Characterization of pulmonary nodules: Effects of size and feature type on reported performance

Artit C. Jirapatnakul[a], Anthony P. Reeves[a], Tatiyana V. Apanasovich[b], Alberto M. Biancardi[a], David F. Yankelevitz[c], Claudia I. Henschke[c]

[a]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY;
[b]School of Operations Research and Information Engineering, Cornell University, Ithaca, NY;
[c]Department of Radiology, Weill Medical College of Cornell University, New York, NY

ABSTRACT

Differences in the size distribution of malignant and benign pulmonary nodules in databases used for training and testing characterization systems have a significant impact on the measured performance. The magnitude of this effect and methods to provide more relevant performance results are explored in this paper. Two- and three-dimensional features, both including and excluding size, and two classifiers, logistic regression and distance-weighted nearest-neighbors (dwNN), were evaluated on a database of 178 pulmonary nodules. For the full database, the area under the ROC curve (AUC) of the logistic regression classifier for 2D features with and without size was 0.721 and 0.614 respectively, and for 3D features with and without size, 0.773 and 0.737 respectively. In comparison, the performance using a simple size-threshold classifier was 0.675. In the second part of the study, the performance was measured on a subset of 46 nodules from the entire subset selected to have a similar size-distribution of malignant and benign nodules. For this subset, performance of the size-threshold was 0.504. For logistic regression, the performance for 2D, with and without size, were 0.578 and 0.478, and for 3D, with and without size, 0.671 and 0.767. Over all the databases, logistic regression exhibited better performance using 3D features than 2D features. This study suggests that in systems for nodule classification, size is responsible for a large part of the reported performance. To address this, system performance should be reported with respect to the performance of a size-threshold classifier.

Keywords: pulmonary nodule characterization, size distribution, classification and classifier design, X-ray CT

1. INTRODUCTION

Lung cancer often presents as a pulmonary nodule in its earliest manifestation. Early detection and diagnosis of such nodules may lead to improved patient care. While advances in CT scanner technology have enabled early detection of these nodules, diagnosis typically requires waiting for several months to obtain a follow-up scan, possibly delaying treatment and exposing the patient to additional radiation. Automated nodule characterization systems promise to enable diagnosis of suspicious lesions from a single CT scan. These systems take as input a set of features, train a classifier to achieve optimal performance using some metric on a subset of data known as the training set, and report generalized performance on a separate testing set. The expectation is that the classification of an unknown case will exhibit similar performance as the performance on the testing set. Many studies have experimented with the use of different kinds of features and classifiers; one method by Suzuki et al[1] utilized pixel values in a local window on a region of interest in a CT image in conjunction with a massively trained artificial neural network to distinguish between malignant and benign nodules. The researchers reported a sensitivity of 1.00 and a specificity of 0.48, with an area under the ROC curve (AUC) of 0.88. A second study by Aoyama et al[2] also used neural networks, but utilized 41 features extracted from regions of interest containing a nodule. The effective diameter of the nodule was included among the features; the authors reported an AUC of 0.85 using data from multiple slices. Aside from neural networks, other popular classifiers include logistic regression and linear discriminant analysis (LDA). Shah et al[3] extracted several two-dimensional features, including size-based features, and tested several classifiers, including a LDA classifier, a

Send correspondence to Artit C. Jirapatnakul, e-mail: acj29@cornell.edu, phone: 1 607 255 0963

logistic regression classifier, a decision tree, and quadratic discriminant analysis. Using LDA, they achieved an area under the ROC curve of 0.92. The authors did another study[4] using a different dataset and a different set of features. Some of the features were based on the entire volumetric region of interest as opposed to a single slice, as in their previous paper. The authors again achieved an area under the ROC curve of 0.92 using a logistic regression classifier.

While many studies have achieved good performance of an AUC of 0.85 or above, several issues which may affect performance have yet to be properly addressed. A primary concern is the impact of the distribution of the sizes of malignant and benign nodules. Nearly all datasets used in the training and testing of characterization systems to date exhibit bias in the size distribution of malignant and benign nodules. This *a priori* size information is a powerful feature,[5,6] but in the evaluation of an automated characterization system, the relevant performance is not the absolute performance, but whether, and by how much, performance is improved over the use of this *a priori* size information.

