

Predictive capabilities of statistical learning methods for lung nodule malignancy classification using diagnostic image features: an investigation using the Lung Image Database Consortium dataset

Matthew C. Hancock^a and Jerry F. Magnan^{a,*}

^aFlorida State University, Department of Mathematics, 208 Love Building, 1017 Academic Way Tallahassee, FL, USA, 32306-4510

ABSTRACT

To determine the potential usefulness of quantified diagnostic image features as inputs to a CAD system, we investigate the predictive capabilities of statistical learning methods for classifying nodule malignancy, utilizing the Lung Image Database Consortium (LIDC) dataset, and only employ the radiologist-assigned diagnostic feature values for the lung nodules therein, as well as our derived estimates of the diameter and volume of the nodules from the radiologists' annotations. We calculate theoretical upper bounds on the classification accuracy that is achievable by an ideal classifier that only uses the radiologist-assigned feature values, and we obtain an accuracy of 85.74 (± 1.14)% which is, on average, 4.43% below the theoretical maximum of 90.17%. The corresponding area-under-the-curve (AUC) score is 0.932 (± 0.012), which increases to 0.949 (± 0.007) when diameter and volume features are included, along with the accuracy to 88.08 (± 1.11)%. Our results are comparable to those in the literature that use algorithmically-derived image-based features, which supports our hypothesis that lung nodules can be classified as malignant or benign using only quantified, diagnostic image features, and indicates the competitiveness of this approach. We also analyze how the classification accuracy depends on specific features, and feature subsets, and we rank the features according to their predictive power, statistically demonstrating the top four to be spiculation, lobulation, subtlety, and calcification.

Keywords: computer-aided diagnosis (CAD), lung nodule classification, LIDC dataset, random forests, logistic regression, machine learning

1. INTRODUCTION

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

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

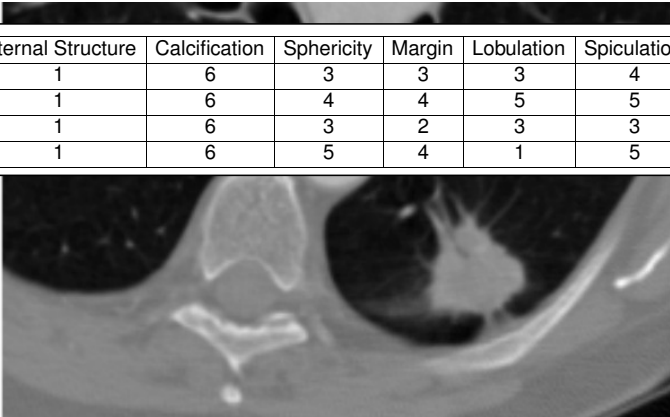
*Send correspondence to: magnan@math.fsu.edu

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

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

2. MATERIALS AND METHODS

2.1 Brief Overview of the LIDC Dataset



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

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

The LIDC dataset¹⁹ is a publicly available set of 1018 lung CT scans collected through various universities and organizations. In addition to the CT image data, manual annotations by anonymous radiologists for each scan are provided. These annotations are made with respect to the following types of structures:

1. Lung nodules whose largest diameter is greater than 3mm.
2. Lung nodules whose largest diameter is less than 3mm.
3. Non-nodule structures whose largest diameter is greater than 3mm.

For each of these types, the location of the structure is given in image coordinates, as determined by each of the four physicians, with no forced consensus about their existence, or location, imposed. It is the first type of structures (i.e., lung nodules with largest diameter ≥ 3 mm) that we analyze in this work. For this type of structure, additional annotations are assigned, including manually-drawn contours of the nodule boundaries in the CT scan slices, quantified values for a variety of nodule features, and a quantified value of the estimation of the nodule's malignancy at the time of assessment. The eight quantified nodule features, and the corresponding malignancy quantification, along with the features' respective rating systems, are listed and described in Table 1. An example nodule from the dataset, along with the assigned diagnostic feature values, is shown in Figure 1.

Table 1. Features annotated by radiologists in the LIDC dataset and associated rating system used.

