AI-based Retrospective Study for Revealing Diagnostic Errors in Chest X-ray Screening

Vitali Liauchuk
Biomedical Image Analysis Dept.
United Institute of Informatics Problems of NAS of Belarus
Minsk, Belarus
vitali.liauchuk@gmail.com

Aleh Tarasau
Republican Research and Practical Center for Pulmonology and Tuberculosis
Minsk, Belarus
novoe1975@gmail.com

Vassili Kovalev
Biomedical Image Analysis Dept.
United Institute of Informatics Problems of NAS of Belarus
Minsk, Belarus
vassili.kovalev@gmail.com

Abstract. **In this paper, we explore the ability of an AI-based computer-aided diagnostic system (CAD) to help to reveal the early signs of probable lung diseases in X-ray images. We use a large screening database which contains natively-digital X-ray images acquired between 2001 and 2014 along with the corresponding diagnostic reports provided by the radiologists. We apply a Deep Learning-based CAD system to the cases from the database which were labeled by the radiologist as a norm and compare the CAD prediction results to the radiologists' diagnostic reports. Our experiments demonstrate the ability of an automated AI-based CAD to reveal discrepancies between the diagnostic reports and the actual state of lungs as conveyed by the X-Ray image. Additionally, in a number of cases the Deep Learning algorithm was able to detect early signs of lung diseases which progressed later according to the patient anamnesis.**

Keywords: **X-ray CAD, AI, Deep Learning, Retrospective**

I. INTRODUCTION

With the recent emergence of Big Data and Deep Learning methods we observe a rapid development of algorithms which are often referred as Alternative Intelligence (AI). State-of-the-art AI-based solutions of different kinds find more and more applications in different business areas including marketing, industry, and modern software.

However, in the field of medicine the process of incorporation of AI-based solutions for diagnosis and treatment is not as rapid as in other fields. The use of AI in medicine is still rather limited due to high responsibility in the decision making, strict protocols and some skepticism with respect to the use of computerized methods [1]. The exact ways of using AI algorithms in medicine are under discussion [2].

In this paper, we assess the potential profit of using an AI-based X-ray CAD system in clinical practice during the screening. One way to do so is to start using an X-ray CAD system in daily clinical routine. After a while, the effect of using the AI algorithms during the preliminary diagnostic process can be quantitatively evaluated in terms of increased sensitivity, reduced time spent for diagnosing, etc. However, such prospective study requires significant efforts and takes a substantial amount of time to gather representative statistics. This is especially true in case of population screening scenarios where most of the examined cases are expected to be normal. Therefore, here we consider a retrospective study which uses the X-ray images and the corresponding diagnostic reports made in the past.

In this work, we present the results of a retrospective study with the use of a large archive X-ray screening database and a domestic X-ray CAD system. Here we address the following two questions.

1) Is a CAD system able to alarm potential miss-classifications of X-ray images?

2) Is a CAD system able to help in detection of early signs of lung diseases?

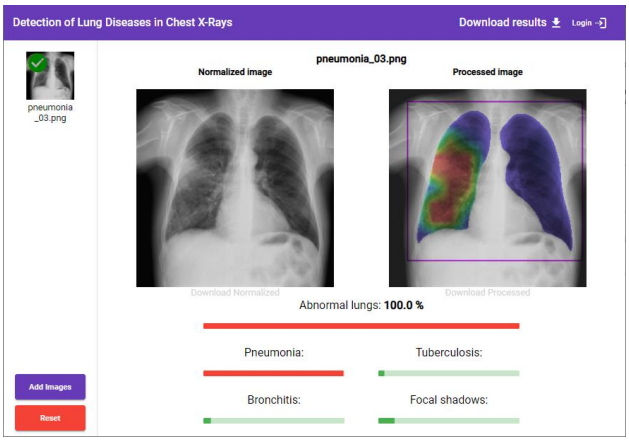

Fig. 1. Visual appearance of the X-ray CAD web-service (https://image.org.by/xray_lung)

To this end, we first automatically detect the cases with potential discrepancy between the radiological description and the X-ray image content. Then, the most probable candidates are manually analyzed by a

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

qualified radiologist to approve or disapprove the revealed discrepancy. Additionally, for those patients which had multiple records in the database we explore whether the automatic CAD could detect the lung abnormalities earlier than the radiologist did according to the anamnesis.

The web-version of the X-ray CAD system used with this study is available for testing at [3]. Visual appearance of the web-service is shown in Fig. 1.

