Bridging the Evaluation Gap Challenge to Diagnostic Labeling of Pulmonary Nodules

William H. Horsthemke, Daniela S. Raicu, Jacob D. Furst
DePaul University
School of Computer Science
Chicago, IL
horsthemke@acm.org, draicu@cs.depaul.edu, jfurst@cs.depaul.edu

Abstract

Diagnosis of focal anomalies in medical images can be aided by automatic labeling and rating of visual features representing medically meaningful diagnostic characteristics as observed by trained image clinicians and radiologists. Prediction performance remains unsatisfactory and significant challenges persist in evaluating the system performance due to the uncertainty about sources of poor performance: whether the problems enter in 1) the measurement and selection of image features, 2) the prediction and classification methods, or 3) the ground truth provided by radiologists. Earlier work explored a diversity of features and modeling methods but no studies have evaluated the ground truth as a cause of poor performance, mainly due to lack of measurement techniques. This paper introduces a disagreement index to systematically study the disagreement on nodules, characteristics, and detection levels towards understanding which characteristics are more predictable and measuring radiologists' agreement patterns towards developing strategies for building and training future models for predicting diagnostic characteristics.

1 Introduction

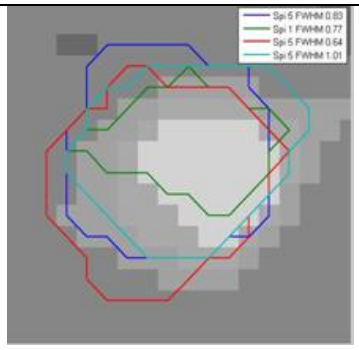
Computer aided detection (CADe) and diagnosis(CADx) aims to augment the radiologist in meeting the increased demand for diagnostic imaging by serving as second reader. While increasingly accurate in detection and diagnosis, CADx rarely offers supporting guidance about the rationale for the diagnosis or supplies descriptive annotations about medically meaningful diagnostic characteristics [1]. A potential opportunity to address this deficiency is the use of semantic mapping to extract image features and build predictive models of diagnostic characteristics for labeling images. For CADe/CADx, semantic mapping promises to add clinically relevant diagnostic evidence to support the medical decision maker and enabling the use of case-based reasoning through the content-based retrieval (CBIR) of similar images [2].

In this paper, we introduce a method for measuring disagreement among radiologists in their ratings of diagnostic characteristics of pulmonary nodules in the Lung Image Database Consortium (LIDC) [1]. This disagreement method is applied to shape-based characteristics (spiculation, lobulation, and sphericity) which are measured using the radial normal index (RNI) recently developed for boundary based shape feature extraction as well as the Fourier descriptor method [19]. These features are measured directly from radiologist-drawn outlines, assuming their outlines best represent their perception of the boundary of the nodules. Through example cases, we verify the approach and show that the shape metric varies accordingly. Comparing the prediction results with the disagreement scores, we show that the failure to predict radiologist

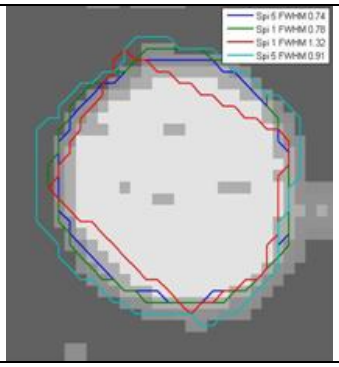
ratings for shape characteristics is the result not of feature selection but of the disagreement about the ratings. This paper concludes with discussion of plans to manage the disagreement and plans for feature extraction methods which are tolerant to radiologist disagreement.

The LIDC [3] has developed a lung nodule collection and reporting protocol for four (4) radiologists to identify, in thoracic CT scans, lesions in one of three (3) categories: 1) nodules between 3 and 30 mm in maximum diameter, 2) nodules less than 3 mm (unless clearly benign), and 3) non-nodules larger than 3 mm. When radiologists identify a nodule in category 1 (3-30 mm), they draw an outline around the nodule and rate a set of nine (9) diagnostics characteristics on a scale of 1 - 5: texture, subtlety, spiculation, sphericity, margin, malignancy, lobulation, internal structure, and calcification (different scale: 1 - 6).