There has been an increasing trend towards the use of 3D features over 2D features, but few studies have directly compared them on the same dataset. Finally, studies tend to use either parametric classifiers, such as logistic regression, or non-parametric classifiers, such as neural networks; few studies have compared the performance of both types of classifiers on the same dataset. This study assesses the effect of unbalanced dataset size distribution on the performance of automated systems for the differentiation between malignant and benign nodules. Both 2D and 3D feature sets are evaluated with respect to two types of classifiers, a parametric (logistic regression) and non-parametric (distance-weighted nearest neighbors) classifier.

2. METHODS AND MATERIALS

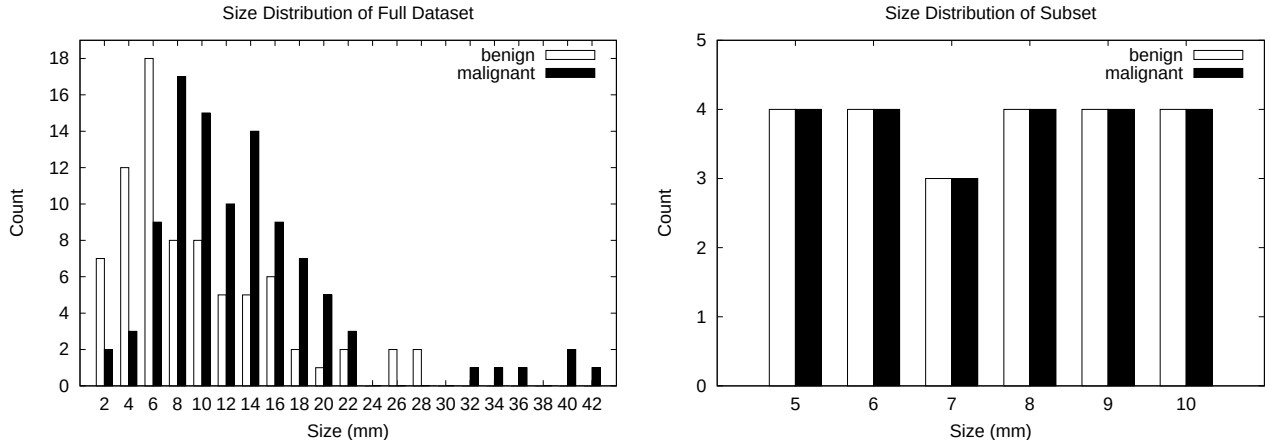
The pulmonary nodule characterization system used in this study was based on the system reported by Jirapatnakul et al.[7] The system will be briefly described, followed by an explanation of the data and experiment methodology.

2.1 Nodule Characterization System

The characterization system is divided into three main sections: segmentation, feature extraction, and classification. Segmentation was performed using an algorithm developed by Reeves et al.[8] Based on a manually specified seed point, the segmentation algorithm estimated the size and center of the nodule, resampled the image into isotropic space, and performed morphological filtering and attached structure removal. Each segmentation was verified visually. The segmentation step resulted in a binary image indicating which pixels were part of the nodule, a 3D model of the nodule, and the grayscale region of interest containing the nodule.

These images were used in the feature extraction step, where 2D and 3D morphological, shape, and CT features were extracted. Two-dimensional features were computed on the slice of the scan which contained the center of mass of each nodule. In contrast to the previous study,[7] not only were both 2D and 3D features used, but features that were dependent on size were also included. Although size-dependent features should not be used, they were included because many other studies make use of these features in their characterization systems.

Once features are extracted for each nodule, the nodules are classified using two different classifiers. One classifier, logistic regression, is a parametric method often used for medical applications. Stepwise feature selection using the Akaike information criterion (AIC) was performed using methods built into the R statistical package. The AIC rewards performance but penalizes additional parameters which should help to prevent overfitting. The second classifier, distance-weighted nearest neighbors (dwNN), was used to assess whether non-parametric methods would have the same behavior as parametric methods. In this implementation of dwNN, each feature is weighted by its information gain ratio, which is a measure of the reduction of entropy of the model contributed by each feature. The classifier result is the weighted average of the class of each neighbor, with weighting done by the distance of each neighbor to the nodule under consideration. To generate an ROC curve, varying thresholds are applied to the classifier result. For both classifiers, a leave-one-out training and testing methodology was used.



