Annotated Lung CT Image Database

David Ivusic, Antun Petrak, Jelena Bozek, Sonja Grgic
University of Zagreb, Faculty of Electrical Engineering and Computing
Zagreb, Croatia
sonja.grgic@fer.hr

2022 International Symposium ELMAR | 978-1-6654-7003-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ELMAR55880.2022.9899805

***Abstract*—Computed tomography (CT) of lungs provides a diagnostic tool for identifying a range of lesions and diseases visible in the obtained scans. For helping radiologists in a timely and efficient assessment of a large number of scans different machine learning methods have been applied for the detection and classification of abnormalities. However, before clinical usage of such algorithms it is necessary to achieve high accuracy of the algorithm. This is achieved through training and testing phases for which it is essential to have a database that would include a range of abnormalities in the lungs. Here we propose a novel database ALCTID (Annotated Lung CT Image Database) with regions of interest (ROI) annotated by an experienced thoracic radiologist. Database includes 170 lung CT images with a total of 307 annotated ROIs comprising a range of abnormalities, from cancerous lesions, enlarged lymph nodes to enlarged heart and edema. To demonstrate the applicability of the novel database we trained and tested convolution neural network based on the YOLO (You Only Look Once) algorithm on 170 images with annotated ROIs and 170 images of healthy lungs. Out of 80 annotated ROIs in the test set, the network correctly detected 56 ROIs, with 12 false positives and 25 false negatives. The new database ALCTID is publicly available at http://www.vcl.fer.hr/alctid.**

***Keywords*—Lung; CT; Database; Region of Interest; Neural Network; YOLO algorithm; ALCTID**

I. Introduction

Computed tomography (CT) is a reliable and commonly used diagnostic tool for identifying a range of lesions and abnormalities. Chest CT scans are acquired to reveal changes in lungs and often have been used for detecting cancerous lesions. Recent pandemic of COVID-19 has resulted in scanning many patients for detecting a broad range of changes in lung tissue in order to classify the stage, severity and progress of the diseases. Huge amount of cases and scans results in an overload imposed on radiologists, and this burden can be decreased by the usage of machine learning algorithm and computer aided detection of abnormalities. For the efficient training and testing of a machine learning algorithm and achieving high accuracy one of the main requirements is having a representative database for the envisioned application. These can either have a set diagnosis and contain healthy and abnormal cases, or can have annotated and/or segmented suspicious lung tissue and regions of interest (ROI).

Some publicly available lung CT databases do not have annotations or segmentations, but they only provide diagnosis [1–3]. On the other hand, many available lung CT databases that do have annotations have specific purposes, i.e. tumor segmentation or detection and thus they contain images with lung cancer, annotated nodules and do not contain other lesions and changes of the lung tissue [4–6].

Recent outbreak of COVID-19 has demonstrated that databases should include a variety of changes in lungs in order to develop efficient machine learning algorithm that would detect changes that are not necessarily cancerous [7]. This resulted in novel databases that include COVID-19 and non-COVID-19 cases [8]. These types of datasets may aid in developing machine learning algorithm that would aid in fast diagnosis of COVID-19. However, in order to develop a generic algorithm that would analyse any kind of an input and detect any kind of change the dataset, used for training and testing the algorithm, should include a range of changes and those changes should be annotated so that the ground truth is known.

In this paper we introduce an annotated lung CT image database (ALCTID). It is a novel database with regions of interest annotated by an experienced thoracic radiologist. The contribution of our annotated database is twofold. First, it contains a variety of lesions and changes of the lung tissue. Second, it contains CT images of different resolutions and contrasts. These two aspects are beneficial in developing efficient algorithms for computer aided detection of abnormalities and deviations from a healthy lung tissue. Our database is publicly available and suitable for use by students learning about the application of deep learning and computer-aided diagnosis for medical image processing. Finally, in this paper we applied a convolutional neural network (CNN) to demonstrate the usability of the database in the automatic detection of changes in the lung tissue.

Section II describes the novel database while Section III presents the examples of ROIs in the database and the results of the detection performed using the CNN. Section IV concludes the paper.

