COMPUTER METHODS FOR PULMONARY NODULE CHARACTERIZATION FROM CT IMAGES

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Artit Chinwattana Jirapatnakul

January 2011

ABSTRACT

Computed tomography (CT) scans provide radiologists a non-invasive method of imaging internal structures of the body. Although CT scans have enabled the earlier detection of suspicious nodules, these nodules are often small and difficult to accurately classify for radiologists. An automated system was developed to classify a pulmonary nodule based on image features extracted from a single CT scan. Several critical issues related to performance evaluation of such systems were also examined.

The image features considered in the system were: statistics from the density distribution, shape, curvature, and boundary features. The shape and density features were computed through moment analysis of the segmented nodule. Local curvature was computed from a triangle-tessellated surface of the nodule; the statistics of the distribution of curvatures were used as features in the system. Finally, the boundary of the nodule was examined to quantify the transition region between the nodule and lung parenchyma. This was accomplished by combining the grayscale information and 3D model to measure the gradient on the surface of the nodule. These methods resulted in a total of 43 features. For comparison, 2D features were computed for the density and shape features, resulting in 26 features. Four feature classification schemes were evaluated: logistic regression, k-nearest-neighbors, distance-weighted nearest-neighbors, and support vector machines (SVM). These features and classifiers were validated on a large dataset of 259 nodules. The best performance, an area under the ROC curve (AUC) of 0.702, was achieved using 3D features and the logistic regression classifier.

A major consideration when evaluating a nodule classification system is whether

the system presents an improvement over a baseline performance. Since the majority of large nodules in many datasets are malignant, the impact of nodule size on the performance of the classification system was examined. This was accomplished by comparing the performance of the system with feature sets that included size-dependent features to feature sets that excluded those features.The performance of size alone, estimated using a size-threshold classifier, was an AUC of 0.653. For the SVM classifier, removing size-dependent features reduced the performance from an AUC of 0.69 to 0.61. To approximate the performance that might be obtained on a dataset without a size bias, a subset of cases was selected where the benign and malignant nodules were of similar sizes. On this subset, size was not a very powerful feature with an AUC of 0.507, and features that were not dependent on size performed better than size-dependent features for SVM, with an AUC of 0.63 compared to 0.52. While other methods have been proposed for performing nodule classification, this is the first study to comprehensively look at the performance impact from datasets with nodules that exhibit a bias in size.

BIOGRAPHICAL SKETCH

Artit Jirapatnakul received a B.S. in Electrical Engineering with a concentration in signal processing from the Pennsylvania State University in 2005. Since 2005, Artit has been working towards a Ph.D at Cornell University. His research has focused on computer methods to aid in the diagnosis of lung cancer, primarily by accurately characterizing and measuring pulmonary nodules from whole-lung, low-dose CT scans.

ACKNOWLEDGEMENTS

Completing the master's thesis has been a long and difficult process. I give much credit to my advisor, Dr. Anthony Reeves, for providing invaluable guidance and advice during the process. I would also like to thank the other members of my special committee, Dr. Peter Doerschuk and Dr. Noah Snavely, for their enlightening discussions and taking time out of their busy schedules to discuss my research. Finally, the collaborations I have had with several doctors at Weill Cornell Medical College, namely Dr. Henschke, Dr. Yankelevitz, and Dr. Cham, have given me insight into how to analyze medical images.

The time would have been much less enjoyable without the friendship of the members of the VIA group. Andrew Browder, Sergei Fotin, Brad Keller, Jaesung Lee, and Jeremiah Wala were always there to converse about academic and personal issues and could always be relied upon to lend a helping hand.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

According to the American Cancer Society, lung cancer is the leading cause of cancer deaths today and is expected to account for 159,390 deaths in 2009 [1]. Early detection and treatment of lung cancer has been shown to improve survival rates [2]. In its earliest manifestation, lung cancer typically presents as a pulmonary nodule which appears in an X-ray computed tomography (CT) image as an area of opacity in the lung parenchyma. The introduction of high-resolution, multi-row detector CT scanners which provide thin-slice images in a single breath-hold has allowed radiologists to detect more small nodules than previously possible with either chest radiographs or thick-slice CT. A majority of these small nodules are benign, but the status of these nodules is often difficult to ascertain, requiring additional physician follow up. This follow up typically consists of an additional CT scan at a later time to assess growth rate; a high growth rate is typically indicative of a malignant nodule. However, growth rate assessment requires a second CT scan which delays the true diagnosis and exposes the patient to a second, possibly unnecessary dose of radiation. Instead, we explore an automated method to diagnose cancer from a single low-dose CT scan used for screening by extracting and classifying various image features from the CT scan to assess the malignancy of a pulmonary nodule. In addition, we analyzed the effect of the underlying size-distribution of the nodules in the dataset on the reported performance of the system.

1.1 Problem Statement

Patients with early stage lung cancer often present no symptoms; thus, early cancers are typically found in CT or X-ray scans. Once a suspicious lesion is detected,

its malignancy may be determined by performing a biopsy or observing the lesion's growth rate. Both these techniques have undesirable characteristics; biopsy requires insertion of a needle into the patient's lung to remove tissue from the lesion which may cause complications such as a collapsed lung, and observing the growth of the lesion requires taking at least one additional scan, prolonging diagnosis and exposing the patient to additional radiation. As an alternative, the malignancy status of suspicious lesions is determined by analyzing features that are able to be assessed from a single CT scan, which we term *pulmonary nodule characterization*.

In pulmonary nodule characterization, various features are computed from the nodule and used to evaluate the probability of the nodule being malignant. Our method uses only features that can be computed from the nodule region on a CT image are considered. Several past studies have attempted to classify nodules using both features estimated by human observers and features computed by image analysis methods. These studies are described in further detail in Section 1.4.

Pulmonary nodule characterization using image features poses several challenges. In addition to computing the features themselves, biases exist in the the size distributions of malignant and benign nodules in the datasets used for system development which poses issues in training and evaluating classifiers that are unique to the task of nodule characterization. This *a priori* size information has been shown to be highly correlated with malignancy [3, 4], but in the evaluation of an automated characterization system, the relevant performance metric is not the absolute performance, but the improvement the system offers over the use of the *a priori* size information.

A major component of characterization systems is the specific classifier used. In general, there are parametric classifiers that assume the data fit a particular model, such as linear regression, and non-parametric classifiers that do not assume any *a*

priori model, such as neural networks. Parametric classifiers assume the data have some underlying probability distribution and, given this distribution, there should be an optimal decision surface to separate the data into different classes. As a result, for situations where the data fit these assumptions, parametric classifiers perform better than non-parametric classifiers since the model is already known. There are many situations where data do not have a known distribution. In these cases, non-parametric classifiers tend to work better than parametric classifiers due to the relaxation of assumptions of the probability distribution of the data. While this makes them more flexible, they are less powerful if there is a parametric model that fits the data, due to the need to learn both the model and the parameters of the model.

There have been many studies in the machine learning field comparing the performance of different classifiers on the same dataset, but there have been few published studies using different classifiers for pulmonary nodule characterization. Given the wide variety of image feature types and the often vague differences between benign and malignant nodules, a non-parametric classifier is likely to offer better performance than a parametric classifier.

In this study, a feature-based classification algorithm for pulmonary nodules in CT images was developed. The performance of this algorithm was evaluated for different types of classifiers. Finally, issues were identified with conventional evaluation methods due to the size bias of most pulmonary nodule datasets, and as a result, a new evaluation method that avoids size bias is proposed and evaluated.

1.2 Computed Tomography Imaging

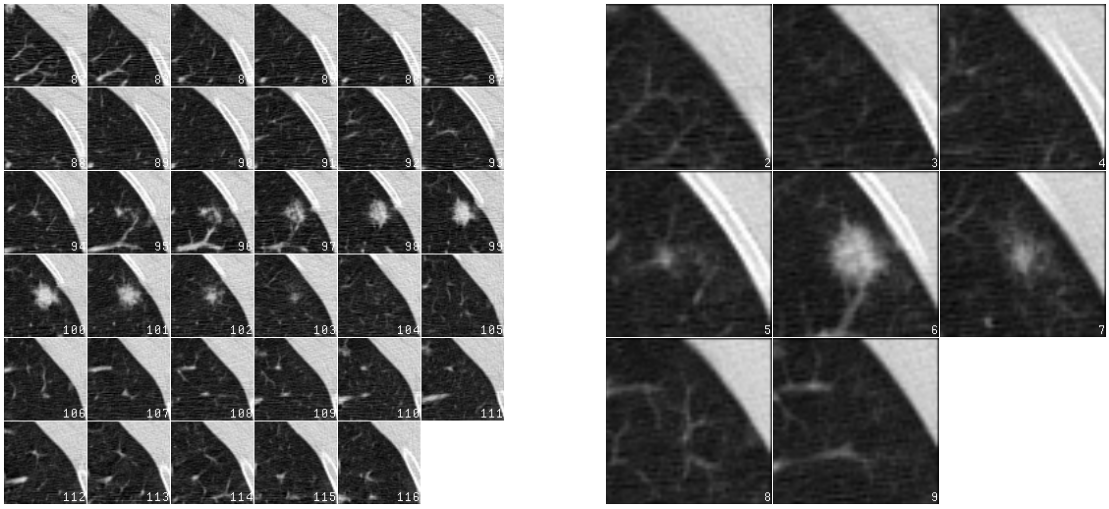
Computed tomography (CT) scanners enable radiologists to view internal body structures in three dimensions. CT scanners make use of an X-ray source and

detector that are rotated around the body. Images are created by reconstructing the X-ray projections. In the resulting CT images, the value of the voxel is related to the density of the tissue; CT scanners are calibrated so that, on the Hounsfield scale, a voxel value of 0 corresponds to water and -1000 to air [5]. Hounsfield units are defined by the following expression:

$$H_{\text{tissue}} = \frac{\mu_{\text{tissue}} - \mu_{\text{water}}}{\mu_{\text{water}}} \times 1000$$

where μ are the linear attenuation coefficients to X-rays. These linear attenuation coefficients quantify the reduction in intensity of an energy beam as it passes through a material.

The quality of a CT scan depends upon several scanner parameters; for the automated analysis of images considered in this work, the most important parameters are radiation dose, slice thickness, and field of view. Higher radiation doses allow for better quality images due to a higher signal to noise ratio, but this has to be balanced against the desire to limit radiation to the patient. The slice thickness specifies the width of each section along the axial direction of the scanner, which is determined by the speed of table movement, the width of each detector, and the amount of overlap between detectors. Thinner slice thickness scans have more detail than scans with thicker slice thickness, but the scan files are larger in size and have more noise than a thick slice scan using the same radiation dosage. Finally, the field of view controls the in-plane size of each voxel. In a whole-lung field of view, the entire lung is in view, resulting in an in-plane resolution of about 0.6 mm per voxel. If the radiologist knows the location of the nodule, a scan with a targeted field of view can be acquired of just the region of interest. These scans typically have an in-plane resolution of 0.18 mm. Although targeted scans have a higher physical resolution, the location of the nodule needs to be known, and thus are not useful for finding new nodules.



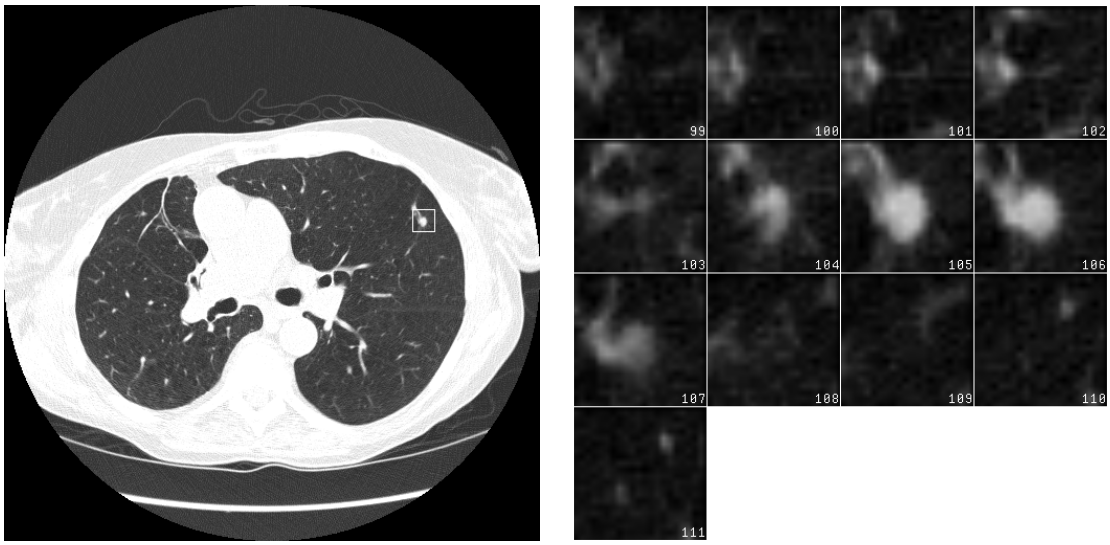
(a) (b)

Figure 1.1: Several slices from a CT scan of a solid pulmonary nodule on a) 1.25 mm thick scan and b) 5.0 mm thick scan. Note that for ease of viewing, the scales are not the same between the two images.

1.3 Pulmonary Nodules on CT Images

In its earliest manifestation, lung cancer often presents as a pulmonary nodule; however, not all pulmonary nodules are malignant – some may be caused by a variety of benign conditions such as inflammation of the airways. A pulmonary nodule appears on a CT scan as a high intensity object within the lung parenchyma which does not belong to any normal anatomical structures such as vessels or airways, as shown in Figure 1.2.

Pulmonary nodules may be categorized according to their density and surrounding attachments. Nodules with a high density, called "solid nodules", have an opaque appearance on CT scans, while nodules with a low density, "non-solid nodules", have a more ill-defined appearance. Nodules with both solid and non-solid components are called "part-solid". The term "subsolid" is often used to refer to both non-solid and part-solid nodules. Examples of these nodules are shown in Figure 1.3. In addition to exhibiting different densities, nodules may either be iso-

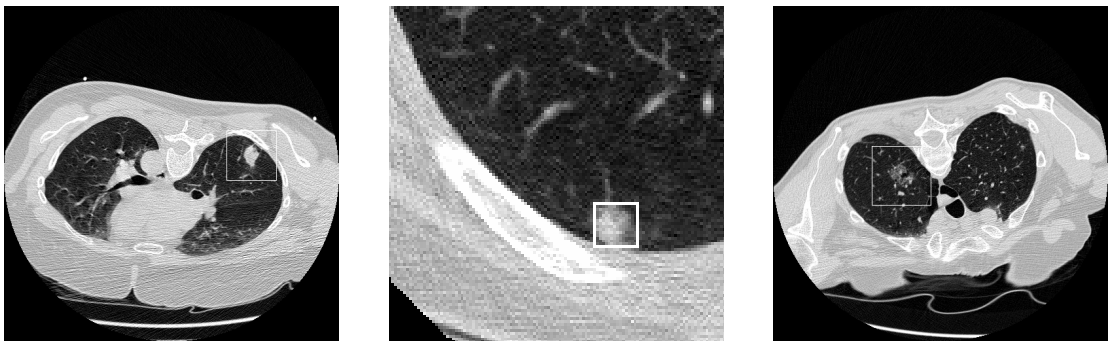


(a) (b)

Figure 1.2: Small solid pulmonary nodule on a) single slice of a CT scan and b) several slices of a CT scan in a small region of interest

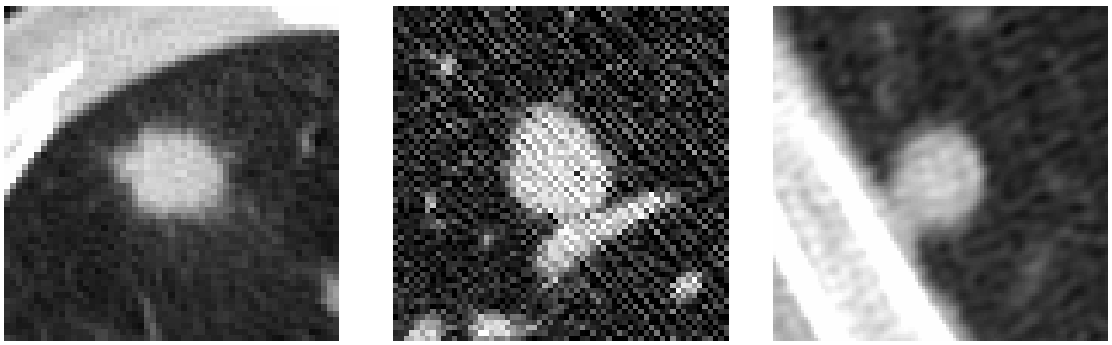
lated in the lung parenchyma or attached to other structures. Isolated nodules are not attached to any other high-intensity structures and are the easiest nodules to segment. Attached nodules may be attached to either blood vessels or airways, and juxtapleural nodules are attached to the chest wall. Enlarged images of the central slices of isolated, juxtapleural, and attached nodules are shown in Figure 1.4.

Only solid and part-solid pulmonary nodules of the three attachment types (isolated, attached, and juxtapleural) were considered in the development of the algorithm, since they comprise the majority of nodules detected during screening and a substantial portion of malignant nodules. A study by Henschke et al. (2002) found that 88.0% (205/233) of the 233 nodules identified during baseline scans in their screening study were solid, while 12% (28/233) were non-solid nodules [6]. A majority of the malignant nodules, 82.8% (24/29), were solid or part-solid. Furthermore, they found that the malignancy types for the subsolid nodules were different than the solid nodules. The predominant malignancy types for subsolid nodules were bronchioloalveolar carcinoma or adenocarcinoma with



(a) (b) (c)

Figure 1.3: Examples of a) solid, b) part-solid, and c) non-solid nodules on a single slice of a CT scan, with the nodule indicated by a white box.



(a) (b) (c)

Figure 1.4: Central slices of regions containing a) isolated, b) attached, and c) juxtapleural nodules

bronchioloalveolar features compared to other subtypes of adenocarcinoma found in solid nodules. This may indicate that subsolid nodules require different sets of features than solid nodules.

1.4 Previous work

Many studies have been made to accurately characterize pulmonary nodules from a single scan. These studies can be divided into two groups: studies that rely on human observations of nodule features and studies using computer methods of image analysis to extract features. Both groups of studies use similar techniques for performing classification. Typically, features are analyzed to determine which have the most discriminating power and the relevant features are used in a classification algorithm. In this section, studies that use human observable features are first described, followed by a review of previously published computer methods for pulmonary nodule characterization.