(a) (b)

Figure 1. Size distribution of a) full dataset and b) subset of dataset selected to have equal size distribution of malignant and benign nodules. For each graph, the number on the x-axis represents the maximum size of the bin.

2.2 Data

The full dataset was comprised of 178 nodules, with 100 malignant and 78 benign nodules. Nodules were selected from the Weill Cornell Medical Center database of solid or part-solid nodules, whose consistency was determined by a radiologist, and both attached and isolated nodules were included. Part-solid nodules were only included if they had a substantial solid component. Malignant nodules had a diagnosis confirmed through biopsy or resection while benign nodules were either biopsied or listed as having 2 years of no clinical change. Metastatic nodules and benign calcifications were excluded from the dataset. Scans were obtained using either GE Medical Systems HiSpeed CT/i, Genesis HiSpeed, LightSpeed QX/i, or LightSpeed Ultra CT scanners using either 1.0 mm, 1.25 mm, 2.5 mm, or 5.0 mm slice thickness.

A subset of nodules in the full dataset were selected to form a dataset of malignant and benign nodules with similar size distributions. Nodules between 4.0 mm and 10.0 mm, the range where most of the malignant and benign nodules overlapped, were selected. These nodules were further pruned to ensure an equal number of malignant and benign nodules in each of the 6 1.0 mm bins; where possible, malignant and benign nodules were chosen that were similar in size. A total of 46 nodules (23 malignant, 23 benign) were selected that fulfilled these criteria. A comparison of the distribution of nodule sizes in the full dataset and the subset of the dataset is shown in Figure 1. Based on the results of the two-tailed t-test, the full dataset had significant difference in size between malignant and benign nodules (P=0.001) , while the subset of nodules had no significant difference in size (P=0.9283).

2.3 Experiment

The focus of this work is not to report on the absolute performance of the characterization system, but to assess the effect of the underlying size distribution of the nodules in the dataset on performance of characterization systems. Preliminary work has suggested that the reported performance of a characterization system is dependent upon the difference in size-distribution between malignant and benign nodules.[9] To attempt to measure this, two sets of data are used: one set comprised of all the nodules in the dataset, and a subset of the data selected to reduce the size bias between malignant and benign nodules, described in the section above. For each dataset, the performance of characterization systems using 2D and 3D features with and without size was determined for both classifiers, resulting in four systems for each dataset and classifier type, for a total of sixteen different performance results. ROC curves were generated for each system by varying the threshold for classification on the system output.

As size may have an effect on the performance of the characterization system, a simple size threshold was applied to each dataset to establish a baseline performance result. Nodules below the size threshold were classified