II. Materials and methods

A. Dataset

We created annotated dataset by annotating ROIs containing lesions in 2D lung CT images. We randomly selected a total of 170 images from the three publicly available databases of lung CT images: i) Chest CT-Scan Images Dataset [3], ii) HRCT Chest Covid Data - CT SCAN [2], iii) Sarscov [1]. It is important to note that images in those databases had no annotations of ROIs and were only marked as being healthy or not.

TABLE I. OVERVIEW OF NUMBER OF IMAGES FROM DIFFERENT DATASETS THAT WERE GATHERED FOR OUR DATABASE AND ANNOTATED (MIDDLE COLUMN) OR WERE HEALTHY FOR THE INCLUSION IN THE TRAINING AND TESTING OF THE CNN (RIGHT COLUMN).

Source dataset	No. of images with lesions	No. of images with healthy lungs
Chest CT-Scan images Dataset	121	0
HRCT Chest Covid Data - CT SCAN	26	67
Sarscov	23	103
Total for ALCTID	170	170

Chest CT-Scan Images Dataset contains 1000 images of healthy and cancerous lungs. It was developed for training artificial neural networks for the task of classifying different types of cancers: adenocarcinoma, large cell carcinoma and squamous cell carcinoma. We randomly selected 121 images with cancerous lungs for the annotation and inclusion in our annotated database.

HRCT Chest Covid Data - CT SCAN dataset comprises 3840 images of healthy lungs and 2242 images of lungs from patients having a confirmed COVID-19. We randomly selected 26 images with lesions and 67 images of healthy lungs.

Sarscov dataset consists of 1229 images of healthy lungs and 1252 images of lungs from patients with confirmed COVID-19. From this dataset we randomly selected 23 images with lesions and 103 images of healthy lungs.

Although images with healthy lungs do not have changes in the tissue that should be annotated, we selected them for the purpose of this work in order to demonstrate the potential usage of the annotated database in the automatic detection of lesions using convolutional neural network (CNN) (see Section II-C.

An overview of number of images gathered from different datasets and that were annotated for the presented new annotated ALCTID database is shown in Table I.

B. Manual lung CT image annotation

Each lung CT image was annotated by an experienced thoracic radiologist (AP) using labelImg application[1]. An example window of the labelImg application is shown in Fig. 1. Annotations are presented as a rectangle over a ROI containing lesions. They are created and stored in the formats compatible with YOLO [9] and Pascal VOC[2].

Annotations were made on 170 images with a total of 307 annotated ROIs.

C. Automatic ROI detection

In order to demonstrate the potential of the new annotated database, we performed automatic detection of the regions of interest containing lesions using convolutional neural networks (CNN). The architecture of the used CNN is based on the YOLO (You Only Look Once) algorithm [9] implemented through Darknet[3], which is an open source framework for real time object detection.

[1]https://github.com/tzutalin/labelImg
[2]http://host.robots.ox.ac.uk/pascal/VOC/
[3]https://github.com/AlexeyAB/darknet

YOLO algorithm scales all input images, regardless of their initial resolution, to 448 x 448 pixels. The algorithm consists of a single forward neural network and predicts bounding boxes and class probabilities directly from full images in one evaluation.

Dataset was separated into training and testing sets. Training set contained a total of 240 images of which 120 annotated images and 120 with healthy lungs. Testing set contained a total of 100 images of which 50 annotated images and 50 with healthy lungs. The 50 annotated images had a total of 80 ROIs, with some images having multiple annotated ROIs.

For every four epochs during testing, a mAP (mean average precision) is being calculated.

III. RESULTS

A. Examples of annotations in the database

Our proposed lung CT database with annotated ROIS with lesions in lungs contains a range of changes in the tissue. In Fig. 2 is an enlarged heart and an edema which characterize an onset of heart failure. Fig. 3 shows a large tumor in mediastinum that pushes trachea. Fig 4 shows a tumor and enlarged lymph nodes which indicate metastasis of the tumor. In Fig 5 is an annotation of a possible tumor or lung abscess. Fig 6 shows annotation with a suspected tuberculosis due to multiple cavities.

B. Results of the automatic ROI detection

Mean average precision of the trained network was 59.14%.

From the 100 images in the test set, the algorithm correctly classified all healthy images as having no lesions. Out of 80 annotated ROIs, the algorithm correctly detected 56 ROIs, while 12 ROIs were false positive and 25 were false negative.

