Reeves, Anthony P. and Xie, Yiting and Jirapatnakul, Artit, "Automated pulmonary nodule CT image characterization in lung cancer screening", International Journal of Computer Assisted Radiology and Surgery, 2016, 11(1): 73-88.

doi=10.1007/s11548-015-1245-7,
url=http://dx.doi.org/10.1007/s11548-015-1245-7

Automated Pulmonary Nodule CT Image Characterization in Lung Cancer Screening

Anthony P. Reeves[1], Yiting Xie[1], and Artit Jirapatnakul[2]

Email: reeves@cornell.edu, yx269@cornell.edu, artit.jirapatnakul@mountsinai.org

[1]School of Electrical and Computer Engineering, Cornell University, Ithaca, New York
[2]Department of Radiology, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Abstract

Purpose
In lung cancer screening, pulmonary nodules are first identified in low-dose chest CT images. Costly follow-up procedures could be avoided if it were possible to establish the malignancy status of these nodules from these initial images. Preliminary computer methods have been proposed to characterize the malignancy status of pulmonary nodules based on features extracted from a CT image. The parameters and performance of such a computer system in a lung cancer screening context are addressed.

Methods

A computer system that incorporates novel 3D image features to determine the malignancy status of pulmonary nodules is evaluated with a large dataset constructed from images from the NLST and ELCAP lung cancer studies. The system is evaluated with different data subsets to determine the impact of class size distribution imbalance in datasets and to evaluate different training and testing strategies.

Results

Results show a modest improvement in malignancy prediction compared to prediction by size alone for a traditional size-unbalanced dataset. Further, the advantage of size binning for classifier design and the advantages of a size-balanced dataset for both training and testing are demonstrated.

Conclusion
Nodule classification in the context of low-resolution low-dose whole-chest CT images for the clinically relevant size range in the context of lung cancer screening is highly challenging, and results are moderate compared to what has been reported in the literature for other clinical contexts. Nodule class size distribution imbalance needs to be considered in the training and evaluation of computer-aided diagnostic systems for producing patient-relevant outcomes.

Keywords- pulmonary nodule characterization; low-dose CT; lung cancer screening; automated computer method.

INTRODUCTION

Compelling clinical studies have shown a benefit of lung cancer screening, which allows for the early diagnosis and treatment of lung cancer. A critical issue is the diagnosis of a pulmonary nodule as benign or malignant. Current lung cancer screening practice is to identify pulmonary nodules on annual low-dose CT scans and to apply a follow-up procedure, such as another CT scan or a fine needle biopsy, to suspicious nodules to determine their malignancy status. We consider here how that malignancy status may be determined from just the initial CT image of the nodule.

Pulmonary Nodule Size

It has been universally recognized that probability of malignancy is correlated with the size of pulmonary nodules; the size of nodules is always noted in radiological reports and size is used in radiological staging and for determining the follow-up in lung cancer screening. The conventional way of recording size is to make a single or the average of two "diameter" measurements across a central image slice through the nodule and is expressed in mm. More recently, with the advent of volumetric measurement methods, the size is represented by the volume of the nodule, which is expressed in mm^3. While the former is more conventional and understandable to most physicians, the latter directly relates to the amount of information (number of pixels) available in the CT image. In this paper, we will use both measures; in discussions, we will relate the equivalent diameter *D* of a nodule given a volume *V* by:

$$D = \left(\frac{6}{\pi} V\right)^{\frac{1}{3}}$$

Therefore, in the context of this paper, diameter refers to a surrogate for the measured volume of the nodule and it does not correspond to any actual single-dimensional measurement made on the nodule image.

Further, the size range of nodules under consideration for a classifier is important for nodule classification. We specify the size range R as the ratio of the largest to smallest volume of nodules in a dataset.

$$R = \frac{V_{largest}}{V_{smallest}}$$

Nodule Size in Lung Cancer Screening

In lung cancer screening, the objective is to identify cancers at the earliest stage; that is, when they are the smallest in size, they are the most curable and simpler treatment options may be available. Small-size nodules have less image information in CT images than large nodules due to the number of fixed-size image pixel elements (pixels) that they span. Further, in screening, CT scans are set at a low-dose as the primary task is detection; therefore, the image noise is much higher than for regular CT scans. For a 1-mm thick slice whole-chest CT scan using the conventional 512 x 512 image size, the volume of each pixel is on the order of 0.5 mm^3; therefore, a very rough estimate of the number of pixels in a nodule image is to double the volume. While nodules may be visible to a physician in an apparent 1-2 mm size range, the image information is limited. For example, a 2mm nodule spans in the order of 8 pixels, a 3mm nodule 27 pixels, a 4mm nodule 64 pixels and a 5mm nodule 620 pixels; further, for all these cases, a large majority of these pixels are partial pixels; that is, they consist of a mixture of the nodule tissue and the surrounding lung tissue.

The larger size limit of interest is 15-20mm. Nodules larger than this generally have a high probability of malignancy and very infrequently occur in the main repeat rounds of screening. Such nodules may be detected in the first baseline screening but should not occur in repeat rounds if appropriate small-nodule follow-up procedures are correctly followed. At the large end of the range scale, we have the most image information – a 15–20mm nodule image has on the order of 10^6 – 10^7 pixels. However, this upper end of the size range is much less clinically interesting since we aspire to identify cancers at an earlier stage and time when they are much smaller in size.

An alternative to characterizing a single CT image is to measure the nodule growth rate from two or more images [25]. However, this approach is not currently supported by volumetrically calibrated CT scanners and also requires a delay in the diagnosis required for a measurable change to occur in the nodule between scans.

Size Bias in Feature Evaluation

Obviously, size is a very important image characteristic for determining the probability of malignancy. In this paper, we explore image features other than size in order to provide an improved probability estimate. Since size is easily determined, the main question of interest is what is the probability of cancer at a given size rather than what is the probability of cancer with respect to distribution of sizes.

A major issue in exploring pulmonary nodule characterization is to acquire a large enough sample of both malignant and benign pulmonary nodule images with known outcomes. It is tempting to use all possible data, but the danger here is that the size distribution for the benign nodules may have a much smaller mean than the size distribution for the malignant nodules. The results of the evaluation then reflect the natural difference in size distribution of the datasets rather than other characteristics of the images.

Our hypothesis is that the size distribution difference may become the largest factor in the performance evaluation of datasets with different distributions. We test this hypothesis in two ways. First, we evaluate a size-based classifier that uses size as the only feature on which to predict malignancy, and second, we have constructed datasets with balanced size distributions. We compare the results of the size classifier and the balanced datasets to the outcome of the traditional size-blind approach [2-3, 17-24]. We also consider the impact of training using size binning. That is, using a set of size-specific classifiers instead of a single size-independent classifier.

Image Features

The general approach for computer-aided classification as applied to malignancy diagnosis is to first establish a dataset of images with known outcomes from both classes. A large number of image features (often termed texture measures) are computed for all images in the dataset, and a subset of the features with the best diagnostic performance are selected for the final classifier. In traditional computer vision for conventional video images, there are a number of "texture" features that are classically used. We placed less importance on these features given the ways that the CT data differs from conventional images; for example, (a) the small number of pixels in a nodule, (b) the large amount of image noise and (c) CT images are 3D and have calibrated pixel values. We included nontraditional image features for evaluation including 3D geometry features, 3D features of the density distribution, surface curvature features and features of the nodule margin.

In our preliminary studies [1], we showed that test set size distribution imbalance had a major impact on the perceived outcomes of other studies [2-3] and that size balancing diminished the ROC AUC. Related work [4] showed that 3D image features based on all the image pixels of a nodule were more effective than 2D image features based on just the central image slice of the nodule.

In this paper, the issues in evaluating nodule characterization by image features in the context of lung cancer screening are explored with a system that includes novel 3D image features. Balanced and unbalanced evaluation datasets are used to determine the impact of size balancing and size binning.

METHODS

Dataset Selection

We combined image data from the two largest lung cancer screening studies, the Early Lung Cancer Action Program (ELCAP) [5] and the National Lung Cancer Screening Trial (NLST) [6]. Malignant nodules were included if there was a pathologically proven cancer diagnosis; benign nodules were included if there was 2 years of no clinical change or a benign pathologic diagnosis.

Pulmonary nodules may be solid, part-solid or nonsolid. Solid nodules are the most common type for cancer and consist of a mass of invasive cells that typically have CT image intensity similar to that of soft tissue. Nonsolid nodules typically have abnormal cells distributed on the epithelial surface of the airways. Hence, the associated lung parenchyma has a higher CT image intensity than normal lung parenchyma but less dense than soft tissue or solid nodules. Little is known about nonsolid nodules compared to the more typical solid nodules. One lung cancer screening study reported that 17% of the cancers were nonsolid nodules [29]. It has been suggested that the part-solid nodules that contain both solid and nonsolid components may occur when the cancer becomes invasive and a more traditional solid nodule is developing. From an image analysis viewpoint, nonsolid nodules have a very different visual presentation compared to solid nodules and are more challenging for image segmentation. Clinically, nonsolid nodules are considered to be more slow growing than solid nodules and also harder to measure; screening protocols usually have a different management for these nodules.

Given the different visual presentation of nonsolid nodules and their small numbers in our databases, only solid nodules or the solid component of part-solid nodules were included in this study. Nonsolid nodules will be considered in a future study when more images are available. It is likely that a separate image analysis system for the nonsolid subtype may produce the best analysis outcomes.

