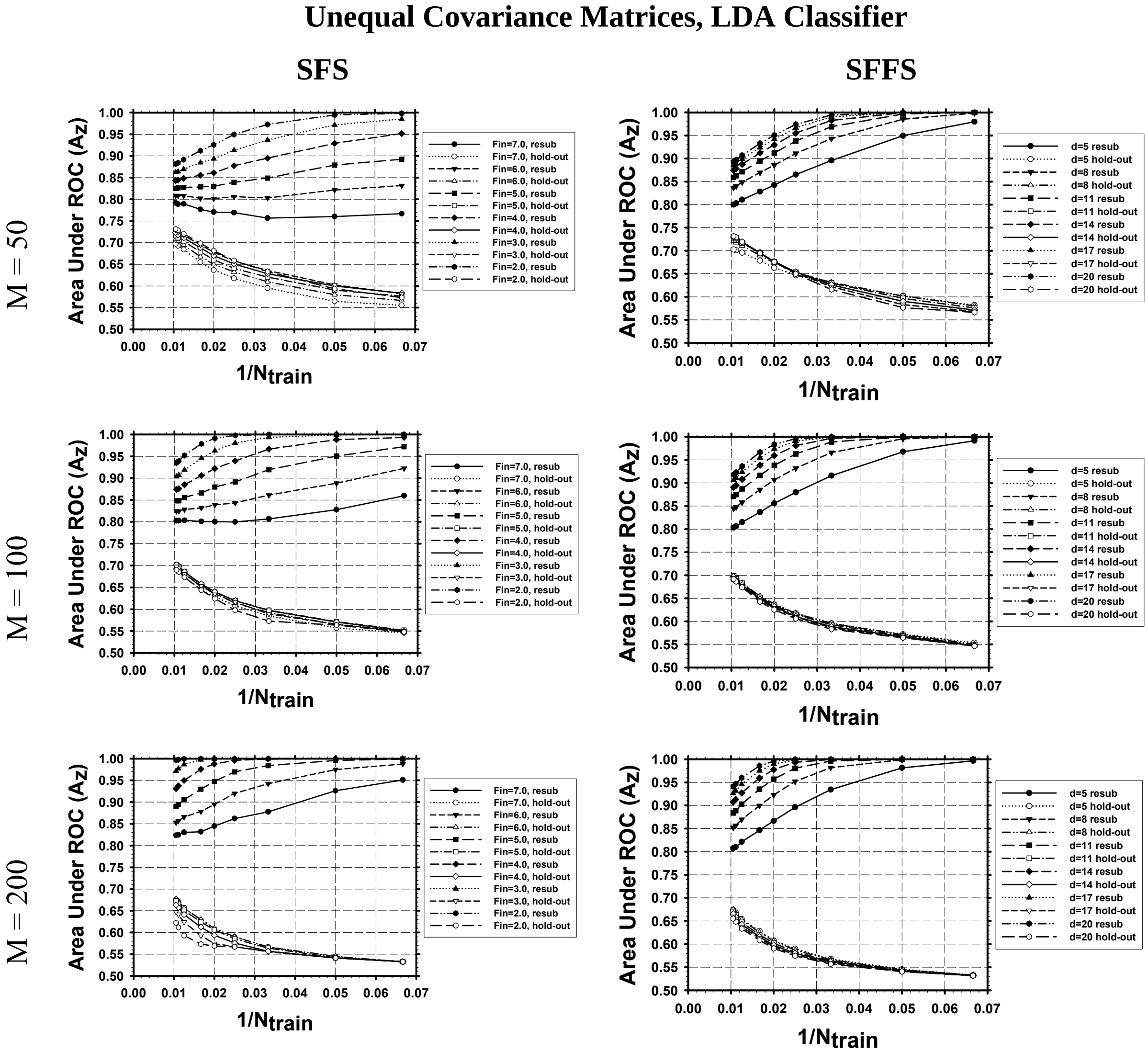


Figure 4.7: Dependence of the LDA classifier performance, A_z, on training sample size. The two class distributions were multivariate normal with unequal covariance matrices and unequal means. The effect of increasing dimensionality of the feature space available for selection (M) is shown in each column. The comparison of the SFS and SFFS methods is shown in each row.

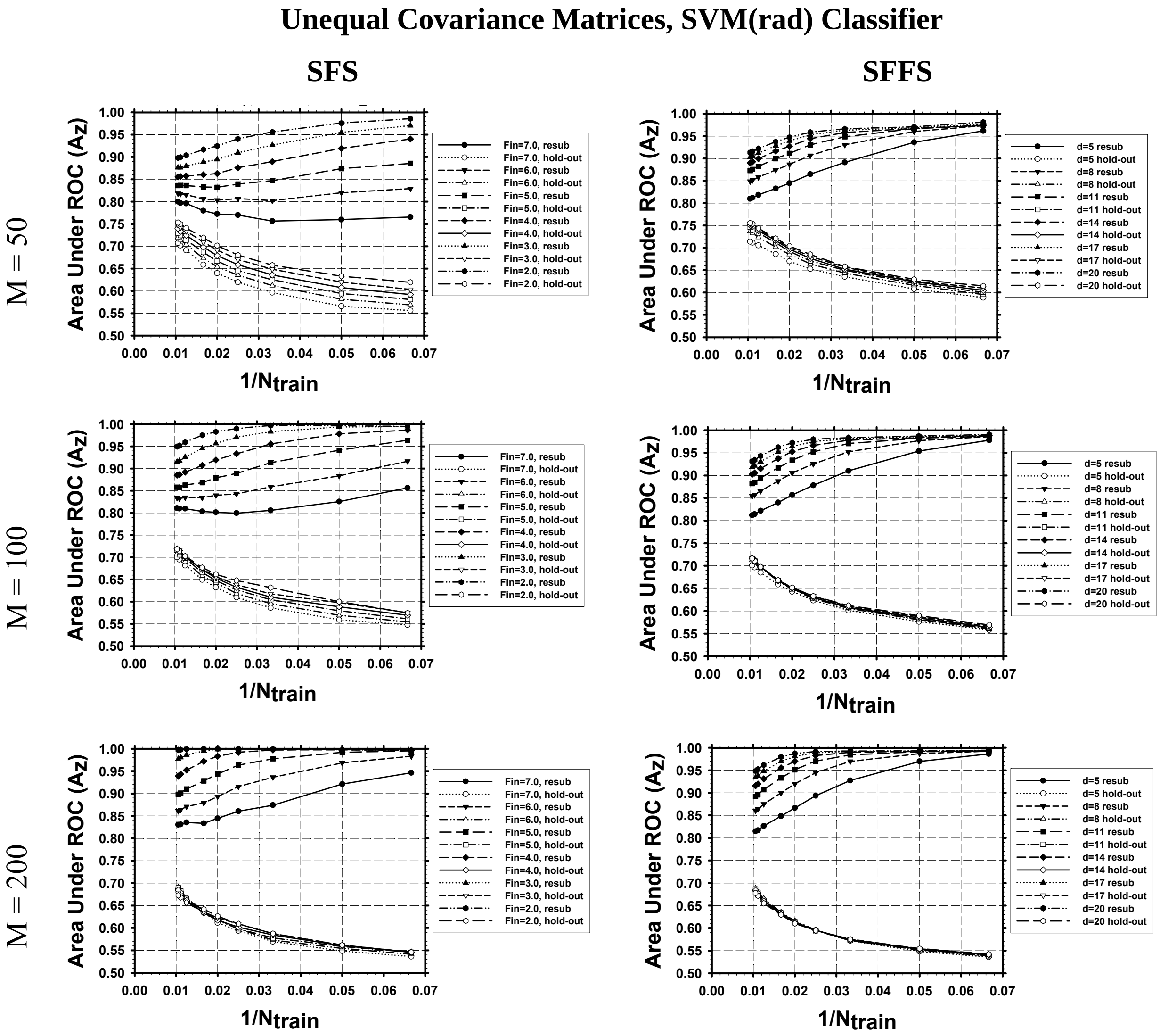


Figure 4.8: Dependence of the performance, A_z, of the SVM classifier with radial kernel on training sample size. The two class distributions were multivariate normal with unequal covariance matrices and unequal means. The effect of increasing dimensionality of the feature space available for selection (M) is shown in each column. The comparison of the SFS and SFFS methods is shown in each row.

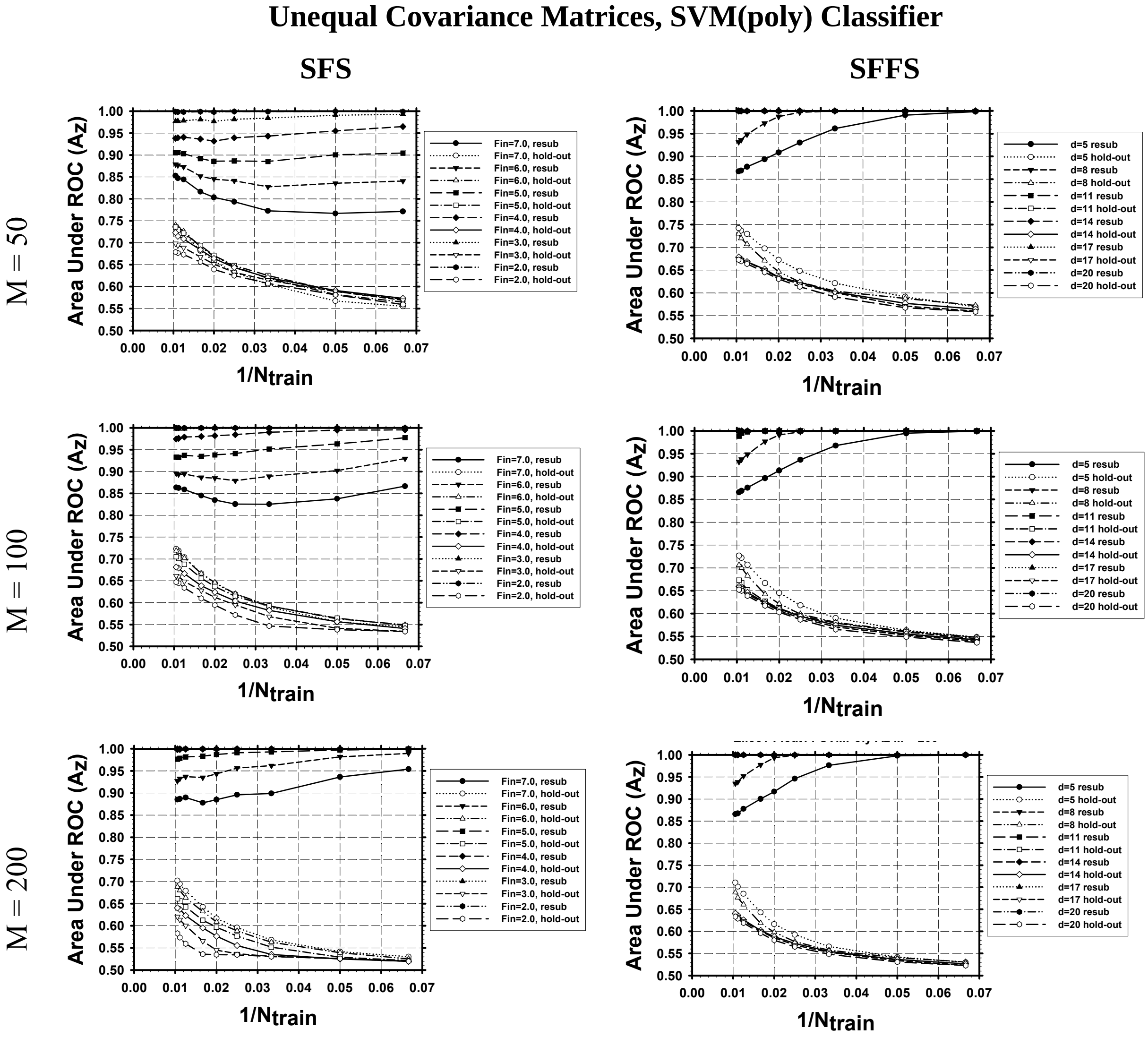


Figure 4.9: Dependence of the performance, A_z, of the SVM classifier with polynomial kernel on training sample size. The two class distributions were multivariate normal with unequal covariance matrices and unequal means. The effect of increasing dimensionality of the feature space available for selection (M) is shown in each column. The comparison of the SFS and SFFS methods is shown in each row.

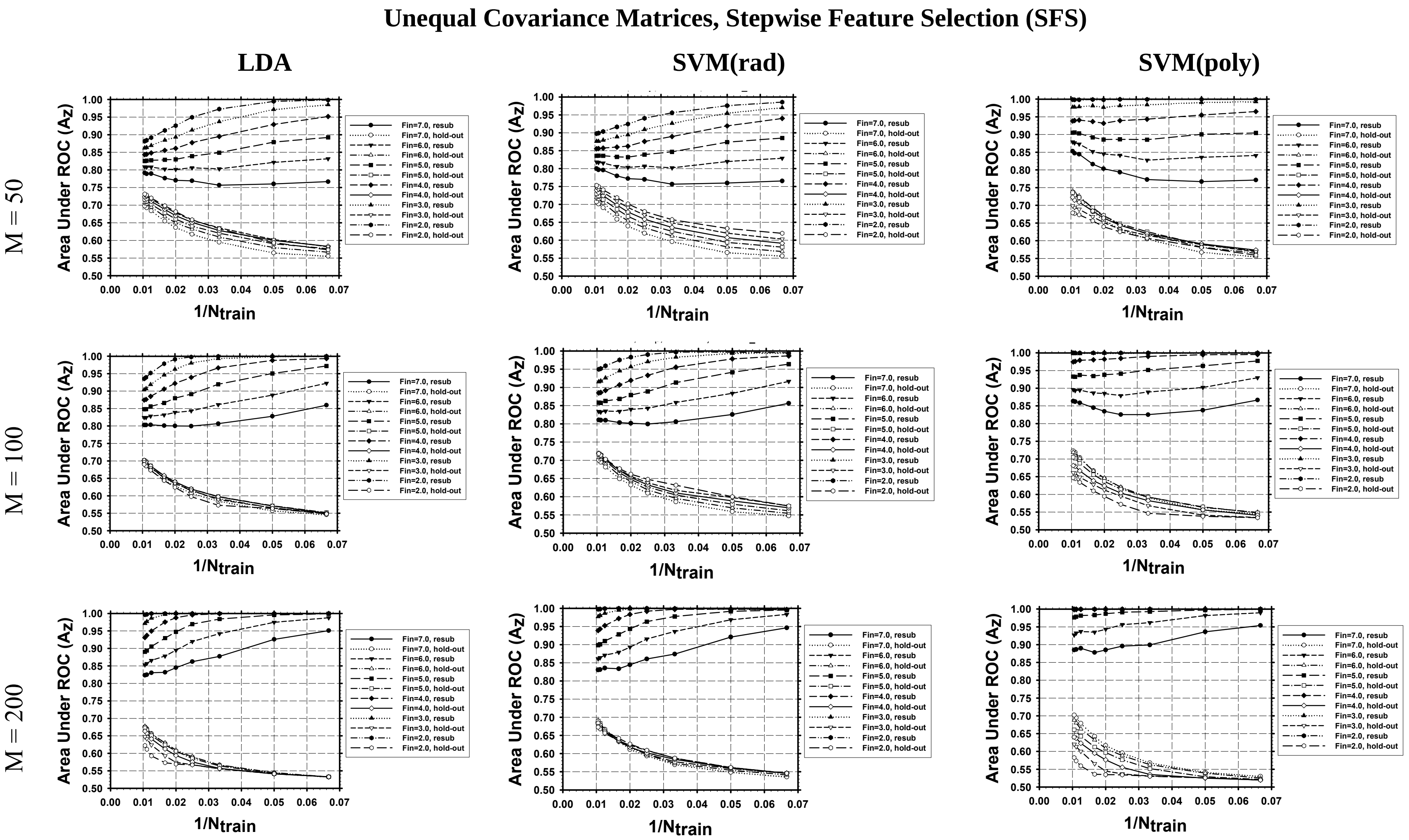


Figure 4.10: Comparison of the LDA, SVM(rad), and SVM(poly) classifiers with the same input features obtained from stepwise feature selection (SFS). The two class distributions were multivariate normal with unequal covariance matrices and unequal means.

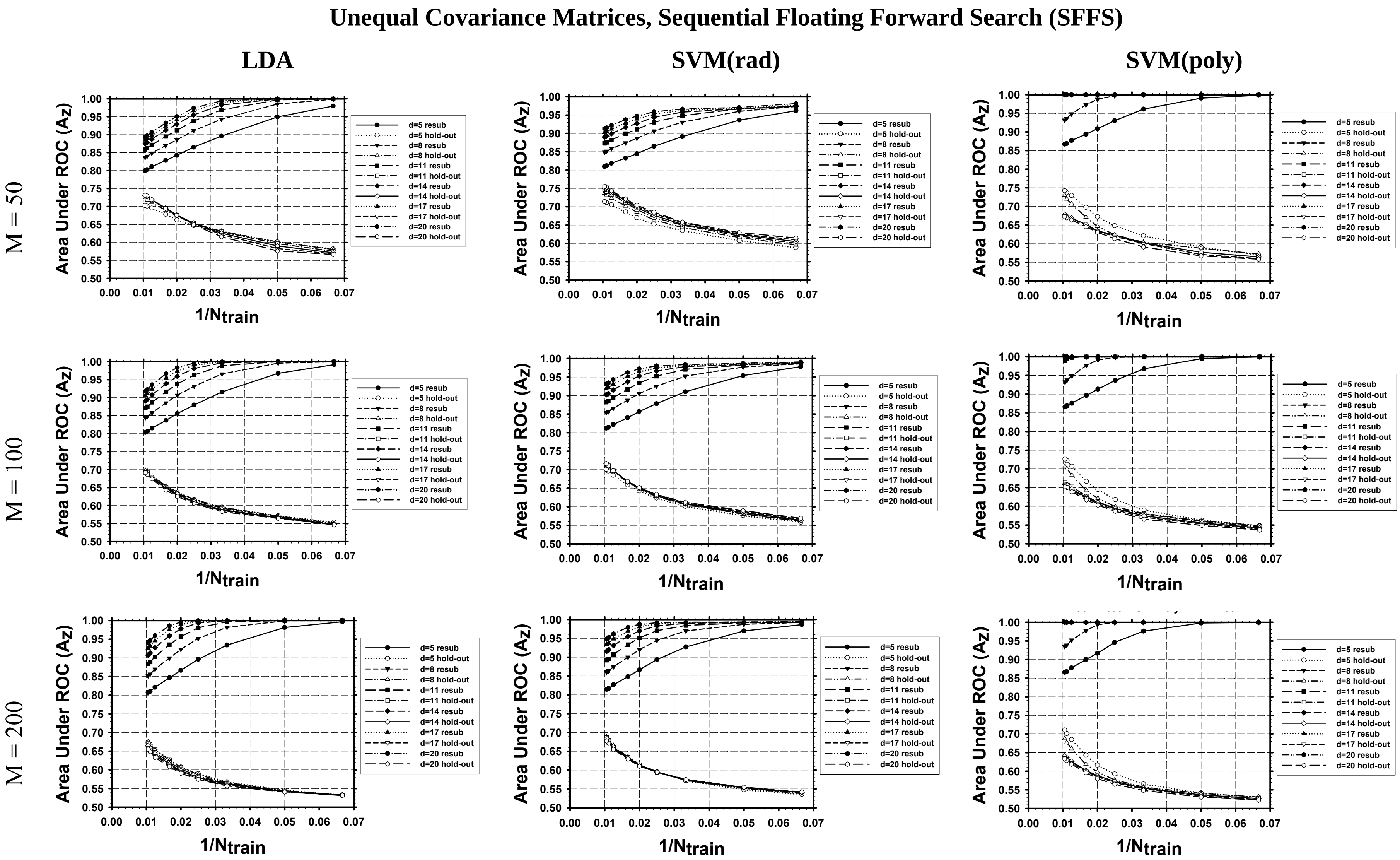


Figure 4.11: Comparison of the LDA, SVM(rad), and SVM(poly) classifiers with the same input features obtained from sequential floating forward search (SFFS). The two class distributions were multivariate normal with unequal covariance matrices and unequal means.

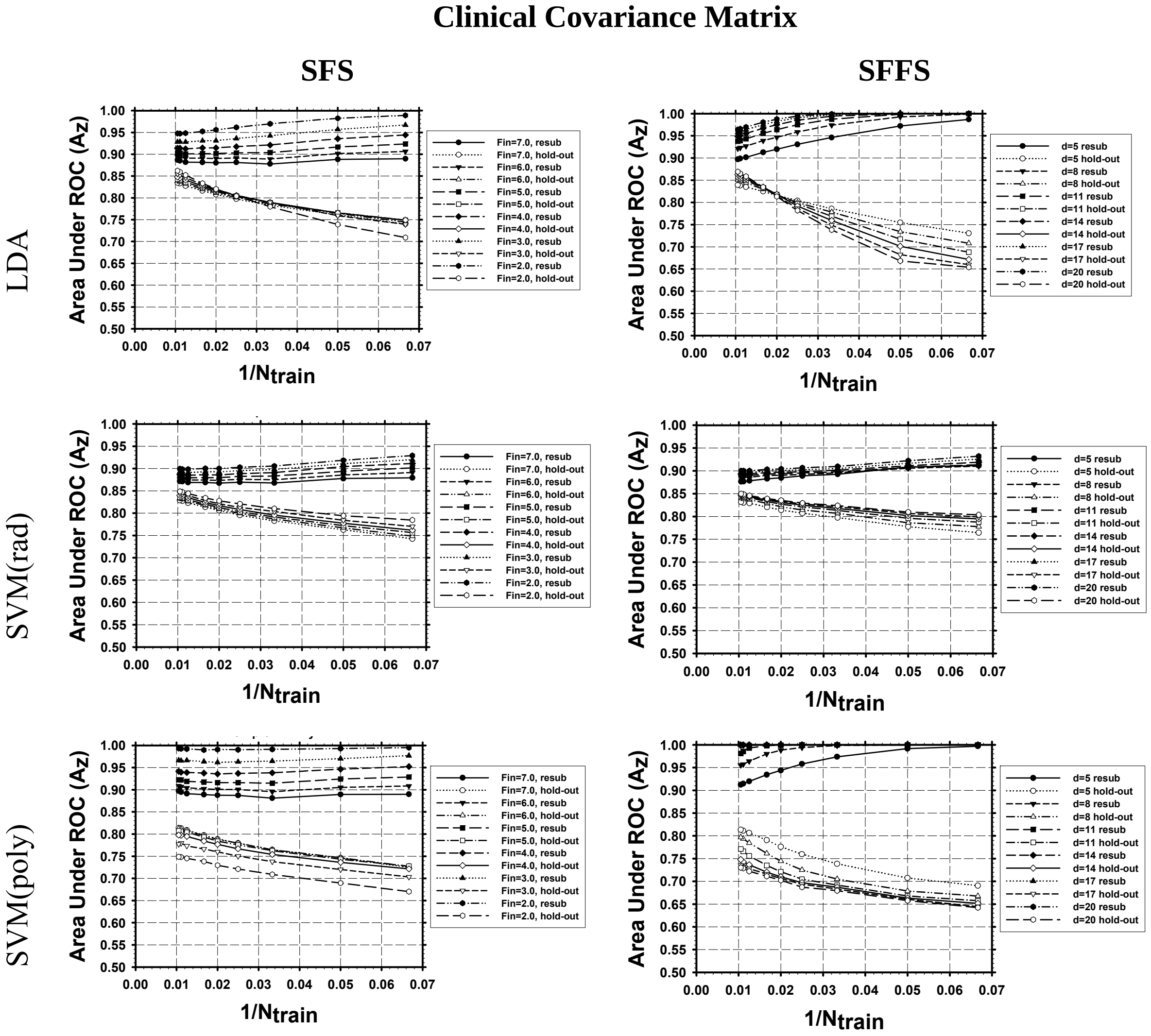


Figure 4.12: Performance of the SFS and SFFS feature selection methods and the LDA, SVM(rad), and SVM(poly) classifiers for simulated multivariate normal class distributions with equal covariance matrices estimated from a clinical data set (M=61).

