Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods

Matthew C. Hancock
Jerry F. Magnan

Matthew C. Hancock, Jerry F. Magnan, "Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods," *J. Med. Imag.* **3**(4), 044504 (2016), doi: 10.1117/1.JMI.3.4.044504.

SPIE.

Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods

Matthew C. Hancock and Jerry F. Magnan*
Florida State University, Department of Mathematics, 208 Love Building, 1017 Academic Way, Tallahassee, Florida 32306-4510, United States

Abstract. In the assessment of nodules in CT scans of the lungs, a number of image-derived features are diagnostically relevant. Currently, many of these features are defined only qualitatively, so they are difficult to quantify from first principles. Nevertheless, these features (through their qualitative definitions and interpretations thereof) are often quantified via a variety of mathematical methods for the purpose of computer-aided diagnosis (CAD). To determine the potential usefulness of quantified diagnostic image features as inputs to a CAD system, we investigate the predictive capability of statistical learning methods for classifying nodule malignancy. We utilize the Lung Image Database Consortium dataset and only employ the radiologist-assigned diagnostic feature values for the lung nodules therein, as well as our derived estimates of the diameter and volume of the nodules from the radiologists' annotations. We calculate theoretical upper bounds on the classification accuracy that are achievable by an ideal classifier that only uses the radiologist-assigned feature values, and we obtain an accuracy of 85.74 (± 1.14)%, which is, on average, 4.43% below the theoretical maximum of 90.17%. The corresponding area-under-the-curve (AUC) score is 0.932 ($\pm$0.012), which increases to 0.949 ($\pm$0.007) when diameter and volume features are included and has an accuracy of 88.08 (± 1.11)%. Our results are comparable to those in the literature that use algorithmically derived image-based features, which supports our hypothesis that lung nodules can be classified as malignant or benign using only quantified, diagnostic image features, and indicates the competitiveness of this approach. We also analyze how the classification accuracy depends on specific features and feature subsets, and we rank the features according to their predictive power, statistically demonstrating the top four to be spiculation, lobulation, subtlety, and calcification. [DOI: 10.1117/1.JMI.3.4.044504]

Keywords: computer-aided diagnosis; lung nodule classification; Lung Image Database Consortium dataset; random forests; logistic regression; machine learning.

Paper 16150R received Jul. 21, 2016; accepted for publication Nov. 14, 2016; published online Dec. 8, 2016.

1 Introduction

A number of features derived from CT scan images of the lung are considered to be diagnostically relevant for the assessment of lung nodules.[1–3] We refer to these as diagnostic image features. Examples include simple features, such as nodule diameter and volume, as well as more complex features, such as spicularity and lobularity. Unfortunately, the current definitions of such complex features are qualitative in nature,[1,4] precluding the widespread use of standard algorithmic quantification of the features for use in clinical practice. Nevertheless, many studies have quantified such features numerically for the purpose of either computer-aided diagnosis (CAD) or computer-aided characterization, by mathematically approximating characteristics of the features (from an interpretation of their respective qualitative definitions) using an assortment of algorithmic methods.[5–12] On the other hand, others have used the algorithmic quantification of image features only as intermediate quantities within a system for classifying nodules as malignant or benign.[13–18] In these approaches, it is not clear how well the quantified features capture the true physical nature of the features themselves because only error metrics for the accuracy of nodule classification are considered rather than the approximation error in quantifying the features themselves.

The development of a CAD system that first accurately quantifies diagnostic image features before classifying that the lung nodule as malignant or benign requires that the following two hypotheses be satisfied:

1. Diagnostic features can be quantified accurately from the image scans alone.
2. Lung nodules can be classified as malignant or benign to within a sufficient degree of accuracy using only the (accurately) quantified diagnostic features as input.

The first hypothesis is not discussed in this paper although it is currently under investigation by the authors. We concentrate

*Address all correspondence to: Jerry F. Magnan, E-mail: magnan@math.fsu.edu

2329-4302/2016/$25.00 © 2016 SPIE

on the analysis and validation of the second hypothesis, i.e., we assume that the diagnostic image features can be, and have been, accurately quantified and then test whether a nodule can be accurately classified as malignant or benign from these features alone when they are used as inputs in statistical learning methods. To undertake this investigation, we employ the radiologist-assigned values of the various diagnostic images features provided in the Lung Image Database Consortium (LIDC) dataset,[19] where the degree of nodule malignancy is also indicated by the radiologist annotators. The remainder of this paper is structured as follows. In Sec. 2, we discuss the related work. In Sec. 3, we describe the LIDC dataset and our experimental setup. We then present our results in Sec. 4, followed by a discussion of the results in Sec. 5, before expressing our conclusions in Sec. 6.

2 Related Work

As noted by Wiemker et al.,[20] two approaches to CAD can be identified. We describe them briefly and then provide a more detailed discussion of work done within each in Secs. 2.1 and 2.2. The two approaches are:

1. Approach one: A large set of image features that may, or may not, have a meaningful radiological interpretation are automatically computed. The computed features are used only as intermediate quantities by an algorithm to produce a final binary-valued label describing the class of the nodule (e.g., as malignant or benign). Such features could be calculated, e.g., by the hidden layers of a neural network or by a wavelet transform.

2. Approach two: Diagnostic image features from various medical–radiological terminology sets are specifically computed and quantified algorithmically. For development of the algorithms, these quantified feature values are validated against the quantifications of the same diagnostic image features made by radiologists. Obtaining quantified diagnostic image features may be the end goal (for the purpose of displaying them to the radiologist for consideration during nodule assessment) or may be used instead (or additionally) as inputs to a larger CAD system to automatically classify the nodule (e.g., as malignant or benign).

We note that there is a third approach that is not to be confused with the validation procedure in approach two. This third, and separate, line of study computes various image features and uses traditional statistical correlation techniques to find associations between a particular image feature and a nodule malignancy category.[21–25] Such an approach can be useful for discovering new diagnostic image features and for increasing confidence in known associations when the computed image feature has an accepted radiological interpretation. However, we do not apply this approach or discuss it further here.

A goal of this present work is to determine how statistically accurate a machine learning method can be in determining the malignancy of a nodule, using quantified and radiologically interpretable diagnostic image features. Thus, we explore the feasibility of a CAD system that employs approach two.

2.1 *Previous Work Related to Approach One*

Here, we focus on studies that mathematically approximate characteristics of the image features, which are then used in an intermediary fashion to classify the nodule but that do not necessarily attempt to validate the accuracy of the approximations made with respect to any diagnostic image feature (Table 1).

We describe a number of approaches that use texture feature extraction methods. For a more general review of these methods, as used in biomedical imaging applications, see the review by Depeursinge et al.[29] To analyze texture properties that are small distance outside the nodule's boundary, Way et al.[26] computed the rubber-band straightening transform (RBST) along the one-dimensional nodule boundary contour in various two-dimensional (2-D) planar intersections with the three-dimensional (3-D) nodule volume. The RBST acts as a mapping from 2-D Cartesian coordinates to a coordinate system, where the abscissa corresponds to the position along the nodule's boundary contour and the ordinate corresponds to the distance outside of the boundary. Thus, the radial spicules projecting normal to a nodule's boundary will appear as approximately vertical lines in the transform. After computing the transformation, a run-length matrix was computed, which quantifies the pixel-value frequencies and hence the texture. Statistics from the matrix were computed and used as inputs to a linear-discriminant classifier for classifying 96 nodules either as malignant or benign, achieving an area under the receiver operating characteristic (ROC) curve [area-under-the-curve (AUC)] score of 0.83. Krewer et al.[13] automatically extracted 219 2-D and 3-D texture and shape features and used a feature-selection method to find significant features, which were then used as inputs to support vector machines (SVMs), decision trees, and nearest-neighbor classification methods on 33 cases from the LIDC dataset. They reported an accuracy of 90.91% in classifying 14 malignant and 19 benign nodules from the LIDC dataset when correlation-based feature-selection was used. Only the five optimal features that result from their feature selection process, which include 3-D-wavelet coefficients, are specified,. Han et al.[14] extracted 2-D and 3-D Haralick features (i.e., statistical features computed from a pixel-intensity co-occurrence matrix), Gabor features (i.e., convolutional responses to the Gabor wavelet), and local binary pattern features (i.e., the histograms of neighborhood intensity comparisons) as inputs to SVMs to classify 1356 nodules from the LIDC dataset as malignant or benign. They reported average AUC scores of 0.89, 0.88, and 0.87, respectively, for the three previously mentioned features.

El-Baz et al.[16] considered a series expansion representation of the nodule surface using spherical harmonics. Using the coefficients in the series as descriptors of the nodule boundary shape, they classified 51 malignant and 58 benign nodules using a nearest-neighbor method, and achieved 94.4% accuracy. Tac and Uur[17] used a combination of shape and texture features for detecting and classifying nodules in the LIDC dataset. The features include 16 shape features, among which are geometrical characterizations such as area, eccentricity, and boundary length, and 22 texture features, which include statistical characterizations of pixel values such as the mean, variance, and entropy. After performing feature selection, they reported a classification accuracy of 95.64% using a generalized linear model classifier on 141 nodules. Dilger et al.[27] investigated the hypothesis that image features derived from the region surrounding the nodule improves classification accuracy. Testing their

Table 1 Previous work related to approach one.