The LIDC protocol does not enforce consensus among the radiologists, allowing each radiologist to review the outlines and ratings of the other (3) radiologists. This is accomplished by an initial blinded reading by each radiologist followed by a second, unblinded reading where the initial reporting is present to each radiologist. On the second reading, each radiologist is free to retain or modify their initial ratings and outlines, including incorporating other radiologist drawn outlines. In addition to not enforcing a consensus among the ratings and outlines of the readers, the radiologists are free to select the overall category of the lesion or provide no markings for the lesion. As a result, the nodules may be marked by up to 4 radiologists. At the time of this study, the LIDC database contained 85 cases overall, with 60 cases containing 147 nodules. Since there can be many slices per nodule with only one rating per radiologist, a bias-limiting approach selects only the largest area slice as the representation of the nodule with area defined by the radiologist outline. Depending upon the number of radiologists agreeing on the existence of the nodule, there can be up to 4 slices and 4 ratings per nodule. In comparing the effects of agreement, the dataset is partitioned by the number of radiologists who rate the nodule with agreements of 2, 3, and 4. In the shape measurement method used in this paper, each radiologist outline forms the basis for measuring the shape and predicting the same radiologists rating for the nodules. Examples of radiologists' diversity of opinion on the same nodules are illustrated in Figure 1. Four markedly different outlines for the same nodule are illustrated in Figure 1-A while Figure 1-B illustrates four similar outlines with markedly different ratings for the shape (spiculation) of the nodule.



A Spiculation Ratings: 5,1,5,5



B Spiculation Ratings: 5,1,1,5

Figure 1 Disagreement between radiologists appears both in outlines and ratings of diagnostic characteristics. For the two example nodules, radiologists fully disagree and rate spiculation at both 1 and 5 on a scale of 1-5.

2 Related Work

Measuring disagreement among radiologists is studied for assessing differences in imaging equipment and diagnostic opinion. In most studies, the radiologist is identified and tracked throughout the study, but the blinded design of the LIDC removes any record of the radiologist, making each case an independent study. Most studies measure interobserver agreement using Kappa statistics [8] or ROC[11]. Most studies consider only binary values such as malignant/benign or present/absent while the LIDC study uses multi-valued ordinal ratings. When reporting multi-valued findings such as disease severity {absent, minimal, moderate, or severe}, a weighted (often quadratic) Kappa method is used [4]. Though widely used, Kappa statistics vary according to disease prevalence and are unsuitable for comparative studies [8]. When ground truth is known, each radiologist's performance can be measured by Az (area under the ROC curve) scores and comparison made using ANOVA methods [12]. These methods measure the findings of known radiologists' examinations of the same set of cases, a requirement not met by the LIDC pulmonary nodule study where the radiologist identity is blinded and potentially different between cases.

Several studies used radiologist rankings of similarity between regions of interest to estimate subjective similarity of image characteristics. Muramatsu et al. [13] studied agreement in similarity for mammographic regions and used Spearman's rank ordered correlation coefficients to assess intra-observer agreement between the first and second readings of the same data. They averaged each observer's similarity rankings then used Pearson's correlation coefficient between all-pairs of observers to assess inter-observer correlation. They concluded that their method for obtaining similarity scores for lesions is robust even though some radiologists were noticeable outliers.

No studies have been found that measure agreement in rating diagnostic characteristics such as the LIDC. In one of the few studies examining radiologists' ratings for image-based diagnostic features, Nakamura [14] qualifies the ratings as varied but does not report any quantitative measure of this variance or other measures of inter-observer agreement for their study group which used radiologists from a single institution, an academic medical center. There are five (5) medical centers participating in the LIDC but due to the blinded study there is no method to identify whether differences in agreement are due to radiologists or institutions.

Inter-observer variability in the drawing of outlines around pulmonary nodules was studied directly by Meyer [10], who compared the relative performance of six radiologists using three outlining methods to define the spatial extent of nodules and concluded that radiologists represent the major source of variance in the final outlines. Their study introduced the combination method of p-map (probability map) to represent the likelihood that a pixel is a member of the nodule. Using the initial LIDC dataset, Opfer [16] estimated a 50% variability in the regions selected by multiple radiologists for the same nodule. Reeves [18] concluded that a high inter-observer variation exists when applying four pulmonary nodule size metrics to the LIDC radiologist outlines. These reports indicate that the variability in outlines presents a significant challenge to the characterization of the pulmonary nodule.

