Heliyon 6 (2020) e04614

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

CellPress

Deep learning for automated classification of tuberculosis-related chest X-Ray: dataset distribution shift limits diagnostic performance generalizability



Helivon

Seelwan Sathitratanacheewin^{a,b}, Panasun Sunanta^{b,c}, Krit Pongpirul^{b,d,e,f,*}

^a Department of Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

^b Thai Health AI Foundation, Bangkok, Thailand

^c True Digital Group Co., Ltd., Bangkok, Thailand

^d Department of Preventive and Medicine, Faculty of Medicine, Chulalongkorn University, Bangkok, Thailand

e Department of International Health and Department of Health, Behavior, and Society, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

^f Bumrungrad International Hospital, Bangkok, Thailand

ARTICLE INFO

Keywords: Computer science Applied computing in medical science Tuberculosis Deep learning Chest X-Ray

ABSTRACT

Background: Machine learning has been an emerging tool for various aspects of infectious diseases including tuberculosis surveillance and detection. However, the World Health Organization (WHO) provided no recommendations on using computer-aided tuberculosis detection software because of a small number of studies, methodological limitations, and limited generalizability of the findings.
Methods: To quantify the generalizability of the machine-learning model, we developed a Deep Convolutional Neural Network (DCNN) model using a Tuberculosis (TB)-specific chest x-ray (CXR) dataset of one population (National Library of Medicine Shenzhen No.3 Hospital) and tested it with non-TB-specific CXR dataset of another population (National Institute of Health Clinical Centers).
Results: In the training and intramural test sets using the Shenzhen hospital database, the DCCN model exhibited an AUC of 0.9845 and 0.8502 for detecting TB, respectively. However, the AUC of the supervised DCNN model in the ChestX-ray8 dataset was dramatically dropped to 0.7054. Using the cut points at 0.90, which suggested 72% sensitivity and 82% specificity in the Shenzhen dataset, the final DCNN model estimated that 36.51% of abnormal

radiographs in the ChestX-ray8 dataset were related to TB. *Conclusion:* A supervised deep learning model developed by using the training dataset from one population may not have the same diagnostic performance in another population. Conclusion: Technical specification of CXR images, disease severity distribution, dataset distribution shift, and overdiagnosis should be examined before implementation in other settings.

1. Introduction

Tuberculosis (TB) is a major health problem in many regions of the world, especially in developing countries. While a majority of patients have their lungs infected with TB, some individuals might have TB infection in other body organs. Hence, unlike other infectious diseases, the diagnosis of TB is relatively more difficult and several tests are usually needed. Although Chest X-ray (CXR) is one of the primary tools for TB screening, a suspected individual requires clinical, biological, and genetic investigations before the actual diagnosis can be made and the medications are readily prescribed. As part of the World Health

Organization (WHO) systematic screening strategy to ensure early and correct diagnosis for all people with TB, CXR is one of the primary tools for triaging and screening for TB because of its relatively high sensitivity, depending on how the CXR is interpreted [1]. However, significant intraand inter-observer variations in the reading of CXR can lead to overdiagnosis or underdiagnosis of tuberculosis.

Deep convolutional neural network (DCNN) has emerged as an attractive technique for TB surveillance and detection. Besides, one study indicated that this approach can accurately detect TB cases in less than 3 min at minimal expense (https://www.digitalcreed.in/ai-for-tb/). This 'supervised' machine learning algorithm learns a mapping from a set of

* Corresponding author. *E-mail address:* doctorkrit@gmail.com (K. Pongpirul).

https://doi.org/10.1016/j.heliyon.2020.e04614

Received 10 June 2019; Received in revised form 26 January 2020; Accepted 29 July 2020

^{2405-8440/© 2020} The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

S. Sathitratanacheewin et al.

covariates to the outcome of interest by using the training data then applies this mapping to the new test data for identification or prediction tasks [2]. In a common deep learning model, the covariates are the color pixel values of the CXR images whereas the outcome is the radiologist's interpretation and impression of the corresponding CXR.

Recently, the National Institute of Health (NIH) released the ChestXray8 dataset with more than 100,000 anonymized CXR images and their associated data which compiled from more than 32,000 patients [3]. These data allow researchers to further develop an algorithm for classifying lung abnormalities labelled from the radiological reports using the National Language Processing technique [4, 5]. Nonetheless, according to the ChestX-ray8 criteria, a CXR image with minimal lung lesions could be incorrectly labelled as normal whereas these radiologic labels are not specific to TB [3].

