Evaluation Gap Challenge to Computer Aided Diagnostic Characterization of Pulmonary Nodules

William H. Horsthemke
Ekarin Varutbangkul
Daniela S. Raicu
Jacob D. Furst
Computer Aided Diagnostic Characterization

CADc

Evidence

Reference Cases

Radiologist Diagnosis

Computer Aided Detection CADe
Computer Aided Diagnosis CADx
Computer Aided Diagnostic Characterization

Diagnostic Characteristics
- Spiculation
- Lobulation
- Sphericity
- Margin
- Texture

Evidence
Retrieval Similar Cases

Diagnostic Support

Features:
- Shape, Boundary, Texture, Intensity...

Detection

Computer Aided Detection and Diagnosis
Goal of Computer Aided Diagnostic Characterization: 
*Medically Meaningful Information*

**Computer Aided Detection (CADe) and Diagnosis (CADx)**

*Binary Results with limited, non-descriptive supporting information*

- **CADe**: Present/Absent (perhaps Segmentation Outline)
- **CADx**: Malignant/Benign (perhaps with statistical likelihoods)

**Computer Aided Diagnostic Characterization (CADc)**

*Support diagnoses with descriptions, ratings, and reference*

- Describe and rate nodules according to disease-specific diagnostic characteristics
  - *(evidence-based radiology)*
- Label nodules for reference and retrieval
  - *(case-base reasoning and content/semantic-based medical image retrieval)*

Prior Work

Psychophysical or subjective measures
• Radiologists rate similarity of images containing focal anomaly (nodule or mass)
• Image features measured to predict similarity ratings
  1. **Mammography:** Li et al. 2003
  2. **Pulmonary Nodules:** Muramatsu et al. 2003

Two-stage method introduced by Nakamura (2000) for diagnosing pulmonary nodules
• 1st step: radiologists rate "subjective features" (spiculation, shape, margin, etc.)
  – Image features extracted to predict ratings (unsuccessful due to variability of radiologists ratings)
• 2nd step: subjective features used to predict malignancy (limited success)
• Concluded that single step CADx outperforms 2 step due to inconsistency of radiologists ratings

Prediction of diagnostic characteristics in multi-institutional study of pulmonary nodules
• Image features extracted to predict LIDC (Raicu et al. 2007)
  – Limited success in Texture prediction using data partitions with high agreement on specific characteristics

Prediction of standardized diagnostic characteristics in mammography
• Image features extracted to predict spiculated and circumscribed masses in mammography (Sahiner et al. 2008)
  – Characteristics defined using BI-RADS standarized scoring
  – Breast Imaging and Reporting Data System (BI-RADS)
Evaluation of CADc  
(prediction of diagnostic characteristics)

Requires Ground Truth:

• Provided by domain expert, such as thoracic radiologist
  – Diagnostic Characteristics (such as Texture, Shape, Nodule Outlines) (Nakamura, LIDC)
  – Image or ROI (mass, nodule) similarities provided by radiologist (Subjective Similarity)

• Obtained from patient pathology or confirmed from other diagnostics
  – Rarely includes image related features

• Multiple ground truths arise when radiologists disagree
  – LIDC disagreement on presence/absence, size scale, size details, outlines, and diagnostic characteristics of nodules

Methodology uses standard statistical or machine learning:

  Measure and select features which correlate with Ground Truth, either numerical diagnostic characteristic (LIDC, Nakamura) or psychophysical, subjective similarity matching (Muramatsu)


Problem:
How to evaluate predictive performance of CADc?

Sources of error:

- Image feature measurement technique
  - Implicit features (texture methods covering a set of possibilities)
  - Explicit model for measuring concept (circle detection)

- Selection of feature in statistical or machine learning model

- Choice of statistical or machine learning model (*Logistic regression and Decision trees*)

- Choice of target representation for ratings \{1, 2, 3, 4, 5\}
  - Interval, ordinal, *categorical* (*Per advice of LIDC*)

- Ground truth of nodule location and extent
  - Multiple radiologist-drawn outlines (each radiologist outline and ratings)

- Ground truth of ratings
  - Radiologist agreement and disagreement, bias, and variability
  - Blinded ratings within and between patient cases
Forms of Ground Truth (Labeled Training Data)

**Full Images** (CT Slice, PA Chest Film, …) or Sub-divided (Quadrants)
Labels range from existence of a nodule, location in a lung, quadrant, or XY, typically not outlined. Diagnostic labels range from nodule presence (CAD) to radiologist diagnostic assessment (CADx) or pathology report (CADx).

