INTERNATIONAL JOURNAL OF SYSTEMS APPLICATIONS, ENGINEERING & DEVELOPMENT
DOI: 10.46300/91015.2021.15.33 Volume 15, 2021

Semantic and Content-Based Medical Image Retrieval with Proven Pathology for Lung Cancer Diagnosis

Preeti Aggarwal[1], H K Sardana[2]
[1]UIET, Panjab University, Chandigarh, India
[2]CSIO, Chandigarh, India
pree_agg2002@yahoo.com
hk_sardana@csio.res.in

[1]Renu Vig
[1]UIET, Panjab University, Chandigarh, India
renuvig@hotmail.com

***Abstract*—In lung cancer computer-aided diagnosis (CAD) systems, having an accurate ground truth is critical and time consuming. Due to lack of ground truth and semantic information, lung CAD systems are not progressing in the manner these are supposed to. In this study, we have explored Lung Image Database Consortium (LIDC) database containing annotated pulmonary computed tomography (CT) scans, and we have used semantic and content-based image retrieval (CBIR) approach to exploit the limited amount of diagnostically labeled data in order to annotate unlabeled images with diagnoses. We evaluated the method by various combinations of lung nodule sets as queries and retrieves similar nodules from the diagnostically labeled dataset. In calculating the precision of this system Diagnosed dataset and computer-predicted malignancy data are used as ground truth for the undiagnosed query nodules. Our results indicate that CBIR expansion is an effective method for labeling undiagnosed images in order to improve the performance of CAD systems while tested on PGIMER data. Also a little knowledge of biopsy confirmed cases can also assist the physician's as second opinion to mark the undiagnosed cases and avoid unnecessary biopsies.**

***Keywords*—Chest CT scan; computer-aided diagnosis; LIDC; cancer detection and diagnosis; biopsy;PGIMER**

I. INTRODUCTION

Lung cancer is the leading cause of cancer death in the United States. Early detection and treatment of lung cancer is important in order to improve the five year survival rate of cancer patients. Medical imaging plays an important role in the early detection and treatment of cancer. In order to improve lung nodule detection, CAD is effective as a second opinion for radiologists in clinical settings [1]. A dataset with ground truth diagnosis information is essential for CAD systems in order to analyze new cases. To assess the high-quality of the data, several researchers and physicians have to be involved in the case selection process and the delineation of regions of interest (ROIs) to cope with the inter- and intra-observer variability, the latter being particularly important in radiology [2]. Efforts for building a resource for the lung imaging research community are detailed in [3] [4]. The pulmonary CT scans used in this study were obtained from the LIDC [4], and we refer to the nodules in this dataset as the LIDC Nodule Dataset. Recently, diagnosis data for some of the nodules were released by the LIDC; however, because the diagnosis is available patient-wise not nodule-wise, only the diagnoses belonging to patients with a single nodule could be reliably matched with the nodules in the
LIDC Nodule Dataset, resulting in 18 diagnosed nodules (eight benign, six malignant, three metastases and one unknown). The 17 nodules with known diagnoses comprise the initial Diagnosed Subset as one case with unknown diagnose cannot be considered as ground truth. Since the diagnoses in the LIDC Diagnosis Dataset are the closest thing to a ground truth available for the malignancy of the LIDC nodules, our goal is to expand the Diagnosed Subset by adding nodules similar to those already in the subset.
To identify these similar nodules and to predict their diagnoses, CBIR with classification is employed. The radiologist's annotation along with LIDC data is also considered as semantic rating to prepare the ground truth from LIDC data. Increasing the number of nodules for which a diagnostic ground truth is available is important for future CAD applications of the LIDC database. With the aid of similar images, radiologists' diagnoses of lung nodules in CT scans can be significantly improved [5]. Having diagnostic information for medical images is an important tool for datasets used in clinical CBIR [6]; however, any CAD system would benefit from a larger Diagnosed Subset as well as the semantic rating, since the increased variability in this set would result in more accurately predicted diagnoses for new patients.

A. State of the Art

Only a limited number of CAD studies have used a pathologically confirmed diagnostic ground truth, since there are few publically available databases with pathological annotations [7]. In CAD applications for which pathological diagnosis data is absent, determining a ground truth is more challenging. Even with LIDC data where biopsy confirmed cases are available still due to the variability in the opinion of four different radiologists made the LIDC data more complex and redundant. In exploring the relationship between content-based similarity and semantic-based similarity for LIDC images, Jabon et al. found that there is a high correlation between image features and radiologists' semantic ratings [8]. Though in this study, the malignancy rating is also considered for patients having multiple nodules by taking the mean of all the four radiologists rating.