Feature	Subtlety (Ordinal)	Internal Structure (Categorical)	Calcification (Categorical)
Rating System	1 Extremely Subtle 2 3 4 5 Obvious	1 Soft Tissue 2 Fluid 3 Fat 4 Air	1 Popcorn 2 Laminated 3 Solid 4 Non-central 5 Central 6 Absent
Feature	Sphericity (Ordinal)	Margin (Ordinal)	Lobulation (Ordinal)
Rating System	1 Linear 2 3 Ovoid 4 5 Round	1 Poorly-defined 2 3 4 5 Sharp	1 No Lobulation 2 3 4 5 Marked Lobulation
Feature	Spiculation (Ordinal)	Texture (Ordinal)	Malignancy (Ordinal)
Rating System	1 No Spiculation 2 3 4 5 Marked Spiculation	1 Non-solid / Ground Glass 2 3 Mixed-solid 4 5 Solid	1 Highly Unlikely for Cancer 2 3 Indeterminate 4 5 Highly Likely for Cancer

2.1.1 Our Use of the LIDC Dataset

Leaving out 399 cases that may contain inconsistent labelings (see our related article²⁰ and the LIDC data-hosting site²¹ for more detail), there are 4384 nodule annotations, which is further reduced to 2817 by excluding indeterminate cases (as described in the following section). Each nodule may have been assigned between one and four annotations, depending on the level of agreement between the four annotators of the nodule belonging to the first type of structure. The physical nodules lack a universal, unique identifier among the many annotations and thus, it is difficult to ascertain which annotations refer to the same physical nodule in a scan without careful visual inspection. Algorithmically, it is possible to roughly determine which annotations refer to identical nodules by comparing the coordinates and overlap of annotations. However, this process requires somewhat arbitrary choices to be made in order to determine when multiple annotations may actually refer to the same nodule. For example, one would need to decide at what percentage of overlap, or at what average distance between annotation coordinates, multiple annotations would be declared to refer to the same physical nodule. For these reasons, we treat each annotation in the LIDC dataset as a unique sample. We describe the two statistical learning methods that we use, and our methodology, in the following section.

2.2 Malignancy Classification via Statistical Learning

We treat the statistical approximation of the malignancy category of nodules as a binary classification problem for ‘malignant’ vs. ‘benign’ by thresholding the radiologist-assigned malignancy values in such a way that malignancy values below 3 (i.e., 1 and 2) are categorized as benign, and values above 3 (i.e., values of 4 and 5) are categorized as malignant. We exclude cases that are labeled by a radiologist as having an indeterminate malignancy (i.e., an assigned value of 3). Thus, excluding from the dataset annotations with a malignancy value of 3, we are left with 2817 annotations, each of which consists of the quantified diagnostic image features (which are the input features for a nodule) and its assigned malignancy category (which is the nodule’s corresponding target label). The distribution of values for each of the eight diagnostic input features and for the distribution of malignancy is shown in Figure 2, with the values for each feature defined in Table 1.

More specifically, we treat the quantified features as random vectors, X , and malignancy values as random variables, Y , and we consider each annotation as an independent draw from the joint distribution, (X, Y) . It may be possible that separate dataset samples refer to the same physical nodule; however, we consider these instances to be different realizations of the random quantity, (X, Y) , where the source of randomness is from noise (e.g., due to accidental mislabelings) and from natural variations of the quantified feature values (e.g., due to varying annotator experience and training). This leads to a consistent view of the data. The statistical learning methods then model the conditional probability distribution, $P(Y|X)$, of malignancy, given the quantified feature values.

To generate the malignancy category from the annotated nodule features algorithmically, we employ two statistical learning techniques for classification. The first, logistic regression, is a linear method, while the

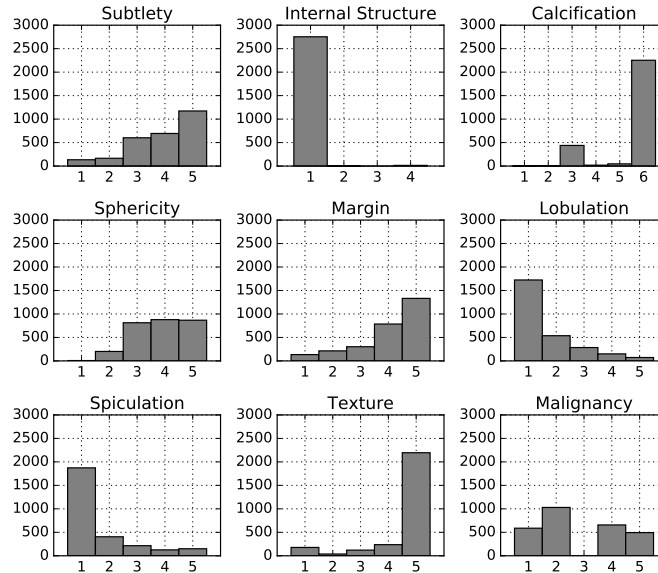


Figure 2. Distribution of annotation values for image features and malignancy. Note the excluded bin for the indeterminate malignancy value of 3.

second, random forests (which is based on decision trees), is a nonlinear method.²² These techniques use a subset of the data to learn a mapping – from the diagnostic image features as inputs, to the malignancy category as output – during the training (or learning) phase of the algorithms. In the testing phase, the accuracy is evaluated on a subset of the data that the algorithms did not use in any way during the training phase, i.e., on the testing data.

