Evaluation Challenges for Bridging Semantic Gap: Shape Disagreements on Pulmonary Nodules in the Lung Image Database Consortium

William H. Horsthemke, Daniela S. Raicu, Jacob D. Furst
Intelligent Multimedia Processing Laboratory, School of Computer Science, Telecommunications, and Information Systems, DePaul University, Chicago, IL 60604, USA;
horsthemke@acm.org, draicu@cs.depaul.edu, jfurst@cs.depaul.edu

Abstract

Evaluating the success of prediction and retrieval systems depends upon a reliable reference standard, a ground truth. The ideal gold standard is expected to result from the marking, labeling, and rating of domain experts of the image of interest. However experts often disagree and this lack of agreement challenges the development and evaluation of image-based feature prediction of expert-defined "truth." This paper addresses the success and limitations in bridging the semantic gap between CT-based pulmonary nodule image features and the ratings of diagnostic characteristics recorded by expert pulmonary radiologists. The prediction of diagnostic characteristics promises to automatically annotate medical images with medically meaningful descriptions directly usable for indexing and retrieving in content-based image retrieval (CBIR) and assisting in computer aided diagnosis (CADx). Successful results in predicting texture characteristics will be contrasted with less successful results for boundary shapes. The two primary differences in agreement between radiologists will be discussed; the first concerns agreement about the existence of a nodule, while the second considers the variability in diagnostic ratings among radiologists who agree on the presence of a nodule.

1. Introduction

Increased utilization of diagnostic imaging imposes a growing demand for improving the efficiency of radiology departments. Picture Archiving and Communication Systems (PACS) technology helps manage the operational demands but new technologies are needed to address the diagnostic demand. Recent research [34] and commercial success [9] with computer-aided detection (CAD) and diagnosis (CADx) have demonstrated the efficacy of using CAD(x) as a second reader to augment the diagnostic process. While increasingly accurate in detection and diagnosis, CAD(x) rarely offers supporting guidance about the rationale for the diagnosis or supplies descriptive annotations about medically meaningful diagnostic characteristics [Doi]. Two areas of research attempt to address this deficiency, one focuses on standardizing diagnostic terminology with notable success in the mammography community [1] while the other, semantic mapping, focuses on automatically labeling images with usable, medically meaningful, diagnostic descriptors. Semantic mapping attempts to extract image features and build predictive models of diagnostic characteristics for labeling images.

Semantic mapping promises to add clinically relevant diagnostic evidence to support the medical decision maker using CAD(x)-based tools as well as content-based retrieval (CBIR) systems [4]. Case-based reasoning is a major aim of CBIR and approaches include the display of relevant images based entirely upon the image contents [11]. Extending this approach to automatically extract clinically relevant diagnostic characteristics will allow for retrieval of past cases with similar characteristics [4].

The semantic mapping approach discussed in this paper follows the methodology introduced by Raicu [31] of combining measurements of low-level image features and radiologists' outlines of pulmonary nodules to predict each radiologists' ratings for a set of diagnostic characteristics such as texture, margin, subtlety, spiculation, etc. These characteristics reflect image-based radiologist-interpreted image features useful for diagnostic decision making. Using radiologist-defined ratings for diagnostic characteristics as the target for classification and prediction models, this approach aims to bridge the semantic gap between low-level image features and medically meaningful descriptions of images. The predicted ratings for diagnostic characteristics promise to enhance medical decision making both with the CADx framework as well as content-based image retrieval. For CADx, the ratings of diagnostic characteristics offer evidence for radiologists in decision making. For CBIR, the ratings will enhance indexing and retrieval relevance by offering queries based upon diagnostic criteria and providing additional diagnostic evidence for retrieved images.

In this paper, we extend previous work on semantic mapping of image features to predict radiologists' ratings of diagnostic characteristics of pulmonary nodules by focusing on building boundary-based shape descriptor models directly from radiologist-drawn outlines on the assumption that 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. After applying the shape feature method, two predictive methods are applied but fail to predict the radiologist ratings for three (3) shape characteristics. In discussing the results we show that the variability in their

ratings for shape characteristics is much higher than for other diagnostics characteristics such as texture which have been successfully mapped by Raicu et al. [31]. We conclude with strategies for combining the radiologists' outlines and ratings in an attempt to produce a usable shape descriptor model.

This work uses a publicly available dataset: The Lung Imaging Database Consortium (LIDC): http://imaging.cancer.gov/programsandresources/InformationSystems/LIDC. The LIDC contain images, outlines, and radiologist ratings of pulmonary nodules for use in developing computer aided tools for nodule identification and characterization.

2. Related Work

Evaluation remains a significant challenge to the development of semantic mapping in medical imaging. The challenges reside in the variability in diagnostic opinions among radiologists [3] and the potential differences in medical meaning of image characteristics. Another major challenge results from the lack of standardized lexicons for medical image diagnostics [16] and the potential disagreement on the proper terminology and the usage of common terminology. These challenges present potential obstacles to the creation and usage of radiologist annotated databases for discovering valid and consistent mappings from raw image features to medically meaningful semantic labels.

2.1 Evaluation Methodology of CAD, CADx, and Semantic Mapping

The evaluation methods of CAD and CADx follow the basic methodology underpinning receiver operating characteristics (ROC), including two states of truth (often labeled positive and negative): lesion present or absent for CAD and disease or not in CADx, and diagnostic-accuracy measures such as sensitivity, specificity, and ROC curves [Metz]. Sensitivity measures the performance of finding a disease when the disease is present and specificity measures the performance of rejecting the finding of a disease when the disease is not present. A false positive is the detection of a disease which is not present and the probability of this occurring is reported as the true positive fraction (TPF); while a false negative is the failure to find an existing disease and its probability is reported as the false positive fraction (FPF). Sensitivity is equivalent to TPF while FPF is equal to 1 - Specificity. ROC curves plot TPF against FPF (or Sensitivity vs 1 - Specificity) and create a curve based upon different settings of the observer's decision criteria or critical confidence level. The area under the ROC curve (A_z) is often used as single index to reflect the overall performance. Studies evaluating CAD(x) systems typically measure the change in radiologist performance (ROC A_z) between readings performed alone and readings performed with CAD(x).

