Predictive Data Mining for Lung Nodule Interpretation

William Horsthemke, Ekarin Varutbangkul, Daniela Raicu, Jacob Furst
DePaul University, Chicago, IL USA
{whorsthe,evarutba}@students.depaul.edu, {draicu, jfurst}@cs.depaul.edu

Abstract

Diagnostic decision-making in pulmonary medical imaging has been improved by computer-aided diagnosis (CAD) systems, serving as second readers to detect suspicious nodules for diagnosis by a radiologist. Though increasing accurate, these CAD systems rarely offer useful descriptions of the suspected nodule or their decision criteria, mainly due to lack of nodule data.

In this paper, we present a framework for mapping image features to radiologist-defined diagnostic criteria based on the newly available data from the Lung Image Database Consortium (LIDC). Using data mining, we found promising mappings to clinically relevant, human-interpretable nodule characteristics such as malignancy, margin, spiculation, subtlety, and texture. Bridging the semantic gap between computed image features and radiologist defined diagnostic criteria allows CAD systems to offer not only a second opinion but also decision-support criteria usable by radiologists. Presenting transparent decisions will improve the clinical acceptance of CAD.

1. Introduction

Diagnostic decision-making in medical imaging by radiologists has been augmented by computer-aided diagnosis (CAD) systems which extract image features and use data mining techniques to classify or predict a detection or diagnosis. Typically, the CAD system marks the location of a suspicious area, such as a pulmonary nodule, signaling the radiologist to investigate and make the final diagnosis. While beneficial as a tireless and increasingly accurate screening tool, CAD systems rarely offer supporting guidance about their decision rationale or this guidance does not match the perceptual tasks used by the radiologist in forming their diagnosis.

This paper focuses on deriving medical decision-support criteria rather than nodule detection and diagnosis. After extracting low-level image features, decision trees are applied to predict the scoring of image-based diagnostic criteria as interpreted by radiologists. This paper predicts nine (9) nodule characteristics from measurements of sixty-five (65) image features representing shape, size, intensity, and texture information. From our experimental results, we successfully (>75% accuracy) predict five (5) radiologist-defined nodule characteristics: malignancy, margin, spiculation, subtlety, and texture.

2. Background and Related Work

Research in pulmonary nodule detection, segmentation, classification, and diagnosis continues in academic and industrial labs to improve each task of CAD. Increasing the sensitivity of detection remains a primary research focus since detection rates remain about 80% (with 3.8 false positives per patient case), though the CAD rates still exceed the estimated 70% detection rate of radiologists [2]. Efforts to improve detection typically reduce the specificity as demonstrated by [1] which reports 95% sensitivity but 6.9 false positives per case. Addressing the tradeoff between detection of suspicious lung nodule candidates and rejecting non-nodule candidates motivates current research ranging from image segmentation and feature extraction to data mining and knowledge discovery.

Using the data mining techniques of ensemble training and hierarchical neural networks, Suzuki and Dachman [4] attempts to reduce false positives through improved classification. Reducing the detection of non-nodules and improving the segmentation of nodule candidates drives several efforts such as the shape representation introduced by Takashima et al. [8]. Liu and Li [10] proposed a new segmentation method based on gradient and intensity combined level set methods that generated stable and accurate segmentation results for lung bronchia and nodules. Opfer and Wiemker [15] designed a general tumor segmentation approach which combines energy minimization methods with radial basis function surface modeling techniques.

Discrimination between actual nodules and false detections as well as the classification of a nodule (for example, malignant versus benign) depends upon measurement of disease-specific nodule characteristics (size, shape, texture, and internal structure), an approach following in this paper. McNitt-Gray et al. [5] used nodule size, shape and co-occurrence texture features as nodule characteristics to design a linear discriminant analysis (LDA) classification system for malignant versus benign nodules. Lo et al. [6] used direction of vascularity, shape and internal structure to build an artificial neural network (ANN) classification system for prediction of the malignancy of the nodules. Armato [7] used nodule appearance and shape to build an LDA classification system to classify pulmonary nodules into malignant versus benign classes.

There are several challenges encountered when creating and evaluating the CAD systems. One of them is the lack of consistent “ground truth" on which the training and generalization of learning-based classification and prediction CAD systems depend upon.

Another challenge results from the design of the ground truth: CAD rarely provides supporting information about its decision criteria. The ground truth typically uses only a binary representation (such as presence/absence and malignant/benign for the detection and diagnosis tasks, respectively) without any further information about the characteristics of the nodule (such as level of calcification, spiculation). Though meeting the goal for detection/diagnosis, the binary representation offers no guidance for explaining the nature of the suspected nodule.