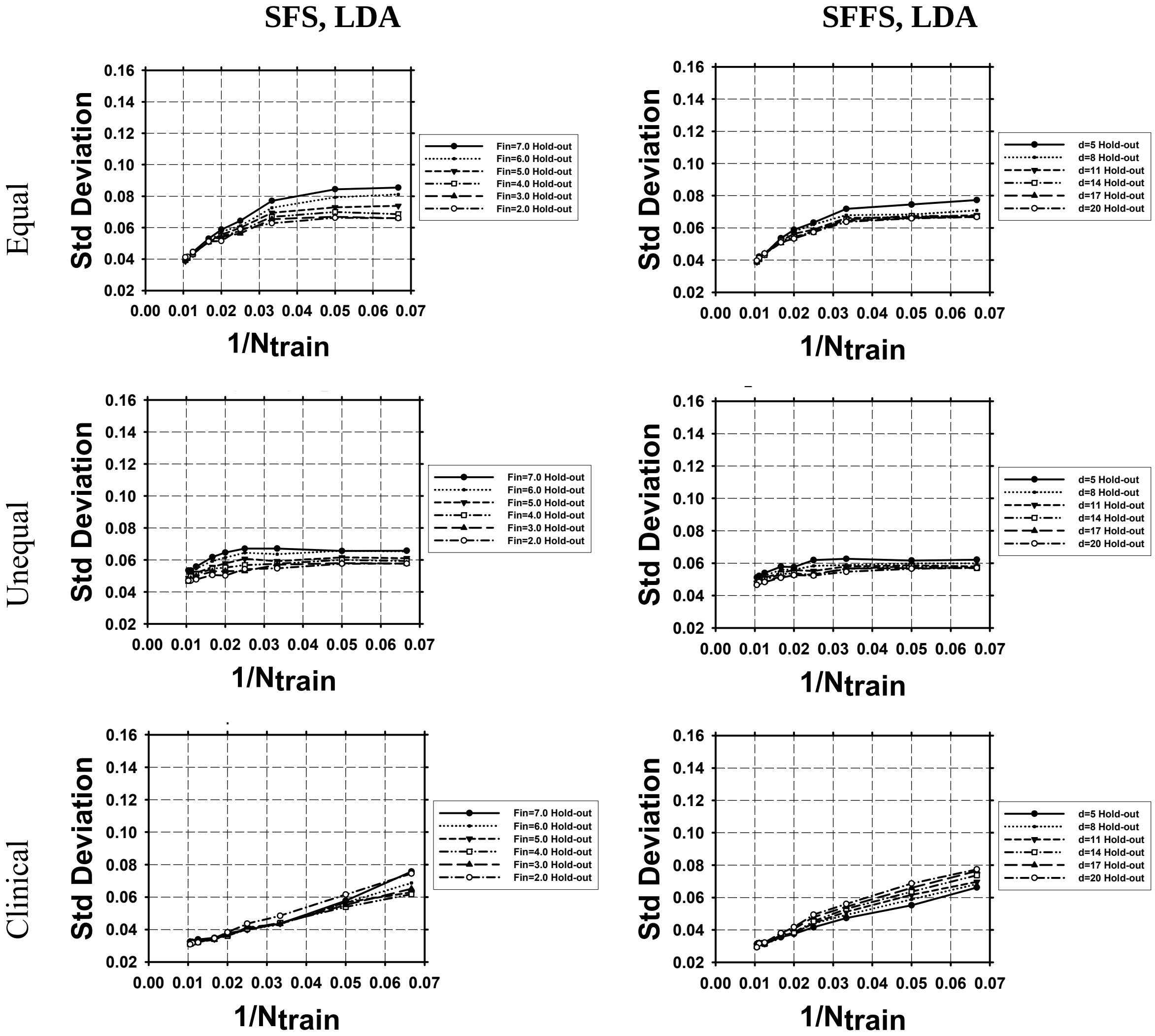


Figure 4.13: Standard deviation as a function of $1/N_{train}$ for the SFS and SFFS feature selection methods and the LDA classifier. The number of features available for selection was M=100 for the equal covariance matrices (first row) and unequal covariance matrices (second row) conditions, and M=61 for the condition with simulated equal covariance matrices estimated from a clinical data set.

Chapter 5
Computer-Aided Diagnosis of Pulmonary Nodules on CT Scans: Improvement of Classification Performance with Nodule Surface Features

5.1 Abstract

We have developed a computer-aided diagnosis (CAD) system to differentiate malignant and benign lung nodules on CT scans. A fully automated system was designed to segment the nodule from its surrounding structured background in a local volume-of-interest (VOI) and to extract image features for classification. Image segmentation was performed with a 3D active contour (AC) method. The initial contour was obtained as the boundary of a binary object generated by *k*-means clustering within the VOI and smoothed by morphological opening. A data set of 256 lung nodules (124 malignant and 132 benign) from 152 patients was used in this study. In addition to morphological and texture features, we designed new nodule surface features to characterize the lung nodule surface smoothness and shape irregularity based on the statistics of the gradient field and the radii segments. The effects of two demographic features, age and gender, as adjunct to the image features were also investigated. A linear discriminant analysis (LDA) classifier built with features from stepwise feature selection was trained using simplex optimization to select the most effective features. A two-loop leave-one-out resampling scheme was developed to reduce the optimistic bias in estimating the test performance of the CAD system. The area under the receiver operating characteristic (ROC) curve, A_z, for the test cases improved significantly ($p<0.05$) from 0.821 ± 0.026 to 0.857 ± 0.023 when the newly developed image features were included with the original morphological and texture features. A similar experiment performed on the data set restricted to primary cancers and benign nodules, excluding the metastatic cancers, also resulted in an improved test A_z, though the improvement did not reach statistical significance ($p=0.07$).

The two demographic features did not significantly affect the performance of the CAD system (p>0.05) when they were added to the feature space containing the morphological, texture, and the new gradient field and radii features. To investigate if a support vector machine (SVM) classifier can achieve improved performance over the LDA classifier, we compared the performance of the LDA and SVMs with various kernels and parameters. Principal component analysis (PCA) was used to reduce the dimensionality of the feature space for both the LDA and the SVM classifiers. When the number of selected principal components was varied, the highest test A_z among the SVMs of various kernels and parameters was slightly higher than that of the LDA in one-loop leave-one-case-out resampling. However, no SVM with fixed architecture consistently performed better than the LDA in the range of principal components selected. This study demonstrated that our proposed segmentation and feature extraction techniques are promising for classifying lung nodules on CT images.

5.2 Introduction

Lung cancer is the leading cause of cancer death in the United States, causing an estimated 160,400 deaths in 2007. At the time of diagnosis, most patients already present with advanced disease. Despite advances in treatment and diagnosis, the five-year overall survival rate is only 15% [3]. As for earlier detection, the "serendipitous discovery of lung cancer in asymptomatic people is currently the principal way in which stage I lung cancer is detected" [4]. Thus, there is great interest in determining whether earlier detection can reduce the mortality rate. Previous trials in the 1970's for screening of lung cancer with chest X-ray and sputum analysis did not result in a significant reduction in mortality [5].

Computed tomography (CT) has been shown to have higher sensitivity in detecting small lung nodules compared to chest X-ray [6-12]. This suggests that CT screening has a strong potential for improving the likelihood of detecting lung cancer at an earlier and potentially more curable stage [13, 120]. A 30-site randomized controlled study (National Lung Screening Trial (NLST)), sponsored by the National Cancer Institute (NCI), has enrolled about 50,000 participants to compare the effect of screening using helical CT or chest x-rays on the mortality rate of lung cancer patients. If CT screening is recommended, however, it would also exacerbate already mounting challenges for detection and diagnosis of lung nodules with CT, namely, interpretation of an ever increasing number of slices and management of a large number of nodules. Despite the increasing spatial resolution of CT, the assessment of the likelihood of malignancy of nodules by visual inspection is difficult. It has been reported that as many as 50% of nodules resected at surgery are benign [12], emphasizing the need to provide radiologists with additional information to improve the accuracy for characterization of nodules and to handle large data sets.

Much work has been reported for development of automated nodule detection methods in CT for computer-aided detection. In this study, we focus on the classification between malignant and benign nodules. Gurney and Swensen [121] conducted a characterization study with a data set of 318 nodules (153 benign and 163 malignant). They trained and tested a neural network in a feature space containing morphological features of the nodule, such as diameter (mm) and appearance of the edge, and demographic features such as age in years and smoking history in pack-years provided by radiologists. They found that the neural network achieved an area under the receiver operator characteristic (ROC) curve A_z of 0.871, but concluded that Bayesian analysis was a better predictor of malignancy with an A_z of 0.894 ($p<0.05$).

Although data sets were smaller for other preliminary studies, the results were encouraging. The features were extracted from the image data, with the goal of quantifying the visual features radiologists typically use to discriminate malignant from benign nodules. Kawata *et al.* [50] used surface curvatures and ridge lines as features for characterization of 62 nodules (47 malignant and 15 benign), and showed good evidence of separation between malignant and benign classes in feature maps; no A_z value was reported. McNitt-Gray *et al.* [47] obtained 90.3% correct classification accuracy between 17 malignant and 14 benign cases. Shah *et al.* [48] achieved A_z values between 0.68 and 0.92 with 48 malignant and 33 benign nodules, using four different types of classifiers in a leave-one-out method. The features were extracted from contours manually drawn on a single representative slice of each nodule. Way *et al.* [1] developed an automated 3D active contour segmentation method and extracted morphological and texture features from the segmented nodule. A leave-one-out test A_z of 0.83±0.04 was achieved in a data set of 44 malignant and 52 benign nodules.

Several classification studies were performed with a larger data set, although the number of malignant nodules was still below 100. Armato *et al.* [49] used an automated detection scheme, then manually separated nodules from non-nodules before the classification step. They achieved an A_z value of 0.79 for 59 malignant and 276 benign nodules using features such as the radius of a sphere of equivalent volume, minimum and maximum compactness, gray-level threshold, effective diameter, and location in the lungs. Li *et al.* [51] reported an A_z of 0.937 for differentiation between 61 malignant and 183 benign nodules in a leave-one-out method, and an A_z of 0.831 for a randomly selected subset consisting of 28 primary lung cancers and 28 benign nodules. The features used included the diameter and contrast of the segmented nodule, and those extracted from the gray-level histograms of pixels inside and outside the segmented nodule. Aoyama *et al.* [52] reported an A_z of 0.846 for classifying 76 primary lung cancers and 413 benign nodules using multiple slices (10 mm collimation and 10 mm reconstruction interval), which was a statistically significant improvement over an A_z of 0.828 when using only single slices. Suzuki *et al*. [53] obtained an A_z of 0.882 by use of a massive training artificial neural network (MTANN) on a data set of 76 malignant and 413 benign nodules.

We are developing a CAD system to assist radiologists in the classification task. Our CAD system automatically segments a nodule from a volume of interest (VOI) on CT images and provides a malignancy rating based on features extracted from the images. Our preliminary results have been reported previously [1]. In this study, we have designed new image features that characterize the nodule boundary and improved the classifier training with an enlarged data set. In addition we investigated the effect of age at the time of CT exam and gender as demographic features. Finally, we compared the performance between the linear discriminant analysis (LDA) and support vector machine (SVM) classifiers.

5.3 Methods and Materials

5.3.1 CT Scan Collection

We retrospectively collected CT scans from the patient files in the Department of Radiology at the University of Michigan with Institutional Review Board (IRB) approval.

The CT scans were acquired with a variety of GE (GE, Waukesha, WI) Genesis HiSpeed and the GE LightSpeed series scanners, including Plus, Power, Pro 16, QX/i, Ultra, and LightSpeed16. Each CT slice was 512 x 512 pixels, with pixel sizes ranging from 0.448 to 0.859 mm and corresponding fields-of-view of 25 to 44 cm. The slice thickness averaged 2.3 ± 1.44 mm (range: 1 to 7.5 mm), and the slice interval averaged 2.0 ± 1.6 mm (range: 0.6 to 7.5 mm). The average values for the scanning parameters were 120 kVp for tube voltage (range:120 to 140 kVp), 209 ± 92 mA for tube current (range: 80 to 500 mA), and 214 ± 141 mAs (range: 40 to 570 mAs).

5.3.2 Lung Nodule Data Set

For this study, 256 lung nodules (124 malignant and 132 benign) were identified by radiologists from 152 patients. Of the 124 malignant nodules, 64 were biopsy-proven to be malignant, and 60 nodules were determined to be malignant either through positive PET scans, being in the same lung as other biopsy-proven malignant nodules, or known metastases from confirmed cancers in other body parts based on the patients' clinical reports. Seventy two were primary and 52 were metastatic cancers. Of the 132 benign nodules, 15 were biopsy-proven and 117 were determined to be benign by two-year follow-up stability on CT. Experienced chest radiologists indicated the location, measured the size, and provided a description of its characteristics and a malignancy rating for each nodule. Of the 256 nodules, 53 were juxta-pleural and 19 were juxta-vascular. A distribution of the longest diameters of the nodules is shown in Fig. 5.1. The nodules had an average longest diameter of 11.7 mm +/- 7.7 (range: 3.0 mm-37.5 mm). Fig. 5.2 shows a distribution of the malignancy ratings of the nodule provided by radiologists, on a scale of 1-5, with 5 indicating most likely malignant. The area, A_z, under the receiver operating characteristic (ROC) curve fitted to the radiologists' malignancy ratings is 0.806±0.028.

5.3.3 CAD System Overview

A detailed description of our CAD system can be found in the literature [1]. A short summary is provided here, and the flowchart is shown in Fig. 5.3. First, the radiologist-identified VOI containing nodule was extracted from the CT scan. We performed linear interpolation in the *z*-direction if the slice interval was greater than the pixel size, or bilinear interpolation in the axial plane otherwise, to obtain isotropic voxels to facilitate initial contour generation and segmentation. To generate the initial contour, *k*-means clustering assigned voxels not part of the mediastinum or chest wall in the VOI to either the object or the background class. Morphological opening was performed with a spherical structuring element that had an automatically calculated size based on the size of the clustered object. The morphological opening may remove attachments such as blood vessels from the object. A 3D active contour (3DAC) model was then used to segment the nodule in the VOI. We estimated the weights for the 3DAC based on the optimization method with classification performance as the figure-of-merit described in our previous study [1]. The segmentation was optimized separately using the feature space with and without the new image features for the performance comparison described below. After optimization, the same set of weights was used to segment all the nodules for the given feature space.

From the nodule contour, 2D and 3D morphological features were extracted. A few examples and descriptions of morphological features are given here, and the rest are described in the literature [1]. The volume was found by multiplying the number of voxels within the contours by the size of one voxel. The longest diameter was the longest distance between two points on a contour. Statistics such as the average, standard deviation, skewness, minimum, and maximum of the CT values (Hounsfeld Units) of the nodule voxels were calculated.

To quantify texture around the nodule, texture features were extracted first from the individual 2-D image slices that intersect the nodule, and the corresponding features were averaged over the nodule slices. For a given slice, the rubber band straightening transform [64] converted the 15-pixel-wide band of pixels surrounding the nodule into a rectangular image. The nodule boundary was mapped to the horizontal dimension of the rectangle while the spiculations emanating radially from the nodule became mapped to an approximately vertical direction. The transformed image was enhanced with Sobel filtering in the vertical and horizontal directions, from which the run-length statistics (RLS) features [65, 66] were calculated.

In this study, we included new features in the feature space, as described in the next section. A feature selection method was then applied to the multidimensional feature space to select the most effective features for the classification task. A feature classifier was trained with the selected features. The performance of the trained classifier was evaluated with test cases and the classification accuracy was quantified by ROC analysis.

5.3.4 Gradient Field and Radiii Features

In addition to the morphological and texture features, we designed three sets of new features to characterize of the nodule surface smoothness and shape irregularity. The first two sets were gradient magnitude and profile features, which were based on the gradient field, and the last set contained statistics of the nodule radii. The gradient vector and its magnitude M_v were computed at each voxel v using a filter-based method as described in Ge *et al.* [93], which was a generalization of the 2D isotropic kernel proposed by Jain [122].

5.3.4.1 Gradient Magnitude

The gradient magnitude features described the sharpness of the nodule boundary. Let F be the set of gradient magnitude values for all voxels on the surface of the nodule segmented by the 3DAC method. We found the mean, standard deviation, variance, minimum, maximum, skewness, kurtosis, and coefficient of variation (standard deviation/mean) for all values in set F. A nodule with well-defined boundary would have a higher mean than a nodule with less distinct boundary.

5.3.4.2 Profile Features

Profile features describe the smoothness of the gradient magnitudes in a shell of voxels just inside and outside the nodule surface. The weighted centroid C, of the segmented nodule was calculated with the weights based on voxel intensity. The segment from this centroid to a surface voxel v is referred to as the radius, $\boldsymbol{r}_v$, where v=1…n surface voxels. The radii lengths from the centroid to each surface voxel were stored. The average radius rad_{avg} of the nodule is defined as the average of all the lengths of $\boldsymbol{r}_v$ from C to each surface voxel v. Along this radial line and centered at surface voxel v, the gradient magnitude values were sampled at one pixel intervals to a distance of (½

rad_{avg}) on the two sides of the surface voxel. Let P_v be the set of sampled gradient magnitude values along $\boldsymbol{r_v}$, $M_{v,i}$ be the i^{th} sample, $|P_v|$ be the cardinality of P_v. Then the average gradient magnitude along one vector is:

$$A_{avg,v} = \left(\frac{1}{|P_v|} \sum_{i=1}^{|P_v|} M_{v,i} \right) \tag{5.1}$$

The features we calculated are listed below, and mathematical formulas for some of the features are given:

- PF1 (Profile feature 1): Average of the average gradient magnitudes over all surface voxels

$$PF1 = \frac{1}{n} \sum_{v=1}^{n} \left(A_{avg,v} \right) \tag{5.2}$$

- PF2: Standard deviation of the average gradient magnitudes over all surface voxels

$$PF2 = \sqrt{\frac{1}{n-1} \sum_{v=1}^{n} \left(A_{avg,v} - PF1 \right)^2} \tag{5.3}$$

- PF3: Variance of average gradient magnitudes over all surface voxels
- PF4: Mean of maxima

$$PF4 = \frac{1}{n} \sum_{v=1}^{n} \left(\max\{P_v\} \right) \tag{5.4}$$

- PF5: Standard deviation of maxima
- PF6: Variance of maxima
- PF7: Mean of minima

$$PF7 = \frac{1}{n} \sum_{v=1}^{n} \left(\min\{P_v\} \right) \tag{5.5}$$

- PF8: Standard deviation of minima
- PF9: Variance of minima

It can be expected that high contrast nodules would have high values of PF1 and PF4. Nodules with mixed-GGO (ground glass opacity) might have high values of PF2.

5.3.4.3 Radii Features

The radii features were calculated based on the lengths of $\mathbf{r}_v$, the segment from weighted centroid C to surface voxel v:

- RA1: The average of all radii

$$RA1 = \frac{1}{n}\sum_{v=1}^{n}\left|\boldsymbol{r}_v\right| \tag{5.6}$$

- RA2: The standard deviation of all radii
- RA3: The variance of all radii
- RA4: The skewness of all radii
- RA5: The kurtosis of all radii

It can be expected that a spherical nodule with a smooth surface would have very low values of RA2, RA3, and RA4, and high values of RA5, since all the radii would be similar. The radii segments of an irregularly-shaped nodule would have varying lengths, with expected high values of RA2 and RA3. These features may therefore be useful in quantifying a nodule's surface smoothness. RA1 is another feature that described the size of the nodule.

5.3.5 Demographic Features

We investigated the effect of patient characteristics including age at the time of the scan and gender as adjunct information for the CAD system. Although most CAD systems only utilize image features, the use of demographic information has been found beneficial [121, 123]. For our data set, we did not obtain smoking history consistently in the patient files so that this potentially useful information cannot be included.

5.3.6 Two-loop Leave-one-case-out Resampling

A feature classifier was trained to differentiate the malignant and benign nodules in the multidimensional feature space described above. We designed a "two-loop" leave-one-case-out resampling scheme to estimate the test performance of the CAD system. In comparison to the commonly used one-loop leave-one-case-out resampling, this method introduces another level of independence and reduces the bias in test A_z. In our data set, the 256 nodules were extracted from 152 patients so that the number of independent cases

N was equal to 152. When a case was left out as a test case in the leave-one-case-out scheme, all nodules from that case are taken out and reserved for testing.

In the two-loop leave-one-case-out resampling scheme (Fig. 5.4), an inner leave-one-case-out loop was nested within the outer leave-one-case-out loop. For a data set with *N* available cases, there were *N* cycles in the outer loop. In each cycle, one case was excluded as the independent test case. The remaining (*N*-1) training cases were used to build the classifier in an inner leave-one-case-out loop that included feature selection and classifier weight determination. Stepwise feature selection (SFS) with LDA was used to select a subset of effective features. In each cycle of this inner loop for feature selection, (*N*-2) cases were used for training while one case was left out as the test case. The best parameters for SFS, namely, the F_{in} and F_{out} for determining whether a feature should be included or removed from the feature space, respectively, and the *tol* threshold for the tolerance on how correlated the selected features can be, were searched by simplex optimization using the test A_z from the (*N*-1) left-out cases in the inner loop as a guide. After the best SFS parameters were determined, they were applied to the (*N*-1) training cases of the outer loop to select a subset of features from the available feature space, and an LDA classifier using the selected features as the input predictor variables was formulated using the (*N*-1) training cases. This classifier was then applied to the independent left-out case in the outer-loop and a test score for each nodule in that case was obtained. The procedure was cycled through the *N* cases of the entire data set, so that each case was left-out in turn, resulting in independent test scores for all the nodules in the data set. These 256 test scores were then evaluated by ROC analysis to obtain the two-loop test A_z. Since the test case was kept out of the SFS parameter estimation, feature selection and classifier weight training processes, the estimated performance using the two-loop resampling scheme was less optimistically biased than the one-loop scheme.