2.1 Shape of Pulmonary Nodules

The shape of pulmonary nodules is visually assessed by radiologists and the appearance of spiculation along the boundary of nodules indicates malignancy [21]. Giger et al. employed geometric methods (effective diameter and degree of circularity) to detect suspicious nodules in chest x-rays [5]. Nakamura et al. used the root-mean-square and first moment of a Fourier transformation of the nodule outline while measuring the radial gradient index (RGI) to detect nodule spiculation [14]. Towards predicting the LIDC diagnostic characteristics for shape, Raicu et al. extended the set of geometric features to measure roughness, eccentricity, solidity, extent, and radial standard deviation [17].

This paper applies the Radial Normal Indexing (RNI) to measure and predict the shape-based diagnostic characteristics in the LIDC. The RNI method was introduced in [6] as an adaptation of Radial Gradient Indexing for use on radiologist-drawn outlines as provided in the LIDC. In addition to the RNI shape features, this paper also applies the Fourier shape descriptor technique [19] which formed the validation benchmark for RNI [6].

The radial normal index (RNI) captures the variability in the boundary (outline) of an object by comparing the perpendicular (normal) along the outline to the radial angle from the center of the object, similar to the RGI method. As illustrated in Figure 2, the boundary (represented by the gradient magnitude) of a smooth object is regular while a spiculated object is quite irregular. Plotting the gradient direction illustrates the differences in angular variation along the boundary of the near-circular and irregular (spiculated) object. The radial gradient index (RGI) method measures this variability by the difference between the radial angle from the center of the object to the boundary and the gradient angle at that location along the boundary. The radial normal index (RNI) mimics this approach by substituting the normal of boundary outline for the direction of the gradient.

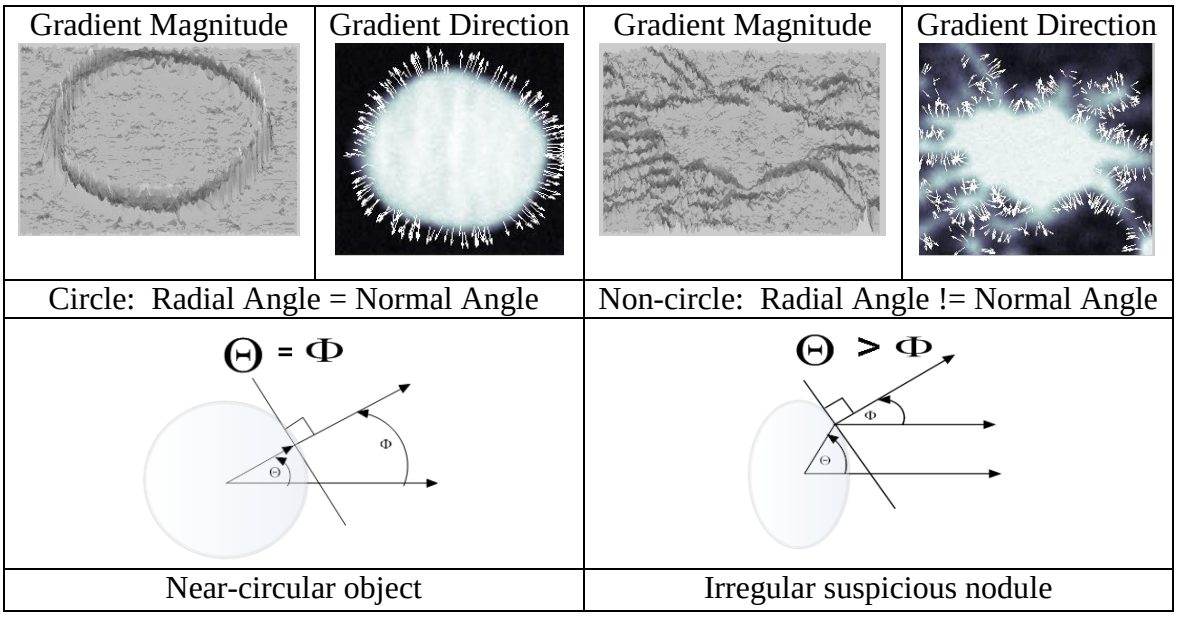


Figure 2. Gradient magnitude and direction along boundary of regular and irregular objects. . RNI method for computing difference between radial angle (θ) - direction of vector from center of object to point on perimeter – and normal angle (φ) -direction of vector normal to object at point on perimeter.

The method of the Radial Normal Index is illustrated in Figure 2. The radial gradient angle (θ) is computed from the center of the object, while the normal angle (φ) is computed as the angle from the object in the direction of the normal. The difference between the radial and normal angles represents the value of the RNI at this value of the radial angle. Sweeping along the 360 degrees of the radial angle and accumulating the differences produces an angular difference distribution (histogram) and a set of bins and measures of central tendency (standard deviation and full-width of the maximum height (FWMH) [7]) for use as feature vectors.