Both CAD and CADx use ROC analysis, but special considerations are often necessary to measure the performance of detection (CAD) when multiple lesions are possible or the location of lesions is important. ROC analysis is suitable for CAD performance when the task is to decide only if the image contains a lesion but not its location. This might include a single or set of images or a region of interest within an image [28]. If the CAD task is to decide not only the presence but also the location, then Localization ROC (LROC) analysis is applicable. For multiple lesions, several methods have been proposed based upon the Free-response operation characteristic (FROC) [7], but many studies choose to report sensitivity and false positives per image (FPI) or per case.

ROC analysis is restricted to two-class data and is not applicable to the ordinal category [31] or interval [26] data used in semantic mapping (subjective similarity) studies. The methods to evaluate performance of semantic mapping will depend upon the type of data but will primarily use overall and per characteristic classification accuracy (e.g. hit ratio: the percentage of instances correctly classified). Since the aim of semantic mapping includes not only labeling images with meaningful diagnostic descriptions but also using those mapping models to retrieve similar images, the evaluation of semantic mapping for CBIR will report information retrieval statistics such as precision and recall [8]. Recall corresponds to the ROC concept of sensitivity and measures the probability that relevant images are retrieved by a query. Precision does not correspond to a ROC concept but measures the positive predictive value of a query or the percentage of retrieved images which are relevant to a query.

2.2 Inter-observer Agreement

Agreement among radiologists remains an active area of research both as a subject of understanding differences between radiologists as well as studying the effects of different imaging modalities, different types or experience of radiologists, the comparison of CAD(x) and radiologist performance, and the effect of CAD(x) as a second reader on radiologist performance. The two primary methods of measuring inter-observer agreement are the Kappa statistics [17] and ROC [23]. Many studies consider only binary categories (disease or not) and report Cohen's Kappa statistics using either a pair of observers or average the Kappa scores for multiple observers [17]. For studies of ranked, multi-category findings such as disease severity {absent, minimal, moderate, or severe}, a

weighted (often quadratic) Kappa method is used [12]. Though widely used, Kappa statistics vary according to disease prevalence and are unsuitable for comparative studies [17]. When ground truth is known, each radiologist's performance can be measured by A_z (area under the ROC curve) scores and comparison made using ANOVA methods [24]. 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. [26] 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 ratings in the LIDC for spiculation, etc. In one of the few studies examining radiologists' ratings for image-based diagnostics features such as spiculation, Nakamura et al. [27] 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.

2.3 Semantic Mapping

CAD(x) can be considered a diagnostic mapping from image features to detection or diagnosis and is increasingly accepted as a successful diagnostic adjunct [9,34]. Semantic mapping attempts to insert 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. Therefore, we can define the semantic mappings as a two step process with the first step representing a mapping from image features to diagnostic characteristics (subjective features [20,27]) and the second step a mapping from diagnostic characteristics (subjective features) to overall diagnosis.

CBIR aims to retrieve images relevant to the query and semantic mapping offers an additional indexing strategy to include medically meaningful, semantically-mapped, image descriptions for filtering retrievals and annotating retrieved images. However, 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 [20]. 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 [27]. 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.

2.4 Reporting Terminology

Burns et al. demonstrated a lack of consistent radiologist reporting of lung nodule characteristics and recommended adoption of a standardized criteria [5]. Efforts to standardize reporting in pulmonary nodules are underway as part of the development of a general purpose radiology lexicon [18].

Within the mammography community, the BiRads lexicon has been developed to standardize and reduce confusing in imaging interpretations [1]. Recent analysis indicates the effectiveness of this approach. Lazarus evaluated inter-observer variability in BiRads reporting and concluded that radiologists showed good agreement (Cohen's Kappa), the ratings had a high predictive value, and the results validate the use of the US BiRads lexicon [19].

2.5 Segmentation

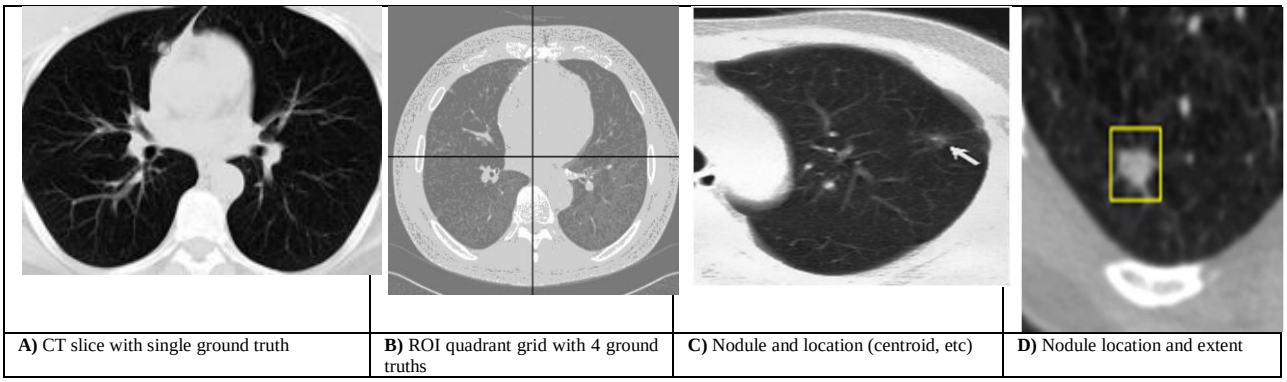
Accurate segmentation of pulmonary nodules plays an important role in measuring shape features [34] as well as ensuring that comparison of foreground nodule and background imagery is preserved. In this study, the radiologist drawn outlines serve as a segmentation template: thus each nodule has up to four (4) outlines. Inter-observer variability in the drawing of outlines around pulmonary nodules was studied directly by Meyer at al. [22]. The Meyer study compared the relative performance of six radiologists using three (3) outlining methods to define the spatial extent of 23 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. Examining the initial LIDC dataset, Opfer [30] estimated a 50% variability in the regions selected by multiple radiologists for the same nodule. Reeves et al. investigated the choice of different metrics for estimating the size of pulmonary nodules on upcoming LIDC data and concluded that a very high inter-observer variation exists for four (4) size metrics [32]. These reports indicate that the significant variability in outlines presents a significant challenge to the characterization of the pulmonary nodule with respect to the background of the image.