This lack of publicly shareable databases for benchmarking pulmonary nodule CAD performance motivated the establishment of the LIDC. The LIDC database includes both the radiologist nodule outlines and the nodule characteristics which we use in this paper to bridge the semantic gap between nodule image features and human-interpretable diagnostic criteria. Given the fact that, in practice, physicians use several perceptual categories to make diagnoses, which are not always reliable and reproducible when used to train CAD systems [17-20], our approach is expected to both increase the accuracy of radiologist’s interpretation and to reduce the variability among radiologists.

The decision tree approach was motivated by our earlier work using linear [3] and logistic [21] regression approaches. Response to our linear regression analysis [3] with an early version of the LIDC indicated that the nodule characteristics might be better represented as ordinal or nominal rather than numeric. Using the dataset presented in this paper, logistic regression analysis was performed with an overall accuracy of less than 63% accuracy. Decision trees offer another approach to predicting non-numeric class labels and permit assigning probabilities to the decisions, a decision support criteria beneficial to a clinical setting. As shown in Table 1, decision trees markedly outperform logistic regression for the cases where at least 3 radiologists agreed.

Table 1: Comparison of logistic regression and decision trees when at least 3 radiologists agree on characteristic

Characteristics	**Logistic Regression**	**Decision Trees**
Lobulation	41.27%	68.25%
Malignancy	31.15%	90.16%
Margin	57.14%	82.14%
Sphericity	58.82%	71.76%
Spiculation	62.07%	78.16%
Subtlety	93.18%	94.32%
Texture	94.17%	98.33%
Overall Accuracy	**62.54%**	**84.64%**

Decision trees include a feature selection mechanism which permits non-linear separation of class boundaries, unlike the earlier linear approaches. This non-linear approach is motivated by the complexity of the nodule data given the significant disagreement between radiologists [9]; for example, in this study, two radiologists agree on the contour of a nodule but one assigns a spiculation (shape) score of five (5) while the other assigns one (1).

3. Methodology

In this section we present our methodology for finding mappings between the image features and nodule characteristics as summarized in Figure 1. The basic procedure uses the radiologist-drawn contour as a template for image segmentation and feature extraction. Since a nodule might appear in many CT slices, only the largest area slice is chosen to represent the nodule as selected by a particular radiologist. The image features extracted from this slice segmentation and the diagnostic characteristics annotated by the radiologist form this nodule-radiologist data point. Up to four (4) data points might represent this nodule.

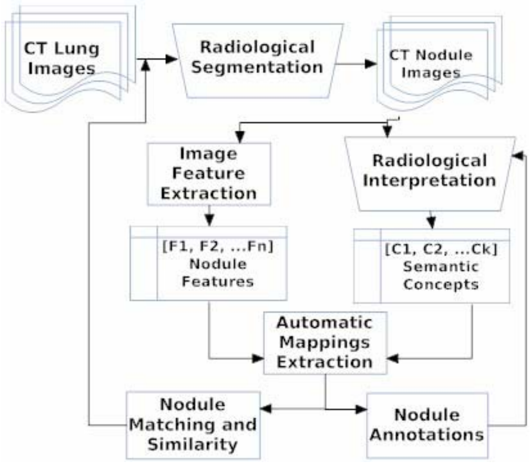

CT Lung Images
Radiological Segmentation
CT Nodule Images
Image Feature Extraction
Radiological Interpretation
[F1, F2, ...Fn] Nodule Features
[C1, C2, ...Ck] Semantic Concepts
Automatic Mappings Extraction
Nodule Matching and Similarity
Nodule Annotations

Figure 1: Diagram of the mapping framework

In Section 3.1 we present the data set. In Section 3.2, we present the image feature extraction process and, in Section 3.3, we present the decision tree approach applied to find mappings from the image features to the radiologists' assessments. In Section 4 we discuss our results and conclude in Section 5 on how the results can be further improved.

3.1. Data Set

The data used in this study were generated from 85 cases of thoracic CT collected by the LIDC. In the marking process, up to 4 radiologists marked the boundary of lung nodules with sizes between 3 mm and 3 cm for every slice on which the nodule appears and rated nine semantic characteristics for each identified nodule. The nine characteristics selected by the LIDC represent physical and diagnostic criteria usable by radiologist for diagnosis and consultation with other physicians and patients.