When applied to LIDC radiologist-drawn nodule outlines, RNI captures increased angular variability along outlines and predicts some of the radiologist ratings for spiculation. As illustrated in the Figure 3, the less spiculated (LIDC Spiculation Rating = 4) outline has a narrower distribution and smaller FWHM than the more spiculated (LIDC Spiculation Rating = 2) nodule on the right. The ratings for Spiculation range from a maximum with rating of 1 to a minimum of 5.

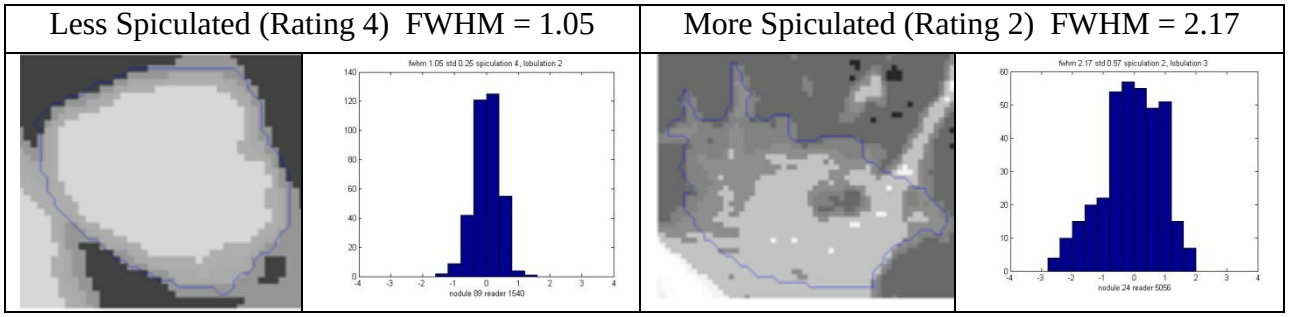


Figure 3. RNI correlates well with spiculation for selected nodules.

2.2 Semantic Mapping

CADe/CADx can be considered a diagnostic mapping from image features to detection or diagnosis [10]. Semantic mapping aims to form an intermediate step in this diagnostic mapping process by creating image-based diagnostic characteristics which are medically meaningful and semantically similar to radiologists' diagnostic interpretations. Diagnostic mapping becomes a two step process with the first step mapping from image features to diagnostic characteristics (subjective features [9,14]) and the second step mapping from diagnostic characteristics (subjective features) to overall diagnosis.

Learning radiologists' interpretations of diagnostic characteristics presents a significant challenge. Li et al. developed a nodule similarity rating based upon a set of extracted image features selected to represent radiologists' diagnostic characterizations of lesions [9]. They performed two studies, one to show the value of a CBIR-like system and the other to evaluate feature performance in predicting image similarity. In the first study, they asked radiologists to diagnose an unknown lesion then presented 6 labeled lesions (3 similar benign and 3 similar malignant with similarity based upon feature similarity) and asked the radiologist to repeat their diagnosis. They report improvement in radiologist diagnostic performance (A_z) using this CBIR-based approach. In the second study, pairs of images are presented to the radiologists who rate their similarity on a scale from 0 (not similar) to 3 (almost identical). They report the correlation of various image features in predicting the radiologist similarity ratings, but chose to predict not the raw ratings but the average similarity rating due to rating variability.

A seminal study on semantic mapping evaluated the use of extracted image features to predict the subjective diagnostic characteristics rated by radiologists [14]. Using pulmonary nodules in chest radiography, Nakamura et al. asked radiologists to rate subjective features such as shape, margin irregularity, spiculation, lobulation, etc. on a scale of 1 to 5. After extracting raw image features such as intensity statistics and geometric, Fourier, and radial gradient indices for shape, they correlated image features to radiologists' subjective ratings. Their results show that radial gradient features are strongly correlated with radiologists' ratings (interpretation) of spiculation while geometric features correlate with nodule shape.

The Nakamura study compared the performance of a single step CADx approach to diagnosis of pulmonary nodules using image features and the second step of the approach described above where the radiologists' ratings for diagnostic characteristics are used to predict diagnosis. They reported that the single step CADx predictive performance exceeded the prediction performance of radiologists' ratings. Their study illustrates the major challenge to semantic mapping and indicates that the design and selection of images features is less important than obtaining consistent ratings from radiologists for diagnostically useful image characteristics.