2.6 Forms of Diagnostic Training Imagery and Ground Truth

Ground truth training data ranges from coarse, binary labeled data to detailed, annotated outlines of suspect regions of interest and includes either expert-defined diagnostic opinion or pathology-confirmed findings. Typically, this data is collected from case-histories at a single institution and labeled by a single or consensus of domain experts; mostly these datasets remain private or not publically available. Few publically available datasets with annotated diagnostic information are available, but include mammography data [14], confirmed by pathology and annotated by a set of radiologists, and the pulmonary nodule database (LIDC). A more general collection of medical images using various modalities and containing numerous diseases and anatomies is available at http://www.casimage.com/ and described by [25]. The Reference Image Database to Evaluate Response (RIDER) is another NCI Cancer Imaging Program effort similar to the LIDC but targeted at serial image histories of lung cancer patient conditions as a method for measuring the image response to cancer treatment and is located at http://www.nibib.nih.gov/Research/Resources/ImageClinData.

Semantic mapping closely matches the forms of image data used by CAD(x) but with the addition of medically meaningful, descriptive information. The image data can be classified according to its specificity to the suspected nodules, either by limiting the region of interest or by using localization marking techniques such as coordinates, bounding boxes, or detailed outlines. Isolation of the suspected region is the major factor driving the need for limiting the region of interest to ensure feature extraction obtains as little of the non-diagnostic background as possible when measuring the features of the diagnostic foreground of the suspected region of interest. The following discussion uses CT-based pulmonary nodule imagery as an example, but applies to other modalities such as posterior-anterior chest x-rays, MRI, and sonography or other anatomies such as mammography.

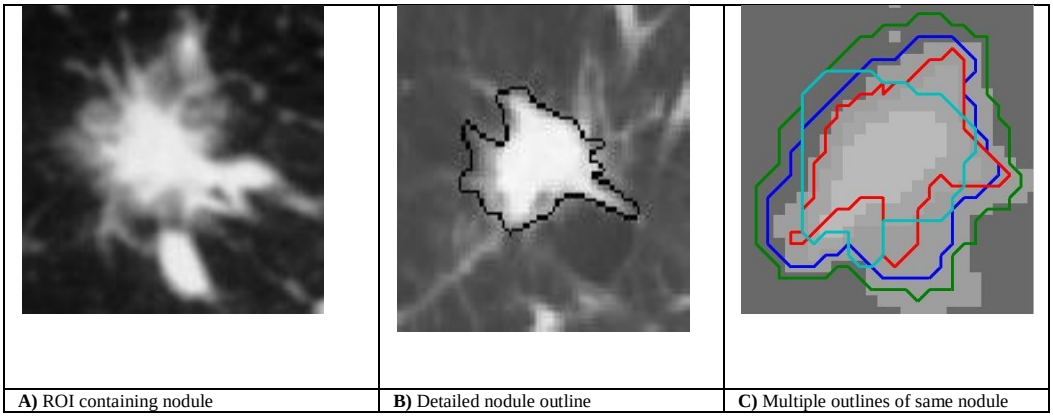
Images without localization of the suspect region provide limited benefit to training and evaluating semantic mapping models, since they offer no method to identify agreement between algorithm and ground truth. Some CAD/CADx studies use global images as the unit of analysis with ground truth labels for a patient case (set of slices) or single slice (Figure 1A). The ground truth labels indicate lesion presence or absence, diagnosis of malignant or benign, and the location, if any, of nodules is not known. Since multiple nodules could be present within a slice, a more localized region of interest approach [28] uses subdivided slices (e.g. quadrants) (Figure 1B) as the unit of analysis with a single ground truth without location for the entire ROI. These designs are motivated by the lack of localization support in traditional ROC analysis [23]. Using other localization evaluation methodologies (LROC/AFROC or false positives per image (FPI)), studies often use datasets with more specific nodule location (centroids) (Figure 1C) or location and extent of nodules (bounding boxes) (Figure 1D).



A) CT slice with single ground truth	**B)** ROI quadrant grid with 4 ground truths	**C)** Nodule and location (centroid, etc)	**D)** Nodule location and extent

Figure 1: Global image ground truth with varying levels of localization

Some studies focus on the nodule as the unit of analysis and create images containing a single nodule or detailed outline(s) of nodule, perhaps with known diagnosis. Several studies on subjective similarity [20,26] present radiologists with sets of images of isolated lesions for rating similarity between images such as Figure 2A. Use of detailed outlines (Figure 2B) is primarily used in segmentation studies, though use of an outline or segmentation is essential for accurate characterization of the nodules [27]. The LIDC and DDSM datasets for lung and mammography contain one or multiple outlines of the same nodule. Figure 2C from the LIDC shows four radiologist-drawn outlines of the same pulmonary nodule. Currently, the LIDC does not contain pathology confirmed diagnoses, but contains radiologist characterizations of nodules in terms of presence/absence, size, outline, and ratings for clinically significant characteristics such as spiculation, subtletly, and texture (among others).



A) ROI containing nodule	**B)** Detailed nodule outline	**C)** Multiple outlines of same nodule

Figure 2: Local ground truth images with increasingly detailed localization.