II. X-RAY CAD SYSTEM EMPLOYED

A. CNN Training data

The majority of the X-ray images used in this study were originally digital images taken from the population screening data storage system. The X-ray images were presented as single-channel 16-bit DICOM images with resolution varying from 520×576 to 2800×2531 pixels. Each X-ray scan from the database had its corresponding textual description provided by the radiologist during the image assessment process. In order to extract the image class labels ("normal", "tuberculosis", "pneumonia", etc.) the textual descriptions were parsed with use of keyword and keyphrase matching. The textual descriptions were assigned by qualified radiologists within the screening procedure.

A relatively small portion of the database images was used to train the Convolutional Neural Network (CNN) model for X-ray image classification. The entire database contained 1,908,926 records. From the screening database, a total number of 33,089 cases were selected to compose the study group. Out of those, 16,594 cases represented healthy subjects ("normal"), and 16,495 represented X-ray cases with visible signs of at least one lung disease including tuberculosis, pneumonia, focal shadows and bronchitis ("abnormal"). The screening-based study group was randomly split into training, validation and testing subsets with ratio 70%, 20% and 10% respectively. The split was performed so that all images that were known to belong to a single patient appeared all together in one subset.

To increase the robustness of the resultant prediction model, data from the 3rd party datasets was added into the training subset. The 3rd party datasets included Montgomery and Shenzhen datasets [4], and an additional subset of normal X-ray images taken from "Normal CXR Module: Train Your Eye" [5]. In total, the training subset was extended by 657 normal and 295 abnormal images from 3rd party sources.

B. CNN model training

The CNN model used for classification of X-ray images and localization of abnormal regions was composed of a convolutional part of VGG16 network as backbone appended by several additional layers including a special Heatmap layer. The details about the CNN architecture employed are described in [6]. During inference, the network outputs the overall abnormality confidence score, partial confidence scores for presence of signs of tuberculosis, pneumonia, focal shadows and bronchitis, and a heatmap indicating the localization of the abnormal and suspicious regions in the target image (see Fig. 1).

The CNN training process included two stages. At the first stage, the CNN backbone was initialized with ImageNet-trained weights and the training was performed for the binary classification task ("normal" vs. "abnormal"). This allows using all the available data including 3rd party images which lack meta-information on the specific diseases (tuberculosis, pneumonia, etc.). The second stage is continued from the checkpoint from the first training stage which minimizes the validation loss value. The second stage considers training the CNN in a multi-class multi-label mode with use of all the data except for abnormal 3rd party cases. The experiments showed that using data from different sources makes the final trained model better usable for 3rd party data.

The classification performance of the trained model is assessed with use of ROC-curves. Results of the performance assessment evaluated on the testing subset are shown in Fig. 2.

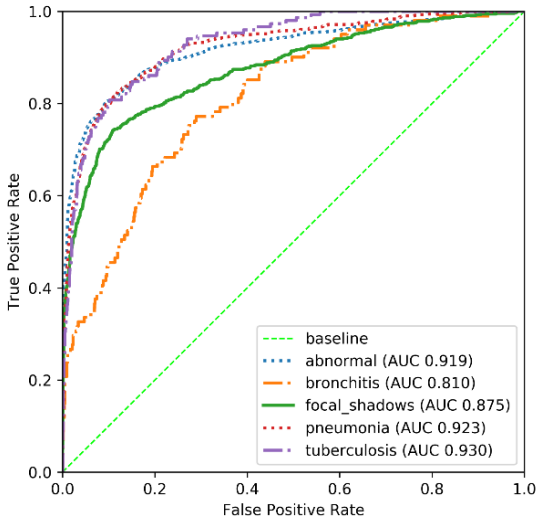

Fig. 2. Roc-curves for prediction of X-ray image with use of the domestic CAD system evaluated on the testing subset

III. RETROSPECTIVE DATABASE ANALYSIS

A. Automatic evaluation of normal X-rays

A total number of 563,495 screening database records were recognized as normal according to their textual descriptions. All the selected images were

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

evaluated with use of the domestic CAD system. The overall abnormality score was used for the subsequent triage of X-ray cases. The statistical results of the evaluation of cases which were marked as normal in the database are shown in Fig. 3. As it can be seen from the histogram, the vast majority of cases had the abnormality score below 0.5, which is well expected for such selection of cases. Still, a large number of images had big abnormality score values. Specifically, 10,007 cases had scores above 0.9, out of which 17 had the maximum possible score of 1.0.