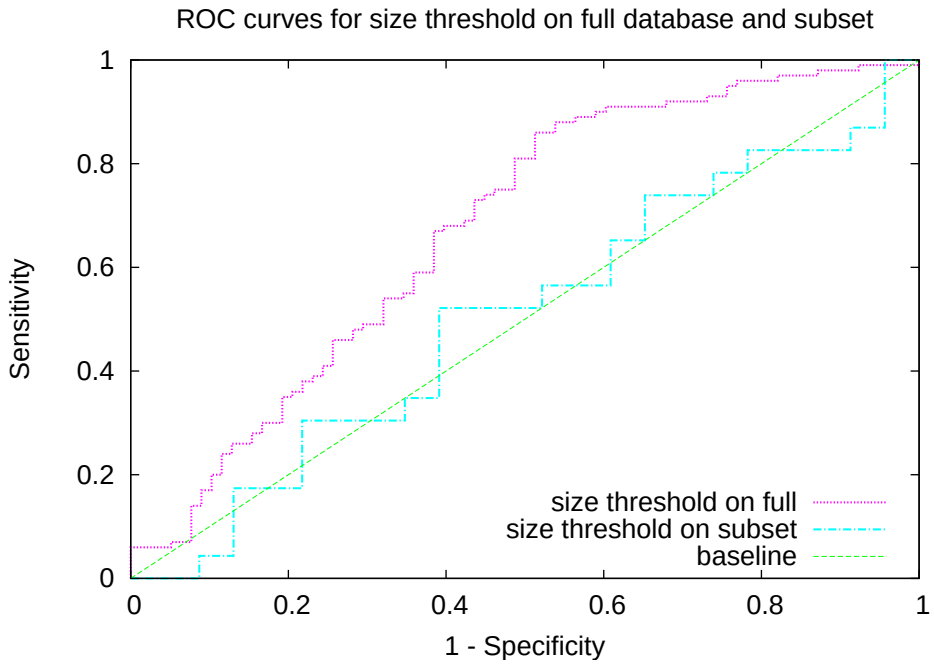


Figure 2. ROC curves of simple size threshold classifier on full dataset and subset of dataset compared to baseline performance. Note that on the full dataset, the performance of size is higher than baseline, while on the subset of nodules with similar size distribution of malignant and benign nodules, the performance of size is nearly baseline.

as benign, while nodules equal to, or above the threshold were classified as malignant. The size threshold was varied across the entire range of nodule sizes in the dataset to generate an ROC curve. Nodule sizes were computed using the product of the maximum length and the perpendicular (pseudo-WHO), on the central slice through the nodule.

3. RESULTS

3.1 Full dataset

In the full dataset, the performance from using the simple size threshold was a sensitivity of 73% with a specificity of 55% with an area under the ROC curve (AUC) of 0.675. The performance of the size threshold classifier is shown in Figure 2 with the performance of the size threshold on the subset of nodules and conventional baseline measure with an AUC of 0.50.

The results for the characterization system using logistic regression, dwNN, and the size threshold classifier are summarized in Table 1. ROC curves for the logistic regression classifier are shown in Figure 3 for both 2D and 3D features. The performance of the size threshold classifier and normal baseline are shown on the plots for reference. For the 2D features, the AUC was 0.721 for the classifier with size features and 0.614 without size, while for the 3D features, the AUC was 0.773 and 0.737 respectively. The dwNN classifier results are shown in Figure 4; the performance was, for the 2D features, AUC of 0.709 and 0.631 for features with size and without size respectively. For 3D features, the AUC was 0.728 and 0.711 for features with size and without size respectively.

Table 1. Summary of performance (AUC) for full dataset for all classifiers and feature combinations

	2D features		3D features	
Classifier	size	no size	size	no size
logistic regression	0.721	0.614	0.773	0.737
distance-weighted nearest neighbor	0.709	0.631	0.728	0.711
size threshold	0.675			