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

2.2.1 Experiment One

The purpose of the first experiment is to determine how well the radiologists’ categorization of malignancy from diagnostic image features can be approximated by statistical learning algorithms. We train both the linear and nonlinear classifier on a randomly chosen subset of the data (with each subset containing approximately 75% of the whole dataset), and test the accuracy on the remaining 25% of the dataset. We repeat this procedure 1000 times to obtain robust statistical results. Note that there is an overall class-imbalance of about 9% towards the ‘benign’ class, and so, a classifier that always chooses the ‘benign’ class, independent of the input vector, will achieve a classification accuracy of approximately 59%. This should be considered the baseline accuracy for comparison with any given classifier’s performance.

In addition, we note that ground-truth estimates of two additional features, the nodule diameter and volume, can be made from the nodule boundary contours provided in each radiologist’s nodule annotation. We treat these two features separately from the eight diagnostic features given in Table 1 for three reasons. First, since the diameter and volume are given only implicitly by the annotation contours, choices must be made as to how to define and algorithmically extract these two quantities. Second, since their values are given by positive real numbers, these two features differ from the others given in Table 1, whose values are restricted to a small, finite set of positive integers. Third, we are able to derive a theoretical upper bound on the classification accuracy when

the diameter and volume features are excluded, as we describe in Section 2.2.3. Hence, in the first experiment, we analyze two cases. The first includes diameter and volume estimates, along with the eight diagnostic image features given in Table 1, and thus, employs a total of ten input features. The second excludes the diameter and volume features, leaving a total of eight input features. We repeat the process described in the preceding paragraph (involving the 75%/25% training/testing dataset split and 1000 trials) with, and without, the nodule diameter and volume estimates. We record the following error metrics averaged over the 1000 trials performed in Experiment One:

1. Classification Accuracy
2. Sensitivity, or True Positive Rate (TPR)
3. Area-under-the-ROC-curve score (AUC)

2.2.2 Experiment Two

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

We generate ad-hoc rankings of the features via the following metrics:

1. Single-feature accuracy – this is defined as the classification accuracy when a specific feature is used by itself.
2. Percent feature-significance – this is defined as the percentage of cases for which the addition of a specific feature to any subset not containing that feature produces a statistically significant increase (according to a paired t -test) in accuracy.
3. The geometric mean of the single-feature accuracy and the percent-feature significance.
4. The Random Forest feature-importance metric.²² This is used for comparison with the previous metrics.

2.2.3 Maximum Attainable Accuracies

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

There is not a one-to-one correspondence between the input values (i.e., the quantified diagnostic feature values) and output labels (i.e., the two malignancy categories) in the annotated dataset since there are instances across the entire dataset where multiple annotations of nodules that were made by the radiologists have identical input feature values, but correspondingly different output labels. We will refer to such sets of multiple annotations as degenerate groups. There are a total of 151 such groups, involving 1441 annotations out of the total 2817 annotations considered in the dataset. The size of each degenerate group varies, as does the number of annotations associated with a particular output label. For example, a particular degenerate group consists of 7 annotations (with identical diagnostic feature inputs), where 6 of the 7 annotations assign a label of malignant and where the remaining annotation assigns a label of benign.

Keeping the degenerate groups in mind, we consider the situation where the training dataset has been selected, and the classifier has determined its parameters from it. Thus, we are at the testing phase. If the classifier is to

generalize well, then it should correctly classify all examples in the unseen testing data. However, the classifier's accuracy cannot theoretically be 100% on the testing dataset because it is limited in the following two ways by the presence of degenerate groups in the datasets:

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

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

3. RESULTS FROM EXPERIMENTS ONE AND TWO

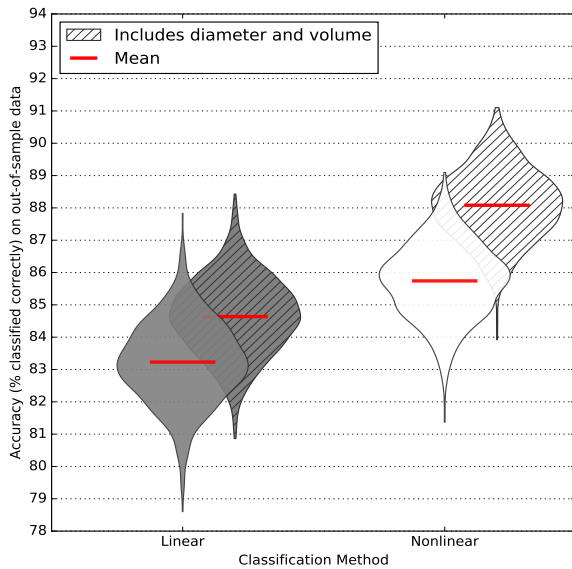


Figure 3. Distribution of accuracies for Experiment One.

