Exploiting Variance in Radiologist-Defined Truth: Opportunities from the LIDC

William H. Horsthemke
Daniela S. Raicu
DePaul University
School of Computer Science
Chicago, IL
horsthemke@acm.com
dstan@cti.depaul.edu

Abstract

New opportunities for research in computer aided diagnosis (CAD) of pulmonary nodules resulted from the emergence of publicly available clinical case studies, containing expert radiologist-defined truth. Access to high-quality labeled data affords non-clinical image analysis groups the opportunity to pursue diagnostic imaging research. Historically, clinical research centers reported results based upon in-house datasets, not typically shared or publicly available. Questions about the generalization of methods to other clinical cases or the comparison of methods on the same cases have been often cited in the literature.

The lack of benchmarked results has led to a federally funded effort to provide a research resource by the National Cancer Institute Lung Image Database Consortium (LIDC) for the development, training, and evaluation of computer aided diagnosis (CAD) for pulmonary nodule detection and diagnosis. The database contains over 100 computed tomography (CT) scans containing over 800 nodules.

This paper introduces the basic process of CAD for pulmonary nodules, the issues of defining ground truth for nodules, the information newly available from the lung image database, and possible research strategies for contributing to the knowledge of the detection of pulmonary nodules.

1 Overview of Computer Aided Diagnosis of Pulmonary Nodules

Reported research on CAD of lung nodule follows a basic detection and classification approach. This approach begins with the detection of candidate pulmonary nodules by selecting (thresholding) the brighter pixels (possible nodules) from the darker background of the lung (air and parenchyma). Neighborhoods with numerous, brighter pixels are combined and isolated from the remaining, background image by various segmentation algorithms. Once segmented, the potential nodule candidates are measured and identified as nodules if they are similar to other known nodules but different from normal lung tissue. Various features of the candidate regions of interest are employed to find (extract) useful properties which can discriminate a "true" nodule from other "false" non-nodule regions of the lung such as scars, non-malignant nodules, vessels, airways, pleura, and artifacts. Using the measured properties of "true" nodules and "false" regions, a statistical or machine learning (classification) technique finds an appropriate combination and range of measurements to discriminate between "true" pulmonary nodules and
non-nodules. This method relies entirely upon accurately defined truth about whether a candidate is a "true" pulmonary nodule or a false, non-nodule. Datasets of nodules are collected, labeled with ground truth, and used in the training and testing of the classifier.

Truth (gold-standard, ground truth) is provided by radiologists when they confirm or reject the results of the CAD system. Rejection occurs in two forms: false positive, when the CAD identifies a pulmonary nodule that the radiologist does not confirm; and, false negative, when the radiologist discovers a pulmonary nodule missed by the CAD. These two error rates form the basis of CAD performance analysis. Sensitivity reports the detection rate of true nodules: did CAD detect all the nodules identified by the radiologist. Specificity shows rate of false positive detections: how often did CAD detect a nodule rejected by the radiologist. An alternate reporting method is often used to represent the CAD ability to reject false positive candidate nodules. False positives are reported as a count per scan (single image slice) or per case (patient). This count of false positives better represents the increased load placed upon the radiologist from the use of CAD as a decision support system.

2 Lung Image Database Consortium

The LIDC pulmonary nodule database [2] contains patient CT cases representing a diverse set of pulmonary nodules as well as normal cases without abnormalities or with non-cancerous focal anomalies. Four (4) radiologists, at different medical centers, read these cases and indicate the presence or absence of nodules. Small (less than 3mm) nodules receive only a location mark and larger nodules are marked with an outline. The radiologists rate these nodule according to nine (9) diagnostic criteria, describing shape, texture, and other visual characteristics. After reading a case, the radiologist submits their findings to the LIDC.

Once read by all four (4) medical centers (one radiologist reading per center), the marked and annotated patient cases are reviewed by all 4 radiologists who may choose to modify their markings, adopt the markings of other radiologists, or maintain their original diagnosis and labeling. No forced or encouraged consensus is suggested by the LIDC. The final LIDC dataset contains the final markings by all four (4) radiologists: location for small nodules, outlines for larger nodules, and nine (9) ranked scores for diagnostic characteristics: calcification, internal structure, lobulation, malignancy, margin, sphericity, spiculation, subtlety, and texture. These new descriptors and outlines expand the possible forms of truth from the binary true/false confirmation previously available.

3 Prior Work in CAD

Research in pulmonary nodule detection, segmentation, classification, and diagnosis continues in academic and industrial labs to improve each task of CAD. Research to improve detection remains active since only 80% of nodules are detected by CAD (with 3.8 false positives per case), though this exceeds the estimated 70% detection rate of radiologists [4]. Efforts to increase the sensitivity (detection rate) typically increases the false positive detection rate as demonstrated by [3] which reports 95% sensitivity but 6.9 false positives per case.