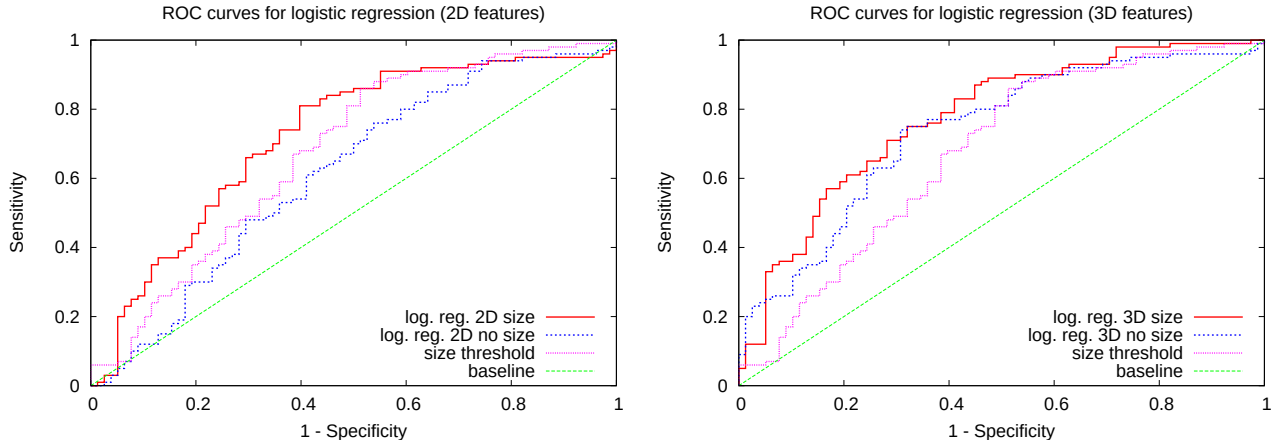
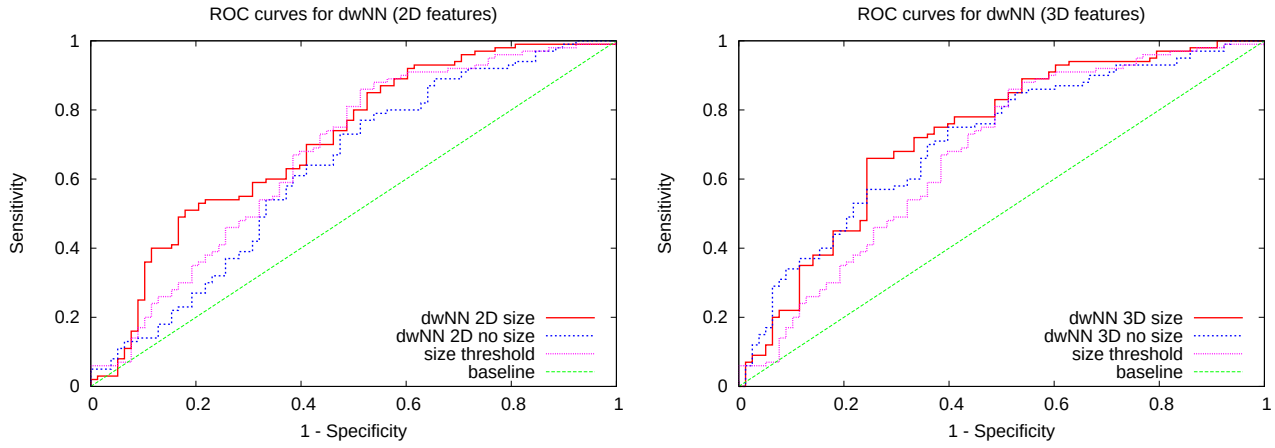


Figure 3. ROC curves of characterization system using logistic regression with a) 2D and b) 3D features on the full dataset. The performance of the size-threshold classifier and conventional baseline are included for reference.



(a) (b)

Figure 4. ROC curves of system using dwNN classifier with a) 2D and b) 3D features on the full dataset

3.2 Subset dataset

On the dataset where the distribution of nodules was balanced in terms of size, the baseline performance was near the conventional baseline on the ROC curve. This corresponded to an AUC of 0.504, with a sensitivity/specificity of 42%/47% with the ROC curve shown in Figure 2. For the logistic regression classifier, different features were selected for this subset based on AIC as compared to the full set of data; for 2D features with size, two features

Table 2. Summary of performance (AUC) for reduced dataset for all classifiers and feature combinations

	2D features		3D features	
Classifier	size	no size	size	no size
logistic regression	0.578	0.478	0.671	0.767
distance-weighted k-NN	0.684	0.616	0.667	0.507
size threshold	0.504			