5.3.7 Evaluation of CAD System on the Entire Data Set and on Primary and Metastatic Nodules

The CAD system without and with the newly developed features described in Section 5.3.4 in addition to the demographic information was evaluated on the entire data set. Furthermore, nodules from primary cancers and metastases have distinctive characteristics. The former are more likely to be irregularly shaped or spiculated whereas the latter are often round and smooth. We therefore also evaluated the performance of the CAD system using two subsets of the data set, one containing primary cancers and benign nodules, and the other metastatic cancers and benign nodules. For each of the two subsets, a new set of weights for the 3DAC segmentation was determined using the procedure described previously [1]. The two-loop test A_z and features selected were compared.

5.3.8 Comparison between LDA and SVM

We compared the classification performance of LDA with that of SVMs. Since the SFS method described above used the LDA classification result as a guide, the selected feature set may be biased towards LDA. We therefore used principal component analysis (PCA), which is a well-known method for dimensionality reduction and is independent of the choice of the classifier, to obtain a reduced set of features as input to both classifiers for this comparison. PCA transforms a number of correlated variables into a number of uncorrelated variables, i.e., the principal components. It performs eigenvalue decomposition of the covariance matrix of the features, projecting the multivariate feature vectors onto the space spanned by the eigenvectors. The order of a principal component represents its importance in accounting for the variance in the data set. The dimensionality of the feature space is reduced by retaining the lower-order (higher-magnitude) principal components that are most important while ignoring the higher-order ones. Retaining only the lower-order principal components is essentially equivalent to approximating the data by a linear subspace using the mean squared error criterion [124].

The SVM works similarly to the LDA by constructing a decision hyperplane to separate classes using training data. A brief overview of the SVM is given below, with more details in the literature [115]. Geometrically, the SVM maps the original data to a higher dimensional Euclidean space H, via a kernel K. A decision hyperplane is constructed in this higher dimensional space such that the distance between the training samples of both classes and the hyperplane is maximized. This distance between a training sample and the hyperplane is called the margin, and the SVM calculates the hyperplane with the largest margin.

Suppose we have labeled training data {$\boldsymbol{x}_\mathrm{i}$, y_i}, $i = 1 \ldots t$, $y_i \in \{-1,1\}$, and $\boldsymbol{x}_\mathrm{i} \in R^u$, where t is the number of samples, u is the dimensionality (number of features), and y_i is the class label of the i^{th} sample that can assume a value of -1 (class 1) or +1 (class 2). The design of the SVM can be shown to consist of a quadratic programming optimization problem. In the dual of the quadratic program, the data appear in the form of dot products, $\boldsymbol{x}_\mathrm{i}\cdot\boldsymbol{x}_\mathrm{j}$. The SVM algorithm uses a mapping, Φ, to a higher-dimensional Euclidean space H, $\Phi : R^u \mapsto H$. Because of the mapping, the algorithm depends only on data through the dot products in H of the form $\Phi(\boldsymbol{x}_\mathrm{i})\cdot\Phi(\boldsymbol{x}_\mathrm{j})$. There exist kernel functions K so that $K(\boldsymbol{x}_\mathrm{i}, \boldsymbol{x}_\mathrm{j}) = \Phi(\boldsymbol{x}_\mathrm{i})\cdot\Phi(\boldsymbol{x}_\mathrm{j})$, and the training algorithm uses only the kernel K and operations in the lower-dimensional space $\boldsymbol{R}^u$, instead of computationally expensive operations in H. A number of different kernels have been proposed in the literature, and we chose ones commonly used for this study. The dot kernel is the inner product: $K(\boldsymbol{x}_\mathrm{i}, \boldsymbol{x}_\mathrm{j}) = \boldsymbol{x}_\mathrm{i}\cdot\boldsymbol{x}_\mathrm{j}$. The polynomial kernel has the parameter degree z: $K(\boldsymbol{x}_\mathrm{i}, \boldsymbol{x}_\mathrm{j}) = (\boldsymbol{x}_\mathrm{i}\cdot\boldsymbol{x}_\mathrm{j} + 1)^z$. The neural kernel has parameters a and b: $K(\boldsymbol{x}_\mathrm{i}, \boldsymbol{x}_\mathrm{j}) = (a\boldsymbol{x}_\mathrm{i}\cdot\boldsymbol{x}_\mathrm{j} + b)$. The radial kernel is defined as

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2) \tag{5.7}$$

with parameter γ. A capacity parameter, *Cap*, is common to all kernels. We implemented the SVM with the freely available software *mySVM* [119].

From the PCA, we selected the r largest eigenvalues and transformed the data with their corresponding eigenvectors. Since it was not known how many principal components were optimal for this classification task, we varied r from 1 to 15. For the LDA and SVM, we performed leave-one-case-out training and testing for each r to arrive at the test A_z. Because we varied r from 1 to 15 and did not have to choose features, a one-loop leave-one-case-out resampling process was used for classification. In addition, we varied the four kernels and their associated parameters of the SVM to investigate the effect of the kernel and parameters on test performance. A total of 120 (16 polynomial kernels of various degrees, 4 dot kernels of various parameters, 36 radial kernels of various γ, and 64 neural kernels of various coefficients) separate leave-one-case-out training and testing processes were performed for each *r*.

5.4 Results

There were four groups of features used in this study: morphological (M), texture features extracted from the RBST images (T), newly developed image features based on the gradient field and radii features (G), and demographic (D) features. In the following results, the subscript denotes which groups of features were included in the feature space, e.g., $Feat_{MTG}$ is the feature space containing the morphological, texture, and newly-developed image features.

5.4.1 Effect of Gradient Field and Clinical Features on Classification

The training and test A_z were calculated from the two-loop procedure described in Section 5.3.6. When $Feat_{MT}$ was used as the feature space, the CAD system achieved an average training A_z of 0.858 ± 0.023 and two-loop test A_z of 0.821 ± 0.026. An average of 5.80 features was selected. The six most consistently selected features were surface area, maximum CT value, variance of nodule gray-level values, and three RLS texture features. When the newly-developed image features were combined with the previous features, i.e., the $Feat_{MTG}$ feature space, the average training A_z was 0.881 ± 0.021, and the two-loop test A_z increased significantly ($p < 0.05$) to 0.857 ± 0.023. An average of 6.62 features was selected. The most consistently selected features were: perimeter, a profile feature (PF2), the skewness of the gradient magnitude values of the surface voxels, two radii features (RA3 and RA4), and two RLS texture features. Four of these features were from the new space. These results are summarized in Table 5.1, and the features that were selected the most times are listed with the total number of times they were selected in the inner leave-one-case-out loop.

When the $Feat_{MTGD}$ space that included the demographic information was used, the average training A_z was 0.892 ± 0.020, and the two-loop test A_z was 0.863 ± 0.022, with an average of 7.50 features selected. The consistently selected features were the same as those when the feature space was $Feat_{MTG}$, with the addition of the patient age. However, the improvement compared to the $Feat_{MTG}$ feature space did not achieve statistical significance ($p = 0.585$). The ROC curves are compared in Fig. 5.5.

Fig. 5.6 shows nodules in which the CAD system performed poorly. The benign nodules that obtained higher malignancy scores were generally larger and not spherical in shape. Some of the nodules were juxta-vascular, emphasizing the need for a more effective vessel-removal method than that used in this study. The malignant nodules that had low malignancy scores were mostly metastatic, with round shapes and smooth, distinct edges. The texture around these nodules was also more homogeneous.

5.4.2 Classification Performance on Primary and Metastatic Nodules

5.4.2.1 Primary Cancers

For classification of primary cancers and benign nodules, the CAD system achieved an average training A_z of 0.895 ± 0.022 and a two-loop test A_z of 0.857 ± 0.026 in the $Feat_{MT}$ feature space. An average of 5.92 features was selected. The six most consistently selected features were minimum CT number and five RLS texture features. When feature selection was performed in the $Feat_{MTG}$ feature space, the average training A_z was 0.902 ± 0.021, and the two-loop test A_z increased to 0.892 ± 0.022, although the improvement fell short of statistical significance ($p = 0.07$). An average of 4.04 features was selected. The most consistently-selected features included one gradient profile feature (PF4), one radii feature (RA4), and two RLS texture features. Two of these features were selected from the new space. When the demographic information were added, the average training A_z was 0.921 ± 0.019, and the two-loop test A_z was 0.900 ± 0.022 for $Feat_{MTGD}$. The improvement compared to $Feat_{MTG}$ feature space again did not achieve statistical significance ($p = 0.7$). An average of 5.01 features was selected. The most consistently selected features were the same as those when the feature space was $Feat_{MTG}$, with the addition of the patient age. These results are summarized in Table 5.1.

5.4.2.2 Metastatic Cancers

On the subset containing metastatic cancers and benign nodules, the CAD system achieved an average training A_z of 0.855 ± 0.027 and two-loop test A_z of 0.822 ± 0.031 when $Feat_{MT}$ was used as the feature space. An average of 2.96 features was selected, with the largest perimeter and two RLS texture features as the three most consistently selected features. When $Feat_{MTG}$ was used as the feature space, the average training A_z was 0.890 ± 0.024, and the two-loop test A_z decreased to 0.803 ± 0.034, though the decrease was not significant (p = 0.45). An average of 6.69 features was selected. Among the features most consistently selected were two texture features and five from the new feature space including three radii features (RA1, RA3, and RA5), the average gradient magnitude of surface voxels, and one gradient profile feature (PF2). In the $Feat_{MTGD}$ feature space, no demographic features were selected, and the performance was the same as that in the $Feat_{MTG}$ feature space. These results are summarized in Table 5.1.

5.4.3 LDA and SVM Comparison

The performance comparison between the test A_z values of the LDA and the SVM classifiers is shown in Fig. 5.7. PCA was applied to the $Feat_{MTG}$ feature space, and the same number of features from PCA was input into each classifier. For a given number of chosen features, a set of 120 different combinations of kernels and parameters for the SVM were studied. The highest test A_z for the SVM for a given number of selected features is shown. The SVM performance using the radial kernel with $\gamma = 0.02$ (Eqn. (5.7)) and $Cap = 1$ is also shown in Fig. 5.7 to demonstrate SVM performance with a fixed kernel and fixed parameters. This SVM was chosen as an example because it provided the best performance among the SVMs evaluated the most times. The classification performance of this SVM was slightly higher or lower than that of the LDA when the number of PCA features was less than 10, and was consistently lower when the number of PCA features increased to greater than 10. The highest test A_z among all SVMs studied was generally higher than the test A_z from LDA except at r=1, but it was still within one standard deviation of the test A_z from LDA. None of the SVM architectures used in our study provided a consistently better performance than the LDA over the range of the number of PCA features investigated (r=1 to 15).

5.5 Discussion

The newly designed features utilized the gradient field to determine whether the nodule edge is distinct or fuzzy. We also designed features that analyzed statistics of the radii segments of a nodule to quantify surface irregularities and size. The profile features examine a shell of voxels on either side of the segmented boundary, and these features are robust to contours that may be close to but not on the nodule boundary. Nevertheless, the segmented boundaries using the 3DAC are reasonable, as evaluated in our previous study [1].

Previous simulation studies using LDA with SFS performed by our group found that increasing the dimensionality of the feature space resulted in more pessimistic hold-out performance estimate [97]. Based on these results, we would expect that adding features that have only small incremental discriminatory power would degrade the classification performance. However, a few of the new gradient field and radii features were selected, and their inclusion significantly ($p<0.05$) improved the test A_z. This demonstrates that the newly-designed features are beneficial in discriminating between malignant and benign nodules when used in conjunction with the other types of features used in this study.

Effective features are important due to the inherent variability in lung nodule appearance. Previous studies that investigated the performance of CAD systems in classifying nodules show that no single feature can perfectly distinguish malignant from benign nodules [1, 47, 48 Armato, 2003 #1469, 50, 52, 53]. Nodule shape, size, margin, and presence of calcifications or fat are major features that are useful, but far from being perfectly accurate in lung nodule characterization. There is substantial overlap in the appearance of malignant and benign nodules [60, 77, 125], which may be one reason that as many as 50% of nodules resected at surgery are benign [12].

The classifier designed to distinguish primary cancers from benign nodules had a higher performance, whereas the one designed to distinguish metastatic cancers from benign nodules had a lower performance compared to the classifier designed to distinguish all cancers (primary and metastatic) from benign nodules. This may be due to the different characteristics that are unique to primary and metastatic cancers. Primary lung cancers tend to be more spiculated and irregular whereas metastatic cancers tend to be rounder with well-defined borders, which are more similar to benign nodules. It is therefore difficult to design features that can distinguish both primary and metastatic cancers from benign nodules. From a screening perspective, radiologists may be more interested in using CAD to detect and classify primary cancers, since they may already be alerted to the possibility of metastatic cancer if the patient has a history of cancer elsewhere in the body. The performance of the CAD system on primary lung cancers may be a more informative indicator of its potential usefulness.

Radiologists use a variety of factors in arriving at a diagnosis, including a patient's gender, age, and smoking history. We investigated the effect of two demographic features, gender and age, at the time of the scan. Other features such as smoking history or presence of other diseases were either incomplete from the patient records or difficult to quantify. Gender was never selected as a feature, but age as a feature improved the accuracy of the CAD system for the entire data set and for the subset of primary cancers and benign nodules, although the improvement did not achieve statistical significance. Demographic and clinical information may not always be available or reported accurately, especially if large-scale screening with CT is performed. Using the objective image data to design a CAD system is more flexible in that the radiologist can use the assessment by the CAD system as a complement to the other clinical information, if available, in the decision-making process. This study showed that although some demographic information is beneficial in diagnosis, the CAD system would perform similarly without the non-image features we investigated.

Currently, researchers are not able to compare the performance of their CAD systems because of the lack of a common test set. If a large test data set with proven diagnoses is available, it will be a useful resource to compare the effectiveness of different approaches to classification of malignant and benign nodules. A publicly available data set would also increase the number of training samples that CAD developers may use for design of their CAD systems.

Because of the relatively small data sets available, we designed the two-loop leave-one-case-out scheme for feature selection and training the classifier weights. A one-loop leave-one-case-out resampling method is sometimes used for the design of an LDA classifier with SFS. In each cycle of the one-loop leave-one-case-out, SFS and LDA classifier weights are determined using N-1 cases and tested on the left-out case. The SFS parameters F_{in}, F_{out}, and *tol* are chosen based on the test A_z from the N test cases. It is desirable to optimize the classifier with respect to SFS parameters because they influence the number of features selected and thus the performance of the classifier. However, the use of the performance of the classifier designed with the one-loop leave-one-case-out resampling for this optimization will introduce an optimistic bias, because the test cases are being used in the optimization process. In other words, in such an optimization scheme, the test A_z is not independent of training. In the two-loop leave-one-case-out resampling process, optimization of the SFS parameters is performed only within the (N-1) training cases in the inner leave-one-case-out cycle. The left-out case in the outer loop is not used either to design the stepwise LDA or to guide the selection of the SFS parameters, so that the test A_z may not be as optimistically biased. However, since we used the same data set to iteratively improve the CAD system, our CAD system may still have been overtrained to suit the characteristics of the nodule samples in this small data set. Further evaluation of its generalizability is needed when an independent test set is available in the future.

We compared the performance between the LDA and SVM classifiers. Because we were only interested in the relative performance between the two, we performed PCA on the extracted $Feat_{MTG}$ features of the entire data set first, and then varied the selected number of features as input to the classifiers based on the highest eigenvalues of the covariance matrix of the features. PCA was used because it is a filter feature selection method such that it does not select features based on the performance of a specific classifier. This is opposed to a wrapper method such as SFS, which selects features guided by the performance of a classifier using those features. The SVM performed slightly better than the LDA when the highest performance was chosen among a large number of combinations of kernels and set of parameters for a given set of input PCA features. This indicated that, for our data set, if the SVM was tuned for a specific set of input features, it could achieve better performance than the LDA. However, none of the SVMs with a fixed kernel and fixed parameters performed consistently better than the LDA for the combinations of kernels and parameters that we investigated.

A CAD system will only be considered useful if radiologists show improvement in diagnostic accuracy when they use the system as a second reader. The effect of our CAD system on radiologists' classification of lung nodules will be investigated in an observer study. To that end, it is important to continually improve the CAD system to provide radiologists with accurate diagnostic information. Future work will also include analyzing interval change information for classification of malignant and benign nodules

[20, 21] and building on our previous work [96, 97] in investigating the effect of sample size on feature selection and classification.

5.6 Conclusion

In this study we designed new image features by analysis of the gradient field and the surface smoothness of the nodules. We have demonstrated that the new features could improve the performance of our CAD system. The test A_z for the entire data set was improved significantly ($p<0.05$) when feature selection was performed in the entire feature space that included the new features in addition to the morphological and texture features. The discrimination of the CAD system between primary lung cancers and benign nodules was higher than that between metastatic cancers and benign nodules, likely because there is a larger overlap between the appearance of benign nodules and metastatic cancers. When the LDA and SVM classifiers used the same feature set obtained by PCA, and the number of features was varied between 1 and 15 by changing the number of selected principal components, our comparison indicated that no single SVM classifier resulted in a consistently higher performance than the LDA in our classification task. Further work is underway to evaluate the usefulness of the CAD system in assisting radiologists in the classification of malignant and benign lung nodules.

5.7 Tables

	Entire Data Set		
	Feat$_{MT}$	**Feat$_{MTG}$**	**Feat$_{MTGD}$**
Two-loop Training Az	0.858 ± 0.023	0.881 ± 0.021	0.892 ± 0.020
Two-loop test Az	0.821 ± 0.026	0.857 ± 0.023	0.863 ± 0.022
Avg # Feat Sel	5.80	6.62	7.50
Feature Name	***# Times Feature Selected***		
Surface area	122		
Max CT	51		
Variance of gray-levels	100		
LR low GL, obl, x, 90	151		
LR high GL, obl, x, 90	138	86	73
LR high GL, obl, y, 90	144	152	152
Perimeter		152	152
PF2		151	152
Skewness of grad mag		152	152
RA3		150	152
RA4		152	152
Age			152

(a)

Table 5.1: Two-loop test Az for (a) entire data set, (b) primary and benign subset, and (c) metastatic and benign subset. The average number of features selected over all inner loop leave-one-case-out cycles and most frequently selected features and their frequency being selected are also shown for the different data subsets and feature spaces. The features that were consistently selected can be considered the effective features for this classification task. For the RLS texture features, SR = short range, LR = long range, GL = gray-level, horiz = the axial plane, obl = the oblique plane, x and y specify which direction Sobel filtering was performed, and 0 or 90 indicates the direction the run-length statistics features were acquired. More details on these features were described in our previous study [1].

	Primary Cancers and Benign Nodules		
	Feat$_{MT}$	**Feat$_{MTG}$**	**Feat$_{MTGD}$**
Two-loop Training Az	0.895 ± 0.022	0.902 ± 0.021	0.921 ± 0.019
Two-loop test Az	0.857 ± 0.026	0.892 ± 0.022	0.900 ± 0.022
Avg # Feat Sel	5.92	4.04	5.01
Feature Name	***# Times Feature Selected***		
LR low GL, y, 0	118		
LR low GL, y, 90	112		
LR, horiz, y, 0	126		
SR, obl, y, 90	105	126	123
GL nonuniformity, x, 0	124		
Min CT	125		
Run length nonuniformity		126	124
RA4		126	126
PF4		125	125
Age			126

(b)

	Metastatic Cancers and Benign Nodules		
	Feat$_{MT}$	**Feat$_{MTG}$**	**Feat$_{MTGD}$**
Two-loop Training Az	0.855 ± 0.027	0.890 ± 0.024	0.890 ± 0.024
Two-loop test Az	0.822 ± 0.031	0.803 ± 0.034	0.803 ± 0.034
Avg # Feat Sel	2.96	6.69	6.69
Feature Name	***# Times Feature Selected***		
Perimeter	100		
LR high, horiz, x, 90	104	67	67
LR high, obl, x, 90	99	83	83
RA1		104	104
RA5		104	104
RA3		103	103
Mean of all surface vox grad		96	96
PF2		58	58

(c)

5.8 Figures

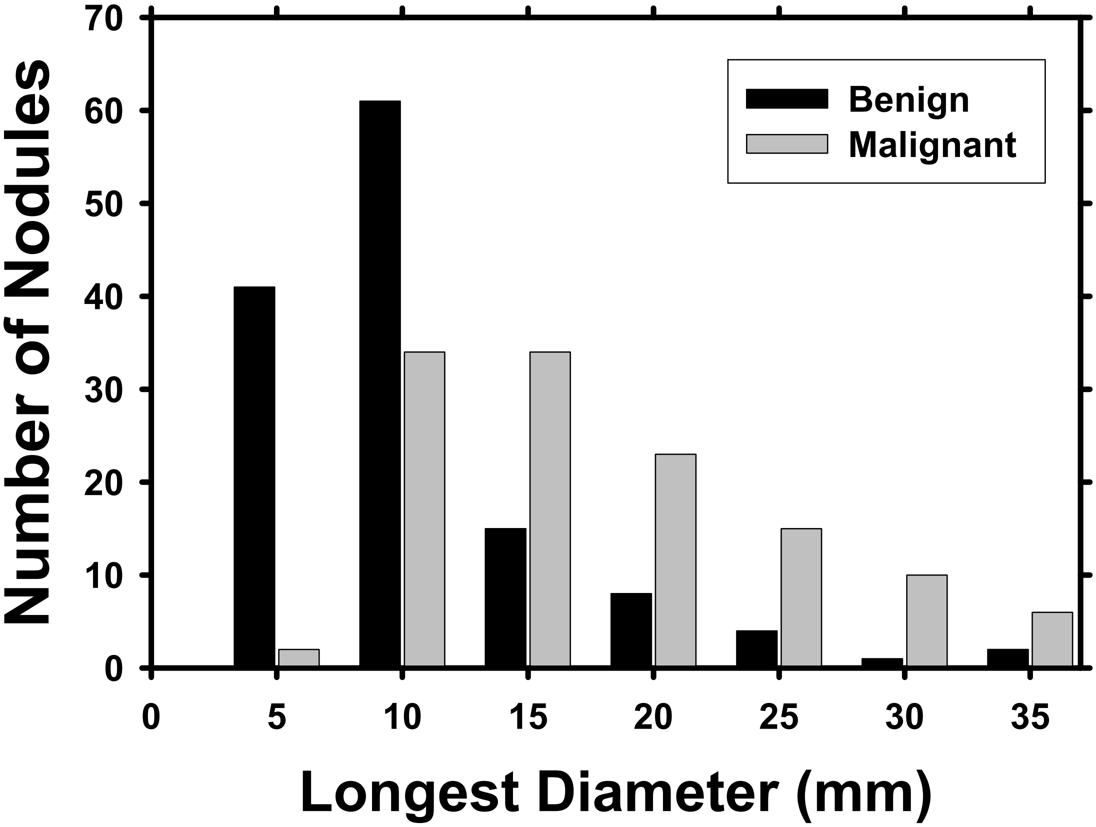


Figure 5.1: Histograms of the longest diameters of the benign and malignant nodules as measured by experienced chest radiologists.

