Combination of computer extracted shape and texture features enables discrimination of granulomas from adenocarcinoma on chest computed tomography

Mahdi Orooji
Mehdi Alilou
Sagar Rakshit
Niha Beig
Mohammad Hadi Khorrami
Prabhakar Rajiah
Rajat Thawani
Jennifer Ginsberg
Christopher Donatelli
Michael Yang
Frank Jacono
Robert Gilkeson
Vamsidhar Velcheti
Philip Linden
Anant Madabhushi

SPIE.

Combination of computer extracted shape and texture features enables discrimination of granulomas from adenocarcinoma on chest computed tomography

Mahdi Orooji,[a,*] Mehdi Alilou,[a] Sagar Rakshit,[b] Niha Beig,[a] Mohammad Hadi Khorrami,[a] Prabhakar Rajiah,[c] Rajat Thawani,[a] Jennifer Ginsberg,[d] Christopher Donatelli,[e] Michael Yang,[f] Frank Jacono,[e] Robert Gilkeson,[g] Vamsidhar Velcheti,[h] Philip Linden,[d] and Anant Madabhushi[a]
[a]Case Western Reserve University, Department of Biomedical Engineering, Cleveland, Ohio, United States
[b]Cleveland Clinic Foundation, Department of Medicine, Cleveland, Ohio, United States
[c]UT Southwestern, Department of Radiology, Dallas, Texas, United States
[d]University Hospitals Cleveland Medical Center, Division of Thoracic and Esophageal Surgery, Cleveland, Ohio, United States
[e]University Hospitals Cleveland Medical Center, Division of Pulmonary and Critical Care, Department of Medicine, Cleveland, Ohio, United States
[f]University Hospitals Cleveland Medical Center, Department of Pathology, Cleveland, Ohio, United States
[g]University Hospitals Cleveland Medical Center, Department of Radiology, Cleveland, Ohio, United States
[h]Cleveland Clinic Foundation, Department of Solid Tumor Oncology, Cleveland, Ohio, United States

Abstract. Differentiation between benign and malignant nodules is a problem encountered by radiologists when visualizing computed tomography (CT) scans. Adenocarcinomas and granulomas have a characteristic spiculated appearance and may be fluorodeoxyglucose avid, making them difficult to distinguish for human readers. In this retrospective study, we aimed to evaluate whether a combination of radiomic texture and shape features from noncontrast CT scans can enable discrimination between granulomas and adenocarcinomas. Our study is composed of CT scans of 195 patients from two institutions, one cohort for training ($N = 139$) and the other ($N = 56$) for independent validation. A set of 645 three-dimensional texture and 24 shape features were extracted from CT scans in the training cohort. Feature selection was employed to identify the most informative features using this set. The top ranked features were also assessed in terms of their stability and reproducibility across the training and testing cohorts and between scans of different slice thickness. Three different classifiers were constructed using the top ranked features identified from the training set. These classifiers were then validated on the test set and the best classifier (support vector machine) yielded an area under the receiver operating characteristic curve of 77.8%. © 2018 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JMI.5.2.024501]

1 Introduction

The vast majority of malignant peripheral lung nodules are adenocarcinomas.[1] Granulomas, which represent small areas of inflammation in tissue as response to infections or foreign agents, represent one of the major confounders to cancer on computed tomography (CT).[2] Most granulomas, which have a characteristic CT spiculated appearance, are also fluorodeoxyglucose avid and, thus, are indistinguishable from carcinomas based upon the current noninvasive modalities.[3,4]

Over one million people in the US are annually subjected to a CT guided or bronchoscopic biopsy, and over 60,000 are subjected to a surgical wedge resection for pathologic confirmation of a pulmonary nodule found on a CT scan.[5] However, more than 26% of the suspicious pulmonary nodules on a CT scan that are biopsied or resected are identified as benign, translating to nearly $600M being spent annually on unnecessary and invasive surgical procedures.[5,6] Some patients with a nodule on a CT scan may not undergo a surgical procedure for diagnostic confirmation and may be followed up with repeat CT scans to evaluate whether the nodule is increasing in size. However, granulomas and slow growing cancers may increase at the same rate, thus rendering the follow-up CT scans largely noninformative.[7]

Employing machine learning technologies for the analysis of pulmonary nodules dates back to the early 1990s,[8–10] and the benefits of computer-aided diagnosis algorithms in localization and characterization of lung lesions on CT are well investigated.[11–15] While a number of papers have looked at radiomic or computerized feature-based analysis of lung CT to distinguish between benign and malignant nodules,[16,17] there has not been much work done to distinguish granulomas from malignancies. Dennie et al.[18] explored the role of texture features for distinguishing granulomas from malignancies on a single cohort of 55 cases. While they looked at the role of image texture for nodule characterization, we are not aware of any work that has attempted to jointly explore the role of computer-extracted features of nodule texture and shape to distinguish granulomas from adenocarcinomas.

In this work, for the first time, we explore the role of a combination of radiomic texture and shape features derived from routine noncontrast CT scans to distinguish granulomas from adenocarcinomas. We used the top discriminating features and combined them with a machine learning classifier, which was

*Address all correspondence to: Mahdi Orooji, E-mail: mxo143@case.edu

optimized on a training set to assign a probability of a nodule being an adenocarcinoma. The classifier was then independently validated on a separate test set from a different institution. Furthermore, we compared this classifier with the manual interpretations of two expert human readers.

The rest of the paper is organized as follows. In Sec. 2, the methods and materials including the patient selection criteria and the description of the employed radiomics is described. The results of the paper are presented in Sec. 3 and they are discussed in Sec. 4. Finally, the concluding remarks are presented in Sec. 5.

2 Methods

2.1 Patient Selection, Annotation, and Chest Computed Tomography Acquisition

In this retrospective study, we included patients from two sites, who had a suspicious lung nodule on CT scan and underwent surgical resection. All scans acquired for this study were collected as part of an Institutional Review Board-approved, HIPAA-compliant protocol. These scans were obtained as part of standard of care clinical management for these patients, all of whom subsequently went on to have a surgical wedge resection for excision of a suspicious nodule. All the information provided was deidentified and the need for an informed consent was waived. Histology was confirmed by an anatomical pathologist based off visual interrogation of the surgical specimen.

One hundred and thirty-nine solitary nodules were obtained from site 1 (70 carcinomas and 69 granulomas) and were used as the training set. Fifty six solitary nodules (34 carcinomas and 22 granulomas) were acquired from site 2 and composed of the independent test set. Patients with multiple solitary nodules were
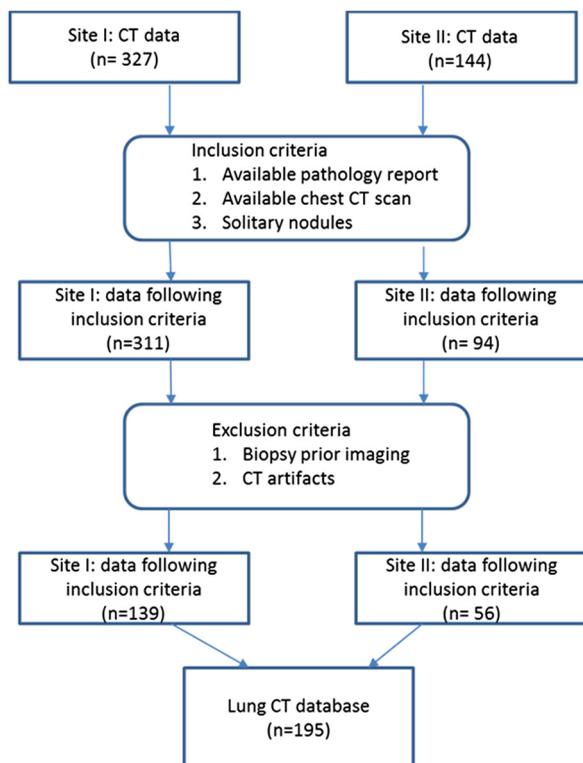
Fig. 1 Inclusion and exclusion criteria for patient selection in two cohorts.

excluded. Figure 1 illustrates the inclusion and exclusion criteria.