<table>
<thead>
<tr>
<th>Pulmonary Nodule CT</th>
<th>Pulmonary Nodule CT</th>
<th>Pulmonary Nodule PA Chest Film</th>
</tr>
</thead>
</table>

**Usage:**

Binary (presence/absence or malignant/benign) data offers a benchmark for screening algorithms rather than CAD/CADx. More information about location and extent of nodules is necessary to validate detection and diagnostic algorithms.

Results from this truth/training data can answer only whether the image is suspicious or not, but cannot address the specific location of suspect lesions.
### Forms of Ground Truth (Labeled Training Data) with Increased Specificity

**Regions Of Interest**  Sub-images (nodule, abnormal non-nodule, or normal tissue). Labels specify location of nodule with possible bounding box, a single outline, or, rarely, several outlines (LIDC). Diagnostic labels range from nodule presence (CAD) to radiologist diagnostic assessment (CADx) or pathology report (CADx). Other information includes similarity with known images (Subjective Similarity) and visual characteristics (LIDC).

<table>
<thead>
<tr>
<th>Nodule Region of Interest</th>
<th>Nodule Extent</th>
<th>Nodule Outline (LIDC)</th>
<th>Multiple Outlines (LIDC)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Usage:</strong> Identification and possible diagnosis of nodule, but no information about the extent of the nodule (no outline).</td>
<td><strong>Usage:</strong> Addition of bounding box adds information about nodule size and extent <em>(Volume Doubling Time is an Important Diagnostic Indicator)</em> Without outline, feature extraction will contain nodule and non-nodule information, unless expertly segmented.</td>
<td><strong>Usage:</strong> Outline provides nodule boundary, a template for feature extraction and measurements, and benchmark for segmentation algorithms.</td>
<td><strong>Usage:</strong> Multiple outlines presents challenges for nodule location and extent, but offers range of medical opinion. Overlapping outlines permits probability maps of pixel membership, useful for segmentation algorithm testing.</td>
</tr>
</tbody>
</table>
### Forms of Ground Truth (Labeled Training Data) with Richer Semantic Content

Texture={1,2,..5}; Subtlety={1,2,..5}; Spiculation={1,2,..5};
Sphericity={1,2,..5}; Lobulation={1,2,..5}; Margin={1,2,..5};
Malignancy={1,2,..5}; & <= Four (4) Outline(s)

<table>
<thead>
<tr>
<th>Image Similarity Ratings</th>
<th>Ratings of Diagnostic Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Usage:</strong> Inductive learning and selection of image features that predict similarity between diagnostic categories of images. Requires segmentation of nodule region. Similarity defined by expert radiologists but without description or rating of characteristics.</td>
<td><strong>Usage:</strong> Build and test deductive, image-feature models for direct prediction of radiologist-defined perception of diagnostic characteristics; outlines provide templates for nodule feature extraction and ground truth for segmentation applications.</td>
</tr>
<tr>
<td>Results identify low-level image features that predict diagnostic similarity but offer no further</td>
<td>Results offer semantic description of the medical diagnostic criteria.</td>
</tr>
</tbody>
</table>


Lung Image Database Consortium (LIDC) 
Computed Tomography (CT)

Lung nodule collection and reporting protocol for up to four (4) radiologists to detect and assess focal anomalies (lesions) into 3 categories:

1. non-nodules larger than 3 mm
2. nodules less than 3 mm (unless clearly benign)
3. nodules between 3 and 30 mm in maximum diameter

For category 3, radiologists outline the nodule and rate (1-5) the diagnostic characteristics:
- texture, subtlety, spiculation, sphericity, margin, malignancy, lobulation, internal structure, and calcification (different scale: 1 - 6).