Year	Author	Method	Result
2006	Way et al.[26]	Linear discriminant trained on run-length features from RBST	AUC of 0.83
2011	El-Baz et al.[16]	Nearest-neighbor trained on spherical harmonic coefficients	Accuracy of 94.40%
2013	Krewer et al.[13]	Nearest-neighbor trained on various image-derived shape and texture features	Accuracy of 90.91%
2015	Dilger et al.[27]	Artificial neural network and linear-discriminant classifiers trained on features including those derived from lung tissue surrounding the lung nodule	AUC of 0.938
2015	Tac and Uur[17]	Generalized linear model trained on various image-derived shape and texture features	Accuracy of 95.64%
2015	Han et al.[14]	SVM trained on Haralick features	AUC of 0.89
2015	Kaya and Can[18]	Random forest trained on various image-derived features and LIDC labels	Accuracy of 84.89%
2016	Firmino et al.[28]	SVM trained on HOG features and LIDC labels	AUC of 0.91

hypothesis on 50 lung nodules, they found that the AUC score improved from 0.918 (excluding features derived from surrounding tissue) to 0.938 (including features derived from surrounding tissue). Kaya and Can[18] used a combination of both quantified features from the LIDC dataset and image-derived features, such as Zernike moments to automatically assess the malignancy of a nodule, and achieved an accuracy of 84.89%. In the work by Firmino et al.,[28] the histogram of oriented gradient (HOG) features were used in conjunction with the radiologist-quantified features as inputs to an SVM classifier, employing the LIDC dataset for the diagnosis phase of a combined lung-nodule detection and diagnosis system; they achieved AUC scores of up to 0.91.

Although these methods may achieve high classification accuracy, it should be noted that direct comparisons are difficult to make because of the varying datasets and dataset sizes used. Moreover, it is not clear whether or not the intermediate features agree with any known diagnostic image features because the quantifications are not validated against ground-truth labels of diagnostic image features assigned by radiologists. We also see from the above studies that there appears to be no standard way of quantifying these diagnostic features. A lack of a standard method could be problematic if different methods exhibit different sensitivities to noise and to the variability of anatomical features present in operational settings.

2.2 *Previous Work Related to Approach Two*

Nodule volume and volume doubling-time are well-established indicators used to assess malignancy,[3] and there are studies that attempt to accurately quantify the volume of a nodule using CT scan image data (Table 2). Compared to other more complex and qualitatively defined diagnostic features of a nodule, such as spiculation, the nodule volume is simpler to define. Thus, its quantification has been more prevalent in studies compared to other diagnostic features. Quantification of nodule volume is often posed as a segmentation problem, which is a well-established subfield of medical imaging research.[33–35] Mullally et al.[30] devised an adaptive-threshold segmentation algorithm to determine nodule growth rate. They tested their method on images from scanned physical phantoms consisting of implanted artificial lung nodules of various known volumes. When their algorithm was tested on CT scans of the phantom, they achieved a volume error to within 23% of the known size, according to a root-mean-square error metric. Reeves et al.[31] used a combination of "pleural segmentation, adaptive thresholding (of Hounsfield units in the CT image data), image registration, and knowledge-based shape matching" to measure nodule volume and determine volume change in consecutive CT scans of the same patient. They validated their algorithm by measuring the variability in the volume measurements on 50 nodules, which were known to be stable (i.e., they demonstrated no change in volume) over a 2-year period. A significant decrease in variability was observed when compared to their previous methods, which employed a fixed rather than adaptive threshold. Way et al.[26] modified the 2-D active-contours algorithm by adding 3-D energy terms for segmenting 23 lung nodules in the LIDC dataset. Employing the radiologist-annotated segmentations in the LIDC dataset, they quantified the performance of their method using a measure of overlap with the given annotations and achieved an average overlap score of up to 0.63. Messay et al.[32] used a regression neural network to guide thresholding parameters for segmenting lung nodules in the LIDC dataset. They tested their system on 66 lung nodules and achieved up to an 80% overlap score with the known nodule annotations.

In addition to nodule volume measurements, a number of papers have attempted to quantify more complex shape and appearance features. Iwano et al.[5] considered 102 nodules classified by radiologists into different categories of shapes (i.e., round, lobulated, spiculated, and four other shape categories) and used computed measures of aspect ratio, circularity, and second central moment to quantify these shape categories. They found that circularity and second-moment features are suitable for differentiating between some of the categories; however, they found difficulty in using these features to separate nodules in the spiculated and "ragged" categories. Raicu[6] argues that quantification of diagnostic features will "reduce the semantic

Table 2 Previous work related to approach two.

Year	Author	Targeted radiological feature	Method	Result
2004	Mullally et al.[30]	Volume	Adaptive thresholding	Within 23% of known size
2005	Iwano et al.[5]	Roundness, lobulation, spiculation	No formal classification method used	No reported metric
2006	Reeves et al.[31]	Volume	Adaptive thresholding with additional knowledge-based techniques	Observed significant decrease in segmentation variability
2006	Way et al.[26]	Volume	Active contours with 3-D energy terms	Average overlap score of 0.63
2009	Raicu[6]	LIDC feature labels	Logistic regression, decision trees, SVM trained on low-level, computed image features	Accuracy ranged from 30% to 100%, depending on target features
2009	Zinovev et al.[7]	LIDC feature labels	Active-learning method, trained on low-level, computed image features	Accuracy up to 73%
2013	Dhara et al.[8]	LIDC spiculation labels	Quantification via differential-geometry methods	Average accuracy of 87.4%
2014	Zhang et al.[10]	LIDC spiculation labels	Quantification via nodule boundary information	Sensitivity of 90%
2015	Ciompi[12]	NELSON and DLCST spiculation labels	Random forest trained with "bag-of-frequencies" features	AUC of 0.9
2015	Messay et al.[32]	Volume	Regression neural network	Average overlap score of 0.80
2015	Niehaus et al.[11]	LIDC spiculation labels	Decision tree trained on various image-derived shape and texture features	AUCs of 0.6 to 0.9, depending on nodule-size grouping

gap" in medical image retrieval and interpretation, and they use a variety of low-level shape, size, texture, and intensity features to map these features to the corresponding diagnostic image feature labeled in the LIDC dataset, using logistic regression, decision trees, and SVMs. Separating out the cases for analysis where various numbers of annotators agree on the quantified values of the features, their accuracies range from 38% to 100%, on average. Zinovev et al.[7] used shape, size, texture, and intensity features, combined with an active learning method, to compute the diagnostic feature labels assigned by radiologists in the LIDC dataset and achieved up to 73% accuracy. Dhara et al.[8] used a differential geometry approach, using Gaussian and mean curvatures, to quantify the spiculation of lung nodules. They validated their quantification method using the expert-derived labels assigned by the radiologists in the LIDC dataset and achieved an average accuracy of 87.4% on 95 nodules. Zhang et al.[10] also use the spiculation labels provided in the LIDC dataset to validate their nodule-spiculation quantification method, which uses edge information computed along the nodule boundary, on 20 cases (10 spiculated and 10 nonspiculated). They achieve a sensitivity of 90%. Niehaus et al.[11] analyzed the association between nodule size and the success of computing accurate spiculation labels from the LIDC dataset. They find that the AUC score increases roughly from 0.6 to 0.9 as the size of the nodules considered increases. Ciompi[12] introduced a bag-of-frequencies descriptor as inputs to a statistical classifier to categorize 255 nodules from the NELSON and DLCST datasets as either spiculated or nonspiculated. They compared the use of the proposed bag-of-frequencies features with the use of scale invariant feature transform and spherical harmonic features, resulting in AUC scores of 0.9, 0.456, and 0.63, respectively.

In approach two, although the computed diagnostic image features are validated against the ground-truth provided by expert annotators, it is clear that, similar to the situation with methods that are employed in approach one, there is no standard method for the algorithmic quantification of any particular feature. The different accuracy of the methods and their potential variation due to scanner noise and parameters make approach two a challenging one to effectively incorporate as a medically useful component of a CAD system.

3 Materials and Methods

3.1 *Brief Overview of the Lung Image Database Consortium Dataset*

The LIDC dataset[19] is a publicly available set of 1018 lung CT scans collected through various universities and organizations. In addition to the CT image data, manual annotations for each scan from anonymous radiologists from four sites are provided. These annotations are made with respect to the following types of structures:

1. Lung nodules whose largest diameter is greater than 3 mm.
2. Lung nodules whose largest diameter is less than 3 mm.
3. Nonnodule structures whose largest diameter is greater than 3 mm.

For each of these types, the location of the structure is given in image coordinates as determined by each of the four

physicians, with no forced consensus about their existence or location imposed. It is the first type of structure (i.e., lung nodules with largest diameter ≥3 mm) that we analyze in this work. For this type of structure, additional annotations are assigned by each of the same radiologists from the four sites. These annotations include manually drawn contours of the nodule boundaries in the CT scan slices, quantified values for a variety of nodule features, and a quantified value of the estimation of the nodule's malignancy at the time of assessment. The eight quantified nodule features and the corresponding malignancy quantification, along with the features' respective rating systems, are listed and described in Table 3. Note that the quantifications are radiological interpretations of the presence of the respective physical features. We emphasize that the malignancy quantification is not pathologically established in the majority of nodules. However, some follow-up data are available for a small subset of the nodules in the dataset (but we have not considered this data in our study). An example nodule from the dataset, along with the assigned diagnostic feature values, is shown in Fig. 1.

We summarize some of the patient demographic and scan information (obtained by inspecting the DICOM file data from the LIDC dataset) in Figs. 2 and 3. In 734 scans, no age or gender information is provided. In the remaining 284 cases where gender information (DICOM Tag ID: 0010,0040) is available, the distribution is 49.3% male and 50.7% female. When age information (DICOM Tag ID: 0010,1010) is also available, the median age is 61 years, as can be seen in Fig. 2. It is never the case that age information is available but gender is not. Figure 3 illustrates the distribution of the spacing of pixels within a slice of the scan (DICOM Tag ID: 0028,0030) and the thickness of each slice (DICOM Tag ID: 0018,0050), having medians of 0.6986 and 2.0 mm, respectively.