McNitt-Gray *et al*. [9] [10] used nodule size, shape and co-occurrence texture features as nodule characteristics to design a linear discriminant analysis (LDA) classification system for malignant versus benign nodules. Armato *et al*. [11] used nodule appearance and shape to build an LDA classification system to classify pulmonary nodules into malignant versus benign classes. Takashima *et al*. [12] [13] used shape information to characterize malignant versus benign lesions in the lung. Samuel *et al*. [14] developed a system for lung nodule diagnosis using Fuzzy Logic. Matsuki *et al.* [15] also used both clinical information and sixteen features scored by radiologists to design an ANN for malignant versus benign classification.

In all these systems the major concern was to distinguish benign nodules from malignant one where as in the current study we have assigned a new class to the nodules metastases, which indicates that the nodule is malignant however the primary cancer is not lung cancer. The cancer has spread from other organ like neck, breast etc. to lung which can definitely further help the physicians in better understanding of cause and diagnosis for those patients. In the current study, we adopted a semi-supervised approach for labeling undiagnosed nodules in the LIDC. CBIR is used to label nodules most similar to the query with respect to Euclidean distance of image features. By evaluating the method with a CAD application, we determined how to effectively expand the Diagnosed Subset with CBIR. Finally precision is calculated for the PGIMER data with the prepared ground truth.

II. MATERIALS AND METHODS

A. Lung Image Database Consortium (LIDC) Dataset; A benchmark

The NIH LIDC database, released in 2009, contains 399 pulmonary CT scans. Up to four radiologists analyzed each scan by identifying nodules and rating the malignancy of each nodule on a scale of 1-5. The boundaries provided in the XML files are already marked using manual as well as semi-automated methods [1] [5]. Both cancerous and non-cancerous regions appear with little distinction on CT scan image. For accurate detection of cancerous nodules, we need to differentiate the cancerous nodules from the noncancerous ones. The nine characteristics are presented in [16] are the common terms physicians consider for a nodule to be benign or malignant. To our best knowledge, this is the first use of the LIDC dataset for the purpose of validating and classifying lung nodule using biopsy report as well as the semantics attached.

B. Lung Nodule Detection and Selection of Slices

Lung nodules are volumetric and almost available in each slice of patient. It is used for nodule diagnosis as well as for monitoring tumor response to therapy. CT scan of chest is the better method to analyze these nodules for detection as well as for diagnosis. Due to multiple slices in CT, the physician has to see each and every slice for better understanding of each nodule, if present. This task is time consuming as well as not deterministic in any way. We presented a CAD system designed to ensure the nodules marked by different radiologists and consider only effective nodules which can lead to lung cancer, if any, present in the patient. This method can further lead to decrease in time needed to examine the patient's scan by a radiologist.

In this work, these marking are used for the nodule detection and segmentation from chest CT scan, see Figure 1. For better results as well to prepare the ground truth the values of annotations are averaged for all the four radiologists. No automatic segmentation is considered as manual segmentation in medical imaging provides better results [17].

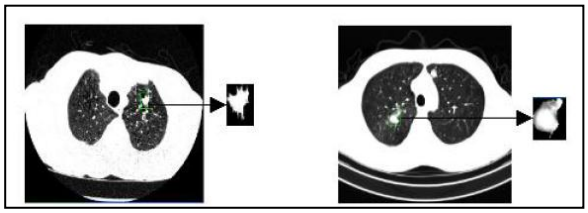

Fig. 1. Comparison of automated and radiologist segmentation

Each slice is read independently to identify its area marked by all the four radiologists and only those slices per nodule is considered to be in the database whose area is maximum [18] and visible in three consecutive slices.