1.4.1 Characterization by Human Observable Features

Many attempts have been made to establish criteria, based on image features, for accurately evaluating the malignancy status of pulmonary nodules by correlating radiologic features with malignancy. One of the most basic image features that can be measured from a CT scan are the density of voxels within the nodule region. An early study by Siegelman et al. (1980) [7] found that using a representative CT number (Hounsfield unit value) from the mean of 32 contiguous voxels on the single slice with the highest CT number was a good indication of malignancy. In their study, of the 45 solid pulmonary nodules under 2 cm with CT numbers below 146 HU, 37 (82.2%) nodules were malignant. The study also found that the

distribution of densities was different between benign and malignant nodules, with benign nodules having either an even distribution of density or concentrated in the center, while in malignant nodules, voxels with the highest attenuation were along the edges of the nodules. While this study used a limited number of nodules (91) and used thick slice scans (10 - 12 mm), the results suggest that density features are useful for differentiating benign and malignant nodules.

Another early study by Gurney (1993) [8] examined previously published literature to determine what radiologic and clinical features were useful for distinguishing malignant nodules. He performed Bayesian analysis to derive likelihood ratio for features from previously published literature and found that several radiologic features were found to have a high likelihood of malignancy: size, edge characteristics, contour, calcification, growth rate, location, and cavitation. Gurney et al. (1993) [9] applied the likelihood ratios to a later study using six radiologists on 66 pulmonary nodules. Four radiologists estimated the probability of malignancy, while two radiologists evaluated the nodules according to the radiologic and clinical features used in the earlier study [8]. The readers using Bayesian analysis with the previously computed likelihood ratios performed better than radiologists alone, with the readers using Bayesian analysis misclassifying fewer malignant nodules as benign (6.5) than the expert readers (16.5). While these studies used manually determined features and clinical history, they show that it was possible to use nodule features and statistical analysis to improve diagnostic performance.

Later studies examined the performance of morphological features often used by radiologists. A study by Seemann et al. (1999) [10] examined several radiologist-determined features in their dataset of 23 benign and 81 malignant solid pulmonary nodules imaged on high-resolution CT scans of 1 mm slice thickness. Most of the features were categorical, for example, radiologists were asked to determine if the

appearance of the edge of a nodule was either smooth or indistinct. The following features were significantly different between benign and malignant nodules ($p < 0.01$): presence of ground-glass attenuation adjacent to the nodule, presence of spicules, length of spicules, bronchus sign, vessel sign, pleural retraction, and circumscribed thickening of the visceral pleura. Based on these features, a sensitivity of 91.4% and a specificity of 56.5% was obtained. While the performance was relatively high, the dataset was small, and all the features were determined by radiologist review. Takashima et al. (2003) [11] also used several radiologist-determined features to classify 25 malignant and 40 benign nodules. The best performance at 100% specificity was achieved with only two features – polygonal shape and a three-dimensional ratio of greater than 1.78. The highest sensitivities of 63% and 60% for both reviewers was achieved using a combined criterion of a predominately solid nodule and peripheral subpleural nodule or polygonal shape or the three dimensional ratio. Polygonal shape was also found to be a significant feature in a study by Li et al (2004) [12], along with a smooth or somewhat smooth margin. Their study had a large number of nodules, with 222 suspicious nodules detected during screening on thin-section CT scans. Polygonal shape was found in only 7% of malignant nodules compared to 35% of benign nodules, and a smooth or somewhat smooth margin was found in 0% of malignant versus 63% of benign nodules.

Although all the works described so far use features manually determined by radiologists, they suggest features that may be useful to extract using automated methods. Several examples of automated methods are described in the following section.

Table 1.1: Summary of features used in previous work

Author	Year	2D/3D	Size	Density	Shape	Texture	Margin	Other
Kawata [13]	2001	3D	N	Y	Y	N	N	Y
Aoyama [14]	2003	2D	Y	Y	Y	N	Y	Y
Shah [15]	2005	2D	Y	Y	Y	Y	N	Y
Shah [16]	2005	3D	Y	Y	Y	N	N	N
Suzuki [17]	2005	2D	Y	Y	N	N	N	N
Way [18]	2006	3D	Y	Y	Y	Y	N	N
This thesis	-	Both	Y	Y	Y	N	Y	N

1.4.2 Characterization by Algorithmic Image Features

Most methods for automated nodule characterization follow a similar feature characterization scheme. Various features are extracted from a large, documented database of nodules in CT images. These features are often analogs to the human observable features described in the previous section. Feature selection is usually performed to reduce the number of features to prevent overfitting. A classifier is trained on the database of nodules, usually using a leave-one-out methodology to make maximal use of the typically very small number of cases in the database. Individual methods differ in the details of feature extraction, feature selection, and classifier. This section will overview general categories of features used by many automated characterization systems then discuss in further detail several selected systems of interest.

There are a variety of features used by automated classification systems. These features can be roughly divided into five categories: density, shape, size, texture, and margin or edge features. Further, these features may be computed in two-dimensions on a single image slice through the center of the nodule or in three-dimensions across all the images on which the nodules appear. A summary of the features in six selected works discussed in this section is presented in Table 1.1.

Density features are computed from the attenuation values from the CT scan

in the region inside the nodule. Typically, several statistics are computed from these values, including mean, minimum and maximum, median, mode, variance, skewness, and kurtosis. Shape features, sometimes called morphological features, are descriptors obtained from the boundary of the segmented nodule. This includes compactness, sphericity, extent ratios, and curvature features.

Size features include volume, diameter, and surface area. It is interesting to note that some features from the other categories are size-dependent, for example curvature. Curvature is defined as the rate of change of the surface normal with respect to the surface length and is described in more detail in Section 2.2.2; for the two-dimensional case of a circle, the curvature is the reciprocal of a circle, and thus becomes smaller with increasing size. While in this work, curvature is normalized so that a sphere has a curvature of 1, it is difficult to remove the size dependence of other features such as the volume to surface area ratio. Furthermore, the size of a nodule and the resolution of the CT scan affect the computation of features due to factors such as the partial voxel effect.

The non-uniformity of tissue density can be measured using texture features, with the idea that malignant tissue tends to have a more irregular density distribution than benign tissue. Margin features measure the abruptness of the transition from the nodule to the lung parenchyma; these features may be computed from the gradient of the image.

A summary of the features used in several methods reviewed in this section is given in Table 1.1. Particularly noteworthy are the features most commonly used by the systems described in the literature – all the systems use density features, and most use size and shape features, all of which have been shown to be good predictors of malignancy. Few methods use texture, and only one other system uses margin features. The performance of these methods are summarized in Table

Table 1.2: Summary of reported performance for published characterization systems

Lead Author	Year	# benign	# malignant	AUC
Kawata	2001	95	33	0.87
Aoyama	2003	413	76	0.85
Shah	2005	16	19	0.92
Shah	2005	33	48	0.92
Suzuki	2005	413	76	0.88
Way	2006	52	44	0.83
This thesis	-	92	167	0.69

1.2, and the results obtained in this thesis are included in the table for comparison; the method and results will be fully described in Chapter 2. For each method, the number of benign and malignant nodules used to evaluate the methods is given along with the area under the ROC curve (AUC). Additional information about the AUC is provided in Section 2.5.2.

The best performing method on a large dataset was the method proposed by Suzuki et al. (2005) [17]. Their method utilizes pixel values in a local region of interest in a CT image in conjunction with a massively trained artificial neural network (MTANN) to distinguish between malignant and benign nodules. For training, their targets were 2D Gaussian functions for malignant nodules and 0 for a benign nodule. Since their system relies solely on pixel values in a region of interest, no segmentation is required, which is an advantage for complex nodules. The researchers reported a sensitivity of 100% with a specificity of 48%, with an area under the ROC curve (AUC) of 0.88 on their dataset of 413 benign and 76 malignant nodules which contained both solid and subsolid nodules. Of these nodules, 10 malignant and 60 benign nodules were used for training, with the rest of the nodules available for evaluating the system. An earlier study by Aoyama et al (2003) [14] also used neural networks on the same dataset, but utilized 41 features extracted from regions of interest containing a nodule. The features included shape,

gradient, density, and histogram features. The effective diameter of the nodule was included among the features; the authors reported an AUC of 0.85 using multiple slices.

Aside from neural networks, other popular classifiers include logistic regression and linear discriminant analysis (LDA) [19]. Kawata et al. (2001) [13] compute curvature to measure the surface irregularity of nodules and histogram features. They used a linear discriminant function and achieved an AUC of 0.87 for their automated method on a separate test set. Another study by Way et al. (2006) [18] also used a linear discriminate analysis classifier. In their study, 3D active contours were used to segment a dataset of 44 malignant and 52 benign nodules. Morphological, gray-level, and texture features were extracted from the segmented nodule, and their system achieved an AUC of 0.83. Unlike other studies, Way et al. suggest that texture features might be useful for the classification of lung nodules. A pair of studies by Shah et al. (2005) [15, 16] also used linear discriminant analysis. In the first study [15], several two-dimensional features, including size-based features, were extracted from a region of interest for each nodule. Several classifiers were tested, including a LDA classifier, a logistic regression classifier, a decision tree, and a quadratic discriminant analysis classifier. Using LDA, they achieved their best performance with an AUC of 0.92. In the second study [16], the authors used 3D features and evaluated their method on a different, larger dataset. The features included density, size, shape, and enhancement features. As this dataset used scans with contrast, enhancement features comparing pre- and post-contrast scans could be extracted. The authors again achieved an area under the ROC curve of 0.92, but in this study with a logistic regression classifier.

Several computer aided diagnosis (CAD) methods have been evaluated by measuring their impact on affecting a radiologist's decision rather than from by their

performance alone. Several studies have used computer aided diagnosis (CAD) algorithms to improve the classification performance of radiologists. These studies report better performance when the radiologist is aided by the CAD scheme than either than CAD system or the radiologist alone [20, 21, 22]. Li et al (2004) [20] found statistically significant improvement for radiologists with the use of their CAD scheme (from an AUC of 0.785 to 0.853), and Awai et al [22] also observed improvement in the AUC achieved by the radiologists alone (0.843) compared to the radiologists assisted by the CAD scheme (0.924).

1.5 Outline

The primary goal of this work is the development of an automated system for the characterization of solid pulmonary nodules and an analysis of the effect of the underlying size-distribution of nodules in the dataset on the reported performance of the system. The nodule characterization system is fully described in Chapter 2, including discussion on the features used in the system and the classifiers. Chapter 3 contains an analysis of the effect of the size-distribution of the nodules in the dataset, comparing the results of the system using feature sets including and excluding size-dependent features on the full dataset and a subset of the cases selected to eliminate any size-distribution bias. A concluding discussion and suggestions for future work are explored in Chapter 4.

CHAPTER 2

THE NODULE CHARACTERIZATION SYSTEM

Most systems for performing image feature classification may be divided into three primary stages, as shown in Figure 2.1. Since features are extracted from segmented nodule images, the first stage segments the nodule from surrounding structures on the CT scan. In the second stage, features are extracted from the segmented image and normalized if necessary. The third and final stage classifies the nodule as malignant or benign. These stages are described in further detail in the sections below.

2.1 Pulmonary Nodule Segmentation

The first stage of the classification system is segmentation of the nodule. Segmenting the nodule separates voxels belonging to the nodule from voxels belonging to surrounding structures and lung parenchyma. Segmentation is performed using an algorithm previously developed by Reeves et al. (2006) [23]; a flowchart of the algorithm is presented in Figure 2.2. To summarize, an approximate size and location for the nodule in the CT image is computed based on an initial user-specified seed point using a Gaussian-weighted spherical template-matching method. From this size and location, a region of interest (ROI) is selected around the nodule, as shown

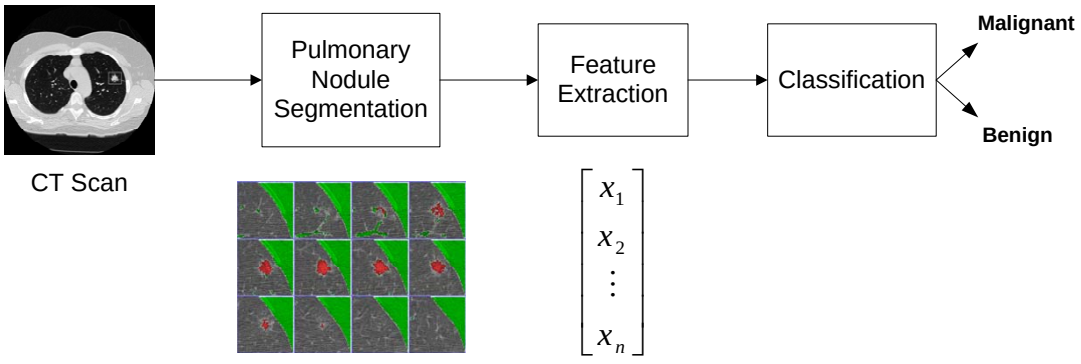


Figure 2.1: Overview of the nodule characterization system.

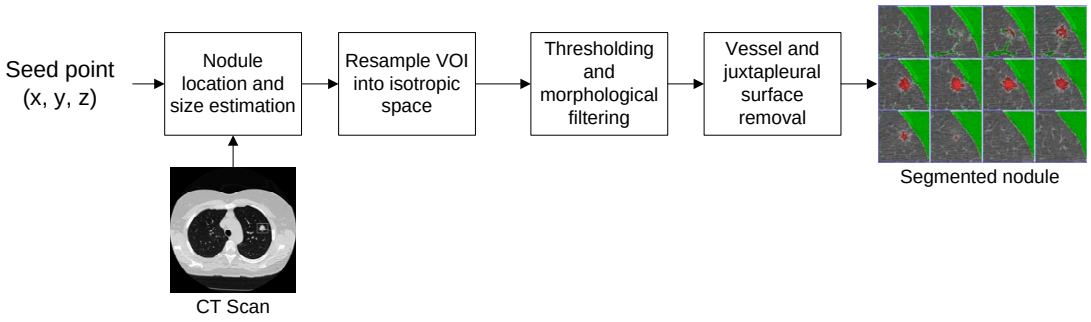


Figure 2.2: Flowchart of the pulmonary nodule segmentation algorithm

in Figure 2.3a. The ROI is re-sampled into isotropic space by trilinear interpolation and a threshold is applied to obtain a binary image. Morphological filtering using an algorithm by Kostis et al. (2003) [24] is performed to remove any attached vessels, followed by juxtapleural detection and, if necessary, segmentation using an iterative algorithm that separates the nodule from the pleural surface. The result of this algorithm is a binary segmented image of the nodule as shown in Figure 2.3b. A gray-scale image for density analysis is obtained by using the binary image as a mask on the selected region of interest; an example is shown in Figure 2.3c. A three-dimensional light-shaded visualization is shown in Figure 2.3d. These images are used in the feature extraction stage described in the following section.

2.2 Image Features

There have been several studies regarding what features best differentiate malignant from benign nodules based on radiologists' observations. In an early paper by Siegelman et al. (1986) [25], the criteria for benign pulmonary nodules included a representative CT number of at least 164 HU and smooth margins. Zwirewich et al. (1991) [26] found that size (mean diameter), coarse spiculation, and lobulation were good indicators of malignancy. The authors also found that homogeneous attenuation within the nodule occurred with significantly more frequency

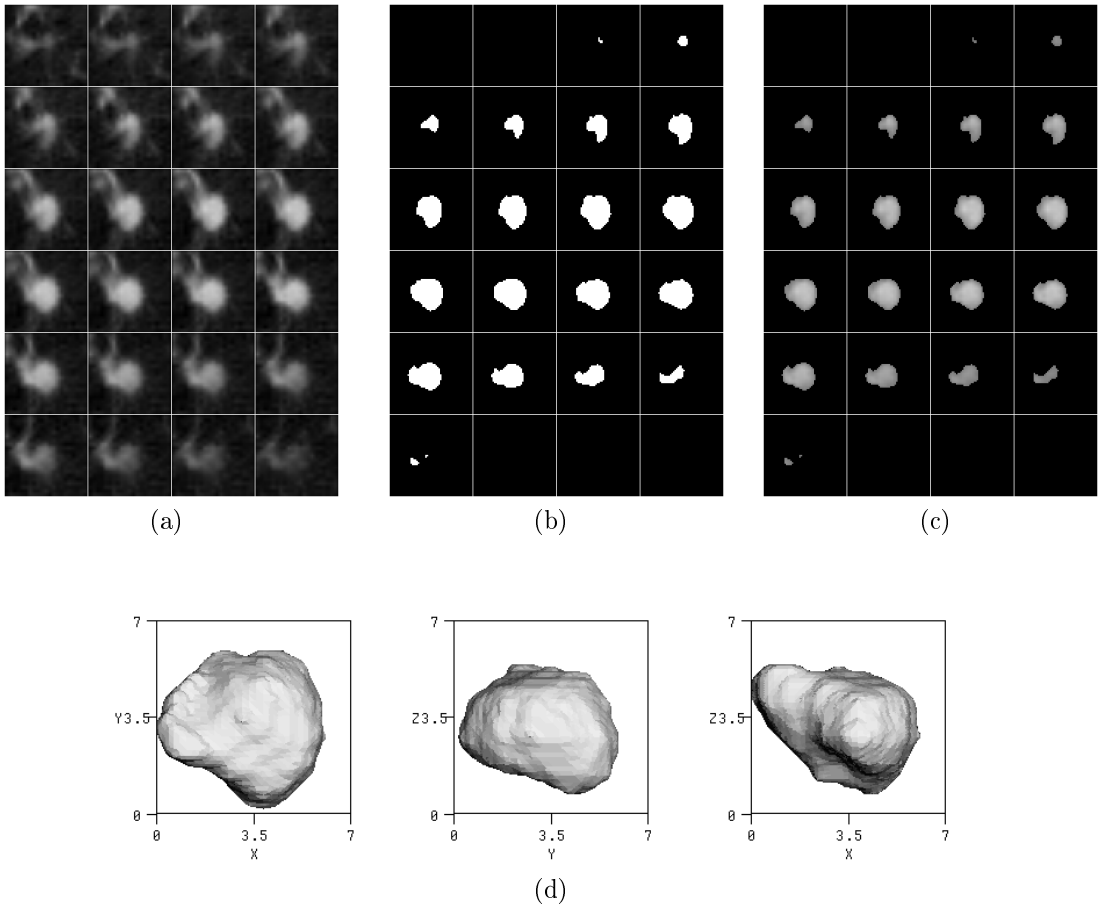


Figure 2.3: Segmentation of nodule, starting with a) region of interest and resulting in b) a binary segmented image and c) grayscale segmented image. d) A 3D light shaded visualization of the axial, sagittal, and coronal views left to right respectively.

among benign lesions compared to malignant lesions. More recent studies using higher resolution CT scans found similar features that were significantly different between malignant and benign nodules, such as the presence of spicules, the presence of ground-grass attenuation, polygonal shape, three-dimensional size ratios, and irregular margins [27, 11, 12, 28]. Although the features noted here are not a comprehensive list of features studied in the literature, they serve to suggest some of the features that should be included in an automated nodule classification system. From the segmented binary and grayscale images, 2D and 3D morphological, shape, and CT density features were computed using moment analysis, curvature estimation, and analysis of CT gray-level data.