2.2 Lesion Segmentation and Computer-Extracted Texture and Shape Features

The region of interest (RoI) containing the lesion was manually segmented across contiguous slices by an expert cardiothoracic radiologist with 20 years of experience in interpreting chest CT scans via a hand-annotation tool in 3D-Slicer® software.[19] To evaluate the interobserver variability in the lesion segmentation and also to show the effect of it on the stability of the top radiomics, a subset of 10 adenocarcinoma and 10 granuloma lesions is randomly picked and resegmented by an expert radiologist with 14 years of experience in cardiothoracic radiology. Note that the second reader was completely blinded to the segmentation of the first reader. A total of 645 two-dimensional (2-D) texture and 24 three-dimensional (3-D) shape features were extracted from the lesion area. Texture features were extracted in 2-D instead of 3-D, since the available retrospective CT volumes were all anisotropic. After extracting per-voxel based features within the nodule of interest, five statistics (mean, variance, minimum, maximum, and entropy) were calculated for each nodule. All feature calculations were implemented using MATLAB® 2014b platform (Mathworks, Natick, Massachusetts). The description of the extracted features is given in the following subsections.

2.2.1 Haralick texture features

Haralick features are based on quantifying the spatial gray-level co-occurrence matrix (GLCM) within local neighborhoods around each pixel in an image.[20] A total of 13 Haralick texture descriptors were calculated from every lesion by computing the median of the statistics derived from the corresponding co-occurrence matrices. The GLCM is calculated within 5×5 windows, and to avoid the sparse a GLCM, the gray-level images were uniformly quantized to 64 intensity levels.

2.2.2 Laws texture features

Laws features use 5×5 separable masks[21] that are symmetric or antisymmetric to extract level (L), edge (E), spot (S), wave (W), and ripple (R) patterns on an image. The convolution of these masks with the image resulted in a total of 25 distinct Laws feature representations.

2.2.3 Laws–Laplacian pyramids, level 2

Laplacian pyramids allow for capturing multiscale edge representations via a set of band-pass filters.[22] First, the original image is convolved with a Gaussian kernel. The Laplacian is then computed as the difference between the original image and the low-pass filtered image. The resulting image is then subsampled by a factor of 2, and the filter subsample operation is repeated recursively. This process is continued to obtain a set of band-pass filtered images (since each is the difference between two levels of the Gaussian pyramid). The output of the second level of the pyramid was then subjected to feature extraction via the Laws operators and to obtain the corresponding Laws–Laplacian feature. Similar to Laws features, and corresponding to level, edge, spot, wave, and ripple patterns, a total of 25 distinct Laws–Laplacian pyramid features were extracted for each representative slice that contained the lesion.

2.2.4 Gray-level features

This feature class is composed of four voxel-level intensity-based features, including mean, median, range, and standard deviation.[23]

2.2.5 Gabor filter features

The Gabor function is obtained as a convolution of a sinusoidal-modulated Gaussian kernel function.[24] These multiscale, steerable filters enable extraction of dominant oriented textures within the nodule. A total of 48 Gabor filters were extracted across different frequencies and scales were extracted for each image.

2.2.6 Gradient features

These features represent the directional change in the intensity values of pixels within the RoI.[25] These gradient features are

obtained via the Sobel edge kernel across each direction (X and Y) and diagonal directions (XY and YX).

2.2.7 Local binary pattern

These features involve comparing the intensity of a pixel under consideration with the pixels within its neighborhood and creating a binary vector based off whether the intensity of the center pixels is greater or less than each of the neighborhood pixels.[26] Thus, the local binary pattern (LBP) operation results in an 8-bit code-word describing local neighborhood around every pixel. Table 1 describes the texture features employed in this work.

2.2.8 3-D shape features

Since irregularities in tumors shape can result from its internal heterogeneity and differences in the growth pattern, we described

Table 1 Texture features evaluated in this work.

Feature category	Descriptor	Intuitive description
Haralick features (repeated occurrence of gray level configuration in the texture represented via the GLCM, which varies rapidly with distance in fine textures and slowly in large textures; total of 65 features)	Inverse difference moment (IDM)	IDM is a reflection of the presence or absence of uniformity, and hence is a measure of local regions of homogeneity
		High IDM: higher presence of locally uniform windows in GLCM
		Low IDM: Higher presence of locally heterogeneous windows in GLCM
	Correlation	Quantifies the linear patterns in an image based on the distance parameter.
	Sum entropy	Measure of GLCM relationship to distribution of intensity with respect to entropy. Entropy is the measure of disorder.
	Sum variance	Measure of GLCM relationship to distribution of intensity with respect to variance
		High sum variance: greater standard deviation of sum average
		Low sum variance: low standard deviation of sum average
Laws features (total of 125 features)	E5, L5, S5,W5,R5 (combination in both X and Y directions)	E-Edges
		L-Level
		S-Spots
		W-Wave
		R-Ripple
Laplacian pyramids (total of 125 features)		Multiresolution filters capture edges at different levels
Gray level features (total of 20 features)		The basic intensity-based features including mean, median, range, and standard deviation.
Gabor features (total of 240 features)		Oriented textures via changes in direction and scale; capture microarchitectures
Gradient features (total of 65 features)		Represent the directional change in the intensity values of pixels in the RoI
Local binary pattern (total of 5 features)		Thresolding the window with the center pixel value.

Features	Description
Size (three features)	Including width, height, and depth of bounding box
Area (two features)	From 2-D slices of each nodule
Perimeter (two features)	From 2-D slices of each nodule
Eccentricity (two features)	Foci of the ellipse and to major axis length
Extend (two features)	Ratio of pixels in the region to pixels in the total bounding box
Compactness (two features)	Ratio of the perimeter squared to the product of 4π and area
Radial distance (two features)	Distances from center of each slice to contour points
Roughness (two features)	Perimeter of slices divided by convex perimeter
Elongation (two features)	From major and minor axis
Convexity (two features)	From convex hull
Equivalent diameter (two features)	Diameter of circle with same area of slices
Sphericity (one feature)	3-D compactness

tumors shape with the following radiomic features. The convexity of the tumor border was calculated as the ratio of the volumes contained within the tumor border to its convex hull (i.e., the smallest convex polygon enclosing a planar tumor region). We also computed the following shape features: width, height, depth, perimeter, area, eccentricity, compactness, radial distance, roughness, elongation equivalent diameter, and 3-D sphericity of the nodule. Note that the width, height, depth as well as sphericity features were calculated in 3-D space. The remaining features were computed in 2-D on a slice by slice basis. The mean and standard deviation of each feature were computed across all the pixels and over all slices containing the tumor. A description of the shape features is presented in Table 2.