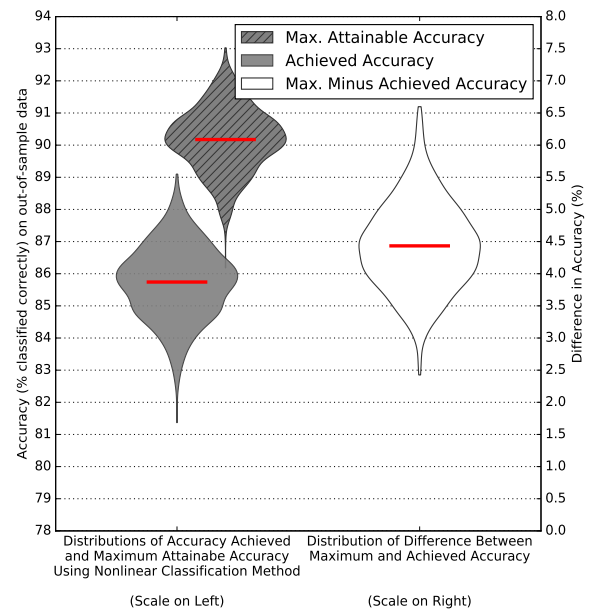


Figure 4. Comparison of maximum attainable and achieved accuracies with nonlinear classification (diameter and volume excluded).

We describe here the results from the experiments specified in Sections 2.2.1 and 2.2.2. In Section 4, we discuss and interpret these results.

The results for Experiment One are shown in violin plots of classification accuracy in Figures 3 and 4 and are summarized in Table 2. Each violin plot is a smooth estimate of the probability density function (for the

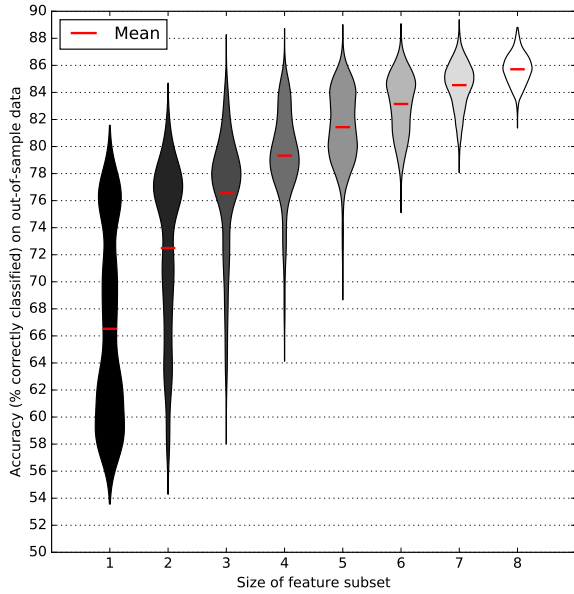


Figure 5. Distribution of accuracies for Experiment Two.

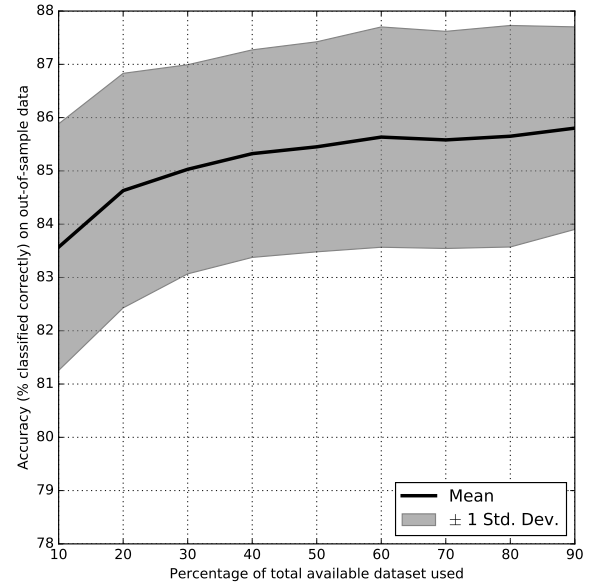


Figure 6. Testing set accuracy as a function of dataset size.

