31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

A Public Image Database to Support Research in Computer Aided Diagnosis

A. P. Reeves, A. M. Biancardi, D. Yankelevitz, S. Fotin, B. M. Keller, A. Jirapatnakul, J. Lee

Abstract— **The Public Lung Database to address drug response (*PLD*) has been developed to support research in computer aided diagnosis (CAD). Originally established for applications involving the characterization of pulmonary nodules, the *PLD* has been augmented to provide initial datasets for CAD research of other diseases. In general, the best performance for a CAD system is achieved when it is trained with a large amount of well documented data. Such training databases are very expensive to create and their lack of general availability limits the targets that can be considered for CAD applications and hampers development of the CAD field. The approach taken with the *PLD* has been to make available small datasets together with both manual and automated documentation. Furthermore, datasets with special properties are provided either to span the range of task complexity or to provide small change repeat images for direct calibration and evaluation of CAD systems. This resource offers a starting point for other research groups wishing to pursue CAD research in new directions. It also provides an on-line reference for better defining the issues relating to specific CAD tasks.**

I. INTRODUCTION

The development of any application of computer aided diagnosis (CAD) requires access to a set of documented data that can be used for the training or the validation of the application algorithms. In general, training a system is particularly demanding because the best performance can be achieved only if the system is trained with a large amount of well documented data that is able to provide the system with enough generality to cope successfully with real life cases. Such training databases are very expensive to create and their lack of general availability limits the targets that can be considered for CAD applications and hampers development of the CAD field. Historically researchers have considered data collected in studies to be a valuable resource investment for future research projects and there has been a reluctance to make such datasets publicly available. Furthermore, there are other barriers including the effort for obtaining the permission to release the data and the cost of data deidentification and documentation for making a public release.

There are a limited number public image databases that available to support research in CAD. In general these follow one of two models (a) a large set of cases uniformly documented for a carefully designed task of (b) a challenge in which a small set of carefully chosen cases are given usually user-restricted access to the public to obtain results from a range of developers in a comparative study that advances the state of the art. In this paper we present a third alternative, a small public database (without use restrictions) that provides example images and various forms of documentation for a variety of different CAD tasks.

The features of this database are as follows:

1) Carefully selected cases that characterize the range of issues for a given task.
2) Multiple forms of documentation both manual and from computer algorithms.
3) The ongoing addition of new content including: data sets for new tasks, new cases for existing tasks, and new documentation for existing tasks.
4) On-line visualization tools that extend the database to make it an interactive teaching file to support the learning of both clinical and technical CAD issues.

The approval of CAD devices for clinical use requires validation on a large set of cases; however, the use of public databases in this context has not yet been resolved. The main development of a CAD method may be accomplished in many situations with a much smaller set of cases than is required for clinical validation; especially if those cases are carefully selected to represent the range of task issues. The Public Lung Database to address drug response (*PLD*) [1] is designed to provide such a set of cases that will enable research groups to explore new ideas for CAD methods and to evaluate their results with the databases documentation.

There have been recent efforts by the National Cancer Institute (NCI) and other government agencies to create public image databases (PIDB), available from the National Cancer Imaging Archive [2]. The NCI Lung Image Database Consortium (LIDC) [3], [4] is one of these that will soon make available 1000 cases of whole-lung scans with a review by four experienced radiologists who document all pulmonary nodules. The LIDC experience highlights the many formidable technical challenges in developing a PIDB. For example, for the LIDC database decisions were made on the size of the lung nodules to be documented, the acceptable ranges of CT scanner parameters, the methods for documentation, and the inclusion criteria for the patient population from which studies are to be collected [5]. The RIDER project [6], on the other hand, has collected a large number of cases, but only provides very limited documentation on a small number of them.