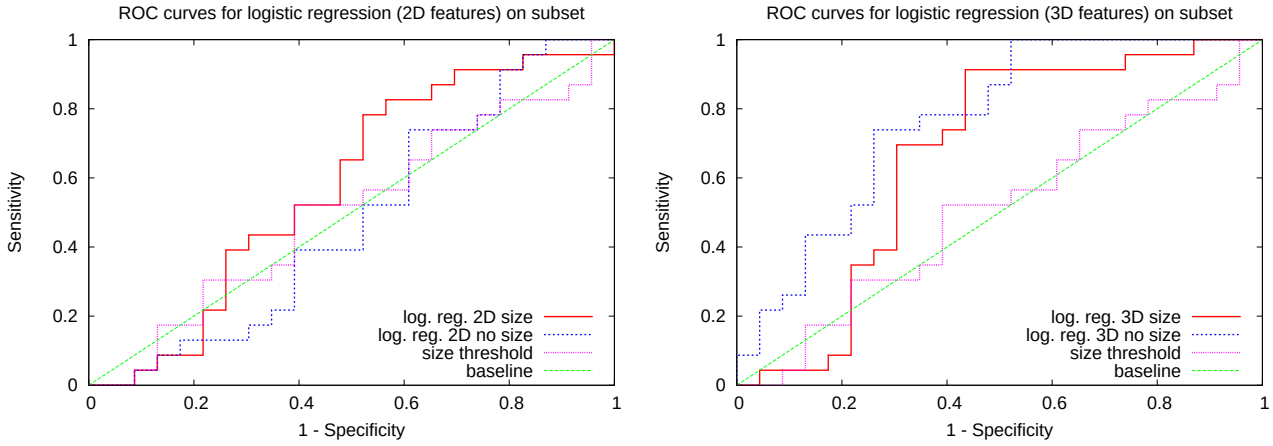
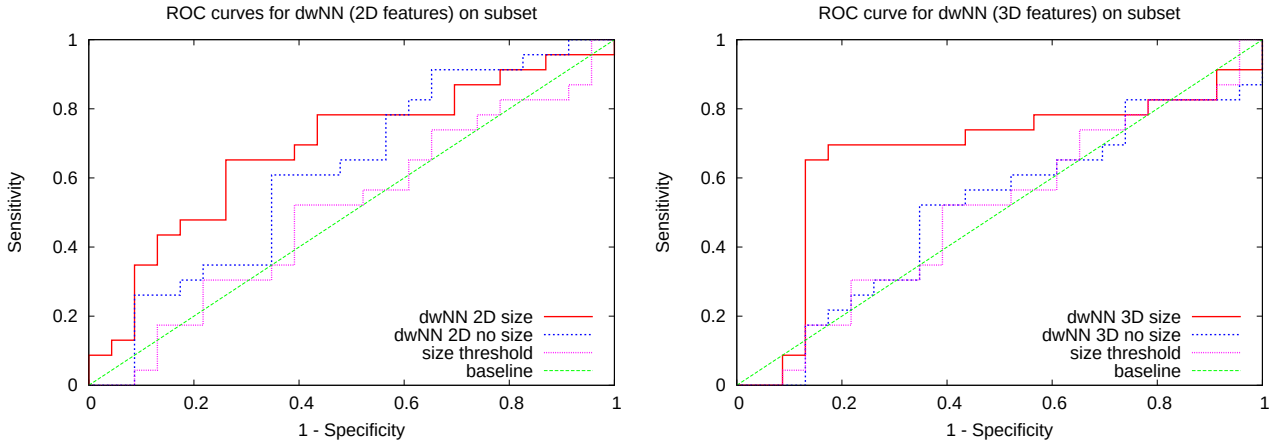


Figure 5. ROC curves for logistic regression on subset of nodules using a) 2D and b) 3D features



(a) (b)

Figure 6. ROC curves for distance-weighted nearest neighbor on subset of nodules using a) 2D and b) 3D features

were selected, but only one when size was excluded. For the 3D features, five non-size features were selected; although size features were provided in the feature selection step, none were selected. To enable comparison with a set of size-based features, the same set of features used on the full dataset was used. ROC curves for both 2D and 3D features are shown in Figure 5a and 5b respectively. The 2D features do very little for the performance, with the classifier including size doing marginally better than baseline while the classifier excluding size performs worse; the AUC were 0.578 and 0.478 respectively. For the set of 3D features, the AUC was 0.671 with size and 0.767 without size. The dwNN classifier had an opposite trend, with 2D features with and without size (AUC of 0.684 and 0.616 respectively) performing better than with 3D features with and without size (AUC of 0.667 and 0.507 respectively). The performance for all classifiers on this subset of nodules is summarized in Table 2.