2.3 Stability Evaluation of Radiomic Features

While feature discriminability is critical to be able to distinguish the two pathologies of interest, an important consideration in constructing generalizable and stable classifiers is to ensure that the features invoked in constructing the machine learning classifier are relatively robust across sites and slice thickness for a specific pathology (i.e., granulomas or adenocarcinomas). The criticality of invoking stable features has been well documented in the radiology domain,[27–29] however, relatively little attention has to be paid to this issue in the context of radiomics.[30,31] In this work, we also investigated the sensitivity of radiomic feature expression to variations in slice thickness and across the different acquisition sites and scanners. In this paper, we use the preparation-induced instability (PI) score defined by Leo et al.[30] to measure robustness of the feature value. The PI

value is a number between 0 and 1 and calculated per radiomics. The PI closer to 0 implies that the corresponding radiomics is more stable between two cohorts of interests.

In this paper, the PI score was calculated between the testing cohort and the training cohort to quantitatively demonstrate the robustness of the top radiomic features. Additionally, the PI score was calculated for a subset of the training data to investigate the variance in radiomic features between scans with a slice thickness of less than 3 mm versus those scans with a slice thickness more than 3 mm.

To provide a better evaluation and more valid conclusion, a subset of 20 lesions (10 from each pathology) is randomly selected and segmented by an independent cardiothoracic radiologist. The Dice coefficient,[32] undersegmentation error, and oversegmentation error between the segmentation of the readers, and also the area under the curve (AUC) of diagnostics of each one, was reported as a measure of inter-reader variability. In addition, the PI value of top radiomics, from the segmentation of two experts, is expressed as a measure of radiomics stability versus interreader variability.

2.4 Statistical Analysis

We employed a consensus clustering approach to graphically compare the within-class and between-classes correlation for the features identified as discriminatory by feature selection within the training set.[33] To perform the consensus clustering, the similarity between different nodules, assessed according to a distance in the space of the top ranked features, was first determined. The closer two nodules were to each other, the higher the likelihood that they both belong to the same cluster. Clusters were determined by invoking the idea that nodules within a cluster should have a high intraclass correlation, whereas nodules belonging to different clusters should have minimal correlation. Figure 2(a) shows the consensus clustering result generated for the nodules in the training set within the space of the top ranked radiomic features. Figure 2(b) shows the corresponding consensus clustering obtained for the nodules in the training set when only invoking the Haralick texture features.

2.5 Features-Based Classification and Validation on Independent Test Set

Following feature extraction, sequential forward feature selection was employed to identify the most discriminating subset of features.[34] To mitigate bias in feature selection and classifier training, a threefold (onefold held-out for testing), patient-stratified, cross-validation scheme was used for classifier constructing using the instances within the training set, and cross validation was repeated a 1000 times.

The top six features identified by the feature selection approach were then used to train and lock down linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and support vector machine (SVM) classifiers. These classifiers were then applied to predict the class labels of the nodules in the validation set. Each nodule was assigned a probability of being an adenocarcinoma. Area under the receiver operating characteristic (AUC) curve was used to evaluate the performance for each of the SVM, LDA, and QDA classifiers on the validation set. A Bayesian cross-validation extensive search on the normalized data was employed to find the best box constraint and the sigma value of radial basis function (RBF) kernel of SVM

(a)   (b)
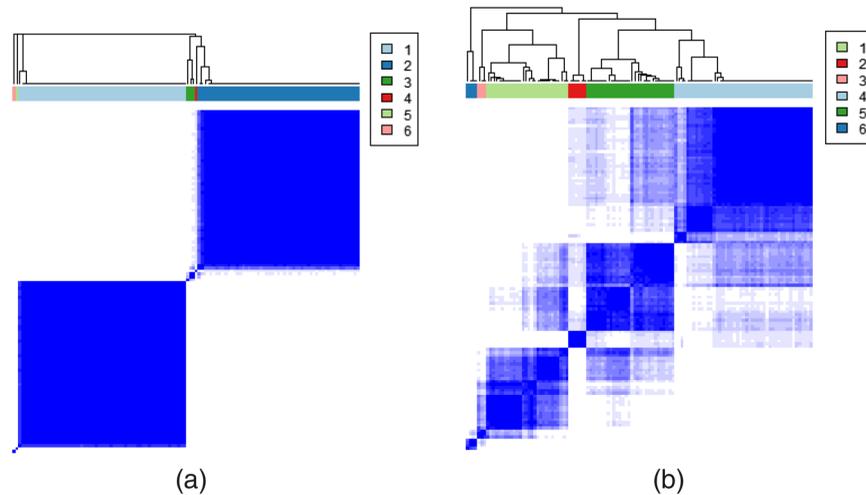
Fig. 2 Consensus clustering of the cases in the training set when the data were clustered into six partitions. The clustering was done (a) on the most predictive shape and texture features identified during feature selection and (b) for Haralick texture features alone. While the result of clustering in panel (a) reveals two primarily distinct clusters of adenocarcinomas and granulomas and (b) reveals multiple disjointed clusters of varying sizes. The comparison of (a) and (b) shows good inherent discriminability between adenocarcinoma and granuloma when a combination of the most predictive shape and texture features is considered, as opposed to Haralick features alone.

classifier. The employed box constraint and sigma are 0.431 and 14.07, respectively.

2.6 Human–Machine Comparison

The classification performance of our SVM classifier was compared against the diagnosis of two experts, one a board certified attending radiologist with 12 years of experience in thoracic radiology and a pulmonologist with 2 years of experience in reading chest CT scans. Both experts were blinded to the true histopathologic diagnosis of the 56 cases, which comprised the test set. Each reader was asked to assign a score between 1 and 5 to each nodule, with 1 referring to a high confidence that the nodule is "benign," 2 referring to a diagnosis of "mostly benign," 3 being "not sure," 4 being "mostly malignant," and 5 being "malignant." To evaluate the performance of the experts, the classifier probability output was compared to diagnostic ground truth determined from the pathology reports. From

these comparisons, a receiver operating characteristic (ROC) curve was obtained and the AUC was calculated. The AUC for the human readers on the test set was then compared against the corresponding values for the SVM classifiers.

3 Results

3.1 Patient Characteristics

Table 3 illustrates the patient characteristics in the two patient cohorts considered in this study.

A single- and double-tailed paired student's t-test was applied between the lesion size of adenocarcinoma and granuloma in the testing set and the training set. The results of this comparison are reported in Table 4. With the double-tailed test, no significant difference between the lesion size of two classes across the two cohorts was observed, though a weak correlation ($p = 0.03$) was observed for the single-tailed paired student's t-test.

Table 3 The patient characteristics in the two cohorts.

Parameters	Site-1		Site-2	
Histology	Adenocarcinoma	Granuloma	Adenocarcinoma	Granuloma
No. of patients ($n = 196$)	70	69	34	23
Gender (male, %)	28, 40%	36, 52%	14, 41%	13, 57%
Nodule size $\pm$ SD (mm)	13.33 ± 6.65	11.15 ± 4.45	11.91 ± 4.36	12.19 ± 6.48
CT acquisition parameters	Siemens, Philips		Siemens	
Exposure	120 to 140 kVp		120 kVp	
Slice thickness	1 to 5 mm		1 to 5 mm	
X-ray tube current	150 mAs		41 to 200 mAs	

Table 4 Results of performing a one-tail and two-tail paired student's t-test between the lesion diameters of two classes in testing and training cohort.

