PULMONARY NODULE CLASSIFICATION: SIZE DISTRIBUTION ISSUES

A.C. Jirapatnakul[a], A.P. Reeves[a], T.V. Apanasovich[b], A.M. Biancardi[a], D.F. Yankelevitz[c], and C.I. Henschke[c]

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

ABSTRACT

Automated nodule classification systems determine a model based on features extracted from documented databases of nodules. These databases cover a large size range and have an unequal distribution of malignant and benign nodules, leading to a high correlation between malignancy and size. For two recent studies in the literature, much of the reported performance of the system may be derived from size based on analysis of their size distributions. We performed experiments to determine the effect of unequal size distribution on a nodule classification system's performance. Preliminary results indicate that the performance across the entire dataset (a sensitivity/specificity of 0.85/0.80) does not generalize to a subset of nodules (0.50/0.80), but performance can be improved by specifically training on that subset (0.60/0.80). Additional testing with larger datasets needs to be performed, but results reported in this area are overly optimistic.

Index terms–Medical diagnosis, biomedical image processing, pulmonary nodule characterization, size distribution

1. INTRODUCTION

The introduction of high-resolution CT scans has allowed radiologists to detect more small lesions than previously possible with either chest radiographs or thick-slice CT. The status of these nodules is often difficult to ascertain, requiring follow up work. Currently, many protocols rely on the assessment of growth rate based on CT scans followed by biopsy to determine if a nodule is malignant. However, growth rate assessment requires a second CT scan which prolongs diagnosis and exposes the patient to a second, possibly unnecessary dose of radiation. To address these issues, automated methods of nodule classification have been developed that differentiate malignant from benign nodules based on features extracted from a single CT scan.

Current approaches to nodule classification exhibit three problems. First, these methods require a large database of documented cases of both malignant and benign nodules, but, in practice, it is difficult to obtain a sufficiently large number of nodules. Second, there is a large size range of nodules in these databases (over 1000 to 1 by volume), where nodule size is generally expressed as the "average" diameter of the lesion in one dimension. Third, for databases constructed from a given population, malignancy is highly correlated to size, with benign nodules dominating the small size range and cancers dominating the large size range. Due to these three factors, automated methods tend to make extensive use of size information, which may not be useful when predicting whether a nodule in an intermediate size range is malignant and may result in misleading performance.

In this paper, we assess the impact of unequal size distribution of malignant and benign nodules on the performance of an automated nodule classification system through the use of datasets with different size ranges and distributions.

2. METHODS

2.1. Analysis of previous work

To better illustrate the problem of using a database with unequal nodule size distribution, we consider two datasets recently reported upon in the literature. In the first example[1], the vast majority of nodules came from a screening study in which the majority of the nodules were small (less than 7 mm) and benign. The database of nodules in this study included 413 benign and 76 malignant nodules ranging in size from 3 mm to 31 mm, with the distribution shown in Figure 1. If we predict malignancy based solely on size by using a criterion of "all nodules greater than 7 mm are malignant", we would achieve a sensitivity and specificity (SS) of (0.80, 0.80). This performance is very similar to that shown on the ROC curve for the trained computer method of Suzuki et al[1], although numerically they reported a SS of (1.00, 0.48). In a dataset used by Shah et al [2], the nodule sizes are much larger, suggesting that the nodules were taken from a clinical population. Their dataset consisted of 33 benign and 48 malignant nodules with the size distribution in Figure 2. The sizes in their size distribution histogram represent the maximum diameter of the nodule. On this database, a size criterion of "all nodules greater than 15 mm are malignant" would achieve a SS of (0.64, 0.79). In their semi-automated method, features were extracted from regions of interest demarcated by boundaries around the nodule created manually on a single slice by a radiologist. In this case, the authors achieved a better re-

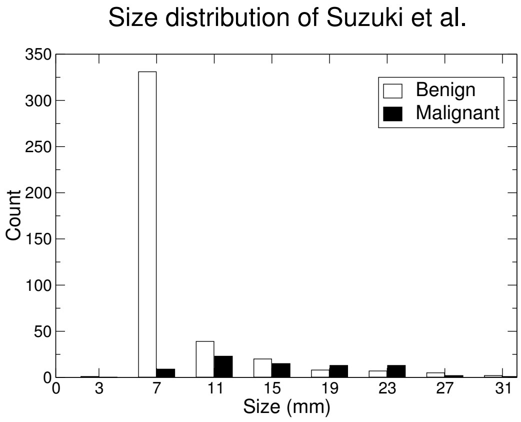

Fig. 1. Histogram of size distribution of dataset used by Suzuki et al.[1]