C. Final Extracted Nodule Dataset

CT scan of 80 biopsy confirmed patients with solitary pulmonary nodules mostly less than 3 cm have been taken from. All the images are of size 512*512 and each having 16 bit resolution. All images are in DICOM (Digital Imaging and Communication in Medicine) format which is well known standard used in medical field. Each patient file is associated with an XML annotated file having details of nodule boundaries as well as physician's annotation is associated. Total of 1737 nodules are marked in 80 patients considering each slice of a patient having area greater than all those marked by four different radiologists. Out of 80 biopsy confirmed cases only 18 cases were available with single nodule. From these 18, only 17 cases were considered further to prepare the ground truth as diagnosis for one patient was unknown and this set will be referred to as the Diagnosed17. The classes assigned to these nodules were malignant, benign and metastases based on the diagnosis report available. Rest 62 patients were assigned the class based on the mean of malignancy rating provided by four different radiologists as no ground truth is available for these 62 patients with multiple nodules and this set will be referred as RadioMarked62. It contains 1677 nodules from 62 patients. 83 well known image features were extracted for each nodule based on texture, size, shape, and intensity [16]. The four feature extraction methods used to obtain these 83 features from the LIDC images were Haralick co-occurrence, GLDM, Gabor filters, and Intensity [16]. The number of nodules was reduced to 210 by removing nodules smaller than five-by-five pixels and multiple slices per nodules because features extracted from these smaller nodules are imprecise.
Four different "undiagnosed" query sets containing subsets of the LIDC Nodule Dataset were used, since neither computer-predicted nor radiologist-predicted malignancy ratings can be considered ground truth due to high variability between radiologists' ratings. Each of these query sets differed in diagnostic ground truth. The first query set (Rad210) used the radiologist-predicted malignancy, the second set (Comp210) used the computer-predicted malignancy, the third set (Comp_Rad_biopsy57) used only those nodules for which the radiologist, computer-predicted as well as biopsy confirmed malignancies agreed and the fourth set used only those nodules form which the radiologist- and computer-predicted malignancies agreed. For each query set, nodules with unknown malignancies were removed, and the set was balanced to contain all the three classes i.e. benign, malignant and metastases. The radiologist-predicted and computer-predicted contained equal number of nodules i.e. 210. and radiologist-computer-biopsy-agreement query set contained 57, and Rad_Comp92 contained 92 nodules after all modifications.

III. METHODS

A. Labeling of the Nodules

Nodules are labeled according to single nodule per patient and patients with multiple nodules. Following sections show the details:

- *Patients with single nodule*

Out of 80, only 18 patient cases were having one nodule whereas 62 patients were having more than one nodule. The diagnostic report of LIDC data is patient-wise not nodule-wise. Due to this limitation, biopsy report is used only for 18 patients with single nodule to prepare the ground truth. Biopsy report for those patients has four classes identified as 0, 1, 2 and 3. The meaning of these terms is as described in following table, Table1:

TABLE I. MALIGNANCY RATINGS AND ITS MEANING IN LIDC DATASET

Diagnosis	**Diagnosis at patient level as per LIDC diagnosis report**	**Class assigned in this work**	**Description**
0	Unknown	I	In-determined
1	Benign	B	Non-Cancerous
2	Malignant	M	Cancerous
3	Metastases	MT	Cancer is spreading from other organ to lung.

17 out of 18 biopsy confirmed cases were having the diagnosis as 1, 2 and 3 whereas only one patient was having the diagnosis as 0 which means unknown or indeterminate. This can decrease the classification results, so was not considered in this study. Consequently, 17 pathologically confirmed cases were assigned three classes malignant (M), benign (B) and metastases (MT). There are eight benign (B) nodules, six malignant (M) nodules and three metastases (MT) nodules present in the initial Diagnosed17 set.

- *Patients with Multiple Nodules*

62 out of 80 biopsy confirmed cases with multiple nodules are assigned classes on the basis of radiologist's malignancy characteristics. The meaning and description of malignancy annotation feature of LIDC data is shown in Table1. Out of nine annotations only malignancy feature is used to assign the class to each nodule marked by radiologists as this is most promising feature to determine the malignancy of a nodule. Also, the other characteristics like margin, spiculation, and calcification are already involved in the medical definition of malignancy, so instead of considering all the nine only malignancy features is considered to assign the class as it approximately covers

almost all the other features too. The method used to label each nodule is as follows

Nodules with malignancy rating >=3 assigned class Malignant (M) whereas
Nodules with malignancy rating <3 assigned class Benign (B)