In our study's two datasets, the first dataset contained cases selected from the Weill Cornell Medical Center database (which is part of the ELCAP study) that had at least one solid or part-solid nodule on at least one thin-slice CT scan. Part-solid nodules were only included if they were comprised primarily of a solid component. The status of malignant nodules were determined by either biopsy or resection, while the status of benign nodules was established through a negative biopsy result or by 2 years of no clinical change by a board certified radiologist. All CT scans had a slice thickness of 2.5mm or less. Metastatic cancer and benign calcified nodules were excluded. A total of 259 nodules (167 malignant and 92 benign) with CT scans of 1.0, 1.25, or 2.5mm slice were included. Approximately 13.9% (36/259) of the nodules were on 1.0mm scans, 73.8% (191/259) on 1.25mm scans and 12.4% (32/259) on 2.5mm scans. Scans were obtained using GE Medical Systems scanners. The Weill Cornell image acquisition time period was 1994-2007, and the majority of the Weill Cornell instances were reconstructed using the BONE kernel.

The second dataset contained cases selected from NLST. Participants underwent three rounds of screening at 1-year intervals. Cancers were identified through the NLST protocol. After three rounds, abnormalities suspicious for lung cancer that were stable across the three rounds were classified as minor abnormalities (i.e., benign). We selected NLST CT scans with a slice thickness less than or equal to 3.2mm. A total of 477 nodules (245 malignant and 232 benign) with CT scans of 1.0, 1.25, 1.3, 2.0, 2.5, 3.0 and 3.2mm slice thickness were chosen. Approximately 2.94% (14/477) of the nodules were on 1.0mm scans, 2.73% (13/477) on 1.25mm scans, 0.63% (3/477) on 1.3mm scans, 37.74% (180/477) on 2.0mm scans, 48.01% (229/477) on 2.5mm scans, 0.21% (1/477) on 3.0mm scans and 7.76% (37/477) on 3.2mm scans. Scans were obtained using a wide range of scanners including Siemens, GE Medical Systems, Philips and Toshiba scanners. For NLST, the screening time period was 2002-2007; NLST images were reconstructed with a variety of reconstruction kernels including STANDARD and BONE (for GE scanners) and B30f and B50f (for SIEMENS scanners).

Nodules were selected to meet the 3D feature image quality criterion; that is that they spanned at least three image slices and preferably four or more. Further, all nodules had a diameter between 3 and 30mm. The volume of each nodule was computed from automated segmentation [7] and the nodule size was represented as the equivalent diameter of a sphere with the equivalent volume as the nodule. Only one instance of a nodule was used per case.

For both datasets we used methods to minimize the size distribution differences between malignant and benign. For the ELCAP dataset, we selected all the large benign nodules that were available to match the sizes of the cancers. For the NLST, we sought to minimize the size of the cancers by selecting the first CT image in a longitudinal sequence where possible.

Figures 1 and 2 show the nodule size distribution for the Weill Cornell and NLST dataset. By combining these two datasets, we created a database with 736 nodules (412 malignant and 324 benign). Figure 3 shows the size distribution for the entire database, and Table 1 gives the statistics for size distribution for malignant and benign nodules.

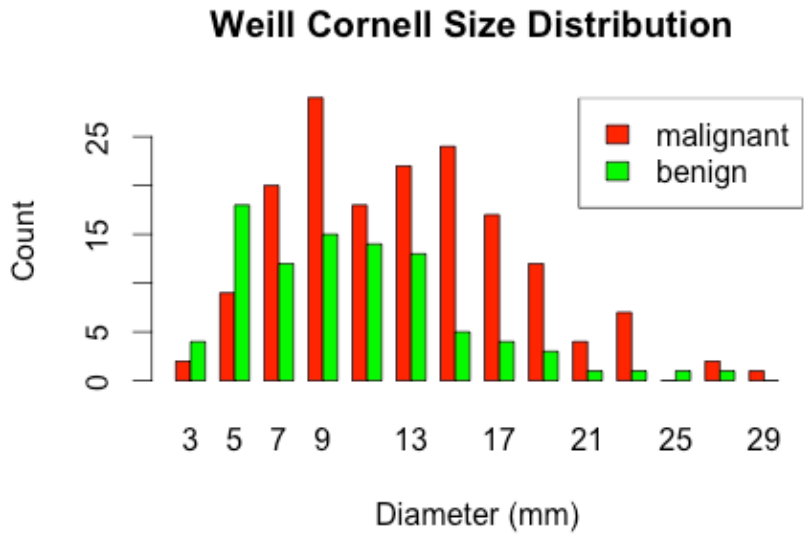


Fig. 1 Weill Cornell nodule subset size distribution

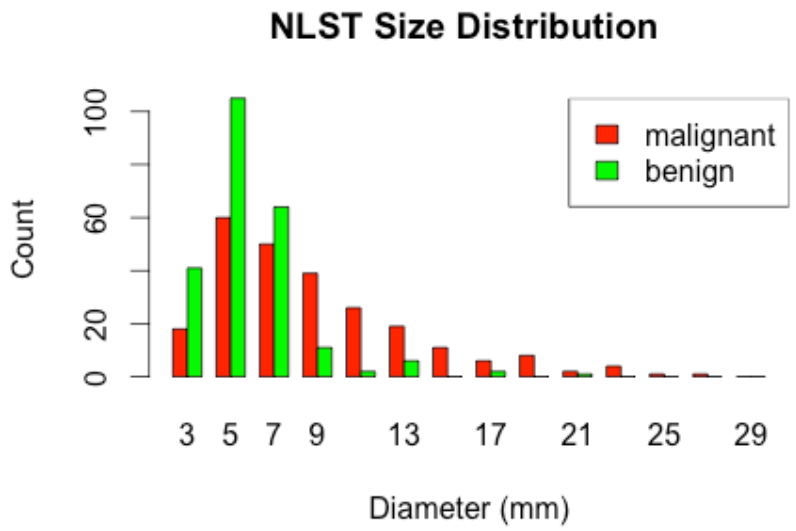


Fig. 2 NLST nodule subset size distribution

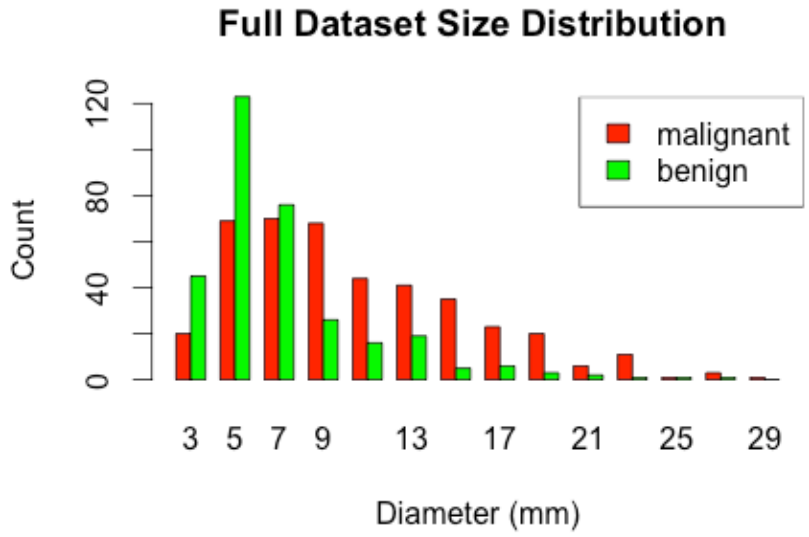


Fig. 3 Full dataset size distribution

Table 1 Size statistics for the main datasets

Source	Type	Number	Size Min (mm)	Size Max (mm)	Size Mean (mm)	Size Median (mm)
Weill Cornell	Malignant	167	3.72	29.14	12.80	12.17
	Benign	92	3.25	27.11	10.21	9.47
	All	259	3.25	29.14	11.88	11.28
NLST	Malignant	245	3.00	27.60	8.93	7.60
	Benign	232	3.11	21.43	5.76	5.15
	All	477	3.00	27.60	7.39	6.15
Combined	Malignant	412	3.00	29.14	10.50	9.21
	Benign	324	3.11	27.11	7.02	5.84
	All	736	3.00	29.14	8.97	7.33

Size-Balanced Nodule Dataset

A size-balanced subset of nodules (GA) was created from the full database to assess the impact of size on the classification result (see Table 2). First, all malignant and benign nodules were divided into bins based on their volumetric derived diameters (3, 4, 5mm, etc). Then, bins smaller than 5mm were discarded since these nodules were too small for the shape related features to be effective. Bins larger than 14mm were discarded due to the lack of data (usually less than three nodules per bin). For the remaining bins (5-14mm), the same number of malignant and benign nodules was randomly selected to maximize the number of nodules in each bin. We explored two binning strategies: the first was to create three bins each with a similar size range and the second was to partition into just two bins (by combining the two largest size bins) so that each bin would have a similar number of nodules (see Table 3). For the first binning strategy, the first bin (G6) only includes nodules with a size from 5.0 to 7.0mm; the second bin (G8) includes nodules with a size from 7.0 to 9.0mm; the third bin (G12) includes nodules with a size greater than 9.0mm. The three bins were designed so that each bin would have a sufficiently large number of nodules and the volume range within each bin would be similar (see Table 3 volume range). For the second binning strategy, the first bin contains G6 nodules and the second bin combines both G8 and G12 nodules.

In total, 163 malignant and 163 benign nodules were selected to have as similar size distribution as possible. In the size-balanced dataset, 44.79% (146/326) nodules had a size between 5.0 and 7.0mm, 28.22% (92/326) nodules had a size between 7.0 and 9.0mm and 26.99% (88/326) nodules had a size between 9.0 and 14.0mm.