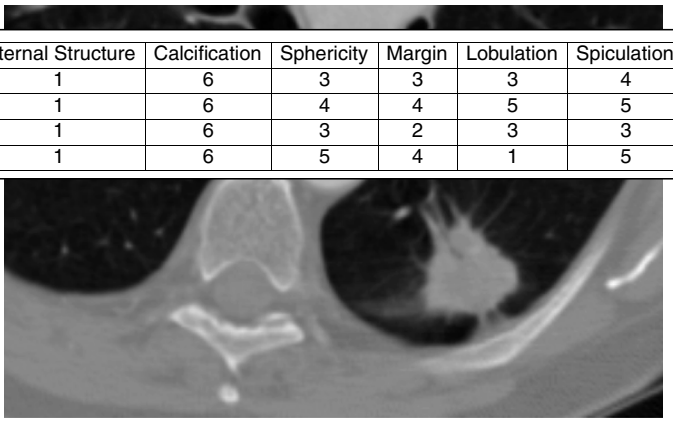
Table 3 Features annotated by radiologists in the LIDC dataset and associated rating system used.

Feature	Subtlety (ordinal)	Internal structure (categorical)	Calcification (categorical)
Rating system	1 Extremely subtle	1 Soft tissue	1 Popcorn
	2	2 Fluid	2 Laminated
	3	3 Fat	3 Solid
	4	4 Air	4 Noncentral
	5 Obvious		5 Central
			6 Absent
Feature	Sphericity (ordinal)	Margin (ordinal)	Lobulation (ordinal)
Rating system	1 Linear	1 Poorly-defined	1 No lobulation
	2	2	2
	3 Ovoid	3	3
	4	4	4
	5 Round	5 Sharp	5 Marked lobulation
Feature	Spiculation (ordinal)	Texture (ordinal)	Malignancy (ordinal)
Rating system	1 No spiculation	1 Nonsolid/ground glass	1 Highly unlikely for cancer
	2	2	2
	3	3 Mixed-solid	3 Indeterminate
	4	4	4
	5 Marked spiculation	5 Solid	5 Highly likely for cancer

3.1.1 *Our use of the Lung Image Database Consortium dataset*

We note a few important subtleties regarding the values and scales used for the spiculation and lobulation feature values. The initial description[36] of the rating systems used to quantify these features specified a value of 1 as highly spiculated (lobulated) and a value of 5 as lacking spiculation (lobulation). However, the present rating system reverses this description,[37] i.e., it designates 1 as a low presence of the feature and 5 as high (as shown in Table 3). Furthermore, it has been reported that there are 399 known cases in the LIDC dataset for which a subset of 100 may have been annotated using the inconsistent rating systems for spiculation and lobulation.[37] It is not known precisely for which 100 of the 399 cases the ratings may have been inconsistently applied (i.e., with a 1 as high and a 5 as low). For this reason, we have omitted these 399 cases in our analysis. However, we observe that there are a number of published articles that employ these physician-quantified labelings of spiculation and lobulation from the LIDC dataset, but none mention the possible mislabelings in the dataset nor the exclusion of these 399 cases from their studies.[6–8,10,11,18,22,32,38–41]

Leaving out these 399 cases, we are left with 4384 nodule annotations that were consistently labeled. The number of annotations used is further reduced from 4384 to 2817 to exclude indeterminate cases (as described in Sec. 3.2). Each nodule may have been assigned between one and four annotations, depending on the level of agreement among the four annotators of the nodule belonging to the first type of structure. The physical nodules lack a universal, unique identifier among the many annotations; thus, it is difficult to ascertain which annotations refer to the same physical nodule in a scan without careful visual inspection. Algorithmically, it is possible to roughly determine which annotations refer to identical nodules by comparing the coordinates and overlap of annotations. However, this process requires somewhat arbitrary choices to be made to determine when multiple annotations may actually refer to the same nodule. For example, one would need to decide at what percentage of overlap, or at what average distance among annotation coordinates, multiple annotations would be declared to refer to the same physical nodule. For these reasons, we treat each annotation as a unique sample for our dataset. More specifically, we treat the quantified features as random vectors, X, and malignancy values as random variables, Y, and we consider each annotation as an independent draw from the joint distribution, (X, Y). Thus, it may be possible that separate dataset samples refer to the same physical nodule; however, we consider these instances to be different realizations of the random quantity, (X, Y), where the source of randomness is from noise (e.g., due to accidental mislabelings) and from natural variations of the quantified feature values (e.g., due to varying annotator



Physician	Subtlety	Internal Structure	Calcification	Sphericity	Margin	Lobulation	Spiculation	Texture	Malignancy
1	5	1	6	3	3	3	4	5	5
2	5	1	6	4	4	5	5	5	5
3	5	1	6	3	2	3	3	5	5
4	5	1	6	5	4	1	5	4	4

Fig. 1 Example nodule from the LIDC dataset with diagnostic feature values from four radiologists.

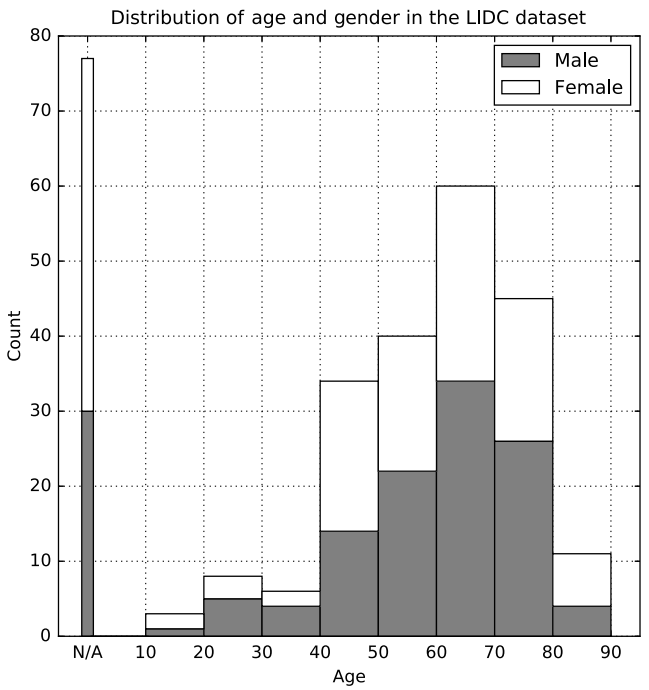


Fig. 2 Distribution of age and sex in the LIDC dataset.

experience and training). This leads to a consistent view of the data. The statistical learning methods then model the conditional probability distribution, $P(Y|X)$, of malignancy, given the quantified feature values. We describe the two statistical learning methods that we use and our methodology in Sec. 3.2.

3.2 *Approximation of Malignancy Category Via Statistical Learning*

We treat the statistical approximation of the malignancy category of nodules as a binary classification problem for "malignant" versus "benign" by thresholding the radiologist-assigned malignancy values so that malignancy values below 3 (i.e., 1 and 2) are categorized as benign and values above 3 (i.e., 4 and 5) are categorized as malignant. We also experimented with treating the output labels as multiple classes (i.e., belonging to one of the 5 values, {1, 2,...,5}) and the output label as a continuous-valued variable (i.e., with real-valued outputs on the interval, [1, 5]), but in both of these cases, the results obtained are inferior to the results presented here. We exclude cases that are labeled by a radiologist as having an indeterminate malignancy (i.e., an assigned value of 3). We also investigated the use of different possible thresholdings, specifically, the two that categorize nodules with a malignancy value of 3 into either the benign or malignant category, and we obtained similar results.

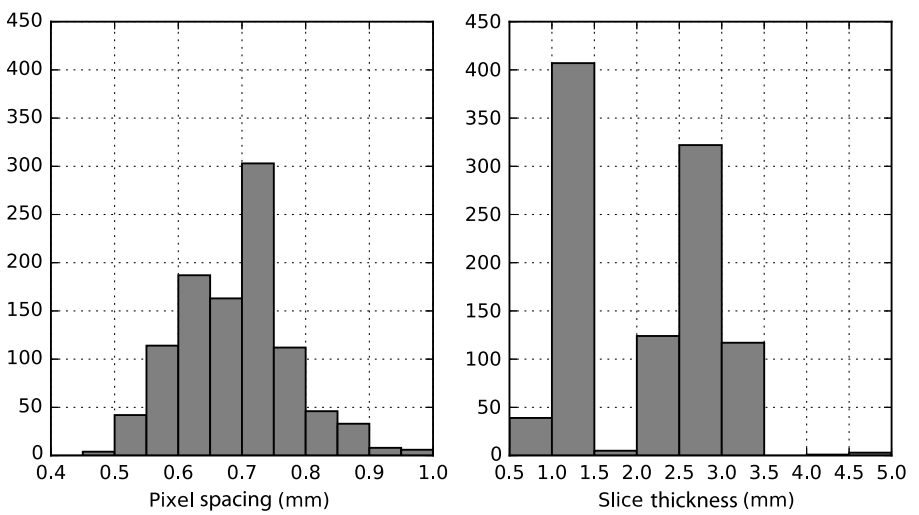


Fig. 3 Distributions of scanner resolutions in the LIDC dataset.

Thus, excluding from the dataset annotations with a malignancy value of 3, we are left with 2817 annotations that remain to be analyzed, each of which consists of the quantified diagnostic image features (which are the input features for a nodule) and the malignancy category (which is the nodule's corresponding target label). The distribution of values for each of the eight diagnostic input features and for the distribution of malignancy is shown in Fig. 4, with the values for each feature defined in Table 3.