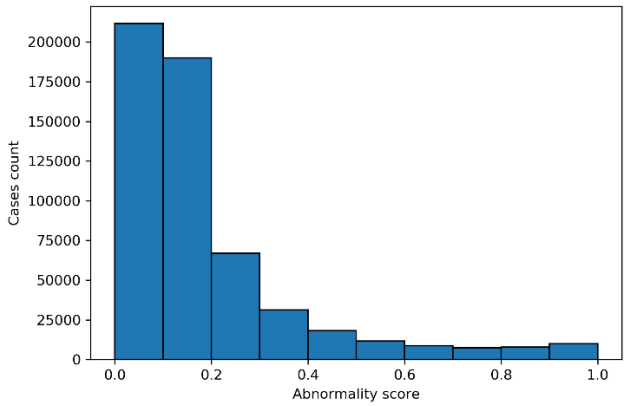

Fig. 3. Abnormality scores automatically evaluated on the screening X-ray images which were labeled as normal

B. Manual evaluation

Due to high costs of manual analysis, only the top-500 X-ray images with the highest abnormality scores were selected for further analysis. Thus, in this preliminary study only a small portion of the potentially interesting data was evaluated. The subsequent analysis included two stages. At the first stage, the automatically selected cases were filtered by a non-radiologist with use of visual analysis to exclude the obvious false-positives (wrong image orientation, unsuitable projection, scanning failures, artifacts, CNN reaction to nipple shadows, etc.). As a result of this stage, 124 out of 500 X-ray cases were selected for the subsequent visual analysis by a qualified radiologist. At the second stage of visual analysis the radiologist had access to all X-ray images of the selected patients available in the screening database. For each image, the corresponding textual description was available. For each analyzed image, the radiologist was to answer two questions: (1) “Is there a discrepancy between the X-ray image content and its description in the database?”, and (2) “Was the X-ray CAD helpful in revealing such discrepancy?”. An additional task was to find those cases in which the automatic CAD could reveal signs of a lung disease earlier than it was done by the radiologists according to the anamnesis.

C. Results

The visual analysis procedure described above revealed 54 cases which had discrepancy between the X-ray image content and its description in the database. In all these cases the CAD-provided heatmaps were helpful in finding the suspicious findings in the X-ray images. In 5 cases the CNN revealed findings which were confirmed later in the anamnesis, in some cases the findings developed into a lung disease.

Fig. 4 shows the examples of X-ray images which have visual signs of abnormalities but were marked as normal in the screening database. Here, the left column shows the original X-ray images, and the right column depicts the corresponding images analyzed with use of the developed X-ray CAD with lung mask and heatmap overlayed on top of the original image.

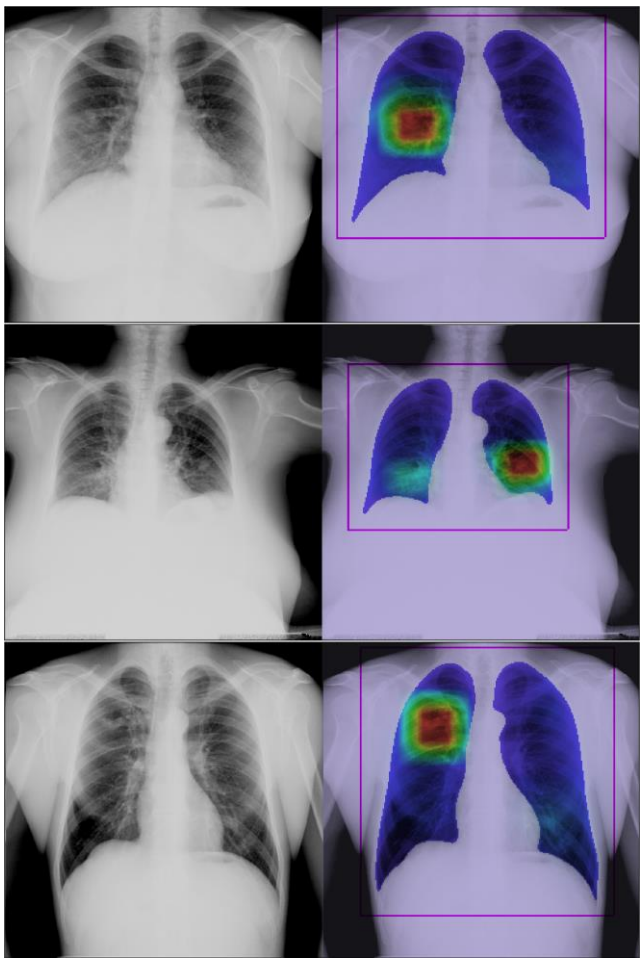

Fig. 4. Examples of X-ray images and the corresponding AI processing results which have visual signs of abnormalities but were marked as normal in the screening database

Fig. 5 shows a case which demonstrates the potential use of the developed X-ray CAD for early detection of signs of disease. Here, the top row shows an X-ray scan of a patient from October 2007, the

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.

textual description reports no visual signs of abnormalities. The bottom row shows another scan of the same patient from October 2008, the description reports a “medium-intensity linear shadow in lower lung field on the right side”. On the other hand, the X-ray CAD applied to these images highlighted the suspicious region with shadow in both cases.