Table 2 Size balanced nodule size distributions

Group GA	Number	Size Min (mm)	Size Max (mm)	Size Mean (mm)	Size Median (mm)	Volume Range
Malignant	163	5.01	14.00	8.05	7.21	21.82
Benign	163	5.02	13.91	8.01	7.27	21.28
All	326	5.01	14.00	8.03	7.22	21.82

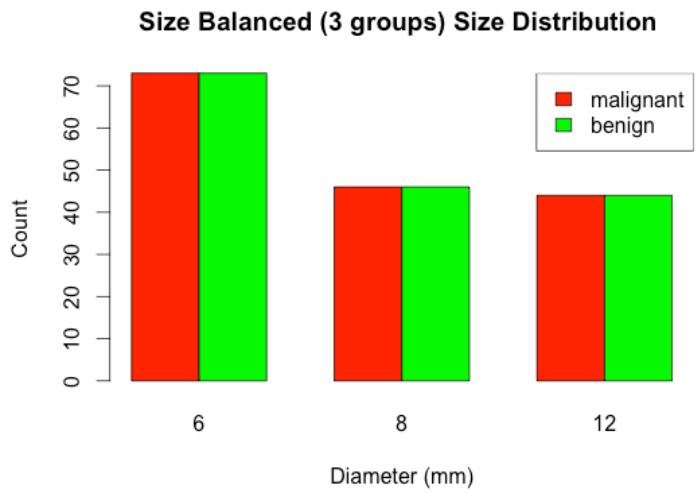


Fig. 4 Size balanced subset nodule size distribution

Table 3 Size-balanced nodule size distributions with binning

Group	Type	Number	Size Min (mm)	Size Max (mm)	Size Mean (mm)	Size Median (mm)	Volume Median (mm^3)	Volume Range
G6	Malignant	73	5.01	6.98	6.04	6.04	115	2.70
	Benign	73	5.02	6.97	5.96	6.06	116	2.66
G8	Malignant	46	7.04	8.96	7.82	7.65	234	2.07
	Benign	46	7.03	8.84	7.83	7.71	239	1.99
G12	Malignant	44	9.05	14.00	11.64	11.72	842	3.70
	Benign	44	9.02	13.91	11.61	11.41	777	3.66
GA	All	326	5.01	14.00	8.03	7.22	197	21.79

Image features

In this work, 46 3D features [8] were computed from the segmented nodule images. These features are grouped into four categories: morphological, density, surface curvature and margin gradient (see Table 1 in "Appendix"). Images were resampled to $0.25mm^3$ isotropic resolution for feature evaluation [8].

Morphological features describe the shape characteristics of the nodule and are derived from standard image moments [9]. Radiologists use the nodule shape as an indicator of malignancy; for example, Takashima et al identified a greater prevalence of polygonal shape and 3D ratios in benign nodules compared to malignancies [10]. The morphological features are: volume, surface area, volume-to-surface area ratio, compactness, sphericity, attachment ratio, length/width/height of the ellipsoid of inertia, ratios of the length/width/height, the roll/pitch/yaw of the ellipsoid of inertia, and the scale-normalized second-order morphological moment.

Since the gray levels of a CT scan are representative of the density of the tissue, density features can be derived from the gray-level voxel values of the image. One of the density characteristics often used by radiologists is the average density of the nodule – whether the nodule is solid, part-solid, or nonsolid has a significant effect on the interpretation of the nodule. The density features analyzed in this work are: density mass, mean density, the standard deviation, skewness and kurtosis of the density histogram, the length/width/height of the density-based ellipsoid of inertia, the ratios of length/width/height, and the scale-normalized second-order densitometric moment.

The surface features of a nodule are often considered by radiologists in determining nodule malignancy status. These features are represented by the surface curvature features, which measure the rate of change of the surface normal to the length of the surface. Although the surface curvature can be computed directly from the gray-level voxels [11], errors are introduced from the fact that the voxels are rectangular approximations of the nodule surface. To address this problem, the surface curvature is estimated from a smoothed polygonal tessellation of the segmented binary nodule image as described by Jirapatnakul [30]. To generate the tessellation, the marching cubes algorithm developed by Lorensen and Cline [12] was used. This algorithm results in triangles located at angles that are multiples of 45 degrees; to improve the surface representation, the polygonal tessellation was smoothed by modifying the position of each vertex as a weighted sum of the neighboring vertices and itself. Once the smoothed polygonal representation is obtained, the surface normal of each triangle can be computed. From the surface normal of each triangle, the surface normal of each vertex can be computed as the average of the surface normals from each triangle of which it is a member. Finally, the curvature for a triangle is computed as the average difference between the surface normals at each vertex. The mean, minimum, maximum, range, standard deviation, skewness and kurtosis of the curvature distribution were included as features.

The final category of features is the nodule margin. The nodule margin refers to the boundary of the nodule and the surrounding lung parenchyma. While the surface curvature features capture the shape of the nodule at the margin, the margin gradient features measure the density changes that occur at the margin. To compute the margin gradient, the surface normals for each triangle in the nodule surface representation are used. These normals are computed in the process of computing the surface curvature, as described in the previous paragraph. In addition to the surface normals, gradient images in each direction (x, y, and z) are created from the resampled isotropic grayscale images using a 3D operator proposed by Monga et al. [13, 31]. At each triangle, ten gradient samples are taken along the surface normal vector through the center of the triangle. The highest gradient value is recorded for the triangle. The mean, minimum, maximum, range, standard deviation, skewness and kurtosis of the distribution of gradients were included as features.

Feature classification

Five different classifiers were evaluated: the distance weighted k-nearest-neighbors classifier (dwNN) [14], the Support Vector Machine (SVM) classifier [15] with a polynomial kernel (SVM-P), SVM with a Radial Basis Function kernel (SVM-R), the logistic regression classifier (LOG) and the size threshold classifier (Size-C). For dwNN, SVM (polynomial and RBF) and LOG classifiers, fivefold cross-validation approach was used for training and testing. In the training stage, training set was further divided into train and validation for parameter optimization using fivefold cross-validation. The final classification outcome was

represented by the average ROC curve and the area under the ROC curve (AUC) obtained using the five ROC curves from fivefold cross-validation. The threshold averaging method was used for ROC averaging (Fawcett et al. [26]).

Compared to the conventional K-Nearest Neighbors classifier, the dwNN classifier weights each neighbor n of a feature vector based on their distance d_n. The weight w_n is computed as follows where σ is a constant that controls the impact of each neighbor on the classification outcome. In the training stage, a grid search was performed to find the optimal σ (see Table2 in "Appendix").

$$w_n = \frac{1}{exp(\sigma * d_n)}$$

The SVM classifier was implemented using the SVM^{light} library [28]. For the SVM with polynomial kernel (SVM-P), the two parameters obtained from training were the order of polynomial kernel d and the trade-off between training error and margin c. The search space for d and c is shown in Table 2 in "Appendix". Joachims [28] stated that c=0.001 is acceptable for most tasks and a larger c leads to considerably longer training time. For SVM with RBF kernel (SVM-R), the two parameters obtained from training were the weighting factor in the polynomial kernel g and the trade-off between training error and margin c. The search space for g and c is also shown in Table 2 in "Appendix".

For the LOG classifier, Peduzzi et al. [16] have shown in a simulation study that for each feature, LOG would require at least 10 positive and 10 negative samples to avoid bias. In the training stage, each feature was ranked based on its individual AUC and the top n features were selected. The search space for n is shown in Appendix Table 2.

In addition, results for the size-only classification scheme were computed. For a given size threshold T, the size classifier indicates that all nodules with a size greater than T are malignant and all nodules with a size less than T are benign. The evaluation metric was the AUC, which was achieved by varying the size threshold T through the size range of the nodules in the dataset; therefore, no training was required for this classification method. The size classifier provides information on the size imbalance within the malignant and benign size distribution of the test dataset – the greater the size imbalance, the higher the AUC.

Experiments

Two main experiments were performed: the first to evaluate the impact of class size distribution imbalance by comparing the size-only classifier to methods using additional image features, and the second to evaluate the impact of using size-balanced datasets. In all experiments, the full set of image features and all five classifier types were considered. The organization of the experiments is illustrated in Fig. 5. The main dataset All Data consists of the two trial cohorts. A size distribution-balanced dataset is selected from All Data for the second experiment, and it is further partitioned into size bins for the binning classifiers.

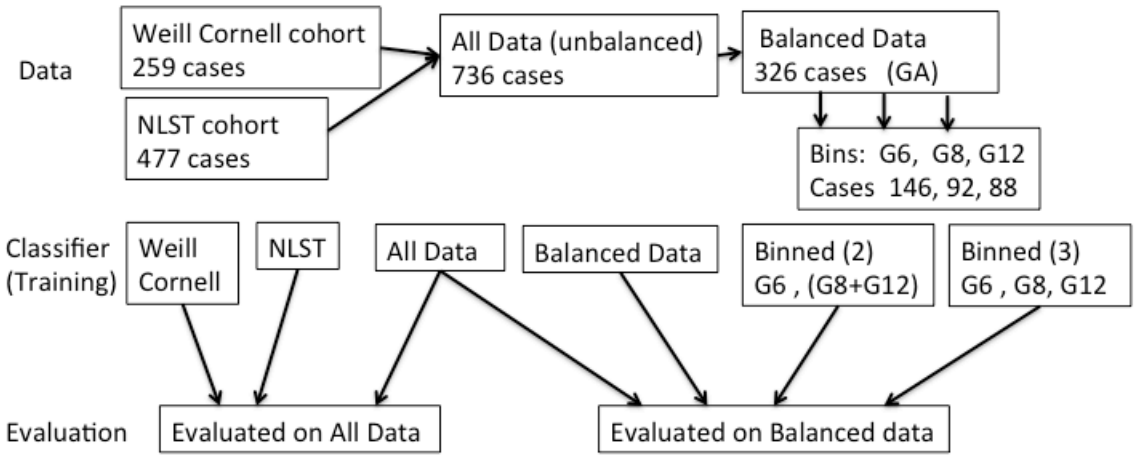


Fig. 5 Overview of the experiment organization.