2.2.1 Moment analysis

Moments have been used to perform shape analysis in computer vision and medical imaging algorithms. In this paper, 2D and 3D geometric and densitometric moments were computed according to the method described by Reeves et al [29]. Several descriptors of the general nodule shape can be easily derived from the moments, including compactness, sphericity, and aspect ratios. These measures were described by Kostis (2001) [30]. The conventional definition of a three-dimensional moment of order (p+q+r) of a function f(x,y,z) is

$$m_{pqr} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q z^r \, f(x,y,z) \, dx \, dy \, dz$$

where $f(x,y,z)$ is a continuous function of three dimensions. In a sampled 3D image, the moment definition becomes

$$m_{pqr} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} \sum_{z=0}^{L-1} x^p y^q z^r \, v(x,y,z)$$

where $v(x, y, z)$ is a discrete function of size (M $\times$ N $\times$ L). The function $v(x, y, z)$ can be of two types: binary or grayscale. If $v(x, y, z)$ is binary, it takes on a value of either 0 or 1, which would be useful for applications where only the shape is of importance. These are referred to in this paper as geometric moments. If the intensity distribution is of interest as well, $v(x, y, z)$ is continuous with a range of values corresponding to the pixel intensities in the image; this corresponds to densitometric moments. Note that density values are only considered for those pixels that are determined to be within the nodule from the segmentation performed in Section 2.1.

A complete moment set of order n, where $n = p+q+r$, is defined to be the set of all moments with order n and lower. A conventional set of moments are sensitive to scale, translation, and rotation of the image. For the purpose of performing shape analysis, the set of moments should be invariant to these transformations. To accomplish this, several normalizing operations described by Reeves et al [29] are applied to the set of moments, resulting in a set of standard moments which are normalized with respect to scale, translation, and rotation.

As with two-dimensional moments, various orders of moments have physical meaning. For example, the zeroth-order moment,

$$m_{000} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} \sum_{z=0}^{L-1} v(x, y, z)$$

gives the number of voxels comprising the object, from which we can compute the volume,

$$\text{Volume} = m_{000} \cdot V_{\text{voxel}}$$

where $v(x, y, z)$ is binary so that it has the value 1 within the nodule and 0 outside the nodule and the voxel size is computed from the resolution of the scan, $V_{\text{voxel}} = x_{\text{res}} \cdot y_{\text{res}} \cdot z_{\text{res}}$ A similar expression can be written for the area in the 2D case:

$$\text{Area} = m_{00} \cdot A_{\text{pixel}}$$

where m_{00} is the 2D moment without z, and $A_{\text{pixel}} = x_{\text{res}} \cdot y_{\text{res}}$. The mass of the nodule is simply m_{000} computed with $v(x, y, z)$ equal to the density of the pixel.

Higher-order moments give the center of mass, principal axes, and moments of inertia. From these metrics, other ad hoc features such as aspect ratios can be computed as well as compactness and sphericity.

A simple descriptor of the shape of a nodule are its aspect ratios. These are ratios of the dimensions of the segmented nodule volume, as computed by the ellipsoid of inertia. The ellipsoid of inertia is determined in the process of computing the standard moments. In order to create the set of standard moments, the orientation of the object must be determined. This is accomplished by solving the following eigenproblem:

$$Ax = \lambda x$$

where

$$A = \begin{pmatrix} m_{200} & m_{110} & m_{101} \\ m_{110} & m_{020} & m_{011} \\ m_{101} & m_{011} & m_{002} \end{pmatrix} \tag{2.1}$$

The eigenvectors (V_x, V_y, V_z) from the solution of this problem form an orthonormal basis which will point in the directions of each of the principal axes of the object. The standard orientation is defined such that the major principal axis (V_x) is aligned with the x-axis, the intermediate principal axis (V_y) is aligned with the y-axis, and the minor principal axis (V_z) is aligned with the z-axis. If the eigenvalues of the system given in Equation 2.1 are sorted such that

$$\lambda_0 \geq \lambda_1 \geq \lambda_2$$

then the lengths of the principal axes are

$$\text{length} = |V_x| = 2\sqrt{\lambda_0} \cdot \sqrt[3]{\frac{3V}{4\pi\sqrt{(\lambda_0\lambda_1\lambda_2}}}$$

$$\text{width} = |V_y| = 2\sqrt{\lambda_1} \cdot \sqrt[3]{\frac{3V}{4\pi\sqrt{(\lambda_0\lambda_1\lambda_2}}}$$

$$\text{height} = |V_z| = 2\sqrt{\lambda_2} \cdot \sqrt[3]{\frac{3V}{4\pi\sqrt{(\lambda_0\lambda_1\lambda_2}}}$$

From these lengths, the aspect ratios can be computed:

$$A_{lh} = \frac{\text{length}}{\text{height}}$$

$$A_{lw} = \frac{\text{length}}{\text{width}}$$

$$A_{wh} = \frac{\text{width}}{\text{height}}$$

Another shape measure of interest is the compactness of a nodule, defined as the ratio of the size of a shape to its surface area. In the three-dimensional case, compactness is expressed as:

$$\text{Compactness}_{\text{3D}} = \frac{6\sqrt{\pi} \cdot V}{S^{3/2}}$$

where V is the volume of the segmented nodule and S is the surface area. In the two-dimensional case, compactness is expressed as a ratio of the area to the perimeter:

$$\text{Compactness}_{\text{2D}} = \frac{4\pi \cdot \text{Area}}{\text{Perimeter}^2}$$

In both cases, constants are introduced so that the compactness of a sphere and circle are equal to 1. Similar measures to compactness are sphericity and circularity for three- and two-dimensions respectively. While both compactness and sphericity compare the volume of the object to its surface area, sphericity also considers the major to minor aspect ratio, which results in a lower sphericity for shapes that significantly differ from a spherical shape.

$$\text{Sphericity} = \frac{\text{Compactness}_{\text{3D}}}{A_{lh}}$$

$$\text{Circularity} = \frac{\text{Compactness}_{\text{2D}}}{A_{lw}}$$

Density statistics are computed using the central statistical moments. These moments are summations of powers of the voxel density values normalized to the mean value, $\overline{\mu}$.

$$\mu_p = \sum_0^{N-1} (v(x,y,z) - \overline{\mu})^p$$

where N is the number of voxels. Thus, the number of voxels is equal to the zeroth-order moment

$$\mu_0 = N$$

and the mean voxel density is equivalent to the first-order moment divided by the first-order moment

$$D_\mu = \frac{1}{N} \sum_0^{N-1} v(x,y,z)$$

and the variance is the second-order moment divided by the zeroth-order moment.

$$D_{\sigma^2} = \frac{\mu_2}{\mu_0} = \frac{1}{N} \sum_0^{N-1} (v(x,y,z) - \mu)^2$$

The standard deviation is defined to be the square root of the variance.

$$D_\sigma = \sqrt{D_{\sigma^2}}$$

Two higher order measures quantify the shape of the distribution, skewness and kurtosis. Skewness measures the shift of the distribution above or below the mean,

$$D_{\text{skewness}} = \frac{\mu^3}{D_\sigma^3} = \frac{\sum_0^{N-1} (v(x,y,z) - \mu)^3}{D_\sigma^3}$$

and kurtosis measures the "peakiness" of the distribution. A narrower distribution has a higher kurtosis value.

$$D_{\text{kurtosis}} = \frac{\mu^4}{D_\sigma^4} - 3 = \frac{\sum_0^{N-1} (v(x,y,z) - \mu)^4}{D_\sigma^4} - 3$$

Note that expression is normalized by subtracting three such that the kurtosis of a normal distribution is 0.

There were several secondary density metrics defined to quantify the regularity of the nodule density distribution compared to a uniformly dense sphere. The eccentricity, ε, of the density distribution measures the displacement between the geometric and densitometric centers of mass (CoM), and is defined as

$$\varepsilon = \text{dist}(\text{Co}M_{geom.}, \text{Co}M_{dens.})$$

where dist indicates the Euclidean distance. To make the metric size-invariant, the eccentricity may be normalized by an estimate of the nodule radius.

$$\hat{\varepsilon} = \frac{\text{dist}(\text{Co}M_{geom.}, \text{Co}M_{dens.})}{\sqrt[3]{V}}$$

Another measure of regularity is the density skew, ϕ_d, which measures the angle between the geometric and densitometric ellipsoids of inertia (EOI) according to the expression

$$\phi_d = \theta_{\text{geom.}} - \theta_{\text{dens.}}$$

where $\theta_{\text{geom.}}$ and $\theta_{\text{dens.}}$ are the orientation of the geometric and densitometric EOI respectively. This metric may not be stable as the nodule becomes more spherical, since the orientations are less stable. To attempt to address this, we want shapes that are more spherical, and therefore with more uncertainty in measurement, to have a smaller density skew. This is accomplished by normalizing by the sphericity.

$$\hat{\phi_d} = \frac{\theta_{\text{geom.}} - \theta_{\text{dens.}}}{\text{Sphericity}}$$

2.2.2 Surface curvature estimation

The margin of a malignant nodule may contain irregularities such as spiculation or lobulation. Such irregularities tend to result in an uneven surface, whereas benign nodules typically have smooth surfaces. The irregularity of the surface can be described through an analysis of the surface curvature.

Surface curvature is defined as the rate of change of the surface normal, ϕ, with respect to the surface length. In the two-dimensional case for a curve, this can be defined as the derivative of the normal vector with respect to the arc length.

$$\kappa = \frac{d\phi}{ds}$$

A straight line has a curvature value of 0, while a circle has a curvature of $\frac{1}{R}$ where R is the radius of the circle.

For measuring the curvature of pulmonary nodules in three-dimensions, we use a discrete piecewise linear model for the nodule surface described in previous work by Kostis (2001) [30]. Thus, the curvature can be estimated as the change in the surface normal between a particular vertex and all of the adjacent vertices. In contrast, a previous method of surface curvature estimation by Kawata et al. (1999) [31] used the values of the gray-level voxels directly; however, estimating the curvature from the gray-level voxels introduces errors due to the fact that voxels are rectangular approximations of the nodule surface.

To address this problem, curvature is estimated on a smoothed tessellated polygonal surface model of the nodule, as described by Kostis (2001) [30]. Similar algorithms, such as one proposed by Rusinkiewicz (2004) [32], have been used to computed curvature for colon polyp detection [33]. A diagram illustrating the curvature estimation using the piecewise linear model is shown in Figure 2.4. In Figure 2.4a, the curve is indicated by a gray dashed line and a piecewise linear model of the curve shown by a solid black line. As described above, curvature is related to the change in the surface normal, so surface normal vectors are drawn at two points on the curve and piecewise model in Figure 2.4b and c. To show the change in the surface normal vectors, they are drawn with the same origin in Figure 2.4d and e. Note that the angular difference between the surface normal vectors is very similar between the actual curve and the piecewise linear model of

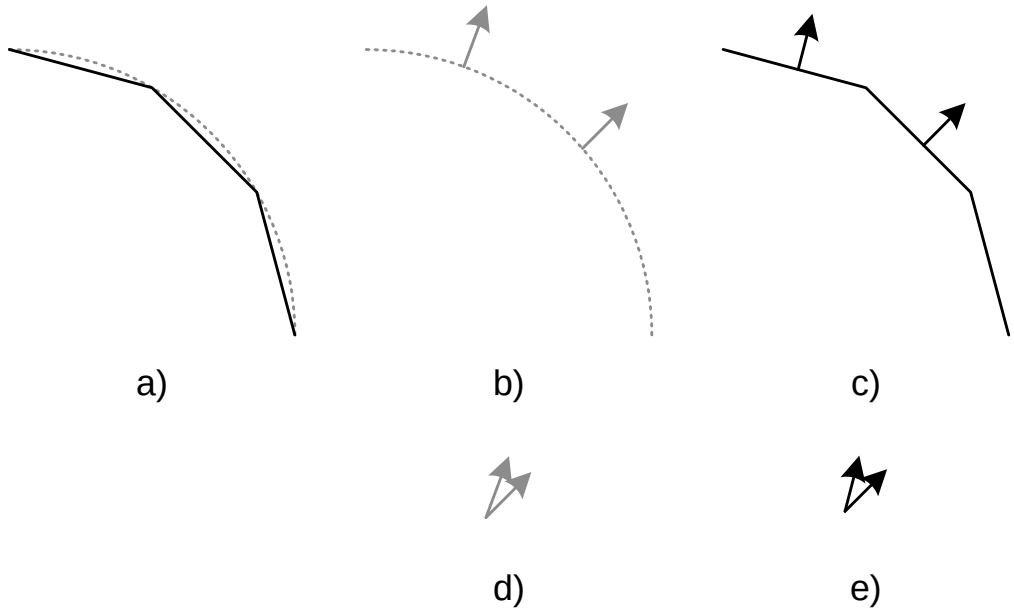


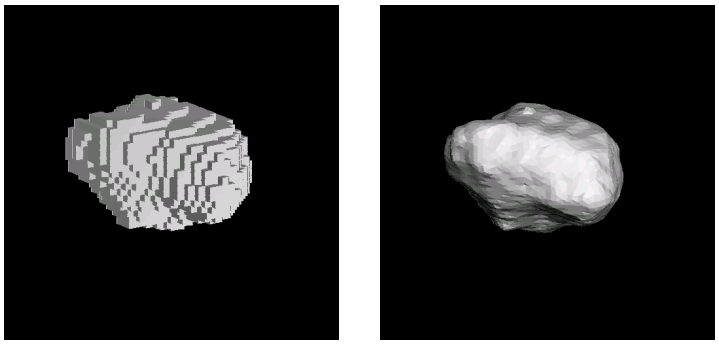
Figure 2.4: Illustration of surface curvature estimation for a 2D curve. In a), the curve (gray dashed line) is represented by a piecewise linear model (solid black line). The surface normals are labeled for b) the curve and c) the piecewise linear estimate. The normals are placed next to each other to indicate the angular difference in d) and e), and note that the differences are nearly the same.

the curve.

Marching cubes was used to generate a polygonal representation of the surface of the segmented nodule [34]. Since marching cubes was only used on binary segmented images, the algorithm could be simplified slightly by removing gradient computations; additional modifications prevented the generation of discontinuities in the tessellation and ensured a consistent ordering of the vertices of each triangle [30].

Due to the way marching cubes tessellates the surface of the segmented region, all triangles are located at angles that are multiples of 45°. To improve the surface representation, the polygonal tessellation was smoothed by replacing the location of a vertex by a weighted sum of neighboring vertices and itself.

The result of applying the marching cubes algorithm to a pulmonary nodule is shown in Figure 2.5, with the nodule surface estimated using voxels from the seg-



(a) (b)

Figure 2.5: Nodule surface represented using a) voxels and b) smoothed, tessellated polygonal surface. Curvature estimated from the tessellated polygonal surface is closer to the actual curvature of the nodule surface than the voxel representation of the surface.

mented image in Figure 2.5a and the surface estimated using a smoothed polygonal surface in Figure 2.5b. By using the smoothed polygonal surface model, we reduce the quantization error, ensuring a more accurate curvature estimate compared to a method based on just the voxels of the segmented image.

A problem that arises when considering regions of the nodule where attached structures, such as vessels, have been removed is that, in these regions, the curvature of the nodule surface is an artifact of the removal algorithm and should not be included in the curvature of the nodule. These regions are ignored by using the binary image of the removed structures as a mask on the polygonal model.

Once we have a 3D polygonal representation of the surface, the next step is to estimate the surface normals in order to compute the curvature. From the surface normals at each triangle, the average surface normal of each vertex can be computed. Finally, the curvature is computed as the average difference between the surface normals at each vertex. These steps are described in further detail below.

An example patch of a polygonal surface is illustrated in Figure 2.6. For each vertex in the polygonal representation, the triangles of which it is a member are

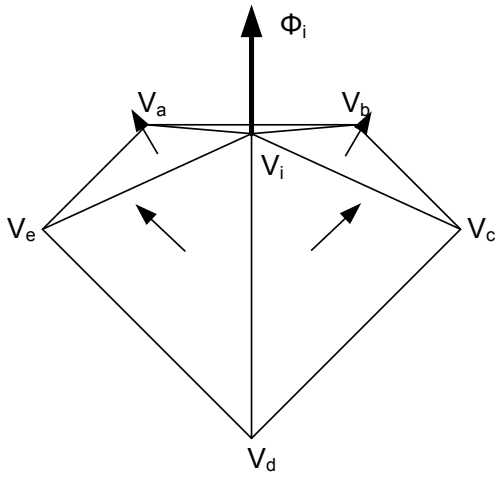


Figure 2.6: An example patch of a 3D polygonal tessellated surface with the surface normal vectors shown for each triangle and the surface normal at the vertex

determined through connectivity analysis. The surface normal for each triangle can be computed as the normalized cross product of two sides, according to the diagram in Figure 2.7:

$$N_i = \frac{\overrightarrow{V_iV_c} \times \overrightarrow{V_iV_d}}{\left|\overrightarrow{V_iV_c} \times \overrightarrow{V_iV_d}\right|}$$

Once the surface normal is computed for each triangle, the surface normal at each vertex can be computed. This is accomplished by averaging the surface normals of the triangles of which the vertex is a member:

$$\phi_i = \frac{\sum_{j=0}^{m} N_i}{|T|}$$

where $|T|$ is the number of triangles of which V_i is a member.