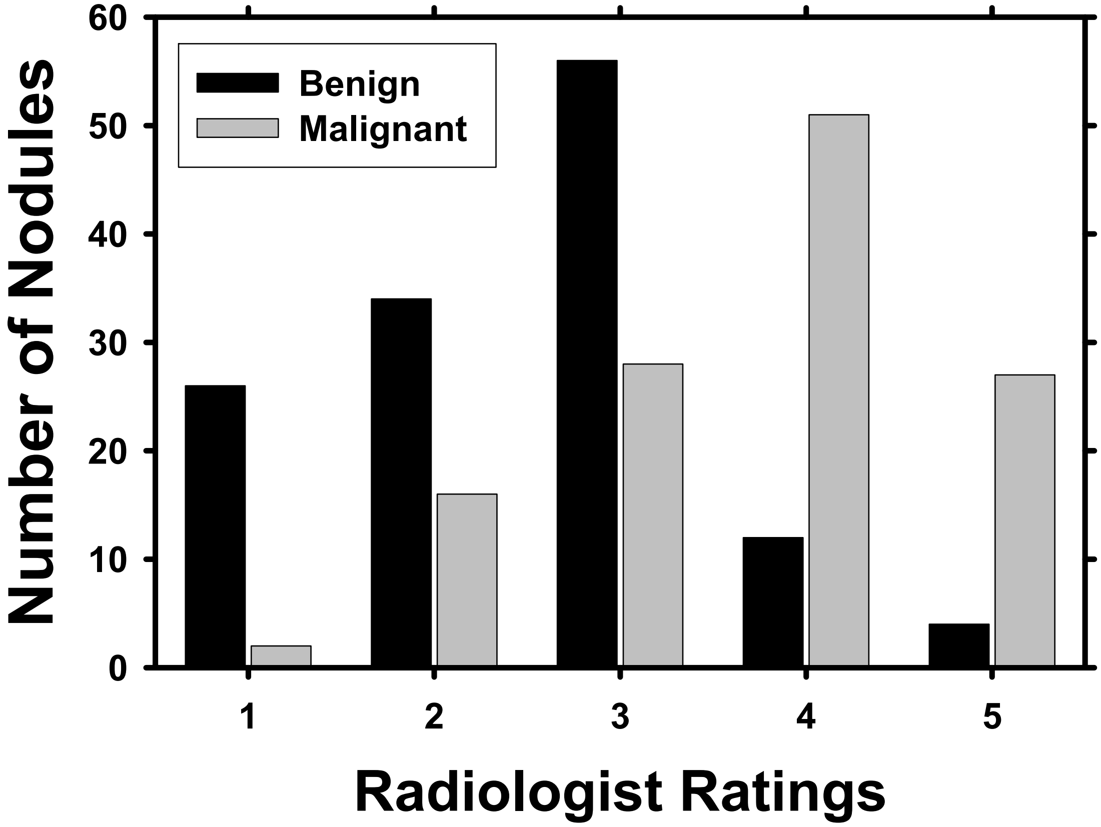


Figure 5.2: Malignancy ratings provided by radiologists on a scale of 1-5, with 5 being most likely malignant. The area, A_z , under the ROC curve fitted to the radiologists' malignancy ratings is 0.806±0.028.

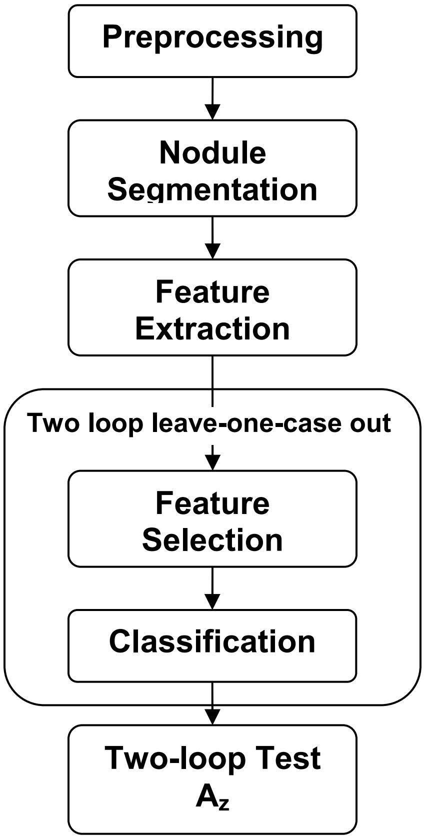


Figure 5.3: A schematic showing the major image processing steps of the CAD system. The two-loop resampling scheme is described in Fig. 5.4.

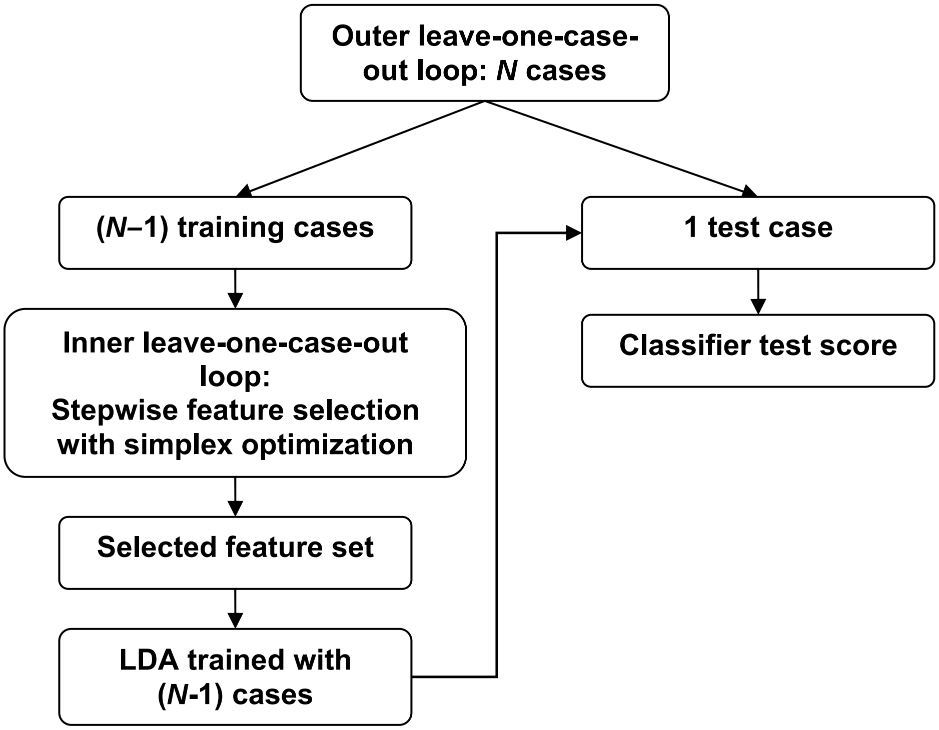

Outer leave-one-case-out loop: *N* cases
(*N*–1) training cases
1 test case
Inner leave-one-case-out loop: Stepwise feature selection with simplex optimization
Classifier test score
Selected feature set
LDA trained with (*N*-1) cases

Figure 5.4: In the outer leave-one-case-out loop, the data set is divided into (*N*-1) training cases and 1 test case. For each (*N*-1) case cycle, an LDA classifier is designed from a set of selected features as a result of an inner leave-one-case-out training and testing scheme. After each case is left-out in turn, the two-loop test A_z is calculated from the malignancy scores of *N* test cases.

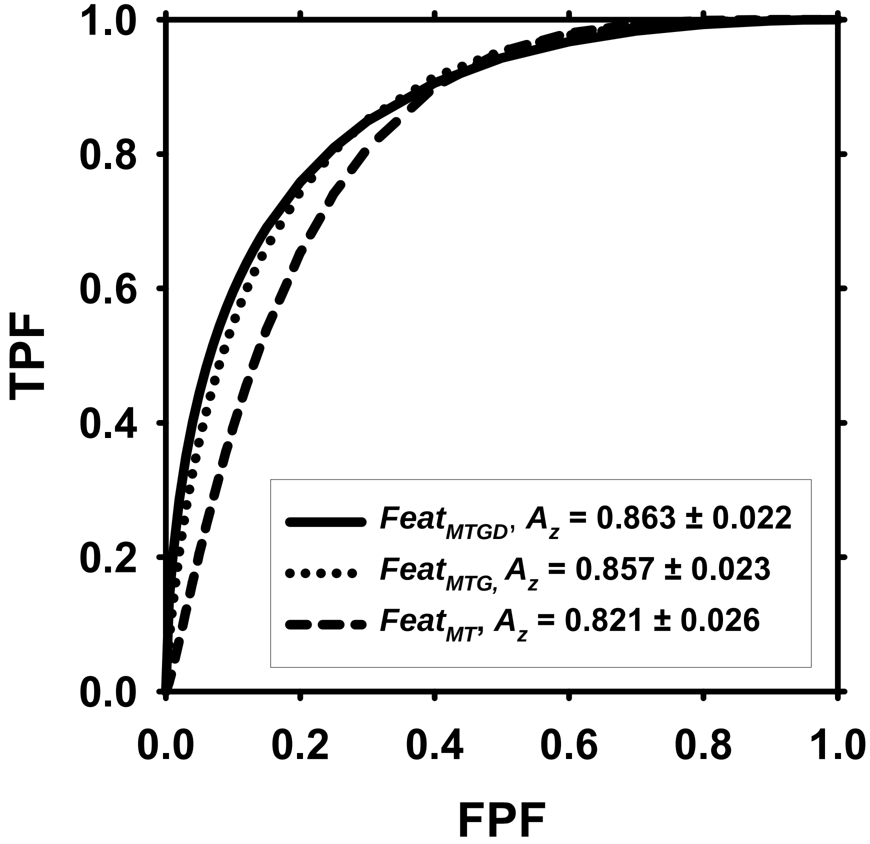
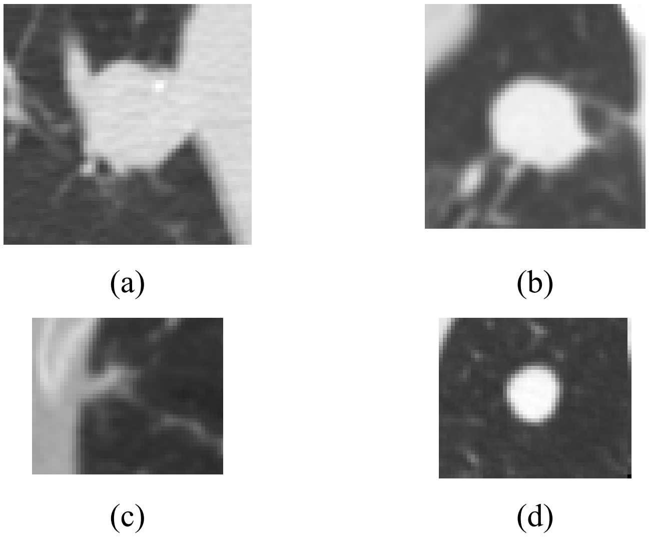


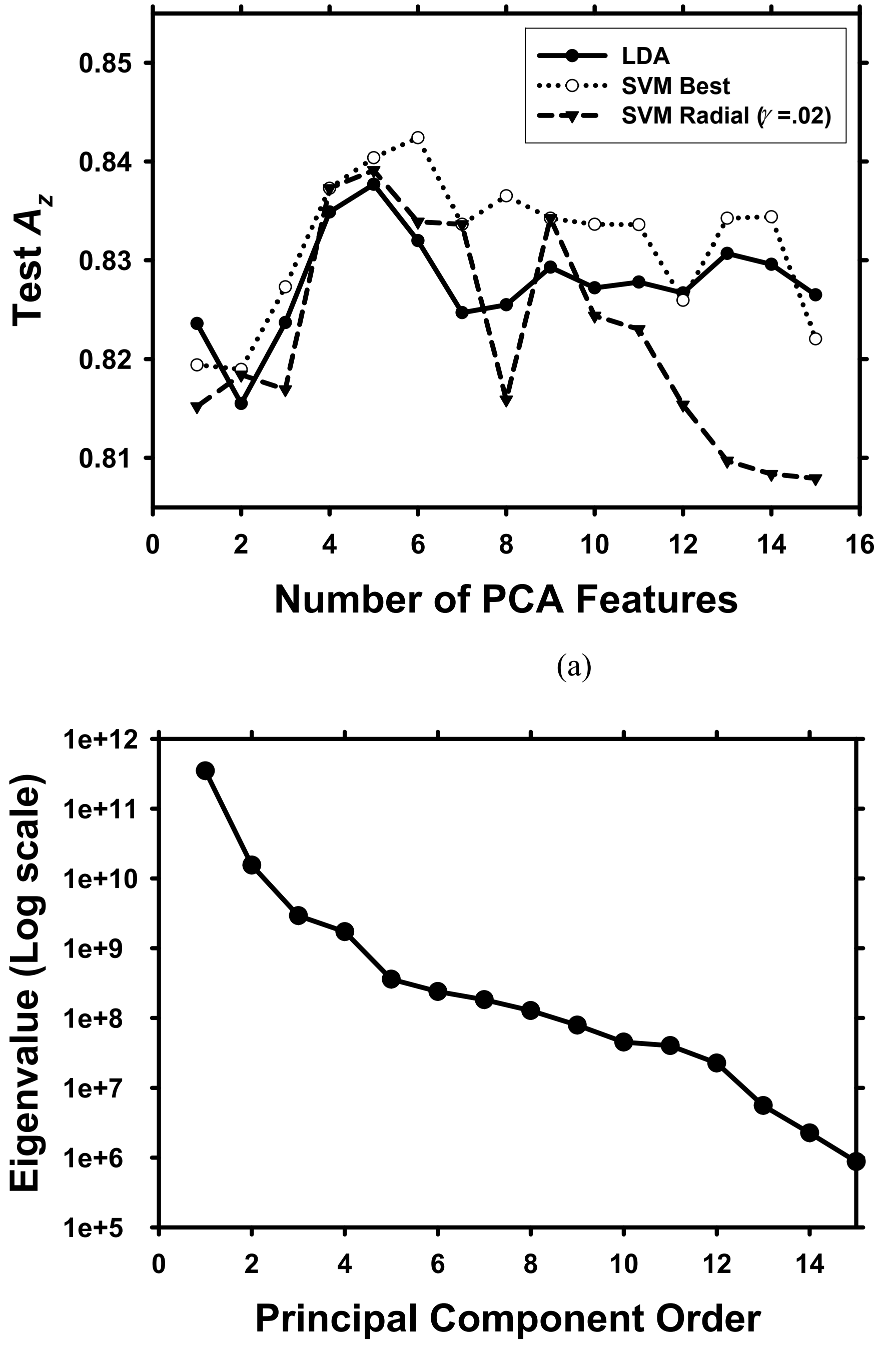
Figure 5.5: ROC curves for the performance of the CAD system based on the two-loop test scores. The two-loop test A_z using features selected from the $Feat_{MTG}$ space was 0.857 ± 0.023, which was significantly higher ($p < 0.05$) than the two-loop test A_z of 0.821 ± 0.026 when features were selected only from the $Feat_{MT}$ space. The addition of demographic information improved the two-loop test A_z to 0.863 ± 0.022, but the difference did not achieve statistical significance ($p = 0.585$).



(a) (b)

(c) (d)

Figure 5.6: Examples of nodules for which the CAD system performed poorly. (a) A large benign nodule that was unchanged over two years, (b) biopsy-proven non-necrotizing benign granuloma, (c) adenocarcinoma that may have been too small for the extraction of useful texture information, (d) metastatic adenoid cystic carcinoma with features that may overlap with many benign nodules, e.g., round shape and distinct boundaries.



(a)

(b)

Figure 5.7: (a) Comparison of test A_z between LDA and SVM. For a given number of selected features, 120 combinations of parameters and kernels were evaluated for the SVM, and the best test A_z is shown. The standard deviations of the test A_z ranged from 0.024 to 0.026 for both classifiers. The test A_z using the radial kernel is also shown as an example of the performance when the kernel and parameters are fixed. (b) The sorted eigenvalues of the covariance matrix of the $Feat_{MTG}$ feature space obtained from PCA.

Chapter 6
Computer-Aided Diagnosis of Lung Nodules on CT Scans: An Observer Study of its Effect on Radiologists' Performance

6.1 Abstract

The purpose of this study is to evaluate the effect of computer-aided diagnosis (CAD) on radiologists' estimates of the likelihood of malignancy of lung nodules on CT. We retrospectively collected 256 lung nodules (124 malignant, 132 benign) from the thoracic CT scans of 152 patients with IRB approval. We developed an automated CAD system to characterize and provide a malignancy rating for lung nodules on CT volumetric images. An observer study was conducted with receiver operating characteristic (ROC) methodology to evaluate the effect of CAD on radiologists' characterization of lung nodules. Six fellowship-trained thoracic radiologists served as readers. The reading order of the individual nodules was randomized differently for each reader. The readers rated the likelihood of malignancy on a scale of 0-100% and recommended appropriate action first without CAD and then with CAD. The observer ratings were analyzed with the Dorfman-Berbaum-Metz multi-reader, multi-case method.

The CAD system achieved a test A_z of 0.857±0.023 using the perimeter, two nodule radii measures, two texture features, and two gradient field features. All 6 radiologists obtained improved performance with CAD, with three reaching statistical significance ($p<0.05$). The average A_z of the radiologists improved significantly ($p = 0.006$) from 0.833 (range: 0.817 to 0.847) to 0.853 (range: 0.834 to 0.887). We conclude that CAD has the potential to increase radiologists' accuracy in assessing the likelihood of malignancy of lung nodules on CT.

6.2 Introduction

Lung cancer is the leading cause of cancer death in the United States, causing an estimated 160,400 deaths in 2007. Despite advances in treatment and diagnosis, the five-year overall survival rate is only 15% [3]. Currently, there is no generally accepted or recommended screening method for lung cancer that has been proven to reduce patient mortality. As a result, patients typically present with clinically advanced stages of disease at diagnosis.

One area of active research in lung cancer screening is the use of computed tomography (CT), which has been shown to be more sensitive to lung nodule detection than chest X-ray (CXR), especially for smaller nodules [6, 8, 9, 13]. Henschke *et al.* [126] reported a 92% survival rate among patients who underwent surgical resection for detected Stage I lung cancers. Sobue *et al.* [127] reported an almost 100% five-year survival rate for patients with nodules less than 9 mm. These data suggest the benefits of earlier intervention with early detection. It is expected that the National Lung Screening Trial, which is a randomized, controlled study of over 50,000 enrolled patients, will provide more definitive results as to whether early detection with CT compared to CXR will lead to reduced patient mortality.

The higher sensitivity of CT results in an increase in the number of nodules detected, and thus, an increase in the nodules that need to be followed-up and managed. This may require expensive diagnostic tests such as follow-up CT scans and biopsy. Multidetector row CT technology has resulted in thinner slices and higher resolution. However, the large number of images that radiologists have to interpret greatly increases their workload. Despite higher-quality images, Swensen *et al.* reported as many as 50% of nodules resected at surgery are benign [12], signifying the difficulty radiologists have in determining whether a lung nodule is malignant or not by CT and other clinical information. This emphasizes the importance of providing radiologists with tools to better characterize nodules and determine the appropriate course of action.

Computer-aided detection and diagnosis software is being developed to address these issues. Computer-aided detection has been shown to increase the sensitivity of lung nodule detection [25, 42, 128-131]. Computer-aided diagnosis (CAD) methods for classification of lung nodules as malignant or benign have been reported by a number of investigators, with the area under the receiver operator characteristic (ROC) curve, A_z, ranging from 0.79 to 0.92 [1, 47-53, 132]. A few observer studies have been performed to evaluate the effects of CAD on radiologists' assessment of the malignancy of lung nodules on CT. Matsuki *et al.* [133] performed a study with 4 radiologists, 4 fellows and 4 residents reading a data set of 25 malignant and 25 benign nodules. They found that the accuracy of each group of observers improved significantly with CAD and the difference in performance among the three groups were reduced. Shah *et al.* [134] conducted a study with eight radiologists reading 28 nodules (15 malignant and 13 benign) and obtained a significant improvement in the average A_z from 0.75 to 0.81 ($p = 0.018$) when a computer aid was used. Li *et al.* [51] found that the average A_z for 16 radiologists significantly increased from 0.785 to 0.853 ($p = 0.016$). In addition, they [135] observed that CAD had a beneficial effect on 68% of their changed recommendations for a data set of 28 malignant and 28 benign nodules. Awai *et al.* [136] reported a significant ($p = 0.021$) improvement for 19 observers from an average A_z of 0.843 ± 0.097 to 0.924 ± 0.043 for 18 malignant and 15 benign nodules. A subgroup analysis showed that the nine radiology residents improved significantly as a group, but the improvement of 10 board-certified radiologists did not achieve statistical significance.

Although the previous studies have demonstrated a trend of improvement in radiologists' classification accuracy with CAD, the data sets in those studies were small. In this study, we collected a relatively large data set of 256 (132 benign and 124 malignant) nodules to evaluate the effect of our CAD system on radiologists' estimates of the malignancy of lung nodules. To make it more challenging for the CAD system to demonstrate a beneficial effect, fellowship-trained experienced thoracic radiologists were recruited as observers. Our data set included both primary and metastatic lung cancers, and we analyzed the assessment of the two groups collectively and separately. This will reveal the effect of CAD in an environment with a heterogeneous case mix in comparison to that of a homogeneous data set.

6.3 Methods and Materials

6.3.1 Collection of CT Studies

We retrospectively collected CT scans from the patient archive in the Department of Radiology with IRB approval. The CT studies were acquired in our clinic with a variety of GE scanners (GE, Waukesha, WI), including the Genesis HiSpeed scanners and the GE LightSpeed series scanner models Plus, Power, Pro 16, QX/i, Ultra, and LightSpeed16. The pixel size ranged from 0.448 to 0.859 mm (with corresponding fields-of-view of 25 to 44 cm). The slice thickness averaged 2.3 ± 1.44 mm (range: 1 to 7.5 mm), and the slice interval averaged 2.0 ± 1.6 mm (range: 0.6 to 7.5 mm). The tube voltage averaged 120 ± 1.8 kVp (range:120 to 140 kVp), tube current averaged 209 ± 92 mA (range: 80 to 500 mA), and mAs averaged 214 ± 141 mAs (range: 40 to 570 mAs).

6.3.2 Nodule Selection

For each patient scan, an expert thoracic radiologist marked locations of nodules by placing a box encompassing the nodule using a graphical user interface developed in our laboratory. This radiologist did not participate as an observer. The nodule inclusion criteria for this study were: (1) diameter greater than 3 mm as measured by the radiologist, (2) appearance of the nodule on at least three slices, and (3) proven diagnosis through biopsy, other known metastatic disease, or two-year follow up.