In the first experiment, the traditional method for training and testing using cross-validation on All Data was evaluated. In addition, to illustrate the impact of the size imbalance, the classifiers trained on just one of the two data cohorts are also evaluated with All Data.

In the second experiment, the performance of three different classifiers trained on only the Balanced Data subset is compared to the traditional classifier trained in All Data using three different training strategies with the data subsets shown in Table 3. First, the performance of this dataset (GA) using the unbalanced full-data-trained classifier from the first experiment was measured. Second, the performance of the balanced data (GA) trained (through cross-validation) with itself was evaluated. Third, a binned training strategy was evaluated where the balanced dataset was partitioned into different sized groups and each classifier was trained for each group using only nodules in the same size group. A two bin grouping using bins G6 and G8+G12 and a three bin strategy using bins G6, G8 and G12 were evaluated.

The usual metric for classifier performance is the area under the curve (AUC). Since in this context much of this performance is attributed to difference in the test set size distribution, an additional metric, the incremental increase in AUC compared to a size classifier (IAUC), was considered to be more relevant. The DeLong test [27] was used to assess pairs of ROCs. It estimates a covariance matrix from two ROC curves, which may also be used to construct confidence regions and compute the statistical significance of the difference between the two AUCs.

RESULTS

In the following tables of AUC results, the mean AUC value for the fivefold cross-validation is reported, together with the standard deviation in parenthesis. Also the p-value of the Delong test with respect to the size classifier is given.

Results for the size-unbalanced dataset

The result for the full unbalanced dataset is shown in Table 4. A comparison of the different training datasets using an SVM-P classifier is shown in Fig. 6. A comparison of the different

classifiers for the full unbalanced dataset is given in Fig. 7. For LOG on the full unbalanced data (Table 4 all row), the optimal set of features is listed in Table 3 in "Appendix". For the full unbalanced data, each classifier's ROC was compared to size classifier's ROC and their p-values are listed in Table 4. Values listed in bold in Table 4 indicate the ROC appears in either Figs. 6 or 7.

Table 4 Classifier performance (AUC) for the unbalanced data sets. σ is the standard deviation. Size-C AUC = 0.725.

Training		dwNN	SVM-P	SVM-R	LOG
Weill Cornell	AUC σ	0.737 (0.046)	**0.731** (0.027)	0.731 (0.027)	0.716 (0.036)
NLST	AUC σ	0.738 (0.032)	**0.766** (0.039)	0.763 (0.036)	0.749 (0.047)
All	AUC σ IAUC p-value	**0.750** **(0.042)** **0.025** (p=0.09)	**0.772** **(0.034)** **0.047** (p<0.001)	**0.772** **(0.031)** **0.047** (p<0.001)	**0.761** **(0.038)** **0.036** (p=0.15)

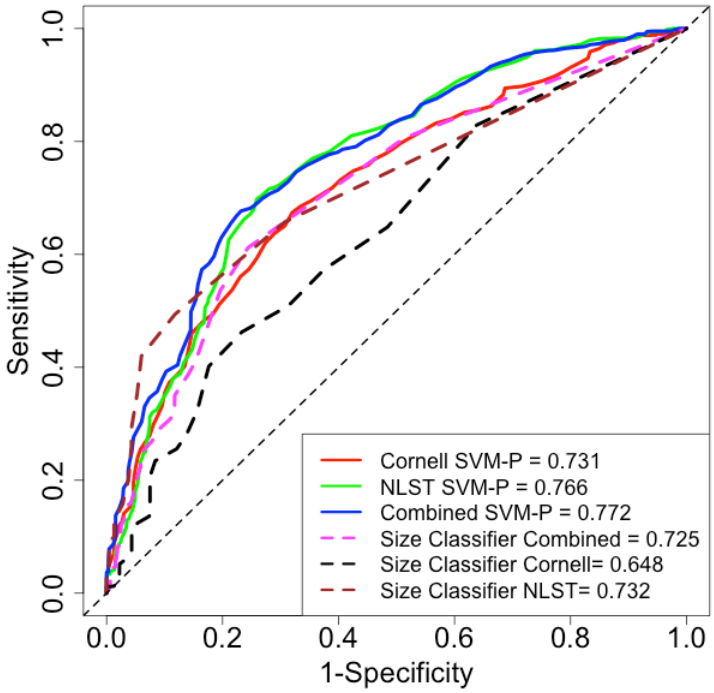


Fig. 6 ROC curve for SVM with Polynomial kernel on the two training dataset separately (red for Weill Cornell and green for NLST) and combined (blue). The size classifiers on Weill Cornell (black), NLST (brown) and combined dataset (magenta) are also shown.

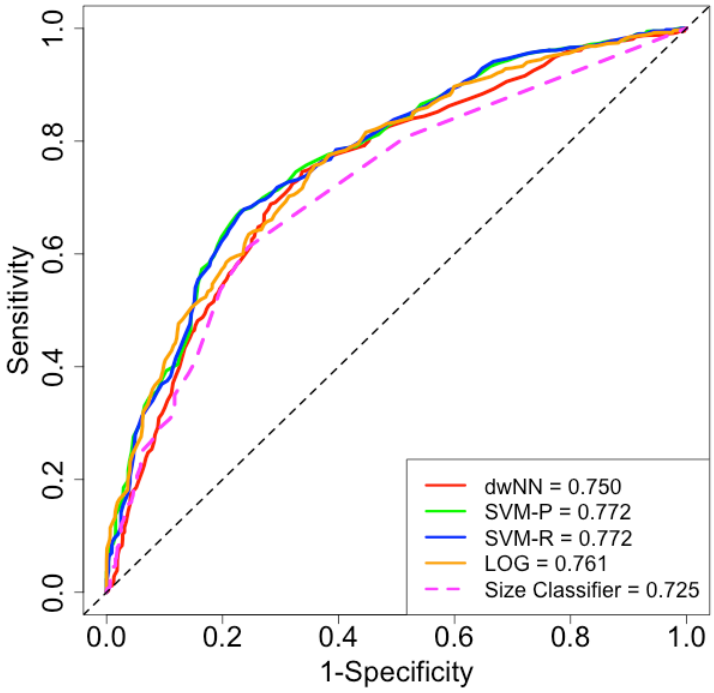


Fig. 7 ROC curve for dwNN (red), SVM with Polynomial kernel (green), SVM with RBF kernel (blue), logistic regression (black) and size classifier (magenta) on the full unbalanced dataset.

Results for the size-balanced dataset

The results for the balanced dataset with the unbalanced training and the balanced training schemes are shown in Table 5. Each classifier was also compared to the size classifier and the p-value is given. Table 6 shows the results using different training conditions for the two binning strategies. The training conditions were: unbalanced binned training and balanced binned training. For each training condition, only the result from the best classifier is shown. The two binning strategies were: three bins (G6, G8 and G12); two bins where the two larger bins G8 and G12 were combined into one large bin and the same experiments were repeated using a small bin (G6) and a large bin (G8+G12), each with a similar number of nodules. Table 7 shows the overall performance for the binned classifiers (three bins and two bins) under different training conditions.

Table 5 Classifier performance (AUC) for the balanced dataset GA trained on balanced and unbalanced data. Standard deviation σ, IAUC and p-value are also listed. Size-C AUC = 0.510.

Training		dwNN	SVM-P	SVM-R	LOG
Unbalanced	AUC	0.584	0.639	0.642	0.564
	σ	(0.014)	(0.050)	(0.048)	(0.035)
	IAUC	0.074	0.129	0.132	0.054
	p-value	(p=0.14)	(p=0.01)	(p=0.009)	(p=0.11)
Balanced	AUC	0.700	0.708	0.699	0.624
	σ	(0.051)	(0.062)	(0.056)	(0.095)
	IAUC	0.190	0.198	0.189	0.115
	p-value	(p<0.001)	(p<0.001)	(p<0.001)	(p=0.003)

The ROC curves for the full balanced dataset with each classifier with balanced training (GA balanced) are shown in Fig. 8. The ROC curves for GA with each classifier with unbalanced training (GA unbalanced) are shown in Fig. 9. Figure 10 shows the ROC curves for the best classifier under each training condition: unbalanced training, balanced training, overall performance using three bins (G6, G8 and G12), and overall performance using two bins (small and large). For testing on GA set using balanced and unbalanced training, the optimal features for LOG are listed in Table 4 in "Appendix".

Table 6 Best performance AUC for different evaluation data sets: G6, G8, G12 and Large (G8+G12). The best classifier and standard deviation σ are also listed.

Training		G6	G8	G12	Large (G8 +G12)
Unbalanced (Binned)	AUC σ Classifier	0.646 (0.050) (dwNN)	0.699 (0.130) (SVM-P)	0.745 (0.093) (SVM-R)	0.740 (0.081) (SVM-R)
Balanced (Binned)	AUC σ Classifier	0.691 (0.078) (LOG)	0.759 (0.141) (SVM-P)	0.759 (0.089) (SVM-P)	0.780 (0.079) (SVM-P)
Size-C	AUC	0.546	0.500	0.507	0.503

Table 7 Best performance AUC for binned classifiers (3-bin and 2-bin). Size-C AUC = 0.510.