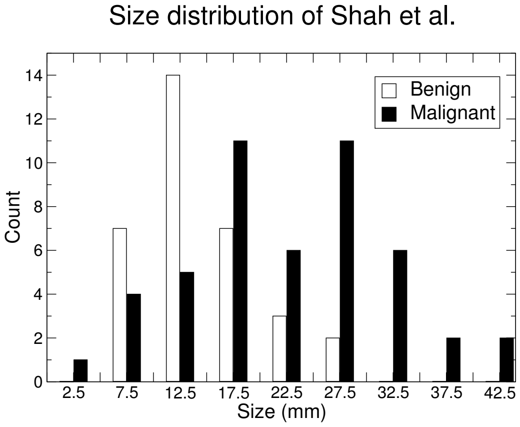

Fig. 2. Histogram of size distribution of dataset used by Shah et al.[2]

sult using their semi-automated method, a SS of (0.90, 0.80) based on their ROC curve, than size alone. As both papers used size-dependent features in their algorithms, each system likely placed a high emphasis on the importance of size distribution for partitioning the dataset. The use of size information is not desirable because it fails to generalize well; as an example, if the size criterion determined from the dataset of Shah et al is used on the database of Suzuki et al, a SS of (0.38, 0.95) would be achieved, and in the reverse case, using the size criterion determined from the dataset of Suzuki et al of 10 mm (the closest interval to 7 mm), a SS of (0.21, 0.90) would be achieved. To date, no studies have analyzed the effect of size distribution on the performance of a classification system.

2.2. Dataset

For this study, a total of 48 malignant and 55 benign solid nodules on whole lung and targeted CT scans were selected from a screening database. Scans were acquired using either GE Medical Systems HiSpeed CT/i, LightSpeed Ultra, LightSpeed QX/i, or Genesis HiSpeed scanners with either 1.0 mm or 1.25 mm slice thickness. Benign nodules were diagnosed by biopsy, histology of resected tissue, or by no clinical change in 2 years. Malignant nodules were diagnosed by biopsy or histology of resected tissue. Nodule sizes ranged from 1.9 mm to 32.2 mm, with the size distribution shown in Figure 3. From the entire dataset, 10 malignant and 10 benign nodules were selected for a size-controlled dataset. Nodules were manually selected to achieve a similar size distribution of the malignant and benign nodules. Only sizes with at least two malignant and benign nodules were used to ensure that the distributions within each size bin were as similar as possible. The nodules ranged in size from 5.13 mm to 8.39 mm, with the size distribution shown in Figure 4.