The network accurately detected common lesions in the training set, but misclassified damaged tissue that was not present often in the dataset. Further, the network misclassified spleen and liver as lesions, which happened due to their presence only in some of the images in the dataset.

ROIs with false negatives most often included regions around heart (e.g. increased lymph nodes, structures misinterpreted as bronchi or visceral fat around heart), smaller damages and lung edema located in the lower areas of the field of view in the image. Further, lower areas on the image can contain an increased blood volume since the patient is lying still. The network can misinterpret this as a lesion since it has not properly learned the difference.

Note the mismatch in the summation of correctly detected ROIs and false positives and false negatives. This happened

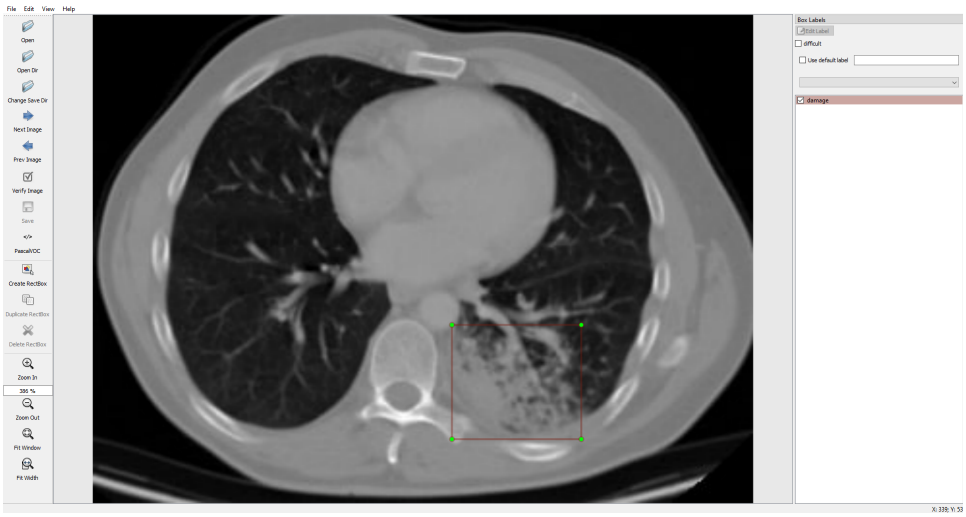


Figure 1. An example screenshot of the application for annotating images with a red box representing annotation made by the radiologist.

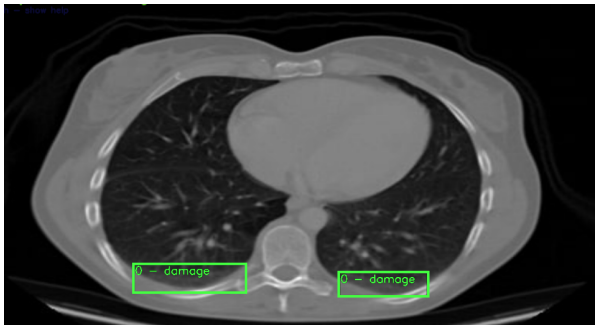


Figure 2. An example annotation of an edema.

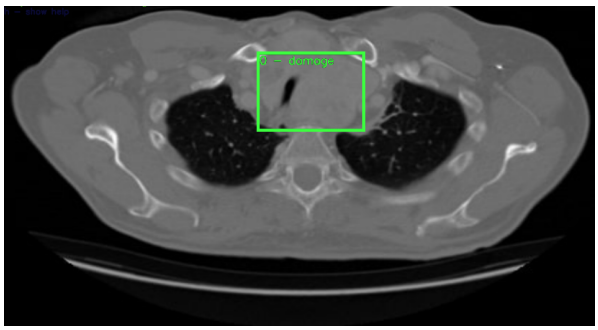


Figure 3. An example annotation of a large tumor in mediastinum.