To generate the malignancy category from the annotated nodule features algorithmically, we employ two statistical learning techniques for classification. The first, logistic regression, is a linear method, while the second, random forests (which is based on decision trees), is a nonlinear method.[42] These techniques use a subset of the data to learn a mapping—from the diagnostic image features as inputs to the malignancy category as output—during the training (or learning) phase of the algorithms. In the testing phase, the accuracy is evaluated on a subset of the data that the algorithms did not use in any way during the training phase, i.e., on the testing data. For all the numerical experiments herein, each random forest employs 100 decision trees constrained to maximum depths of 8, which was found to be a good parameter value through preliminary testing. We also experimented with various other linear and nonlinear statistical classifiers applied to our dataset, and obtained results similar to those obtained with the logistic regression and random forest classifiers.

In summary, our dataset consists of $N = 2817$ samples, which belong to 530 of the available total of 1018 scans. This results from: (1) removing the 399 possibly inconsistently labeled data and (2) removing the annotations with an "indeterminate" malignancy rating. The binary output-label for the statistical learning algorithms, Y_i $(i = 1, \ldots, N)$, is the thresholded, radiological, malignancy quantification, while the input vector, X_i $(i = 1, \ldots, N)$, is a subset (of length L) of the available radiologist-quantified image features (with $1 \leq L \leq 10$). The particular value of L depends on the numerical experiment being performed. We describe the numerical experiments conducted in this study in Secs. 3.2.1 and 3.2.2.

3.2.1 *Experiment one*

The purpose of the first experiment is to determine how well the radiologists' categorization of malignancy from diagnostic image features can be approximated by statistical learning algorithms. We train both the linear and nonlinear classifier on a randomly chosen subset of the data (with each subset containing approximately 75% of the whole dataset) and test the accuracy on the remaining 25% of the dataset. We repeat this procedure 1000 times to obtain robust statistical results. Examples for the training and testing sets were selected using a uniform random distribution. Thus, the average percentage of malignant cases in both the randomly generated training sets (41.04%) and testing sets (41.02%) is approximately equal to the percentage of

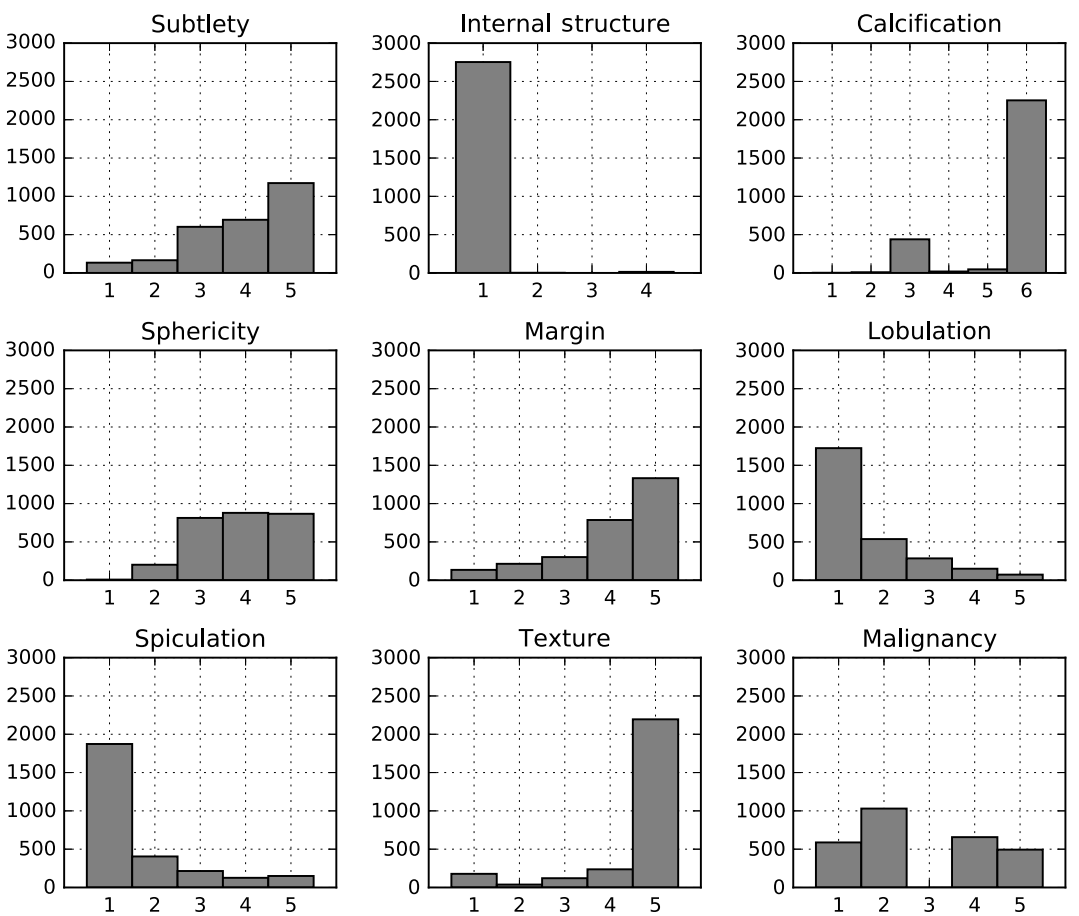


Fig. 4 Distribution of annotation values for image features and malignancy. Note the excluded bin for the indeterminate malignancy value of 3.

malignant cases in the entire dataset (41.04%). Note that because of the overall class imbalance of about 9%, a classifier that always chooses the "benign" class, independent of the input vector, will achieve a classification accuracy of approximately 59%. Therefore, this should be considered the baseline accuracy for comparison with any given classifier's performance.

In addition, we note that ground-truth estimates of two additional features, the nodule diameter and volume, can be made from the nodule boundary contours provided in each radiologist's nodule annotation. We treat these two features separately from the eight diagnostic features given in Table 3 for three reasons. First, since the diameter and volume are given only implicitly by the annotation contours, choices must be made as to how to define and algorithmically extract these two quantities. Second, since their values are given by positive real numbers, these two features differ from the others given in Table 3, whose values are restricted to a small, finite set of positive integers. Third, we are able to derive a theoretical upper bound on the classification accuracy when the diameter and volume features are excluded, as we describe in Sec. 3.2.3. Hence, in the first experiment, we analyze two cases. The first includes diameter and volume estimates, along with the eight diagnostic image features given in Table 3 and, thus, employs a total of 10 input features. The second excludes the diameter and volume features, leaving a total of eight input features. We repeat the process described in the preceding paragraph (involving the 75%/25% training/testing dataset split and 1000 trials) with and without the nodule diameter and volume estimates.

To estimate the diameter of a nodule from a nodule's boundary contour annotations, we first find the maximum pairwise-distance among contour coordinates in each image slice that contains the nodule's boundary contour; then we adjust each of these distances by taking into account the axial-plane scan resolution (which is the same within each slice). The diameter is taken to be the maximum of these distances over all image slices that contain a boundary contour belonging to the nodule. We do not exclude cases where the corresponding maximum-length line segment in each slice passes outside of the nodule or through cavities of the nodule.

To estimate the volume of a nodule using the radiologist-assigned, nodule boundary contours, we first find the area inside of each axial-plane contour (accounting for the axial-plane resolution of the respective scan) using Green's theorem. Next, we multiply the area enclosed by each contour by its respective slice thickness to obtain the volume of each slice. Finally, to arrive at the total volume, we sum each slice volume of the contours annotated as "included" (i.e., the contours marked as enclosing the nodule) and subtract the slice-volumes of those contours that are annotated as "excluded" (e.g., contours that mark the presence of cavities within nodules).

We rely on the following error metrics (defined below) averaged over the 1000 trials performed in experiment one:

1. classification accuracy,
2. sensitivity or true positive rate (TPR), and
3. area-under-the-ROC-curve score (AUC).

The classification accuracy is the percentage of correctly labeled examples in a test dataset when the probabilistic output of the classifier is thresholded at $t = 1/2$. For a given threshold, t, the number of occurrences where the classifier-assigned label and ground-truth label are both malignant is denoted by $\mathrm{TP}(t)$, which represents the number of true positives. Similarly, for a given threshold, t, the number of occurrences where the classifier predicts malignant and the correct label is benign is denoted by $\mathrm{FP}(t)$, which represents the number of false positives. The number of true and false negatives (having their expected meanings) is likewise denoted by $\mathrm{TN}(t)$ and $\mathrm{FN}(t)$, respectively. Sensitivity, or the TPR, is a function of the threshold, t, and is defined as

$$\mathrm{TPR}(t) = \frac{\mathrm{TP}(t)}{\mathrm{TP}(t) + \mathrm{FN}(t)}.$$

Sensitivity is an empirical estimate of the probability that the classifier predicts malignant cases correctly as malignant. We record the average sensitivity, with $t = 1/2$, over the trials. The ROC curve requires the calculation of the false positive rate (FPR), which is defined for a particular probability threshold, t, as

$$\mathrm{FPR(t)} = \frac{\mathrm{FP}(t)}{\mathrm{FP}(t) + \mathrm{TN}(t)}.$$

The FPR is an empirical estimate of the probability that the classifier incorrectly predicts benign cases as malignant. The ROC curve is the parametric curve with coordinates, $[\mathrm{FPR}(t), \mathrm{TPR}(t)]$, for $0 \leq t \leq 1$. This curve is sampled by discretizing the interval, [0, 1], with 101 points, $t_i = i/100$ with $i = 0,1,\ldots,100$. Finally, we calculate the AUC score, which is the area under the ROC curve. It represents an empirical estimate of the probability that the probabilistic output of the classifier is greater for a malignant example than for a benign example. We approximate the AUC score by employing the trapezoidal rule, a standard numerical integration method.