The Computer-Aided Detection for Tuberculosis (CAD4TB), SemanticMD, and Qure.ai are selected examples of currently available computer-aided detection (CAD) software. CAD4TB is TB specific and had demonstrated a good diagnostic performance [6, 7] but still inferior to that of expert readers [8]. As of 2016, WHO provided no recommendations on using CAD for TB because of the small number of studies, methodological limitations, and limited generalizability of the findings [1]. The other two software cover a broader range of clinical conditions but available only for investigational use in the U.S.

Unlike other diagnostic tests, the technical specification of CXR images and disease severity distribution could affect the diagnostic performance of a supervised machine learning model. That is, the picture archive and communication system (PACS) data has more information but requires more time, storage, and processing power than low bit-depth formats (.jpg) whereas the supervised deep learning model developed based on CXR images of hospitalized patients would be smarter at detecting severe case than another community-based model. To quantify the extent of dataset specificity that limits the generalizability of CAD for TB, we developed a DCNN model using a TB-specific CXR dataset of one population and tested it with non-TB-specific CXR dataset of another population.

2. Methods

Two de-identified Health Insurance Portability and Accountability Act (HIPAA)-compliant datasets, the National Library of Medicine (NLM) Shenzhen No.3 Hospital X-ray set, and the NIH ChestX-ray8 database were included in this study. The Shenzhen dataset collected 336 normal and 326 abnormal CXR of bacteriologically and/or clinically confirmed TB with various manifestations as part of the standard TB care at Shenzhen No.3 Hospital in Shenzhen, Guangdong providence, China. The Shenzhen dataset was provided in PNG format with a resolution of 3000×3000 pixels. The ChestX-ray8 comprises of CXR images acquired as a part of routine care at NIH Clinical Center, Bethesda, Maryland, USA. The images were directly extracted from the DICOM file and resized as 1024×1024 bitmap images without significantly losing the detailed contents. It comprised of 112,120 frontal-view CXR images (60,360 normal and 51,760 abnormal images) of 30,805 unique de-identified patients with the text mined eight disease image labels (including normal, atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, and pneumothorax), from their corresponding radiological reports using natural language processing.

Firstly, the Shenzhen dataset was split into training (75%), validation (15%), and intramural test (10%) sets. Based on the TensorFlow framework, Inception V3, a novel pre-trained DCNN, was augmented with color-space, crop-flip, rotational methods to classify each image as having TB characteristics or as healthy. The learning rate was 0.01, training and validating batch sizes were both 100 with five EPOCHs (https://gith ub.com/panasun/dac4tb). The desktop computer was equipped with Intel i7-7700 CPU @ 3.60GHz, 32G RAM, and NVIDIA GeForce GTX 1080 @ 8GB GPU.

Given no classification of TB within the Chest X-ray8 dataset, we adopted the WHO guideline on Chest Radiography in Tuberculosis Detection which described the standard characteristics of CXR associated with active TB disease by using only 5,000 images of normal CXR and 5,000 images of CXR with active TB associated features according to the WHO guideline (infiltration, pneumonia, atelectasis, and effusion) (1) from Chest X-ray8 dataset to create a valid extramural test set to examine



Figure 1. ROC Curves of DAC4TB in the training Shenzhen No.3 Hospital set.



Figure 2. ROC Curves of DAC4TB in the intramural Shenzhen No.3 Hospital test set.

the generalizability of the DCNN model to classify normal and other CXR in addition to test set from intramural (Shenzhen) dataset.

Lastly, the prevalence of TB-associated CXR in the ChestX-ray8 dataset was estimated by using the final DCNN model. Receiver operating characteristic (ROC) curves and areas under the curve (AUC) were used to assess model performance and to define the optimal cut point for TB detection. detecting TB, respectively (Figures 1 and 2). However, the AUC of the supervised DCNN model in the ChestX-ray8 dataset was dramatically dropped to 0.7054 (Figure 3). Using the cut points at 0.90 which suggested 72% sensitivity and 82% specificity in the Shenzhen dataset, the final DCNN model estimated that 36.51% of abnormal radiographs in the ChestX-ray8 dataset were related to TB.

3. Results

In the training and intramural test sets using the Shenzhen hospital database, the DCNN model exhibited an AUC of 0.9845 and 0.8502 for

4. Discussion

Our findings suggested that the diagnostic performance of a supervised machine learning model is dataset specific, because of the varying technical specification of CXR images and disease severity distribution in



Figure 3. ROC Curves of DAC4TB in the extramural NIH ChestX-ray8 test set.

S. Sathitratanacheewin et al.

different populations. In this study, the Inception V3 TensorFlow model maintained the diagnostic accuracy for the training, validation, and intramural test images in the Shenzhen dataset. This ability to correctly classified unseen test images within the same dataset is known as generalization. However, the significant drop in the diagnostic performance in the extramural test images reflects poor generalizability.