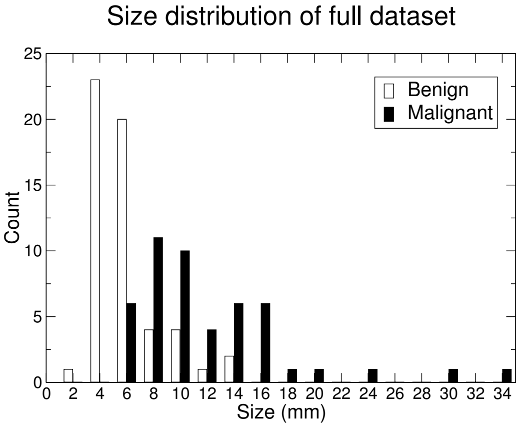

Fig. 3. Histogram of size distribution of entire dataset

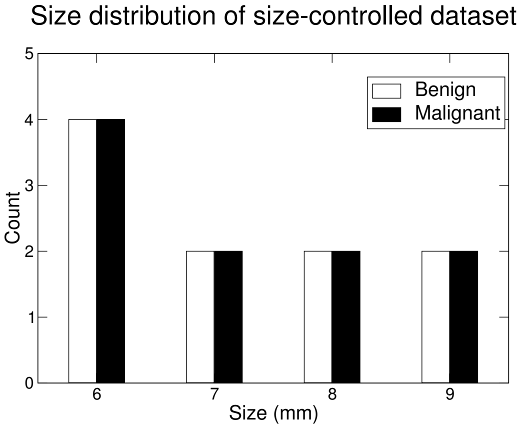

Fig. 4. Histogram of size distribution of size-controlled dataset

2.3. Nodule classification system

Our automated system for nodule classification consisted of three stages: nodule segmentation, feature extraction and selection, and classification. In our scheme, nodule segmentation was performed based on a seed point selected from reports by a radiologist using the algorithm described in [3]. In brief, this algorithm determines the location and size of the nodule, and then performs thresholding, morphological filtering, and juxtapleural segmentation to isolate the nodule from other structures, resulting in a binary segmented image. Using this image as a mask on the original image data, 2D and 3D features are computed based on geometric and densitometric moments. Surface shape information is extracted from a 3D reconstruction of the binary segmented image. The process of computing these features is described further in [4, 5]. Examples of features include compactness, sphericity, x-y extent ratio, and curvature. Features that were obviously size-dependent, such as size and volume, were eliminated from consideration.

Past studies have used a wide variety of classifiers for the task of nodule classification. These can be divided into two classes, parametric and non-parametric. Parametric classifiers are characterized by the specification of a model a priori and include methods such as logistic regression. Non-parametric classifiers determine a model from the data and include techniques such as artificial neural networks and k-nearest-neighbors. This study tested two classifiers, logistic regression and k-NN, to ascertain which type of algorithm would be better on this problem and whether both types of classifiers would be similarly affected by an unequal size distribution. Logistic regression was chosen because it is often used in medical classification tasks due to its statistical foundation and availability of methods and tools to interpret its results. k-NN was selected due to its simplicity and adaptability to irregular feature spaces. Neural networks have also been used for nodule classification and are likely to have similar performance to k-NN for this problem. The k-NN algorithm used Euclidean distance in feature space as a similarity measure. Features were normalized to have a standard deviation of 1. Forward stepwise feature selection was used to select the best set of features for the k-NN algorithm while stepwise feature selection was used for the logistic regression algorithm.

2.4. Experiment design

In order to determine the impact of the nodule size distribution of the training dataset on our nodule classification system, a series of three experiments were performed. For all experiments, leave-one-out cross validation was used to test the algorithms, with feature selection (training) performed on either the entire dataset or the size-controlled dataset.

In the first experiment, A, the algorithm was trained and tested using nodules from the entire dataset. This type of analysis is typical of studies reported in the literature. In the second experiment, B, the system was trained using the entire dataset, but results were reported only on the nodules on the size-controlled dataset. This would reflect the performance of the system for a nodule within the size range of the size-controlled dataset. In the last experiment, C, the system was trained and tested using the size-controlled dataset. This experiment represents the performance of a system tuned specifically for the size range of the size-controlled dataset, without the benefit of having useful size information. If the nodule size distribution is not a factor in the performance of the classification system, the results of these three experiments should be similar.

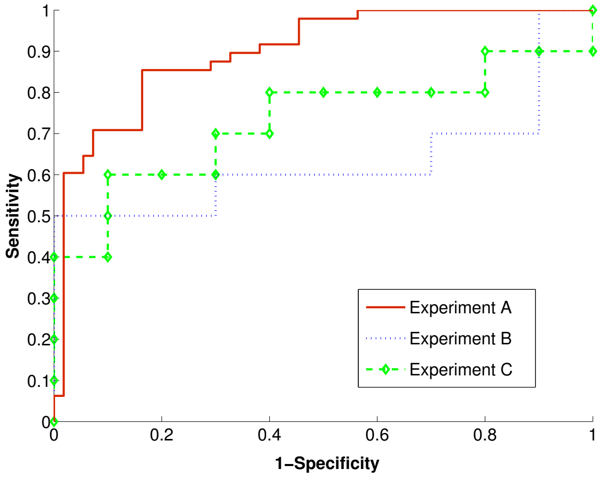

Fig. 5. ROC curves of the logistic regression classifier for experiments A (training and testing on the entire dataset) , B (training on the entire dataset, but testing only on the size-controlled dataset), and C (training and testing on the size-controlled dataset). The best performance is on experiment A, while the worst performance is obtained on experiment B.