The LIDC attempts to capture the differences of radiologist diagnostic opinion and does not force any consensus. This lack of agreement results in nodules marked and characterized by up to 4 radiologists. Nodule boundary outlines are marked on each CT slice where the nodule is present, though each radiologist scores the characteristics only once per nodule. Since the database is designed for benchmarking CAD performance, some cases contain no nodules and the dataset at the time of our experiment contained 149 nodules from 60 of the total 85 cases. Given the diversity of radiologist opinion, we examine the semantic mappings for various levels of agreement, using an N-of-4 approach, where N is the number of agreeing and four (4) is the total number of radiologists, similar to [13], and report results for the nodules where at least two (2) or at least three (3) radiologists agree. To eliminate bias, we select only one nodule image (the largest) from the set of nodules marked by a radiologist. The number of nodule images, nodules, and cases of the reduced data used for predict each characteristic are presented in Table 2.

Table 2: Number of images, nodules, and cases, respectively, for each characteristic

Characteristics	Entire dataset	At least 2 agreed	At least 3 agreed
Lobulation	379 (images), 149 (nodules), (60 cases)	180, 73,42	63, 20, 19
Malignancy		187,73, 42	61, 19, 15
Margin		186,77, 42	56, 17, 12
Sphericity		197,77, 45	85,27,20
Spiculation		200,77, 43	87, 27, 24
Subtlety		194,70, 41	88, 25, 21
Texture		222,79, 43	120,34, 24

3.2. Feature Extraction

In order to quantify the image content, we calculated four types of image features for each nodule: size, shape, intensity, and texture; this feature extraction stage generated 64 image features as presented below. The choice of these features was based on a literature review of the most common image features used for pulmonary nodule detection and diagnosis by existing CAD systems [5-8].

Shape measurements include circularity measured by the ratio of nodule area over area of a circle with the same convex hull. Roughness is the perimeter ratio of the object to the convex hull. Eccentricity is the ratio of distance between the foci and major axis length of an ellipse with the same second-moments as the region. Solidity and extent measure the percentage of the convex hull and bounding box covered by the nodule. The RadialDistanceSD is the standard deviation of the distances from every boundary pixel to the centroid of the region. Size measurements include area, perimeter, convex hull perimeter, diameter of an equivalent area circle, and the major and minor axis length of ellipse with the same normalized second central moments as the region. Intensity features capture the absolute and relative brightness of the pixel in both the foreground (nodule) and background (bounding box around nodule) regions; for this study we use the min, max, mean, and standard deviation of each region as well as the absolute difference between the mean of the segmented (foreground) and the background.

We applied three well-known texture analysis techniques: co-occurrence matrices (a statistical-based method), Gabor filters (a transform-based method), and Markov Random Fields (a model based approach).

Co-occurrence matrices represent the conditional joint probability of gray level pairs within a window based upon their separation and orientation [11]. Our implementation computes the Haralick texture descriptors for the twenty (20) matrices formed from five (5) separations (distances of 1-5 pixels) over the half-plane of four (4) angles (0º, 45º, 90º, and 135º). We averaged the descriptors by distance then selected the minimum value by direction to produce eleven (11) Haralick features per nodule. Gabor filters [12] capture localized texture frequencies by convolving an image with Gaussian modulated frequency transform. Our method uses the resulting means and standard deviations of the twelve (12) responses from applying four (4) orientations (0º, 45º, 90º, and 135º) and three (3) frequencies to a 9x9 image window. Markov Random Fields (MRFs) model the local contextual information of an image [14] using a window approach. Using the Cesmeli [15] algorithm, we convolved the image with a 9x9 MRF window and computed the mean response for 4 angular rotations (0°, 45°, 90°, 135°) and variance, for a total of 5 MRF features. Figure 2 shows an example of feature values for a nodule representation.