Table 2. Summary of results from Experiment One.


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

distribution of classification accuracy over the 1000 trials), which is symmetrically mirrored across a vertical line. Thus, the area of a region within a violin plot that is located between any two chosen values of the classification accuracy on the vertical axis is proportional to an empirical estimate of the probability of observing a value of the classification accuracy between the two bounding values chosen. In Figure 3, the results for the linear and nonlinear classifiers are shown in gray and white, respectively; and, the results for the cases when both the diameter and volume features are included, and excluded, are shown with cross-hatching, and without, respectively. In Figure 4, for the nonlinear method when the diameter and volume features are excluded, the distribution of the theoretical maximum accuracy and of the achieved accuracy are shown on the left in gray, with and without cross-hatching, respectively. The distribution of their respective differences is shown on the right, in white. Note that, in Figure 4, the scale for the accuracy plots is given on the left, while the scale for the difference-in-accuracy plot is given on the right.

The results for Experiment Two are shown in Figure 5, and in Table 3. Figure 5 shows, for particular fixed

feature-subset sizes, varying from one to eight, the obtained corresponding distribution of accuracies in a violin plot (as described previously). Table 3 ranks the features according to the measures described in Section 2.2.2, i.e., by the single-feature accuracy and percent feature-significance, by the geometric mean of the two, and lastly, by the average RF feature-importance metric, with the latter two providing an overall ranking of a feature's significance for classification.

Table 3. Rankings (see Section 2 for details) of the diagnostic features used.

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

4. DISCUSSION

The results for Experiment One are summarized in Table 2. When diameter and volume features are excluded, the mean accuracies for the linear and nonlinear classifiers are 83.23 (± 1.25)% and 85.74 (± 1.14)%, respectively, as can be seen in Figure 3. The nonlinear classifier performs slightly better (by 2.51%) on average, indicating that the malignancy category is slightly better explained by a nonlinear transformation of the input features than by a linear combination (i.e., by a weighted sum) of the features. Correspondingly, we compute average sensitivities (for $t = \frac{1}{2}$) of 0.801 (± 0.022) and 0.843 (± 0.024), and we calculate average AUC scores of 0.916 (± 0.009) and 0.932 (± 0.012). The classification accuracy improves when the diameter and volume features are included, with the mean accuracy for the linear classifier increasing by 1.41%, to 84.64 (± 1.18)%, and for the nonlinear classifier, increasing by 2.34%, to 88.08 (± 1.11)%, as shown in Figure 3. Correspondingly, average sensitivities (at $t = \frac{1}{2}$) are calculated to be 0.791 (± 0.024) and 0.846 (± 0.022), with average AUC scores of 0.930 (± 0.008) and 0.949 (± 0.007). The observed increase in these metrics, when diameter and volume are included, demonstrates the relevance and importance of these features for classifying nodule malignancy more accurately.

As shown in Figure 4 in the white violin plot, when the diameter and volume features are excluded, the classification accuracy achieved over the 1000 trials with the nonlinear classifier varies from a minimum of 2.43% below the theoretical maximum attainable accuracy to a maximum of 6.60% below it. On average, the achieved accuracy is 4.43 (± 0.68)% below the theoretical maximum. We also observe that

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

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

This accuracy gap between the theoretical upper-bound for the classification accuracy and the actual classification accuracy achieved is due to overlap in the class-conditional probability distributions, i.e., in the distributions, $P(X|Y)$. Clearly, one source of overlap in the class-conditional distributions are the degenerate groups described in Section 2.2.3, i.e., groups of identical input feature vectors with conflicting malignancy-category output labels. A related overlap can occur when there are data samples that are close in feature space (under

some distance metric), without input features being necessarily identical, but whose malignancy categories conflict, i.e., when there are “near-degenerate” groups of input features with conflicting malignancy-category output labels. The overlap caused by either type of these data degeneracies reduces the theoretical, maximum-attainable classification accuracy which, in turn, reduces and limits the accuracy achievable with a real classifier.

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

1. Inherent variability in quantification of diagnostic features.
2. Lack of additional informative features.
3. Inadequate sample size and diversity, leading to non-optimal statistical estimates.

The first potential cause may be difficult to assess and quantify. However, in a study of the related task of lung nodule detection, Armato et al. found substantial variability across radiologists in the LIDC dataset.²³ Presently, the feature values are assigned by radiologists, but if the features were quantified algorithmically, then the variability (a potential source of classification error) would decrease.