We collected 256 nodules from the CT scans of 152 patients. There were 132 benign and 124 malignant nodules. Seventy-two of the malignant nodules were primary and 52 were metastatic cancers. There were 218 solid nodules, 17 ground glass opacity (GGO), and 21 mixed attenuation types. Of the 124 malignant nodules, 64 was established by biopsy, and 60 were determined to be malignant either through positive PET scans, being in the same lung as other biopsy-proven malignant nodules, or known metastases from confirmed cancers in other body parts based on the patients' clinical reports. Of the 132 benign nodules, 15 were biopsy-proven and 117 were determined to be benign by two-year follow-up stability on CT. Of the 256 nodules, 53 were juxta-pleural and 19 were juxta-vascular. A distribution of the longest diameters of the nodules measured by radiologists is shown in Fig. 6.1. The nodules had an average longest diameter of 11.7 ± 7.7 mm (range: 3.0 -37.5 mm). Eight of the nodules in our data set had longest diameters greater than 30 mm but less than 38 mm. Although nodules are generally defined as less than 30 mm in diameter, we included these masses in our data set because their edges and surrounding texture may contribute to the training of the CAD system. Although they seem highly suspicious, one was proven benign by biopsy. Test results with and without these 8 masses were compared.

6.3.3 CAD System

Our CAD system is summarized as follows, while further details can be found in the literature[1]. First, a volume-of-interest (VOI) containing the nodule was extracted based on the box placed by the expert radiologist. Since our CAD system was designed to classify whether a nodule was malignant or benign, the input to the system was assumed to be a VOI that contained a lung nodule. If this system is combined with an automated nodule detection system in future developments, improvement in the classification method to include false-positive nodules will be needed. The system performed linear interpolation in the *z*-direction to reduce the *z*-dimension of the voxel to that of the axial plane if the slice interval was greater than the pixel size, or bilinear interpolation in the axial plane to reduce the pixel size to that of the slice interval if pixel size was greater than the slice interval, to obtain isotropic voxel dimensions. The isotropic voxels facilitated the implementation of the 3D active contour (3DAC) segmentation and feature extraction operations in the CAD system. The interpolation to smaller voxel dimensions did not increase the spatial resolution of the image data.

To generate the initial contour, *k*-means clustering was used to assign voxels in the VOI to the object or the background class based on voxel intensity. Morphological opening was then performed with a spherical kernel that had an automatically calculated kernel size based on the size of the clustered object. The purpose of the morphological opening operation was to remove blood vessels that might be attached to the nodule. The nodule in the VOI was then segmented using the 3DAC method. Our approach to the determination of the weights for the various energies in the 3DAC model was described previously [1].

Morphological features including volume, largest perimeter, and statistics based on the CT values (Hounsfeld Units) inside the nodule were then extracted from the segmented nodule. To quantify tissue texture around the nodule, the Rubber Band Straightening Transform (RBST) [64] was used to convert a 15-voxel-wide band surrounding the nodule on each slice into a rectangular image. In the RBST image, the nodule boundary was transformed to the horizontal direction and the spiculations emanating radially from the nodule were oriented approximately in the vertical direction. After performing Sobel filtering on the RBST images, run-length statistics (RLS) features[65, 66] were extracted. In addition, gradient field features[93] were extracted from the gradient magnitude value at each voxel. The statistics of the gradient magnitudes of all surface voxels and along the rays tracing from the nodule centroid to the surface voxels were used to describe the smoothness of the nodule surface.

We used a "two-loop" leave-one-case-out resampling method to train and test the CAD system using the *n* available cases. In each cycle of the outer leave-one-case-out loop, we reserved one case, including all nodules from this case, as the independent test case. The remaining (*n*-1) cases were used to train the classifier in a process that included feature selection and classifier weight determination. A subset of most effective features was selected by stepwise feature selection. An "inner" leave-one-case-out scheme was performed within the (*n*-1) training cases to determine the best thresholds for stepwise feature selection. These thresholds were F_{in} and F_{out} for deciding whether a feature should be included or removed from the feature space, respectively, and the *tol* threshold for setting the tolerance on the correlation of the selected features. In each cycle of this inner leave-one-case-out scheme for feature selection, (*n*-2) cases were available for training while 1 case was left out as the test case. The best set of F_{in}, F_{out} and *tol* thresholds was searched by simplex optimization using the test A_z of the (*n*-1) left-out cases from the inner loop as a guide. After the feature selection thresholds were determined, a set of features was selected from the (*n*-1) cases and a linear discriminant analysis (LDA) classifier with proper weights for the features was built. This classifier was then applied to all nodules of the original independent left-out case and a test score for each nodule was obtained. This procedure was cycled through the *n* cases of the entire data set in the outer loop, so that each case was left-out in turn, resulting in test scores for all the nodules in the data set.

A histogram with 10 bins was generated from the test scores of the entire data set. Each bin was further separated into benign and malignant classes. For each class, a Gaussian curve was fitted (SigmaPlot 9.0, SysStat; San Jose, CA, USA). Both curves were normalized so that the area under each curve was unity. The two fitted Gaussian curves represented the probability density functions of the malignancy ratings for test lung nodules estimated by the CAD system. The original bin value was mapped as the malignancy rating on a scale of 1 to 10. The Gaussian curves with the malignancy rating scale are shown in Fig. 6.2.

6.3.4 Observer study

We conducted an ROC study with a sequential reading method in which the radiologist was asked to estimate the likelihood of malignancy (LM) of a nodule, view the malignancy rating by the CAD system, and then modify his/her LM estimate if desired. The sequential reading method emulated the use of CAD as a second opinion, in which the radiologist first made his/her own judgment without CAD and then made a refined decision after taking the malignancy rating of the CAD system into consideration. Six fellowship-trained thoracic radiologists participated as observers.

The reading order of the nodules was “randomized” for each reader such that, on average, no nodule would be read more often in a certain order in the reading sessions than the other nodules. In addition, different nodules from the same case were separated by a number of nodules from other cases. The purpose of the randomization was to minimize the effects of fatigue, learning, memorization, and nodule correlation on the results of observer performance [137].

We developed a graphical user interface (GUI) (Fig. 6.3) to display the CT scan and record the observer ratings in this study. For a given nodule to be read, the entire CT scan was loaded but the slice in which the nodule appeared to be the largest was shown first. The nodule was enclosed in a box, previously marked by an expert thoracic radiologist who did not participate as a reader. The observer was free to scroll through the available slices from the scan, but he/she was instructed to focus on the visual characteristics of the nodule of interest. No clinical or demographic information about the patient was provided. The original reconstructed CT slices without interpolation were shown and the slice interval was displayed. Readers were free to adjust the brightness and contrast of the image, and a zoom function was available. A 3 mm x 3 mm box displayed in the upper left corner of the image served as a size reference. A rendered volume for each nodule was available should the observer choose to look at its surface characteristics.

Each radiologist was asked to rate the LM of the marked nodule on a scale of 0 to 100% and provide the recommended action (no action; CT follow-up; or immediate action, such as biopsy, PET, or surgery). In addition, he/she marked the presence of cavitation, calcification, nodule edge (smooth, lobulated, or spiculated/irregular), and attenuation type (solid, GGO, or mixed).

The GUI prevented the reader from viewing the rating of the CAD system until the assessments listed above were completed. The classifier result was presented as an integer rating on a scale of 1 to 10, as described in the previous section and shown in Fig. 6.2. The probability density distributions of malignant and benign classes as estimated by the CAD system were shown on the GUI to provide a reference for the observer. After viewing the CAD system rating, the radiologist had the option of adjusting the LM estimate of the nodule and the recommended action.

Each observer underwent a training session with nodules not part of the data set to become familiar with the GUI and the experimental process before the actual reading session would start. We instructed the observers to utilize the entire range of the rating scale and to interpret the CAD system rating by reference to the two-class distributions of the classifier. The radiologists were informed of the total number of nodules and the number of patients. They were not told the proportion of malignant and benign nodules, only that the prevalence of malignant nodules was enriched compared to what they would see in clinical practice. No time limit was imposed to assess each nodule.

6.3.5 Statistical Analysis

We analyzed the radiologists' malignancy ratings with ROC methodology. The classification accuracy was quantified by A_z, which was estimated using the Dorfman-

Berbaum-Metz (DBM) method for analysis of multi-reader multi-case data[138]. The *a* and *b* parameters of the individual observers' ROC curves were averaged and these average parameters were used to derive an average ROC curve. We also calculated the partial A_z, $A_z^{0.9}$, which is the area under the ROC curve above a true-positive fraction (TPF) value of 0.9. A larger $A_z^{0.9}$ indicates a higher specificity in the high sensitivity region [70, 139]. The DBM method uses the maximum likelihood estimation of the binormal distributions to fit the observer rating data and provides an estimate of the statistical significance of the difference in the two conditions, without and with CAD, taking into account the multi-reader multi-case readings. In addition, we compared the individual observer's A_z values without and with CAD using Student's two-tailed paired *t*-test.

Because primary cancers and metastatic cancers have somewhat different characteristics and may be distinguished from benign nodules in different ways, we separately analyzed the classification accuracy for two subsets of lung cancers; one subset contained only the primary cancers and the benign nodules, and the other subset contained the metastatic cancers and the benign nodules. In addition, we analyzed the performance when the 8 masses in the data set larger than 30 mm in diameter were excluded to evaluate the classification of lesions that were considered to be nodules (≤30 mm) by radiologists.

Differences in feature descriptors provided by the radiologists for cavitation, calcification, nodule margin, attenuation, and recommended action were analyzed with single-factor analysis of variance (ANOVA). Numerical values were assigned to the responses of each feature. For example, we assigned solid = 1, GGO = 2, and mixed = 3 for attenuation type. We then used ANOVA to determine whether there was a significant difference in the descriptors that radiologists provided.

6.4 Results

The CAD system achieved a leave-one-case-out test A_z of 0.857 ± 0.023 for the 256 nodules and a partial A_z, $A_z^{0.9}$, of 0.476. The selected features were very consistent with only a slight variation among the 152 (total number of cases) cycles in the leave-one-case-out process. Overall, an average of 6.62 features was selected. The most frequently selected features were the nodule perimeter, two radii measures, two texture features, and two gradient field features. In the 152 cycles, these features were each selected between 150 and 152 times, except for one texture feature that was selected in 86 cycles. This shows that the minor variations in the training set did not drastically change the set of features that would be selected.

For the radiologists, the average A_z without CAD was 0.833 (range: 0.817 to 0.847), and it improved significantly to 0.853 (range: 0.834 to 0.877) with CAD ($p < 0.01$). In addition, the $A_z^{0.9}$ improved significantly from 0.390 to 0.456 ($p=0.043$). All radiologists showed improvement in terms of A_z, with three reaching statistical significance ($p < 0.05$). The A_z values for the radiologists are shown in Table 6.1. The differences in scores without CAD and with CAD for the individual radiologists are shown in Fig. 6.4. The *a* and *b* parameters of the individual observers' ROC curves were averaged and these average parameters were used to derive an average ROC curve. The average ROC curves for the radiologists without and with CAD, in addition to the ROC curve of the computer classifier are compared in Fig. 6.5.

Of the 256 nodules, the radiologists modified their LM estimates after the use of CAD an average of 126.0 ± 46.8 times (range: 57 to 192). We define a "correct" LM change when a radiologist increased the LM estimate for a malignant nodule or reduced the LM estimate for a benign nodule with CAD, and vice versa for an "incorrect" change. The radiologists made correct LM changes an average of 95.0 ± 34.0 (range: 37 to 126) times out of the 126 average changes, modifying their estimates by an average of 10.2 ± 2.8 (range: 6.6 to 13.4) points. The radiologists made incorrect LM changes an average of 31.0 ± 18.2 (range: 16 to 66) times, changing their estimates by an average of 10.9 ± 3.6 (range 6.7 to 16.2) points. These changes are summarized in Table 6.2.

The radiologists also changed their recommended actions an average of 10.8 ± 5.8 times (range: 5 to 18). We consider a change to be "correct" for a malignant nodule when the recommended action was changed from "no action" to either "CT follow-up" or "immediate action", or "CT follow-up" was changed to "immediate action"; and vice-versa for "incorrect" change. Correct recommended action changes were made an average of 6.8 ± 2.5 (range: 4 to 10) times, while an average of 4 ± 3.6 (range: 1 to 9) incorrect recommended action changes were made. These changes are also summarized in Table 6.2.

In the relative scale of our CAD system, a classifier score of 5 indicated that the nodule was estimated to be about equally likely malignant or benign, thus providing no indication one way or the other to the nodule's malignancy. We consider a "correct" classifier score as greater than 5 for malignant nodules and less than 5 for benign nodules. The effect of the classifier score on the radiologists, specifically on the subset of nodules for which modifications to the LM were made and the classifier score was not 5, can provide an indication of how useful CAD may be. As shown in Fig. 6.6, there are four possible scenarios. The classifier score could be correct or incorrect, and for each classifier outcome, the radiologist's LM modification could be correct or incorrect. We found that an average of 78.0% of radiologists' modifications was correct and the classifier score was correct, signifying CAD's benefit. There was an average of 13.6% incorrect radiologist modification and incorrect classifier scores, suggesting that radiologists were misled by the CAD system's assessment. On the other hand, 6.4% of the modifications were incorrect although the classifier score was correct, and 2.0% of the modifications were correct despite the incorrect CAD system score.

We analyzed observer performance on two subsets of the data (Table 6.3): (1) primary cancers and benign nodules, and (2) metastatic cancers and benign nodules. For the primary cancer subset, the average A_z of the radiologists improved significantly ($p < 0.01$) from 0.823 (range: 0.805 to 0.837) without CAD to 0.848 (range: 0.823 to 0.866) with CAD. Their average $A_z^{0.9}$ improved significantly from 0.338 to 0.415 ($p=0.045$). For the metastatic cancer subset, the average A_z of the radiologists also improved significantly ($p = 0.01$) from 0.849 (range: 0.813 to 0.877) without CAD to 0.861 (range: 0.834 to 0.895) with CAD. Their average $A_z^{0.9}$ improved significantly from 0.493 to 0.535 ($p=0.01$).

There was considerable inter-observer variability in the nodule feature assessment. The null hypothesis is that all radiologists would give the same feature descriptors for every nodule (e.g., all radiologists would consider one nodule to be solid). Using ANOVA, the null hypothesis was accepted only for the presence of cavitation ($p = 0.71$).

For every other feature, the null hypothesis was rejected ($p < 0.05$), which showed a significant difference among radiologists.

There were 8 masses with diameters greater than 30 mm in the data set. If the test scores of the eight masses were removed, the test A_z of the CAD system was 0.849 ± 0.024. When the scores of the eight masses were removed from each observer's data, the average A_z for the observers still improved significantly ($p = 0.005$) from 0.832 (range 0.813 to 0.853) without CAD to 0.850 (range 0.837 to 0.879) with CAD. The average $A_z^{0.9}$ improved significantly ($p = 0.024$) from 0.392 (range: 0.311 to 0.446) without CAD to 0.455 (range: 0.402 to 0.548) with CAD. These results are also summarized in Table 6.3.

The effects of CAD on radiologist performance are demonstrated by the following examples. Fig. 6.7 is an example of the beneficial influence of CAD. This was a biopsy-proven non-small cell lung cancer. Radiologists gave an average LM of 45.8% (range 5-70%) without CAD, but increased it to 58.3% after seeing the classifier score of 7. Fig. 6.8 is another example of the beneficial influence of CAD. This nodule was determined to be benign after no changes were observed over two years. Radiologists gave an average LM of 53.3% (range 20-65%) without CAD, but reduced it to 48.3% after seeing the classifier score of 4. Fig. 6.9 is an example for which the CAD system gave an incorrect score that did not adversely affect radiologists. This nodule was found to be adenoid cystic carcinoma by biopsy. Radiologists gave an average LM of 57.5% (range 35-90%), but they did not modify their rating substantially after seeing the classifier score of 4, as the average LM with CAD was 55.8% (range 35-90%).

6.5 Discussion

Our results indicate that CAD can benefit radiologists in characterizing lung nodules on CT. The improvement was modest, though significant, possibly because the radiologists participating in the study are experienced fellowship-trained thoracic radiologists. Experience level could influence how beneficial CAD is for a radiologist. Awai[136] has reported that radiology residents showed significant improvement from an average A_z of 0.768 ± 0.078 to 0.901 ± 0.036 ($p = 0.009$), but there was no significant improvement for board-certified radiologists.

As with any other second opinion, the radiologist may or may not concur with the suggestion and may even change for the worse on second thought. The radiologists were cautious in making changes, as the classifier score affected them enough to modify their malignancy assessment for only half the nodules on average. Even when the CAD system assessment did give them more confidence, they changed their scores by an average of only 10 points. This could be attributed to several reasons. First, the observers were all thoracic radiologists experienced in chest CT interpretation. The CAD system test A_z was only comparable to their assessment without CAD. Second, this study was the first experience of using CAD for all observers so that they might not have strong confidence in the CAD system. Finally, the opinion of a CAD system may not be as effective as a second radiologist because it is not interactive. In our implementation, the CAD system only provided a relative malignancy rating between 1 and 10 in reference to the rating distributions for the data set. In a consensus double reading setting, the radiologist can interact with a second radiologist to understand his/her opinion and reasons for arriving at a diagnosis, but this is not possible with CAD. It may be beneficial for a CAD system to provide examples of similar nodules with known diagnosis to justify its decision, as in a content-based image retrieval CAD system [140]. The effectiveness of these two CAD approaches will warrant comparison in future studies.

Our CAD system extracts a variety of features to analyze the nodules. The diversity of selected features, including morphological, texture, and gradient field, indicates that many different characteristics of the nodule contain useful information about its malignancy. In particular, the size, edge, and texture surrounding the nodule have been found to be effective discriminators. This is consistent with the findings by other investigators in characterization of nodules [60]. The advantage of computerized image analysis is that it can extract features such as the skewness of gray level histogram or texture descriptors that may not be readily visualized, thus allowing the computer-extracted information to complement the radiologist's assessment. In clinical practice, radiologists use other patient information such as age, gender, and smoking history to make a diagnosis. We have compared the performances of the CAD system without and with patient age and gender as available features. Age was selected consistently as an input feature but the improvement in the test A_z value was less than 0.01. Because the gain in performance is minimal and the chance of erroneous or missing input information may increase if CAD is recommended for use in a screening setting, we did not include patient information in the CAD system.

We included both primary and metastatic cancers in our data set. Primary cancers were the majority of the malignant nodules (72/124=58%). These are the cancers that would be of main concern, since CAD would be beneficial when primary cancers are correctly characterized in otherwise asymptomatic patients. There was significant improvement (p<0.05) in radiologists' classification of primary cancers with CAD. Metastatic cancers were included in our data set because they would also appear on clinical scans, though radiologists may be more vigilant and suspicious of metastatic diseases in patients with other cancers. Nevertheless, CAD also significantly improved the radiologists' characterization accuracy for metastatic cancers.

One issue that concerns all CAD researchers is the lack of a large training and independent test set. It is difficult to reserve an independent set for testing because of the countering priority to have as many samples as possible for training. Before larger sets from the Lung Image Database Consortium (LIDC) or other sources are available, resampling schemes such as the leave-one-out method may be used to estimate the test performance. Our study used such an approach to produce test scores for the nodule set. The results of the ROC study indicate that, given the use of a CAD system with the level of performance as that in this study, the observers could achieve significantly higher accuracy with CAD than without CAD. This relative improvement demonstrates the benefit of using CAD. Whether or not the malignancy ratings for the nodules were obtained from independent testing should not affect the estimated relative change in observer performance in an ROC study.

Large inter- and intra-observer variabilities have been reported on such tasks as segmentation [81, 83]. We asked the radiologists to rate nodule features in this study. Considerable inter-observer variability was observed among the nodule feature descriptors provided by radiologists. This demonstrates that the radiologists may have perceived differences in whether a nodule had solid or GGO components, for example. Because features such as the margin or the presence of calcification are also useful indicators of the likelihood of malignancy of a nodule [60], when radiologists do not agree on their perception of these features, it may follow that radiologists' assessments of malignancy would be substantially different

CAD systems for detecting lung nodules have been approved by the Food and Drug Administration and are already in clinical use. CAD for characterization may follow some time in the future. Computerized image analysis by the CAD system will likely provide useful information complementary to, but not in place of, the diagnostic and clinical information that the radiologists routinely used for assessment of nodule malignancy. Radiologists should be well informed of the performance of the specific CAD system before implementing it for clinical practice. They should also evaluate the CAD system over time based on their own experience in assessing nodules without and with CAD. Only if they understand the benefits and limitations can they take best advantage of the information provided by the CAD system and use it properly as a second opinion.

There are limitations in our study. The participating readers were fellowship-trained thoracic radiologists and do not accurately reflect the population of radiologists in general. Thus, these results should not be extrapolated to the performance of radiologists as a whole, though the significant improvement of these experienced radiologists is encouraging. We plan to conduct an observer study for radiologists that vary in experience to evaluate the effect of CAD on their diagnostic performance. A second limitation is the fact that this study was a controlled laboratory experiment, where radiologists knew they would be reading cases containing nodules in succession. The prevalence of lung nodules on CT, especially malignant ones, was not reflective of what radiologists would typically see in clinical practice. Since their assessment of the nodule in this retrospective observer study would not affect patient care, there is a possibility that their response to the second opinion by CAD could be different from that in an actual clinical setting.

Thirdly, the low number of changes in recommended action could have been due to the very broad range of options each choice encompassed. For example, three-, six-, and twelve-month follow-up all fall under "CT follow-up," while "immediate action" could be anything from sputum analysis, PET, or surgery. A change of less than 20% in the LM estimates would therefore not make a difference in the recommended action in most cases. Furthermore, even if the CAD system is highly accurate, radiologists will have to develop strong confidence through their experience with the CAD system before they will be willing to change a recommended action in clinical situations because of medicolegal issues. Until CAD is deployed in a real-world situation, it is impossible to know whether it would be truly beneficial.

Finally, we have taken steps to prevent overtraining and bias in our classifier by using the two-loop leave-one-case-out training and testing method so that there is an independent test case for each training cycle. However, we are aware that there is a possibility of inadvertent overtraining by virtue of using the same data set many times to improve our CAD system. We will continue to expand the database from our patient files and also expect to use the LIDC public data set if pathological results of the lung nodules are made available in the future.

In conclusion, we performed an observer study to evaluate the effects of CAD on the diagnostic performance of radiologists for lung nodules on CT. We found that radiologists obtained significant improvement in diagnostic accuracy with the computer aid. The recommended action changes as a result of CAD were also mostly beneficial. These results suggest that CAD may be helpful as a second opinion in increasing diagnostic confidence for radiologists. Future work includes expanding the data set to increase the number of training samples, improving the performance of the CAD system, and evaluating the CAD system with a previously unseen test set. Further studies are also needed to determine whether similar improvement in diagnostic accuracy will be realized for radiologists of different experience levels and in clinical practice.

6.6 Tables

Table 6.1: The individual and average performance of the observers in terms of the area under the ROC curve (A_z) without and with CAD. The improvement for three of the radiologists (indicated by *) and the average A_z achieved statistical significance ($p<0.05$).