3. RESULTS

Using the full dataset of 48 malignant and 48 benign nodules in experiment A, the classification system achieved good performance using either of the classifiers, with a sensitivity and specificity (SS) of 0.85 and 0.80 for the logistic regression classifier, and a SS of 0.75 and 0.80 for the k-NN classifier. The ROC curve for the logistic regression classifier is indicated by the solid line in Figure 5. In experiment B, we see that the performance drops significantly, with a SS of 0.50 and 0.80 for logistic regression, and a corresponding shift to the right (dotted line) of the ROC curve, and a SS of 0.50 and 0.70 for the k-NN classifier. The nodules of the size-controlled dataset were included in feature selection and testing (through leave-one-out) of the model, yet the per-

Table 1. Performance of k-NN and logistic regression classifiers. LR = Logistic Regression

	Experiment A		Experiment B		Experiment C	
	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
LR	0.85	0.80	0.50	0.80	0.60	0.80
k-NN	0.75	0.80	0.50	0.70	0.80	0.80

formance is significantly worse for this subset of nodules than the entire set of nodules. To ascertain whether this drop in performance could be rectified, new models for both classifiers were trained on just the size-controlled dataset in experiment C. This improved performance for both classifiers, shifting the ROC curve to the left (dashed line) slightly for the logistic regression classifier, with a SS of 0.60 and 0.80. The k-NN classifier achieved slightly better performance in experiment C than in experiment A, with a SS of 0.80 and 0.80. Both classifiers showed an improvement in performance compared to experiment B, suggesting that a different set of features may be necessary when analyzing nodules in a limited size range. The sensitivity and specificity for both classification algorithms are summarized in Table 1.

4. DISCUSSION

Performance results from systems trained on datasets containing different size distributions will likely produce misleading results for realistic situations in which the size of the lesion being considered is known. Analysis of recent publications show that much of the claimed performance could be achieved by thresholding with size. The influence of the nodule size distribution of the training/testing dataset on performance is reinforced by the differences in performance of our classification system across different experiments. Due to the natural distribution of nodule sizes, small nodules have a greater probability of being benign, while large nodules are more likely to be cancer. Including size information improves the performance of any system due to correctly classifying very large or small nodules, but the performance improvement is not constant across a large size range, as is shown by the decrease in performance in experiment B. This drop in performance is not limited to one type of classifier; both the k-NN and logistic regression classifiers had similar reductions in performance.

One method to address this issue is limiting the training set to a small size range with a similar distribution of malignant and benign nodules, as we have done in experiment C. Among all available features, many are size-dependent, but under this condition, size no longer provides a benefit to the classification performance. This enables the feature selection algorithm used by the automated method to choose features that might be less discriminating than size on the entire dataset, but more discriminating on the smaller subset. We tested this hypothesis in experiment C, and found that different features were selected which improved performance on the size-controlled dataset as compared to experiment B. One limitation of our study was the small number of cases used in our dataset. With this small number, the differences between classification methods were not considered to be significant.

5. CONCLUSION

This preliminary work has identified several issues with using datasets with malignant and benign nodules of different size ranges and distributions. While testing was only performed using our classification system with two different classification methods, we expect the effects will be similar on any automated classification system. If no adjustment is made for size distribution, misleading, overly optimistic results are reported, with much of the performance derived from very small benign nodules and very large malignant nodules. Constructing classification systems at a number of size ranges is one method to improve real performance; however, this requires much larger datasets with enriched populations. Additional testing needs to be performed with a larger dataset to determine if compensating for the nodule size distribution is possible.

6. REFERENCES

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

[2] 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*, vol. 12, pp. 570–575, May 2005.

[3] A. Reeves, A. Chan, D. Yankelevitz, C. Henschke, B. Kressler, and W. Kostis, “On measuring the change in size of pulmonary nodules,” *IEEE Transactions on Medical Imaging*, vol. 25, pp. 435–450, April 2006.

[4] W. J. Kostis, *Three-dimensional computed tomographic image analysis for early cancer diagnosis in small pulmonary nodules*. PhD thesis, Cornell University, January 2001.

[5] A. P. Reeves, R. J. Prokop, S. E. Andrews, and F. P. Kuhl, “Three-dimensional shape analysis using moments and fourier descriptors,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, pp. 937–943, November 1988.