In most of studies, malignancy rating equal to three is considered as unknown however in our study, we have considered that nodule also as malignant which had made the system more sensitive than others. Nodules are having multiple markings by four radiologists on different slices; therefore to reduce the variability among radiologists, the mean of the radiologists' ratings was used. In this way, 1677 nodules from 62 patients were assigned the malignancy class as above. These 1677 nodules contain multiple slices per nodule also and assigned to RadioMarked62 set, which further have been reduced to 210 and assigned to QueryNoduleSet210. If the same nodule appears in the multiple slices, then only those slices are considered in which nodule are having maximum area [18]. This method definitely reduces the database of nodules as well as makes the complexity of volumetric data simpler and effective to analyze.QueryNoduleSet210 further assigned to various categories like Rad210, Comp210 and Comp_Rad_biopsy210 as explained earlier.

B. Summary of CBIR method of Expanding the Diagnosed Subset17; CBIR Expansion Occurs Iteratively

As ground truth for only 17 patients were available, there is a need to expand the diagnostically labeled database. In the absence of diagnostic information, labels can be applied to unlabeled data using semi-supervised learning (SSL) approaches. In SSL, unlabeled data is exploited to improve learning when the dataset contains an insufficient amount of labeled data [19]. CBIR can be used as a machine learning process that trains a system to classify images as relevant or irrelevant to the query. Using available datasets and by evaluating the method with a CAD application, we determined how to effectively expand the Diagnosed17 with CBIR and assist the physicians in the final diagnosis. Each nodule in the QueryNoduleSet210 was then used as a query to retrieve the ten most similar images from the remaining nodules in the Diagnosed17 using CBIR with Euclidean distance. The query nodule was assigned predicted malignancy ratings based on the retrieved nodules (e.g., if the maximum retrieved nodules belong to class malignant then the query nodule was assigned the class M), Figure 2. The newly identified nodule was considered candidates for addition to the Diagnosed17.

C. Diagnosed Subset Evaluation

In the current study, we adopted a semi-supervised approach for labeling undiagnosed nodules in the LIDC. CBIR was used to label nodules most similar to the query with respect to Euclidean distance of image features. Nodules to be added to the Diagnosed17 were selected from the candidates described above. For verifying the addition of a candidate nodule in the Diagnosed17, a reverse mechanism is adopted. Diagnosed17 nodules acted as query and nodules to be retrieved are from QueryNoduleSet210, see Figure 2.

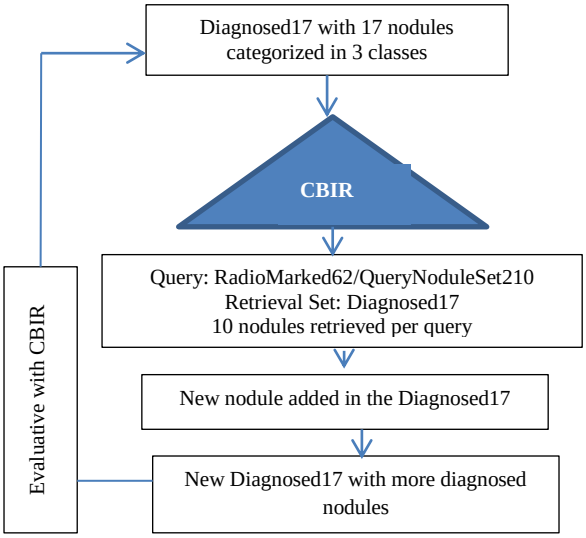

Fig. 2. Selection of candidate nodules using CBIR and Diagnosed17

The first three similar nodules are assigned the same malignancy as the query nodule if they were previously assigned as candidate nodules (i.e. if the query nodule is benign then the top three retrieved nodules are also assigned the class benign if previously are assigned as candidate nodule). Finally based on CBIR and CAD nodules are added in the Diagnosed17. With this mechanism Diagnosed17 in expanded to Diagnosed74, which means that now 74 nodules have the confirmed diagnosis and can be treated as LIDC ground truth. Predicted diagnosis with the pathologically-determined diagnosis, this process guarantees the accuracy of the CBIR-based diagnostic labeling.