Observer	Az without CAD	Az with CAD
1*	0.817 ± 0.026	0.848 ± 0.024
2	0.845 ± 0.025	0.857 ± 0.024
3	0.843 ± 0.024	0.847 ± 0.024
4*	0.829 ± 0.025	0.853 ± 0.023
5*	0.847 ± 0.024	0.877 ± 0.021
6	0.817 ± 0.026	0.834 ± 0.025
Average	0.833	0.853

Table 6.2: The number of times on average that observers changed their likelihood of malignancy (LM Change) estimate and recommended action (Action Change) with CAD, with the percentage relative to the total number of nodules in parenthesis.

	LM Change	Action Change
Total	126 ± 46.8 (49±17%)	10.8 ± 5.8 (4±2%)
Correct	95 ± 34.0 (37±13%)	6.8 ± 2.5 (3±1%)
Incorrect	31 ± 18.2 (12±7%)	4 ± 3.6 (2±1%)

Table 6.3: The average performance of the observers in terms of A_z and $A_z^{0.9}$ for the entire data set, the primary and metastatic subsets, and the data set excluding the eight masses (>30 mm). All improvements with CAD were statistically significant ($p < 0.05$).

Data set	Avg. Az without CAD	Avg. Az with CAD	Az0.9 without CAD	Az0.9 with CAD
All 256 nodules	0.833	0.853	0.390	0.456
Primary cancers	0.823	0.848	0.338	0.415
Met cancers	0.849	0.861	0.493	0.535
Excl (>30 mm) diam	0.832	0.85	0.392	0.455

6.7 Figures

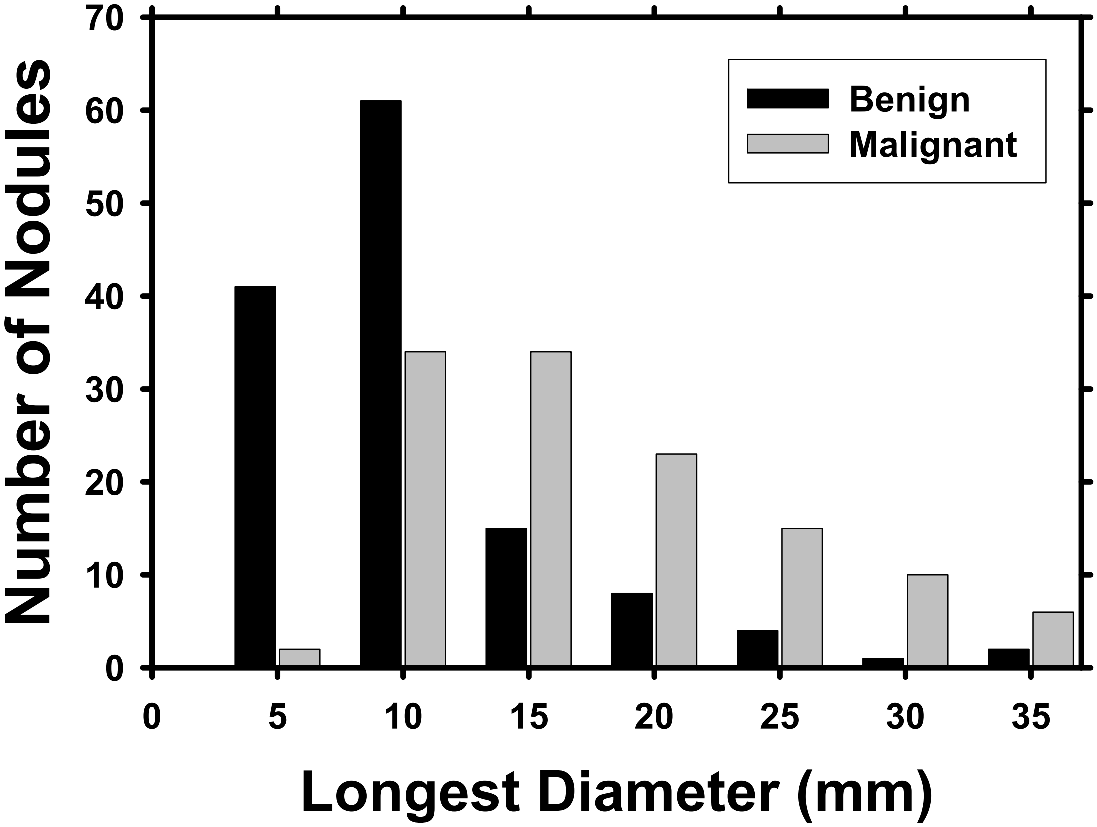
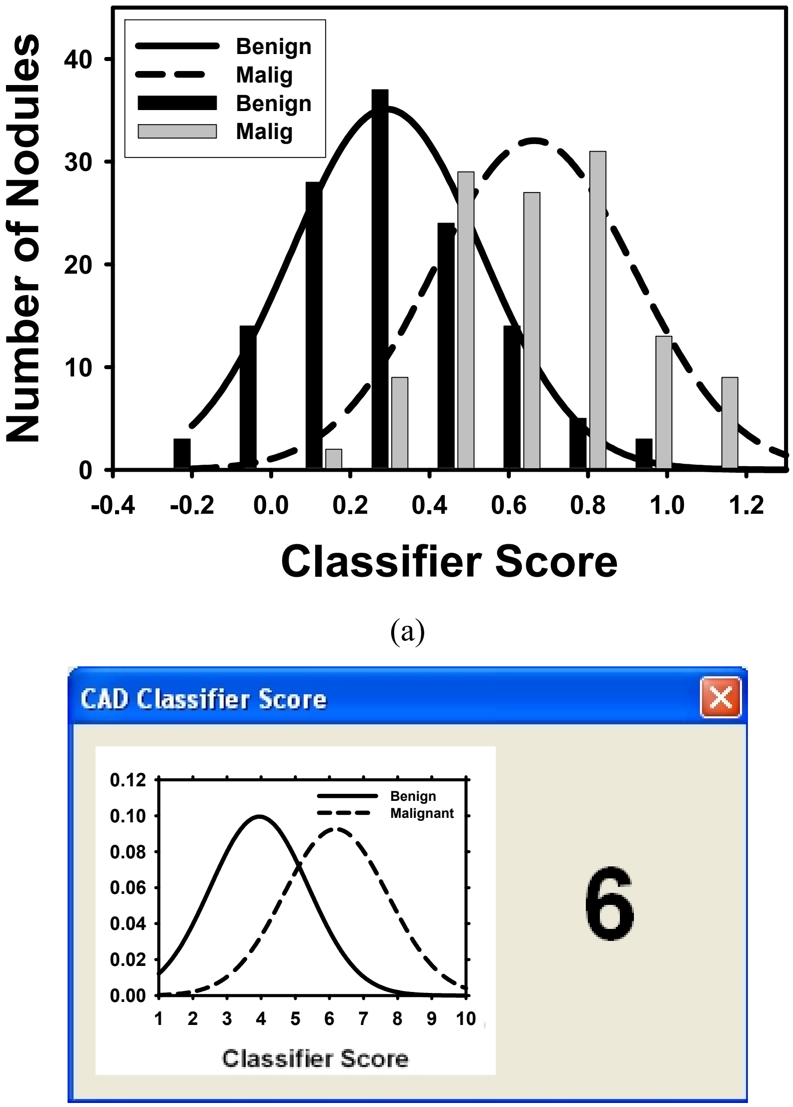


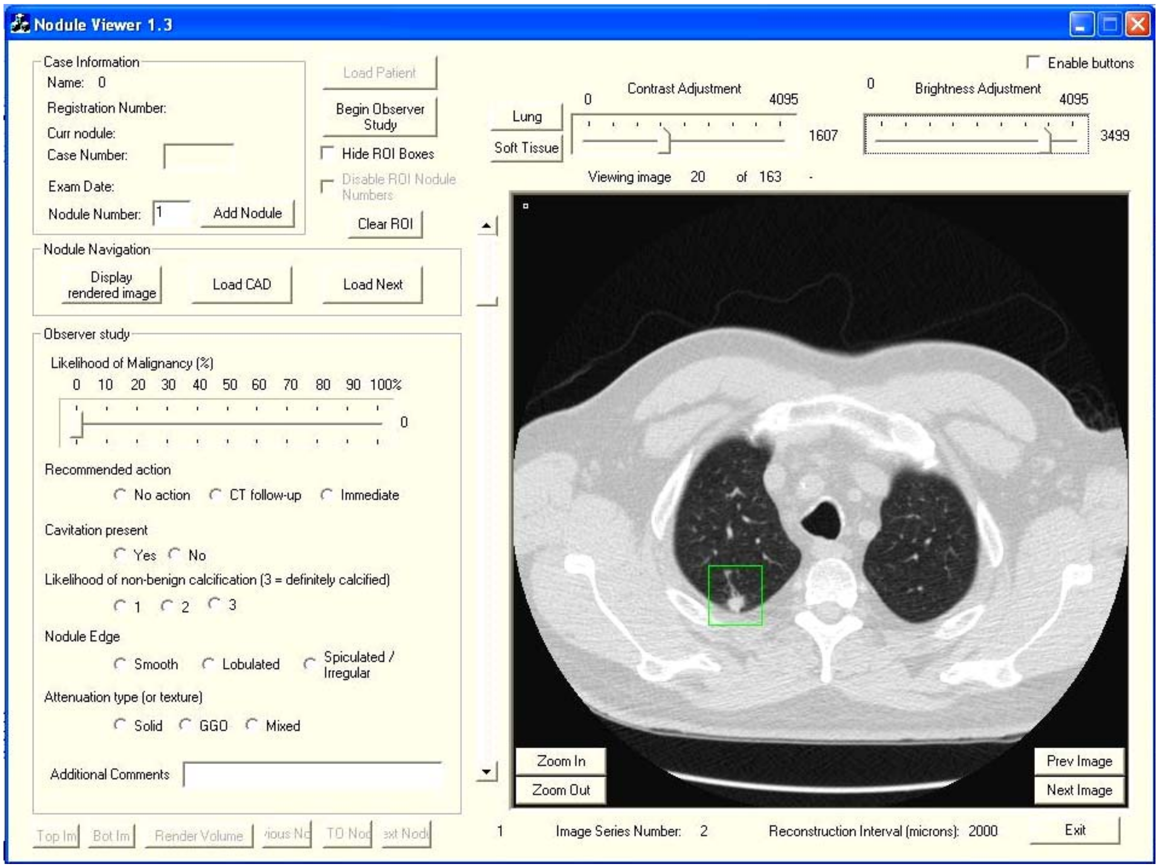
Figure 6.1: Histograms of the longest diameters of the benign and malignant nodules as measured by experienced chest radiologists.



(a)

(b)

Figure 6.2: (a) The 10-bin histogram of classifier scores with fitted Gaussian distributions for the malignant and benign classes. (b) Sample malignancy rating for a nodule shown to an observer in reference to the class distributions (solid line = benign, dashed line = malignant). In this relative scale, a score of 5 yielded a likelihood ratio of about 1.



Case Information
Name: 0
Registration Number:
Curr nodule:
Case Number:
Exam Date:
Nodule Number: 1

□ Hide ROI Boxes
□ Disable ROI Nodule Numbers
□ Enable buttons

Contrast Adjustment 0 4095 1607
Brightness Adjustment 0 4095 3499
Viewing image 20 of 163

Nodule Navigation

Observer study

Likelihood of Malignancy (%)
0 10 20 30 40 50 60 70 80 90 100% 0

Recommended action
○ No action ○ CT follow-up ○ Immediate

Cavitation present
○ Yes ○ No

Likelihood of non-benign calcification (3 = definitely calcified)
○ 1 ○ 2 ○ 3

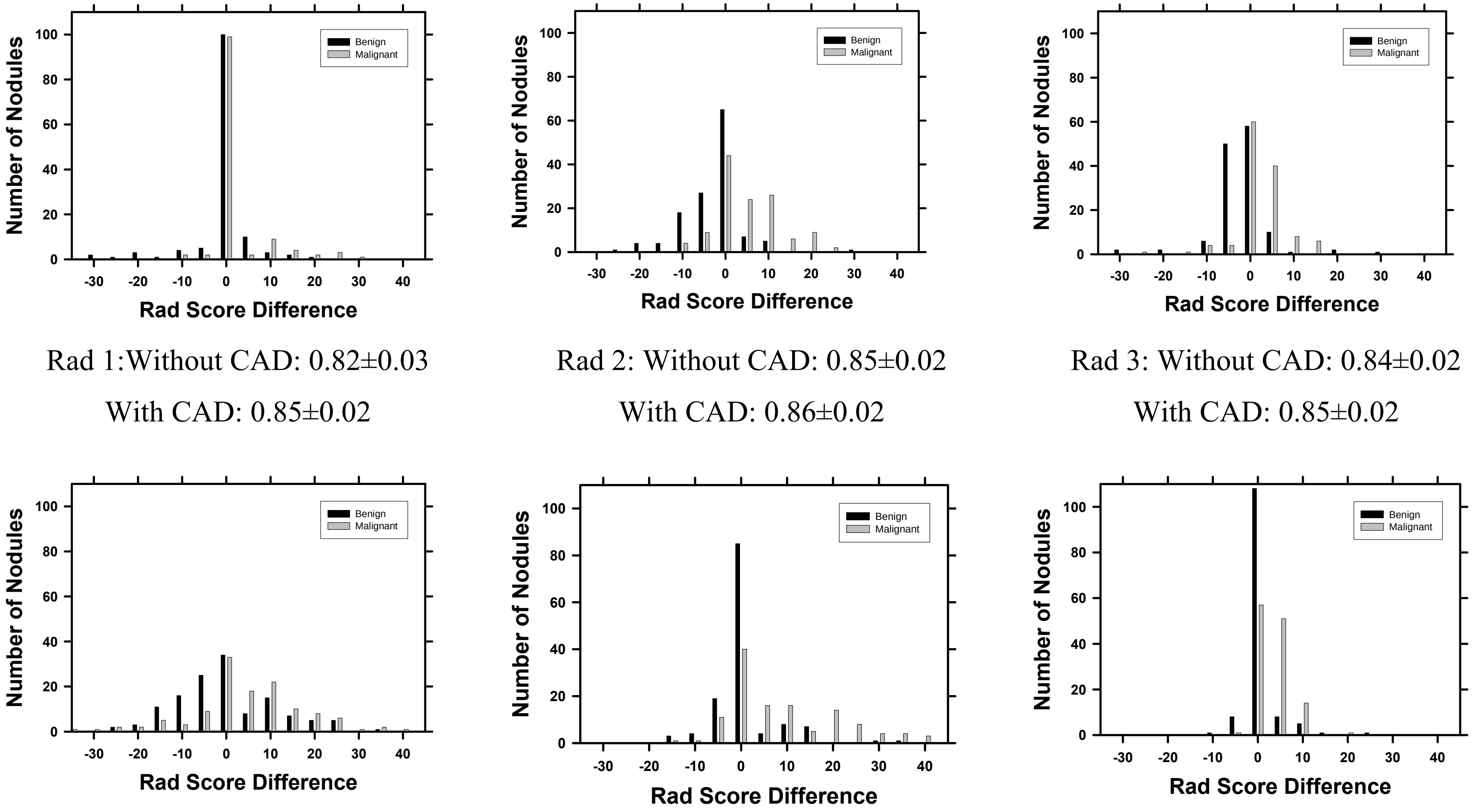
Nodule Edge
○ Smooth ○ Lobulated ○ Spiculated / Irregular

Attenuation type (or texture)
○ Solid ○ GGO ○ Mixed

Additional Comments

1 Image Series Number: 2 Reconstruction Interval (microns): 2000

Figure 6.3: The graphical user interface used by radiologists in the observer study. The first slice of a scan presented is the one containing the nodule marked in a box. The CAD system score (Fig. 6.2(b)) would appear in the upper middle of the screen after the user clicks "Load CAD."



Rad 1:Without CAD: 0.82±0.03

With CAD: 0.85±0.02

Rad 2: Without CAD: 0.85±0.02

With CAD: 0.86±0.02

Rad 3: Without CAD: 0.84±0.02

With CAD: 0.85±0.02

Rad 4: Without CAD: 0.83±0.03

With CAD: 0.85±0.02

Rad 5: Without CAD: 0.85±0.02

With CAD: 0.88±0.02

Rad 6: Without CAD: 0.82±0.03

With CAD: 0.83±0.02

Figure 6.4: Difference in scores for individual radiologists without and with CAD.

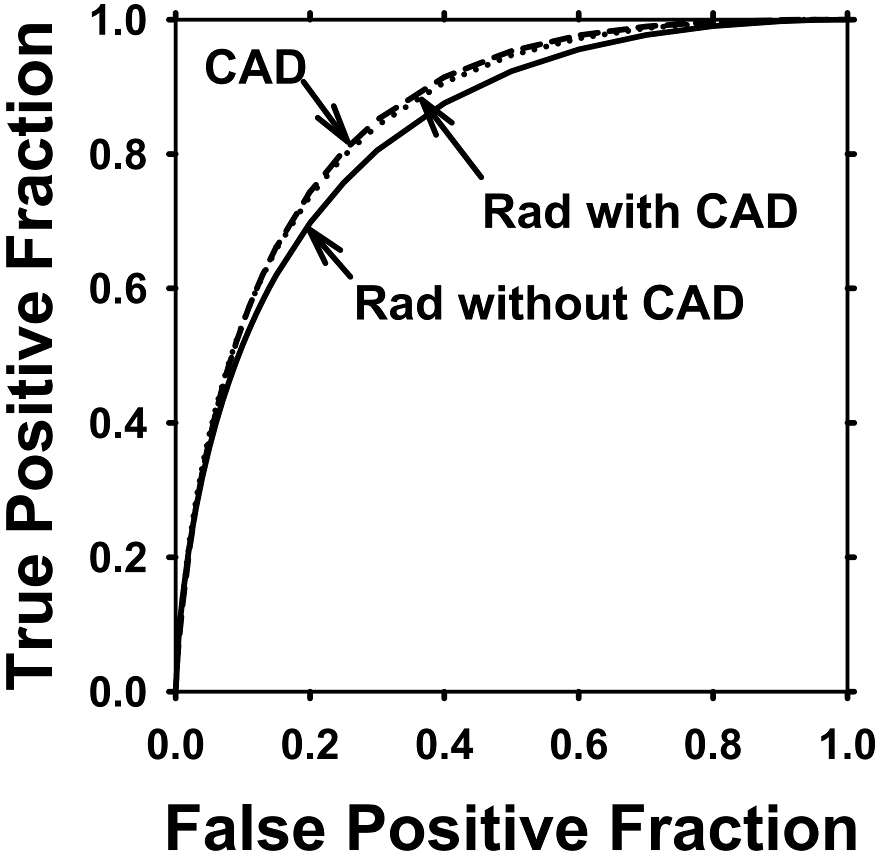


Figure 6.5: Average ROC curves for the six radiologists without (A_z = 0.833) and with CAD (A_z = 0.853) (p<0.01), and the CAD system performance (test A_z = 0.857 ± 0.023).

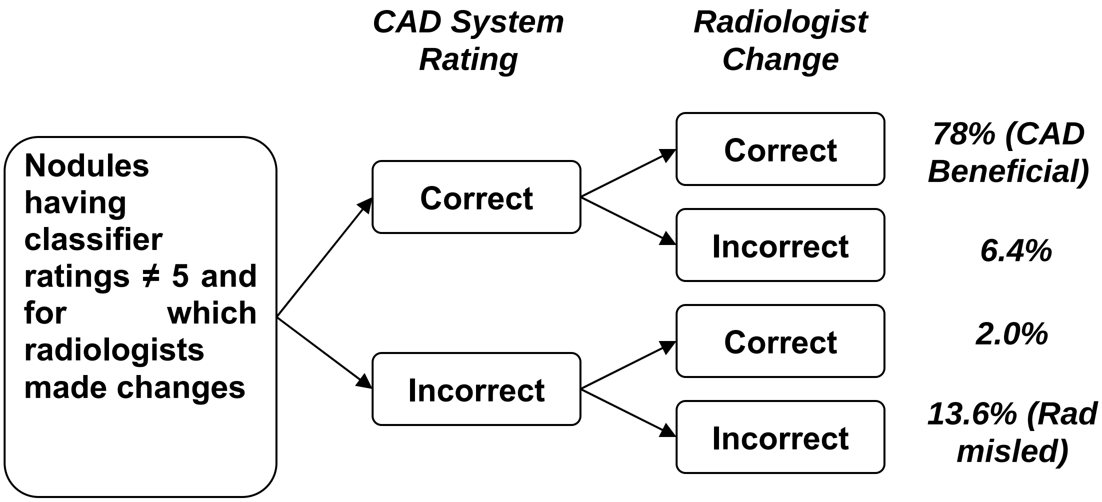


Figure 6.6: Influence of CAD system's malignancy rating on radiologists' LM estimate of nodules.

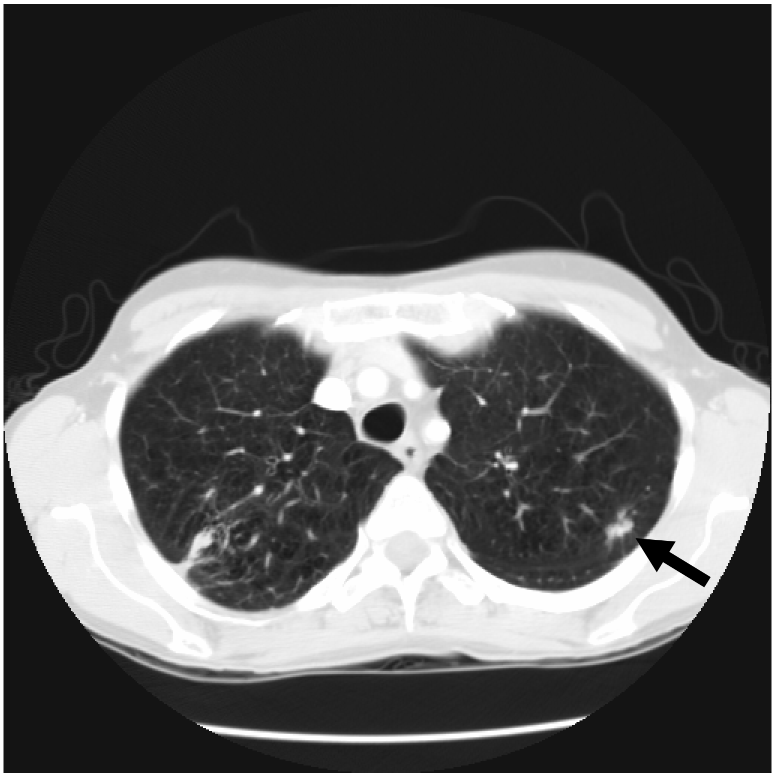


Figure 6.7: Example of a non-small cell lung cancer that radiologists gave an average likelihood of malignancy of 45.8%, but increased it to 58.3% after seeing the classifier rating of 7, showing the beneficial effect of CAD.