2.7 Shape of Pulmonary Nodules

The shape of pulmonary nodules is used by radiologists in clinical practice and the appearance of spiculation along the boundary of nodules is a strong indication of malignancy [37]. Shape measurement algorithms play an important role in detection and diagnosis in pulmonary CAD(x) [33] and formed the basis for the LIDC to include the diagnostic shape characteristics of spiculation, lobulation, and sphericity [LIDC]. Giger et al. employed geometric methods (effective diameter and degree of circularity) to detect suspicious nodules in chest x-rays [13]. Nakamura et al. used the root-mean-square and first moment of a Fourier transformation of the nodule outline and the radial gradient index (RGI); the RGI was chosen to measure nodule spiculation [27]. 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 [31]. Their work showed promising results for predicting the LIDC diagnostic characteristics for texture, sublety, and malignancy, but had less success in predicting shape characteristics such as spiculation, lobulation, and sphericity, and indicated that additional shape features are necessary. The work presented in this paper attempts to address this issue by applying specific boundary-based shape features directly on radiologist-drawn outlines.

This paper introduces the Radial Normal Indexing (RNI) method (an adaption of the RGI technique) and applies the RNI to the radiologists drawn outlines primarily to measure spiculation and predict the radiologists ratings of spiculation. The RNI is also applied to measure two other shape characteristics: lobulation, and sphericity.

Additionally, based upon the effectiveness of Fourier boundary shape measurements in [27], the Fourier shape descriptor technique is applied to the outlines to assess its performance in predicting the radiologists ratings for the shape characteristics. The implementation of the Fourier shape descriptor is described by Svoboda [35].

3. Materials and Methods

3.1 Data Set

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.

3.2 Methods

Earlier work [31] showed promising results in predicting non-shape characteristics such as texture and malignancy and motivated this study to focus on boundary-based shape descriptors found in the literature [15,27] to attempt to predict the shape-based characteristics: spiculation, lobulation, and sphericity. A spiculation-specific boundary shape descriptor, radial gradient index (RGI), was developed for mammography [15] and used the direction of image gradient along the segmented boundary of the mass. [27] showed the effectiveness of RGI features in classifying the diagnosis of pulmonary nodules. In both papers, the RGI was applied to find the direction perpendicular to the outline (the direction of maximum slope of the image gradient (similar to Canny edge detection [6]) along the segmentation of the mass or nodule. In this study, the outlines are provided by radiologists and assumed to represent the radiologists' interpretation of the "segmentation" of the nodule and an approach is developed to measure the perpendiculars along the outlines.

The boundaries of the nodules, represented by each radiologists outline, were measured with the RNI and Fourier shape descriptor. The features extracted from the RNI included a 16-bin histogram and two measures of central tendency, the standard deviation of the histogram and the Full Width of the Half Maximum (FWHM) (described below). The choice of number of bins (=16) was informed by [15] and represents a tradeoff between angular resolution in measuring variation along the boundary and a goal of limiting the number of features extracted. Similarly, fifteen (15) Fourier descriptors (Fourier coefficients) were chosen per guidance [35].

Two approaches for predicting the ordinal categorical radiologist ratings from raw shape features (RNI and Fourier descriptors) were applied: logistic regression and decision trees (both implemented in Weka [36]). Both methods are appropriate for categorical response variables but differ in their use of the ordering inherent in the ratings. Logistic regression uses the ordering information while decision trees can separate non-linear data. Feature selection is automatic in decision trees but requires additional support in logistic regression. Both methods can assign probabilities to their predictions. Model building and performance uses a 10-fold cross validation technique. The prediction of multiple categories (radiologists' ratings from 1 to 5 for most characteristics) is most appropriately measured using overall accuracy in prediction rather than reporting sensitivity and specificity scores for each category of ratings. For the results in this paper, only overall accuracy will be reported where accuracy is the number of correct predictions over the number of predictions.

These shape measurements were applied to the full LIDC dataset then three (3) progressively smaller subsets each representing an increasingly level of agreement about the existence of a nodule. This level of agreement partitioning was suggested by Ochs et al. for reporting LIDC results where reader agreement can be influenced by outliers [29]. The predictive performance of RNI and Fourier shape descriptors are reported by level of agreement to identify differences or trends.

Inter-observer agreement cannot be measured using traditional methods and this paper adopts a raw disagreement scoring system where the absolute differences between radiologists per characteristic are measured and the mean difference represents the disagreement for the particular nodule. These differences are only computed for the radiologists who rated the nodule and only computed for nodules marked by at least two radiologists. These mean differences per nodule can be accumulated in a histogram to represent the level of disagreement for a specific diagnostic characteristic.

3.3 Radial Normal Index

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. This captures an angular variation along the boundary similar to the RGI method which uses the direction of the image gradient along the boundary. As illustrated in Figure 3, 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.

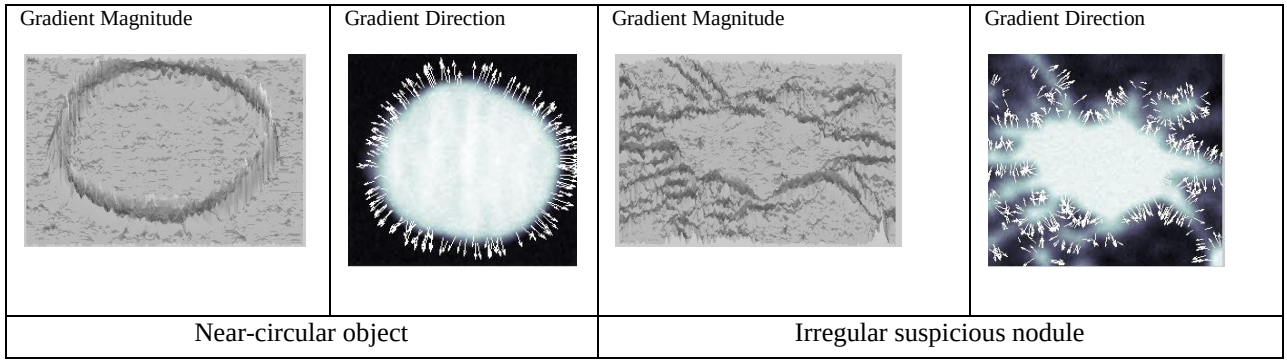


Figure 3: Gradient magnitude and direction along boundary of regular and irregular objects

3.3.1 Computing the Radial Normal Index

The method of the Radial Normal Index is illustrated in Figure 4. 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 which represents the shape of that object. The angular difference distribution is represented using a histogram.