D. CBIR Mapping of Multiple Nodules Database with Single Nodule Database

An independent CBIR framework is implemented to increase the Diagnosed17 using CBIR from the QueryNoduleSet210. QueryNoduleSet210 is having multiple nodules per patient. 210 different nodules are present in this set. One by one each nodule is taken as query nodule and matched against Diagnosed17 using CBIR with Euclidean distance. As patient-wise diagnosis is available for QueryNoduleSet210, hence the top retrieved result is matched with this diagnosis. If top retrieved nodule class matches with the patient-wise

diagnosis of query nodule then it is added in Diagnosed17 else discarded. With multiple iterations in this manner, Diagnosed17 is finally increased to Diagnosed121, see Figure 3. Predicted diagnosis with the pathologically-determined diagnosis, this process guarantees the accuracy of the CBIR-based diagnostic labeling.

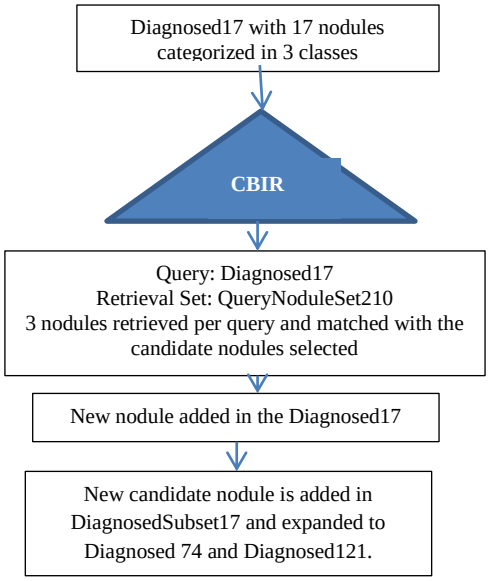

Fig. 3. Expansion of Diagnosed17 to Diagnosed74 and Diagnosed121

E. Query and Retrieval Sets Concluded

In this CAD scenario, two ways process is implemented as discussed earlier. Once the nodules in Diagnosed17 were used as query and QueryNoduleSet210 was used for retrieving the nodules based on CBIR and Euclidean distance and expanded the ground truth to 74 nodules first and then to Diagnosed121. Actually 74 nodules set is prepared from 17 confirmed cases having eight malignant and nine benign cases whereas 121 nodules are prepared from 17 confirmed nodules with six malignant, eight benign and three metastasis cases are there. Secondly, nodules in QueryNoduleSubset210 were treated as query and Diagnosed17 set was used to retrieve most similar nodules to assign the malignancy class accordingly and expanded the Diagnosed dataset to 121. Since neither computer-predicted nor radiologist-predicted malignancy ratings can be considered ground truth due to high variability between radiologists' ratings [7]. This mechanism guarantees the preparation of LIDC ground truth and accuracy of CBIR based diagnostic labeling. All the nodules can be classified in three class benign, malignant and metastases. Various query sets were formed and their precision are compared and shown in Figure 4.

IV. RESULTS

Using the query and retrieval sets as described above, average precision after 3, 5, 10, and 15 images retrieved was calculated. A retrieved nodule was considered relevant if its diagnosis matched the malignancy rating (either radiologist-predicted, computer-predicted, or both) of the query nodule. Initial precision values were obtained by using the 17 nodules in the initial Diagnosed17 as the retrieval set. Then, nodules were added to this set as described in sections 2.2 and 2.3. Precision was recalculated, and the nodule addition process was repeated iteratively using the new Diagnosed17. In each subsequent iteration, only the newly added nodules in the Diagnosed17 were used to identify new candidates. This process repeated until no candidate nodules were added to the Diagnosed17 following an iteration. Various experiments were setup for the validation of nodules examined.

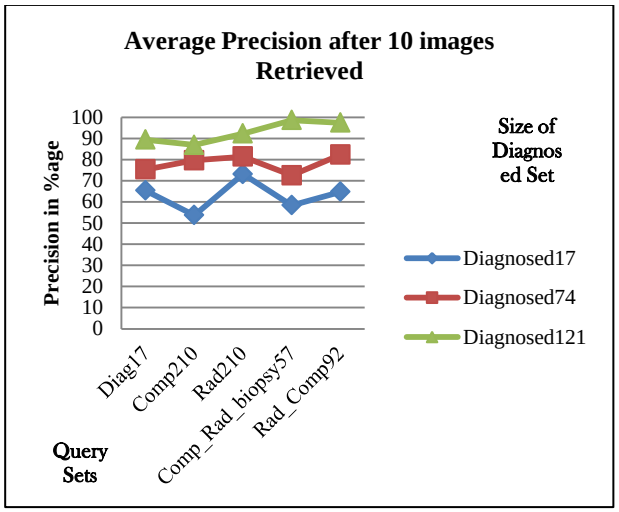

Fig. 4. Comparison of precision for different query sets at x-axis and different retrieval sets at y-axis.