4. DISCUSSION

Studies have shown that the size of a lesion is a good predictor of malignancy.[5,6] However, the use of size as a feature in characterization systems has several caveats:

1. The size measurement has a very large range; a 3 mm to 30 mm range in lesion diameter corresponds to a volume range of 1000 to 1.
2. Many of the the other features are dependent on the size of the nodule; for example, average curvature.
3. Given 1 above, the size is related to the accuracy and detail that other feature measurements can be made on a nodule. As an example, given a typical voxel size for a CT scan of 0.7 x 0.7 x 1.25 mm, a 3 mm nodule has a volume equivalent to about 23 voxels; given partial volume effects, noise, etc., this is inadequate to provide meaningful values for some of the complex shape-based features.
4. For a dataset with a biased size distribution of malignant and benign nodules, the size (or a size derived feature) is always the most useful feature.
5. In all published datasets used for training and evaluating nodules for which the size distribution is given, there is a difference in the size of benign and malignant nodules in which small benign and large malignant nodules predominate. This skewness in the distribution of the dataset reflects the natural history of lesions found in lung scans; however, the actual distribution is very sensitive to the population subset from which the data was acquired; e.g. screening scans would be expected to have a different distribution compared to clinical scans.

Therefore, in any pulmonary nodule dataset, there is an intrinsic classification performance that can be achieved by use of a size feature alone that is dataset-specific. In general we are interested in a system performance evaluation that is not highly dependent on a population feature of a particular dataset. With this in mind, many ROC results that have been published in the literature look very promising but are actually largely characterizing the size skewness in the training dataset.

To assess the impact that biases in the size distribution of nodules may have on performance, this study used two datasets with different size distributions. The full dataset of 178 nodules reflects nodule sizes more typical of characterization studies. Forty-six nodules were selected for the smaller dataset from the full dataset so that the size distributions would be as similar as possible. On the full dataset, the simple size-threshold classifier achieved an AUC of 0.675, showing improvement over baseline AUC of 0.50. This suggests that the distribution of malignant and benign nodules in our dataset are more similar than other datasets with a higher sensitivity and specificity from size (such as those analyzed by Jirapatnakul et al[9]). This reduced bias makes this a more challenging dataset to characterize than most others reported in the literature. The best performance on the full dataset was an AUC of 0.774 achieved by logistic regression with 3D features including size, which is a large improvement over the baseline performance, but a smaller improvement compared to the size threshold AUC. Additionally, in all feature sets on the full dataset, removing size-dependent features reduces performance. This suggests that size is responsible for a portion of the reported performance of all classifiers on the full dataset. On the smaller dataset, the size threshold classifier achieved an AUC of 0.504, which is near baseline performance, as expected from a dataset with an equal distribution of sizes of malignant and benign nodules. Accordingly, performance of the logistic regression classifier that included size features was reduced compared to the full dataset; as one example, consider that the characterization system, using the logistic regression classifier with 2D features that included size, exhibited a reduction in AUC from 0.721 on the full dataset to 0.578 on the subset of nodules, despite similar levels of optimization performed for both datasets. The logistic regression classifier which used 3D features but excluded size-dependent features had the best performance, nearly similar to the results on the full dataset. However, for the dwNN classifier, the feature sets that included size both performed better than without. This may be due to the small number of cases in the subset of the data. The dwNN classifier performed worse on the subset of nodules than the full dataset, suggesting the size distribution affects the performance of the dwNN classifier as well.

Aside from the issue of size, this study considered the performance of 2D and 3D features, both with and without size-dependent features. On the full dataset, both logistic regression and dwNN classifiers performed better with 3D features compared to the 2D features, suggesting that 3D features are more effective. This is likely due to the additional data offered by the use of additional slices for the 3D features. Both classifiers also achieved higher performance when size-dependent features were included, which is reasonable considering the effectiveness of size for discriminating nodules in this dataset. On the subset of cases with an equal size distribution, for logistic regression, 3D features were again more effective than 2D features, with the best performance offered by 3D features without using size-dependent features. For the 2D features, the system using size-dependent features performed better, but this may be do to the lack of power of the non-size-dependent features. Results for the dwNN classifier differed somewhat; the set of 2D features including size achieved the best performance, with 2D performing better than 3D. The dwNN classifier us likely harmed by the fact that, for the set of 3D features, there are nearly as many features as nodules in the dataset, decreasing the effectiveness of information gain weighting of the features.