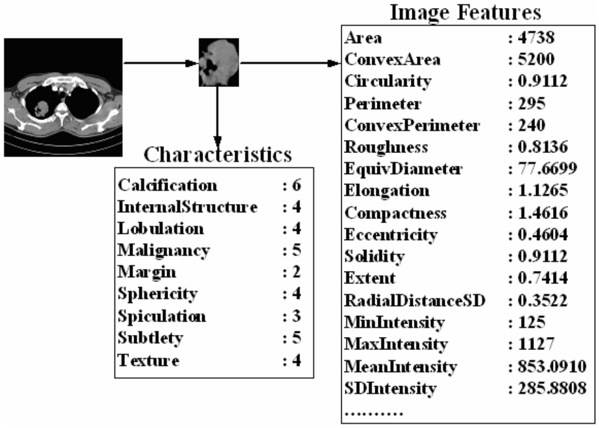


Image Features

Area	: 4738
ConvexArea	: 5200
Circularity	: 0.9112
Perimeter	: 295
ConvexPerimeter	: 240
Roughness	: 0.8136
EquivDiameter	: 77.6699
Elongation	: 1.1265
Compactness	: 1.4616
Eccentricity	: 0.4604
Solidity	: 0.9112
Extent	: 0.7414
RadialDistanceSD	: 0.3522
MinIntensity	: 125
MaxIntensity	: 1127
MeanIntensity	: 853.0910
SDIntensity	: 285.8808
....	

Characteristics

Calcification	: 6
InternalStructure	: 4
Lobulation	: 4
Malignancy	: 5
Margin	: 2
Sphericity	: 4
Spiculation	: 3
Subtlety	: 5
Texture	: 4

Figure 2: An example of nodule characteristics assigned by a radiologist and features extracted from the segmented nodule.

3.3. Learning the Semantic Mappings

This paper applies decision trees, an information-theoretic machine learning [16] technique, to predict the value (target category) of the human-interpretable diagnostic characteristic of a nodule from the image-extracted nodule features. Decision trees learn target categories by splitting training data based on attribute values into lower entropy sub-trees using information gain and can classify data which are not linearly separable. We employ the C4.5 algorithm variant (J48 in Weka). The final tree represents a sequence (conjunction) of decision criteria based upon feature (attribute) values with the leaves of the tree representing the classifications.

In this study we consider only seven (7) characteristics and eliminate calcification and internal structure (each received only one rating). The model was validated by using leave-one-out cross validation. The classification accuracies are reported in Table 3, where accuracy is defined as the total number of correctly predicted values for the characteristic divided by the total number of samples. Overall, the total accuracy is 84.64% (with > 3 radiologist agreements).

Table 3: Classification Accuracy

Characteristics	At least 2 agreed	At least 3 agreed
Lobulation	57.22%	68.25%
Malignancy	68.98%	90.16%
Margin	61.83%	82.14%
Sphericity	63.45%	71.76%
Spiculation	69.50%	78.16%
Subtlety	65.45%	94.32%
Texture	81.08%	98.33%

Five (5) characteristics are predicted with greater than 75% accuracy, when at least three (3) radiologists agree: malignancy, margin, spiculation, subtlety, and texture, with lobulation and sphericity about 70% accurate.

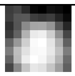
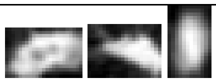
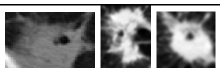
When analyzing the importance and contribution of the image features (Table 4) for predicting radiologists' assessments, we found that only 37.5% of the features (24 of 64) were needed to form the decision tree rules for these seven (7) characteristics, with only three (3) required for malignancy but ten (10) for spiculation.

Table 4: Number of attributes and maximum rule depth (# of decisions required to predict the characteristic) indicate the complexity of the decision tree.

Characteristic	# of Attributes	Max Rule Depth
Malignancy	3	2
Margin	5	4
Spiculation	10	6
Subtlety	2	2
Texture	3	2
Lobulation	7	5
Sphericity	9	6

The attributes selected by the decision tree correspond to the features included in our literature review, with some anomalies. For instance, MRFs and Gabor filters are selected for texture, as expected. However, if the area is large, texture is always 4. Spiculation attributes include radial distance, compactness, perimeter, and roughness, as anticipated; but also include Gabor (texture frequency). Malignancy attributes include intensity (bright nodules are less likely malignant), but also texture features. Subtlety attributes indicate darker, elongated nodules which are less obvious. Margin uses intensity of nodule and background as expected. Table 5 shows an example of mappings.

Table 5: Rules for subtlety; second column shows examples of nodules classified based on the learned mappings

Rules for subtlety	Nodule examples
IF (minorAxisLength <= 0.15) AND (maxIntensity <= 0.23) THEN subtlety = 1	 Nodule ID: 65
IF (minorAxisLength <= 0.15) AND (maxIntensity > 0.23) THEN subtlety = 4	 Nodule IDs: 42, 47, 105
IF (minorAxisLength > 0.15) THEN subtlety = 5	 Nodule IDs: 24, 46, 68

4. Discussion of the results

Using the decision tree classification approach, we were able to predict malignancy, subtlety, and texture with over 90% accuracy for the nodules on which at least three (3) radiologists agree with respect to the corresponding characteristic.