3 Methods

This paper examines disagreement and ratings prediction for the shape diagnostic characteristics according to various levels of detection of nodules agreement where detection represents a rating by a radiologist for a nodule, such as a nodule detected by two radiologists receives two ratings for each characteristic. This allows for five (5) major partitions representing three (3) pure groups where either 2, 3, or 4 radiologists rate the nodule, or combined partitions where at least two (2) or three (3) ratings are recorded as suggested by Ochs [15]. For each dataset partition, the prediction result is estimated using the best combination of feature extraction (RNI or FD) and machine learning method (decision tree or logistic regression) while the disagreement is measured using the Radiologist Disagreement Index (RDI) introduced in this paper.

The Radiologist Disagreement Index (Equation 1) takes the absolute difference between ratings per characteristic per nodule as the unit of interest for measuring the extent of radiologist disagreement. This produces a mean nodule disagreement (MND) score as the raw metric but the range of these raw results differs with the range and number of ratings due to the all-pairs approach and requires normalization with the maximum disagreement (MD_R) possible for R ratings. The maximum disagreement is computed by considering all possible combinations of R ratings where R= {2, 3, or 4} represents the number of radiologists rating the characteristic and computing the maximum of mean nodule disagreement for each combination. The maximum disagreement for the shape characteristics is 4 when only 2 radiologists rate the nodule and 2.67 when either 3 or 4 ratings are given. The maximum disagreement depends upon the range of the ratings which is 1-5 for the shape characteristics. Normalizing the disagreement provides a 0-1 scaled metric ranging from 0 for no disagreement to 1 for full disagreement and produces a Radiologist Disagreement Index for each diagnostic characteristic per nodule. For example, a nodule rated by three (3) radiologists with ratings of {1,2,4} has three (3) pairs of absolute differences {|1-2|, |1-4|, |2-4|} for a total difference of 6, a mean nodule disagreement (MND) of 2, a maximum disagreement of 2.67, and a normalized disagreement of 2/2.67 = 0.74

Mean Nodule Disagreement (MND) for R ratings =

$$\sum_{i=1}^{R} * \sum_{i+1}^{R} abs(ratings(i) - ratings(i+1)) \div (R \text{ choose } 2)$$

R choose 2 = number of pairs of ratings

Normalized Disagreement=MND/MaxDisagreement$(MD)_R$: MD_R = 4 for R=2 and 2.67 for R={3,4}

Equation 1. Method for computing normalized, mean all-pairs absolute ratings difference per characteristic per nodule.

4 Prediction and Disagreement Results

In Figure 4 (top panel), disagreement on shape-based diagnostic characteristics is compared with predictive performance (highest accuracy among the FD and RNI methods) for five groupings of the detection levels of agreement. The five groupings include three pure categories with only two, three, or four detection agreements (raters) and two aggregations combining levels of agreement where at least 2 or at least 3 radiologists detect and rate the nodule. Using these groupings, the group labeled "at least 2" represents the entire set of nodules detected by multiple radiologists. The nodules detected by only one radiologist are excluded from this study. As shown in Figure 4, prediction accuracy for spiculation and lobulation is greater for three (3) raters than 2 or 4, but accuracy for sphericity increases with the number of raters and is greatest for four (4) raters. The level of disagreement for spiculation and lobulation increases with the number of raters but sphericity follows the pattern of most other diagnostic characteristics (Figure 5) where the disagreement is greatest for three (3) raters. Disagreement on texture (Figure 5-D) differs by decreasing with number of raters.