The main common aspect of current PIDBs is their size, providing (or aiming at providing) several hundreds of cases. The issues shared by large PIDBs may be summarized as follows:

A. P. Reeves, A. M. Biancardi, S. Fotin, B. M. Keller, A. Jirapatnakul, J. Lee are with the School of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14853, USA reeves@cornell.edu

D. Yankelevitz is with the Department of Radiology, Weill Cornell Medical College, New York, NY 10065

1) many decisions need to be made at the outset of data collection that may prove to not be optimal by the time that the database is completed (e.g. technology development may impact the imaging systems while data is still being collected). Further such decisions of necessity tend to be application specific and therefore limit the useful range of lesion characteristics and imaging protocols considered;
2) there is a high cost in documenting a large number of cases;
3) a significant amount of time (years) is required to establish a protocol, acquire the cases, and complete the documentation of the cases.

The high development cost limits the number of different databases that can be created and consequently limits the number of possible CAD applications that can be developed.

The Challenge model provides a set of image cases to be processed by different developers. The results are then sent by the developers to the challenge sponsor for analysis. This can be a very effective method of advancing the state of the art by evaluating competing methods on a single data set. These data sets could provide a rich resource of data for other research but, in general, use-restrictions made on this data prohibit any activity that is not the specific challenge itself.

In this paper we consider a model where a limited set of well documented cases are made publicly available as a service to the community. In addition to basic disease documentation we include other documentation used in our research. For example, in the case of pulmonary nodules we may include both manual and computer made measurements so that a user of the database can review both of these annotations in developing an algorithm. Furthermore, we make available on-line visualization tools that permit user to explore the image data and associated documentation in the context of understanding the issues of the task through a teaching file. We have created two such databases to date. The first, made available in 2003, is the "VIA/ELCAP public lung Image database" [7] contains 50 whole-lung low-dose CT scans from a lung cancer screening study with the locations of all pulmonary nodules identified. The second, "The public lung database to address drug response" was established in 2005 and contains over 100 documented cases [1]. This is described in more detail in the following section. We have also made available an on-line list of current PIDBs and Challenge databases [8].

II. The *PLD* Project

The *PLD* database grew out of the need to gain better understanding about using imaging as a biomarker in the evaluation of drug therapy in the context of pulmonary nodules [13]. There had been considerable interest in diagnosing lung cancer from the growth rate of pulmonary nodules [14,15] but the context for therapy typically considered lesions that were significantly larger than those that were the basis of early diagnosis. One goal of the *PLD* was to provide a range of examples that would span the full range of issues for both small and large nodules. This has now been expanded to a number of different tasks and corresponding datasets as shown in Table I. On entering the website the user can access or download all the cases in a given category. Each image data set may be viewed from the web browser with a fully functional java-applet image viewer as shown in Fig. 1. This viewer permits full medical image set manipulation including zooming, panning, windowing, animation, and image measuring tools. In addition, if there are image annotations in the database for a given image they can be interactively selected for visualization. For many specific CAD tasks there are additional presentations that include task specific visualizations and numerical information; some of these are discussed in the following section. The on-line visualization capability allows the database to be used as a teaching file. The data and corresponding documentation is carefully selected on a task dependent basis as outlined below:

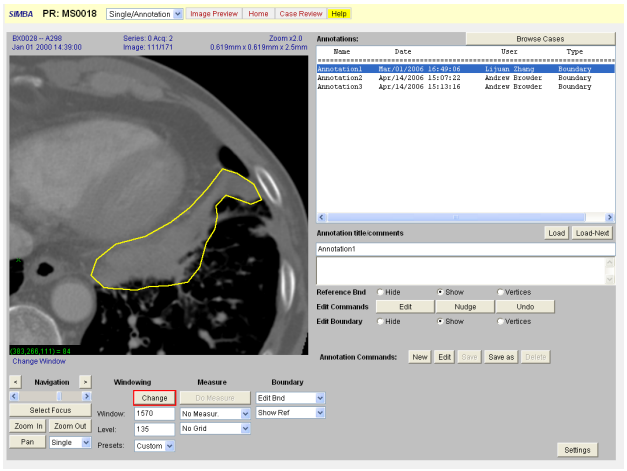


Fig. 1. The interactive tool for image viewing and marking

1) Broad range of disease presentations: The goal of an image-based CAD application is to extract useful diagnostic information from images. To understand a CAD task it is important that the database contains representatives from across the spectrum of cases that occur in clinical practice. For example, for the case of pulmonary nodules there are many different possible presentations to consider: isolated nodules, attached nodules, large tumors, metastatic tumors, benign lesions. For any task as many presentation subtypes as possible should be represented given the limitation in the total number of cases.