	Testing set		Training set	
	Adenocarcinoma	Granuloma	Adenocarcinoma	Granuloma
Mean	13.333	11.408	11.915	12.190
Variance	44.232	17.821	18.981	42.014
Pooled variance	31.993		27.938	
t stat	1.89799883		−0.1901731	
$P\ (T \leq t)$ one-tail	0.0300199		0.42494329	
t critical one-tail	1.6573364		1.67356491	
$P\ (T \leq t)$ two-tail	0.0600398		0.84988659	
t critical two-tail	1.97943869		2.00487929	

3.2 Top Features Identified by Feature Selection

The top features selected during the cross-validation process in the training phase of cohort-1 and three different classifiers (LDA, QDA, and SVM) are shown in Table 5. Based on empirical evaluation, the RBF kernel was considered the optimal kernel for use in conjunction with the SVM, and polynomial kernels of degrees 1, 2, and 3 were also evaluated. The Table suggests that among texture features, "skewness of Laws features (L5 × E5) and (L5 × R5)," "skewness of gradient features (Sobel in both direction and diagonal)," and "Gabor texture

Table 5 The most discriminative texture and shape features (based off AUC) when training the LDA, QDA, and SVM classifiers with cohort-1 ($N = 139$) and testing on cohort-2.

	LDA classifier		QDA classifier		SVM classifier	
	Features	AUC% ± SD	Features	AUC% ± SD	Features	AUC% ± SD
Texture	Skewness of Law L5 × E5	81.2 ± 1.37	Skewness of Law L5 × E5	78.9 ± 1.29	Skewness of Law L5 × E5	81.9 ± 0.89
	Skewness of Law L5 × R5	79.4 ± 0.95	Variance of Gabor, $S = 1/4$, $\Theta = 3\Pi/8$	77.8 ± 1.26	Skewness of Law L5 × R5	81.4 ± 1.15
	Sum of standard deviation	79.0 ± 1.41	Variance of energy	77.6 ± 2.01	Variance of Gabor, $S = 1/4$, $\Theta = 3\Pi/8$	80.2 ± 1.33
	Variance of energy	78.8 ± 2.10	Skewness of Law L5 × R5	77.4 ± 1.76	Variance of sum variance	79.3 ± 2.02
	Skewness of Sobel (X-direction)	78.5 ± 1.24	Sum of standard deviation	76.8 ± 1.35	Skewness of Gabor, $S = 1/4$, $\Theta = \Pi/4$	78.9 ± 1.57
	Variance of Gabor, $S = 1/4$, $\Theta = 3\Pi/8$	78.3 ± 1.44	Skewness of diagonal gradient	76.1 ± 1.88	Skewness of Sobel (X-direction)	78.3 ± 2.10
	Skewness of Sobel (Y-direction)	77.8 ± 1.73	Mean of sum variance	75.1 ± 1.34	Skewness of Sobel (XY-direction)	78.2 ± 1.43
Shape	Mean of extend	67.9 ± 1.57	Variance of eccentricity	67.7 ± 3.16	Mean of extend	69.3 ± 0.92
	Variance of eccentricity	67.7 ± 1.58	Mean of convexity	62.8 ± 3.12	Mean of convexity	68.8 ± 1.61
	Mean of convexity	67.6 ± 1.75	Mean of extend	62.5 ± 4.35	Mean of roughness	67.9 ± 2.11

Table 6 The most discriminative texture and shape features (based off AUC) when training the LDA, QDA, and SVM classifiers with cohort-2 ($N = 56$) and testing on cohort-1.

	LDA classifier		QDA classifier		SVM classifier	
	Features	AUC% ± SD	Features	AUC% ± SD	Features	AUC% ± SD
Texture	Variance of correlation	84.24 ± 1.41	Variance of correlation	83.27 ± 1.47	Variance of correlation	83.69 ± 1.52
	Kurtosis of correlation	76.42 ± 1.61	Mean of Gabor, $S = \sqrt{2}/8, \Theta = 7\Pi/8$	73.96 ± 1.52	Kurtosis of correlation	74.89 ± 1.97
	Skewness of measure of correlation	72.63 ± 1.95	Kurtosis of Gabor, $S = \sqrt{2}/4, ? = \Pi/4$	73.02 ± 2.04	Skewness of measure of correlation	71.1 ± 1.37
	Variance of Laws–Laplacian W5 × S5	72.49 ± 0.99	Skewness of measure of correlation	71.47 ± 1.92	Variance of Laws–Laplacian W5 × S5	69.6 ± 0.89
	Variance of sum average	71.28 ± 1.22	Skewness of difference entropy	71.31 ± 1.43	Mean of Gabor, $S = \sqrt{2}/2, ? = 3\Pi/8$	69.44 ± 1.92
	Variance of Gabor, $S = 1/2, \Theta = \Pi/8$	71.25 ± 1.66	Mean of Gabor, $S = 1/4, ? = 5\Pi/8$	71.14 ± 1.86	Variance of Gabor, $S = 1/2, ? = 7\Pi/8$	68.93 ± 1.25
	Variance of sum variance	70.9 ± 1.07	Mean of Gabor, $S = 1/2, ? = 3\Pi/4$	70.77 ± 1.06	Variance of Gabor, $S = 1/2, ? = \Pi/8$	68.67 ± 1.74
Shape	Variance of roughness	67.9 ± 1.19	Variance of compactness	63.8 ± 1.23	Variance of convexity	66.51 ± 1.66
	Variance of compactness	65.3 ± 1.71	Variance of roughness	63.4 ± 0.85	Variance of compactness	65.49 ± 1.79
	Mean of radial distance	62.4 ± 1.88	Variance of convexity	62.7 ± 1.74	Variance of roughness	65.17 ± 2.14

features" were the most consistent and top performing texture features across the three different classifiers. The top performing shape features were identified as "mean of extend," "mean of convexity," and "variance of eccentricity," though the texture features substantially outperformed the shape features. The most informative statistics were skewness and variance, and none of the top features include kurtosis or entropy of texture features.