Training		3-bin	2-bin
Unbalanced	AUC σ IAUC p-value	0.666 (0.036) 0.156 (p=0.002)	0.684 (0.048) 0.174 (p<0.001)
Balanced	AUC σ IAUC p-value	0.726 (0.056) 0.216 (p<0.001)	0.742 (0.057) 0.232 (p<0.001)

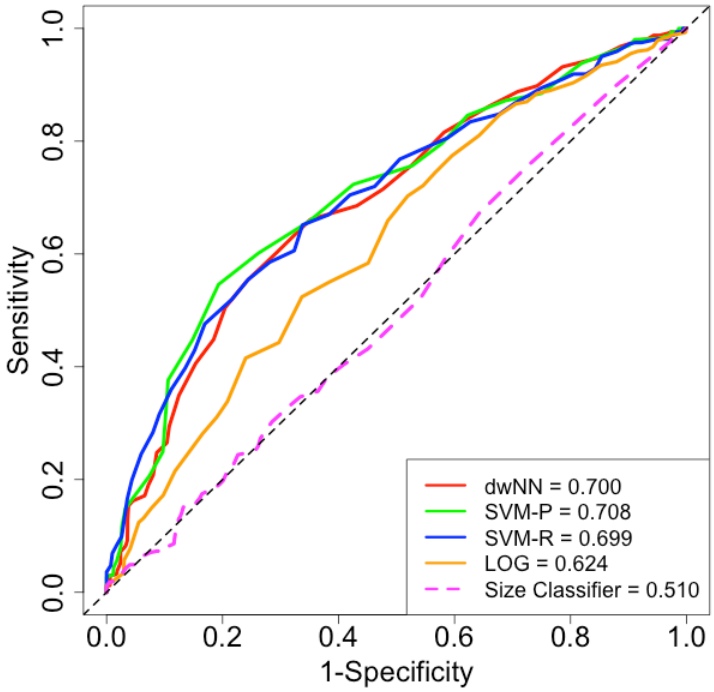


Fig. 8 Comparison of classifiers on the size-balanced dataset GA using balanced training.

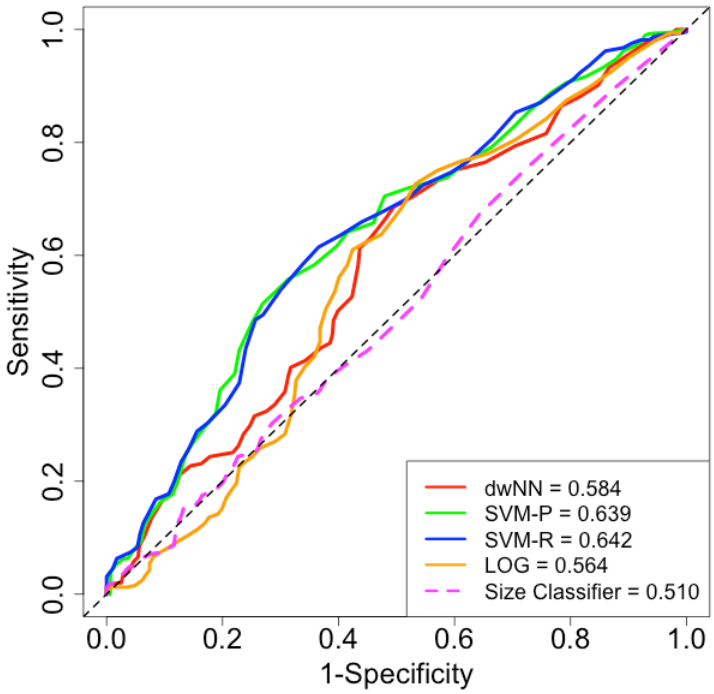


Fig. 9 Comparison of classifiers on the size-balanced dataset GA using unbalanced training.

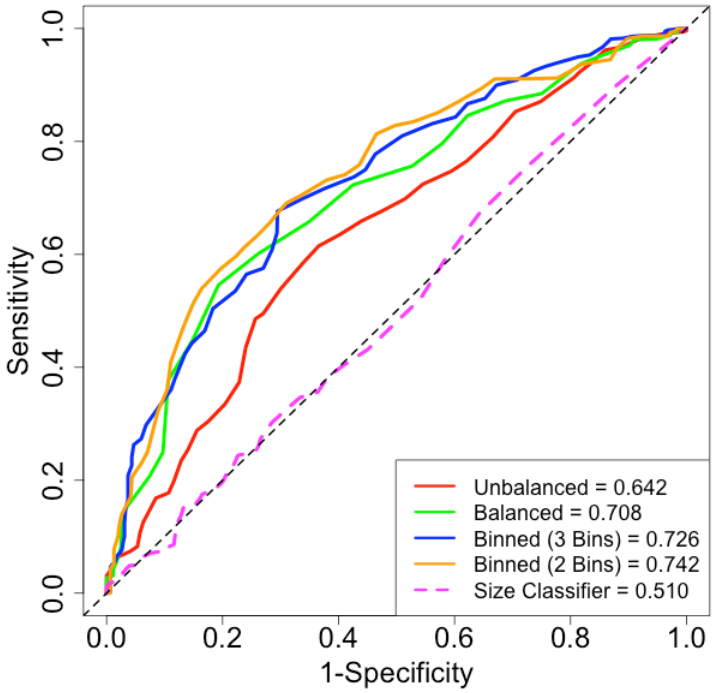


Fig. 10 Comparison of the best classifier under different training conditions when tested on the size-balance dataset GA.

DISCUSSION

Due to the data selection methods for size balancing and the image quality requirements neither of the size distributions for the Weill Cornell nor NLST data accurately reflect the size distributions of the subjects in lung cancer screening studies; however, the general distribution for the cancers is representative as we selected all usable malignant nodule images. This is not the case for the benign nodules since these were selected with a view to size balancing. In the full studies, there are many more small benign nodules.

Pulmonary nodule classification from screening CT images acquired for nodule detection is a very challenging task given the small size of the nodules and the large amount of image noise. From Table 4 and Figs. 6, 7, we see that the size classifier, which is only sensitive to the difference in the size distributions for benign and malignant nodules, provides an AUC of 0.725 for the combined dataset. This number would have been much higher (and comparable to other published studies) if we had included the very large number of small benign nodules that were documented in the full screening studies. The size classifier ROC curves in Fig. 6 for the two individual study datasets show very similar properties with a slightly larger size imbalance for the NLST dataset. Note, in Fig. 6, the best evaluation results are superior to but follow most closely the size evaluation curve of the All Data test set even when the classifier is trained only by a single cohort.

In Fig. 7, we see a comparison of the different classification methods used for the combined full unbalanced dataset. Very little difference is noted; the best classifiers (SVM-P and SVM-R) have an IAUC of only 0.047 over the size classifier. The average improvement is 0.039. From Table 4 we see that the IAUC is only statistically significant for the two SVM classifiers ($p<0.05$).

The results for the size full balanced dataset are shown in Table 5, 6, 7 and Figs. 8, 9, 10. The size classifier has an AUC of 0.510; for a perfectly balanced dataset the value would be 0.500. For the balanced data test set GA, the unbalanced classifier (which claimed an AUC of 0.772 when evaluated on all nodules) only achieved an AUC of 0.642 (IAUC of 0.132) compared to an AUC of 0.708 (IAUC of 0.198) using the balanced classifier. This difference in performance was statistically significant ($p=0.01$).

In Table 6 and 7 and Fig. 10, the AUC results for binning are shown. While all AUCs were statistically significant with respect to the size classifier, none of the balanced binned classifier was statistically significantly different with respect to the unbalanced binned classifier. However, the binned results show better AUC values compared to balanced training overall 0.742 (two bins), 0.726 (three bins) versus 0.708 (balanced training). Further, the small-nodule bin (G6) shows a lower AUC than the others under all training conditions. For the binned training, the IAUC for G6 was 0.145, while for the other bins, it was much higher (G8, 0.259, G12, 0.252; (G8+G12) 0.277). This implies the image features are less effective for these small nodules. An improvement of performance of the 2-bin classifier is noted (0.742 vs. 0.708) although this is not statistically significant $p=0.35$.

Figures 11 through 13 provide examples of some of the image issues and demonstrate the range of presentations shared by both malignant and benign nodules. Figure 11 shows the impact of the image reconstruction filter on image quality, Fig. 12 shows malignant and benign nodules with similar complex presentations and Fig. 13 shows the impact of structured image noise.