With the surface normals computed for each vertex, the curvature is computed by taking the angular difference between a vertex and an adjacent vertex. In this example, the angular difference between the surface normal vectors ϕ_i and ϕ_a is:

$$\theta_i = \cos^{-1}\left(\frac{\phi_i \cdot \phi_a}{|\phi_i|\,|\phi_a|}\right)$$

For each vertex, there are several adjacent vertices; to generate a single curvature estimate, the average curvature is computed for the vertex. In the example depicted

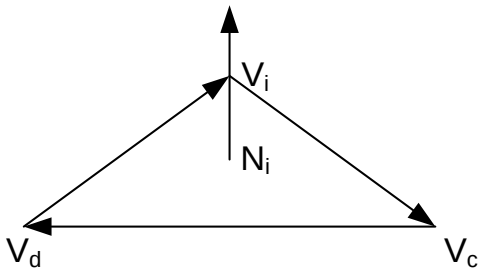


Figure 2.7: Diagram of surface normal calculation for a triangle

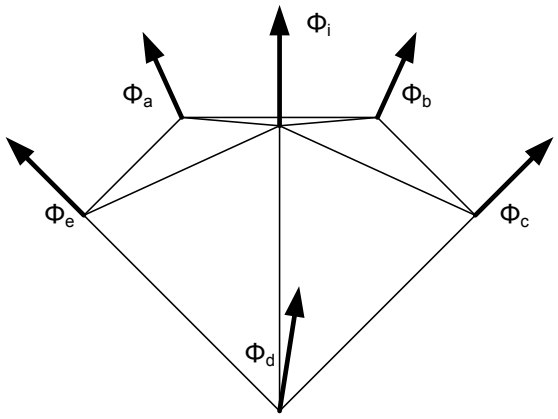


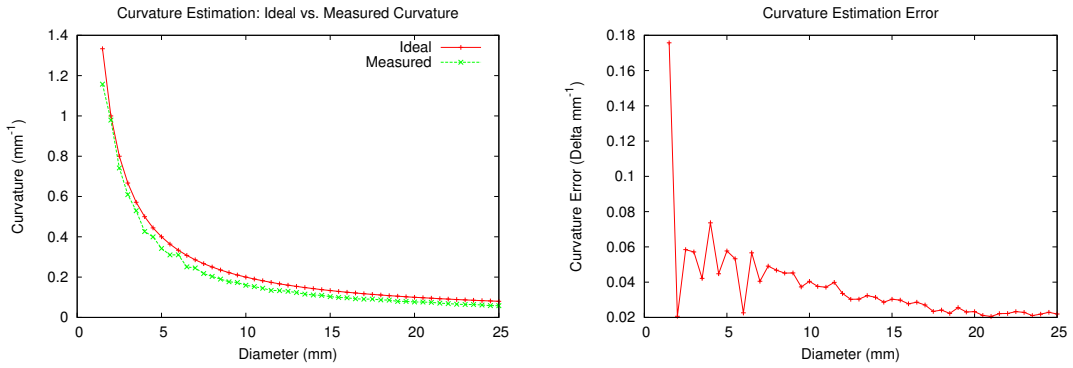
Figure 2.8: Curvature estimation from vertex surface normals

in Figure 2.8, the average curvature for vertex V_i would be computed as:

$$C_{V_i} = \frac{\sum_{m \in \{a,b,c,d,e\}} \cos^{-1}\left(\frac{\phi_i \bullet \phi_m}{|\phi_i||\phi_m|}\right)}{n}$$

where n is the number of adjacent vertices. Finally, each triangle in the polygonal representation is assigned a curvature value based on the average of the curvatures of the vertices which comprise the triangle. Basic statistics of the distribution of curvatures over the entire nodule surface were used as features.

An experiment was performed to quantify the error in curvature estimation. Synthetic images of spheres of diameters from 1.5 mm to 25 mm were generated and the curvatures measured using the algorithm described in this section. The curvature of a sphere is defined to be the inverse of its radius; thus, the error between the true curvature and estimated curvature can be computed. A plot of the ideal curvature compared to the measured curvature is shown in Figure



(a) (b)

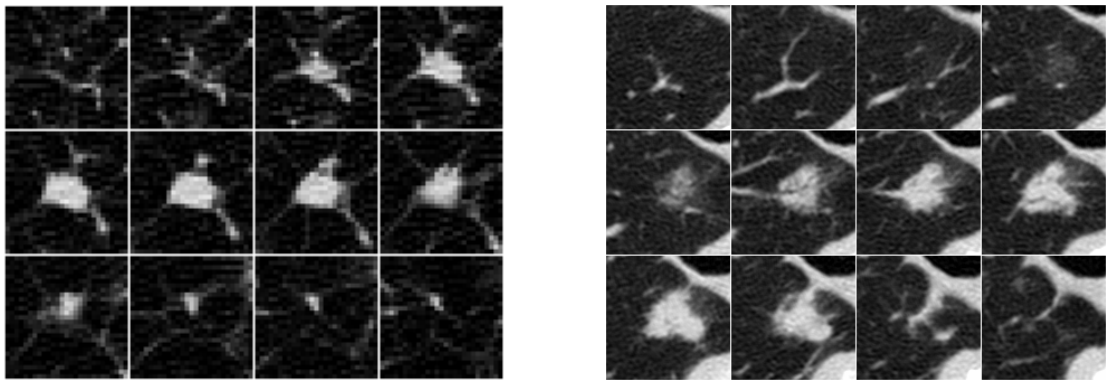
Figure 2.9: Plot of a) curvature estimation compared to ideal curvature value and b) curvature estimation error

2.9a and the error between the ideal and measured curvatures is show in the plot in Figure 2.9b. Note that the method has some slight estimation error, and the estimation error tends to decrease for larger spheres.

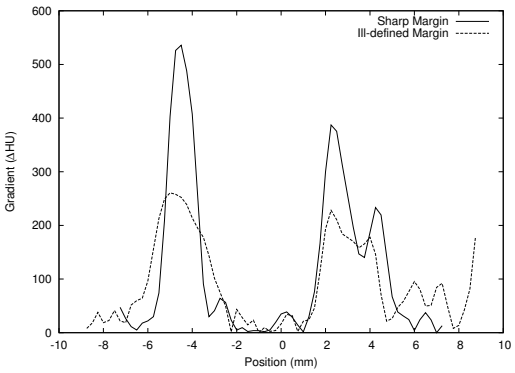
2.2.3 Margin analysis

The margin of a nodule is defined as the region along the boundary of the nodule and lung parenchyma. Nodule margins may be sharply demarcated or ill-defined, with an example of each in Figure 2.10a and 2.10b; previous work has suggested that some margin types are more correlated with malignancy than others [35].

To measure the nodule margin, the gradient was measured at the boundary between the nodule and lung parenchyma. A method developed by Monga and Deriche was used to compute the gradient [36]. In two-dimensions, the gradient is measured by determining the gradient in the x- and y-directions, which can then be used to compute the gradient in any direction. The gradient operator can be divided into a smoothing function, $l(z)$, and a derivative function $d(z)$. The optimal derivative function exhibits good localization, robust detection of edges,



(a) (b)



(c)

Figure 2.10: Examples of nodules with a) sharp margin and b) ill-defined margin with c) the gradients sampled along a horizontal ray through the center of each nodule on the central slice. Note that the nodule with a sharp margin in a) has a much higher maximum gradient in the gradient plot in c).

and has a single response to an edge, and was determined to be:

$$d(z) = -cze^{-\alpha|z|}$$

The smoothing function $l(z)$ is chosen to be the integral of the derivative function in order to provide for an efficient recursive implementation

$$l(z) = s(\alpha\,|z| + 1)e^{-\alpha|z|}$$

There are two convolution masks, one for the x- and y-directions, with the derivative function computed parallel to each direction and the smoothing function applied in the orthogonal direction:

$$\begin{aligned} X(i,j) &= d(i)l(j) \\ &= -cie^{-\alpha|i|}s(\alpha\,|j| + 1)e^{-\alpha|j|} \\ &= -ci(s\alpha\,|j| + s)e^{-\alpha(|i|+|j|)} \end{aligned}$$

$$\begin{aligned} Y(i,j) &= l(i)d(j) \\ &= -cj(s\alpha\,|i| + s)e^{-\alpha(|i|+|j|)} \end{aligned}$$

The constants c and s are fixed by the normalization requirements on $d(z)$ and $l(z)$. Let $d(n)$ be samples from $d(z)$ and $D(Z)$ its z-transform:

$$D(Z) = \sum d(n)z^{-n}$$

for $n = -\infty, \ldots, \infty$. Then the normalization requirement:

$$\{\sum f(n) \text{ for } n = 0, \ldots, \infty\} = -\{\sum f(n) \text{ for } n = -\infty, \ldots, 0\}$$

leads to

$$c = \frac{(1 - e^{-\alpha})^2}{e^{-\alpha}}$$

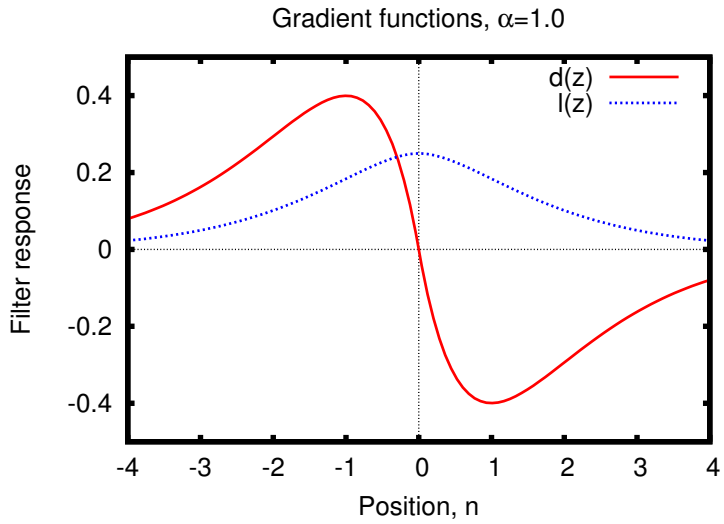


Figure 2.11: Plot of functions used for gradient estimation

and the requirement that

$$\{\sum l(n) \text{ for } n = -\infty, \ldots, \infty\} = 1$$

results in

$$s = \frac{(1 - e^{-\alpha})^2 \cdot \alpha^2}{1 + 2 \cdot \alpha \cdot e^{-\alpha} - e^{-2 \cdot \alpha}}$$

The gradient functions $d(z)$ and $l(z)$ are shown in Figure 2.11 for $\alpha = 1.0$ with the constants c and s computed as described.

This can be extended to the 3D case with three convolution masks:

$$\begin{aligned} X(i,j,k) &= d(i)l(j)l(k) \\ &= -cie^{-\alpha|i|}s(\alpha\,|j|+1)e^{-\alpha|j|}s(\alpha\,|k|+1)e^{-\alpha|k|} \\ &= -ci(s\alpha\,|j|+s)(s\alpha\,|k|+s)e^{-\alpha(|i|+|j|+|k|)} \end{aligned}$$

$$\begin{aligned} Y(i,j,k) &= l(i)d(j)l(k) \\ &= -cj(s\alpha\,|i|+s)(s\alpha\,|k|+s)e^{-\alpha(|i|+|j|+|k|)} \end{aligned}$$

$$
\begin{aligned}
Z(i,j,k) &= l(i)l(j)d(k) \\
&= -ck(s\alpha\,|i| + s)(s\alpha\,|j| + s)e^{-\alpha(|i|+|j|+|k|)}
\end{aligned}
$$

To convolve an image $I(i,j,k)$ with the mask $X(i,j,k)$, a derivative filter is applied in the x-direction followed by a smoothing filter in the y- and z-direction. The algorithm has a parameter, α, that controls the amount of smoothing applied to the image, which in turn controls the tradeoff between localization and noise suppression. Lower values of α cause more smoothing, decreasing localization, but suppressing more noise.

In the plot shown in Figure 2.10c, 2D gradients are sampled along a horizontal ray through the center of the nodule on the central slice. The nodule with a sharp margin has a higher gradient than the nodule with an ill-defined margin.

Although the boundary may be obtained from the segmented image, small errors in the exact location of the boundary, while having little effect on volume measurement, may alter the gradient by a large amount, as evidenced by the plot of the gradient distribution. To address this, the gradient is sampled in the local vicinity of the estimated boundary, in the direction of the surface normal, to determine the location with the maximum gradient. The surface normal was determined for each triangle comprising the 3D polygonal surface representation, as described in Section 2.2.2. At each triangle, ten gradient samples are taken along the surface normal vector through the center of the triangle. The maximum gradient is recorded for each triangle, and statistics regarding the distribution of these maximum gradients are used as features for the nodule characterization system. This is illustrated for the two-dimensional case in Figure 2.12.

The method to compute the two-dimensional gradient feature is similar to the above method, with the exception of the use of surface normals. Instead, vectors

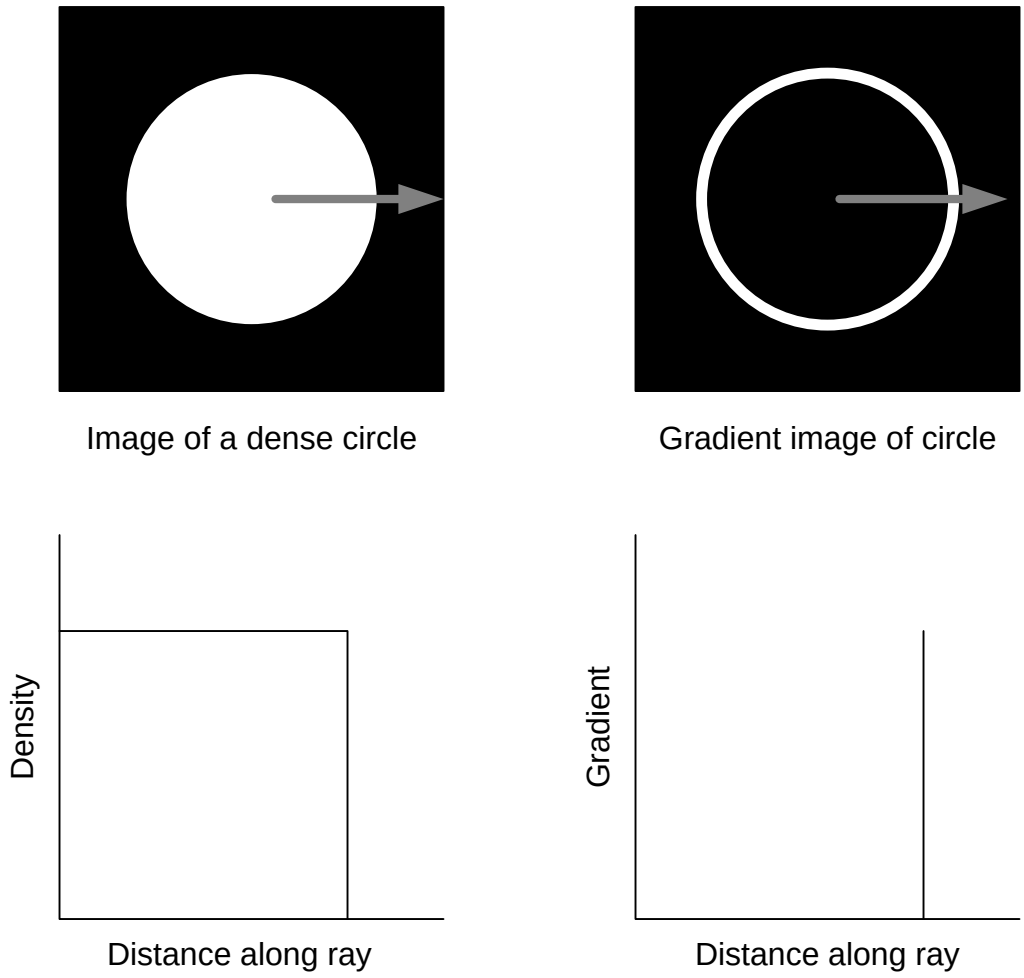


Figure 2.12: Illustration depicting gradient sampling method in 2D. The density image of a dense circle is shown in the upper left, with the gradient image shown in the upper right. A ray, indicating the surface normal, is shown on each image. The plots in the lower left and right show the values sampled along the rays for the original image and gradient image respectively.

are cast radially outward from the center of the nodule on the central slice at 10° intervals, which provides for 36 gradient samples. Again, samples are taken along the vector and the maximum gradient recorded.

2.2.4 Feature summary

A total of 43 3D features and 26 2D features were computed for this characterization system; the 2D features are listed in Table 2.1 and the 3D features in Table 2.2. The table lists the feature, whether the feature is dependent on size, and the type of feature. Features computed from the CT histogram are indicated by an "H", from moments or binary image analysis techniques by a "M", curvature features are indicated by a "C", and finally features computed from the gradient on the margin of the nodule are indicated by a "G" in the table.

2.3 Feature Classification

The goal of classification is to determine the best class (malignant or benign) to assign to a given feature vector corresponding to a nodule. There have been many classifiers developed in the field of statistics and machine learning; in this work, several classifiers were compared: logistic regression, support vector machines, and nearest-neighbors. These methods were selected to represent several different techniques that are often used for classification problems in medical image analysis. Logistic regression represents a parametric classification method that is often used for decision making in the medical field. Support vector machine is also a parametric method that has been shown to be very effective for high dimensional classification problems. Nearest-neighbors is a non-parametric method that does not assume a priori a model for separating the data and offers very fast training

Table 2.1: List of 2D features. Type indicates whether the feature was computed using the CT density histogram (H), moments or binary image analysis (M), curvature estimation (C), or gradient analysis (G)

Feature	Size dependent?	Type
Area	Y	M
Mass	Y	M
D_{μ}	N	H
D_{σ}	N	H
D_{σ^2}	N	H
D_{skewness}	N	H
D_{kurtosis}	N	H
EOI_{L_g}	Y	M
EOI_{W_g}	Y	M
EOI_{L_d}	Y	M
EOI_{W_d}	Y	M
A_{lw}	N	M
A_{lw_d}	N	M
ε	Y	M
$\hat{\varepsilon}$	N	M
circularity	N	M
$\text{compactness}_{\text{2D}}$	N	M
diameter	Y	M
∇_{min}	Y	G
∇_{max}	Y	G
∇_{range}	Y	G
∇_{μ}	Y	G
∇_{σ}	N	G
∇_{σ^2}	N	G
∇_{skewness}	N	G
∇_{kurtosis}	N	G

Table 2.2: List of 3D features. Type indicates whether the feature was computed using the CT density histogram (H), moments or binary image analysis (M), curvature estimation (C), or gradient analysis (G)

Feature	Size dependent?	Type
Volume	Y	M
Surface Area	Y	M
VSR	Y	M
Mass	Y	M
D_{μ}	N	H
D_{σ}	N	H
D_{σ^2}	N	H
D_{skewness}	N	H
D_{kurtosis}	N	H
compactness_{3D}	N	M
sphericity	N	M
EOI_{L_g}	Y	M
EOI_{W_g}	Y	M
EOI_{H_g}	Y	M
EOI_{L_d}	Y	M
EOI_{W_d}	Y	M
EOI_{H_d}	Y	M
A_{lw_g}	N	M
A_{wh_g}	N	M
A_{lh_g}	N	M
A_{lw_d}	N	M
A_{wh_d}	N	M
A_{lh_d}	N	M
ε	Y	M
$\hat{\varepsilon}$	N	M
ϕ_d	N	M
$\hat{\phi_d}$	N	M
κ_{min}	Y	C
κ_{max}	Y	C
κ_{range}	Y	C
κ_{μ}	Y	C
κ_{σ}	N	C
κ_{σ^2}	N	C
κ_{skewness}	N	C
κ_{kurtosis}	N	C
∇_{min}	Y	G
∇_{max}	Y	G
∇_{range}	Y	G
∇_{μ}	Y	G
∇_{σ}	N	G
∇_{σ^2}	N	G
∇_{skewness}	N	G
∇_{kurtosis}	N	G

performance. Each of these methods is further described in the following sections.

2.3.1 Logistic regression

Logistic regression is a classification method often used in the social sciences and medicine. Binomial logistic regression is used when the dependent variable is binary, as is the case here for classifying nodules into benign and malignant classes. The dependent variable is transformed into a **logit** variable; the logit function is the log of the odds of the dependent variable:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$$

where each nodule i has probability p_i of malignancy. The logistic regression equation is:

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}$$

where $x_{k,i}$ is the value of feature k for item i and $\beta_0 \ldots \beta_k$ are the unknown parameters to be found. Maximum likelihood estimation is used to solve for the coefficients $\beta_0 \ldots \beta_k$. Logistic regression was implemented using the generalized linear models function in the statistics toolbox in MATLAB (The Mathworks Inc., Natick, MA).