3.2.2 *Experiment two*

The purpose of the second experiment is twofold. First, it is to test the simple hypothesis that increasing the number of diagnostic features used by the classifier improves its ability to assess a nodule's malignancy; second, it is to determine which of the diagnostic features in Table 3 are most useful for assessing the malignancy. In experiment two, we exclude the diameter and volume estimates that were used in experiment one. Rather than using all eight of the remaining features as input, we use subsets of features of size $n = 1,2,\ldots,8$ as input. Thus, with eight input features, there are 255 possible unique subsets, excluding the empty set; each of which is tested. For each possible subset, we train the nonlinear classifier on a random subset of the total data (containing approximately 75% of the 2817 annotations in the dataset);then we test the classification on the remaining 25% of the data to determine classification accuracy. This procedure is repeated 1000 times for each possible subset. Not only will it reveal how the accuracy varies by using an increasing number of features for classification, but it will also allow us to find the most (and least) relevant diagnostic image features for classifying nodules as malignant or benign.

To analyze the effect of increasing the number of features used, we use both a forward and a backward feature-selection process.[42] Forward selection sequentially chooses the best possible feature to add to the subset to improve classification accuracy, starting from the best, single feature. Backward selection, on the other hand, sequentially removes the worst possible feature, starting from all eight features. Thus, they are both a greedy

process. The minimum, mean, and maximum accuracies are calculated at each step of these two selection processes.

We also generate an ad-hoc ranking of the features via the following two metrics:

1. Single-feature accuracy is defined as the accuracy when a specific feature is used by itself.
2. Percent feature-significance is defined as the percentage of cases for which the addition of a specific feature to any subset not containing that feature produces a statistically significant increase in accuracy.

To compute the percent feature-significance for the j'th feature, we gather every subset that does not contain the j'th feature. Since there are eight features being used, this leaves 127 subsets. For each of these subsets, there is a corresponding subset that results from adding the omitted j'th feature. To determine statistical significance, we perform a paired t-test among the classification accuracies obtained over the 1000 trials for these two subsets. The percent feature-significance is determined by counting the number of times, out of the total of 127 subsets that exclude the j'th feature, that the addition of the j'th feature results in a statistically significant increase in accuracy. To complete the ranking, we sort the features by (1) their single-feature accuracy, (2) their percent feature-significance, and (3) the geometric mean of these two metrics.

Finally, for comparison with these, we also list the average feature-importances (or RF feature-importances) computed by the random forest algorithm when all eight features are used. The RF feature-importance metric is found by randomly permuting the values for the j'th feature on the out-of-sample training data (i.e., training data not used in the bootstrap sampling procedure of the random forest algorithm), recording the increase in error due to the permutations, and averaging the error increase over the out-of-sample data.[42] Larger RF feature-importances correspond to features whose value-permutations cause larger increases in error, on average; thus, a larger RF feature-importance signifies a more informative feature. We note that the RF feature-importances are normalized such that the sum over the features is equal to one.

3.2.3 *Maximum attainable accuracies*

Before proceeding to the results of experiments one and two, we describe how a theoretical upper bound for the classification accuracy on the testing data is calculated.

There is not a one-to-one correspondence between the input values (i.e., the quantified diagnostic feature values) and output labels (i.e., the two malignancy categories) in the annotated dataset since there are instances across the entire dataset where multiple annotations of nodules that were made by the radiologists have identical input feature values but correspondingly different output labels. We will refer to such sets of multiple annotations as degenerate groups. There are a total of 151 such groups, involving 1441 annotations out of the total 2817 annotations considered in the dataset. The number of associated annotations in each degenerate group may vary. For example, one degenerate group consists of seven annotations (with identical diagnostic feature inputs), with six of the seven annotations assigning a label of malignant and the remaining one indicating benign.

Keeping the degenerate groups in mind, we consider the situation where the training dataset has been selected and the classifier has determined its parameters from it. Thus, we are at the testing phase. If the classifier is to generalize well, then it should correctly classify all examples in the testing data. However, the classifier's accuracy cannot theoretically be 100% on the testing dataset because it is limited in the following two ways by the presence of degenerate groups in the datasets:

1. If an example in the testing dataset is a member of a degenerate group that has one or more members in the training set, then the classifier is constrained to output the same output value it predicted in the training set. Any example in the testing dataset that is thus mislabeled by the classifier lessens the maximum attainable accuracy that can be achieved on the testing dataset.
2. If an example in the testing dataset is a member of a degenerate group and has members only in the testing set, then a classifier that performs ideally would predict the majority class for the group, thus maximizing the overall classification accuracy. For example, a member of a degenerate group, whose members are only in the testing dataset and whose class label is the minority class for its group, lessens the maximum attainable accuracy achievable on the testing dataset.

Therefore, in the calculation of the theoretical upper bound of the classification accuracy, for any selected partition of the dataset into training and testing datasets, we compute the maximum attainable accuracy for each testing dataset by considering all possible occurrences of the two types above involving degenerate groups. Note that the upper bound on the attainable accuracy does not apply when the diameter and volume features are included as input features due to their real-valued nature.

4 Results

We describe here the results from the experiments described in Secs. 3.2.1 and 3.2.2. In Sec. 5, we discuss and interpret these results.

The results for experiment one are shown in violin plots of classification accuracy in Figs. 5 and 6 and are summarized in Table 4. Each violin plot is a smooth estimate of the probability density function (for the distribution of classification accuracy), which is symmetrically mirrored across a vertical line. Thus, the area of a region within a violin plot that is located between any two chosen values on the vertical axis is proportional to an empirical estimate of the probability of observing a value of the classification accuracy between the two bounding values chosen. In Fig. 5, the results for the linear and nonlinear classifiers are shown in gray and white, respectively, and the results for the cases when both the diameter and volume features are included and excluded are shown with cross-hatching and without, respectively. In Fig. 6, for the nonlinear method when the diameter and volume features are excluded, the distribution of the theoretical maximum accuracy and the achieved accuracy are shown on the left in gray, with and without cross-hatching, respectively. The distribution of their respective differences is shown on the right, in white. Note that, in Fig. 6, the scale for the accuracy plots is given on the left, while the scale for the difference-in-accuracy plot is given on the right.

The results for experiment two are shown in Fig. 7 and in Tables 5 and 6. Figure 7 shows, for a particular fixed

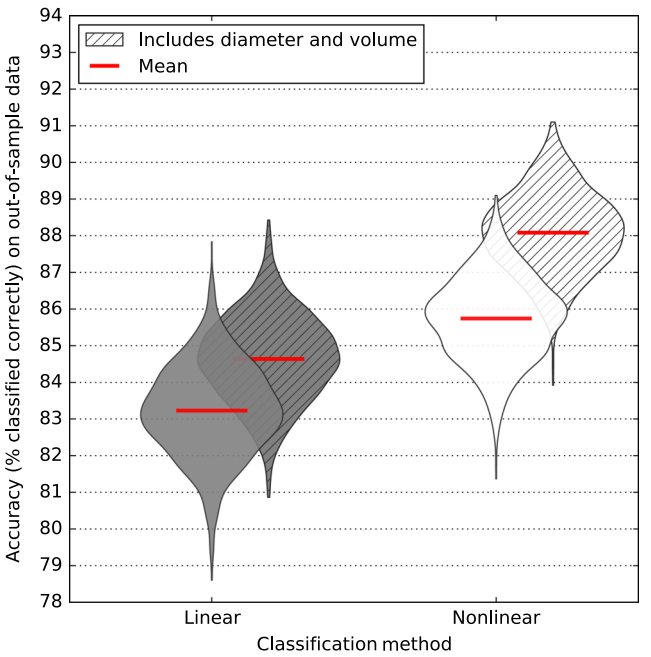


Fig. 5 Distribution of accuracies for experiment one.

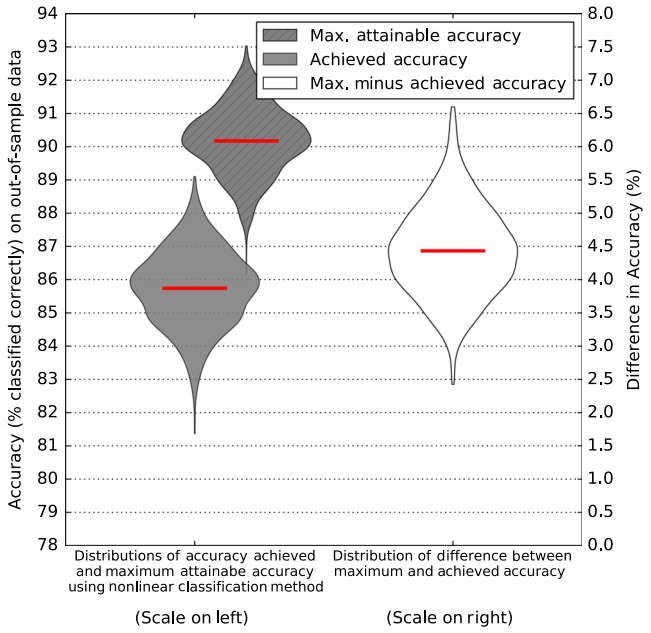


Fig. 6 Comparison of maximum attainable and achieved accuracies with nonlinear classification (diameter and volume excluded).

feature-subset sizes varying from one to eight, the obtained corresponding distribution of accuracies in a violin plot (as described previously). Table 5 lists the feature subsets (and their corresponding classification accuracy) that were obtained at each step of the forward- and backward-selection processes for selecting subsets of size one-greater (that increase accuracy the most) or one-lesser (that decrease accuracy the least), respectively. Table 6 ranks the features according to the measures described in Sec. 3.2.2, i.e., by the single-feature accuracy, percent feature-significance, geometric mean of the two, and the

Table 4 Summary of results from experiment one.