We found that most of the misclassified nodules are those whose ratings were either equal to 2 or 3; these findings correspond to the lack of agreement among radiologists themselves. In the LIDC dataset [13], agreement about nodule existence between two radiologists occurs for only 84% of the nodules and it drops to 70% for three (3) and to only 43% of the nodules for four (4) radiologists. Figure 3 shows there is little agreement for many of the characteristics, with agreement occurring for only low or high ratings.

Furthermore, not only do they disagree about the presence of nodules; when they agree, they draw significantly different contours around the nodules. R. Opfer and Wiemker [13] estimated a 50% variability in the regions selected by multiple radiologists for the same nodule. Our feature extraction technique uses each radiologist outline to measure image features. Differences between radiologists' outlines may generate significantly different feature measurements and predictions, even though the radiologists might agree on the characteristic. We intend to investigate independent methods for determining an outline for the region of the nodule to de-couple this potential conflict; incorporating a nodule segmentation algorithm may be required for this task.

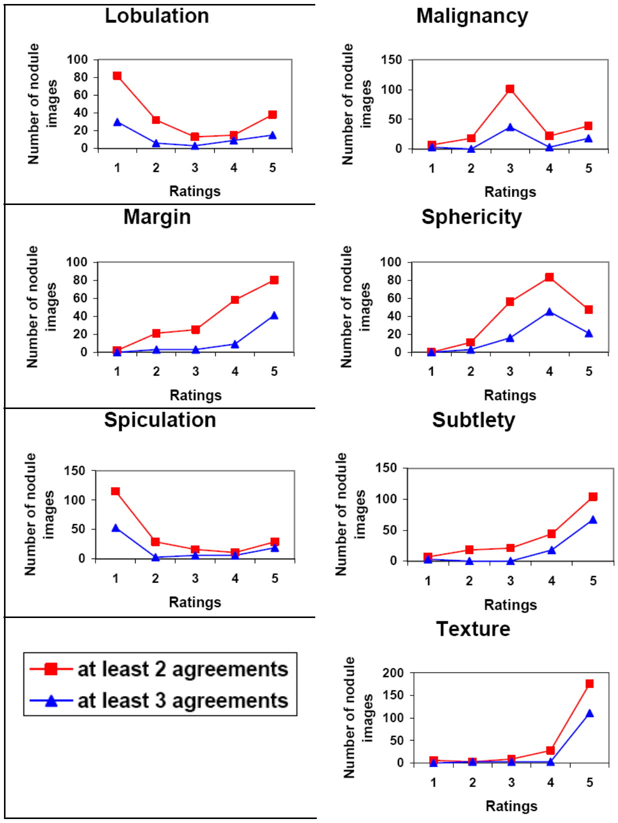


Figure 3: Distributions of ratings for characteristics when only images with at least 2 agreements on the same ratings are included or only images with at least 3 agreements.

The three characteristics with the poorest results represent perceptions of shape. The worst two, lobulation and sphericity, can be considered low frequency or smooth shapes, while spiculation represents a higher frequency change in shape. New shape measurement features are being developed to capture these characteristics.

5. Conclusions

Most CAD systems mimic domain knowledge by extracting image features and training the systems by some algorithms based on the image features against the ground truth provided by domain experts. However, the image features used in the CAD systems are tenuously related to human perception.

In this paper, we proposed a supervised (decision-tree) learning approach for finding the mappings between low-level image features and high-level human concepts used for lung cancer diagnosis. From the preliminary results, we found that the radiologists' perception with respect to malignancy, margin, spiculation, subtlety, and texture can be captured with high accuracy (higher than 75%) based on low-level image features.

Our preliminary results are promising and serve as a CAD framework to support diagnostic decision making in lung nodule diagnosis. Acting as a consulting second reader, this approach provides an estimated rating for each nodule characteristic to offer guidance to the radiologist in their diagnosis as well as conveying nodule assessments for improving the consistency among multiple readers. Beyond this work, we intend to investigate other image features and construct visual concept ontologies for lung nodule interpretation.

6. References

[1] K. T. Bae, Jin-Sung Kim, Yong-Hum Na, Kwang Gi Kim, and Jin-Hwan Kim, "Pulmonary Nodules: Automated Detection on CT Images with Morphologic Matching Algorithm-Preliminary Results," Radiology 2005 236: 286-293.

[2] K. Doi, "Current status and future potential of computer-aided diagnosis in medical imaging," The British Journal of Radiology 2005.