As a new experiment, without causing any contamination to the validation set for subsequent experiments, we swapped the training and the testing cohorts. In this new experiment, three classifiers—SVM, LDA, and QDA—were trained on the new training (previously validation) and testing (previously training) sets. The top features for this experiment are listed in Table 6. While the top features for this experiment are not identical to the
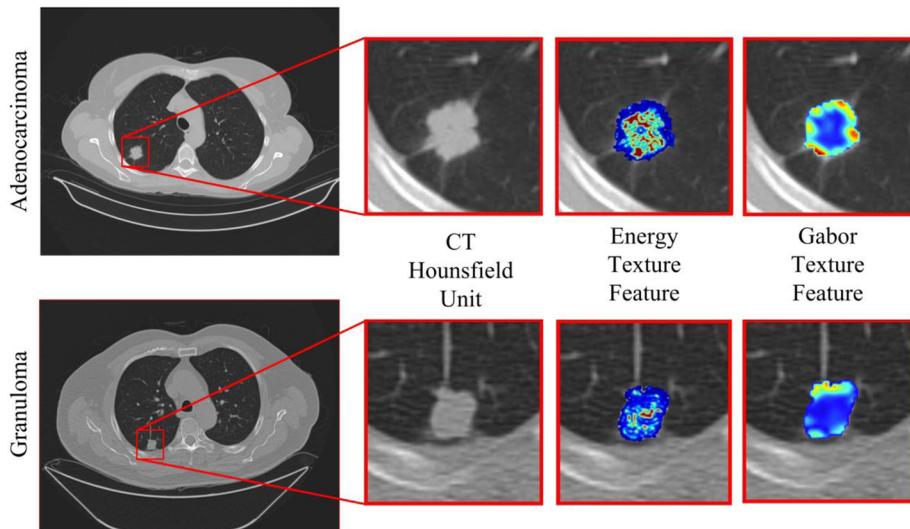
Fig. 3 An illustration of the discriminability of the texture features. Despite the homogeneity of HU in CT of the adenocarcinoma and granuloma, the texture features demonstrate informative heterogeneity in adenocarcinoma in comparison to the granuloma.
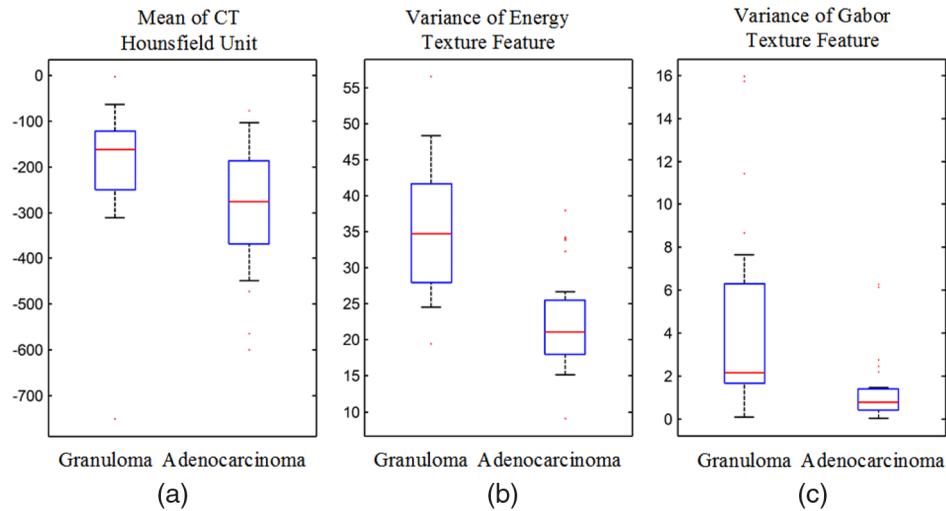
Fig. 4 Box and whisker plots corresponding to the mean of (a) CT Hounsfield units, (b) variance of energy, and (c) variance of Gabor texture features extracted from within the nodules in the training set.
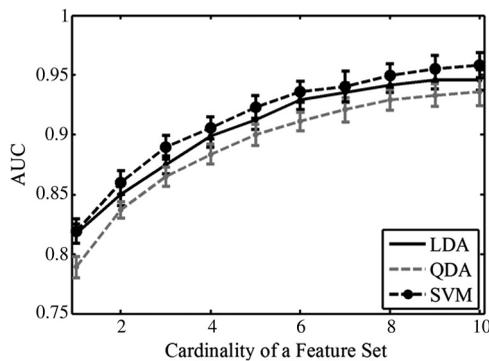
Fig. 5 The AUC values (with the standard deviation bars) for three different machine learning classifiers (LDA, QDA, and SVM) versus the cardinality of the top features subset achieved via sequential feature selection method on the training set ($N = 139$).

top features identified using the original training set, the same feature classes (Gabor and Haralick) and statistics (skewness and kurtosis) were again represented.

Figure 3 visually illustrates the discriminability of the top two features, i.e., "energy" and "Gabor $S = 1/4$, $\Theta = 3\pi/8$" for a granuloma and an adenocarcinoma. The texture heat maps of the nodule in Fig. 3 appear to suggest more heterogeneity in adenocarcinomas compared to granulomas, patterns not immediately obvious on the original CT scans in Fig. 3. This trend also appears reflected in the box and whisker plots shown in Fig. 4, where the average Hounsfield units of the nodules within the CT scans do not appear to show any clear separation

Table 7 The values of PI instability measure for top ranked radiomics.

Features	Training versus testing	Less than 3 mm versus greater than 3 mm
Skewness of Law L5 × E5	0.654	0.288
Skewness of Law L5 × R5	0.431	0.155
Variance of Gabor, $S = 1/4$, $? = 3\Pi/8$	0.145	0.23
Variance of sum variance	0.963	0.792
Skewness of Gabor, $S = 1/4$, $? = \Pi/4$	0.113	0.214
Skewness of Sobel (X-direction)	0.541	0.785
Skewness of Sobel (XY-direction)	0.781	0.627
Mean of extend	0.851	0.923
Mean of convexity	0.42	0.783
Mean of roughness	0.379	0.51

Fig. 6 A 2-D scatter plot of adenocarcinomas (red dots) and granulomas (black dots) in the training set plotted in the space of the top two texture features (Laws L5 × E5 and sum of standard deviation) identified by the feature selection algorithm. The green line was identified as the optimal linear boundary separating the granulomas from the adenocarcinomas in the training set.

between granulomas and adenocarcinomas. Note that the corresponding texture features shown in Fig. 3 do appear to show statistically significant separation on the validation set.

Figure 5 shows the AUC values for the top features identified by the sequential feed-forward feature selection algorithm for the cases in the training set. As the number of employed features increases, the AUC value increases. However, to avoid overfitting,[35] the cardinality of the feature set was restricted to 10 features. Figure 6 illustrates the 2-D scatter plot of top features, sum of standard deviation and skewness of Laws L5 × E5 texture features for the adenocarcinomas and granulomas on the training set. The green line suggests the false-negative free domain for adenocarcinoma (equivalent to false omission rate = 0) in the training set. In this scenario, the classifier can achieve a positive predictive value as high as 72%.

3.3 PI Stability Evaluation of Top Radiomic Features

As a measure of segmentation agreements, the Dice coefficient between the segmentation of two radiologists, $R1$ and $R2$, is defined as follows:

$$S = 2 \times \frac{|R1 \cap R2|}{|R1| + |R2|},$$

where |.| is the cardinality of a set (the number of voxel in the current context). The Dice coefficient and the corresponding overestimation and underestimation error of two segmentations are 0.79 ± 0.12, 0.19 ± 0.16, and 0.18 ± 0.08, respectively.

The results of the stability experiments for the top ranked radiomic features are illustrated in Table 7. Table 7 suggests that the top discriminating features are not necessarily the most stable and reproducible features across the different sites and between scans with greater than and less than 3-mm slice thickness. Interestingly, the shape feature "mean of extend" appeared to be the most stable feature while the "variance of sum variance" feature was the most stable texture feature.