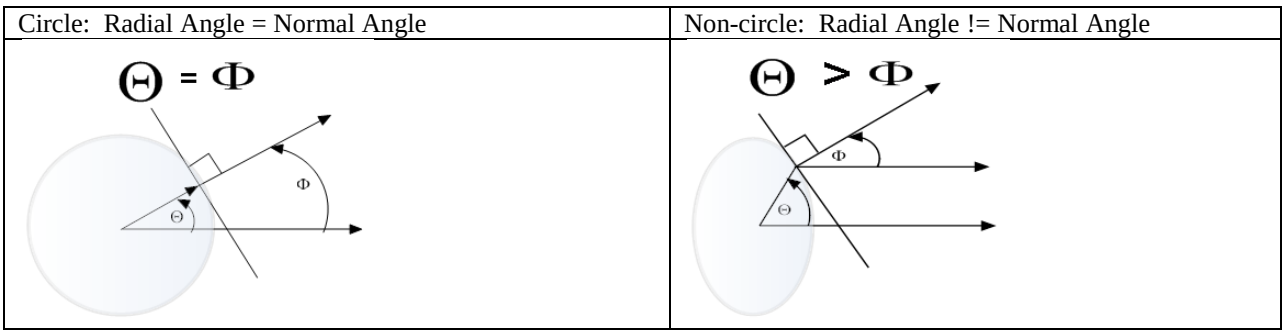


Figure 4: RNI method for computing difference between radial angle (θ) - direction of vector from center of object to point on perimeter – and normal angle (φ) -direction of vector normal to object at point on perimeter

3.3.2 Computing Radial Normal Index Features

After accumulating the differences between the radial and normal angles in a distribution histogram, the radial normal index is computed as a measure of dispersion of this distribution. Though standard deviation is the canonical dispersion statistic, the use of full-width of the maximum height (FWMH) was suggested for spiculation measurements in the radial gradient indexing method [15]. Figure 5 illustrates the computation of the FWHM using a line (versus bar) graph representation of the histogram.

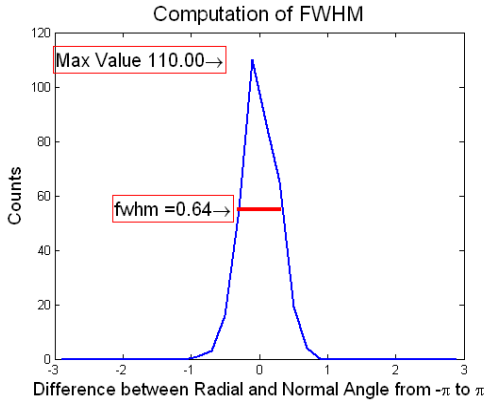


Figure 5: An example of FWHM computation

3.4 RNI Results for Exemplar Objects

Results for near-circular, reference type object, and spiculated nodule illustrate how RNI captures the variability in the shape of an object. The near-circular object has a small RNI and a narrow distribution while the spiculated object has a much larger RNI as reflected in its wider distribution. As shown in Figure 6, the standard deviation also tracks the differences in the dispersion of the angle differences and will be investigated along with the FWHM to assess their performance in describing shape.

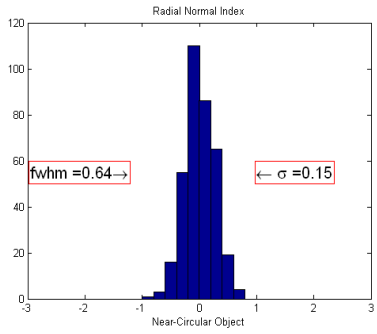
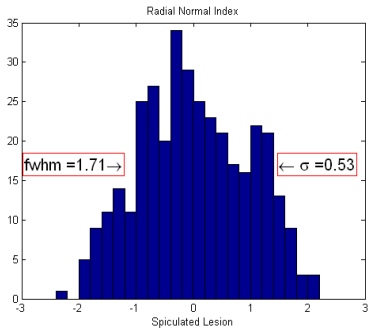
RNI angle difference distribution and measures of dispersion: standard deviation and FWHM	
	
near-circular object	irregular suspicious nodule

Figure 6: RNI results for representative shapes

3.5 RNI Captures Increased Angular Variability along Radiologist-Drawn Contours

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 7, 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.

Less Spiculated (Rating 4) Nodule FWHM = 1.05	More Spiculated (Rating 2) Nodule FWHM = 2.17

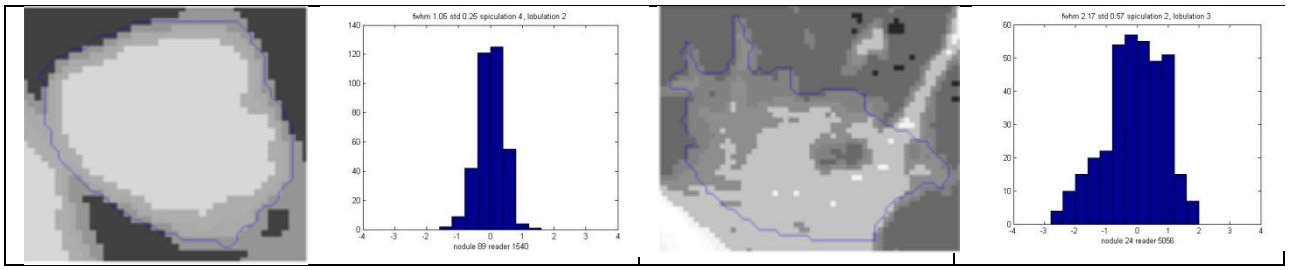


Figure 7: RNI correlates well with spiculation for selected images

4. Results

Table 1 illustrates the poor the performance of RNI features for predicting radiologist ratings for spiculation, lobulation, and sphericity using all nodules and the three (3) progressive levels of agreement [Ochs] among radiologists on the presence/absence of a particular nodule. Overall the predictions are poor (less than 50%) and a couple (~24%) are slightly better than guessing (20%). With respect to agreement, the performance is slightly better for the at-least-3 agreement. Sphericity is somewhat better predicted than the other characteristics. Logistic regression is somewhat better than decision trees and might perform better due to its use of the order in the ratings.