	Accuracy ($t = 1/2$) (%)	TPR ($t = 1/2$)	AUC
Linear classifier, diameter and volume features excluded	83.23 ($\pm$1.252)	0.8013 ($\pm$0.0216)	0.9164 ($\pm$0.0087)
Linear classifier, diameter and volume features included	84.64 ($\pm$1.184)	0.7906 ($\pm$0.0218)	0.9302 ($\pm$0.0079)
Nonlinear classifier, diameter and volume features excluded	85.74 ($\pm$1.141)	0.8430 ($\pm$0.0239)	0.9322 ($\pm$0.0123)
Nonlinear classifier, diameter and volume features included	88.08 ($\pm$1.109)	0.8461 ($\pm$0.0218)	0.9492 ($\pm$0.0070)

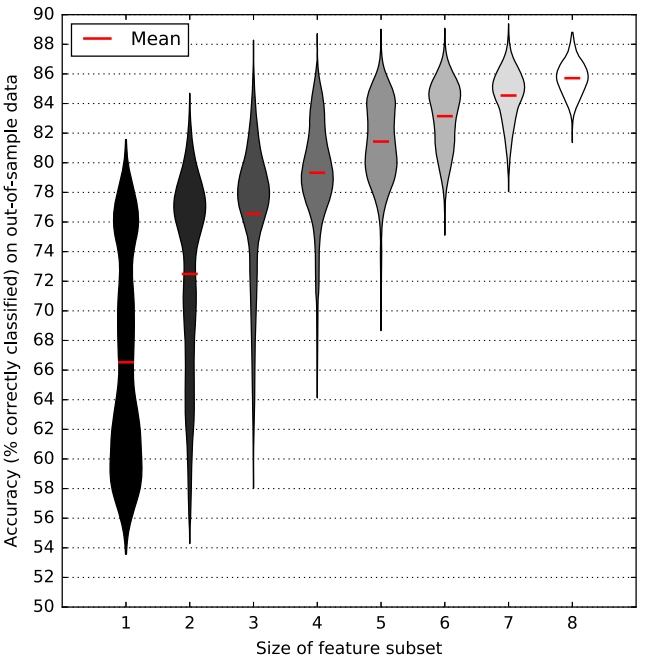


Fig. 7 Distribution of accuracies for experiment two.

average RF feature-importance metric, with the latter two providing an overall ranking of a feature's significance for classification.

5 Discussion

The results for experiment one are summarized in Table 4. When diameter and volume features are excluded, the mean accuracies for the linear and nonlinear classifiers are 83.23 ($\pm$1.25)% and 85.74 ($\pm$1.14)%, respectively, as can be seen in Fig. 5. The nonlinear classifier performs slightly better (by 2.51%) on average, indicating that the malignancy category is slightly better explained by a nonlinear transformation of the input features than by a linear combination (i.e., by a weighted sum) of the features. Correspondingly, we compute average sensitivities (for $t = 1/2$) of 0.801 ($\pm$0.022) and 0.843 ($\pm$0.024), and we calculate average AUC scores of 0.916 ($\pm$0.009) and 0.932 ($\pm$0.012). Our results are comparable to current results in the literature that use features computed from images that may or

Table 5 Results of sequentially choosing the best possible features starting from the single best feature (forward selection) and results of sequentially removing the worst possible feature starting from all features (backward selection).

	Step	Features(s) chosen	% Accuracy		
			Min	Mean (±Std.Dev.)	Max
Forward selection	1	Spiculation	72.82	77.12 (±1.36)	81.57
	2	Spiculation and calcification	74.54	78.76 (±1.31)	82.87
	3	Spiculation, calcification, and subtlety	78.02	82.48 (±1.23)	86.02
	4	Spiculation, calcification, subtlety, and lobulation	80.20	84.77 (±1.15)	88.69
	5	Spiculation, calcification, subtlety, lobulation, and texture	81.08	85.24 (±1.13)	88.86
	6	Spiculation, calcification, subtlety, lobulation, texture, and sphericity	81.21	85.69 (±1.12)	89.08
	7	Spiculation, calcification, subtlety, lobulation, texture, sphericity, and margin	81.37	85.72 (±1.12)	89.19
	8	Spiculation, calcification, subtlety, lobulation, texture, sphericity, margin, and internal structure	81.37	85.71 (±1.13)	88.81
Backward selection	1	Subtlety, calcification, lobulation, spiculation, texture, sphericity, margin, and internal structure	81.37	85.71 (±1.13)	88.81
	2	Subtlety, calcification, lobulation, spiculation, texture, sphericity, and margin	81.37	85.72 (±1.12)	89.19
	3	Subtlety, calcification, lobulation, spiculation, texture, and sphericity	81.21	85.69 (±1.12)	89.08
	4	Subtlety, calcification, lobulation, spiculation, and texture	81.08	85.24 (±1.13)	88.86
	5	Subtlety, calcification, lobulation, and spiculation	80.20	84.77 (±1.15)	88.69
	6	Subtlety, calcification, and lobulation	79.18	83.75 (±1.17)	87.54
	7	Subtlety and calcification	77.14	81.07 (±1.23)	84.69
	8	Subtlety	62.43	67.30 (±1.58)	72.81

Table 6 Rankings (see Sec. 3 for details) of the diagnostic features used.

	Single-feature accuracy	Percent feature significance	Geometric mean	RF feature-importance
Best	(77.12%) Spiculation	(100.00%) Subtlety	(87.12%) Spiculation	(0.2173) Subtlety
	(75.56%) Lobulation	(99.21%) Calcification	(86.24%) Lobulation	(0.2147) Spiculation
	(70.90%) Margin	(98.43%) Spiculation	(82.04%) Subtlety	(0.1818) Lobulation
	(67.30%) Subtlety	(98.43%) Lobulation	(75.72%) Calcification	(0.1737) Calcification
	(63.01%) Texture	(83.46%) Texture	(75.46%) Margin	(0.1116) Margin
	(61.27%) Sphericity	(80.31%) Margin	(72.52%) Texture	(0.0529) Sphericity
	(59.26%) Internal structure	(71.65%) Sphericity	(66.26%) Sphericity	(0.0437) Texture
Worst	(57.79%) Calcification	(62.20%) Internal structure	(60.71%) Internal structure	(0.0044) Internal structure

may not be medically interpretable (as discussed in Sec. 2). For example, Firmino et al.[28] reported an AUC of 0.91 using a combination of image-derived features and LIDC features. Our results yield a similar score but use only the LIDC features. The classification accuracy improves when the diameter and volume features are included, with the mean accuracy for the linear classifier increasing by 1.41%, to 84.64 $(\pm 1.18)\%$ and for the nonlinear classifier increasing by 2.34%, to 88.08 $(\pm 1.11)\%$, as can be seen in Fig. 5. Correspondingly, average sensitivities (at $t = 1/2$) are calculated to be 0.791 ($\pm$0.024) and 0.846 ($\pm$0.022), with average AUC scores of 0.930 ($\pm$0.008) and 0.949 ($\pm$0.007). The observed increase in these metrics,

when diameter and volume are included, demonstrates the relevance and importance of these features for classifying nodule malignancy more accurately.

As shown in the white violin plot in Fig. 6, when the diameter and volume features are excluded, the classification accuracy achieved with the nonlinear classifier varies from a minimum of 2.43% below the theoretical maximum attainable accuracy to a maximum of 6.60% below it. On average, the achieved accuracy is 4.43 (± 0.68)% below the theoretical maximum. We also observe that

$$1 - \text{Maximum Attainable Accuracy} = (1 - \text{Achieved Accuracy}) \times p_d,$$

where p_d is the percentage of the error due to degenerate groups for each of the trials. We find empirically that $p_d = 68.90\%$, on average. In other words, approximately 69% of the errors made are due to degenerate examples in the LIDC dataset, leaving the remaining 31% of the errors to other sources, e.g., to mislabels, insufficiency of the data, or insufficiency of the classifier. The existence of a significant classification accuracy shortfall of 4.43% (on average) raises the question of its potential causes and remedies. It is expected that in other similar and analogous datasets a gap of this type (i.e., a shortfall in classification accuracy from that which is attainable by an ideal classifier on the dataset) would also exist.

This accuracy gap between the theoretical upper-bound for the classification accuracy and the actual classification accuracy achieved is due to overlap in the class-conditional probability distributions, i.e., in the distributions, $P(X|Y)$. Clearly, one source of overlap in the class-conditional distributions is the degenerate groups described in Sec. 3.2.3, i.e., groups of identical input feature vectors with conflicting malignancy-category output labels. A related overlap can occur when there are data samples that are close in feature space (under some distance metric), without input features being necessarily identical, but whose malignancy categories conflict, i.e., when there are "near-degenerate" groups of input features with differing and conflicting malignancy-category output labels. Thus, we define near-degenerate, or ϵ-degenerate, groups as follows: degenerate groups consist of examples where $d(X_i, X_j) = 0$ and $Y_i \neq Y_j$, where $i \neq j$; $d(\cdot,\cdot)$ is some distance metric and X_i is the input vector of example i with corresponding label Y_i. An ϵ-degenerate group, on the other hand, has members, X_i, that are close but not necessarily identical, i.e., $d(X_i, \bar{X}_j) \leq \epsilon$, where $\bar{X}_j$ is some exemplar for the j'th ϵ-degenerate group and $\epsilon \geq 0$ is some chosen nonnegative distance threshold. Such groups could be obtained, e.g., by a clustering algorithm (where $\bar{X}_j$ would correspond to the centroid of the j'th cluster in a centroid-based method). The overlap caused by either type of these data degeneracies reduces the theoretical, maximum-attainable classification accuracy which, in turn, reduces and limits the accuracy achievable with a real classifier.