The second potential cause is supported by the demonstrated increase in classification accuracy due to the addition of diameter and volume feature estimates, which clearly helps to discriminate better between malignant and benign lung nodules, and thus reduces the class-conditional overlap. Other image-derived local and global features used (that are not among the features in the LIDC dataset) may have a similar effect. For example, a radiologist could potentially take into account the relative location of the lung nodule when assigning a malignancy class to the nodule. In the LIDC dataset, neither a nodule’s relative location within the lung, nor its proximity to anatomical structures, was annotated by the radiologists viewing and analyzing the CT images, although such information could have an effect on assigning a nodule’s malignancy category, since certain locations are more probable for lung cancer.²⁴ We note that estimates of upper bounds for classification accuracy, or estimates of class-conditional distribution overlap, must be rederived for different sets of input features used for classification, and that this may be more difficult to do when real-valued input features, such as diameter and volume, are employed.

As a test of the third potential cause, we run experiments with the nonlinear classifier. For generating training datasets, we take random samples of the dataset of increasing sizes from 10% to 90% (of the total dataset of 2817 samples) in increments of 10%, setting the remaining data aside. For each randomly chosen training dataset, we randomly choose a fixed percentage of 10% of the remaining data to use as the corresponding testing dataset. We repeat this 1000 times for all the training set sizes. We observe in Figure 6 that the mean classification accuracy shows an overall increase of approximately 2.23%, with only a minor increase of about 0.16% after training set subsample sizes greater than 60% of the total dataset are used. This suggests that the maximum amount of diversity and information that the classifier can extract from the data for making classifications is already present in samples of sizes of at least 60% of the total dataset size. Hence, the third potential cause of class-conditional overlap is not likely playing a substantial role in our study.

In Experiment Two (see Figure 5, and Table 3), we observe that including more of the input features generally improves classification accuracy, and that there can be a significant variation in classification accuracy for input feature subsets of particular, fixed sizes (see Figure 5). We find that the standard deviation of the accuracy monotonically decreases with increasing subset size (except in the seven-to-eight case, where it increases very slightly, by 0.01%). Furthermore, we can see in Figure 5 that the distributions of accuracies can be non-Gaussian. For example, for feature subsets of size five, the distribution of classification accuracy is bimodal and varies from a low of 68.7% to a high of 89.0%. The upper mode is near the 84% mark, and the lower mode is near the 80% mark (which are above and below, respectively, the mean accuracy of 81.5% achieved with subsets of size five). In other words, certain combinations of five input-features lead more frequently to classification accuracies near 84%, while other combinations of five input-features achieve accuracy near 80% more often. This indicates that some feature combinations are better than others for classification, which we discuss in more detail below.

Finally, in Table 3, we observe that, according to the geometric mean of the single-feature accuracy and percent feature-significance, the four most relevant features (in decreasing order) are spiculation, lobulation, subtlety, and calcification; and, the four least relevant features (also, in decreasing order) are margin, texture, internal structure, and sphericity. Interestingly, some features, while not performing as well as others when used individually (see column one), have a tendency to improve accuracy when used in addition to other features, i.e., they have a positive synergistic effect for the classification of lung nodule malignancy. For example, calcification is quite poor as an individual predictor (and, in fact, is the worst individual malignancy indicator) since it ranks last in single-feature accuracy, but the calcification feature shows a very high tendency to significantly increase accuracy when used with other features, since it is second best, as measured by the percent feature-significance (see column two). Subtlety, spiculation, and lobulation, when individually added to a feature subset, all significantly increase classification accuracy over 98% of the time. Thus, each demonstrates the positive synergistic effect. We also observe in Table 3 that the RF feature-importance metric generally agrees with the geometric mean of the single-feature accuracy and the percent feature-significance, indicating the usefulness of the RF feature-importance metric as a possible surrogate for the combined feature relevance, which requires more work to determine.

5. CONCLUSION

For the LIDC dataset, we have shown that the malignancy label of a nodule can be accurately classified (4.14%, on average, below the theoretical maximum accuracy of an ideal classifier) when only quantified, diagnostic image features are used as inputs to standard statistical learning methods. We have also shown that the accuracy of the linear and nonlinear classifiers can be improved by 1.41% and 2.34% (from 83.23% to 84.64%, and from 85.74% to 88.08%) respectively, by including diameter and volume estimates. Using only the radiologist-quantifications of image features, we have also achieved an average AUC score of 0.932 with a nonlinear classifier, which improves to an average AUC score of 0.949 when the diameter and volume features are included.