Table 1: Poor performance of RNI for predicting radiologists' diagnostic characteristics for shape

Prediction of Shape Characteristics Based Upon RNI *(FWHM, StdDev, and Bin Counts)* Agreement means # Radiologists' Outlines *(Nodule existence)*								
Characteristic	All radiologists (1,2,3,or 4) 147 Nodules		>= 2 radiologists 105 Nodules		>= 3 radiologists 74 Nodules		All 4 radiologists 35 Nodules	
	Decision Tree	*Logistic Reg*	*Decision Tree*	*Logistic Reg*	*Decision Tree*	*Logistic Reg*	*Decision Tree*	*Logistic Reg*
Spiculation	33.33%	42.35%	32.10 %	40.74 %	34.73 %	45.03 %	25.00 %	26.61 %
Lobulation	30.87%	39.34%	24.07 %	37.96 %	24.81 %	37.79 %	29.84 %	29.84 %
Sphericity	34.97%	36.88%	33.64 %	42.28 %	38.17 %	44.66 %	43.55 %	42.74 %

Table 2 shows the results of the Fourier descriptors which also poorly predict the radiologists' ratings of shape characteristics with performances very similar to the RNI and some of the same patterns where logistic regression outperforms decision trees and sphericity is better predicted.

Table 2: Poor performance of Fourier descriptors for predicting radiologists' diagnostic characteristics for shape

Prediction of Shape Characteristics Based Upon Fourier Descriptors Agreement means # Radiologists' Outlines *(Nodule existence)*								
Characteristic	All radiologists (1,2,3,or 4) 147 Nodules		>= 2 radiologists 105 Nodules		>= 3 radiologists 74 Nodules		All 4 radiologists 35 Nodules	
	Decision Tree	*Logistic Reg*	*Decision Tree*	*Logistic Reg*	*Decision Tree*	*Logistic Reg*	*Decision Tree*	*Logistic Reg*
Spiculation	33.88%	41.80%	35.60 %	41.79 %	37.54 %	44.82 %	27.77 %	25.00 %
Lobulation	28.96%	33.3%	24.14 %	33.43 %	27.58 %	33.33 %	25.69 %	32.63 %
Sphericity	34.70%	39.89%	32.19 %	40.55 %	37.93 %	41.76 %	40.27 %	38.88 %

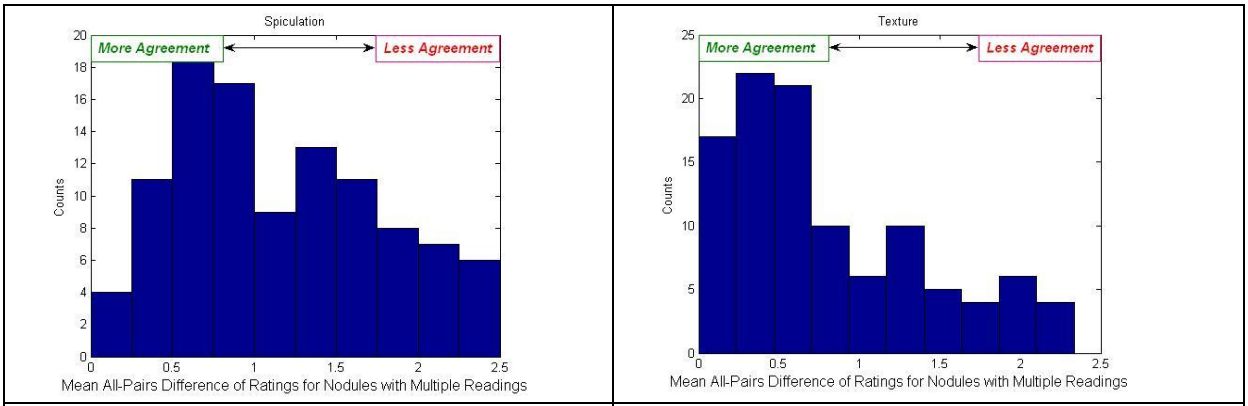
5. Discussion of Results

Two primary causes can result in the poor performance of the boundary shape descriptors in predicting radiologist shape characteristic ratings. Either the shape descriptors (or the selected features) cannot sufficiently measure the image features used by radiologists in rating nodule shape, or there exists too much disagreement among the radiologists' ratings of these shape characteristics. While the verification of the RNI techniques employed only a small number of sample outlines and more research is needed to confirm the approach, the methodology closely matches the RGI technique successfully used in both the pulmonary nodule [Nakamura] and

mammography [15] communities. If the radiologists cannot agree on ratings, then statistical and pattern recognition models will not perform well in discriminating between rating categories. Earlier work on semantic mapping of LIDC diagnostic characteristics [31] showed much better performance for texture and subtlety characteristics than for the shape characteristics in this paper. In addition to poor performance, their decision trees showed other signs of poor generalization such as much more complex decision trees for shape characteristics than for texture and the feature selection process for decision trees included many non-shape features such as texture descriptors.