60 patient cases containing 147 nodules detected by one or more radiologists, partitioned into groups of 2, 3, 4, at least 2, and at least 3 ratings

Data sample contains 1 radiologist ratings of characteristics and largest-area outline from set of CT slices marked by radiologist (single slice data)

Up to 4 samples per nodule depending upon number of radiologists who agree on category of lesion

Predict each radiologist rating based upon radiologist-defined outline
Radiologist Variability on Nodule Detection and Nodule Size and Shape

50% variability in regions selected by multiple radiologists (Opfer 2007)
Ratings Variability for Similar Outlines
Direct effect on current outline->ratings prediction method

Spiculation Ratings \{1,1,2,5\}

Spiculation Ratings \{1,1,5,5\}

Spiculation Ratings \{2,3,4,5\}

Green (Spiculation 2) and Blue (Spiculation 5) Outlines Differ by only 1 segment
Methodology

- Focus on Shape-based diagnostic characteristics
  - Applied to radiologist-drawn outlines
  - Spiculation, Lobulation, and Sphericity

- Apply shape-specific feature extraction algorithms
  - Based upon research literature
  - Radial Gradient Index adapted for outlines (Radial Normal Index)
  - Fourier Descriptors

- Classify Diagnostic Characteristics
  - Select best results of Decision Tree and Logistic Regression

- Assess Classification/Prediction Success
  - Radiologist Disagreement (Radiologist Disagreement Index)
    - (mean all-pairs ratings differences)

- Compare detection agreement groupings based upon number of ratings
  - # of Radiologists detecting and rating the nodule: 2, 3, 4, >=2, >=3 ratings
Nodule Boundary Shape
Clinical Indicator of Disease

Diagnostic Characteristics:  *Spiculation, Lobulation, and Sphericity*

Shape Measurement Methods:
- Shape geometry based: area, perimeter, convexity, elongation, orientation, etc.
- Invariant moments: Hu, affine, and Zernike
- Polygon approximation
- Deformable shape based
- Multi-scale shape representation
  - *Fourier transform based*
  - *Radial Gradient indexing*
Radial Normal Indexing

*RGI applied to outlines rather than image gradients*

- Motivated by the Radial Gradient Index (RGI) developed for analyzing spiculation of masses in mammography by Huo [1995].
- RGI features correlated with radiologist rated spiculation [Nakamura 2000]
- RGI-based features have been applied to pulmonary nodule measurements by Zhao [1999], Li [2003], Roy [2006].


Intensity Image and Gradient Magnitude
Near-circular reference Object
Suspicious nodule
Gradient Direction

Near-Circular Reference

Irregular Suspicious Nodule

Red arrows indicate direction of near-maximal image gradients
Radial Normal Index

Circle: Radial Angle = Normal Angle

Non-circle: Radial Angle ≠ Normal Angle

Radial Angle: \( \theta \) Direction of vector from center of object to point on perimeter
Normal Angle: \( \phi \) Direction of vector normal to object at point on perimeter

Methodology: Measure difference between radial and normal angles at various locations along the outline of an object. Accumulate results as a
RNI Results

![Histogram of Near-Circular Object]

- $fwhm = 0.64$  
- $\sigma = 0.15$

![Histogram of Spiculated Lesion]

- $fwhm = 1.71$  
- $\sigma = 0.53$
RNI Captures Increasing Angular Variability along Contour

Nodule 96 Reader 1636 Spiculation 2 Lobulation 2
RNI with FWHM = 0.77

Nodule 96 Reader 6 Spiculation 1 Lobulation 1
RNI with FWHM = 0.86

Nodule 70 Reader 16333 Spiculation 4 Lobulation 4
RNI with FWHM = 1.80

Nodule 139 Reader 0 Spiculation 2 Lobulation 2
RNI with FWHM = 2.52
Shape Prediction Accuracy