2) Special types of subsets: In any spectrum for characterizing a task we should also consider the possibility of special data subsets that can address specific CAD issues. For example, one very important subclass of cases is those for which there are repeat measurements made on the same subject taken in a short time interval. These cases are important since any difference in outcome from the CAD system for such a pair of images is a reflection of the measurement precision of that CAD system. The *PLD* database currently contains two such data sets. The first is

TABLE I
MAIN DATASETS IN THE *PLD*

Dataset	No. of Cases	Documentation	Main Aim
Single Small Nodules	16	Manual and Automated Segmentations	Size Estimation, Nodule Characterization
Single Large Nodules	12	Manual and Automated Segmentations	Size Estimation, Nodule Characterization
Repeat Single Session	23	Automated Segmentations	Measurement Tool Precision, Size Estimation, Nodule Characterization
Sequential Scans	24	Manual Segmentations	Response to Therapy Assessment
Multiple Nodules (Metastatic Cases)	3	Locations, Manual and Automated Segmentations	Example of Computer Aided Detection, Metastatic Nodules
Emphysema	15	Emphysema Index, Severity Map	Emphysema Severity Estimation, Precision of Severity Estimation Tools

the *Zero-Change Dataset* which contains pairs of CT images of pulmonary nodules. These images were recorded during a biopsy procedure before the needle had reached the lesion. Therefore, there was no change in the lesion itself therefore and size changes measured by a CAD system would be due to system variation. The second such special subset is the *Emphysema (Short interval)*, where 5 subjects, spanning the full range of emphysema severity, were imaged within a short interval of time (approximately 30 days). Given the slow progression of emphysema, the degree of the emphysema in the scans can be considered constant and therefore the scans can be used to measure the intrinsic variation in an emphysema measuring system.

3) Careful selection of example cases: In order to well represent a task with a limited amount of resources we have carefully selected several examples from different categories within a task.

III. RICH DOCUMENTATION

For each of the CAD tasks in the we include not just a single form of ground truth that is characteristic of other databases but an enriched documentation which may contain several different versions of markings, visualizations, and numerical analysis. This provides a more detailed context in which a CAD developer can gain appreciation of the task requirements and better evaluate the CAD device performance.

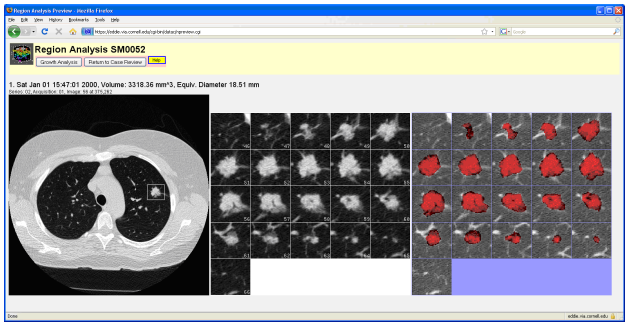
For the task of pulmonary nodule location and size estimation, some cases have manual volumetric markings, where an experienced radiologist has delineated the boundary of the lesion in every scan image where the lesion is visible; some complex lesions have multiple manual volumetric markings to show different approaches in the determination of a lesion boundary and their effect on the lesion volume estimation; the nodules in the Zero-Change dataset have a documentation computed by an automated segmentation algorithm to highlight the challenges facing automated measuring tools. All the documentations were processed to provide also a three-dimensional rendering of the marked or segmented regions and, whenever two or more time points are available, a growth analysis was also performed. Figure 2a shows a visualization of the documentation for case SM0052. There are three image panes: the first from the left shows a central slice of the nodule with the rectangular region of interest, the second shows a montage of the images slices through the lesion, and the third shows a montage in which the regions selected by the marker are highlighted in red. A shaded surface visualization of the lesion is also available and is shown in Figure 2b. The first image is an axial view the second image is a sagittal view and the third is a coronal view. Figure 3 shows an example of growth analysis. The first table shows the size measurements for each scan; for this case there are in fact three scans rather than the usual two. The second table shows the growth analysis for each pair of scans; in this case there are three comparisons and for each of these the doubling time and growth index are computed. Since this lesion is reducing in size with time these growth measures have negative values. Finally, on the same page, shaded surface visualizations rendered at the same scale are presented; the axial view is shown in Figure 3, other views are also presented.