The falling of performance when the distribution of the data changed is known as 'distribution shift', and still an active area of research in the machine learning community. One might argue that the distribution shift becomes more apparent in our study because of insufficient model preparation and optimization. However, we had augmented our model with several techniques that have been known to improve generalization of deep learning in the other area of research including color-space, cropflip, and rotational methods. Although, there are other techniques [9,10] that may help improve our model performance. By far the best, the discrepancy in CXR readings pattern between the two datasets according to differences in prevalence, manifestations of common lung abnormalities, as well as health care system in China and the U.S. are unavoidably having an important role in performance shift in this study. Thus, it is more likely our findings indicated that dataset distribution shifts in deep learning in some specific research area is inevitable. Careful implementation of AI and further researches to examine the problem will improve the development and adoption of AI in medicine and public health.

The 36.51% of abnormal CXR in the NIH ChestX-ray8 dataset was associated with TB might demonstrate the 'overdiagnosis' of deep learning since many CXR abnormalities that are compatible with pulmonary TB are seen also in several lung pathologies and, therefore, are indicative not only of TB but also of other pathologies.

Some limitations of this study should be noted. Like other deep learning studies, the images were resized to a manageable dimension before being fed into the model as a larger file will increase the training time and will require more robust central and graphical processing power. Accuracy may be improved by using higher resolution images, particularly for subtle findings, and more research is needed in this regard. Secondly, we did not comparatively explore the effect of using other pre-trained models such as the VGGNet or ResNet as the former is a very slow pre-trained sequential model and the latter is a relatively less efficient model than the Inception V3. Lastly, this retrospective study was conducted using the datasets that were available at the time. Further investigation on the use of deep learning in a real-world large-scale screening program for pulmonary TB in prevalent regions is essential.

5. Conclusions

A supervised deep learning model developed by using the training dataset from one population may not always have the same diagnostic performance in another population. Technical specification of CXR images, disease severity distribution, dataset distribution shift, and overdiagnosis should be examined before implementation in other settings.

Declarations

Author contribution statement

K. Pongpirul, S. Sathitratanacheewin: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

P. Sunanta: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Funding statement

This study was supported by the Health Systems Research Institute (HSRI 62-103) and Ratchadapiseksompotch Matching Fund, Faculty of Medicine, Chulalongkorn University (RA-MF-12/62).

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

We also thank Dr. Ekapol Chuangsuwanich, Faculty of Engineering, Chulalongkorn University and Dr. Phalin Kamolwat, Bureau of Tuberculosis, Department of Disease Control, Ministry of Public Health for kind advice.

References

- World Health Organization, Chest Radiography in Tuberculosis Detection: Summary of Current WHO Recommendations and Guidance on Programmatic Approaches, World Health Organization, Geneva, 2016.
- [2] J. Wiens, E.S. Shenoy, Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology, Clin. Infect. Dis. 66 (1) (2018) 149–153.
- [3] X. Wang, et al., ChestX-ray8: Hospital-Scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, 2017 arXiv:1705.02315.
- [4] P. Rajpurkar, et al., CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, 2017 arXiv:1711.05225.
- [5] L. Yao, et al., Learning to Diagnose from Scratch by Exploiting Dependencies Among Labels, 2017 arXiv:1710.10501.
- [6] P. Maduskar, et al., Performance evaluation of automatic chest radiograph reading for detection of tuberculosis (TB): a comparative study with a clinical officers and certified readers on TB suspects in sub-Saharan Africa, Insight Imag. 4 (Suppl 1) (2013), S322.
- [7] M. Muyoyeta, et al., The sensitivity and specificity of using a computer aided diagnosis program for automatically scoring chest X-rays of presumptive TB patients compared with Xpert MTB/RIF in Lusaka Zambia, PloS One 9 (4) (2014), e93757.
- [8] A. Steiner, et al., Screening for pulmonary tuberculosis in a Tanzanian prison and computer-aided interpretation of chest X-rays, Public Health Action 5 (4) (2015) 249–254.
- [9] P. Zhou, J. Han, G. Cheng, B. Zhang, Learning compact and discriminative stacked autoencoder for hyperspectral image classification, IEEE Trans. Geosci. Rem. Sens. 57 (7) (July 2019) 4823–4833.
- [10] G. Cheng, C. Yang, X. Yao, L. Guo, J. Han, When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs, IEEE Trans. Geosci. Rem. Sens. 56 (5) (May 2018) 2811–2821.