2.3.2 Support vector machines

Support vector machines (SVM) were originally proposed by Vapnik et al [37] as a method to solve two-group classification problems. They are designed to provide the best performance possible for cases where the two groups are non-separable, as is often the case in real data. Conceptually, an SVM maps an input vector into a

high-dimensional space and tries to find the optimal hyperplane that will separate the two groups. We wish to classify a set of labeled training data

$$(y_1, \mathbf{x}_1), \cdots, (y_l, \mathbf{x}_l) \qquad y_i \in \{-1, 1\} \tag{2.2}$$

using a hyperplane that separates the positive from negative examples (separating hyperplane). The points $\mathbf{x}$ which lie on the hyperplane satisfy $\mathbf{w} \cdot \mathbf{x} + b = 0$, where the vector $\mathbf{w}$ represents the normal vector perpendicular to the plane and b is an offset used to shift the plane. Let d_+ and d_- represent the shortest distance from the separating hyperplane to the closest positive and negative example respectively. The margin of the hyperplane is then $d_+ + d_-$. In the linearly separable case, the SVM algorithm searches for the hyperplane with the largest margin. In other words, given that the training data are linearly separable, they satisfy the following constraints:

$$\mathbf{w} \cdot \mathbf{x_i} + b \geq 1 \qquad \text{if } y_i = 1 \tag{2.3}$$

$$\mathbf{w} \cdot \mathbf{x_i} + b \leq 1 \qquad \text{if } y_i = -1 \tag{2.4}$$

The above equations can be rewritten as a single expression:

$$y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1, \qquad i = 1, \ldots, l \tag{2.5}$$

The training examples for which the equality in Eq. 2.3 holds lie on the hyperplane $H_1 : \mathbf{x}_i \cdot \mathbf{w} + b = 1$, with normal $\mathbf{w}$ and perpendicular distance to the origin $|1 - b| / \|\mathbf{w}\|$, with similar expressions for the training examples for which the equality in Eq. 2.4 holds. Thus, $d_+ = d_- = 1/\|\mathbf{w}\|$, and the margin is then $2/\|\mathbf{w}\|$. These are illustrated in Figure 2.13. To find the optimal hyperplane, we minimize $\|\mathbf{w}\|^2$ subject to the constraint of Eq. 2.5. Those training examples for which the equality in Eq. 2.5 holds and whose removal would change the solution are called *support vectors*.

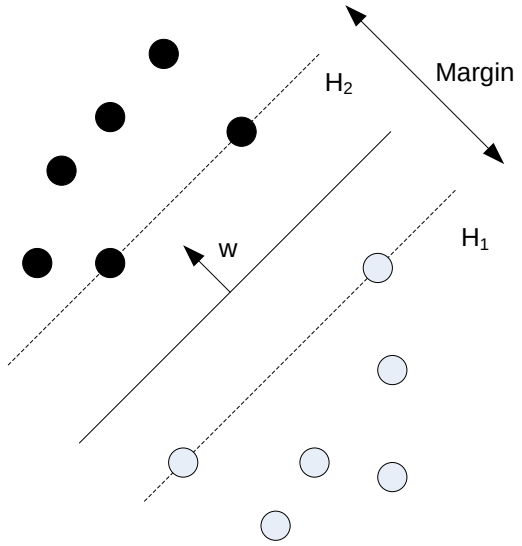


Figure 2.13: Example of a linearly separable SVM, with negative examples indicated by filled black circles and positive examples by open circles.

While this works for the case of linearly separable data, for data that is not linearly separable, there is no such optimal hyperplane. In the case of non-linearly separable data, the optimal hyperplane is one which minimizes classification error. We define a new variable to represent the error in classification, $\xi_i \geq 0, i = 1, \ldots, l$, and now the goal is the minimize the total error:

$$\mathbf{\Phi}(\xi) = C\sum_{i=1}^{l}\xi_i^{\sigma}$$

for $\sigma > 0$, subject to the constraints

$$y_i(\mathbf{w} \cdot \mathbf{x_i} + b) \geq 1 - \xi_i, \qquad i = 1, \ldots, l$$

$$\xi_i \geq 0, \qquad i = 1, \ldots, l$$

We define a parameter, C that controls the penalty for an error, and now we seek to solve the following optimization problem

$$\frac{1}{2}\left|\mathbf{w}\right|^2 + C\sum_{i=0}^{l}\xi_i$$

Using these expressions, the optimal hyperplane that minimizes errors can be computed. Solving these expressions requires some additional detail not described here, but can be found in papers by Vapnik [37] and Burges [38].

There are often cases where the decision function is not a linear function of the data. For these situations, a variation of the conditions above can be used that transforms the data into a non-linear space

$$y_i(\mathbf{w} \cdot \phi(\mathbf{x_i}) + b) \geq 1 - \xi_i, \qquad i = 1, \ldots, l$$

$$\xi_i \geq 0, \qquad i = 1, \ldots, l$$

where $\phi(\mathbf{x_i})$ is a mapping function. A kernel function, $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x_i}) \cdot \phi(\mathbf{x_i})$, can be defined, and this function is responsible for transforming the data. Examples of commonly used kernels are:

- Linear:$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$
- Polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \mathbf{x}_i^T \mathbf{x}_j + r)^d, \gamma > 0$
- Radial Basis Function: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right), \gamma > 0$

For this work, polynomial kernels were evaluated. Note that a polynomial kernel of order 1 ($d = 1$) is similar to a linear kernel. The SVM classifier was implemented using SVM*light*[1], a software package by Joachims et al [39]. The γ and r parameters, which are the scaling coefficient and offset respectively, were fixed at 1 to reduce the size of the parameter space, and the polynomial order (d) was varied from 1 to 4.

[1] Available from http://svmlight.joachims.org/

2.3.3 Nearest-neighbors classifier

The nearest-neighbors classifier is a non-parametric classifier that, in its basic form, assigns an test example the class of the closest example in the training dataset, the "nearest neighbor". The nearest neighbor is defined according to the Euclidean distance in feature space. While using a single example to make a classification decision is effective in datasets with good separation of the classes, in datasets with noisy data, basing the decision on a larger number of examples may be more robust. In a k-nearest-neighbors classifier, the majority class of k nearest examples are used to make the classification decision for an example. The number of examples, k, is determined during training.

The nearest-neighbors algorithm does not take into account the distance to the closest examples when making a decision. However, in many cases, it may be beneficial to give closer examples more weight in the classification decision than examples that are further away. A variation of nearest-neighbors that takes the distance of examples into account is the distance weighted nearest-neighbors (dwNN) classifier. This method was described by Paredes and Vidal in 2006 [40], and in their work, distances between training vectors $T = \{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ and an arbitrary vector $\mathbf{y}$ are computed as a weighted Euclidean distance

$$d(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{m} w_{ij}(y_j - x_j)^2$$

where there are m features, and $i = \text{index}(\mathbf{x}), \mathbf{x} \in T$. Paredes and Vidal described a method for optimizing the weights that attempts to maximize the margin, similar to the optimization goal of SVM. However, we utilized a simpler method of obtaining weights by computing the information gain ratio for each feature. The information gain ratio is described in further detail in Section 2.4.2. The inverse exponential distances from each vector in T to the arbitrary vector $\mathbf{y}$ are used as

weights on the class

$$c_{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{e^{\sigma \cdot d(\mathbf{x}^i, \mathbf{y})}} c_{\mathbf{x}^i}$$

where c indicates the class and σ is a parameter that essentially controls the size of the neighborhood – a large value will cause a lower weight to be assigned to the example. This parameter is optimized during training.

2.4 Feature Selection

Once the features for the nodules have been obtained, a subset of relevant features may need to be selected, depending on the classifier. Feature selection reduces the dimensionality of the feature space, which speeds up training. In addition, many classifiers require significantly more training examples than features. For example, an analysis of the nearest-neighbor algorithm by Langley and Iba (1993) [41] found that, for a single relevant feature, about 15 training examples were required before the accuracy began to asymptote. They also found that as the number of irrelevant features increased, the required number of training examples to reach the asymptote of performance increased exponentially. In the same vein, Blumer et al. (1993) [42] found in a theoretical analysis that by reducing the size of the hypothesis space (reducing the number of features), there was a reduction in the number of training examples required to obtain good generalizability .

Separate feature selection steps were performed for logistic regression and nearest-neighbors since neither algorithm handles irrelevant features well. Nearest-neighbors does not utilize any feature ranking method, and although logistic regression does learn feature weights, it requires a large number of training examples for the full dataset. A simulation study by Peduzzi et al. (1996) [43] suggested that at least 10 positive and 10 negative training examples were required per feature to obtain unbiased regression coefficients for logistic regression analysis. For our feature sets

of 23 2D and 37 3D features, that means at least 460 training examples would be required for the 2D feature set, and 740 examples for the 3D feature set. Given the size of the dataset used in this study (167 malignant and 92 benign nodules, described in detail in Section 2.5.1), nine features were to be selected for logistic regression. The features were selected by a simple ranking of features by discriminative performance for the logistic regression and nearest-neighbor classifiers, and information gain ratio was used in the distance weighted nearest neighbors classifier.

2.4.1 Discriminative performance based on ROC area

In the first feature selection scheme, features are ranked according to their discriminative performance based on the hypothesis that features which are better able to discriminate between benign and malignant nodules should be more useful. To assess discriminative performance, the area under the ROC curve (AUC) was computed for each feature on the entire dataset. The results for the top ten 2D features and top twenty 3D features are given respectively in Table 2.3 and 2.4. For the logistic regression classifier, ten features with an AUC greater than 0.60 were selected. Selecting features by their discriminative performance is easy to perform but has several limitations. Since only the performance of single features were computed, interactions between features were not considered. Also, the number of features as well as the AUC threshold to use for feature selection were established empirically.

2.4.2 Information gain ratio

Information gain, more formally called Kullback–Leibler divergence, is a measure of the reduction in entropy of a system gained by the use of a feature. The equation

Table 2.3: Discriminative performance for each 2D feature (only top 10 shown)

Feature	Area under ROC Curve
Area	0.674
Mass	0.663
∇_{skewness}	0.652
∇_{σ^2}	0.619
$\text{compactness}_{\text{2D}}$	0.617
∇_{range}	0.616
EOI_{w_d}	0.609
∇_{max}	0.604
A_{lw_d}	0.603
A_{lw_g}	0.600

Table 2.4: Discriminative performance for each 3D feature (only top 20 shown)

Feature	Area under ROC Curve
κ_σ	0.709
κ_{σ^2}	0.709
∇_μ	0.689
A_{lh_d}	0.685
∇_{range}	0.682
A_{lh_g}	0.682
EOI_{H_d}	0.679
EOI_{H_g}	0.678
∇_{range}	0.673
κ_{max}	0.670
κ_{min}	0.669
EOI_{W_d}	0.664
EOI_{W_g}	0.663
Volume	0.647
κ_μ	0.646
A_{lw_d}	0.646
Surface Area	0.646
∇_{skewness}	0.643
A_{lw_g}	0.641

Table 2.5: Illustrative example for information gain calculation, listing the objects and associated feature and class values

Color (A)	Size (B)	Class
red	big	+1
green	big	+1
red	big	+1
green	big	+1
green	big	+1
green	small	-1
red	small	-1
green	small	-1
red	small	-1
green	small	-1

for the entropy of a random variable X is:

$$H(X) = -\sum_{i=1}^{n} P(X = x_i) \log_2 P(X = x_i)$$

where X can take on any of n values $x_1, \ldots, x_n$, and $P(X = x_i) = \frac{|x_i|}{|X|}$. A high entropy indicates that X is from a uniform distribution, while a low entropy indicates a varied distribution. Given that we are trying to predict an output Y from input variable X, a related measure is the specific conditional entropy:

$$H(Y|X) = \sum_{i=1}^{n} P(X = x_i) H(Y|X = x_i)$$

where $H(Y|X = x_i)$ is the entropy of Y for only those examples in which X has the value x_i. Thus, $H(Y|X)$ gives the average specific conditional entropy of Y. Finally, the information gain is

$$IG(Y|X) = H(Y) - H(Y|X)$$

Information gain ranges from for a feature where no information is added to 1 where the feature perfectly separates the classes. The information gain computation is best illustrated with an example. Suppose we are trying to classify ten objects into two classes using features color (A) and size (B), and to simplify this

example, both color and size are binary features. The feature and class values of the ten objects are included in Table 2.5. We compute $H(Y)$ using the expression above:

$$H(Y) = -\left(\frac{|y_{-1}|}{|Y|}\log_2\frac{|y_{-1}|}{|Y|} + \frac{|y_{+1}|}{|Y|}\log_2\frac{|y_{+1}|}{|Y|}\right)$$

$$H(Y) = -\left(\frac{5}{10}\log_2\frac{5}{10} + \frac{5}{10}\log_2\frac{5}{10}\right) = 1.00$$

Next, we compute $H(Y|A = \text{red})$ and $H(Y|A = \text{green})$ in order to compute $H(Y|A)$:

$$\begin{aligned} H(Y|A = \text{red}) \quad = \quad & -\left(\frac{|y_{-1}|A = \text{red}|}{|Y|A = \text{red}|}\log_2\frac{|y_{-1}|A = \text{red}|}{|Y|A = \text{red}|} + \right. \\ & \left. \frac{|y_{+1}|A = \text{red}|}{|Y|A = \text{red}|}\log_2\frac{|y_{+1}|A = \text{red}|}{|Y|A = \text{red}|}\right) \end{aligned}$$

$$H(Y|A = \text{red}) = -\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right) = 1.00$$

$$H(Y|A = \text{green}) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = 1.00$$

$$H(Y|A) = P(A = \text{red})H(Y|A = \text{red}) + \text{P(A = green)H(Y|A = green)}$$

$$H(Y|A) = \frac{4}{10} * 1.00 + \frac{6}{10} * 1.00 = 1.00$$

Thus the information gain is

$$IG(Y|A) = H(Y) - H(Y|A) = 1.00 - 1.00 = 0.00$$

which indicates that the color feature provides no reduction in entropy, that is, using the color feature does not separate the objects into their classes. We perform the same computation for the size feature

$$H(Y|B = \text{small}) = -\left(\frac{5}{5}\log_2\frac{5}{5} + \frac{0}{5}\log_2\frac{0}{5}\right) = 0.00$$

Table 2.6: Information gain ratio of the top 10 features (all features, all data)

Feature	Gain ratio
∇_{range}	0.255
EOI_{H_g}	0.240
EOI_{H_d}	0.240
κ_{min}	0.232
κ_{range}	0.215
∇_{σ^2}	0.205
∇_{skewness}	0.205
∇_{σ}	0.205
A_{wh_d}	0.205
D_{σ}	0.193

$$H(Y|B = \text{big}) = -\left(\frac{0}{5}\log_2\frac{0}{5} + \frac{5}{5}\log_2\frac{5}{5}\right) = 0.00$$

$$H(Y|B) = \frac{5}{10} * 0.00 + \frac{5}{10} * 0.00 = 0.00$$

and the information gain for the size feature is

$$IG(Y|B) = H(Y) - H(Y|B) = 1.00 - 0.00 = 1.00$$

which indicates that the size feature does well in separating the objects into their respective classes. Table 2.6 lists top 10 features ranked by information gain.

2.5 Nodule Characterization Experiment

Two feature sets were evaluated to test the hypothesis that 3D features would be more effective for classification – a set of two-dimensional features and a separate set of three-dimensional features. This hypothesis was based on the fact that 3D features make use of additional data compared to 2D features. The nodule characterization system was trained and tested using a nested leave-one-out methodology. The dataset, performance metric, training and testing experiments are described in further detail in the following sections.

2.5.1 Nodule dataset

Cases were selected from the Weill Cornell Medical Center database that had at least one solid or part-solid nodule on at least one thin-slice CT scan. Part-solid nodules were only included if they were comprised primarily of a solid component. The status of malignant nodules was determined by either biopsy or resection, while the status of benign nodules was established through a negative biopsy result or by two years of no clinical change by a board certified radiologist. Nodules were included if they met the following criteria:

- Size greater than 3.0 mm and less than 30 mm, as measured by an automated algorithm
- CT scans with slices of 2.5 mm or less

Nodules were excluded if they met the following criteria:

- Metastatic cancers
- Benign calcifications

A total of 259 nodules (167 malignant and 92 benign) with CT scans of 1.0 mm, 1.25 mm, or 2.5 mm slice thickness fulfilled these criteria and were included in the dataset. Approximately 13.9% (36/259) of the nodules were on 1.0 mm scans, 73.8% (191/259) on 1.25 mm scans, and 12.4% (32/259) on 2.5 mm scans. Scans were obtained using either GE Medical Systems HiSpeed CT/i, Genesis HiSpeed, Genesis Zeus, LightSpeed Pro 16, LightSpeed QX/i, or LightSpeed Ultra CT scanners.

The three-dimensional automated segmentations for all 259 nodules were verified by visual inspection. The volume of each nodule was computed from each segmentation, and the nodule size was represented as the equivalent diameter of a sphere with the equivalent volume as the nodule. The 259 nodules in the dataset

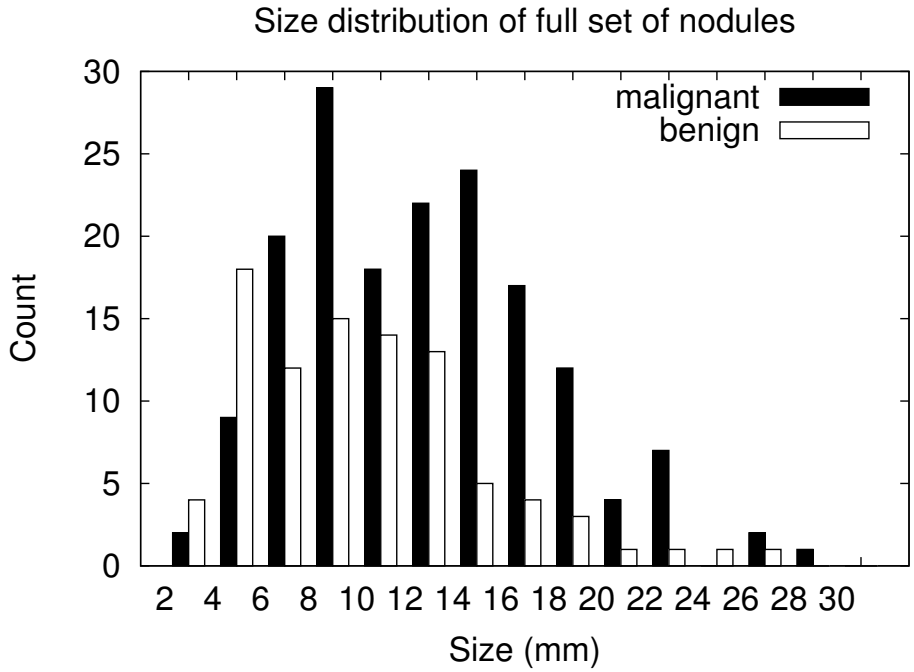


Figure 2.14: Size distribution of nodules in the dataset where size was determined through automated 3D segmentation.

ranged in size from 3.3 mm to 29.1 mm (median 11.3 mm), distributed as shown in Figure 2.14. The 167 malignant nodules ranged in size from 3.7 mm to 29.1 mm (median 12.2 mm) while the 92 benign nodules ranged in size from 3.3 mm to 27.1 mm (median 9.5 mm). A two-sided t-test showed a significant difference in the mean size between the malignant and benign nodules ($p < 0.01$).