A data degeneracy analysis (i.e., examining the degree of class-conditional overlap) is useful for determining the theoretical maximum-attainable accuracy; however, this type of analysis is unconcerned with the possible origins of the overlap itself. Thus, to further our understanding of data degeneracy, we identify three potential causes of class-conditional distribution overlap:

1. inherent variability in quantification of diagnostic features,
2. lack of additional informative features, and
3. inadequate sample size and diversity, leading to nonoptimal statistical estimates.

The first potential cause may be difficult to assess and quantify. However, in a study of the related task of lung nodule detection, Armato et al.[43] found substantial variability across radiologists in the LIDC dataset. Currently, the feature values are assigned by radiologists, but if the features were quantified algorithmically, then the variability (a potential source of classification error) would decrease. The second potential cause is supported by the demonstrated increase in classification accuracy due to the addition of diameter and volume feature estimates, which clearly helps to discriminate better between malignant and benign lung nodules and thus reduces the class-conditional overlap. Other used image-derived local and global features (that are not among the features in the LIDC dataset) may have a similar effect. For example, a radiologist could potentially take into account the relative location of the lung nodule when assigning a malignancy class to the nodule. In the LIDC dataset, neither a nodule's relative location within the lung nor its proximity to anatomical structures were annotated by the radiologists viewing and analyzing the CT images, although such information could have an effect on assigning a nodule's malignancy category since certain locations are more probable for lung cancer.[44] We note that estimates of upper bounds for classification accuracy or estimates of class-conditional distribution overlap must be rederived for different sets of input features used for classification and that this may be more difficult to do when real-valued input features, such as diameter and volume, are employed.

As a test of the third potential cause, we run experiments with the nonlinear classifier. For generating training data, we take random samples of the dataset of increasing sizes from 10% to 90% (of the total dataset of 2817 samples) in increments of 10%, setting the remaining data aside. For each randomly chosen training dataset, we randomly choose 10% of the remaining data as testing data. We repeat this 1000 times for all the training set sizes. We observe in Fig. 8 that the mean classification accuracy shows an overall increase of approximately 2.23%, with only a minor increase of about 0.16% after training set subsample sizes greater than 60% of the total dataset are used. This suggests that the maximum amount of information that the classifier can extract from the data for making classifications is already present in samples of sizes of at least 60% of the total dataset size. Hence, the third potential cause of class-conditional overlap is not likely playing a substantial role in our study.

In experiment two (see Fig. 7, and Tables 5 and 6), we observe that including more of the input features generally improves classification accuracy and that there can be a significant variation in classification accuracy for input feature subsets of particular, fixed sizes (see Fig. 7). We find that the standard deviation of the accuracy monotonically decreases with increasing subset size (except in the case subset sizes increasing seven to eight, where it increases very slightly, by 0.01%). Furthermore, we can see in Fig. 7 that the distributions of accuracies can be non-Gaussian. For example, for feature subsets of size five, the distribution of classification accuracy is bimodal and varies from a low of 68.7% to a high of 89.0%. The

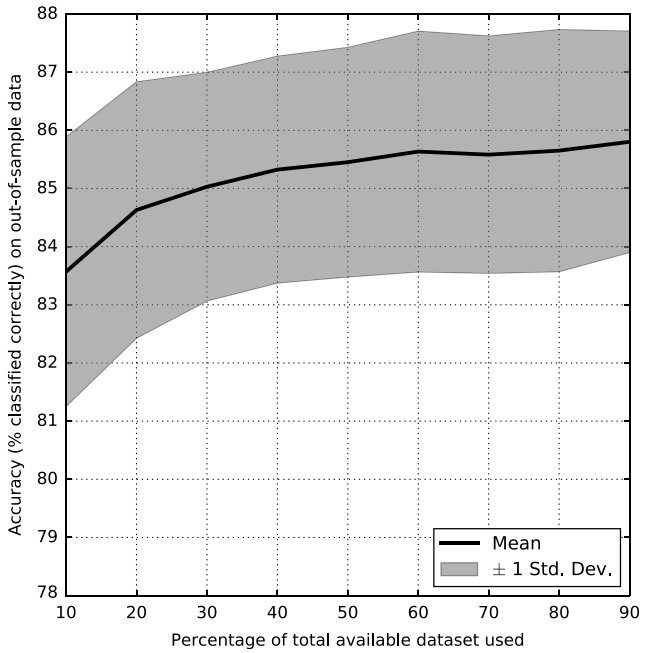


Fig. 8 Validation set accuracy as a function of dataset size.

upper mode is near the 84% mark, and the lower mode is near the 80% mark (which are above and below, respectively, the mean accuracy of 81.5% achieved with subsets of size five). In other words, certain combinations of five input-features lead more frequently to classification accuracies near 84%, while other combinations of five input-features achieve accuracy near 80% more often. This indicates that some feature combinations are better than others for classification, which we discuss in more detail below.

To see which subsets have optimal features with the highest predictive power, we turn to Table 5, which shows the results obtained for the two forward- and backward-selection processes described in Sec. 3.2.2. In the forward-selection process, we find that spiculation is the single best feature for classification. In steps two and three of the process, the calcification and subtlety features are added and, thus, are the two next best features to use. The maximum accuracy (86.02%) that is achieved with just these three features is within one standard deviation (1.13%) of the mean accuracy (85.71%) obtained using all eight features. We observe that by adding one more feature (i.e., lobulation) in step four to this feature triplet, the mean accuracy increases to 84.77% from 82.48%; by adding two additional features (i.e., texture and sphericity) in steps five and six, the accuracy increases to 85.69%, which is 0.02% short of the mean accuracy when using all eight features. Adding the two remaining features (i.e., margin and internal structure) in steps seven and eight does not substantially increase or decrease accuracy. The backward-selection process produces similar results. Generally, the mean accuracy is less with fewer features, and the standard deviation of the accuracy decreases with more features added. We note that there is a partial reflection-symmetry in Table 5 since steps one through five of the backward-selection process result in identical feature subsets as in steps four through eight of the forward-selection process.

Finally, in Table 6, we observe that, according to the geometric mean of the single-feature accuracy and percent feature-significance, the four most relevant features (in decreasing order) are spiculation, lobulation, subtlety, and calcification; the four least relevant features (also, in decreasing order) are margin, texture, internal structure, and sphericity. Interestingly, some features, while not performing as well as others when used individually, have a tendency to improve accuracy when used in addition to other features, i.e., they have a positive synergistic effect. For example, calcification is quite poor as an individual predictor (and, in fact, is the worst individual malignancy indicator) since it ranks last in single-feature accuracy, but the calcification feature shows a very high tendency to significantly increase accuracy when used with other features since it is second best as measured by the percent feature-significance. When individually added to a feature subset, subtlety, spiculation, and lobulation all significantly increase classification accuracy over 98% of the time. Thus, each demonstrates the positive synergistic effect. We also observe in Table 6 that the RF feature-importance metric generally agrees with the geometric mean of the former two features (i.e., the single-feature accuracy and the percent feature-significance), indicating the usefulness of the RF feature-importance metric as a possible surrogate for the combined feature relevance, which requires more work to determine.

6 Conclusion

For the LIDC dataset, we have shown that the malignancy label of a nodule can be accurately classified (4.14%, on average, below the theoretical maximum accuracy of an ideal classifier) when only quantified, diagnostic image features are used as inputs to standard statistical learning methods. We have also shown that the accuracy of the linear and nonlinear classifiers can be improved by 1.41% and 2.34% (from 83.23% to 84.64% and from 85.74% to 88.08%) by including diameter and volume estimates, respectively. Using only the radiologist quantifications of image features, we have also achieved an average AUC score of 0.932 with a nonlinear classifier, which improves to an average AUC score of 0.949 when diameter and volume features are included. Our results are comparable to those obtained with other approaches currently reported in the literature. These other approaches, by contrast, often use image-based features that are extracted algorithmically, which, in some cases, may not be medically interpretable. Our positive results both support and motivate the further consideration of the CAD paradigm that first extracts and approximates a set of radiologist-interpretable diagnostic image features from a CT scan and then uses these features as inputs to a statistical learning method for classifying the corresponding nodule as malignant or benign. By design, this approach is both modular and transparent in the following senses:

1. Modularity comes from the independent quantification of constituent features used for malignancy classification and allows a radiologist the choice to only use the desired parts.
2. Transparency comes from the restriction of input features used for malignancy classification to only those which are medically interpretable, which allows for the components of the CAD system to be medically informative and intuitive for radiologists.

This CAD approach could, therefore, be useful as an informative second-reader. Furthermore, we have ranked the lung nodule features given in the LIDC dataset according to their predictive power (when used both singly and in combination

with other features), showing that the four most relevant features for classification are (in decreasing order of relevance) spiculation, lobulation, subtlety, and calcification.

Future work for such a CAD approach will focus on the quantification and accurate approximation of interpretable, diagnostic image features extracted from CT scans (such as the most relevant four mentioned above). We believe the task of algorithmically quantifying such features is much more challenging than the determination of a lung nodule's malignancy category from already accurately quantified nodule-features due to the limited amount of validation data available by which to measure the error in quantifying these physical features and the current qualitative and imprecise definitions of these features. The former requires relatively large datasets of lung nodule images with diagnostic image features quantified by experts to provide ground-truth labels, while the latter requires the standardization of the diagnostic image features to be used, according to some standard terminology set, e.g., the RadLex ontology.[45] Indeed, mentions of the need for standardization of image features have been made recently in the literature.[3,46]

Although challenging to implement, the CAD approach discussed above has the potential to facilitate the diagnostic work of a radiologist by automatically quantifying relevant and interpretable features of lung nodules and offering an accurate, medically useful summary of a region of interest within a CT image for a radiologist's consideration.