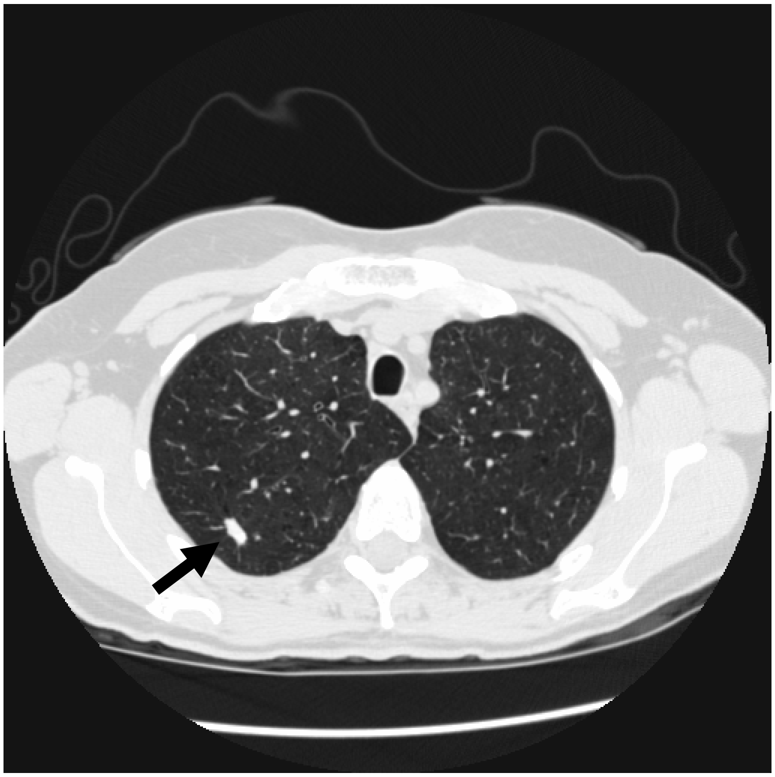


Figure 6.8: Example of a benign nodule that radiologists gave an average likelihood of malignancy of 53.3% but was reduced to 48.3% after seeing the classifier score of 4, showing the beneficial effect of CAD.

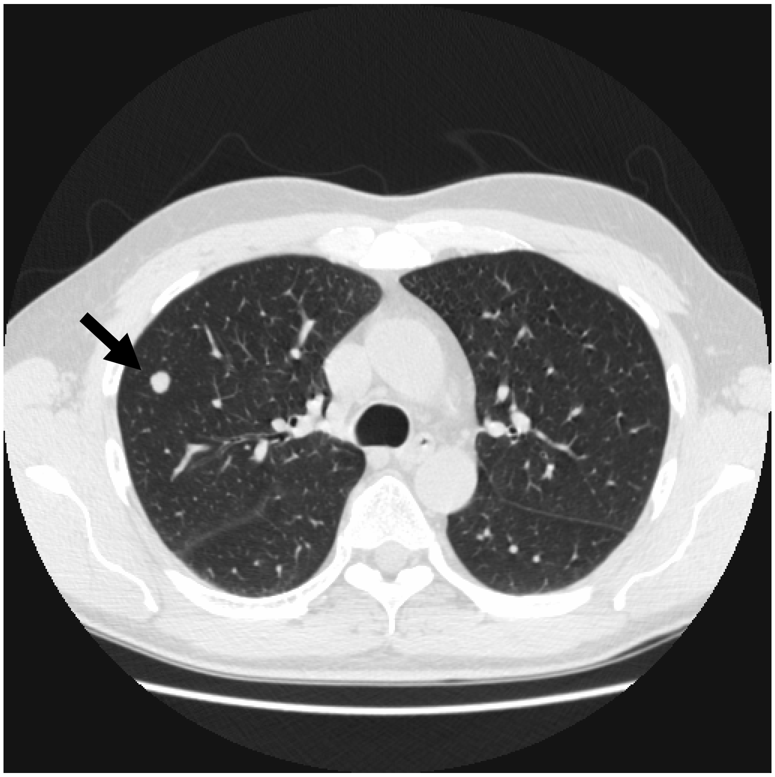


Figure 6.9: Biopsy determined that this was adenoid cystic carcinoma, and radiologists gave it an average LM of 57.5%. Though the classifier score of 4 was incorrect, the radiologists changed the likelihood to an average of 55.8%, showing that radiologists are not easily misled by the CAD system if they believe CAD is incorrect.

Chapter 7
Summary and Future Work

7.1 Summary

We have investigated and improved the various components of a computer-aided diagnosis (CAD) system, Our CAD system pre-processes a volume-of-interest (VOI) containing the nodule to obtain an initial contour. The active contour model is used to segment the nodule, from which features may then be extracted. Feature selection chooses the best set of features to build the classifier. The classifier's output is the malignancy rating. We demonstrated a statistically significant improvement in the malignancy assessment of 256 nodules by six fellowship-trained thoracic radiologists with CAD.

Effective segmentation of the nodule is vital for extracting accurate feature descriptors. We adapted the active contour model for its speed and accuracy in segmentation. We added new energy terms to the cost function to use 3D information from volumetric CT data. Manually drawn contours by radiologists serve as the gold standard for nodule segmentation, but even these contours vary between experts. Figures-of-merit such as overlap or distance between points on the contours have been proposed. Since the ultimate goal of segmentation in our application is for nodule classification, we used the classification accuracy quantified by the area under the receiver operating characteristic (ROC) curve, A_z, as the figure-of-merit. This figure-of-merit was effective in providing high classification performance during leave-one-out test evaluation. We further tested our segmentation on an independent data set provided by the Lung Image Database Consortium and compared them with the gold standard segmented boundaries provided by 6 expert chest radiologists to verify its performance. In addition to using test A_z as a measure of effectiveness, we also proposed metrics to quantify segmentation results.

CT scanning technology has been improving rapidly; comprehensive and quantitative evaluation on the effects of scanning parameters on segmentation performance in terms of volumetric measurements has not been conducted to our knowledge. When patients receive follow-up CT scans, often the scans are acquired with a different set of parameters that are determined depending on the need at the time. For example, a low-dose, low-resolution scan at baseline may be followed by a scan at higher dose and resolution in subsequent repeat scans. If CAD is to be used to track indicators of malignancy such as interval change in volume, then the scanning parameters may affect the accuracy of segmentation and thus the change assessment. Therefore, we investigated the effect of slice thickness, tube current, pitch, and field-of-view on automated segmentation. We found that varying slice thickness had a significant effect on volume measurement, and we concluded that patients should be scanned using identical parameters at follow-up. The investigation also provided insight in how the partial-volume effect influences volume measurements. Knowledge of the potential biases in volume change assessments could help radiologists interpret volume estimates more appropriately.

The number of features that could be extracted from image analysis is essentially unlimited. However, the only features that are useful are those that are effective in discriminating between malignant and benign nodules. The presence of useless features in the feature space can be detrimental because they may obscure the function of the truly effective features, especially when the training sample size is small. Given the limited amount of clinical data with ground truth that is available, we conducted a simulation study to investigate the effect of finite sample size on CAD system performance. Many feature selection methods have been discussed in the literature. We chose to evaluate several commonly used methods for our application, including the stepwise and floating search algorithms that use a classifier to guide the search for features. The feature selection methods were combined with linear discriminant analysis (LDA) and support vector machine (SVM) classifiers. The relative performance of these combinations was compared systematically in a range of training sample sizes and different feature space distributions.

We designed an improved CAD system for lung nodule classification by designing new feature descriptors based on the gradient field. Applying a two-loop leave-one-case-out resampling method to the limited clinical data set, we sought to reduce the optimistic bias on the test results that could occur when a one-loop leave-one-case-out scheme is used. We showed that classification performance improved significantly with the addition of the new features. Furthermore, we performed PCA on the extracted features and varied the order, i.e., number of principal components used, as input to the LDA and SVM classifiers. Using various kernels and parameters of the SVM, we found that there was not a single kernel and set of parameters that consistently performed better than the LDA. Only when the best-performing SVM, out of all the kernels and parameters studied, was chosen for a given PCA order did the SVM have higher test A_z than the LDA.

To test the effectiveness of the CAD system, we conducted an observer ROC study with six fellowship-trained thoracic radiologists serving as readers to compare nodule classification without and with CAD. They provided a malignancy rating for a nodule on a 0-100% scale without CAD. Immediately following the viewing of the CAD system's assessment, they were allowed to modify their rating. Using the Dorfman-Berbaum-Metz method for multi-reader multi-case ROC analysis, we found that the radiologists on average achieved statistically significant improvement in malignancy assessment with CAD. These results indicate the potential usefulness of the CAD system as a second reader for radiologists in characterizing lung nodules.

7.2 Future Work

Ideally, a CAD system would have 100% accuracy in characterizing lung nodules, so that no patient would have to endure a biopsy for a benign nodule. Radiologists and physicians would need to only focus on determining the type of cancer the patient has and deciding on the best course of treatment. Barring that, a more realistic goal would be for the CAD to be accurate enough to provide a reliable second opinion. Extending the investigations conducted in this dissertation towards that goal would involve:

- Reducing the number of energy terms in the 3DAC total energy equation –

The weights for the eight current energies need to be determined for segmentation to be robust when applied to new nodules. Searching for the set of eight ideal weights with a small data set may lead to sub-optimal results. A smaller search space would reduce complexity and overfitting.

- Exploring other segmentation methods –
  The 3D AC is faster compared to level set segmentation methods and more accurate compared to adaptive thresholding methods. Accurate segmentation is important for measuring features such as volume. Further comparison with other segmentation methods to search for faster and more accurate methods than the 3D AC may be warranted.
- Extending the feature space beyond what can be extracted from the VOI –
  Only image features that may be extracted from the VOI are used in the current system. A radiologist makes use of other information to make a decision. A study in our laboratory is currently underway on the use of temporal features from repeated CT scans over time, such as change in volume or in tissue texture around the nodule. Clinical and demographic information could also be important indicators as to the overall health of the patient. We added age at the time of the CT examination and gender to the feature space, but those features did not significantly improve the classifier's performance. The number of other detected nodules and their features could be combined to provide an overall view of the patient's lungs.
- Incorporating other clinical information –
  For example, the ability to process PET/CT fusion scans would add another dimension in imaging. Radiologists read through a patient's clinical history to understand what other radiologists have found in the past, or to determine any factors that may contribute to the patient's state of health. Natural language processing techniques to mine these reports may benefit the decision making of the CAD system.
- Interactivity with radiologists –
  Currently the output of the CAD system for a malignancy rating is an integer on a scale of 1 to 10, and that may not give radiologists an understanding of how the

CAD system estimated the likelihood of malignancy of a nodule. If the CAD system could provide feature values or other evidence to support its decision such as showing nodules with similar characteristics that have been biopsy-proven, the radiologists may have more confidence in the CAD system's assessment. If the radiologist disagrees with the classifier result, this information could also prevent the radiologist from second-guessing himself when he knows how the CAD system may have calculated an incorrect assessment.

- Improvement with new data –

  A large training set is essential for the CAD system to learn from the characteristics of malignant and benign nodules. A method to collect nodules with proven diagnoses and periodic retraining of the CAD system is needed for improvement in accuracy.

- Performing an observer study with various groups of radiologists –

  The readers who participated in our observer study were fellowship-trained thoracic radiologists. It will be important to evaluate the effects of CAD on general radiologists, community hospital-based radiologists, and residents.

Bibliography

[1] T. W. Way, L. M. Hadjiiski, B. Sahiner, H.-P. Chan, P. N. Cascade, E. A. Kazerooni, N. Bogot, and C. Zhou, "Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours," *Medical Physics*, vol. 33, pp. 2323-2337, 2006.

[2] T. W. Way, H.-P. Chan, M. M. Goodsitt, B. Sahiner, L. M. Hadjiiski, C. Zhou, and A. Chughtai, "Effect of CT scanning parameters on volumetric measurements of pulmonary nodules by 3D active contour segmentation: a phantom study," *Physics in Medicine and Biology*, vol. 53, pp. 1295-1312, 2008.

[3] L. A. G. Ries, D. Harkins, M. Krapcho, A. Mariotto, B. A. Miller, E. J. Feuer, L. Clegg, M. P. Eisner, M. J. Horner, N. Howlader, M. Hayat, B. F. Hankey, and B. K. Edwards, "SEER Cancer Statistics Review, 1975-2003," National Cancer Institute, Bethesda, MD 2006.

[4] M. Unger, "A Pause, Progress, and Reassessment in Lung Cancer Screening," *N Engl J Med*, vol. 355, pp. 1822-1824, 2006.

[5] "Early lung cancer detection: summary and conclusions," *Am Rev Respiratory Disease*, vol. 130, pp. 565-70, 1984.

[6] S. Diederich, D. Wormanns, M. Semik, M. Thomas, H. Lenzen, N. Roos, and W. Heindel, "Screening for early lung cancer with low-dose spiral CT: Prevalence in 817 asymptomatic smokers," *Radiology*, vol. 222, pp. 773-781, 2002.

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

[8] T. Nawa, T. Nakagawa, S. Kusano, Y. Kawasaki, Y. Sugawara, and H. Nakata, "Lung cancer screening using low-dose spiral CT: Results of baseline and 1-Year follow-up studies," *Chest*, vol. 122, pp. 15-20, 2002.

[9] S. Sone, S. Takashima, F. Li, F. Yang, T. Honda, Y. Maruyama, M. Hasega, T. Yamanda, K. Kubo, K. Hanamura, and K. Asakura, "Mass screening for lung cancer with mobile spiral computed tomography scanner," *The Lancet*, vol. 352, pp. 1242-1245, 1998.

[10] S. Sone, F. Li, Z. G. Yang, T. Honda, Y. Maruyama, S. Takashima, M. Hasegawa, S. Kawakami, K. Kubo, K. M. Haniuda, and T. Yamanda, "Results of three-year mass screening programme for lung cancer using mobile low-dose spiral computed tomography scanner," *British J. of Cancer*, vol. 84, pp. 25-32, 2001.

[11] S. J. Swensen, J. R. Jett, T. E. Hartman, D. E. Midthun, S. J. Mandrekar, S. L. Hillman, A.-M. Sykes, G. L. Aughenbaugh, and A. O. B. L. Allen, "CT screening for lung cancer: Five-year prospective experience," *Radiology*, vol. 235, pp. 259-265, 2005.

[12] S. J. Swensen, J. R. Jett, T. E. Hartman, D. E. Midthun, J. A. Sloan, A. M. Sykes, G. L. Aughenbaugh, and M. A. Clemens., "Lung cancer screening with CT: Mayo Clinic experience," *Radiology*, vol. 226, pp. 756-761, 2003.

[13] C. I. Henschke, D. I. McCauley, D. F. Yankelevitz, D. P. Naidich, G. McGuinness, O. S. Miettinen, D. M. Libby, M. W. Pasmantier, J. Koizumi, N. K. Altorki, and J. P. Smith, "Early lung cancer action project: overall design and findings from baseline screening," *The Lancet*, vol. 354, pp. 99-105, 1999.

[14] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models," *International Journal of Computer Vision*, vol. 1, pp. 321-331, 1987.

[15] D. J. Williams and M. Shah, "A fast algorithm for active contours and curvature estimation," *CVGIP: Image Understanding*, vol. 55, pp. 14-26, 1992.

[16] T. W. Way, B. Sahiner, L. Hadjiiski, H.-P. Chan, N. Bogot, P. Cascade, E. Kazerooni, and J. A. Fessler, "Segmentation of pulmonary nodules with 3D active contour model for computer-aided diagnosis," presented at RSNA Program Book, Radiological Soc. N. Amer., Chicago, IL, 2003.

[17] T. W. Way, L. M. Hadjiiski, B. Sahiner, H. P. Chan, P. Cascade, N. Bogot, E. A. Kazerooni, J. A. Fessler, and Z. Ge, "Classification of Pulmonary Nodules Using Automated 3D Segmentation and Feature Classification for Computer-aided Diagnosis on CT Scans," presented at RSNA Program Book, Radiological Soc. N. Amer., Chicago, IL, 2004.

[18] T. W. Way, B. Sahiner, L. Hadjiiski, H.-P. Chan, P. Cascade, E. Kazerooni, N. Bogot, and C. Zhou, "Computer-aided Diagnosis (CAD) of Malignant and Benign Lung Nodules on CT Scans: The Effect of the Primary versus Metastatic Status of the Malignancy," presented at RSNA Program Book, Radiological Soc. N. Amer., Chicago, IL, 2005.

[19] T. W. Way, L. Hadjiiski, B. Sahiner, H.-P. Chan, M. Goodsitt, and C. Zhou, "Evaluation of Volumetric Measurement of CT Phantom Spheres and LIDC Nodules by 3D Active Contour Segmentation in a Computer-Aided Diagnosis (CAD) System " presented at RSNA Program Book, Radiological Soc. N. Amer., Chicago, IL, 2005.

[20] L. M. Hadjiiski, B. Sahiner, H.-P. Chan, N. Bogot, P. N. Cascade, E. A. Kazerooni, and T. W. Way, "Computer-aided diagnosis of lung cancer: Interval change analysis of nodule features in serial CT examinations," presented at RSNA, Radiological Soc of N America, Chicago, IL, 2004.

[21] L. M. Hadjiiski, T. W. Way, B. Sahiner, H. P. Chan, P. N. Cascade, N. Bogot, E. A. Kazerooni, and C. Zhou, "Computer-aided diagnosis for interval change analysis of lung nodule features in serial CT examinations," *Proc. SPIE*, vol. 6514, pp. 111-117, 2007.

[22] T. W. Way, L. M. Hadjiiski, B. Sahiner, H.-P. Chan, P. N. Cascade, E. A. Kazerooni, N. Bogot, and C. Zhou, "Classification of CT lung nodules by a computer-aided diagnosis (CAD) system with texture and gradient field features," presented at RSNA Program Book, Radiological Soc. N. Amer., Chicago, IL, 2006.

[23] T. W. Way, H.-P. Chan, J. Stojanovska-Nojkova, L. Frank, T. K. Song, E. A. Kazerooni, P. N. Cascade, A. Chughtai, A. Attili, B. Sahiner, and L. Hadjiiski, "Effect of computer-aided diagnosis (CAD) on radiologists' characterization of

lung nodules on CT: An ROC study," *RSNA Program Book*, vol. 2007, pp. 267, 2007.

[24] T. W. Way, L. M. Hadjiiski, B. Sahiner, A. Chughtai, A. Attili, C. Poopat, J. Stojanovska-Nojkova, L. Frank, T. K. Song, E. A. Kazerooni, P. N. Cascade, and H. P. Chan, "Computer-aided diagnosis for classification of malignant and benign lung nodules on thoracic CT scans: Hands-on experience with an interactive system," presented at RSNA Program Book, Rad. Soc. N. Amer., Chicago, IL, 2007.

[25] B. Sahiner, L. M. Hadjiiski, H. P. Chan, J. Shi, T. W. Way, P. N. Cascade, E. A. Kazerooni, C. Zhou, and J. Wei, "The effect of nodule segmentation on the accuracy of computerized lung nodule detection on CT scans: Comparison on a data set annotated by multiple radiologists," *Proc. SPIE*, vol. 6514, pp. 65140L-1,7, 2007.

[26] M. M. Goodsitt, H.-P. Chan, S. C. Larson, E. G. Christodoulou, J. Kim, and T. W. Way, "CT number accuracy of simulated lung nodules imaged with a multi-detector CT scanner," presented at AAPM 47th Annual Meeting, Seattle, WA, USA, 2005.

[27] M. M. Goodsitt, H.-P. Chan, T. W. Way, S. C. Larson, E. G. Christodoulou, and J. Kim, "Accuracy of the CT numbers of simulated lung nodules imaged with multi-detector CT scanners," *Medical Physics*, vol. 33, pp. 3006-3017, 2006.

[28] M. M. Goodsitt, H.-P. Chan, T. W. Way, S. C. Larson, and E. G. Christodoulou, "SU-EE-A4-01: CT Number Accuracy of Lung Nodules: Effect of Patient Body Size and Lung Size," *Medical Physics*, vol. 33, pp. 1995, 2006.

[29] M. Goodsitt, H.-P. Chan, T. W. Way, S. Larson, and E. Christodoulou, "SU-FF-I-07: Single- and Dual-Energy CT Calibration Lines for Assessing the Calcium Content of Lung Nodules: Effects of Patient Body and Lung Nodule Size," *Medical Physics*, vol. 34, pp. 2339, 2007.

[30] "American Cancer Society, www.cancer.org 2005, "Cancer Facts & Figures 2005"," 2005.

[31] S. J. Swensen, R. W. Viggiano, D. E. Midthun, N. L. Müller, A. Sherrick, K. Yamashita, D. P. Naidich, E. F. Patz, T. E. Hartman, J. R. Muhm, and A. L. Weaver, "Lung Nodule Enhancement at CT: Multicenter Study," *Radiology*, vol. 214, pp. 73-80, 2000.

[32] T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center," *Radiology*, vol. 220, pp. 781-786, 2001.

[33] S. G. Armato, M. L. Giger, C. J. Moran, J. T. Blackburn, K. Doi, and H. MacMahon, "Computerized detection of pulmonary nodules on CT scans," *RadioGraphics*, vol. 19, pp. 1303-1311, 1999.

[34] M. S. Brown, M. F. McNitt-Gray, J. G. Goldin, R. D. Suh, J. W. Sayre, and D. R. Aberle, "Patient-specific models for lung nodule detection and surveillance in CT images," *IEEE Trans Med Imag*, vol. 20, pp. 1242-1250, 2001.

[35] P. Croisille, M. Souto, M. Cova, S. Wood, Y. Afework, J. E. Kuhlman, and E. A. Zerhouni, "Pulmonary nodules: improved detection with vascular segmentation and extraction with sprial CT," *Radiology*, vol. 197, pp. 397:401, 1995.

[36] R. Malladi, J. A. Sethian, and B. C. Vemuri, "Shape modeling with front propagation: a level set approach," *IEEE Pat. Analysis and Mach. Intel.*, vol. 17, pp. 158 - 175, 1995.

[37] J. Yang, L. H. Staib, and J. S. Duncan, "Neighbor-constrained segmentation with level set based 3-D deformable models," *IEEE Trans Med Imag*, vol. 23, pp. 940-948, 2004.

[38] M. B. Cuadra, C. Pollo, A. Bardera, O. Cuisenaire, J.-G. Villemure, and J.-P. Thiran, "Atlas-Based Segmentation of Pathological MR Brain Images Using a Model of Lesion Growth," *IEEE Trans Med Imag*, vol. 23, pp. 1301-1314, 2004.

[39] M. N. Gurcan, B. Sahiner, N. Petrick, H. P. Chan, E. A. Kazerooni, P. N. Cascade, and L. Hadjiiski, "Lung nodule detection on thoracic computed tomography images: preliminary evaluation of a computer-aided diagnosis system," *Medical Physics*, vol. 29, pp. 2552-2558, 2002.