2.5.2 Performance of the nodule characterization system

Evaluating classification systems necessitates the selection of an appropriate performance metric. For the task of pulmonary nodule classification, there are five relevant metrics that will be discussed: accuracy, sensitivity, specificity, ROC curves, and area under the ROC curve. Accuracy is a ratio of the number of examples correctly classified to the total number of examples and is sometimes expressed as

a percentage:

$$\text{Accuracy} = \frac{\text{number of cases correctly classified}}{\text{total number of cases}} \bullet 100\%$$

The best accuracy, obtained when all cases are classified correctly, is 100% and the worst accuracy is 0%. This metric is simple to calculate and summarizes the system's performance in a single number, but fails to distinguish between the different classes. To illustrate the last point, consider a set of 10 cases, 8 malignant and 2 benign. If all the cases were classified as malignant, the system would achieve 80% accuracy despite misclassifying all the benign cases. To address this problem, accuracy can be computed for each class separately. Accuracy for malignant cases is defined as sensitivity:

$$\text{Sensitivity} = \frac{\text{number of malignant cases correctly classified}}{\text{total number of malignant cases}}$$

and accuracy for the benign cases is defined as specificity:

$$\text{Specificity} = \frac{\text{number of benign cases correctly classified}}{\text{total number of benign cases}}$$

Sensitivity and specificity are bounded between 0 and 1, with the best performance indicated by 1.

Given the same example as before, a dataset with 8 malignant and 2 benign cases, with all cases classified as malignant, would result in:

$$\text{sensitivity} = \frac{8}{8} = 1.0$$

$$\text{specificity} = \frac{0}{2} = 0.0$$

While sensitivity and specificity are able to capture the performance of the system for each class, we now have two numbers to interpret, which makes comparing different methods more complicated, since one system could have a higher specificity but lower sensitivity than the other or vice versa.

Many classifiers generate a real-valued output instead of a binary classification. For these classifiers, it is possible to pick different thresholds on the output to get different performance results. Both accuracy and sensitivity/specificity fail to demonstrate the possible trade-offs in performance possible by varying this threshold. Receiver-operating characteristic (ROC) curves are used to graphically demonstrate these trade-offs.

An ROC curve plots sensitivity on the y-axis and 1-specificity on the x-axis. The best performance of 1.0 sensitivity and 1.0 specificity is indicated by the upper left corner of the plot and random chance is the diagonal line connecting (0,0) and (1,1) on the plot. The sensitivity and specificity are plotted for each threshold on the classification output. ROC curves can be used to compare different classification systems by assessing whether one is better at all levels of sensitivity and specificity, or the levels of importance. However, this relies on visual observation, and again, if the curves intersect it becomes more difficult to determine if one method is better.

A common metric derived from the ROC curve is the area under the ROC curve (AUC). The AUC provides a single number to represent the performance of a classification method. We computed the AUC by computing the sum of the areas of rectangles under the curve using `perf`[2], a program developed for the 2004 ACM Knowledge Discovery and Data Mining competition. Given n operating points, and a set of true positive fractions (sensitivity) T_k for each operating point $1...n$, and a set of false positive fractions (1-specificity) F_k, the AUC is estimated as:

$$\text{AUC} = \sum_{k=1}^{n} \frac{T_k + T_{k-1}}{2} * (F_k - F_{k-1})$$

where $T_0 = 0$ and $F_0 = 0$.

[2]Available from http://www.sigkdd.org/kddcup/index.php?section=2004&method=soft

Table 2.7: Example of training and testing sets using leave-one-out

Training Set	Testing Set
A B C D	E
A B C E	D
A B D E	C
A C D E	B
B C D E	A

2.5.3 Classifier Training and Evaluation Methodology

A leave-one-out (LOO) methodology was used to evaluate each classifier. In LOO, the system is trained on all examples except for one, and the one example is used to evaluate the system. This is repeated until all examples have been used for testing. This has the advantage of giving the system the largest number of examples from which to derive a model for the data, while preventing the system from overfitting the data. An example showing the training and testing sets for a case with five examples, labeled A - E, is shown in Table 2.7. In the case of the dataset used in this study, there are 259 examples, which will result in 259 iterations of training and testing the system.

While it is possible to optimize parameters for the system using the full training set (258 examples) within each leave-one-out iteration, without the use of a validation set, the system may overfit to the training data. Overfitting to the training data results in very high performance on the training set, but much lower performance on the testing data. To address this, for those classifiers that had parameters to adjust, the training data was further divided into a training set and an optimization set.

The goal of separating the training data from the data used to evaluate parameters is to reduce the possibility of choosing parameters that result in overfitting. To maximize the use of the training data, we again use either leave-one-out within

the training data, or for those classifiers where leave-one-out was computationally expensive, five-fold cross-validation was applied to the training set. In five-fold cross-validation, the dataset is divided into five sets; in each iteration, four sets are used for training while one set is used for validation. As with leave-one-out, the process is repeated until all sets have been used for validation.

For each iteration of leave-one-out for the test set, the testing example was classified using the classifier with the best set of parameters obtained in training. The AUC was computed across all the test examples. The performance was measured for two- and three-dimensional features separately. The areas under the ROC curves [44] were used to evaluate the performance of the algorithms, and the ROC curves were plotted for visual comparison.

The issue of class imbalanced datasets

In this dataset, the number of malignant and benign nodules is not equal, which resulted in a dataset with a class imbalance. Studies in the area of machine learning have suggested that this class imbalance may negatively affect the training of many classifiers [45], such as neural networks [46], decision trees [47], nearest-neighbors [48], and SVM [49]. To understand why this may be a problem, consider the case of an extreme class imbalance where there are 99 negative examples and a single positive example. In this case, the best performance would be achieved by classifying everything as a member of the negative class; however, this does not allow for any learning by the classification method. There have been several techniques proposed in the literature for addressing this problem, which can be divided into two general classes: undersampling and oversampling. In undersampling, examples from the class with the larger number are removed from the training set, while in oversampling, examples from the class with the smaller number of examples are duplicated. Of these two methods, oversampling tends to give better performance

[50, 51]. We perform oversampling by randomly duplicating benign nodule examples to equalize the number of cases in the training dataset. This was done for the SVM and nearest-neighbors classifiers.

Logistic Regression Training

Logistic regression does not have any parameters that require optimization. Thus, after the full dataset was divided using leave-one-out, the classifier simply used all of the examples in the training sets.

SVM Training

The polynomial SVM used in this study had two parameters to optimize – the order of the polynomial and the tradeoff between training error and margin. The polynomial order was varied from 1 to 4, while the tradeoff value was varied logarithmically from 0.0001 to 100.0 in order to capture the value with the greatest performance. Since running the SVM is computationally expensive, once leave-one-out was applied to the whole dataset, five-fold cross-validation was used on the training data to optimize the parameters.

Nearest-Neighbors and dwNN Training

The k-nearest-neighbors algorithm (kNN) has a single parameter, k, that controls the number of neighbors to consider when making a classification decision. This parameter was varied from 1 to 15, using only odd numbers to avoid tie situations that could occur with an even number of neighbors. The output of the kNN classifier is a binary output, in contrast to all the other classifiers evaluated in this thesis, and as a result, an AUC could not be computed because there is no threshold to vary on the classifier output. Instead, accuracy was used as the performance metric.

Table 2.8: Performance of characterization systems using 2D and 3D features. Sensitivity (Sens.) and specificity (spec.) are chosen at points with similar specificity.

Classifier	2D Features AUC	3D Features AUC
Logistic Regression	0.713	0.702
dwNN	0.641	0.700
SVM	0.624	0.686

The distance weighted nearest neighbor classifier (dwNN) had a single parameter, σ, that affected the size of the neighborhood to consider. This parameter was varied from 0.1 to 16.

Since leave-one-out is easy to perform for both kNN and dwNN, once leave-one-out was applied to the whole dataset, leave-one-out was also used on each training set to optimize the parameters described above.

2.6 Results: Evaluation of performance with two- and three-dimensional features

Two sets of features were used in this experiment – 2D features and 3D features. The results for each classifier are given in Table 2.8.

For logistic regression, given the set of 3D features, eight features were selected using the feature ranking criterion:

- volume of the nodule
- the height of the ellipsoid of inertia (EOI)
- the minimum, range, and standard deviation of the local curvature distribution
- the minimum and mean gradient along the margin of the nodule.

For the set of 2D features, eight features were selected:

- area
- mass
- length and width of the ellipsoid of inertia (EOI) of the nodule computed from geometric moments
- kurtosis of the density distribution
- length to width ratio of the EOI computed from the geometric moments
- length and width of the ellipsoid of inertia computed from the densitometric moments

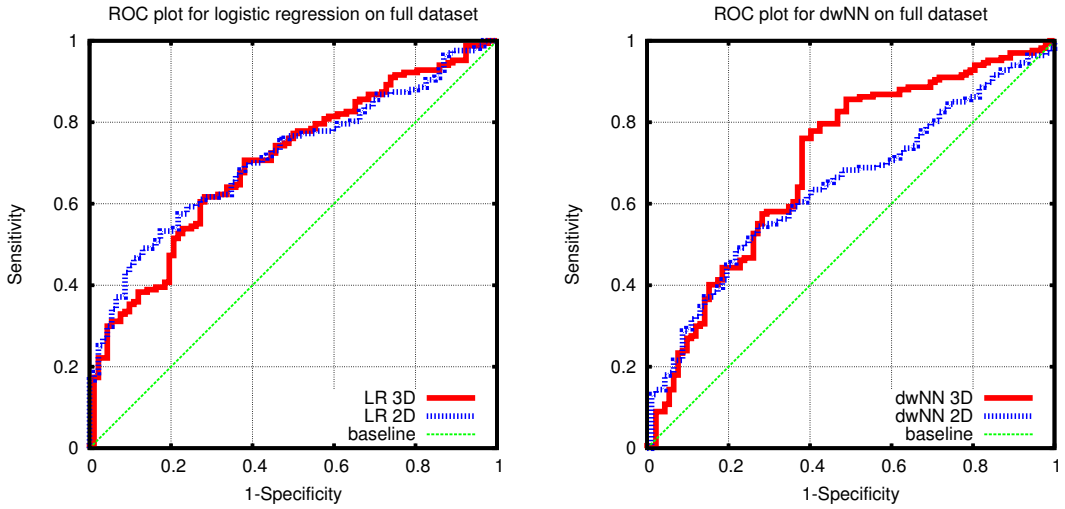
The ROC curves for the logistic regression classifier for both 2D and 3D features are shown in Figure 2.15a.

The dwNN classifier used all the features, but weighted each feature by its information gain. The ROC curves for the dwNN classifier using 2D and 3D features are shown in Figure 2.15b and the curves for the SVM classifier are shown in Figure 2.15c.

2.7 Discussion

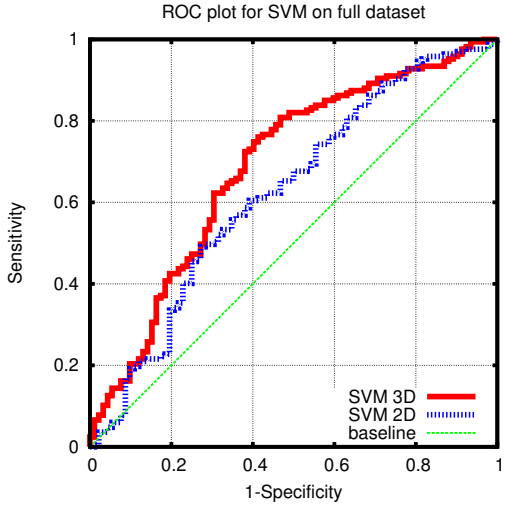
We conducted this study to assess the performance of the features developed for nodule characterization, determine if there is a difference between using 2D and 3D features, and to compare the performance of different classifiers.

Many features were not useful according to the feature selection algorithms performed. Ranking the features by discriminative performance indicated that many features were only marginally better than random chance. On this dataset of nodules, 3D features performed better than 2D features. However, looking at the results of the feature-ranking selection method, the best individual features



(a)

(b)



(c)

Figure 2.15: ROC curves for characterization systems using 2D and 3D features on full dataset with a) logistic regression (LR), b) dwNN, and c) SVM.

performed similarly between the 2D and 3D features, but more 3D features had high discriminative performance compared to the 2D features.

Distance-weighted nearest neighbors (dwNN) and support vector machines (SVM) performed the best on this dataset, for both 2D and 3D features. These two classifiers performed similarly on both the 2D and 3D sets of features. All methods did achieve higher performance on the training set than any individual feature alone.

The best performance achieved by this system was an AUC of 0.71, which is a measureable improvement in AUC over random chance. However, an AUC of 0.71 is disappointing for use in a clincial scenario. Previous studies have obtained AUC values in the range of 0.83 to 0.92 as shown in Table 1.2, though the datasets used by these studies had many more large malignant nodules than small malignant nodules, and vice versa for the benign nodules. This size bias was a major factor in the performance these systems were able to achieve and provides an overly optimistic evaluation of clinical usefulness; this issue is more closely examined in the following chapter.

CHAPTER 3

THE IMPACT OF NODULE SIZE DISTRIBUTION ON THE PERFORMANCE OF THE NODULE CHARACTERIZATION SYSTEM

The primary performance measure used in evaluating nodule characterization systems is the area under the ROC curve (AUC); in the ideal case with an area of 1.0, the system is able to completely distinguish between all benign and malignant nodules in the evaluation dataset. The performance is typically compared to a baseline performance of random chance, which would yield an AUC of 0.50. Thus, while the system presented in Chapter 2 has performance better than random chance, it does not appear to be as effective as previously published systems.

Studies have shown that the size of a lesion is a good predictor of malignancy [4, 3, 52]. However, the use of size as a feature in characterization systems is complicated by several factors:

1. Nodules in most datasets have a very large range; for example, a 3 mm to 30 mm range in lesion diameter corresponds to a volume range of 1000 to 1.
2. In addition to size features, such as volume, many of the other features are dependent on the size of the nodule, such as average curvature.
3. Due to 1 above, the size is related to the accuracy and detail that other feature measurements can be made on a nodule. As an example, given a typical voxel size for a CT scan of 0.7 x 0.7 x 1.25 mm, a 3 mm nodule has a volume equivalent to about 23 voxels; given partial volume effects, noise, etc., this is inadequate to provide meaningful values for some of the complex shape-based features.
4. For a dataset with a biased size distribution of malignant and benign nodules, the size (or a size derived feature) is often the most useful feature.

5. In all published datasets used for training and evaluating nodules for which the size distribution is given, there is a difference in the size of benign and malignant nodules in which small benign and large malignant nodules predominate. This skewness in the distribution of the dataset reflects the natural history of lesions found in lung scans; however, the actual distribution is very sensitive to the population subset from which the data was acquired; e.g. screening scans would be expected to have a different distribution compared to clinical scans.

Therefore, in any pulmonary nodule dataset, there is an intrinsic classification performance that can be achieved by use of a size feature alone that is dataset-specific. In general, we are interested in a system performance evaluation that is not highly dependent on a population feature of a particular dataset. With this in mind, many ROC results that have been published in the literature look very promising but are actually largely characterizing the size skewness in the training dataset. All of the studies mentioned thus far, with the exception of the study by Awai et al [22], either have different size distributions of malignant and benign nodules or no information regarding the size distribution.

This chapter investigates the effect of the difference in size distribution of benign and malignant nodules on the performance of the nodule characterization system. The system is evaluated on both the full dataset of nodules and an enriched dataset of nodules which were selected to maintain a similar size distribution of benign and malignant nodules. Finally, a new baseline performance is proposed that takes into account the size distribution.

3.1 Nodule Size Distribution Experiments

To evaluate the effect of the difference in nodule size distribution between benign and malignant nodules, we examined three related issues:

1. Performance that results from the difference in the size distributions
2. Features that are dependent on size
3. Behavior of the system for datasets with different size distributions

To address the first issue, we measure the performance of a simple size-threshold classifier to estimate the baseline performance from the size-distribution; this classifier is explained in detail in Section 3.1.1. In the case of a dataset where the benign and malignant nodules have the same size distribution, using a size-threshold classifier should not perform much better than random chance, but as the size distributions begin to differ, the size-threshold classifier will perform better.

The second issue can be addressed by comparing the performance from systems that include or exclude features that are size-dependent; see Tables 2.1 and 2.2 for a list of features and their dependency on size. The system was trained and evaluated according to the same methodology as in Section 2.5.3.

Finally, the third issue can be evaluated by comparing the performance of systems trained and tested on a subset of nodules selected to have the same size-distribution for malignant and benign nodules; this dataset is described in Section 3.1.2. Due to the small size of the dataset, a leave-one-out evaluation methodology was performed. In leave-one-out, all nodules in the dataset except for one are used for training, and the one nodule is used for testing. Optimal parameters for each classifier were selected based on each training set. For the SVM classifier, the optimal parameters were determined on the entire training set, while for the

k-nearest neighbors and distance-weighted nearest neighbors classifiers, leave-one-out was used on just the training set to determine the optimal parameters.

3.1.1 Size threshold classifier

A size-threshold classifier was included to establish the performance that can be obtained from the difference in size distributions of benign and malignant nodules. The rationale for the size-threshold classifier was based on the empirical observation that malignant nodules in most datasets tend to be larger than benign nodules; therefore, size should offer high discriminative performance. The probability of malignancy given a nodule's size can be determined for the dataset based on the size distribution. A threshold can then be set to classify nodules as either malignant or benign. The size-threshold classifier uses size as the sole discriminating feature – nodules below the size threshold are benign while nodules above the size threshold are classified as malignant. For this classifier, the size was represented by the equivalent diameter given the volume of the nodule, though volume could have also been used since the order of the nodules would remain the same.

3.1.2 Size-balanced subset of nodules

The size-balanced, enriched subset was created to help assess how eliminating the discriminating power of size from the dataset would affect the performance of classification systems. Thirty benign and malignant nodules were selected from the full dataset to have as similar size distributions as possible. In this dataset, the nodules ranged in size from 7.04 mm to 12.91 mm (median 10.01 mm), with similar ranges for both malignant (7.06 mm to 12.85 mm, median 9.96 mm) and benign (7.04 mm to 12.91 mm, median 10.01 mm) nodules, with the distribution shown in Figure 3.1.