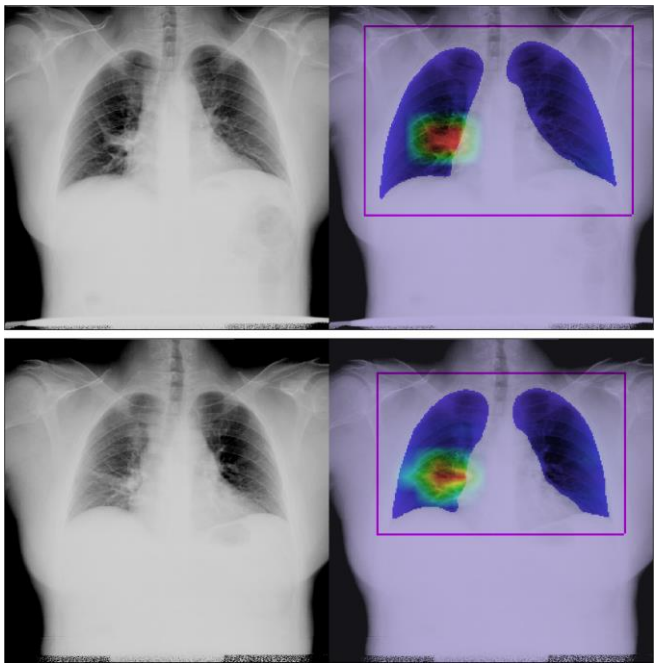

Fig. 5. Two subsequent X-ray scans of a single patient with ~1 year difference; the first image was diagnosed as normal, whilst the second one is reported to have an abnormality in the right lung; automatic X-ray CAD correctly localized the abnormality in both cases

IV. Conclusions

The experimental results presented above suggest drawing the following conclusions.

1) The X-ray CAD system employed with this study is capable of alarming the potential miss-classifications of X-ray images.

2) The employed X-ray CAD system is helpful in detection of early signs of lung diseases.

It should be noticed that in this preliminary study only a small portion of the potentially misclassified X-ray images was visually verified. Out of only 500 cases with the highest abnormality confidence scores we found 54 images which were indeed annotated with an error. This is roughly 10% of the suspicious cases examined. Presumably, a large-scale analysis of tens of thousands of suspicious X-ray images could reveal hundreds or even thousands of misclassified cases.

In general, the results of this retrospective study suggest that using a high-precision X-ray CAD system in routine population screening can increase the diagnostic sensitivity and, in some cases, detect early signs of lung diseases.

Acknowledgment

This study was partly supported by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, U.S. Department of Health and Human Services, USA through the CRDF project DAA9-19-65987-1 “Year 8: Belarus TB Database and TB Portal”.

References

[1] V. H. Buch, I. Ahmed, and M. Maruthappu, “Artificial intelligence in medicine: current trends and future possibilities”, The British journal of general practice : the journal of the Royal College of General Practitioners, Vol. 68, pp. 143–144. March 2018.

[2] A. S. Ahuja, “The impact of artificial intelligence in medicine on the future role of the physician.”, PeerJ, Vol. 7, pp. e7702, October 2019.

[3] “Detection of Lung Diseases in Chest X-Rays”: https://image.org.by/xray_lung – last visited August 7th, 2021.

[4] S. Jaeger, S. Candemir, S. Antani, Y. X. Wáng, P. X. Lu, and G. Thoma, “Two public chest X-ray datasets for computer-aided screening of pulmonary diseases”, Quantitative imaging in medicine and surgery, Vol. 4, pp. 475–477, December 2014.

[5] ”Chest X-ray.com”: http://www.chestx-ray.com/ - last visited August 7th, 2021.

[6] V. Kovalev, V. Liauchuk, D. Voynov, and A. Tuzikov, “Biomedical Image Recognition in Pulmonology and Oncology with the Use of Deep Learning”, Pattern Recognition and Image Analysis, Vol. 31, no. 1, pp. 133–151, April 2021.

Pattern Recognition and Information Processing (PRIP'2021) : Proceedings of the 15th International Conference, 21–24 Sept. 2021, Minsk, Belarus. – Minsk : UIIP NASB, 2021. – 246 p. – ISBN 978-985-7198-07-8.