DRAFT

Lung Image Database Consortium

Subcommittee on Evaluation Metrics

Report #2

A. Role of LIDC

Rather than supplying software tools for CAD evaluation, the LIDC should prepare a comprehensive document that describes evaluation metrics that are valid for various CAD tasks as reported in the literature. Since it is recognized that no standards exist in the literature regarding the "best" evaluation metrics, this document will describe the benefits and pitfalls of each evaluation metric, the appropriate application of each, and LIDC recommendations regarding the utility of each in the context of the Database. This document will also provide an overview of general issues and caveats involved in the evaluation of CAD schemes and guiding principles for reporting results in the literature. The document may also direct researchers to publicly available software (e.g., ROC and FROC software).

B. The Issues

When reporting results based on the Database, researchers should explicitly state in detail the portion of the dataset used to perform the study. Enough detail regarding the query parameters and exclusion criteria used should be provided to allow for the extraction of the exact same subset of image data by other investigators. The training and testing parameters should be fully disclosed along with the manner in which the dataset was divided between training and testing cases. The LIDC may decide to recommend several subsets of cases for the training and testing of various CAD tasks. For example, the web interface might include an option such as "Click here to download LIDC recommended detection task training set #1." The subcommittee, however, discourages the "policing" of the Database and the collection of "secret" test cases.

Researchers also need to specify the metric used to establish "truth" (e.g., lesion centroid, lesion boundary as supplied in the Database, center-of-mass derived from the lesion boundary) and the criterion used to indicate agreement between CAD output and "truth" (e.g., for the detection task, greater than 50% area overlap between true nodule and detected structure or inclusion of the detected structure's centroid within the boundary of the true nodule). The rationale behind this issue was presented in the first report of the Evaluation Metrics Subcommittee:

> The greatest discrepancy in the literature is the manner in which results are reported. This involves both the criteria established for identifying a "hit" (in a detection task) as well as the definition of "truth." For example, a computer detection may be scored as a "true positive" if (a) the center-of-mass of the computer detection is spatially separated from the center-of-mass of the true lesion by less than some specified distance, or (b) the area of the computer detection overlaps the area of the true lesion by more than some specified area measure. Clearly, the reported performance of a detection algorithm will depend on which scoring metric was used and on the precise value of the limiting distance or area.

The evaluation of CAD performance is a multi-stage process. First, images are input to the system to generate output, which generally will be highly non-uniform and will depend greatly on the specific task and the idiosyncrasies of different algorithms. This output is then passed through some "criteria filter" to determine a category for that output (e.g., nodule or non-nodule, benign or malignant). Finally, the modified output must be further manipulated to accommodate performance evaluation (e.g., the output must be binarized for ROC analysis). Researchers must describe aspects of each step in this process, including the parameter being altered to achieve ROC analysis, if appropriate.

C. "Truth"

To summarize a topic from the first report, the identification of truth and the choice of possible scoring metrics are strongly related. For example, a scoring method based on area overlap would only be possible with truth identified through the manual delineation of lesion margins. The efforts of this subcommittee should be coordinated with the Spatial Truth Subcommittee.

D. Action Items

1) The Steering Committee must reach consensus on the role of the LIDC with regard to evaluation metrics. Should the approach presented in this report be adopted, or should the LIDC take a more active role by supplying software capable of taking a researcher through the stages of performance evaluation in a consistent manner?