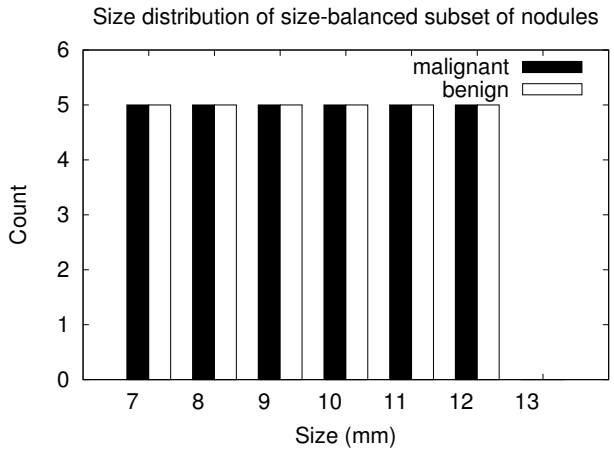


Figure 3.1: Size distribution of nodules in the subset selected to have similar size distributions. Labels on the axis represent the range of nodule sizes included in the bin.

3.2 Results

To examine the impact of the underlying size distribution of the dataset of pulmonary nodules, we first measured the performance of a size-threshold classifier on the full dataset. Next, we measured the performance of systems where size-dependent features were excluded and compared the performance to systems using all available features on both the full dataset and the subset of nodules selected to have similar size distributions.

3.2.1 Performance of size-threshold classifier

The size-threshold classifier used the diameter of each nodule estimated by the automated 3D segmentation method, as described in Section 3.1.1. The size-threshold classifier achieved an area under the ROC curve (AUC) of 0.653. Note that the AUC for the size-threshold classifier is above the conventional baseline AUC of 0.50, and the ROC curve, shown as a solid line in Figure 3.2, is above the baseline performance indicated by the the diagonal dashed line.

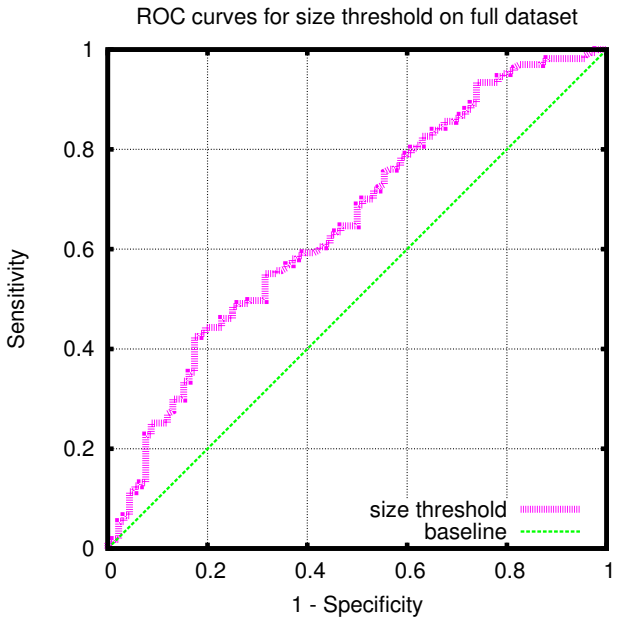


Figure 3.2: ROC curve of size-threshold classifier on full dataset

Table 3.1: Summary of AUC performance on full dataset on feature sets including and excluding size-dependent features

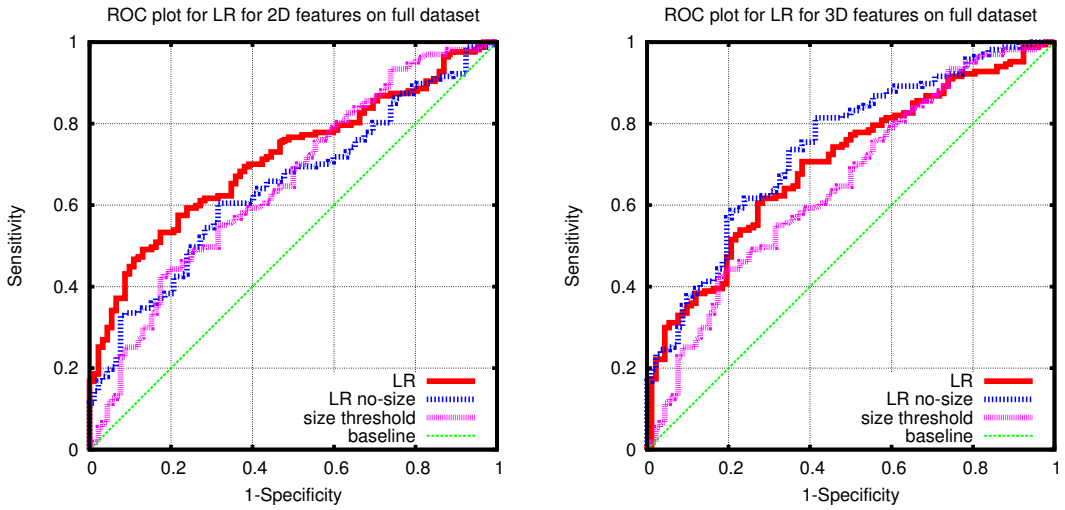
	2D features		3D features	
classifier	all	no size	all	no size
logistic regression	0.713	0.647	0.702	0.743
distance-weighted nearest neighbor	0.641	0.620	0.700	0.704
SVM	0.624	0.581	0.686	0.614
size threshold	0.653			

3.2.2 The impact of size dependent features

For this experiment, the characterization system was limited only to features considered to be size-independent in Table 2.1 and 2.2. The AUC for each system is given in Table 3.1; for ease of comparison, the performances for the systems using all features from Section 2.6 are also listed in the table. The performance for the k-nearest-neighbors classifier is given in Table 3.2. The ROC plots for the logistic regression, distance-weighted nearest neighbors, and SVM classifiers are presented in Figures 3.3, 3.4, and 3.5 respectively.

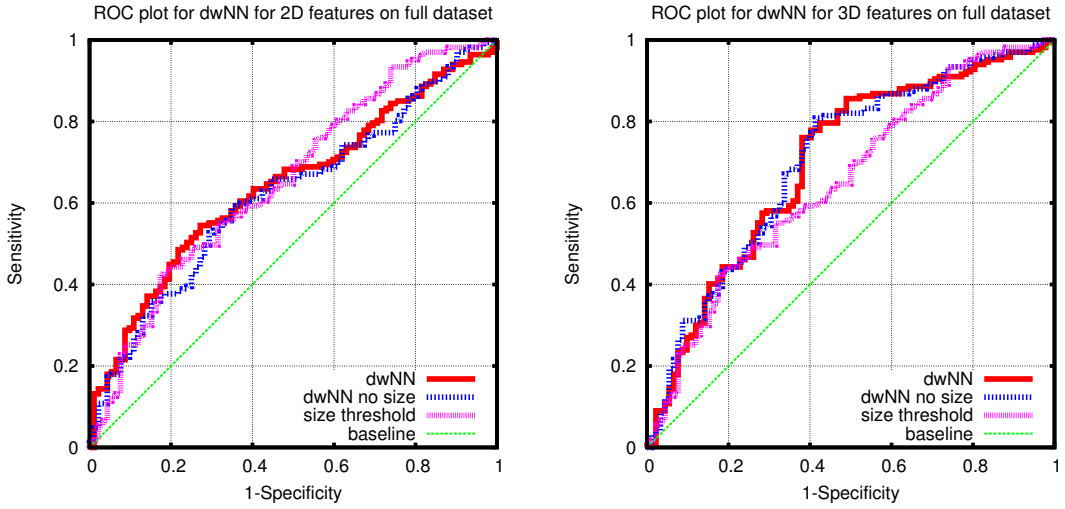
Table 3.2: Performance of k-nearest-neighbors classifier on full dataset on feature sets including and excluding size-dependent features

	Size features			No size features		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
2D	0.567	0.695	0.337	0.622	0.832	0.239
3D	0.618	0.778	0.326	0.664	0.832	0.359



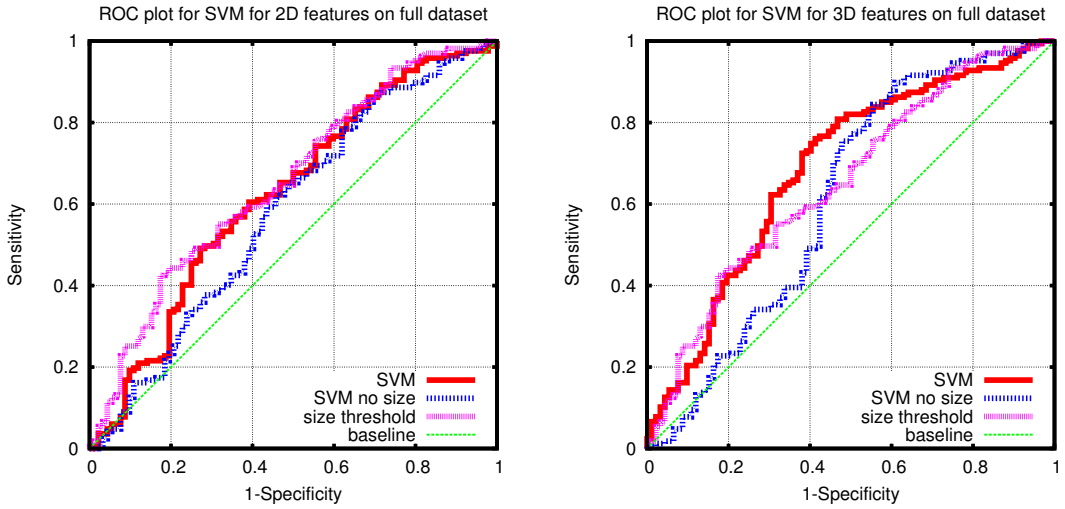
(a) (b)

Figure 3.3: ROC curves for logistic regression (LR) classifier on full dataset with a) 2D features and b) 3D features, both including and excluding size. For reference, the ROC curve for the size-threshold classifier is shown as well.



(a) (b)

Figure 3.4: ROC curve for dwNN classifier on full dataset with a) 2D features and b) 3D features, both including and excluding size. For reference, the ROC curve for the size-threshold classifier is shown along with the conventional baseline indicated by a diagonal line.



(a) (b)

Figure 3.5: ROC curve for SVM classifier on full dataset with a) 2D features and b) 3D features, both including and excluding size. For reference, the ROC curve for the size-threshold classifier is shown along with the conventional baseline indicated by a diagonal line.

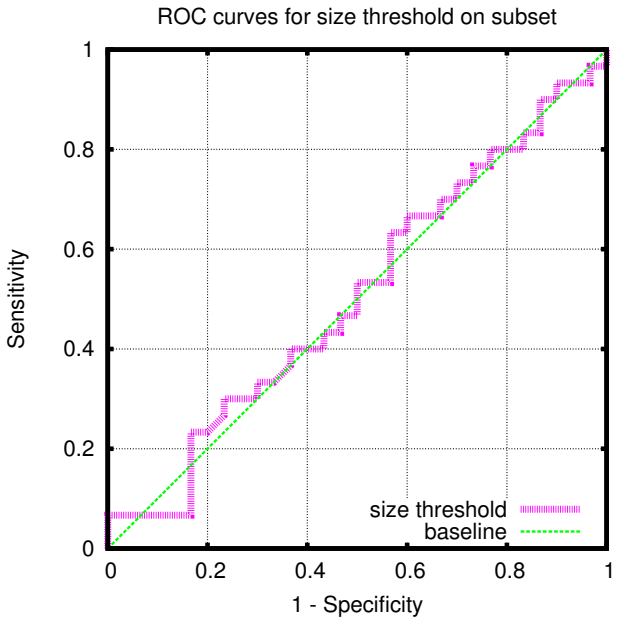


Figure 3.6: ROC curve for size-threshold classifier on size-balanced subset

Table 3.3: Performance of classification system (AUC) on the size-balanced subset

	2D features		3D features	
classifier	all	no size	all	no size
logistic regression	0.508	0.566	0.579	0.583
distance-weighted nearest neighbor	0.618	0.608	0.503	0.561
SVM	0.638	0.681	0.523	0.634
size threshold	0.507			

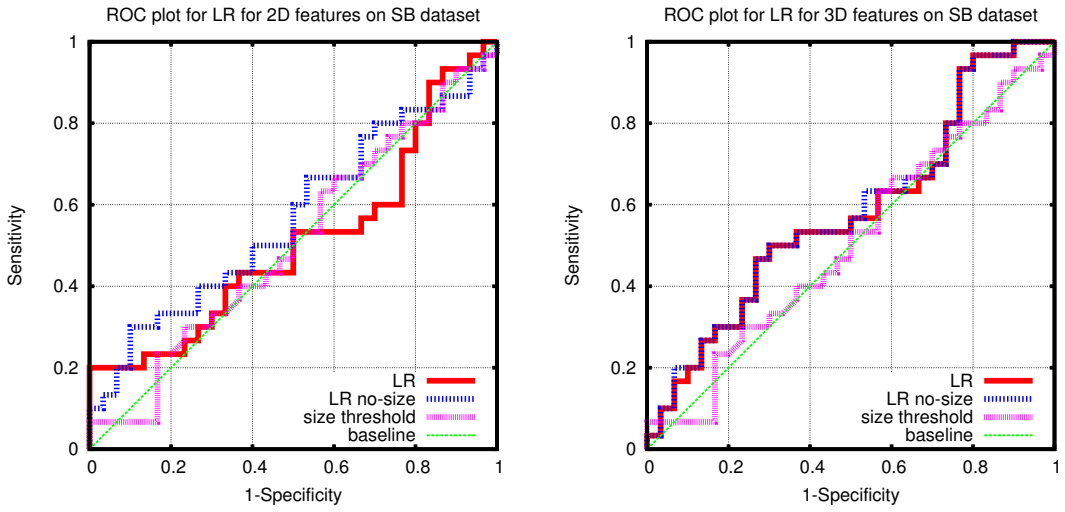
3.2.3 Size-balanced subset results

The size-threshold classifier achieved an AUC of 0.507, near the conventional baseline of 0.500. This is reflected in the ROC curves shown in Figure 3.7 by the dashed line. This suggests that size offers little benefit over random chance on this dataset, in contrast to the full dataset where size offered considerable improvement over random chance.

The ROC curves are shown for the logistic regression classifier in Figure 3.7. Note that the feature selection step chose different sets of features than in the pre-

Table 3.4: k-nearest neighbor performance on size-balanced subset

	Size features			No size features		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
2D	0.617	0.700	0.533	0.550	0.533	0.567
3D	0.550	0.533	0.567	0.550	0.533	0.567



(a) (b)

Figure 3.7: ROC curves for logistic regression classifier on size-balanced subset with a) 2D features and b) 3D features, both including and excluding size. The ROC curve for the size threshold classifier and the conventional baseline are shown on the plots as well.

vious experiments; Tables 3.5 and 3.6 rank the features according to discriminative performance on this dataset. In accordance with the results presented by Peduzzi et al. (1996), only 3 features were selected for the logistic regression classifier [43].

ROC curves for the dwNN and SVM classifiers are shown in Figures 3.8 and 3.9 respectively. AUC values for all the classifiers are presented in Table 3.3, with separate performance results for k-NN in Table 3.4.

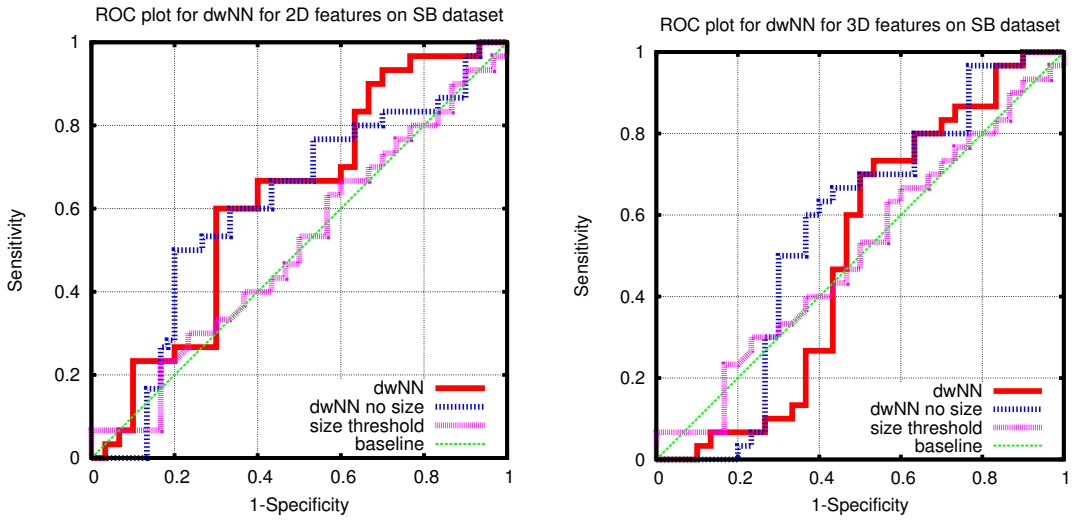
For the dwNN and logistic regression classifiers, better performance was achieved using 3D features with size-dependent features excluded than with all features. The

Table 3.5: Discriminative performance on the size-balanced subset of nodules for each 2D feature (only top 10 shown)

Feature	Area under ROC
LWR_G	0.703
LWR_D	0.687
EOI_{L_G}	0.629
D_μ	0.602
D_{kurtosis}	0.601
EOI_{L_D}	0.598
Area	0.570
Mass	0.558
D_{skewness}	0.544
$\nabla_{\mathrm{kurtosis}}$	0.540

Table 3.6: Discriminative performance on the size-balanced subset of nodules for each 3D feature (only top 10 shown)

Feature	Area under ROC
∇_μ	0.688
LHR_D	0.661
LHR_G	0.657
D_μ	0.650
κ_σ	0.647
LWR_D	0.629
∇_{min}	0.623
LWR_G	0.622
EOI_{L_G}	0.601
κ_{max}	0.600



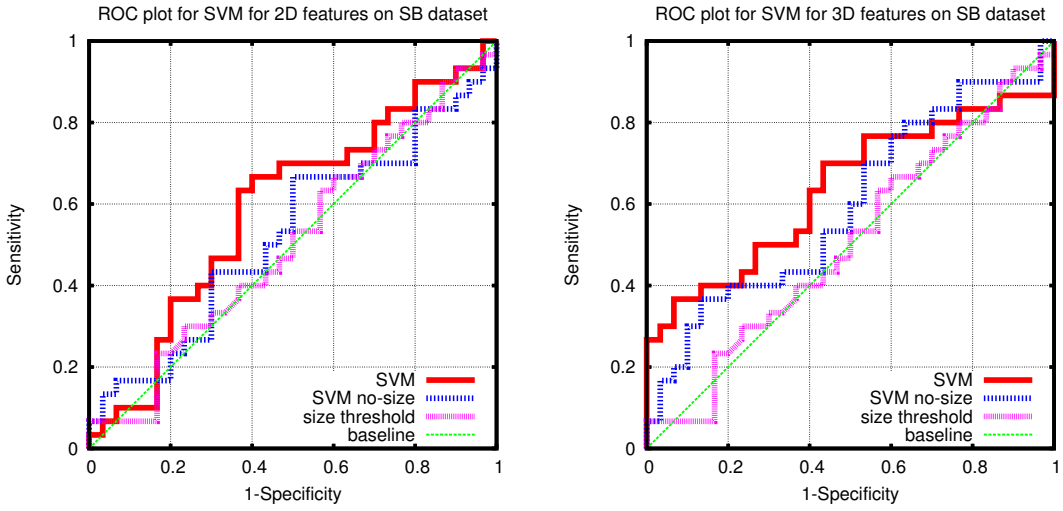
(a) (b)

Figure 3.8: ROC curves for distance-weighted nearest neighbors classifier on size-balanced subset with a) 2D and b) 3D features, both including and excluding size.

performance of the SVM classifier was better with all 3D features. The results for the 2D features were less definitive, with the logistic regression classifier achieving better performance on the 2D feature set without size compared to the 2D feature set with all features, while the situation was reversed for the remaining classifiers. These results seem to suggest that size features offer an advantage, even in the absence of any size-bias; however, these results should be interpreted carefully due to the limited size of the dataset.