The stability of the top radiomics features across the segmentation of two radiologists is also measured by PI values. The PI values of the most stable shape feature and the most stable texture feature are 0.891 and 0.924, respectively. The comparison between the PI values of the most stable radiomics in the different sites and thickness with the PI values for different readers
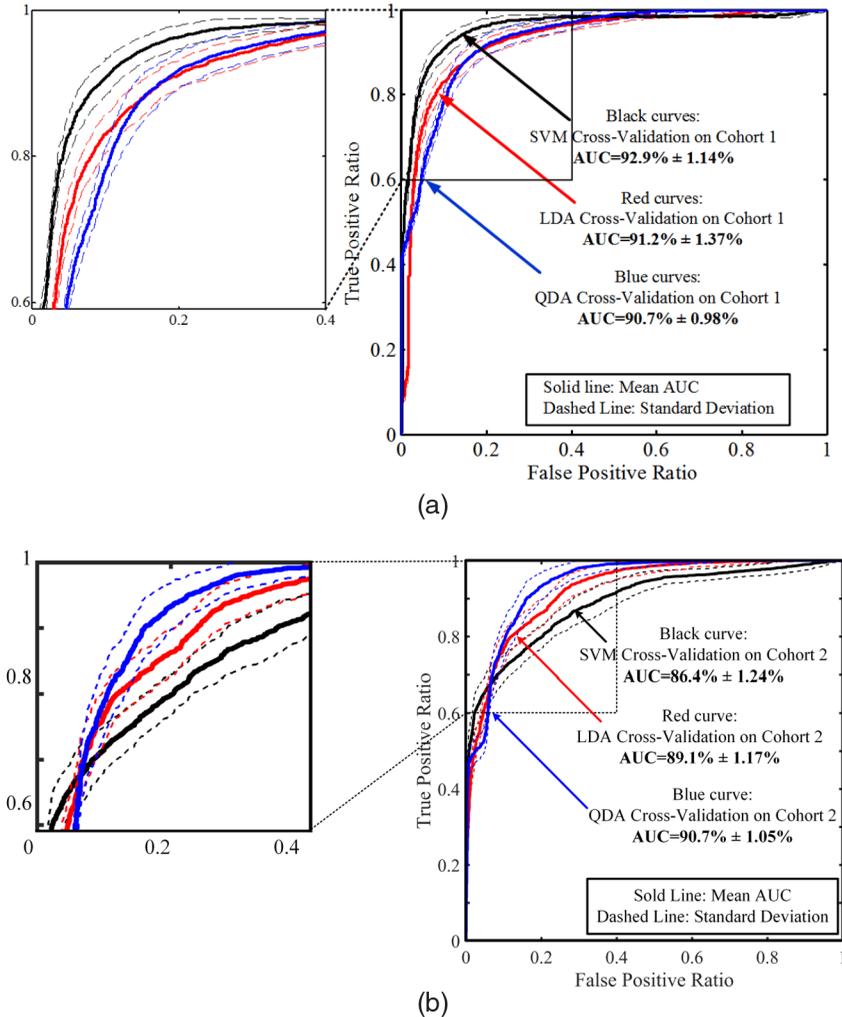
Fig. 7 (a) Receiver operating characteristic curves and corresponding AUC values for the LDA, QDA, and SVM classifiers for discriminating adenocarcinoma from granulomas on the training set on cohort-1 ($n = 139$). The left panel shows a zoomed in version of the ROC curves for the different classifiers in the specificity and sensitivity range of 60% to 100%. (b) Receiver operating characteristic curves and corresponding AUC values for the training set on cohort-2.

(a)

(b)
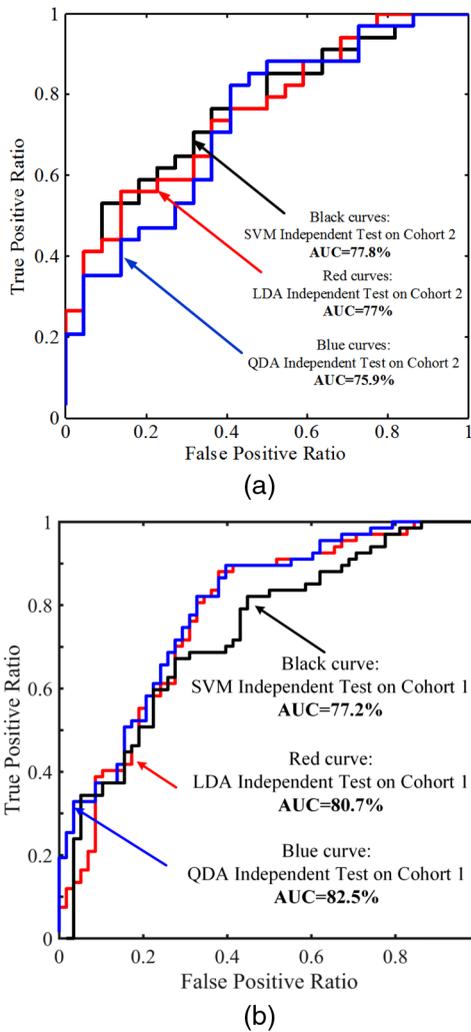
Fig. 8 (a) Receiver operating characteristic curves and corresponding AUC values using the LDA, QDA, and SVM classifiers for discriminating adenocarcinomas from granulomas on the test set (n = 56). (b) Receiver operating characteristic curves and corresponding AUC values for training on cohort-2 and testing on cohort-1.

reveals that the stable features across sites and thickness were remained stable across different readers.

3.4 Statistical Analysis

Figure 2 shows the consensus clustering for radiomic features extracted from the cases in the training set. This figure suggests

that the combination of top texture and shape features yields two distinct clusters, the individual clusters corresponding almost exclusively to either granulomas or adenocarcinomas. However, the corresponding consensus clustering results obtained using the Haralick features employed by Dennie et al.[18] yield multiple disjointed clusters of varying sizes.

Figure 7(a) illustrates the discriminability of the combination of texture and shape features for the LDA, QDA, and SVM classifiers in the training set. Results for these three different classifiers on the testing set are shown in Fig. 8. The best AUC, corresponding to the SVM classifier, on the training set for a combination of four texture and two shape features was 92.9% $\pm$ 1.14%. On the independent test set, the resulting AUC of the locked down SVM classifier was 77.8%. When the training and testing sets were swapped, the QDA classifier was identified as the top ranked classifier with an AUC = 82.5% on the test set. To extend the generalizability of the most informative radiomics, their stability between the training and testing cohorts and between different slice thicknesses was evaluated.

For the human–machine comparison in the same holdout set, the AUCs for an attending radiologist with over 12 years of experience and a pulmonology fellow were found to be 69.72% and 72.39%, respectively. Pearson's correlation and single- and double-tailed paired student's t-test were performed between the prediction results on the test set of the SVM classifier and reader 1 (radiologist). However, no statistically significant differences were found. The results of this comparison are reported in Table 8.

4 Discussion

Differentiating adenocarcinomas from granulomas is one of the most challenging dilemmas faced by thoracic radiologists, due to the similar appearance of the two conditions on CT. Noninvasive differentiation of benign granulomas versus malignant pulmonary nodules could potentially allow for (1) early interventions in patients identified with a high likelihood of having a malignant nodule like adenocarcinoma and (2) prevent unnecessary invasive interventions, such as surgical resection in patients with benign infection.