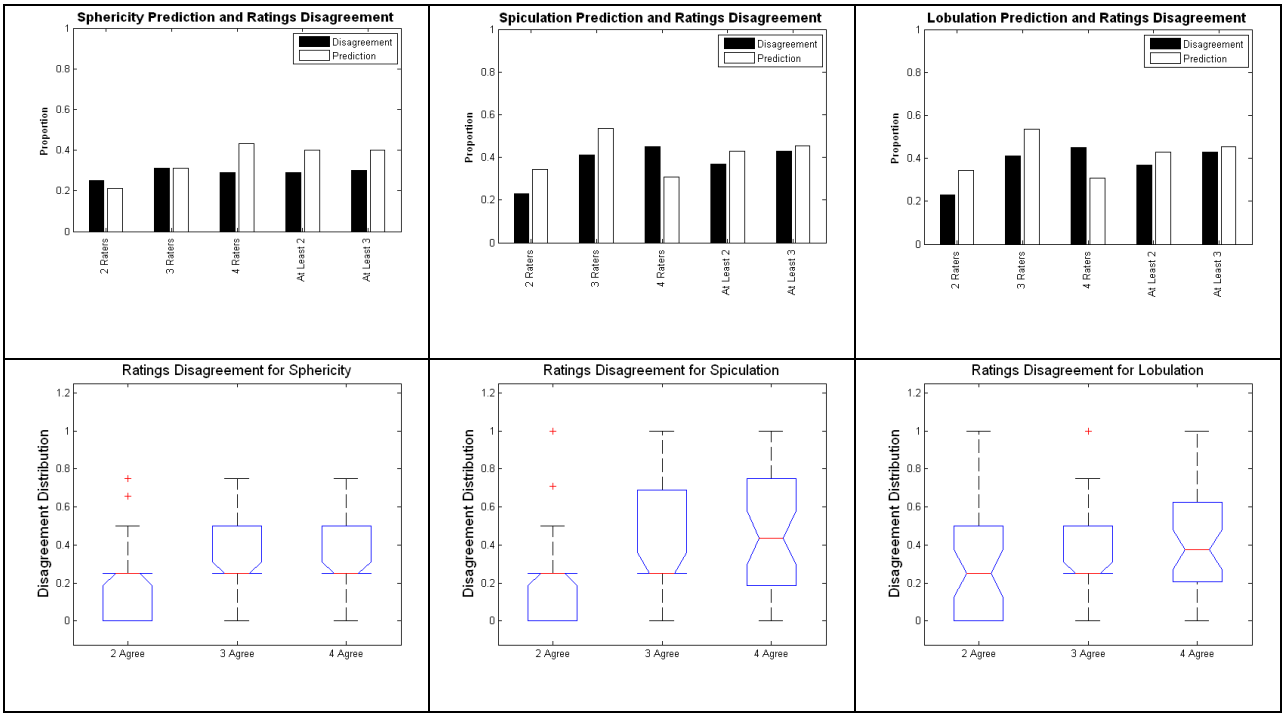


Figure 4. Prediction accuracy (*white*) and disagreement index *(black)* comparison Radiologists' disagreement on diagnostic characteristics; the major elements are the height of median disagreement (bar in middle of box), the spread of disagreement between the 25 and 75th percentile (the box), and the height of the 25 percentile (bottom of box).