[3] D. S. Raicu, Ekarin Varutbangkul, Janie G. Cisneros, Jacob D. Furst, David S. Channin, Samuel G. Armato III, "Semantics and image content integration for pulmonary nodule interpretation in thoracic computed tomography," SPIE Medical Imaging 2007.

[4] K. Suzuki, A. H. Dachman, "Mixture of expert artificial neural networks with ensemble training for reduction of various sources of false positives in CAD," SPIE Medical Imaging 2007

[5] M. F. McNitt-Gray, N.Wyckoff, J.W. Sayre, J. G. Goldin, and D. R. Aberle, "The effects of co-occurrence matrix based texture parameters on the classification of solitary pulmonary nodules imaged on computed tomography," Computerized Medical Imaging and Graphics, vol. 23, no. 6, pp. 339-348, 1999.

[6] S.C. B. Lo, M. T. F. Li-Yueh Hsu and, Y.-M. F. Lure, and H. Zhao, "Classification of lung nodules in diagnostic CT: An approach based on 3-D vascular features, nodule density distributions, and shape features," in SPIE Proc., vol. 5032, 2003, pp. 183-189.

[7] Armato SG III, Altman MB, Wilkie J, Sone S, Li F, Doi K, Roy AS, “Automated lung nodule classification following automated nodule detection on CT: A serial approach”, Medical Physics 30: 1188-1197, 2003.

[8] S. Takashima, S. Sone, F. Li, Y. Maruyama, M. Hasegawa et al., "Small solitary pulmonary nodules detected at population-based CT screening for lung cancer: reliable high-resolution CT features of benign lesions," American Journal of Roentgenology, vol. 180, no. 4, pp. 955-964, 2003.

[9] S. G. Armato, G. McLennan, M. F. McNitt-Gray, C. R.Meyer, D. Yankelevitz, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, A. P. Reeves, B. Y. Croft, and L. P. Clarke, "Lung Image Database Consortium: Developing a resource for the medical imaging research community," Radiology, vol. 232, no. 3, pp. 739-748, 2004.

[10] S. Liu, and J. Li, "Automatic Medical Image Segmentation Using Gradient and Intensity Combined Level Set Method," The 28th IEEE EMBS Annual International Conference, 3118-3121, 2006.

[11] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," IEEE Trans. On Systems, Man, and Cybernetics, vol. 3, no. 6, 610-621, 1973.

[12] T. Andrysiak and M. Choras, "Image retrieval based on hierarchical Gabor filters," International Journal Applied Computer Science, vol. 15, no. 4, 471-480, 2005.

[13] R. Opfer, R. Wiemker, "A new general tumor segmentation framework based on radial basis function energy minimization with a validation study on LIDC lung nodules," SPIE Medical Imaging Conference, San Diego, CA, February 2007.

[14] C. Bouman, "Markov random fields and stochastic image models," in 1995 IEEE International Conference on Image Processing, 1995. Tutorial notes.

[15] E. Cesmeli and D. Wang, "Texture segmentation using Gaussian-Markov random fields and neural oscillator networks," IEEE Transactions on Neural Networks, vol 12, pp. 394-404, March 2001.

[16] T.M. Mitchell, Machine Learning, McGraw-Hill, 1997.

[17] G. D. Rubin, John K. Lyo, D.S. Paik, A. J. Sherbondy et al. “Pulmonary Nodules on Multi–Detector Row CT Scans: Performance Comparison of Radiologists and Computer-aided Detection”, Radiology 2005

[18] F. Li, Q. Li, R. Engelmann, M. Aoyama, S. Sone, H. MacMahon, K. Doi, “Improving Radiologists’ Recommendations With Computer-Aided Diagnosis for Management of Small Nodules Detected by CT”, Academic Radiology, Volume 13, (8) 943-950, 2006.

[19] R Wiemker, P Rogalla, T Blaffert, D Sifri, O Hay, et al. “Aspects of computer-aided detection (CAD) and volumetry of pulmonary nodules using multislice CT” British Journal of Radiology (2005) 78, S46-S56.

[20] J. Burns, L. B. Haramati, K. Whitney, and M. N. Zelefsky, “Consistency of reporting basic characteristics of lung nodules and masses on computed tomography,” Academic Radiology, vol. 11, pp. 233–237, 2004.

[21] Ekarin Varutbangkul, D. Raicu and J Furst, “A Computer-Aided Diagnosis Framework for Pulmonary Nodule Interpretation in Thoracic Computed Tomography”, DePaul CTI Research Symposium 2007.