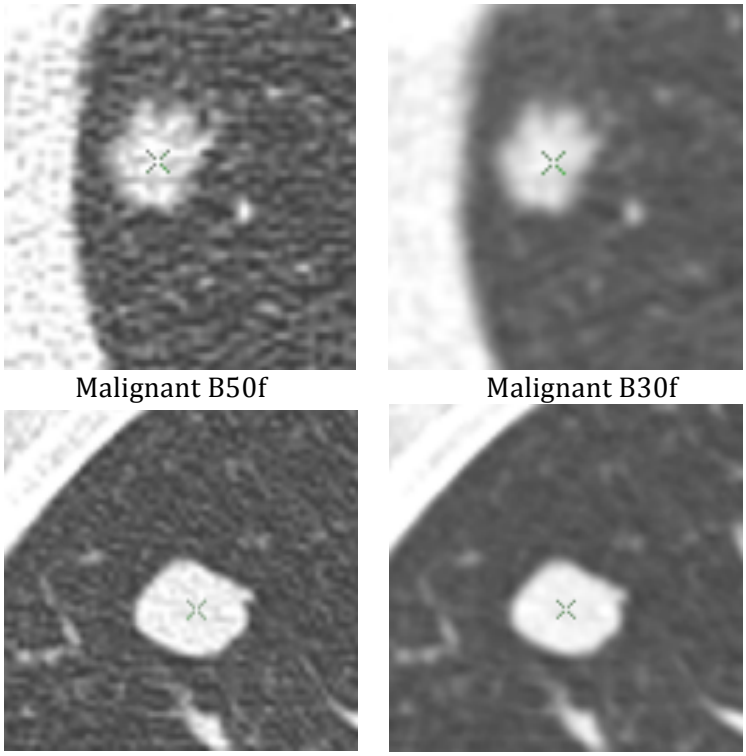


Fig. 11 The effect of the CT image reconstruction filter

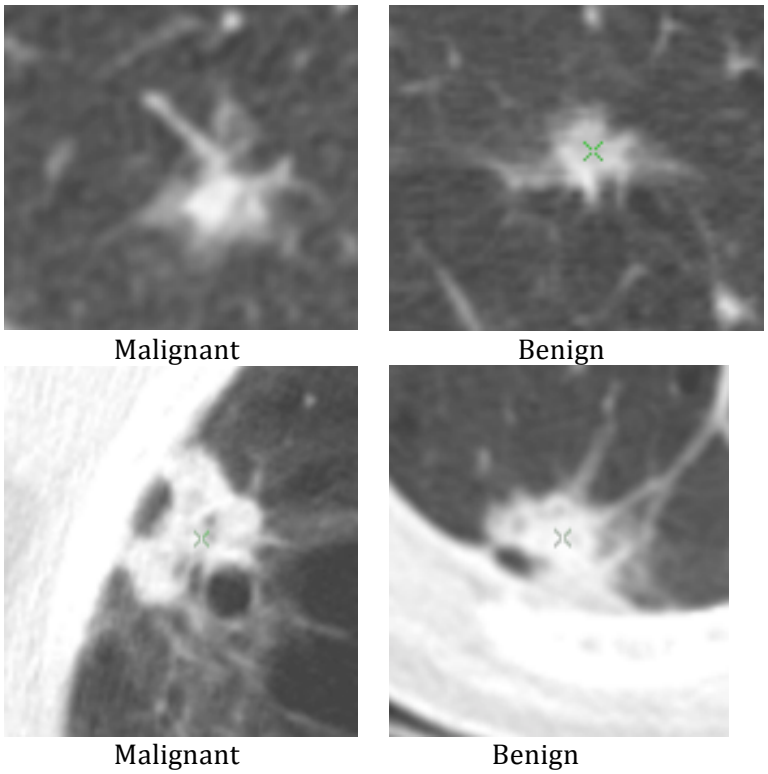
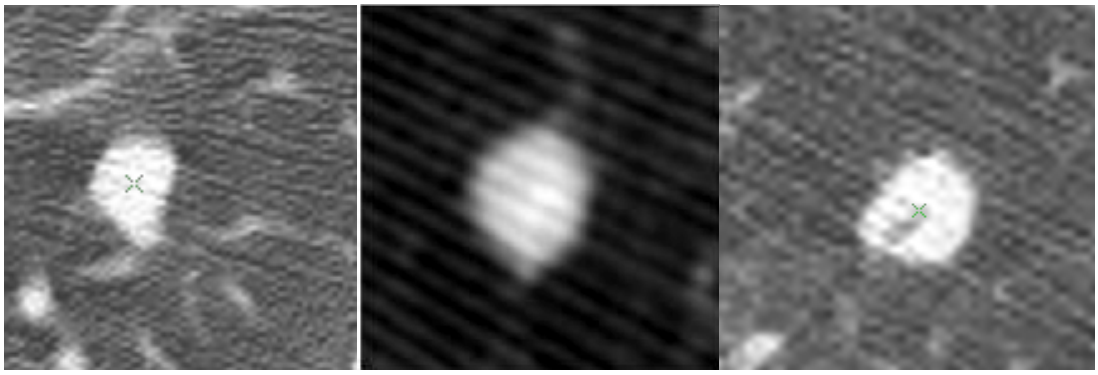


Fig. 12 Examples of nodules with similar complex presentations



Malignant Benign Benign

Fig. 13 Examples of images in which the nodules intensity is impacted by structured scanner noise

There have been a number of studies on characterizing the malignancy status of pulmonary nodules from CT images reported in the literature [2-3, 17-24]; performances have been reported in terms of area under the curve (AUC) for ROC analysis in the range of 0.79-0.92. Of these studies, only three have used nodules from a lung cancer screening study [2, 18, 21]. These three studies all used the same dataset that has over 400 benign nodules and less than 80 malignant nodules and is dominated by a large number of very small benign nodules (smaller than any malignant nodule); which is a major factor in determining the AUC performance [1].

Limitations of this study

For this retrospective study, the data are on the order of 10 years old and does not reflect the impact of recent changes in CT technology. Many of the scans (65%) especially those from the NLST (94%) had a slice thickness greater than or equal to 2mm which is half the image resolution specified by current lung cancer screening guidelines. We did not consider nonsolid nodules since these are a different phenotype and lack representation in our database in sufficient numbers; these nodules should be the subject of a future study.

The scans for this study span a wide range of CT models and parameter settings. No image preprocessing was performed to compensate for different image scanner parameters, especially with respect to image reconstruction filtering and image noise (see Figs. 11, 13); however, image resampling to isotropic space was performed for feature evaluation.

CONCLUSION

The task of nodule classification in the context of lung cancer screening has the following distinguishing characteristics: (a) low image resolution, (b) high image noise (c) tremendous size range of nodules, (d) different size distributions for benign and malignant nodules and (e) a large variation in CT scanner acquisition parameters. For a classification system to be relevant to lung cancer screening, all these issues need to be considered. Ignoring size issues may result in overly optimistic performance results that reflect only the imbalance in the test set size distribution. This imbalance causes the system to confound the population-based difference in size distribution with the patient-specific image features of

the nodule. The predictive power associated with the nonsize-impacted image features may be determined by using a size-balanced dataset.

In this study, we have explored the size issues using a large size-enriched dataset of 736 nodules by combining images from the two largest lung cancer screening studies. Our results indicate that there is a measurable improvement in the prediction of malignancy by using image features over size alone; however, the main predictor is size and this must be carefully accounted for when attributing the benefit of other image features. The overoptimistic performance and biased learning due to class size distribution differences can be avoided by using a size–balanced evaluation dataset. The tremendous size range of pulmonary nodules encountered in screening may be addressed by binning, that is, training a set of classifiers on a small nodule size ranges and selecting the size-specific classifier for a given case. In any case, appropriate representation of the large size range will require much larger data sets than the 736 cases that we used in this study.

In this study, the incremental improvement of the AUC over size was only 0.047 for the full unbalanced dataset. The balanced data test set had a statistically significant improved performance ($p=0.01$) with the IAUC increasing from 0.132 to 0.198 when trained on the balanced data, which was further increased to 0.232 by using two bins. This provides a modest improvement over size information alone.

The population-based probability of malignancy based on size is a major prediction factor that is known a priori from the analysis of cancer screening studies and practice. The essential issue for a patient-based nodule characterization system is to determine the probability of malignancy conditioned on a size, rather than the joint probability of malignancy and size.

There are several technical improvements that may lead to improved classification performance including higher resolution images, standardization on scanner parameters and reduction in image noise.

ACKNOWLEDGMENTS

We gratefully acknowledge the contributions of Robert Gillies and Yoganand Balagurunathan of Moffitt Cancer Center and support by U01 CA143062-04 in the preparation of the NLST image dataset. We gratefully acknowledge the support of Claudia Henschke and David Yankelevitz of Icahn School of Medicine at Mount Sinai, in making the Weill Cornell dataset available to this project. This work was funded in part by the Flight Attendants' Medical Research Foundation (FAMRI) and NSF award CBET-1014813.

Conflict of interest Yiting Xie, and Artit Jirapatnakul declare that they have no conflict of interest. Anthony Reeves financial and research disclosures: *Financial:* (1) General Electric: Dr. Reeves is a co-inventor on a patent and other pending patents owned by Cornell Research Foundation (CRF) which are nonexclusively licensed and related to technology involving computer-aided diagnostic methods, including measurement of pulmonary nodules in CT images. (2) D4Vision Inc.: Dr. Reeves is the owner of D4Vision Inc. a company that licenses software for image analysis. *Research:* Dr. Reeves receives research support in the form of grants and contracts from: NSF and the Flight Attendants' Medical Research Institute.

Ethical standard All the image data was de-identified, and was obtained retrospectively from two major multi-center trials (ECLAP and NLST), which had appropriate IRB approval for the use of human subjects.

REFERENCES

[1] Jirapatnakul A, Reeves AP, Apanasovich TV, and Biancardi A, Yankelevitz DF, Henschke CI (2007) Pulmonary nodule classification: size distribution issues, ISBI, pp 1248-1251.

[2] Suzuki K, Li F, Sone S, Doi K (2005) Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network, IEEE Trans Med Imaging 24(9): 1138–1150.