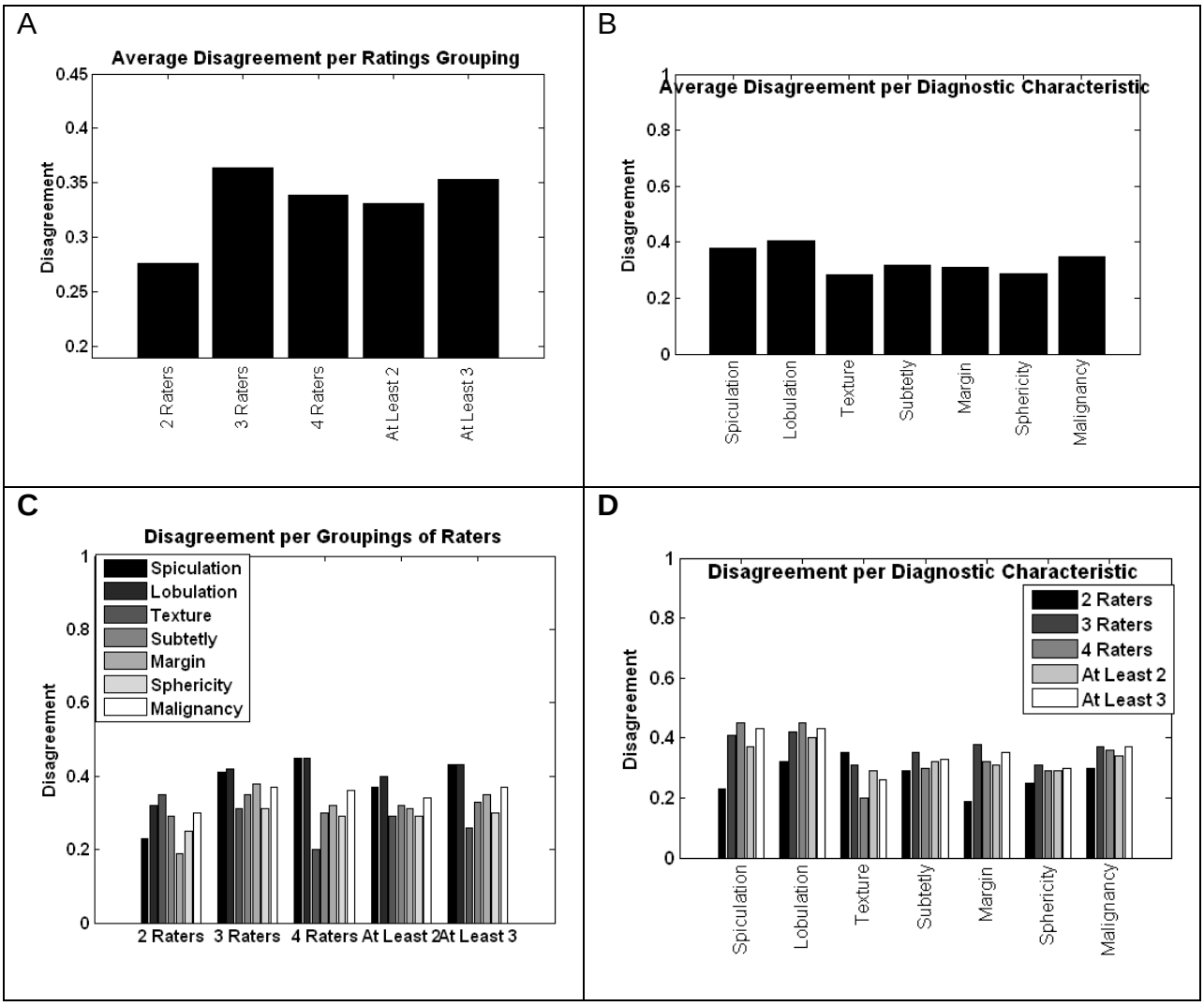


Figure 5. Disagreement per number of raters (detection level of agreement) and diagnostic characteristic

5 Discussion of Results and Conclusion

Diagnostic characteristic prediction accuracy varies markedly between the levels of detection agreement, though no single pattern holds for all characteristics. When considering only the pure detection levels (2, 3, or 4 raters), the predictive accuracy of spiculation and lobulation is greatest for three (3) raters, while sphericity is best predicted using four (4) raters. Sphericity prediction increases with the number of raters, while the prediction of lobulation and spiculation declines with four raters. In contrast to prediction, disagreement increases with the number of raters though at different rates per characteristic. As a result, the change in prediction is not clearly explained by the change in disagreement, nor does it follow the expected inverse relationship where an increase in disagreement corresponds to a decrease in prediction accuracy.

The box plots of Figure 4 (bottom panel) offer some insight into this variability by illustrating the distribution of disagreement for each characteristic at each level of pure detection agreement. Sphericity has a more consistently compact distribution while the range varies more

for lobulation and spiculation both overall and within the 25-75% percentiles. The median disagreement for all characteristics is near the lower range for the 3 raters. Overall these differences in the distribution of agreement suggest that the use of average disagreement might fail to capture important differences in disagreement which affect the predictive performance. The disagreement index uses the mean of the absolute differences but the median might better represent the predictive performance.

This methodology for measuring disagreement varies with the range and distribution of the ratings and predictive methods such as linear regression, a least squares approach, would tend to vary along with the distribution, but the predictive modeling methods employed for categorical prediction/classification (decision trees and logistic regression) are less influenced by the distribution of ratings. In this work, the ratings are treated as categorical within decision trees but ordinal for logistic regression, but neither categorical prediction method considers the ratings as interval values where the distance between values is considered. Using the current formulation, the RDI can only inform the future models in strategies for exploiting radiologist groupings with low (or high) disagreement, rather than an indicator for predictive performance.

Future work will pursue three agendas: 1) investigate whether median representation of disagreement indexes better correlates with prediction accuracy; 2) explore methods for combining individual radiologists' outlines and ratings into composite features and labels for group prediction, such as using a median rating and combining outlines using the probability map method, an intersection/union ratio method described in [10]; and 3) experiment with training strategies to exploit groupings with low disagreement and research other modeling methods for managing inconsistent datasets.