[40] M. S. Brown, J. G. Goldin, R. D. Suh, M. F. McNitt-Gray, J. W. Sayre, and D. R. Aberle, "Lung micronodules: Automated method for detection at thin-section CT - Initial experience," *Radiology*, vol. 226, pp. 256-262, 2003.

[41] K. G. Kim, J. M. Goo, J. H. Kim, H. J. Lee, B. G. Min, K. T. Bae, and J.-G. Im, "Computer-aided Diagnosis of Localized Ground-Glass Opacity in the Lung at CT: Initial Experience," *Radiology*, vol. 237, pp. 657-661, 2005.

[42] S. Armato, F. Li, M. Giger, H. MacMahon, S. Sone, and K. Doi, "Lung cancer: Performance of automated lung nodule detection applied to cancers missed in a CT screening program," *Radiology*, vol. 225, pp. 685-692, 2002.

[43] K. T. Bae, J.-S. Kim, Y.-H. Na, K. G. Kim, and J.-H. Kim, "Pulmonary Nodules: Automated Detection on CT Images with Morphologic Matching Algorithm--Preliminary Results," *Radiology*, vol. 236, pp. 286-293, 2005.

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

[45] K. Kanazawa, Y. Kawata, N. Niki, H. Satoh, H. Ohmatsu, R. Kakinuma, M. Kaneko, N. Moriyama, and K. Eguchi, "Computer-aided diagnosis for pulmonary nodules based on helical CT images," *Computerized Medical Imaging and Graphics*, vol. 22, pp. 157-167, 1998.

[46] K. Suzuki, S. G. Armato, F. Li, S. Sone, and K. Doi, "Massive training artificial neural network (MTANN) for reduction of false positives in computerized detection of lung nodules in low-dose computed tomography," *Med. Phys*, vol. 30, pp. 1602-1617, 2003.

[47] M. F. McNitt-Gray, E. M. Hart, N. Wyckoff, J. W. Sayre, J. G. Goldin, and D. R. Aberle, "A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution CT: Preliminary results," *Medical Physics*, vol. 26, pp. 880-888, 1999.

[48] S. K. Shah, M. F. McNitt-Gray, S. R. Rogers, J. G. Goldin, R. D. Suh, J. W. Sayre, I. Petkovska, H. J. Kim, and D. R. Aberle, "Computer-aided Diagnosis of the Solitary Pulmonary Nodule," *Academic Radiology*, vol. 12, pp. 570-575, 2005.

[49] S. G. Armato, M. B. Altman, and J. Wilkie, "Automated lung nodule classification following automated nodule detection on CT: a serial approach," *Medical Physics*, vol. 30, pp. 1188-1197, 2003.

[50] Y. Kawata, N. Niki, H. Ohmatsu, R. Kakinuma, K. Eguchi, M. Kaneko, and N. Moriyama, "Quantitative surface characterization of pulmonary nodules based on thin-section CT images," *IEEE Trans Nuclear Science*, vol. 45, pp. 2132-2138, 1998.

[51] F. Li, M. Aoyama, J. Shiraishi, H. Abe, Q. Li, K. Suzuki, R. Engelmann, S. Sone, H. MacMahon, and a. K. Doi, "Radiologists' performance for differentiating benign from malignant lung nodules on high-resolution CT using computer-estimated likelihood of malignancy," *AJR Am J Roentgenol*, vol. 183, pp. 1209-15, 2004.

[52] M. Aoyama, Q. Li, S. Katsuragawa, F. Li, S. Sone, and K. Doi, "Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose CT images," *Med. Phys.*, vol. 30, pp. 387-394, 2003.

[53] K. Suzuki, F. Li, S. Sone, and K. Doi, "Computer-Aided Diagnostic Scheme for Distinction Between Benign and Malignant Nodules in Thoracic Low-Dose CT by Use of Massive Training Artificial Neural Network," *IEEE Transactions on Medical Imaging*, vol. 24, pp. 1138-1150, 2005.

[54] S. G. Armato, G. McLennan, M. F. McNitt-Gray, C. R. Meyer, D. Yankelevitz, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, A. P. Reeves, B. Y. Croft, and L. P. Clarke, "Lung Image Database Consortium: Developing a Resource for the Medical Imaging Research Community," *Radiology*, vol. 232, pp. 739-748, 2004.

[55] B. Sahiner, H. P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue on mammograms," *Medical Physics*, vol. 23, pp. 1671-1684, 1996.

[56] C. S. Poon and M. Braun, "Image segmentation by a deformable contour model incorporating region analysis," *Physics in Medicine and Biology*, vol. 42, pp. 1833-1841, 1997.

[57] B. Sahiner, N. Petrick, H. P. Chan, L. M. Hadjiiski, C. Paramagul, M. A. Helvie, and M. N. Gurcan, "Computer-Aided Characterization of Mammographic Masses: Accuracy of Mass Segmentation and its Effects on Characterization," *IEEE Transactions on Medical Imaging*, vol. 20, pp. 1275-1284, 2001.

[58] L. D. Cohen, "On active contour models and balloons," *CVGIP: Image Understanding*, vol. 53, pp. 211-218, 1991.

[59] W. J. Kostis, A. P. Reeves, D. F. Yankelevitz, and C. I. Henschke, "Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images," *IEEE Trans. Med. Imaging*, vol. 22, pp. 1259-1274, 2003.

[60] J. W. Gurney, "Determining the likelihood of malignancy in solitary pulmonary nodules with Bayesian analysis - Part I. Theory," *Radiology*, vol. 186, pp. 405-413, 1993.

[61] J. J. Erasmus, J. E. Connolly, H. P. McAdams, and V. L. Roggli, "Solitary pulmonary nodules: Part I. Morphological evaluation for differentiation of benign and malignant lesions," *Radiographics*, vol. 20, pp. 43-58, 2000.

[62] S. Siegelman, N. Khouri, J. WW Scott, F. Leo, U. Hamper, E. Fishman, and E. Zerhouni, "Pulmonary hamartoma: CT findings," *Radiology*, vol. 160, pp. 313, 1986.

[63] K. Ledor, B. Fish, L. Chaise, and S. Ledor, "CT diagnosis of pulmonary hamartomas," *J Comut Assit Tomography*, vol. 5, pp. 343-344, 1981.

[64] B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mammograms: The rubber band straightening transform and texture analysis," *Medical Physics*, vol. 25, pp. 516-526, 1998.

[65] M. M. Galloway, "Texture classification using gray level run lengths," *Computer Graphics and Image Processing*, vol. 4, pp. 172-179, 1975.

[66] B. R. Dasarathy and E. B. Holder, "Image characterizations based on joint gray-level run-length distributions," *Pattern Recog. Letters*, vol. 12, pp. 497-502, 1991.

[67] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," *Physics in Medicine and Biology*, vol. 40, pp. 857-876, 1995.

[68] W. Spendley, G. R. Hext, and F. R. Himsworth, "Sequential application of simplex designs in optimisation and evolutionary operation," *Technometrics*, vol. 4, pp. 441-461, 1962.

[69] C. E. Metz, "ROC methodology in radiologic imaging," *Investigative Radiology*, vol. 21, pp. 720-733, 1986.

[70] B. Sahiner, H. P. Chan, N. Petrick, M. A. Helvie, and M. M. Goodsitt, "Design of a high-sensitivity classifier based on a genetic algorithm: Application to computer-aided diagnosis," *Physics in Medicine and Biology*, vol. 43, pp. 2853-2871, 1998.

[71] M. M. Goodsitt, H. P. Chan, J. T. Lydick, C. R. Gandra, N. G. Chen, M. A. Helvie, J. Bailey, M. A. Roubidoux, C. E. Blane, B. Sahiner, and N. Petrick, "An observer study comparing spot imaging regions selected by radiologists and a computer for an automated stereo spot mammography technique," *Medical Physics*, vol. 31, pp. 1558-1567, 2004.

[72] C. E. Metz, B. A. Herman, and J. H. Shen, "Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," *Statistics in Medicine*, vol. 17, pp. 1033-1053, 1998.

[73] B. B. Tan, K. R. Flaherty, E. A. Kazerooni, and M. D. Iannettoni, "The solitary pulmonary nodule," *Chest*, vol. 123, pp. 89-96, 2003.

[74] D. Ost and A. Fein, "Evaluation and Management of the Solitary Pulmonary Nodule," *Am. J. Respir. Crit. Care Med*, vol. 162, pp. 782-787, 2000.

[75] F. Fischbach, F. Knollmann, V. Griesshaber, T. Freund, E. Akkol, and R. Felix, "Detection of pulmonary nodules by multislice computed tomography: improved detection rate with reduced slice thickness," *Eur. Radiol.*, vol. 13, pp. 2378–2383, 2003.

[76] H. MacMahon, J. H. M. Austin, G. Gamsu, C. J. Herold, J. R. Jett, D. P. Naidich, E. F. Patz, Jr., and S. J. Swensen, "Guidelines for management of small pulmonary nodules detected on CT scans: a statement from the Fleischner Society," *Radiology*, vol. 237, pp. 395-400, 2005.

[77] D. Ost, A. M. Fein, and S. H. Feinsilver, "The solitary pulmonary nodule," *N Engl J Med*, vol. 348, pp. 2535-2542, 2003.

[78] S. K. Shah, M. F. McNitt-Gray, S. R. Rogers, J. G. Goldin, R. D. Suh, J. W. Sayre, I. Petkovska, H. J. Kim, and D. R. Aberle, "Computer aided characterization of the solitary pulmonary nodule using volumetric and contrast enhancement features," *Academic Radiology*, vol. 12, pp. 1310-1319, 2005.

[79] G. A. Lillington and C. I. Caskey, "Evaluation and management of solitary and multiple pulmonary nodules," *Clin Chest Med*, vol. 14, pp. 111-119, 1993.

[80] K. Usuda, Y. Saito, M. Sagawa, M. Sato, K. Kanma, S. Takahashi, C. Endo, Y. Chen, A. Sakurada, and S. Fujimura, "Tumor doubling time and prognostic assessment of patients with primary lung cancer," *Cancer*, vol. 74, pp. 2239-44, 1994.

[81] N. R. Bogot, E. A. Kazerooni, A. M. Kelly, L. E. Quint, B. Desjardins, and B. Nan, "Interobserver and intraobserver variability in the assessment of pulmonary nodule size on CT using film and computer display methods," *Academic Radiology*, vol. 12, pp. 948–956, 2005.

[82] J. Erasmus, G. Gladish, L. Broemeling, B. Sabloff, M. Truong, R. Herbst, and R. Munden, "Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response," *J Clin Oncol*, vol. 21, pp. 2574-82, 2003.

[83] C. R. Meyer, T. D. Johnson, G. McLennan, D. R. Aberle, E. A. Kazerooni, H. MacMahon, B. F. Mullan, D. F. Yankelevitz, E. J. R. van Beek, and S. G. Armato, III, "Evaluation of lung MDCT nodule annotation across radiologists and methods," *Academic Radiology*, vol. 13, pp. 1254-1265, 2006.

[84] K. Harris, H. Adams, D. Lloyd, and D. Harvey, "The effect on apparent size of simulated pulmonary nodules of using three standard CT window settings," *Clin Radiol*, vol. 47, pp. 241-4, 1993.

[85] H. T. Winer-Muram, S. G. Jennings, C. A. Meyer, Y. Liang, A. M. Aisen, R. D. Tarver, and R. C. McGarry, "Effect of varying CT section width on volumetric measurement of lung tumors and application of compensatory equations," *Radiology*, vol. 229, pp. 184-194, 2003.

[86] D. F. Yankelevitz, R. Gupta, B. Zhao, and C. I. Henschke, "Small pulmonary nodules: evaluation with repeat CT - Preliminary experience," *Radiology*, vol. 212, pp. 561-566, 1999.

[87] D. Yankelevitz, A. Reeves, W. Kostis, B. Zhao, and C. Henschke, "Small pulmonary nodules: volumetrically determined growth rates based on CT evaluation," *Radiology*, vol. 217, pp. 251-6, 2000.

[88] E. A. Zerhouni, J. F. Spivey, R. H. Morgan, F. P. Leo, F. P. Stitik, and S. S. Siegelman, "Factors influencing quantitative CT measurements of solitary pulmonary nodules," *JCAT*, vol. 6, pp. 1075-1087, 1982.

[89] J. Im, G. Gamsu, D. Gordon, M. Stein, W. Webb, C. Cann, and L. Niklason, "CT densitometry of pulmonary nodules in a frozen human thorax," *AJR Am J Roentgenol*, vol. 150, pp. 61-6, 1988.

[90] J. P. Ko, H. Rusinek, E. L. Jacobs, J. S. Babb, M. Betke, G. McGuinness, and D. P. Naidich, "Small pulmonary nodules: volume measurement at chest CT—phantom study," *Radiology*, vol. 223, pp. 864-870, 2003.

[91] J. M. Goo, T. Tongdee, R. Tongdee, K. Yeo, C. F. Hildebolt, and K. T. Bae, "Volumetric measurement of synthetic lung nodules with multi–detector row CT: effect of various image reconstruction parameters and segmentation thresholds on measurement accuracy," *Radiology*, vol. 235, pp. 850–856, 2005.

[92] I. Petkovska, M. S. Brown, J. G. Goldin, H. J. Kim, M. F. McNitt-Gray, F. G. Abtin, R. J. Ghurabi, and D. R. Aberle, "The effect of lung volume on nodule size on CT," *Academic Radiology*, vol. 14, pp. 476-485, 2007.

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

[94] J. M. Bland and D. G. Altman, "Statistics notes: Multiple significance tests: the Bonferroni method," *BMJ*, vol. 310, pp. 170, 1995.

[95] D. Wormanns, G. Kohl, E. Klotz, A. Marheine, F. Beyer, W. Heindel, and S. Diederich, "Volumetric measurements of pulmonary nodules at multi-row detector CT: in vivo reproducibility," *Eur Radiol*, vol. 14, pp. 86–92, 2004.

[96] H. P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers," *Medical Physics*, vol. 26, pp. 2654-2668, 1999.

[97] B. Sahiner, H. P. Chan, N. Petrick, R. F. Wagner, and L. M. Hadjiiski, "Feature selection and classifier performance in computer-aided diagnosis: The effect of finite sample size," *Medical Physics*, vol. 27, pp. 1509-1522, 2000.

[98] B. Sahiner, H. P. Chan, and L. Hadjiiski, "Classifier performance prediction for computer-aided diagnosis using a limited data set," *Medical Physics*, vol. 35, pp. 1559-1570, 2008.

[99] B. Sahiner, H. P. Chan, and L. M. Hadjiiski, "Classifier performance estimation under the constraint of a finite sample size: Resampling schemes applied to neural network classifiers," *Neural Networks*, vol. 21, pp. 476–483, 2008.

[100] Q. Li and K. Doi, "Analysis and minimization of overtraining effect in rule-based classifiers for computer-aided diagnosis," *Medical Physics*, vol. 33, pp. 320–328, 2006.

[101] Q. Li and K. Doi, "Comparison of typical evaluation methods for computer-aided diagnostic schemes: Monte Carlo simulation study," *Medical Physics*, vol. 34, pp. 871-876, 2007.

[102] S. V. Beiden, M. A. Maloof, and R. F. Wagner, "A general model for finite-sample effects in training and testing of competing classifiers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1561-1569, 2003.

[103] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample size performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 153-158, 1997.

[104] P. Pudil, J. Novovicová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, pp. 1119-1125, 1994.

[105] M. Kudo and J. Sklansky, "Comparison of algorithms that select features for pattern classifiers," *Pattern Recognition*, vol. 33, pp. 25-41, 2000.

[106] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic Press, 1990.

[107] D. J. Hand, *Discrimination and Classification*. New York: Wiley, 1981.

[108] T. Marill and D. Green, "On the effectiveness of receptors in recognition systems," *Information Theory, IEEE Transactions on*, vol. 9, pp. 11-17, 1963.

[109] A. W. Whitney, "A Direct Method of Nonparametric Measurement Selection," *Computers, IEEE Transactions on*, vol. C-20, pp. 1100-1103, 1971.

[110] N. R. Draper, *Applied regression analysis*. New York: Wiley, 1998.

[111] M. M. Tatsuoka, *Multivariate Analysis, Techniques for Educational and Psychological Research*, 2nd ed. New York: Macmillan, 1988.

[112] *SPSS for Windows Release 6 Professional Statistics*. Chicago, IL: SPSS Inc., 1993.

[113] S. D. Stearns, "On selecting features for pattern classifiers," presented at Third International Conference on Pattern Recognition, Coronado, CA, 1976.

[114] P. A. Lachenbruch, *Discriminant Analysis*. New York: Hafner Press, 1975.

[115] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, pp. 121-167, 1998.

[116] Y. Arzhaeva, M. Prokop, D. M. J. Tax, P. A. D. Jong, C. M. SchaeferProkop, and B. v. Ginneken, "Computer-aided detection of interstitial abnormalities in chest radiographs using a reference standard based on computed tomography," *Medical Physics*, vol. 34, pp. 4798-4809, 2007.

[117] P. Campadelli, E. Casiraghi, and D. Artioli, "A Fully Automated Method for Lung Nodule Detection From Postero-Anterior Chest Radiographs," *Medical Imaging, IEEE Transactions on*, vol. 25, pp. 1588-1603, 2006.

[118] A. K. Jerebko, J. D. Malley, M. Franaszek, and R. M. Summers, "Support vector machines committee classification method for computer-aided polyp detection in CT colonography," *Academic Radiology*, vol. 12, pp. 479-486, 2005.

[119] S. Ruping, "Incremental learning with support vector machines," presented at Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on, 2001.

[120] The International Early Lung Cancer Action Program Investigators, "Survival of Patients with Stage I Lung Cancer Detected on CT Screening," *N Engl J Med*, vol. 355, pp. 1763-1771, 2006.

[121] J. W. Gurney, "Solitary pulmonary nodules: determining the likelihood of malignancy with neural network analysis," *Radiology*, vol. 196, pp. 823-829, 1995.

[122] A. K. Jain, *Fundamentals of digital image processing*. New Jersey: Prentice-Hall, 1989.

[123] J. Shiraishi, H. Abe, R. Engelmann, M. Aoyama, H. MacMahon, and K. Doi, "Computer-aided Diagnosis to Distinguish Benign from Malignant Solitary Pulmonary Nodules on Radiographs: ROC Analysis of Radiologists' Performance--Initial Experience," *Radiology*, vol. 227, pp. 469-474, 2003.

[124] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 4-37, 2000.

[125] H. T. Winer-Muram, "The Solitary Pulmonary Nodule," *Radiology*, vol. 239, pp. 34-49, 2006.

[126] C. I. Henschke, D. F. Yankelevitz, D. M. Libby, M. W. Pasmantier, J. P. Smith, and O. S. Miettinen, "Survival of patients with stage I lung cancer detected on CT screening," *New England Journal of Medicine*, vol. 355, pp. 1763-1771 2006.

[127] T. Sobue, N. Moriyama, M. Kaneko, M. Kusumoto, T. Kobayashi, R. Tsuchiya, R. Kakinuma, H. Ohmatsu, K. Nagai, H. Nishiyama, E. Matsui, and K. Eguchi, "Screening for lung cancer with low-dose helical computed tomography: Anti-lung cancer association project," *Journal of clinical oncology*, vol. 20, pp. 911-920, 2002.

[128] M. Das, G. Muhlenbruch, A. H. Mahnken, T. G. Flohr, L. Gundel, S. Stanzel, T. Kraus, R. W. Gunther, and J. E. Wildberger, "Small Pulmonary Nodules: Effect of Two Computer-aided Detection Systems on Radiologist Performance," *Radiology*, vol. 241, pp. 564-571, 2006.

[129] M. S. Brown, J. G. Goldin, S. Rogers, H. J. Kim, R. D. Suh, M. F. McNitt-Gray, S. K. Shah, D. Truong, K. Brown, and J. W. Sayre, "Computer-aided Lung Nodule Detection in CT Results of Large-Scale Observer Test," *Academic Radiology*, vol. 12, pp. 681-686, 2005.

[130] K. Awai, K. Murao, A. Ozawa, M. Komi, H. Hayakawa, S. Hori, and Y. Nishimura, "Pulmonary nodules at chest CT: Effect of computer-aided diagnosis on radiologists' detection performance " *Radiology*, vol. 230 pp. 347-352, 2004.

[131] F. Li, H. Arimura, K. Suzuki, J. Shiraishi, Q. Li, H. Abe, R. Engelmann, S. Sone, H. MacMahon, and K. Doi, "Computer-aided Detection of Peripheral Lung Cancers Missed at CT: ROC Analyses without and with Localization," *Radiology*, vol. 237, pp. 684-690, 2005.

[132] S. C. B. Lo, L. Y. Hsu, M. T. Freedman, F. Lure, and H. Zhao, "Classification of Lung Nodules in Diagnostic CT: An Approach Based on 3-D Vascular Features, Nodule Density Distributions, and Shape Features," *Proc. SPIE*, vol. 5032, pp. 183-189, 2003.

[133] Y. Matsuki, K. Nakamura, H. Watanabe, and T. Aoki, "Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution CT: Evaluation with receiver operating characteristic analysis," *American Journal of Roentgenology*, vol. 178, pp. 657-663, 2002.

[134] S. K. Shah, M. F. McNitt-Gray, K. R. De Zoysa, J. W. Sayre, H. J. Kim, P. Batra, A. Behrashi, K. Brown, L. E. Greaser, and J. M. Park, "Solitary pulmonary nodule diagnosis on CT Results of an observer study," *Academic Radiology*, vol. 12, pp. 496-501, 2005.

[135] F. Li, Q. Li, R. Engelmann, M. Aoyama, S. Sone, H. MacMahon, and K. Doi, "Improving radiologists' recommendations with computer-aided diagnosis for management of small nodules detected by CT," *Academic Radiology*, vol. 13, pp. 943-950, 2006.

[136] K. Awai, K. Murao, A. Ozawa, Y. Nakayama, T. Nakaura, D. Liu, K. Kawanaka, Y. Funama, S. Morishita, and Y. Yamashita, "Pulmonary Nodules: Estimation of Malignancy at Thin-Section Helical CT--Effect of Computer-aided Diagnosis on Performance of Radiologists," *Radiology*, vol. 239, pp. 276-284, 2006.

[137] C. E. Metz, "Some practical issues of experimental design and data analysis in radiological ROC studies," *Investigative Radiology*, vol. 24, pp. 234-245, 1989.

[138] D. D. Dorfman, K. S. Berbaum, and C. E. Metz, "ROC rating analysis: Generalization to the population of readers and cases with the jackknife method," *Investigative Radiology*, vol. 27, pp. 723-731, 1992.

[139] Y. Jiang, C. E. Metz, and R. M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," *Radiology*, vol. 201, pp. 745-750, 1996.

[140] Q. Li, F. Li, J. Shiraishi, S. Katsuragawa, S. Sone, and K. Doi, "Investigation of new psychophysical measures for evaluation of similar images on thoracic computed tomography for distinction between benign and malignant nodules," *Medical Physics*, vol. 30, pp. 2584-2593, 2003.