In this study, we investigated the role of computerized image analysis to identify a set of image texture and shape features that best distinguish adenocarcinomas from granulomas on noncontrast CT scans of the chest. Our study revealed that the Laws features (L5 × E5), (L5 × R5), gradient features (Sobel in both direction, and diagonal), and Gabor texture features were the most predictive and discriminating texture features. The adenocarcinomas tend to have a more chaotic microarchitecture and, hence, substantially more heterogeneity compared

Table 8 Results of performing a Pearson correlation and one-tail and two-tail paired student's t-test between the prediction results of the machine classifier (support vector machine) with the two human readers (radiologist and pulmonology fellow) on the test set (n = 56).

	Machine and radiologist	Machine and pulmonology fellow	Radiologist and pulmonology fellow
Pearson's correlation	0.106933025	0.429644601	0.360786475
t stat	−0.94185815	1.992253037	0.771795225
P (T ≤ t) one-tail	0.175191999	0.025658181	0.221769651
t critical one-tail	1.673033965	1.673033965	1.673033965
P (T ≤ t) two-tail	0.350383998	0.051316363	0.443539302

to granulomas, which is what the Gabor and Sobel features (both gradient related features) might have been capturing. Laws features tend to capture patterns, such as speckle and ripples, which are most likely reflective of the differences in microarchitecture and heterogeneity between the adenocarcinomas and granulomas.

To date, there have been no quantitative studies on shape differences between granulomas and adenocarcinomas. However, our study revealed that shape features, specifically nodule convexity, eccentricity, and the extend features, were strongly discriminative of granulomas and adenocarcinomas. Interestingly, no significant differences in nodule volume or area were found between granulomas and adenocarcinomas across the cohorts considered in this study via a double-tailed paired student's t-test, though a weak correlation was observed when invoking the single-tailed paired student's t-test. The inclusion of shape along with texture measurements appeared to further improve the predictive performance of the SVM classifier compared to the use of texture features alone, with an AUC of 92.9% and 77.8% on the training and validation sets, respectively. When the training and testing sets were swapped, the QDA classifier was identified as the top ranked classifier with an AUC = 82.5% on the validation set.

We also performed an initial study of the stability of the top ranked, most discriminating radiomic features. Clearly, this was a preliminary experiment and additional work needs to be done to evaluate whether the combination of stability and discriminability can yield classifiers, which are predictive and robust. Additionally, this initial experiment focused on only site variations and slice thickness. A more robust evaluation of other parameters (e.g., reconstruction kernels) will need to be undertaken to evaluate feature stability more comprehensively.

We also found that skewness and variance of the shape and texture features were the most discriminating attributes. These results are intuitive considering that a bright object increases the mean value and results in positive skewness, whereas a dark object decreases the mean value and produces negative skewness. Given the fact that adenocarcinomas have increased heterogeneity, the corresponding texture features tend to over-express (see Fig. 3) resulting in positive skewness, whereas granulomas, which tend to have a more coherent microarchitecture, result in a more muted response from the texture filters and, consequently, lower skewness values.

Dennie et al.[18] employed Haralick-related texture features on 55 nodules to discriminate granulomas from primary lung cancer (including adenocarcinoma and squamous cell cancer). Interestingly, in our study, consensus clustering of the nodules in the learning set within the space of Haralick features resulted in multiple fragmented clusters. On the other hand, the combination of Laws features, gradient features, Gabor texture features, and convexity followed by eccentricity yielded two fairly distinct and disjointed clusters corresponding primarily to granulomas and adenocarcinomas. While the approach by Dennie et al.[18] reported an AUC = 90.2%, it was not validated on an independent test set. Interestingly, our AUC on the learning set was 91.2%, which is marginally higher than the AUC reported by them. Additionally, our model yielded a 0% false negative rate on the training set with a positive predictive value of 72%.

In addition, we compared the performance of the classifier with a radiologist with more than 12 years of experience and a pulmonology fellow. We found the classifier marginally outperformed the two human readers.

Our study did have its limitations, which included using datasets from only two institutions. While the two sites were kept independent of each other for training and validation, an obvious question is the generalizability of the classifier to multiple different sites. A second limitation was that we limited this study to one specific type of benign and malignant pathology, namely granulomas and adenocarcinomas. Another limitation of our study was that we did not provide the radiologists with the clinical history associated with the patient, which could have negatively influenced the diagnosis of the human readers. Additionally, a couple of recent papers[36,37] have rigorously and quantitatively investigated the influence of convolution kernels, reconstruction algorithms, and slice thickness on radiomic features for characterization of lung nodules on CT. We did not explicitly consider the influence of these parameters on the extracted texture and shape features, though our classification results did not appear to be significantly affected by variations in slice thickness. Clearly, one of the avenues for future work will need to involve a more rigorous investigation of the influence of slice thickness, convolution kernels, and reconstruction algorithms on the radiomics classifier. An additional avenue for future work will entail evaluating the discriminability of the features and the classifier in distinguishing other benign conditions, such as hamartoma, fibrosis, broncholiths, and inflammation from other types of nonsmall cell lung cancers like squamous cell and large cell carcinomas.

5 Concluding Remarks

In this radiomics study, we investigated the role of texture and shape features in distinguishing adenocarcinomas from granulomas on routine noncontrast CT scans of the chest. Our results suggest that computer-extracted texture and shape descriptors of the nodule can discriminate between these two pathological conditions. Following additional larger scale validation, the classifier could potentially serve as a decision support tool for thoracic radiologists.