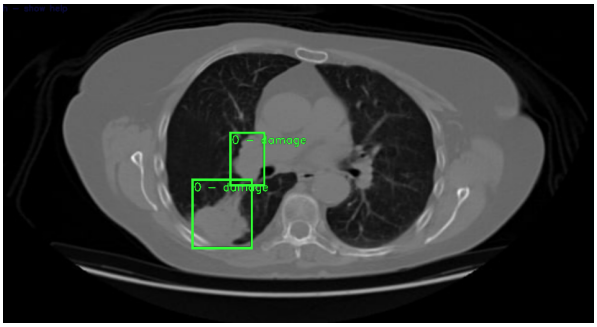


Figure 4. An example annotations of a tumor and enlarged lymph nodes.

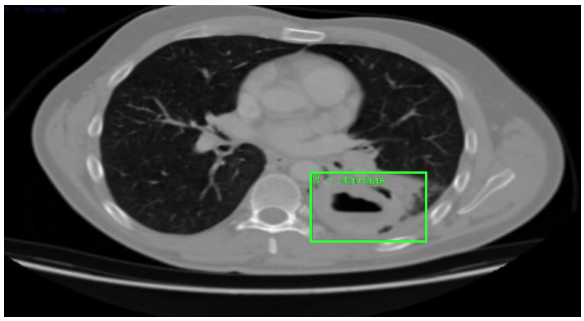


Figure 5. An example annotation of a possible tumor or lung abscess.

due to several combinations of the algorithm output compared to the annotated regions. Some images had a single big area annotated, and the network output were two detected ROIs that together cover the annotated area, and thus both are correctly detected. Also, some images had two annotated ROIs, but the network output was only one of the ROIs, and the other was false negative. In Fig. 7 is an example of such complex outputs. Fig. 8a shows input image with two annotated ROIs, while Fig. 8b is the output of the network. It shows two ROIs correctly detected that represent a single annotation, denoted as "damage:0.88" and "damage: 0.69". Further, network detected false positive on the left side, denoted as "damage: 0.45" , and produced a false negative on the upper right side by missing the area next to the heart.

An example of a correct automatic detection of ROIs that match annotations is presented in Fig. 8.

IV. Conclusion

This paper presents novel database ALCTID with lung CT images containing annotations. Annotated abnormalities include cancerous tissue, enlarged lymph nodes, enlarged heart etc. The database is publicly available at http://www.vcl.fer.hr/alctid.

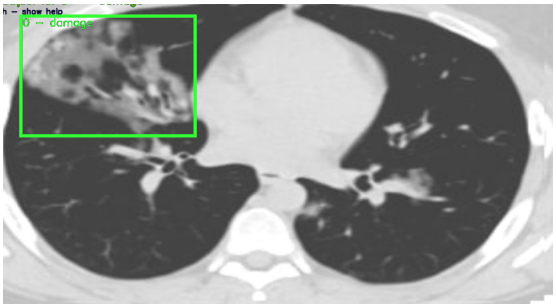
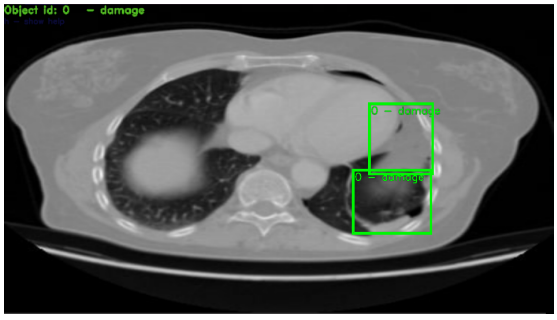
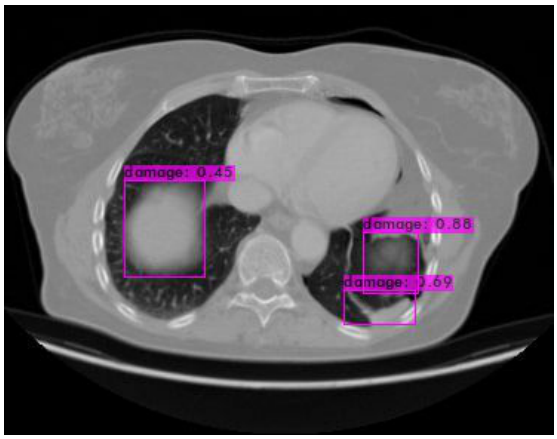


Figure 6. An example annotation of multiple cavities.