Finally, the comparison between logistic regression and dwNN was inconclusive. On the full dataset, logistic regression using 3D features with size achieved the best performance with an AUC of 0.773, while the best performance using the dwNN classifier was 0.728 for the same set of features. The logistic regression classifier performed better than the dwNN classifier for three out of the four sets of features. However, the performance difference ranged from 0.012 to 0.045, which is a small amount compared to the difference between the feature sets. Results were also mixed on the subset of nodules, with neither classifier offering a consistent advantage over the other.

5. CONCLUSION

Pulmonary nodule characterization systems are all trained on datasets with a bias in the size distribution of malignant and benign nodules. This bias leads to size being a very predictive feature. In the dataset used in this study, size alone is responsible for a significant improvement (AUC 0.675) over the conventional baseline (AUC of 0.50). However, the performance from size is available with a simple threshold without the use of a complex automated characterization system; therefore, the relevant measure of a characterization system is the performance above and beyond the performance from the *a priori* size information.

This study showed that size had a positive effect on the performance of a system whose dataset had a bias in the size distributions of malignant and benign nodules. Performance of the system was reduced when trained and tested on a dataset where the bias in the size distributions was reduced. Comparing 2D and 3D features, on the full dataset in this study, 3D features were more effective than 2D features, with classifiers using 3D features performing better than baseline, with or without size. Of the classifiers using 2D features, only those with size performed better than baseline. The improved performance of 3D features can likely be attributed having additional information contained in the additional slices. This result suggests that 3D features are likely to be superior for nodule characterization, however additional testing is required on larger datasets of similar size distributions of malignant and benign nodules.

As classification performance is heavily dependent upon the underlying size-distribution of the training and testing datasets, measurement of system performance should take into account the skewness of the size distributions of the dataset. A more reliable way of reporting the system performance is as the improvement with respect to just a size feature classifier, as opposed to the conventional random chance.

REFERENCES

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

[2] M. Aoyama, Q. Li, S. Katsuragawa, F. Li, S. Sone, and K. Doi, "Computerized scheme for determination of the likelihood measure of malignancy for pulmonary nodules on low-dose ct images," *Medical Physics* **30**, pp. 387–394, March 2003.

[3] S. K. Shah, M. F. McNitt-Gray, S. R. Rogers, J. G. Goldin, R. D. Suh, J. W. Sayre, I. Petkovska, H. J. Kim, and D. R. Aberle, "Computer-aided diagnosis of the solitary pulmonary nodule," *Academic Radiology* **12**, pp. 570–575, May 2005.

[4] S. K. Shah, M. F. McNitt-Gray, S. R. Rogers, J. G. Goldin, R. D. Suh, J. W. Sayre, I. Petkovska, H. J. Kim, and D. R. Aberle, "Computer aided characterization of the solitary pulmonary nodule using volumetric and contrast enhancment features," *Academic Radiology* **12**, pp. 1310–1319, October 2005.

[5] E. A. Zerhouni, F. P. Stitik, S. S. Siegelman, D. P. Naidich, S. S. Sagel, A. V. Proto, J. R. Muhm, J. W. Walsh, C. R. Martinez, and R. T. Heelan, "CT of the pulmonary nodule: a cooperative study," *Radiology* **160**(2), pp. 319–327, 1986.

[6] D. M. Libby, J. P. Smith, N. K. Altorki, M. W. Pasmantier, D. Yankelevitz, and C. I. Henschke, "Managing the Small Pulmonary Nodule Discovered by CT," *Chest* **125**(4), pp. 1522–1529, 2004.

[7] A. C. Jirapatnakul, A. P. Reeves, T. V. Apanasovich, M. D. Cham, D. F. Yankelevitz, and C. I. Henschke, "Characterization of solid pulmonary nodules using three-dimensional features," *SPIE International Symposium on Medical Imaging 2007* **6514**, p. 65143E, Feb. 2007.

[8] W. J. Kostis, A. P. Reeves, D. F. Yankelevitz, and C. I. Henschke, "Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical ct images," *IEEE Transactions on Medical Imaging* **22**, pp. 1259–1274, Oct. 2003.

[9] A. Jirapatnakul, A. Reeves, T. Apanasovich, A. Biancardi, D. Yankelevitz, and C. Henschke, "Pulmonary nodule classification: Size distribution issues," in *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1248–1251, 2007.