Our results are comparable to those obtained with other approaches currently reported in the literature. For example, Firmino et al.,²⁵ reported an AUC of 0.91 using a combination of image-derived features and LIDC features. Our results yield a similar score, but use only the LIDC features. These other approaches, by contrast, often use image-based features that are extracted algorithmically, which, in some cases, may be not be medically-interpretable. This supports and motivates the further consideration of the CAD paradigm that first extracts and approximates a set of radiologist-interpretable diagnostic image features from a CT scan, and then, uses these features as inputs to a statistical learning method for classifying the corresponding nodule as malignant or benign. Furthermore, we have ranked the lung nodule features given in the LIDC dataset according to their predictive power (when used both singly and in combination with other features), showing that the four most relevant features for classification are (in decreasing order of relevance) spiculation, lobulation, subtlety, and calcification.

Future work for such a CAD approach will focus on the quantification and accurate approximation of interpretable, diagnostic image features extracted from CT scans (such as the most relevant four mentioned above). We believe this task, of algorithmically quantifying such features, is much more challenging than the determination of a lung nodule's malignancy category from already accurately quantified, nodule features because this task requires the resolution of the following problems:

1. Which features are considered to be “diagnostic” or “medically-relevant” is not currently standardized. The quantified image features in the LIDC dataset are not necessarily standard. The mention of the need for standardization of image features have been made recently in the literature.^{3,26} A terminology set such as the RadLex ontology might prove useful in this respect.²⁷
2. Precise, quantitative definitions of the features do not currently exist. To alleviate this, it first requires the understanding of any potential variability in quantifications of these features made by experts, which, in turn, requires large datasets of annotations made by many experts in order to analyze the distribution of the quantified feature-values. Alternately, a “bottom-up” approach could be taken, where the features are first defined mathematically. In this case, artificial nodules with a precise degree of the presence of a particular diagnostic image feature (e.g., spiculation) could be created and implanted in either digital or

physical phantoms. It would remain to be demonstrated empirically, in this “bottom-up” approach, that the mathematical definition agrees with the medically-assessed presence of any such feature.

3. Numerical and statistical analyses of methods used for quantification of any particular feature must be made to ensure that the algorithmic quantification of any diagnostic image feature is robust in clinical settings.

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

REFERENCES

- [1] Erasmus, J., Connolly, J., McAdams, H., and Roggli, V., “Solitary pulmonary nodules: part I. Morphologic evaluation for differentiation of benign and malignant lesions,” *Radiographics* (2000).
- [2] Erasmus, J., McAdams, H., and Connolly, J., “Solitary pulmonary nodules: part II. Evaluation of the indeterminate nodule,” *Radiographics* (2000).
- [3] Kim, H., Park, C. M., Goo, J. M., Wildberger, J. E., and Kauczor, H.-U., “Quantitative Computed Tomography Imaging Biomarkers in the Diagnosis and Management of Lung Cancer:,” *Investigative Radiology* **50**, 571–583 (Sept. 2015).
- [4] “NCI Dictionary of Cancer Terms - National Cancer Institute.” <http://www.cancer.gov/publications/dictionaries/cancer-terms> (accessed Mar. 2016).
- [5] Iwano, S., Nakamura, T., Kamioka, Y., and Ishigaki, T., “Computer-aided diagnosis: A shape classification of pulmonary nodules imaged by high-resolution CT,” *Computerized Medical Imaging and Graphics* **29**, 565–570 (Oct. 2005).
- [6] Raicu, D., “Modelling semantics from image data opportunities from LIDC,” *International Journal of Biomedical Engineering and Technology* **3**(1) (2009).
- [7] Zinovev, D., Raicu, D., Furst, J., and Armato III, S. G., “Predicting Radiological Panel Opinions Using a Panel of Machine Learning Classifiers,” *Algorithms* **2**, 1473–1502 (Nov. 2009).
- [8] Dhara, A. K., Mukhopadhyay, S., Alam, N., and Khandelwal, N., “Measurement of spiculation index in 3d for solitary pulmonary nodules in volumetric lung CT images,” in [*SPIE medical imaging*], 86700K–86700K, International Society for Optics and Photonics (2013).
- [9] Li, G., Kim, H., Tan, J. K., Ishikawa, S., Hirano, Y., Kido, S., and Tachibana, R., “Semantic characteristics prediction of pulmonary nodule using Artificial Neural Networks,” in [*Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*], 5465–5468, IEEE (2013).
- [10] Zhang, G., Xiao, N., and Guo, W., “Spiculation Quantification Method Based on Edge Gradient Orientation Histogram,” in [*Virtual Reality and Visualization (ICVRV), 2014 International Conference on*], 86–91, IEEE (2014).
- [11] Niehaus, R., Stan Raicu, D., Furst, J., and Armato, S., “Toward Understanding the Size Dependence of Shape Features for Predicting Spiculation in Lung Nodules for Computer-Aided Diagnosis,” *Journal of Digital Imaging* **28**, 704–717 (Dec. 2015).
- [12] Ciompi, F., “Bag-of-frequencies: a descriptor of pulmonary nodules in computed tomography images,” *IEEE Transactions on Medical Imaging* **34** (Apr. 2015).
- [13] Krewer, H., Geiger, B., Hall, L. O., Goldgof, D. B., Gu, Y., Tockman, M., and Gillies, R. J., “Effect of Texture Features in Computer Aided Diagnosis of Pulmonary Nodules in Low-Dose Computed Tomography,” 3887–3891, IEEE (Oct. 2013).
- [14] Han, F., Wang, H., Zhang, G., Han, H., Song, B., Li, L., Moore, W., Lu, H., Zhao, H., and Liang, Z., “Texture Feature Analysis for Computer-Aided Diagnosis on Pulmonary Nodules,” *Journal of Digital Imaging* **28**, 99–115 (Feb. 2015).
- [15] Depeursinge, A., Chin, A. S., Leung, A. N., Terrone, D., Bristow, M., Rosen, G., and Rubin, D. L., “Automated classification of usual interstitial pneumonia using regional volumetric texture analysis in high-resolution computed tomography,” *Investigative radiology* **50**(4), 261–267 (2015).