(a)



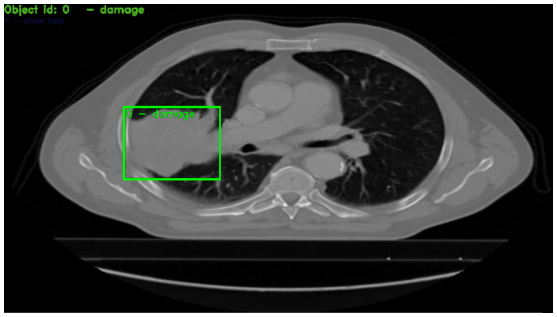
(b)

Figure 7. Example of a) ROIs annotated by experienced radiologist and b) ROIs detected by the algorithm. Note: YOLO algorithm rescales all input images to the resolution of 448 x 448 pixels.

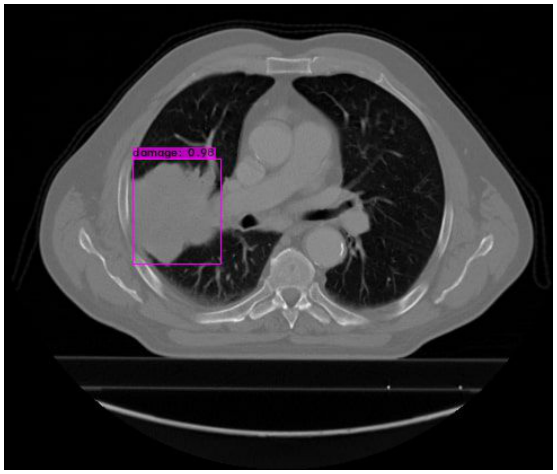
In order to demonstrate the usability of the proposed database we applied YOLO algorithm to detect ROIs and it showed satisfying performance. Out of 80 regions in the test set, the algorithm correctly detected 56 regions, with 12 false positive and 25 false negative ROIs. The algorithm performance could be improved by extending the database with more annotated cases. In the interim, the increase of dataset size for training and testing the machine learning algorithm may be achieved through data augmentation (e.g. TorchIO [10]).

ACKNOWLEDGMENT

This research has been partially funded through the Croatian Science Foundation (HRZZ) project IP-2019-04-1064.



(a)



(b)

Figure 8. Example of a) ROIs annotated by experienced radiologist and b) ROIs detected by the algorithm. Note: YOLO algorithm rescales all input images to the resolution of 448 x 448 pixels.

REFERENCES

[1] "Sarscov," 2021. Accessed 3 May 2022.
[2] M. N. Islam, M. Hasan, A. K. M. Masum, M. Z. Uddin, and M. G. R. Alam, "Demystify the black-box of deep learning models for covid-19 detection from chest ct radiographs," 2021.
[3] "Chest ct-scan images dataset," 2020. Accessed 3 May 2022.
[4] R. V. Adiraju and S. Elias, "A survey on lung ct datasets and research trends," *Research on Biomedical Engineering*, vol. 37, no. 2, pp. 403–418, 2021.
[5] J. Kalpathy-Cramer, S. Napel, D. Goldgof, and B. Zhao, "Multi-site collection of lung ct data with nodule segmentations," 2015.
[6] M. Dolejsi, J. Kybic, M. Polovincak, and S. Tuma, "The Lung TIME: annotated lung nodule dataset and nodule detection framework," in *Medical Imaging 2009: Computer-Aided Diagnosis* (N. Karssemeijer and M. L. Giger, eds.), vol. 7260, pp. 538 – 545, International Society for Optics and Photonics, SPIE, 2009.
[7] H. Jiang, S. Tang, W. Liu, and Y. Zhang, "Deep learning for covid-19 chest ct (computed tomography) image analysis: A lesson from lung cancer," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 1391–1399, 2021.
[8] M. Rahimzadeh, A. Attar, and S. M. Sakhaei, "A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset," *Biomedical Signal Processing and Control*, vol. 68, p. 102588, 2021.
[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015.
[10] F. Pérez-García, R. Sparks, and S. Ourselin, "Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Computer Methods and Programs in Biomedicine*, vol. 208, p. 106236, 2021.