The second question on the radiologist disagreement is measureable in terms of raw scoring differences rather than traditional intra- and interobserver agreement using standard methods for observer studies: (Cohen's kappa statistic, Spearman's correlation coefficient, Interclass coefficients, or ROC) where the identity of the observer is known and consistent throughout the study. In the LIDC data, the observer identity is unknown and might differ between cases, thus the differences between ratings per nodule is the unit of interest and the mean difference between ratings is the primary metric. This metric is computed by measuring the absolute difference between all pairs of readers then dividing by the number of pairs. For example, three readers with ratings of {1,2,4} will have three (3) pairs of differences {|1-2|, |1-4|, |2-4|} for total difference of 6 and mean of 2, and a set of four readers will have six (6) pairs of differences. This measures the disagreement on a single characteristic for a single nodule, while a histogram of the accumulated nodule-level disagreements represents the overall disagreement for the characteristic. Using a normalized histogram, a single index of overall disagreement for a characteristic can be computed using a weighted summation of bin counts where the bin positions form the weights. For example, the range of most LIDC characteristics (ratings 1 to 5) is four (4) and the bin-positions (=weights) are {0, 1, 2, 3}; note that using a weight of 0 removes all counts where the disagreement is zero (0). This produces a raw disagreement metric for characteristics which ranges from 0 (full agreement) to 1 (total disagreement). Selection of a few extreme examples illustrate the extent of disagreement for spiculation: one nodule received four ratings of {1,1,5,5} with mean disagreement of 2.67 while another nodule received nearly identical outlines by two radiologists but nearly opposite ratings {2,5} even though only one vertex on the outline differed and only be a couple of pixels.

Comparing the histograms in Figure 8 illustrates larger agreements on texture (high counts on left side of distribution) while large disagreements occur on spiculation (high counts on right side). Using the bin-weighted index method for overall characteristic agreement, texture receives a disagreement score of 0.19 while spiculation receives a much higher score of 0.29 (Table 3). This corresponds with the results shown in [31] who showed promising prediction for texture, but less success at predicting spiculation. As illustrated, the lack of agreement among radiologists is a major factor in the inability of semantic mapping algorithms to predict their ratings.



A) Spiculation B) Texture

Figure 8: Radiologist disagreement using all-pairs differences

Box plot analysis illustrates direct comparisons of radiologist disagreement (variability) for all nine (9) diagnostic characteristics as shown in Figure 9. Boxes closer to the origin indicate lower disagreement while shorter boxes indicate less variance among the ratings. The more predictable texture and subtlety characteristics appear as a shorter box near the origin, while spiculation and lobulation have much longer boxes farther from the origin. The lack of boxes for internal structure and calcification result from their near total agreement. The cause of the agreement for internal structure is uncertain, while the cause for agreement on calcification is its high negative predictive value where the presence of calcium in a nodule is a strong indication than the nodule is benign [37].

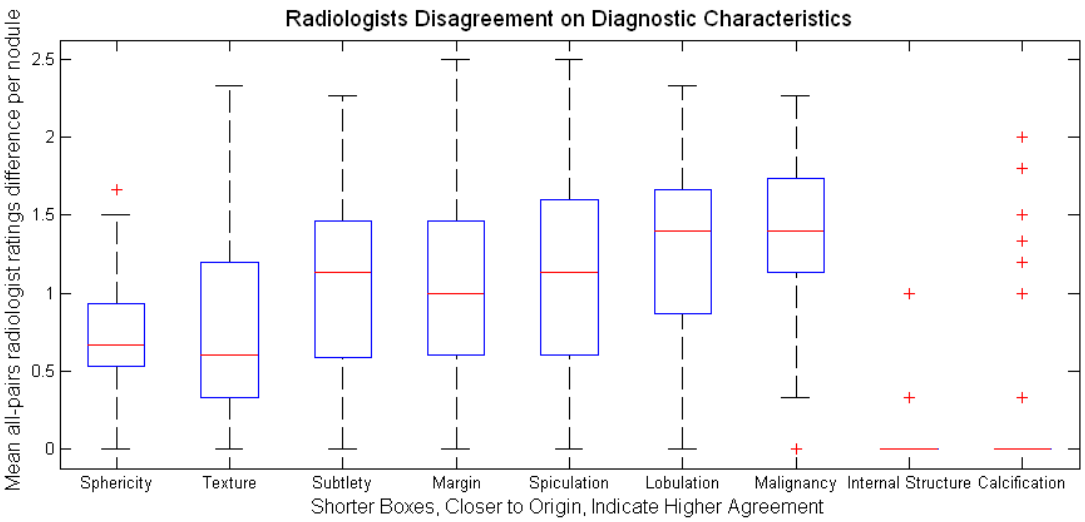


Figure 9: Radiologists' disagreement on diagnostic characteristics for all nodules with 2 or more readers; the major elements are the height of median disagreement (bar in middle of box), the spread of disagreement between the 25 and 75th percentile (the box), and the height of the 25 percentile (bottom of box).

Table 3: Characteristic Disagreement Indexes

Characteristic	**Disagreement**
Internal structure	0.004
Calcification	0.016
Sphericity	0.184
Texture	0.192
Subtlety	0.264
Margin	0.269
Spiculation	0.287
Lobulation	0.318
Malignancy	0.354

6. Conclusions and Future Work

In addressing the boundary shape feature questions posed by previous work [31], this study shows that the inability to predict radiologists' shape characteristics in the LIDC is due more to the variability in radiologists' ratings rather than selection of shape features (either geometric or boundary) of the radiologist defined outlines. Given the persistence of this diversity of opinion from the original 23 LIDC cases studies examined by Raicu to the current 85 cases, the future work must focus on managing this diversity of ratings.

In these studies, the method has attempted to predict individual radiologist's ratings based only upon their definition of the boundary of the nodule as represented by their own outlines. Given the poor success of specific predictions, a method of combining ratings and outlines might yield better results. Recent research on analyzing multiple outlines suggests two similar approaches. One method, developed at the National Institutes of Health (NIH) and reported by [21], is the Simultaneous Truth and Performance Level Estimation (STAPLE). This method offers a method to analyze intra- and inter-expert variability in drawing segmentation outlines as well as generating segmentation maps from the combined outlines. Without knowledge of the radiologists (each reading is blinded in the final dataset), there is no clear method for applying the STAPLE algorithm to the LIDC dataset. Another approach, proposed by the LIDC, uses probability maps, p-maps, as the probability that a pixel is a member of the nodule [22] and offers a method of defining an outlines containing a range of pixels with varied likelihoods of membership. This approach is similar to various methods of overlap, intersection, union, and other similarity

metrics. Future research for predicting semantic descriptors can attempt a range of likelihoods to determine whether this improves the predictive performance of shape descriptors when the nodule boundary is represented by only one outline.