Figure 4 shows that with five query sets and three retrieval sets Diagnosed17, Diagnosed74 and Diagnosed121, the precision increases respectively. Nodules in Comp_Rad_biopsy57 have provided the best precision i.e. 98% which is the best precision achieved in the history of medical CBIR with bets of our knowledge. Finally, the whole system was tested with three cases provided by PGIMER. The PGIMER data is provided with a seed point showing the location of nodule and hence nodule is extracted and shown in Figure 5.

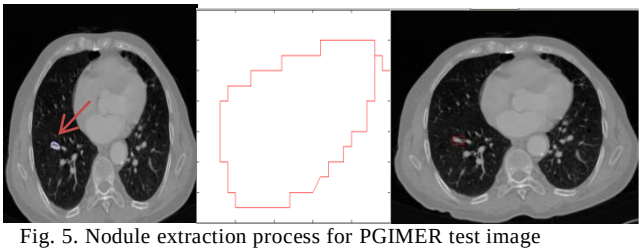

Fig. 5. Nodule extraction process for PGIMER test image

Figure 5 shows the original image with seed point and then shows the plot of nodule extracted from the original image. The third image the bounding box of the extracted nodule (marked in red). Figure 6 shows the exact nodule of the original image.
Precision and recall and calculated for the two queries and are shown in Figure 7. Top 20 images are retrieved and tested for each case. Results indicate that the precision for the malignant query is more than the benign query.

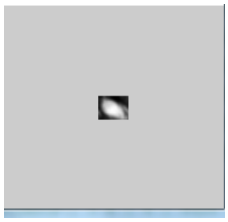

Fig. 6. Nodule extracted from PGIMER test image

Fig. 6. Performance evaluation on PGIMER test images

V. CONCLUSION AND FUTURE WORK

CBIR is an effective method for expanding the Diagnosed Subset by labeling nodules which do not have associated diagnoses. As LIDC is having lack of ground truth, CBIR techniques works tremendously better to prepare the ground truth. This method outperforms control expansion, yielding higher precision values when tested with a potential CAD application [17] that requires a diagnostically accurate ground truth. By increasing the size of the Diagnosed Subset from 17 to 74 and finally to 121 nodules, CBIR expansion provides greater variability in the retrieval set, resulting in retrieved nodules that are more similar to undiagnosed queries. The proposed CBIR expansion method can be applied to differentiate benign, malignant as well as metastases nodules. The third class metastases have not been introduced in the history of CBIR and medical imaging. An expanded set of diagnosed images is also useful for non-CBIR CAD systems, which require large datasets for robust and unbiased training and testing. In future studies, we will investigate using different distance metrics for nodule similarity when identifying candidates with the CBIR expansion method. More test images from PGIMER can provide better precision and recall values and make the system more efficient for use. We also plan to add more classes of malignancy as well as benign to further assist the physicians in more accurate diagnosis.

REFERENCES

[1] D. Wormanns, M. Fiebich, M. Saidi, S. Diederich, and W. Heindel, "Automatic detection of pulmonary nodules at spiral CT: clinical application of a computer-aided diagnosis system," European Radiology, vol. 12, pp. 1052-1057, 2002.

[2] A. Blum and T. Mitchell, "Combining Labelled and Unlabelled Data with Co-Training," Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT '98), pp. 92-100, 1998.

[3] Armato SG, McLennan G, McNitt-Gray MF, Meyer CR, Yankelevitz D, Aberle DR, et al. Lung image database consortium: developing a resource for the medical imaging research community. Radiology 2004; 232(3):739–48.

[4] McNitt-Gray MF, Armato SG, Meyer CR, Reeves AP, McLennan G, Pais RC, et al. The lung image database consortium (LIDC) data collection process for nodule detection and annotation. Academic Radiology 2007;14(12):1464–74.

[5] Armato SG III, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hoffman EA, Kazerooni EA, MacMahon H, van Beek EJR, Yankelevitz D, et al.: The Lung Image Database Consortium (LIDC) and Image Database Resources Initiative (IDRI): A completed reference database of lung nodules on CT scans. Medical Physics; 38: 915–931, 2011.

