Deep Learning Classification of Active Tuberculosis Using Chest X-Rays: Efficacy of Transfer Learning and Generalization Performance of Cross-Population Datasets

James Devasia
Jawaharlal Institute of Post Graduate Medical Education and Research
Hridyanand Goswami
Marwari Hospitals
Subitha Lakshminarayanan (subitha.l@gmail.com)
Jawaharlal Institute of Post Graduate Medical Education and Research
Manju Rajaram
Jawaharlal Institute of Post Graduate Medical Education and Research
Subathra Adithan
Jawaharlal Institute of Post Graduate Medical Education and Research
Ambalavanan Bharanidharan
Sri Ramakrishna Engineering College

Research Article

Keywords:

Posted Date: January 14th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1235165/v1

Abstract

Chest X-ray based diagnosis of active Tuberculosis (TB) is one of the oldest ubiquitous tests in medical practice. Artificial Intelligence (AI) based automated detection of abnormality in chest radiography is crucial in radiology workflow. Most deep convolutional neural networks (DCNN) for diagnosing TB by transfer learning from natural images and using the same dataset to evaluate the model performance and diagnostic accuracy. However, dataset shift is a known issue in predictive models in AI, which is unexplored. In this work, we fine-tuned, validated, and tested two benchmark architectures and utilized the transfer learning methodology to measure the diagnostic accuracy on cross-population datasets. We achieved remarkable calcification accuracy of 100% and area under the receiver operating characteristic (AUC) 1.000 [1.000 – 1.000] (with a sensitivity 0.985 [0.971 – 1.000] and a specificity of 0.986 [0.971 – 1.000]) on intramural test set, but significant drop in extramural test set. Accuracy on various extramural test sets varies 50% - 70%, AUC ranges 0.527 – 0.865 (sensitivity and specificity fluctuate 0.394 – 0.995 and 0.443 – 0.864 respectively). Diagnostic performance on the intramural test set observed in this study shows that DCNN can accurately classify active TB and normal chest radiographs, however the external test set shows DCNN is less likely to generalize well on models trained on specific population dataset.

Introduction

Tuberculosis (TB) is a deadly infectious disease caused by the bacillus Mycobacterium tuberculosis. In 2019, World Health Organization (WHO) estimated and reported 10 million active cases (8.9 -11.0) 95% UI and 1.2 million (1.1 -1.3) 95% UI people died from TB and an extra 208 000 (177,000-242,000) 95% UI deaths (with HIV) in a single infectious agent despite being preventable and curable[1]. The goal of the United Nations General Assembly meeting held in September 2018 was to detect and treat 40 million people with Tuberculosis by 2022[2]. One of the strategies to achieve this goal was to improve and expedite the screening and triage procedure for active Tuberculosis. Sputum smear microscopy, rapid molecular tests, and culture tests are various methods for diagnosing active TB and drug-resistant TB; however, these approaches are relatively expensive and not easily available or applicable in low-resourced regions. Tuberculosis unveiled in chest radiography (CXR) images are broadly classified into parenchymal and pleural involvement. Most common parenchymal pathologies are consolidation, cavitation, reticular opacity, fibrosis, bronchiectasis, calcification, hilar adenopathy, and collapsed lung. In pleural cavity, the most commonly exhibited pathologies are pleural effusion, thickening, calcification, and pneumothorax[3]. Therefore, Posteroanterior (PA) chest radiography as a screening tool for detecting Tuberculosis plays a vital role in many algorithms as it is a fast and economically viable solution despite low specificity and high sensitivity[4,5].

In emerging economic countries, the lack of trained clinicians or radiologists in remote areas and inter/ intra reader variability leads to delay in diagnosing and missing out on the active Tuberculosis case[6,7]. Sometimes another pathology diagnosis as TB. In this scenario, there has been an increased interest in developing and using Computer-Aided Detection (CAD) for radiology interpretation using diverse methods, which show promising results[8–12]. Recent (2021) guidelines released by WHO recommended using CAD packages for automated screening and triage of active TB diseases among individuals above 15 years populations using the interpretation of digital CXR[6] where 56% of active TB cases developed in 2019 were individuals aged >=15 years and old[1].

The Deep Learning research community was attracted to radiology interpretation for detecting many diseases after the breakthrough innovation and success of the winning software AlexNet[13] in ImageNet Large Scale Visual Recognition Competition (LSVRC) based on Convolutional Neural Networks (CNN) in 2012. Before 2012 CAD packages for detecting TB were handcrafted Machine Learning approaches and used TB-specific textural feature selection and classification[14–17]. Since 2012 classification of radiology images gained much attention by the state-of-the-art Deep Convolutional Neural Networks (DCNN) technologies, transfer learning, and publicly available large datasets with interpretations. This resulted in remarkable classification accuracy in lung segmentation[18] using Total Variation-based Active Contour algorithm, lung nodule detection[19] with RetinaNet and modified U-Net, cardiothoracic diseases[20] with localization and multi-label classification on par with experienced radiologists and clinicians.