For cases where many nodules are presents, a simple textual list of the nodule locations would be confusing. To exemplify this situation one of the metastatic cases was processed by an automated nodule detection algorithm. Figure 4 shows how much more effective a graphical approach to the nodule listing. Clicking on one of the detected nodules will take you to the interactive image viewer where the nodule can be reviewed in detail.

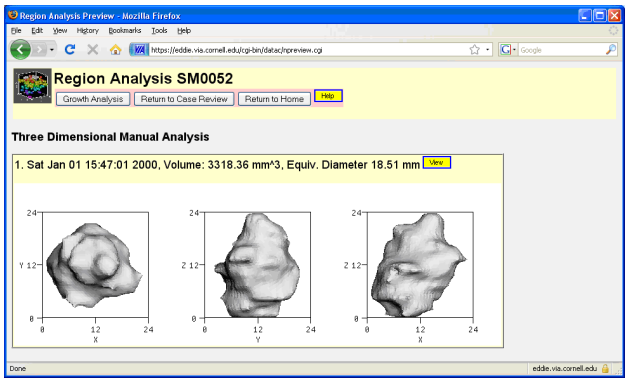
Finally, for the task of emphysema evaluation from CT images, a graphical representation of the severity of the disease is provided, as shown in Figure 5.

IV. CONCLUSIONS

The *PLD* provides a public open learning environment for the development of CAD systems. It is novel in that it provides both a set of documented image data to aid the development of CAD methods and also provides a visual teaching resource. In contrast to the large expensive public image databases, the *PLD* model can provide a critical timely resource for CAD developers with multiple forms of documentation and an update policy that can track the latest technology developments. The *PLD* is an open resource without use restrictions to the community. We believe that adoption of this model by other members of the community would play an important role in the development of CAD. We plan to expand the *PLD* with new tasks, cases, and documentations including documentation submitted by users.



(a)



(b)

Fig. 2. Manual marking of a small pulmonary lesion (a) and its shaded surface visualization (b).