References

[1] Armato S.G., McLennan G., McNitt-Gray M.F., Meyer C.R., Yankelevitz D., Aberle D.R., Henschke C.I., Hoffman E.A., Kazerooni E.A., MacMahon H., Reeves A.P., Croft B.Y., & Clarke L.P. (2004).Lung Image Database Consortium: Developing a resource for the medical imaging research community, *Radiology*, 232(3): 739-748.

[2] Bui, A., Taira, R., Dionisio, J., Aberle, D., El-Saden S., Kangarloo H. (2002) Evidence-Based Radiology - Requirements for Electronic Access, *Academic Radiology*, Volume 9, Number 6.

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

[4] Fleiss, J. (1981). *Statistical methods for rates and proportions* 2nd Ed. Wiley 212-236.

[5] Giger M L, Doi K, MacMahon H. Metz C E, & Yin F F. (1990). Pulmonary nodules: computer-aided detection in digital chest images. *RadioGraphics* 17:861-865.

[6] Horsthemke, W.H., Raicu, D.S., & Furst, J.D. Evaluation Challenges for Bridging Semantic Gap: Shape Disagreements in the LIDC. International Journal of Healthcare Information Systems and Informatics. Invited Paper In Print.

[7] Huo Z., M. L. Giger, C. J. Vyborny, U. Bick, P. Lu, D. E. Wolverton, & R. A. Schmidt, (2005). Analysis of spiculation in the computerized classification of mammographic masses, Medical Physics 22, 1569–1579.

[8] Kundel, H. & Polansky, M. (2003). Measurement of Observer Agreement, *Radiology*.

[9] Li Q, Li F, Shiraishi J, Katsuragawa S, Sone S, & Doi K. (2003). Investigation of new psychophysical measures for evaluation of similar images on thoracic CT for distinction between benign and malignant nodules. *Medical Physics* 30:2584 -2593.

[10] Meyer CR, Johnson TD, McLennan G, et al. (2006). Evaluation of lung MDCT nodule annotation across radiologists and methods. *Academic Radiology* 13(10): 1254–1265.
[11] Metz CE (2008). ROC analysis in medical imaging: a tutorial review of the literature. *Radiological Physics and Technology*: 2-12.
[12] Miller, D., Wood, S., O'Shaughnessy, K., Castellino, R., (2004). Gold standards and expert panels: a pulmonary nodule case study with challenges and solutions, *Proceedings of SPIE Medical Imaging.*
[13] Muramatsu C., Li Q., Suzuki K., Schmidt R. A., Shiraishi J., Newstead G. M., & Doi K. (2005). Investigation of psychophysical measures for evaluation of similar images for mammographic masses: Preliminary results. *Medical Physics* 32: 2295-2304.
[14] Nakamura K, Yoshida H, Engelmann R, MacMahon, H., Katsuragawa, S., Ishida, T., Ashizawa, K., & Doi, K. (2000). Computerized analysis of the likelihood of malignancy in solitary pulmonary nodules with use of artificial neural networks. *Radiology,* 214:823–830.
[15] Ochs, R., Kim, H, Angel, E., Panknin, C.; McNitt-Gray, M., & Brown, M. (2007). Forming a reference standard from LIDC data: impact of reader agreement on reported CAD performance, *Proceedings of SPIE Medical Imaging.*
[16] Opfer R. & Wiemker R., (2007). A new general tumor segmentation framework based on radial basis function energy minimization with a validation study on LIDC lung nodules, *Proceedings of SPIE Medical Imaging*.
[17] Raicu D.S, Varutbangkul E., Cisneros J.G., Furst J.D., Channin D.S., & Armato III S.G. (2007), Semantics and Image Content Integration for Pulmonary Nodule Interpretation in Thoracic Computed Tomography, *Proceedings of SPIE Medical Imaging*.
[18] Reeves, A., Biancard, A., Apanasovich, T., Meyer,C., MacMahon, H, Van Beek, J., Kazerooni, E., Yankelevitz, D., McNitt-Gray, M., McLennan, G., G. Armato III, S., Henschke, C., Aberle, D., Croft, B. & Clarke, L. (2007). The lung image database consortium (LIDC): A comparison of different size metrics for pulmonary nodule measurements. *Academic Radiology*.
[19] Svoboda T., Kybic J., Hlavac V. (2007) Image Processing, Analysis, and Machine Vision, A MATLAB Companion, Thomson Learning, Toronto.
[20] Witten IH & Frank E, (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition, Morgan Kaufmann.
[21] Wormanns, D. & Diederich, S., (2004). Characterization of small pulmonary nodules by CT, *European Radiology*, pp 1380-1391.