Appendix

In the process of working with the LIDC dataset, we have developed an object-relational mapping software library for (1) querying scans, annotations, and contours in a simple query language style, (2) performing common operations on retrieved objects (e.g., computing diameter or volume), and (3) viewing annotation data on top of CT scan data. We have made the software freely available at Ref. 47.

Disclosures

No conflicts of interest, financial or otherwise, are declared by the authors.

Acknowledgments

The authors wish to thank Dr. Juan Luis-Jorge, M.D., for helpful discussions.

References

1. J. Erasmus et al., "Solitary pulmonary nodules: part I. Morphologic evaluation for differentiation of benign and malignant lesions," *RadioGraphics* **20**(1), 43–58 (2000).
2. J. Erasmus, H. McAdams, and J. Connolly, "Solitary pulmonary nodules: part II. Evaluation of the indeterminate nodule," *RadioGraphics* **20**(1), 59–66 (2000).
3. H. Kim et al., "Quantitative computed tomography imaging biomarkers in the diagnosis and management of lung cancer," *Invest. Radiol.* **50**, 571–583 (2015).
4. "NCI dictionary of cancer terms—National Cancer Institute," 2015, http://www.cancer.gov/publications/dictionaries/cancer-terms (March 2016).
5. S. Iwano et al., "Computer-aided diagnosis: a shape classification of pulmonary nodules imaged by high-resolution CT," *Comput. Med. Imaging Graphics* **29**, 565–570 (2005).
6. D. Raicu, "Modelling semantics from image data opportunities from LIDC," *Int. J. Biomed. Eng. Technol.* **3**(1), 83–113 (2009).
7. D. Zinovev et al., "Predicting radiological panel opinions using a panel of machine learning classifiers," *Algorithms* **2**, 1473–1502 (2009).
8. A. K. Dhara et al., "Measurement of spiculation index in 3D for solitary pulmonary nodules in volumetric lung CT images," *Proc. SPIE* **8670**, 86700K (2013).
9. G. Li et al., "Semantic characteristics prediction of pulmonary nodule using artificial neural networks," in *35th Annual Int. Conf. of the IEEE, Engineering in Medicine and Biology Society (EMBC'13)*, pp. 5465–5468, IEEE (2013).
10. G. Zhang, N. Xiao, and W. Guo, "Spiculation quantification method based on edge gradient orientation histogram," in *Int. Conf. on Virtual Reality and Visualization (ICVRV'14)*, pp. 86–91, IEEE (2014).
11. R. Niehaus et al., "Toward understanding the size dependence of shape features for predicting spiculation in lung nodules for computer-aided diagnosis," *J. Digital Imaging* **28**, 704–717 (2015).
12. F. Ciompi, "Bag-of-frequencies: a descriptor of pulmonary nodules in computed tomography images," *IEEE Trans. Med. Imaging* **34**, 962–973 (2015).
13. H. Krewer et al., "Effect of texture features in computer aided diagnosis of pulmonary nodules in low-dose computed tomography," in *IEEE Int. Conf. on Systems, Man, and Cybernetics (SMC'13)*, pp. 3887–3891, IEEE (2013).
14. F. Han et al., "Texture feature analysis for computer-aided diagnosis on pulmonary nodules," *J. Digital Imaging* **28**, 99–115 (2015).
15. A. Depeursinge et al., "Automated classification of usual interstitial pneumonia using regional volumetric texture analysis in high-resolution computed tomography," *Invest. Radiol.* **50**(4), 261–267 (2015).
16. A. El-Baz et al., "A novel shape-based diagnostic approach for early diagnosis of lung nodules," in *IEEE Int. Symp. on Biomedical Imaging: From Nano to Macro*, pp. 137–140, IEEE (2011).
17. E. Tac and A. Uur, "Shape and texture based novel features for automated juxtapleural nodule detection in lung CTs," *J. Med. Syst.* **39**, 46 (2015).
18. A. Kaya and A. B. Can, "A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics," *J. Biomed. Inf.* **56**, 69–79 (2015).
19. S. G. Armato, III et al., "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," *Med. Phys.* **38**(2), 915–931 (2011).
20. R. Wiemker et al., "Repeatability and noise robustness of spicularity features for computer aided characterization of pulmonary nodules in CT," *Proc. SPIE* **6915**, 691511 (2008).
21. S. Kido et al., "Fractal analysis of small peripheral pulmonary nodules in thin-section CT: evaluation of the lung-nodule interfaces," *J. Comput. Assisted Tomogr.* **26**(4), 573–578 (2002).
22. R. Wiemker et al., "Correlation of emphysema score with perceived malignancy of pulmonary nodules: a multi-observer study using the LIDC-IDRI CT lung database," *Proc. SPIE* **7263**, 726310 (2009).
23. B. Ganeshan et al., "Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage," *Cancer Imaging* **10**(1), 137–143 (2010).
24. S. H. Lee et al., "Usefulness of texture analysis in differentiating transient from persistent part-solid nodules (PSNs): a retrospective study," *PLoS One* **9**, e85167 (2014).
25. J. Y. Son et al., "Quantitative CT analysis of pulmonary ground-glass opacity nodules for the distinction of invasive adenocarcinoma from pre-invasive or minimally invasive adenocarcinoma," *PLoS One* **9**, e104066 (2014).
26. T. W. Way et al., "Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours," *Med. Phys.* **33**(7), 2323 (2006).
27. S. K. Dilger et al., "Improved pulmonary nodule classification utilizing quantitative lung parenchyma features," *J. Med. Imaging* **2**(4), 041004 (2015).
28. M. Firmino et al., "Computer-aided detection (CADe) and diagnosis (CADx) system for lung cancer with likelihood of malignancy," *Biomed. Eng. Online* **15**, 2 (2016).
29. A. Depeursinge et al., "Three-dimensional solid texture analysis in biomedical imaging: review and opportunities," *Med. Image Anal.* **18**, 176–196 (2014).

30. W. Mullally et al., "Segmentation of nodules on chest computed tomography for growth assessment," *Med. Phys.* **31**(4), 839 (2004).
31. A. Reeves et al., "On measuring the change in size of pulmonary nodules," *IEEE Trans. Med. Imaging* **25**, 435–450 (2006).
32. T. Messay, R. C. Hardie, and T. R. Tuinstra, "Segmentation of pulmonary nodules in computed tomography using a regression neural network approach and its application to the lung image database consortium and image database resource initiative dataset," *Med. Image Anal.* **22**, 48–62 (2015).
33. T. McInerney and D. Terzopoulos, "Deformable models in medical image analysis: a survey," *Med. Image Anal.* **1**(2), 91–108 (1996).
34. J. S. Suri et al., "Shape recovery algorithms using level sets in 2-D/3-D medical imagery: a state-of-the-art review," *IEEE Trans. Inf. Technol. Biomed.* **6**(1), 8–28 (2002).
35. T. Heimann and H.-P. Meinzer, "Statistical shape models for 3D medical image segmentation: a review," *Med. Image Anal.* **13**, 543–563 (2009).
36. M. F. McNitt-Gray et al., "The lung image database consortium (LIDC) data collection process for nodule detection and annotation," *Acad. Radiol.* **14**, 1464–1474 (2007).
37. The Cancer Imaging Archive, "Lung image database consortium—reader annotation and markup—annotation and markup issues/comments," 2015, https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI (December 2015).
38. W. H. Horsthemke, D. S. Raicu, and J. D. Furst, "Bridging the evaluation gap challenge to diagnostic labeling of pulmonary nodules," in *DePaul CDM Research Symp.* (2008).
39. R. Wiemker et al., "Agreement of CAD features with expert observer ratings for characterization of pulmonary nodules in CT using the LIDC-IDRI database," *Proc. SPIE* **7260**, 72600H (2009).
40. P. Opulencia et al., "Mapping LIDC, RadLex, and lung nodule image features," *J. Digital Imaging* **24**, 256–270 (2011).
41. D. Zinovev et al., "Probabilistic lung nodule classification with belief decision trees," in *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society, EMBC*, pp. 4493–4498, IEEE (2011).
42. J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, Springer Series in Statistics, Vol. **1**, Springer, Berlin (2001).
43. S. G. Armato et al., "The lung image database consortium (LIDC): an evaluation of radiologist variability in the identification of lung nodules on CT scans," *Acad. Radiol.* **14**, 1409–1421 (2007).
44. H. T. Winer-Muram, "The solitary pulmonary nodule," *Radiology* **239**, 34–49 (2006).
45. Radiological Society of America, "RadLex," 2016, http://www.rsna.org/RadLex.aspx (February 2016).
46. M. J. Nyflot et al., "Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards," *J. Med. Imaging* **2**(4), 041002 (2015).
47. M. C. Hancock, "Pylidc-An object relational mapping for the LIDC dataset using sqlalchemy,"https://github.com/pylidc/pylidc/ (1 July 2016).

Matthew C. Hancock is a PhD candidate in applied and computational mathematics at Florida State University, pursuing research involving pattern-recognition, image- and signal-processing, and data visualization in the realm of medical imaging analysis.

Jerry F. Magnan is an associate professor at the Mathematics Department, Florida State University, in the area of applied and computational mathematics. He received his PhD in physics from the University of Miami and was a visiting assistant professor at Northwestern University in the Department of Engineering Sciences and Applied Mathematics. His research interests involve the mathematical modeling and analysis of nonlinear phenomena in natural and man-made systems, and the use of machine learning in scientific and industrial problems.