[6] H. Müller and J. Kalpathy-Cramer, "Putting the Content Into Context: Features and Gaps in Image Retrieval," In J. Tan, New Technologies for Advancing Healthcare and Clinical Practices, IGI Global, Hershey PA, pp. 105-115, 2011.

[7] W. H. Horsthemke, D. S. Raicu, J. D. Furst, and S. G. Armato III, "Evaluation Challenges for Computer-Aided Diagnostic Characterization: Shape Disagreements in the Lung Image Database Consortium Pulmonary Nodule Dataset," In J. Tan, New Technologies for Advancing Healthcare and Clinical Practices, IGI Global, Hershey PA, pp. 18-43, 2011.

[8] S. A. Jabon, D. S. Raicu, and J. D. Furst, "Content-based versus semantic-based similarity retrieval: a LIDC case study," SPIE Medical Imaging Conference, Orlando, February 2009.

[9] McNitt-Gray, M.F.; Hart, E.M.; Wyckoff, N.; Sayre, J.W.; Goldin, J.G.; Aberle, D.R. A pattern classification approach to

characterizing solitary pulmonary nodules imaged on high resolution CT: Preliminary results. Med. Phys. 1999, 26, 880–888.

[10] McNitt-Gray, M.F.; Wyckoff, N.; Sayre, J.W.; Goldin, J.G.; Aberle, D.R. The effects of co-occurrence matrix based texture parameters on the classification of solitary pulmonary nodules imaged on computed tomography. Comput. Med. Imaging Graph.1999, 23, 339–348.

[11] Armato, S.G., III; Altman, M.B.; Wilkie, J.; Sone, S.; Li, F.; Doi, K.; Roy, A.S. Automated lung nodule classification following automated nodule detection on CT: A serial approach. Med. Phys.2003, 30, 1188–1197.

[12] Takashima, S.; Sone, S.; Li, F.; Maruyama, Y.; Hasegawa, M.; Kadoya, M. Indeterminate solitary pulmonary nodules revealed at population-based CT screening of the lung: using first follow-up diagnostic CT to differentiate benign and malignant lesions. Am. J. Roentgenol. 2003, 180, 1255–1263.

[13] Takashima, S.; Sone, S.; Li, F.; Maruyama, Y.; Hasegawa, M.; Matsushita, T.; Takayama, F.; Kadoya, M. Small solitary pulmonary nodules (<1 cm) detected at population-based CT screening for lung cancer: reliable high-resolution CT features of benign lesions. Am. J. Roentgenol. 2003, 180, 955–964.

[14] Samuel, C.C.; Saravanan, V.; Vimala, D.M.R. Lung nodule diagnosis from CT images using fuzzy logic. In Proceedings of International Conference on Computational Intelligence and Multimedia Applications, Sivakasi, Tamilnadu, India, December 13−15, 2007; pp. 159−163.

[15] Matsuki, Y.; Nakamura, K.; Watanabe, H.; Aoki, T.; Nakata, H.; Katsuragawa, S.; Doi, K. Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution CT: Evaluation with receiver operating characteristic analysis. Am. J.Roentgenol. 2002, 178, 657–663.

[16] Raicu, Daniela S; Varutbangkul, Ekarin; Furst, Jacob D, Modelling semantics from image data: opportunities from LIDC, International Journal of Biomedical Engineering and Technology, Volume 3, Numbers 1-2, 30 November 2009 , pp. 83-113(31)

[17] Anne-Marie Giuca, Kerry A. Seitz Jr., Jacob Furst, Daniela Raicu, Expanding diagnostically labeled datasets using content-based image retrieval, IEEE International Conference on Image Processing 2012, September 30 - October 3, Lake Buena Vista, Florida.

[18] Preeti Aggarwal, Renu Vig, and H K Sardana, Largest Versus Smallest Nodules Marked by Different Radiologists in Chest CT Scans for Lung Cancer Detection, International conference on image engineering, ICIE-2013 organized by IAENG at Hong Kong.

[19] Z.-H. Zhou, "Learning with Unlabeled Data and Its Application to Image Retrieval," PRICAI'06 Proceedings of the 9th Pacific Rim International Conference on Artificial Intelligence, 2006.