- [16] El-Baz, A., Nitzken, M., Vanbogaert, E., Gimel'farb, G., Falk, R., and El-Ghar, M. A., "A novel shape-based diagnostic approach for early diagnosis of lung nodules," in [*Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*], 137–140, IEEE (2011).
- [17] Tac, E. and Uur, A., "Shape and Texture Based Novel Features for Automated Juxtapleural Nodule Detection in Lung CTs," *Journal of Medical Systems* **39** (May 2015).
- [18] Kaya, A. and Can, A. B., "A weighted rule based method for predicting malignancy of pulmonary nodules by nodule characteristics," *Journal of Biomedical Informatics* **56**, 69–79 (Aug. 2015).
- [19] Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., Zhao, B., Aberle, D. R., Henschke, C. I., Hoffman, E. A., and others, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," *Medical physics* **38**(2), 915–931 (2011).
- [20] Hancock, M. C. and Magnan, J. F., "Lung nodule malignancy classification using only radiologist-quantified image features as inputs to statistical learning algorithms: probing the Lung Image Database Consortium dataset with two statistical learning methods," *Journal of Medical Imaging* **3**, 044504 (Dec. 2016).
- [21] "Lung image database consortium – reader annotation and markup – annotation and markup issues/comments." <https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI> (accessed Dec. 2015).
- [22] Friedman, J., Hastie, T., and Tibshirani, R., [*The elements of statistical learning*], vol. 1, Springer series in statistics Springer, Berlin (2001).
- [23] Armato, S. G., McNitt-Gray, M. F., Reeves, A. P., Meyer, C. R., McLennan, G., Aberle, D. R., Kazerooni, E. A., MacMahon, H., van Beek, E. J., Yankelevitz, D., Hoffman, E. A., Henschke, C. I., Roberts, R. Y., Brown, M. S., Engelmann, R. M., Pais, R. C., Piker, C. W., Qing, D., Kocherginsky, M., Croft, B. Y., and Clarke, L. P., "The Lung Image Database Consortium (LIDC): An Evaluation of Radiologist Variability in the Identification of Lung Nodules on CT Scans," *Academic Radiology* **14**, 1409–1421 (Nov. 2007).
- [24] Winer-Muram, H. T., "The Solitary Pulmonary Nodule," *Radiology* **239**, 34–49 (Apr. 2006).
- [25] Firmino, M., Angelo, G., Morais, H., Dantas, M. R., and Valentim, R., "Computer-aided detection (CADE) and diagnosis (CADx) system for lung cancer with likelihood of malignancy," *BioMedical Engineering On-Line* **15** (Dec. 2016).
- [26] Nyflot, M. J., Yang, F., Byrd, D., Bowen, S. R., Sandison, G. A., and Kinahan, P. E., "Quantitative radiomics: impact of stochastic effects on textural feature analysis implies the need for standards," *Journal of Medical Imaging* **2**(4), 041002–041002 (2015).
- [27] "RadLex." <http://www.rsna.org/RadLex.aspx> (accessed Feb. 2016).