Research to improve specificity is often considered the "classification task" and includes methods to better segment or measure the nodules; or improve the classification by different statistical or machine learning methods. One segmentation approach was discussed in [1] where they investigate a "region melting technique" in 3D to better isolate and measure suspected nodules. [9] segmented nodules by minimizing the energy of radial basis functions while others have reported on active contours. Finding feature measurements that differ between nodules and
non-nodules improves the classification as shown by [6,7] which measured nodules by co-
ocurrence textures and shape. [10] suggest wavelet coefficients. Experiments with different
machine learning strategies have resulted in the massive artificial neural network approach by
[14]. Numerous other approaches attempt to improve on one or many steps of the CAD process.

4 Prior Work in LIDC Research

Several research reports are available but mainly focus on extending existing research to
LIDC datasets. [15] compare their segmentation to the union of the contours of the radiologists
marking the nodule. [9] reports sensitivity of 89% with 2 false positives per patient, while [12]
reports a CAD observer study where radiologists read cases alone, then with the CAD, and
report a sensitivity of 79% with almost 5 false positives per patient. Research specific to the
types of data available only through the LIDC has been performed by few researchers, most

[11] bridges the semantic gap between the conceptual descriptive data ranked by radiologists
and computer measurable image features of the actual nodules. The semantic mappings allow for
computer measurement and prediction of the radiologists' descriptions of the nodules. Much
work remains in this area due to the significant lack of agreement among the radiologists. Work
in this area may lead to a greater understanding of the differences in radiologist assessments of
pulmonary nodule diagnostic truth or assist in improving the agreement among radiologists.
Other inter-observer agreement research on the LIDC was performed by [8], comparing the
agreement among radiologists for varying size nodules.

5 Variation of Radiologists Outlines of Pulmonary Nodules

Radiologists often disagree about whether or not a nodule exists, its likelihood of
malignancy, and, now, about the descriptive features of pulmonary nodules [11]. Analyzing
disagreement formed the basis for a suite of receiver operator characteristics (ROC) studies and
methodologies, now employed to analyze CAD or compare CAD to radiologists. Within the
pulmonary nodule detection, the primary disagreement focused on the existence or not of
pulmonary nodules. With the LIDC, the analysis of disagreement contains more detail about the
descriptions and the outlines of pulmonary nodules. As observed in the LIDC, radiologists draw
markedly different outlines for the "same" pulmonary nodule. A potential area of research will
attempt to characterize these differences and exploit this understanding for improving CAD
performance. The characterization of radiologists' outlines (markings) will attempt to determine
whether radiologists are outlining different diagnostic opinions about the pulmonary nodule,
such as isolating a central region, while ignoring the outer region included by other radiologists
(see Figure 1). This disagreement illustrates of a conservative versus an aggressive approach to
radiology, a common disagreement of medical opinion. Other disagreements might result from
the variable selection of adjacent regions even among aggressive radiologists (Figure 2), or other
less useful disagreements which might result from inconsistent marking or human computer
interactions rather than intentional choice.
Figure 1 Radiologist Opinion Difference [1]: Black contour represents conservative core, white indicates the aggressive, non-solid component.

Figure 2 Unexplained difference between radiologist outlines where disagreement appears between extent of non-solid component not between solid/non-solid opinion [LIDC document].

6 Potential Research Strategies

A first step towards exploiting the differences requires a quantitative analysis and characterization of the closed curve outlines marked by the radiologists. This research will primarily assist in the segmentation task where previous use of radiologists’ contours either used a union of all contours [12], created a mean contour for ground truth [5], or examined the variation in nodule diameter among radiologists markings [13].

Exploratory analysis of the different markings will begin with a test to determine whether a difference exists between radiologist with a preference for marking the only the inner, conservative core of a nodule while others aggressively choose to include the outer non-solid component. In this difference, the area enclosed by one marking is entirely within another marking, forming a proper subset. Set differencing between the areas can identify this form of variance. If this type of opinion difference is observed, correlation with the nine (9) nodule characteristics will be explored. Other variabilities might be measured by shape features.

Shape measurements were examined [11], but other shape similarity measurements could be explored, such as Pratt's figure of merit and the similarity angle [5]. Measurements of shape such as circularity, eccentricity, compactness, etc. offer a variety of notions of smoothness which might indicate drawing preference. Characterizing and explaining differences, if significant, depends greatly upon the correlation and technique.

Feature selection such as Hessian might be applied to the regions outlined to determine the extent of region differences or use of registration methods such as mean squared differences to determine the extent of difference between radiologists’ "segmentation."
7 Conclusion
Exploratory data analysis of the variability between outlines of pulmonary nodules by expert radiologists will attempt to characterize the difference(s) and suggest strategies of exploiting this knowledge to improve tasks in computer aided diagnosis. Potential discoveries include identifying and quantifying a bias between the opinions of conservative and aggressive radiologists in the diagnosis of pulmonary nodule disease.

References