[3] Shah SK, McNitt-Gray MF, Rogers SR, Goldin JG, Suh RD, Sayre JW, Petkovska I, Kim HJ, Aberle DR (2005) Computer-aided diagnosis of the solitary pulmonary nodule, Acad Radiol, 12(5): 570–575.

[4] Jirapatnakul AC, Reeves AP, Apanasovich TV, Biancardi AM, Yankelevitz DF, Henschke CI (2008) Characterization of pulmonary nodules: effects of size and feature type on reported performance. In SPIE International Symposium on Medical Imaging, pp 69151E.

[5] The International Early Lung Cancer Action Program Investigators (2006) Survival of patients with stage I lung cancer detected on CT screening, N Engl J Med 355: 1763-1771.

[6] The National Lung Screening Trial Research Team (2011) Reduced lung-cancer mortality with low-dose computed tomographic screening, N Engl J Med, 365(5): 395-409.

[7] Reeves AP, Chan AB, Yankelevitz DF, Henschke CI, Kressler B, Kostis WJ (2006) On measuring the change in size of pulmonary nodules, IEEE Trans Med Imaging 25(4): 435-450.

[8] Jirapatnakul AC, Reeves AP, Apanasovich TV, Cham MD, Yankelevitz DF, Henschke CI (2007) Characterization of solid pulmonary nodules using three-dimensional features. In SPIE International Symposium on Medical Imaging, pp 65143E.

[9] Prokop RJ, Reeves AP (1992) A survey of moment-based techniques for unoccluded object representation and recognition, *CVGIP: Graphical Models and Image Processing*, 54(5): 438-460.

[10] Takashima S, Sone S, Li F, Maruyama Y, Hasegawa M, Matsushita T, Takayama F, Kadoya M (2003) Small solitary pulmonary nodules (≤1 cm) detected at population-based CT screening for lung cancer: reliable high-resolution CT features of benign lesions, AJR, 180(4): 955-964.

[11] Kawata Y, Niki N, Ohmatsu H, Kusumoto M, Kakinuma R, Mori K, Nishiyama H, Eguchi K, Kaneko M, Moriyama N (1999) Curvature based characterization of shape and internal intensity structure for classification of pulmonary nodules using thin-section CT images, SPIE Medical Imaging, 3661: 541.

[12] Lorensen WE, Cline HE (1987) Marching cubes: A high resolution 3D surface construction algorithm, ACM SIGGRAPH Computer Graphics, 21(4): 163-169.

[13] Monga O, Deriche R, Rocchisani JM (1991) 3d edge-detection using recursive filtering—application to scanner images. CVGIP-Image Understanding, 53(1): 76–87, doi:10.1016/1049-9660(91)90006-B.

[14] Dudani SA (1976) The distance-weighted k-nearest-neighbor rule, IEEE SMC, 6(4): 325-327.

[15] Joachims T (1999) Making large-scale SVM learning practical, MIT Press.

[16] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996), A simulation study of the number of events per variable in logistic regression analysis, Journal of Clinical Epidemiology, 49(12): 1373-1379.

[17] Kawata Y, Niki N, Ohmatsu J (2001) Curvature-based internal structure analysis of pulmonary nodules using thoracic 3D CT images, Systems and Computers in Japan 32(11): 9-19.

[18] Aoyama M, Li Q, Katsuragawa S, Li F, Sone S, Doi K (2003) Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose ct images, Medical Physics 30(3): 387–394.

[19] Shah SK, McNitt-Gray MF, Rogers SR, Goldin JG, Suh RD, Sayre JW, Petkovska I, Kim HJ, Aberle DR (2005) Computer aided characterization of the solitary pulmonary nodule using volumetric and contrast enhancement features, Academic Radiology 12(10): 1310–1319.

[20] Way TW, Hadjiiski LM, Sahiner B, Chan HP, Cascade PN, Kazerooni EA, Bogot N, Zhou C (2006) Computer-aided diagnosis of pulmonary nodules on CT scans: Segmentation and classification using 3D active contours, Medical Physics, 33(7): 2323-2337.

[21] Armato III SG, Altman MB, Wilkie J, Sone S, Li F, Doi K, Roy AS (2003) Automated lung nodule classification following automated nodule detection on CT: A serial approach, Med. Phys, 30(6): 1188-1197.

[22] El-Baz A, Nitzken M, Khalifa F, Elnakib A, Gimel'farb G, Falk R, El-Ghar MA (2011) 3D shape analysis for early diagnosis of malignant lung nodules, Information Processing in Medical Imaging, 6801: 772-783.

[23] Wu H, Sun T, Wang J, Li X, Wang W, Huo D, Lv P, He W, Wang K, Guo X (2013) Combination of radiological and gray level co-occurrence matrix textural features used to distinguish solitary pulmonary nodules by computed tomography, J Digit Imaging, 26(4): 797-802.

[24] Han F, Wang H, Song B, Zhang G, Lu H, Moore W, Zhao H, Liang Z (2013) A new 3D texture feature based computer-aided diagnosis approach to differentiate pulmonary nodules, SPIE Med Imaging, pp 86702Z.

[25] Jirapatnakul AC, Reeves AP, Biancardi AM, Yankelevitz DF, Henschke CI (2009) Semi-automated measurement of pulmonary nodule growth without explicit segmentation. In: IEEE International Symposium on Biomedical Imaging, pp 855–858.

[26] Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn. Lett. 27(8): 861–874.

[27] DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach, Biometrics, 44(3): 837-845.

[28] Joachims T (1999) Making large-scale SVM learning practical, advances in kernel methods - support vector learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, pp 169-184.

[29] Henschke CI, Yankelevitz DF, Mirtcheva R, McGuinness G, McCauley D, Miettinen OS and the ELCAP Group (2002) CT screening for lung cancer frequency and significance of part-solid and nonsolid nodules, AJR, 178(5): 1053-1057.

[30] Jirapatnakul AC (2011) Computer methods for pulmonary nodule characterization from CT images, Master's thesis, Cornell University.

[31] Deriche R (1987) Using Canny's criteria to derive a recursively implemented optimal Edge Detector, IJCV, pp 167-187.

APPENDIX

The appendix includes four tables. Table 1 lists all 46 3D image features used in the experiments, their abbreviations, and their mathematical definitions. They are divided into four categories: morphological, density, curvature and margin gradient. Morphological and density features are based on standard moments [9] while other features including curvature and gradient features that involve a polygon surface representation are described in [30]. Table 2 gives the parameter search space in the training stage for the four different classifiers: dwNN, SVM with polynomial kernel (SVM-P), SVM with Radial Basis Function kernel (SVM-R) and logistic regression (LOG). Table 3 gives the optimal features for LOG classifier on the full unbalanced dataset using fivefold cross-validation. Table 4 gives the optimal features for LOG classifier on the balanced dataset with balanced training and unbalanced training conditions.

Features

Table 1 Nodule feature description.

Name	Definition	Equation
Morphological Features		
gvol	volume (mm^3)	$m_{000} \cdot V_{voxel}$
gsa	surface area (mm^2)	$\sum (V_{xy} \cdot x_{res} \cdot y_{res} + V_{xz} \cdot x_{res} \cdot z_{res} + V_{yz} \cdot y_{res} \cdot z_{res})$
gvsr	volume to surface area ratio	$\frac{gvol}{gsa}$

gcmp	compactness	$\frac{4\pi \cdot gvol}{gsa^{\frac{3}{2}}}$
gdiml	length of ellipsoid of inertia	$\lvert V_x \rvert$
gdimw	width of ellipsoid of inertia	$\lvert V_y \rvert$
gdimh	height of ellipsoid of inertia	$\lvert V_z \rvert$
garlh	length to height ratio	$\frac{gdiml}{gdimh}$
garlw	length to width ratio	$\frac{gdiml}{gdimw}$
garwh	width to height ratio	$\frac{gdimw}{gdimh}$
gsph	sphericity	$\frac{gcmp}{garlh}$
grr	orientation angle: roll (degrees)	$\cos^{-1}\left(\frac{(0, V_z(y), V_z(z)}{\lvert V_z \rvert}\right) \cdot \mathrm{sgn}(V_z(y))$
grp	orientation angle: pitch (degrees)	$\cos^{-1}\left(\frac{(V_x(x), 0, V_x(z))}{\lvert V_x \rvert}\right) \cdot \mathrm{sgn}(V_x(z))$
gry	orientation angle: yaw (degrees)	$\cos^{-1}\left(\frac{(V_x(x), V_x(y), 0)}{\lvert V_x \rvert}\right) \cdot \mathrm{sgn}(V_x(y))$
m_{200}	normalized morphological moment of order (2,0,0)	M_{200}
m_{020}	normalized morphological moment of order (0,2,0)	M_{020}
m_{002}	normalized morphological moment of order (0,0,2)	M_{002}
attach	attachment ratio	$\frac{\sum V_{Di}}{\sum V_{Ri}}$
Density Features		
dmass	density mass	$\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}\sum_{z=0}^{L-1} I(x,y,z)$
dmd	mean density	$\frac{1}{N_{all}}\sum_{0}^{N_{all}-1} I(x,y,z)$
dsd	density histogram standard deviation	$\sqrt{\frac{1}{N_{all}}\sum_{0}^{N_{all}-1}(I(x,y,z)-\bar{\mu})^2}$
dskew	density histogram skewness	$\frac{\sum_0^{N_{all}-1}(I(x,y,z)-\bar{\mu})^3}{dsd^3}$
dkurt	density histogram kurtosis	$\frac{\sum_0^{N_{all}-1}(I(x,y,z)-\bar{\mu})^4}{dsd^4} - 3$
ddiml	density based length of ellipsoid of inertia	$\lvert D_x \rvert$
ddimw	density based width of ellipsoid of inertia	$\lvert D_y \rvert$