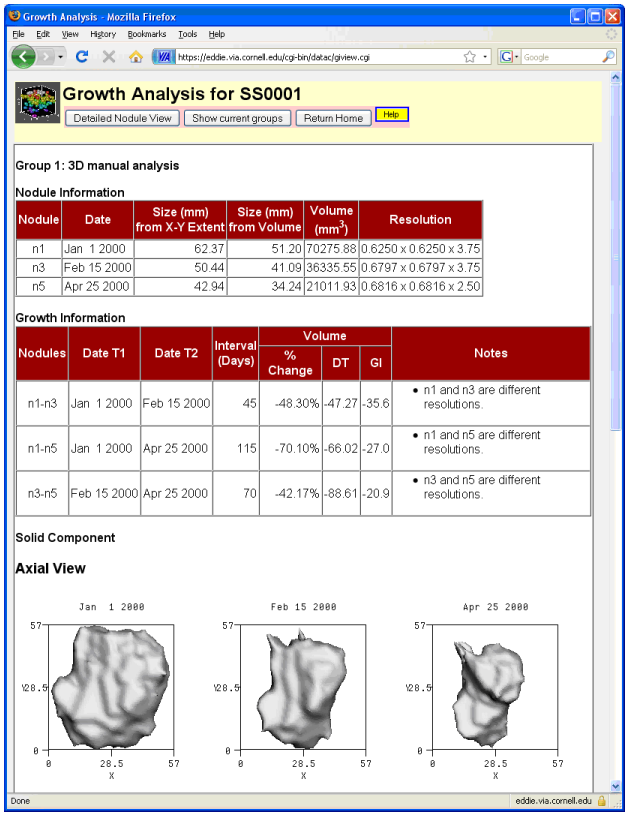


Fig. 3. Growth analysis of a lesion.

V. Acknowledgments

The authors gratefully acknowledge funding from the Prevent Cancer Foundation.

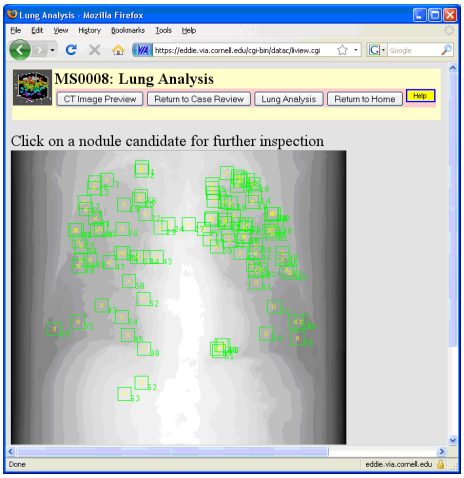


Fig. 4. Nodule detection output on a metastatic case.

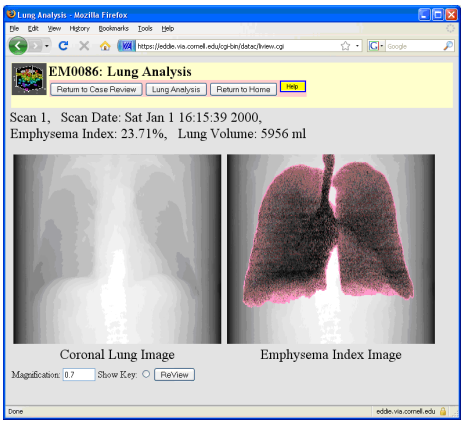


Fig. 5. Graphical rendering of an emphysema severity map.

References

[1] Vision and I. A. Group, "Public lung database to address drug response," http://www.via.cornell.edu/databases/crpf.html, accessed April 7, 2009.

[2] National Cancer Institute, "National cancer imaging archive," https://imaging.nci.nih.gov/ncia/, accessed Jan 9, 2009.

[3] National Institutes of Health, "Lung image database resource for imaging research," http://grants.nih.gov/grants/guide/rfa-files/RFA-CA-01-001.html, April 2000, accessed Jan 9, 2009.

[4] National Cancer Institute, "Lung imaging database consortium (LIDC)," http://imaging.cancer.gov/programsandresources/InformationSystems/LIDC, accessed Jan 9, 2009.

[5] S. G. e. a. Armato, "Lung image database consortium: developing a resource for the medical imaging research community," *Radiology*, vol. 232, no. 3, pp. 739–748, 2004.

[6] S. Armato, C. Meyer, M. McNitt-Gray, G. McLennan, A. Reeves, B. Croft, and L. Clarke, "The reference image database to evaluate response to therapy in lung cancer (RIDER) project: A resource for the development of change-analysis software," *Clinical Pharmacology and Therapeutics*, vol. 84, no. 4, pp. 448–456, 2008.

[7] Vision and I. A. Group, "VIA/I-ELCAP public access research database," http://www.via.cornell.edu/databases/lungdb.html, accessed April 7, 2009.

[8] ——, "Public image databases," http://www.via.cornell.edu/databases/, accessed April 7, 2009.