After combining the outlines, research on combining radiologists' ratings is necessary. When analyzing the performance of radiologists' subjective similarity ratings between unknown and known images, [26] found the inter-observer variability too high and chose to use the average ratings. Averaging is not appropriate for the ordinal LIDC data and median methods represent an alternative. Mode represents another voting approach where the category most chosen is the representative. Another option to explore is based upon the LIDC instructions for recording the shape descriptors. For lobulation and spiculation, the LIDC provided guidance for ratings of 1 and 5 but left blank the ratings for 2-4; while for sphericity, the LIDC labels only 1, 3, and 5 without guidance for 2 and 4. This suggests future experimentation with binary methods for lobulation and spiculation and trinary methods for sphericity.

Measured disagreement on individual pulmonary nodules allows for the design of training and testing datasets according to the confidence of the combined radiologists' results. Future experiments in database design and alternative machine learning strategies will attempt to both the confidence of low disagreement nodules and the uncertainty of high disagreement nodules for improving the overall performance of the semantic mapping approach to predicting diagnostic characteristics for applications in CADx and CBIR.

References

1. American College of Radiology (2003). *Breast Imaging Reporting and Data System (BI-RADS), 4th ed.* Reston, VA: American College of Radiology.
2. Anthony P. Reeves, Alberto M. Biancardi, et al. (2007). The Lung Image Database Consortium (LIDC): A Comparison of Different Size Metrics for Pulmonary Nodule Measurements. *Academic Radiology*. 14 (12):1475-85.
3. 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.
4. 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, pp. 662-669(8).
5. Burns J, Haramati LB, Whitney K, & Zelefsky MN. (2004). Consistency of reporting basic characteristics of lung nodules and masses on computed tomography. *Academic Radiology*,1:233-7.
6. Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE PAMI*, 8(6).
7. Chakraborty D. (2002). Statistical power in observer performance studies: comparison of the receiver operating characteristic and free-response methods in tasks involving localization. *Academic Radiology*, 9:147–156
8. Datteri, R., Raicu, D., & Furst, J. (2008). Local versus Global Texture Analysis for Lung Nodule Image Retrieval, *Proceedings of SPIE Medical Imaging*.
9. Doi K., (2005). Current status and future potential of computer-aided diagnosis in medical imaging, *The British Journal of Radiology*.
10. Dreyer K. (2005). The Alchemy of Data Mining, *Imaging Economics*.
11. El-Naqa, I., Yongyi Y., Galatsanos, N., Nishikawa, R., Wernick, M., (2004). A similarity learning approach to content-based image retrieval: application to digital mammography**.** *IEEE Transactions on Medical Imaging*, 1233 – 1244.
12. Fleiss, J. (1981). *Statistical methods for rates and proportions* 2nd Ed. Wiley 212-236.
13. 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.
14. Heath, M., Bowyer, K., Kopans, D., Moore, R. & Kegelmeyer, W (2001). The Digital Database for Screening Mammography, *Proceedings of the Fifth International Workshop on Digital Mammography* 212-218.
15. 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.
16. Kahn CE, Channin DS, & Rubin DL. (2006), An Ontology for PACS Integration, *Journal of Digital Imaging*, 19(4): 316-327.
17. Kundel, H. & Polansky, M. (2003). Measurement of Observer Agreement, *Radiology* 228-303.
18. Langlotz, C.P. (2006). RadLex: A New Method for Indexing Online Educational Materials, *RadioGraphics*, 26::1595-1597.

19. Lazarus, E., Mainiero, M.,Schepps, B., Koelliker, S., & Livingston, L. (2006). BI-RADS Lexicon for US and Mammography: Interobserver Variability and Positive Predictive Value, *Radiology* 239:385-391.
20. 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.
21. Lotenberg, S., Gordon, S., Long, R., Antani, S., Jeronimo, J., & Greenspan, H. (2007). Automatic Evaluation of Uterine Cervix Segmentations. *Proceedings of SPIE Medical Imaging*.
22. 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.
23. Metz CE (2008). ROC analysis in medical imaging: a tutorial review of the literature. *Radiological Physics and Technology*: 2-12.
24. 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.*
25. Müller, H., Rosset, A., Vallée, J., Terrier, F., & Geissbuhler A. (2004). A reference data set for the evaluation of medical image retrieval systems. *Computerized Medical Imaging and Graphics*, Volume 28, Issue 6, Pages 295-305.
26. 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.
27. 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.
28. Obuchowski N., Lieber M., & Powell K. (2000). Data analysis for detection and localization of multiple abnormalities with application to mammography. *Academic Radiology*, 7:516-525.
29. 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.*
30. 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.*
31. 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*.
32. 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*, 14(12):1475-1485.
33. Sahiner B., L. Hadjiiski, Chan, H. Shi, J.,Way, T., Cascade, P., Kazerooni, E., Zhou, C, & Wei, J (2007). The Effect of Nodule Segmentation on the Accuracy of Computerized Lung Nodule Detection on CT scans: Comparison on a Data Set Annotated by Multiple Radiologists, *Proceedings of SPIE Medical Imaging*.
34. Sluimer I., Schilham A., Prokop M., & Ginneken B. (2006). Computer Analysis of Computed Tomography Scans of the Lung: A Survey, *IEEE Transactions on Medical Imaging*, 25(4).
35. Svoboda T., Kybic J., Hlavac V. (2007) Image Processing, Analysis, and Machine Vision, A MATLAB Companion, Thomson Learning, Toronto.
36. Witten IH & Frank E, (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition, Morgan Kaufmann.
37. Wormanns, D. & Diederich, S., (2004). Characterization of small pulmonary nodules by CT, *European Radiology*, pp 1380-1391.