ddimh	density based height of ellipsoid of inertia	$\lvert D_z \rvert$
darlh	density based length to height ratio	$\frac{\lvert D_x \rvert}{\lvert D_z \rvert}$
darlw	density based length to width ratio	$\frac{\lvert D_x \rvert}{\lvert D_y \rvert}$
darwh	density based width to height ratio	$\frac{\lvert D_y \rvert}{\lvert D_z \rvert}$
d_{200}	normalized densitiometric moment of order (2,0,0)	D_{200}
d_{020}	normalized densitiometric moment of order (0,2,0)	D_{020}
d_{002}	normalized densitiometric moment of order (0,0,2)	D_{002}
Curvature Features		
cmin	minimum curvature	$\min\{C_{T_i}\}$
cmax	maximum curvature	$\max\{C_{T_i}\}$
cran	range of curvature	$\max\{C_{T_i}\} - \min\{C_{T_i}\}$
cmean	mean curvature	$mean\{C_{T_i}\}$
csd	standard deviation of curvature	$sd\{C_{T_i}\}$
cskew	skewness of curvature	$\frac{\sum_i (C_{T_i} - mean\{C_{T_i}\})^3}{sd\{C_{T_i}\}^3}$
ckurt	kurtosis of curvature	$\frac{\sum_i (C_{T_i} - mean\{C_{T_i}\})^4}{sd\{C_{T_i}\}^4} - 3$
Margin Gradient Features		
tmin	minimum gradient	$\min\{G_{T_i}\}$
tmax	maximum gradient	$\max\{G_{T_i}\}$
tran	range of gradient	$\max\{G_{T_i}\} - \min\{G_{T_i}\}$
tmean	mean gradient	$mean\{G_{T_i}\}$
tsd	standard deviation of gradient	$sd\{G_{T_i}\}$
tskew	skewness of gradient	$\frac{\sum_i (G_{T_i} - mean\{G_{T_i}\})^3}{sd\{G_{T_i}\}^3}$
tkurt	kurtosis of gradient	$\frac{\sum_i (G_{T_i} - mean\{G_{T_i}\})^4}{sd\{G_{T_i}\}^4} - 3$

All the features, with the exception of the attachment ratio feature, the curvature and the margin features were derived from 3D image moments defined in [9]. The 3D image moment of order (p+q+r) is defined as:

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

where $v(x,y,z)$ is a discrete function of size ($M \times N \times L$) and can be binary or grayscale. All moment-related features in this paper are standard moments, which are normalized with

respect to scale, translation and rotation [9]. When $v(x, y, z)$ is binary, the moments are normalized morphological moments M_{pqr}. When $v(x, y, z)$ is grayscale (voxel intensity), the moments are normalized densitiometric moments D_{pqr}.

The voxel size is defined as:

$$V_{voxel} = x_{res} \cdot y_{res} \cdot z_{res}$$

where x_{res}, y_{res}, z_{res} are the x, y, z image resolutions. The intensity of V_{voxel} is $I(x, y, z)$.

A surface voxel can be:
V_{xy}: surface voxel in the surface perpendicular to the z-axis;
V_{xz}: surface voxel in the surface perpendicular to the y-axis;
V_{yz}: surface voxel in the surface perpendicular to the x-axis.

The orientation is derived from the solution of the eigenproblem: $Av = \lambda v$, where v is a vector and A is defined as:

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

The solution of this problem, eigenvectors (V_x, V_y, V_z) and eigenvalues $\lambda_0 > \lambda_1 > \lambda_2$, form the major principal axis V_x, the intermediate principal axis V_y, the minor principal axis V_z [9, 30]. When m_{pqr} is the geometric moment, the major, intermediate and minor axes are geometric: V_x, V_y, V_z. When m_{pqr} is the density moment, the axes are density based: D_x, D_y, D_z.

The roll, pitch, yaw angles are defined as the rotation of an object about the standard x-y-z axes. The roll angle γ is a rotation of γ about the x-axis. The pitch angle β is a rotation of β about the y-axis after the first rotation. The yaw angle α is a rotation of α about the z-axis after the first two rotations [9].

The attachment ratio is the ratio of the number of surface voxels along the border of the removed vessels and the nodule, V_{Di}, to the number of surface voxels of the segmented nodule, V_{Ri}.

The density statistics are computed using the central statistical moments, which are the summations of powers of the voxel density values (intensity) normalized to the mean value, $\bar{\mu}$.

$$\mu_p = \sum_{0}^{N_{all}-1} (v(x, y, z) - \bar{\mu})^p$$

where N_{all} is the number of voxels.

Surface curvature is defined as the rate of change of the surface normal ϕ with respect to the surface length. For 3D curvature measurement, a discrete piecewise linear model for the nodule surface is used. Curvature is estimated on a smoothed tessellated polygonal surface model of the segmented nodule, generated using the marching cubes algorithm. The resulting triangular polygonal surface representation is smoothed by replacing the location of a vertex by a weighted sum of neighboring vertices and itself. The nodule surface regions where attached structures such as vessels have been removed are deleted.

The surface curvature is estimated for each pair of vertices in the remaining triangular mesh. First, the normal of each triangle is computed: given vertices $\{V_l\, V_c\, V_d\}$ of a triangle T_i the surface normal N_i is given by:

$$N_i = \frac{\overrightarrow{V_l V_c} \times \overrightarrow{V_l V_d}}{|\overrightarrow{V_l V_c} \times \overrightarrow{V_l V_d}|}$$

The surface normal at each vertex is computed by averaging the surface normal of the triangles of which the vertex is a member:

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

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

The curvature is computed by taking the angular difference between the surface normals at a vertex and an adjacent vertex. The angular difference between the surface normal vectors ϕ_i and ϕ_a is:

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

For each vertex, the average curvature corresponding to all adjacent vertices is computed. For example, the average curvature for vertex V_i can be computed as:

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

where n is the number of adjacent vertices.
Finally, each triangle in the polygonal representation is assigned a curvature value based on the average of the curvatures of the vertices which comprise the triangle:

$$C_{T_i} = \frac{(C_{V_i} + C_{V_d} + C_{V_c})}{3}$$

where V_i, V_d, V_c are the vertices of the triangle T_i.

Descriptive statistics of the distribution of curvatures over the entire nodule surface are used as curvature features.

The gradient features are used to measure the nodule margin, which is defined as the region along the nodule boundary and lung parenchyma. The 3D gradient is measured in the x, y, z directions as defined by Deriche [31].

To optimize the gradient estimate, at each triangle, 10 gradient samples are evaluates along the surface normal vector through the center of the triangle and the maximum gradient is recorded:

$$G_{T_i} = \max\{G_{T_{i0}}, G_{T_{i2}}, \dots, G_{T_{i9}}\}$$

where $G_{T_{ij}}$ is a gradient sample for triangle T_i.

Descriptive statistics of the distribution of these maximum gradients over the entire nodule surface are used as gradient features.

Parameter optimization

Table 2 Parameter search space for each classifier.

Classifier	Parameter	Search space
dwNN	σ	0.1, 0.15, 0.2, 0.3, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 2.0, 2.4, 2.8, 3.2, 4.0, 4.8, 5.4, 6.4, 8.0, 9.6, 11.2, 12.8, 14.4, 16.0
SVM-P	*d*	1, 2, 3, 4
SVM-P	*c*	10e-6, 2×10e-6, 4×10e-6, …, 0.262, 0.524, 1.049
SVM-R	*g*	10e-3, 2×10e-3, 4×10e-3, …, 0.512, 1.024
SVM-R	*c*	10e-4, 2×10e-4, 4×10e-4, …, 3.277, 6.554
LOG	*n*	1, 2, 3, …, min(number of positive samples, number of negative samples)/10

Table 3 LOG classifier optimal features on full unbalanced dataset (fold one to fold five listed from 5-fold cross validation). The number of features is *n*=25.

fold	Features
1	gdimh, ddimh, gsa, gvol, cran, dmass, cmin, ddimw, gdimw, cmean, gvsr, ddiml, gdiml, tmin, cmax, gcmp, csd, ckurt, dskew, tran, dmd, dkurt, dsd, garwh, tskew
2, 3	gdimh, ddimh, gsa, gvol, dmass, cran, ddimw, gdimw, cmin, gvsr, ddiml, gdiml, cmean, tmin, cmax, gcmp, csd, ckurt, dskew, tran, dmd, tskew, tmean, garwh, darwh
4	cran, gdimh, ddimh, tmin, cmin, gsa, gvol, dmass, cmax, ddimw, gdimw, gdiml, ddiml, cmean, gvsr, csd, gcmp, ckurt, tran, dskew, tskew, garwh, tmean, darwh, tsd
5	cran, gdimh, ddimh, gsa, cmin, gvol, tmin, dmass, gdiml, ddimw, gdimw, ddiml, cmean, gvsr, gcmp, cmax, csd, ckurt, tmean, tran, dskew, tskew, gsph, dmd, cskew

Table 4 LOG classifier optimal features on full balanced dataset GA using balanced training and unbalanced training (fold one to fold five listed from 5-fold cross validation). The number of features is *n*=3.

fold	training	features
1, 3	balanced	tmean, dmd, tmax
2	balanced	tmean, dmd, csd
4	balanced	tmean, dmd, dsd
5	balanced	tmean, dmd, garwh
1-5	unbalanced	gdimh, ddimh, cran