References

1. J. Mathew and R. A. Kratzke, "Lung cancer and lung transplantation: a review," *J. Thorac. Oncol.* **4**(6), 753–760 (2009).
2. C.-Y. Zhang et al., "Diagnostic value of computed tomography scanning in differentiating malignant from benign solitary pulmonary nodules: a meta-analysis," *Tumor Biol.* **35**(9), 8551–8558 (2014).
3. H. T. Winer-Muram, "The solitary pulmonary nodule," *Radiology* **239**(1), 34–49 (2006).
4. T. Bunyaviroch and R. E. Coleman, "PET evaluation of lung cancer," *J. Nucl. Med.* **47**(3), 451–469 (2006).
5. D. E. Wood et al., "Lung cancer screening," *J. Natl. Compr. Cancer Network* **10**(2), 240–265 (2012).
6. H. Thorsteinsson et al., "Resection rate and outcome of pulmonary resections for non-small-cell lung cancer: a nationwide study from Iceland," *J. Thorac. Oncol.* **7**(7), 1164–1169 (2012).
7. J. R. H. Klein et al., "One hundred consecutive granulomas in a pulmonary pathology consultation practice," *Am. J. Surg. Pathol.* **34**(10), 1456–1464 (2010).
8. G. D. Tourassi et al., "Acute pulmonary embolism: artificial neural network approach for diagnosis," *Radiology* **189**(2), 555–558 (1993).
9. D. D. Maki, W. B. Gefter, and A. Alavi, "Recent advances in pulmonary imaging," *Chest* **116**(5), 1388–1402 (1999).
10. I. S. G. Armato et al., "Computerized detection of pulmonary nodules on CT scans," *RadioGraphics* **19**(5), 1303–1311 (1999).
11. M. C. B. Godoy et al., "Benefit of computer-aided detection analysis for the detection of subsolid and solid lung nodules on thin- and thick-section CT," *Am. J. Roentgenol.* **200**(1), 74–83 (2013).
12. K. G. Kim et al., "Computer-aided diagnosis of localized ground-glass opacity in the lung at CT: initial experience," *Radiology* **237**(2), 657–661 (2005).
13. M. G. Penedo et al., "Computer-aided diagnosis: a neural-network-based approach to lung nodule detection," *IEEE Trans. Med. Imaging* **17**(6), 872–880 (1998).
14. M. N. Gurcan et al., "Lung nodule detection on thoracic computed tomography images: preliminary evaluation of a computer-aided diagnosis system," *Med. Phys.* **29**(11), 2552–2558 (2002).
15. J. G. Goldin, M. S. Brown, and I. Petkovska, "Computer-aided diagnosis in lung nodule assessment," *J. Thorac. Imaging* **23**(2), 97–104 (2008).
16. W. Shen et al., "Multi-scale convolutional neural networks for lung nodule classification," *Lect. Notes Comput. Sci.* **9123**, 588–599 (2015).
17. B. J. Bartholmai et al., "Pulmonary nodule characterization, including computer analysis and quantitative features," *J. Thorac. Imaging* **30**(2), 139–156 (2015).
18. C. Dennie et al., "Role of quantitative computed tomography texture analysis in the differentiation of primary lung cancer and granulomatous nodules," *Quant. Imaging Med. Surg.* **6**(1), 6–15 (2016).
19. A. Fedorov et al., "3D slicer as an image computing platform for the quantitative imaging network," *Magn. Reson. Imaging* **30**(9), 1323–1341 (2012).
20. R. M. Haralick, "Statistical and structural approaches to texture," *Proc. IEEE* **67**(5), 786–804 (1979).
21. K. I. Laws, "Textured image segmentation," PhD Dissertation, University of Southern California (1980).
22. P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.* **31**(4), 532–540 (1983).
23. N. Ahuja, A. Rosenfeld, and R. M. Haralick, "Neighbor gray levels as features in pixel classification," *Pattern Recognit.* **12**(4), 251–260 (1980).
24. J. M. H. du Buf and P. Heitkämper, "Texture features based on Gabor phase," *Signal Process.* **23**(3), 227–244 (1991).
25. R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, p. 793, Addison-Wesley Longman Publishing Co., Inc., Boston, Massachusetts (2001).
26. D.-C. He and L. Wang, "Texture unit, texture spectrum, and texture analysis," *IEEE Trans. Geosci. Remote Sens.* **28**(4), 509–512 (1990).
27. J. A. Oliver et al., "Variability of image features computed from conventional and respiratory-gated PET/CT images of lung cancer," *Transl. Oncol.* **8**(6), 524–534 (2015).
28. X. Fave et al., "Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer?" *Med. Phys.* **42**(12), 6784–6797 (2015).
29. 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).
30. P. Leo et al., "Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images," *J. Med. Imaging* **3**(4), 047502 (2016).
31. C. Parmar et al., "Machine learning methods for quantitative radiomic biomarkers," *Sci. Rep.* **5**, 13087 (2015).
32. T. J. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons," *K. Dan. Vidensk. Selsk.* **5**, 1–34 (1948).
33. S. Monti et al., "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Mach. Learn.* **52**(1), 91–118 (2003).
34. J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: a review," in *Data Classification: Algorithms and Applications*, C. Aggarwal, Ed., pp. 37–64, CRC Press, Boston, Massachusetts (2014).
35. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York (2013).
36. H. Kim et al., "Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: analysis of intra- and inter-reader variability and inter-reconstruction algorithm variability," *PLoS One* **11**(10), e0164924 (2016).
37. L. He et al., "Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule," *Sci. Rep.* **6**, 34921 (2016).

Mahdi Orooji received his BSc degree in electrical engineering from the University of Tehran, Iran, in 2003, and his MSc and PhD degrees in digital systems from Louisiana State University, Baton Rouge, Louisiana, USA, in 2010 and 2013, respectively. From 2013 to 2016, he was a postdoctoral fellow in the Center for Computational Imaging and Personalized Diagnostics, Cleveland, Ohio, USA. Currently, he is an assistant professor of biomedical engineering in Tarbiat Modares University, Tehran, Iran.

Mehdi Alilou is a senior research associate in the Center for Computational Imaging and Personalized Diagnostics (CCIPD), Department of Biomedical Engineering at Case Western Reserve University. His current work focuses on utilizing machine learning, mathematical modeling, and machine vision algorithms to develop novel imaging biomarkers for computer-aided diagnosis of lung cancer.

Sagar Rakshit is a PGY-2 internal medicine resident at the Cleveland Clinic. He wants to pursue a career in oncology and has interests in personalized cancer treatment and immunotherapy.

Niha Beig is a graduate research assistant in the Biomedical Engineering Department of Case Western Reserve University, Cleveland, Ohio. Her research interests cover areas of machine learning, medical image analysis, and clinical/genomic informatics.

Prabhakar Rajiah is an associate professor of radiology, cardiothoracic imaging, and the associate director of cardiac CT and MRI in UT Southwestern Medical Center, Dallas, Texas, USA. He has authored over 90 peer-reviewed publications and nine books. His current work focuses on clinical translation of advanced techniques in cardiothoracic CT and MRI.

Rajat Thawani is currently an internal medicine resident at Maimonides Medical Center, Brooklyn, New York. He was working at CCIPD as a research associate before his residency. He is interested in lung cancer and medical education.

Michael Yang is the lead thoracic pathologist at University Hospitals Cleveland Medical Center and an assistant professor of pathology at the Case Western Reserve University School of Medicine. His current practice includes a variety of lung, pleural, and thymic neoplasms, as well as nonneoplastic diseases of the lung. His current research interests include lung small cell carcinoma and nonsmall cell carcinoma diagnostic and prognostic biomarkers.

Frank Jacono is an associate professor of medicine at Case Western Reserve University School of Medicine, University Hospitals Cleveland Medical Center and the Louis Stokes Cleveland VA Medical Center, where he serves as the chief of pulmonary and critical care medicine. In addition to his administrative responsibilities and clinical practice, he has a VA, DoD, and NIH-funded basic science and translational research program.

Robert Gilkeson is the vice chairman of research in the Department of Radiology and director of cardiothoracic imaging at the University Hospitals of Cleveland and professor of radiology, Case Western Reserve University School of Medicine. He has authored or coauthored 150 articles or book chapters, and delivered over 200 scientific abstracts or presentations.

Vamsidhar Velcheti is a medical oncologist with expertise in thoracic oncology and cancer immunotherapy. Currently, he is a staff physician and an associate director of Center for Immuno-Oncology Research at the Cleveland Clinic in Cleveland, Ohio, USA. His current work focuses on novel immunotherapy strategies to treat lung cancer and biomarker discovery. He is interested in clinical trial design and incorporation of biomarker studies into early drug trials.

Anant Madabhushi is the director of the Center for Computational Imaging and Personalized Diagnostics (CCIPD) and the F. Alex Nason professor II in the Departments of Biomedical Engineering, Pathology, Radiology, Radiation Oncology, Urology, General Medical Sciences, and Electrical Engineering and Computer Science at Case Western Reserve University. He has authored over 140 peer-reviewed journal publications and over 160 conference papers and delivered over 200 invited talks and lectures both in the United States and abroad.

Biographies for the other authors are not available.