Prediction Rate of Shape Characteristics

- Spiculation
- Sphericity
- Lobulation

Prediction Accuracy

- 2 Raters
- 3 Raters
- 4 Raters
- At Least 2
- At Least 3
Radiologist Disagreement Index (RDI)

Introduced recently, RDI measures average, absolute difference between ratings for all pairs of radiologists rating the nodules, normalized from 0 to 1 (max disagreement)

Mean Nodule Disagreement (MND) for R ratings = \[ \frac{\sum_{i=1}^{R} \sum_{i+1}^{R} \text{abs}(\text{ratings}(i) - \text{ratings}(i+1))}{(R \text{ choose } 2)} \] = \# Pairs of Ratings

Normalized Disagreement = \[ \frac{\text{MND}}{\text{MaxDisagreement (MD)}_R} \] = \frac{\text{MND}}{\text{MD}_R} = 4 \text{ for 2 ratings or 2.67 for 3 or 4 ratings}

Radiologist Disagreement Index = Sum(Normalized Disagreement)/number of nodules (ranges from 0 -1)

Example: 3 Radiologists rate nodule with ratings: \{2,2,1\}
   Absolute differences = ( 2-2=0, 2-1=1, 2-1=1 ) Sum of Absolute difference = 0+1+1 = 2,
   Number of Pairs = 3 (3 choose 2), and Mean Nodule Disagreement (MND) = 2/3 = 0.67,
   Normalized Disagreement = 0.67/2.67 = 0.25 (Low Disagreement)

Example: 4 Radiologists rate nodule with 4 scores \{5,5,1,1\}
   Absolute differences = (5-5=0, 5-1=4, 5-1=4, 5-1=4, 1-1=0) Sum of Absolute difference = 0+4+4+4+0 = 16,
   Number of Pairs = 6, and Mean Nodule Disagreement (MND) = 16/6 = 2.66,
   Normalized Disagreement = 2.6/2.6 = 1.0 (Maximum Disagreement)

Radiologist Disagreement Indexes
Per Characteristic and Number of Raters

Average Disagreement per Diagnostic Characteristic

Average Disagreement per Ratings Grouping
Comparison of Prediction and Disagreement

Spiculation Prediction and Ratings Disagreement

Lobulation Prediction and Ratings Disagreement

Sphericity Prediction and Ratings Disagreement
Comparison of Prediction and Disagreement

![Bar chart showing comparison of prediction and ratings disagreement for different numbers of raters.](chart.png)
Comparison of Prediction and Disagreement

Lobulation Prediction and Ratings Disagreement

- Prediction
- Disagreement

Proportion

2 Raters | 3 Raters | 4 Raters | At Least 2 | At Least 3
Comparison of Prediction and Disagreement
Future Work (based upon research of Ekarin Varutbangkul)
Training and Test Database Design
Machine Learning Methods to Admit High and Low Confidence Training Examples

Use Disagreement Index to select nodules with high and low confidence
• Low Disagreement provides High Confidence
• High Disagreement provides Uncertainty
Future Work

Predict only 1 set of ratings per nodule instead of 1 per radiologist

• Modify outline and ratings to combine radiologist guidance
  • Proportional map representation for combining outlines
  • Majority/consensus (mode) method for ratings (similar to clinical practice)
• Design boundary region for shape and margin measurements
  • Develop Magnitude-Weighted Radial Gradient
  • Experiment with other region-based shape features
Effect of Selection of Composite Rating: {Mode, Median, or Mean}

Mean-Median

Mean-Mode

Mode-Median
Future Work

• Modify data sample representation to combine radiologist guidance
  – Proportional map representation for combining outlines
  – Majority/consensus (mode) method for ratings (similar to clinical practice)
  – 1 nodule represented by 1 data sample versus nodules with multiple instances

• Develop new methods for measuring shape using a boundary region image gradient approach and explore other shape methods

• Investigate methods of learning from data samples with high agreement and incorporating samples with conflicting ratings to learn uncertainty

• Investigate machine learning strategies for combining classifiers such as multi-view, co-training, and active learning
References