3.3 Discussion

To assess the impact that biases in the size distribution of nodules may have on performance, this study used two datasets with different size distributions. The full dataset of 259 nodules reflects nodule sizes more typical of characterization studies. Sixty nodules were selected from the full dataset so that the size distributions of the



(a) (b)

Figure 3.9: ROC curves for SVM classifier on size-balanced subset with a) 2D and b) 3D features, both including and excluding size.

malignant and benign nodules would be as similar as possible. On the full dataset, the simple size-threshold classifier achieved an AUC of 0.653, showing improvement over the baseline AUC of 0.507. This suggests that the distribution of malignant and benign nodules in our dataset are more similar than other datasets with a higher sensitivity and specificity from size. This reduced bias makes this a more challenging dataset to characterize than most others reported in the literature.

In the second experiment, the effect of including features that were dependent on size was evaluated. The best performance on the full dataset was an AUC of 0.743 achieved by logistic regression with 3D features excluding size, which is a large improvement over the baseline performance, but a smaller improvement compared to the size threshold AUC. Additionally, in all feature sets on the full dataset, removing size-dependent features reduces performance, with the exception of logistic regression using 3D features. This suggests that size is responsible for a portion of the reported performance of all classifiers on the full dataset.

For third experiment using the size-balanced dataset, the size threshold classifier achieved an AUC of 0.507, which is near baseline performance, as expected from a dataset with an equal distribution of sizes of malignant and benign nodules. Accordingly, performance of the logistic regression classifier that included size features was reduced compared to the full dataset. As one example, consider that the characterization system, using the logistic regression classifier with 2D features that included size, exhibited a reduction in AUC from 0.713 on the full dataset to 0.508 on the size-balanced subset of nodules, despite similar levels of optimization performed for both datasets. The logistic regression classifier which used 3D features, but excluded size-dependent features, had the best performance. However, for the dwNN classifier, there was no clear pattern; the 2D feature set had better performance with size, while the 3D feature set had better performance without size. This may be due to the fact that many of the 2D features were not very discriminative, so that the additional features in the set of features including size aided performance. The dwNN classifier performed worse on the subset of nodules than the full dataset, suggesting that size distribution affects the performance of the dwNN classifier as well.

The higher performance of size relative to the conventional baseline performance is not unique to this dataset; the same occurrence can be observed in other characterization studies that have published a size distribution plot for their datasets. For several such studies, performance of a size-threshold classifier was estimated using the counts and sizes listed on each study's size distribution plot. Note that since the plots generally had few size bins, performance of the size-threshold classifier may be better than reported here.

In the study by Suzuki et al. (2005) [17], the vast majority of nodules were small (less than 7 mm) and benign. The database of nodules in their study included 413

benign and 76 malignant nodules ranging in size from 3 mm to 31 mm. Using a size-threshold classifier set to approximately 7 mm, an estimated performance of a sensitivity and specificity (SS) of (0.80, 0.80) would be achieved. This performance is very similar to that shown on the ROC curve for their trained computer method, although numerically they reported a SS of (1.00, 0.48). In a dataset of 33 benign and 48 malignant nodules used by Shah et al. (2005) [15], the nodule sizes are much larger than the dataset by Suzuki et al. (2005)[17], suggesting that the nodules were taken from a clinical population. On this database, the size-threshold classifier using a size of 15 mm would achieve an approximate SS of (0.64, 0.79) based on the author's size distribution chart [15]. The authors achieved a SS of (0.90, 0.80) using their automated method, as measured from their ROC curve. Another study by Way et al. (2006) [18] on a database of 96 nodules reveals similar performance results from size. A size-threshold classifier on their database at a threshold of 20 mm was able achieve a sensitivity of 70% (31/44) with a specificity of 80% (42/52). In these studies, the size-threshold alone was able to achieve much of the reported performance of the respective automated systems. A large database of nodules alone is not sufficient to address the issue of unequal size distributions; even the study by Suzuki et al, which had nearly 500 nodules, was affected by the size-distribution because of the significant difference in size between the malignant and benign nodules.

The use of size information is not desirable because it fails to generalize well. As an example, if the size criterion determined from the dataset of Shah et al is used on the database of Suzuki et al, a SS of (0.38, 0.95) would be achieved, and in the reverse case, using the size criterion determined from the dataset of Suzuki et al of 10 mm (the closest interval to 7 mm), a SS of (0.21, 0.90) would be achieved. In a clinical setting, the sizes of nodules do not have the same distribution of sizes as

nodules in the datasets used to train and test automated characterization systems; thus, reported performance that does not take into account the size distribution of nodules in the dataset has a high likelihood of being over-optimistic. A more relevant measure of the effectiveness of a system is the improvement in performance over use of a simple size-threshold classifier. By reporting the performance of the size-threshold classifier, the improvement in classification performance for the nodule characterization system can be computed by simply subtracting the AUC of the system from the AUC of the size-threshold classifier. For example, a system that achieves an AUC of 0.80 on a dataset with a size-threshold classifier performance of 0.50 would have an improvement of 0.30, while if the size-threshold classifier performance was 0.65, the improvement would only be 0.15. Using this performance metric will minimize some of the effect of the underlying size distribution of the dataset.

CHAPTER 4

CONCLUSION

A system for the classification of benign and malignant solid pulmonary nodules has been presented. The best performance was an area under the ROC curve of 0.756, achieved using logistic regression and 3D features. Features were computed from the density of the nodule, geometric and density moments, local surface variations, and the gradient at the edge of the nodule. Although the performance was better than both baseline and the performance of the size-threshold classifier which only used size as a feature, it is lower than previously published studies. Further analysis of the results suggested that the size-distribution of the dataset used for training and testing plays a large role in the reported performance of a classification system, and thus comparing results across different datasets is difficult. In this study, the malignant and benign nodule sizes were more similar than in some previously published studies, which is one explanation for the reduced performance.

4.1 Contributions

The primary contribution of this work is the observation that the datasets used for training and testing nodule characterization systems has a large impact on the performance that can be obtained from these systems. In datasets with a large difference in size between the malignant and benign nodules, size is a very effective discriminative feature. However, the size distribution in these datasets may not be an accurate reflection of the true size distribution of nodules; while it is true that malignant nodules tend to be larger, there should be more small malignant nodules as well, but we do not observe this in most datasets due to the fact that malignant nodules tend to already be large by the time they are identified. Additionally, the

most useful application of a nodule characterization system is for those nodules that are too small to be biopsied; these nodules tend to be of the size where the malignant and benign nodules overlap, where size is not as useful a feature.

Given these limitations of the use of size, the best way to report the performance of nodule characterization systems is using datasets where the size distribution of malignant and benign nodules is as similar as possible. However, it is very difficult to build the large databases necessary for proper training and testing of characterization systems while following that requirement. An alternative is to publish not the absolute performance, but the relative increase that the system provides over size; this is what has been done in this work.

4.2 Future Work

Additional work can be done on refining the more complex three-dimensional features, such as local surface variation and the nodule margin measurement features. An increase in the number of nodules in the database would enable the use of separate training and testing sets for the size-balanced subset, which would eliminate any possibility of overfitting to the data and allow for additional parameter optimization. This system could also be applied to characterizing nodules in other areas of the body, such as the liver or breast.

BIBLIOGRAPHY

[1] American Cancer Society, *Cancer Facts & Figures 2009*. American Cancer Society, 2009.

[2] 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.," *The New England journal of medicine*, vol. 355, pp. 1763–71, Oct. 2006.

[3] D. M. Libby, J. P. Smith, N. K. Altorki, M. W. Pasmantier, D. Yankelevitz, and C. I. Henschke, "Managing the Small Pulmonary Nodule Discovered by CT," *Chest*, vol. 125, no. 4, pp. 1522–1529, 2004.

[4] E. A. Zerhouni, F. P. Stitik, S. S. Siegelman, D. P. Naidich, S. S. Sagel, A. V. Proto, J. R. Muhm, J. W. Walsh, C. R. Martinez, and R. T. Heelan, "CT of the pulmonary nodule: a cooperative study," *Radiology*, vol. 160, no. 2, pp. 319–327, 1986.

[5] R. A. Brooks and G. D. Chiro, "Principles of computer assisted tomography (CAT) in radiographic and radioisotopic imaging," *Physics in Medicine and Biology*, vol. 21, no. 5, pp. 689–732, 1976.

[6] C. I. Henschke, D. F. Yankelevitz, R. Mirtcheva, G. McGuinness, D. McCauley, and O. S. Miettinen, "CT screening for lung cancer: frequency and significance of part-solid and nonsolid nodules," *AJR. American journal of roentgenology*, vol. 178, pp. 1053–7, May 2002.

[7] S. Siegelman, E. Zerhouni, F. Leo, N. Khouri, and F. Stitik, "CT of the solitary pulmonary nodule," *Am. J. Roentgenol.*, vol. 135, no. 1, pp. 1–13, 1980.

[8] J. W. Gurney, "Determining the likelihood of malignancy in solitary pulmonary nodules with bayesian analysis. part I. theory," *Radiology*, vol. 186, no. 2, pp. 405–413, 1993.

[9] J. W. Gurney, D. M. Lyddon, and J. A. McKay, "Determining the likelihood of malignancy in solitary pulmonary nodules with bayesian analysis. part II. application," *Radiology*, vol. 186, no. 2, pp. 415–422, 1993.

[10] M. D. Seemann, A. Staebler, T. Beinert, H. Dienemann, B. Obst, M. Matzko, C. Pistitsch, and M. F. Reiser, "Usefulness of morphological characteristics for the differentiation of benign from malignant solitary pulmonary lesions using HRCT," *European Radiology*, vol. 9, no. 3, pp. 409–417, 1999.

[11] S. Takashima, S. Sone, F. Li, Y. Maruyama, M. Hasegawa, T. Matsushita, F. Takayama, and M. Kadoya, "Small solitary pulmonary nodules (<= 1 cm) detected at population-based CT screening for lung cancer: Reliable high-resolution CT features of benign lesions," *American Journal of Roentgenology*, vol. 180, no. 4, pp. 955–964, 2003.

[12] F. Li, S. Sone, H. Abe, H. MacMahon, and K. Doi, "Malignant versus Benign Nodules at CT Screening for Lung Cancer: Comparison of Thin-Section CT Findings 1," 2004.

[13] Y. Kawata, N. Niki, and J. Ohmatsu, "Curvature-based internal structure analysis of pulmonary nodules using thoracic 3D CT images," *Systems and Computers in Japan*, vol. 32, pp. 9–19, September 2001.

[14] 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," *Medical Physics*, vol. 30, pp. 387–394, March 2003.

[15] 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, May 2005.

[16] 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 enhancment features," *Academic Radiology*, vol. 12, pp. 1310–1319, October 2005.

[17] K. Suzuki, F. Li, S. Sone, and K. Doi, "Computer-aided diagnostic scheme for disctinction between and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network," *IEEE Transactions on Medicial Imaging*, vol. 24, pp. 1138–1150, Sept. 2005.

[18] 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, no. 7, pp. 2323–2337, 2006.

[19] P. A. Lachenbruch and M. Goldstein, "Discriminant analysis," *Biometrics*, vol. 35, no. 1, pp. 69–85, 1979.

[20] F. Li, M. Aoyama, J. Shiraishi, Q. Li, K. Suzuki, R. Engelmann, S. Sone, H. MacMahon, and K. Doi, "Radiologists' performance for differentiating benign from malignant lung nodules on high-resolution CT using computer-estimated likelihood of malignancy," *American Journal of Roentgenology*, vol. 183, pp. 1209–1215, November 2004.

[21] S. K. Shah, M. F. McNitt-Gray, K. R. D. Zoysa, P. Batra, A. Behrashi, K. Brown, L. E. Greaser, J. M. Park, D. K. Roback, C. Wu, E. Zaragoza, J. G. Goldin, R. D. Suh, M. S. Brown, and D. R. Aberle, "Solitary pulmonary nodule diagnosis on CT: Results of an observer study," *Academic Radiology*, vol. 12, pp. 496–501, April 2005.

[22] 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, no. 1, pp. 276–284, 2006.

[23] A. Reeves, A. Chan, D. Yankelevitz, C. Henschke, B. Kressler, and W. Kostis, "On measuring the change in size of pulmonary nodules," *IEEE Transactions on Medical Imaging*, vol. 25, pp. 435–450, April 2006.

[24] 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 Transactions on Medical Imaging*, vol. 22, pp. 1259–1274, October 2003.

[25] S. Siegelman, N. Khouri, F. Leo, E. Fishman, R. Braverman, and E. Zerhouni, "Solitary pulmonary nodules: CT assessment," *Radiology*, vol. 160, no. 2, pp. 307–312, 1986.

[26] C. Zwirewich, S. Vedal, R. Miller, and N. Muller, "Solitary pulmonary nodule: high-resolution CT and radiologic-pathologic correlation," *Radiology*, vol. 179, no. 2, pp. 469–476, 1991.

[27] M. D. Seemann, A. Staebler, T. Beinert, H. Dienemann, B. Obst, M. Matzko, C. Pistitsch, and M. F. Reiser, "Usefulness of morphological characterisitics for the differentiation of benign from malignant solitary pulmonary lesions using HRCT," *European Radiology*, vol. 9, no. 3, pp. 409–417, 1999.

[28] R. Lindell, T. Hartman, S. Swensen, J. Jett, D. Midthun, H. Tazelaar, and J. Mandrekar, "Five-year Lung Cancer Screening Experience: CT Appear-

ance, Growth Rate, Location, and Histologic Features of 61 Lung Cancers," *Radiology*, vol. 242, no. 2, p. 555, 2007.

[29] A. P. Reeves, R. J. Prokop, S. E. Andrews, and F. P. Kuhl, "Three-dimensional shape analysis using moments and fourier descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, pp. 937–943, November 1988.

[30] W. J. Kostis, *Three-dimensional computed tomographic image analysis for early cancer diagnosis in small pulmonary nodules.* PhD thesis, Cornell University, January 2001.

[31] Y. Kawata, N. Niki, H. Ohmatsu, M. Kusumoto, R. Kakinuma, K. Mori, H. Nishiyama, K. Eguchi, M. Kaneko, and N. Moriyama, "Curvature based characterization of shape and internal intensity structure for classification of pulmonary nodules using thin-section CT images," in *Medical Imaging 1999: Image Processing* (K. Hanson, ed.), vol. 3661, pp. 541–552, May 1999.

[32] S. Rusinkiewicz, "Estimating curvatures and their derivatives on triangle meshes," *3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004. Proceedings. 2nd International Symposium on*, pp. 486–493, 6-9 Sept. 2004.

[33] P. Sundaram, A. Zomorodian, C. Beaulieu, and S. Napel, "Colon polyp detection using smoothed shape operators: Preliminary results," *Medical Image Analysis*, Aug 2007.

[34] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *SIGGRAPH Comput. Graph.*, vol. 21, no. 4, pp. 163–169, 1987.

[35] K. Furuya, S. Murayama, H. Soeda, J. Murakami, Y. Ichinose, H. Yabuuchi, Y. Katsuda, M. Koga, and K. Masuda, "New classification of small pulmonary nodules by margin characteristics on high-resolution ct," *Acta Radiol*, vol. 40, no. 5, pp. 496–504, 1999.

[36] O. Monga, R. Deriche, and J.-M. Rocchisani, "3D edge detection using recursive filtering: application to scanner images," *Computer Vision and Pattern Recognition, 1989. Proceedings CVPR'89., IEEE Computer Society Conference on*, pp. 28–35, 1989.

[37] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, September 1995.

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

[39] T. Joachims, *Making large-Scale SVM Learning Practical*. MIT Press, 1999.

[40] R. Paredes and E. Vidal, "Learning weighted metrics to minimize nearest-neighbor classification error," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, pp. 1100–10, July 2006.

[41] P. Langley and W. Iba, "Average-Case Analysis of a Nearest Neighbor Algorithm," in *IJCAI'93: Proceedings of the 13th International Joint Conference on Artificial Intelligence*, vol. 13, (San Francisco, CA, USA), pp. 889–889, Morgan Kaufmann Publishers, 1993.

[42] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, "Occam's razor," *Information processing letters*, vol. 24, no. 6, pp. 377–380, 1987.

[43] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein, "A simulation study of the number of events per variable in logistic regression analysis," *Journal of Clinical Epidemiology*, vol. 49, no. 12, pp. 1373 – 1379, 1996.

[44] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[45] N. Japkowicz and S. Stephen, "The class imbalance problem : A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, 2002.

[46] M. a. Mazurowski, P. a. Habas, J. M. Zurada, J. Y. Lo, J. a. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance.," *Neural networks : the official journal of the International Neural Network Society*, vol. 21, no. 2-3, pp. 427–36, 2008.

[47] G. M. Weiss and F. Provost, "Learning When Training Data are Costly : The Effect of Class Distribution on Tree Induction," *New York*, vol. 19, pp. 315–354, 2003.

[48] S. Tan, "Neighbor-weighted K-nearest neighbor for unbalanced text corpus," *Expert Systems with Applications*, vol. 28, pp. 667–671, May 2005.

[49] R. Akbani, S. Kwek, and N. Japkowicz, "Applying Support Vector Machines

to Imbalanced Datasets," in *Machine Learning: ECML 2004* (J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, eds.), pp. 39–50, Springer Berlin / Heidelberg, 2004.

[50] N. V. Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE : Synthetic Minority Over-sampling TEchnique," *Artificial Intelligence*, vol. 16, pp. 341–378, 2002.

[51] G. E. a. P. a. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newsletter*, vol. 6, p. 20, June 2004.

[52] S. Diederich, M. Thomas, M. Semik, H. Lenzen, N. Roos, A. Weber, W. Heindel, and D. Wormanns, "Screening for early lung cancer with low-dose spiral computed tomography: results of annual follow-up examinations in asymptomatic smokers.," *European radiology*, vol. 14, no. 4, pp. 691–702, 2004.