Dataset shift, also known as a shift in the distribution of the variable, is a known issue in predictive models where training and testing data differ in the distribution of single or multiple features or class itself[21–23]. Reported works[8,9,11,14,16,17,20,24] on classification of Tuberculosis use training and testing from the same dataset to measure the performance matrics to evaluate the model. Few predictive models in Biomedical research addressed this phenomenon[10,25,26], and few reported and incorporated changes to tackle the situation[27,28]. Reported works[25,29,30] established that dataset shift leads to poor generalizability in deep learning predictive models. There is a paucity of evidence regarding cross-regional or cross-population train/ test and the effect in diagnostic accuracy in detecting TB. In this work, we assess the efficacy of transfer learning and diagnostic generalization performance of DCNN using a cross-population train /test dataset to detect active Tuberculosis.

Materials And Methodology

A. Study design

Retrospective study with model creation using transfer learning and analyzing diagnostic generalization to detect active Tuberculosis.

B. Dataset sources and curation

The following four datasets were used for cross-population diagnostic classification accuracy of active Tuberculosis, two publicly available Health Insurance Portability and Accountability (HIPPA) compliant datasets maintained by the National Library of Medicine, Maryland, USA: (a) Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China (SH) collection[31]; (b) Department of Health and Human Services, Montgomery County, Maryland, USA (MC) collection[31]. The SH and MC collections are in in Portable Network Graphics (PNG) format. The other two datasets, (c) TB Portals Program data (TBPP), National Institute of Allergy and Infectious Diseases (NIAID), Bethesda, Maryland, USA, (NIAID) collection[32] , which follows the HL7 Fast Healthcare Interoperability Resources (HL7 FHIR) standard in Digital Imaging and Communications in Medicine (DICOM) format. The NIAID dataset comprises primarily drug-resistant TB cases. PA CXR comes from TB Portals Consortium and participating institutions in India, Belarus, Romania, Georgia, and Azerbaijan, with a significant share from Belarus. (d) Jawaharlal Institute of Postgraduate Medical Education and Research (JIPMER Hospital), Puducherry, India (IN) collection, use HL7 integration interface between Picture Archival and Communication System (PACS) and Health Information System. For the IN collection, patient demography was collected from the Revised National Tuberculosis Control (RNTCP) program referral register, where patients were confirmed as active TB using sputum or culture tests from 2017-2020. CXR was downloaded from PACS using the patient identifier in Tag Image File Format (TIFF) format. Healthy control subject's demography and CXR were collected from an ongoing TB project.

All CXR used in the study were de-identified using system generated study identifier and any overlay information in CXR were removed to protect the privacy of the patients. There were no missing data on patient demographics and CXR. The basic characteristic of the datasets is in Table 1 which also reveals final number of abnormal and healthy images for training, validation, and testing the models.

C. Pre-processing of CXR

The input data were pre-processed before feeding to the model with the following: (1) segmentation of the region of interest (ROI) using U-net[33] architecture, we capture the extreme points from lung area ROI and padded with extra 50 pixels to make sure the bounding box maintained the ROI (e.g., TB patients with Pleural Effusion), (2) extract the ROI to the size of 224x224 and saved in Joint Photographic Experts Group (JPEG) format. Supplementary figure S1 and Supplementary figure S2 shows the pre-processing pipeline of all CXR images. (3) All saved images were pre-processed with pre-processing functions of Resnet50[34] and Densenet121[35]architecture during training, validation, and testing.

D. Ground Truth

All PA TB CXR in IN collection were read by single radiologist (HG) for various TB manifestations and reported in a specified format approved by the institute committee. Peer validation of 10 % CXR was done by the radiologist (SA) and pulmonologist (MR). The interobserver agreement[36] between HG and SA and HG and MR was almost perfect (k=0.83, 95% CI [0.72 – 0.93] and k=0.80, 95% CI [0.71 – 0.94] respectively). The Normal CXR in IN collection was read by the clinician and peer validated by MR, with inter-rater reliability, was almost perfect (k = 0.90 95% CI [0.87 – 0.98]. The MC, SH, and NIAID dataset's ground truth and subject's demography was obtained from the corresponding dataset source.

E. Data Partitions

We randomly selected 55% (1672) of IN collection healthy control images used in conjunction with NIAID dataset for training, validation, and testing, a similar attempt made by Lakhani, P. & Sundaram [24]. Each dataset was split into training (80%), validation (10%), and intramural holdout test (10%) sets. Repeated images of the same patients were included only in training set to avoid data leakage in holdout test set. Extramural test sets are the datasets were not used for training and validation.

F. Model

Resnet50 and Densenet121 pre-trained networks on ImageNet which are widely used in medical image classification were used in this work as base models. The input image size of Resnet50 and Densenet121 was set to 224x224. We used weights from ImageNet (transfer learning) to initialize the network. The classifier in base model is replaced with the following; (1) a Global Average Pooling layer, (2) a Dense layer with ReLU activation, (3) a Dropout layer, and (4) a classifier layer with two outputs and softmax activation. Based on the TensorFlow framework (version 2.7.0, Google Brain Team, CA, USA https://tensorflow.org), Keras (version 2.7.0, https://keras.io) and Python (version 3.7, Python Software Foundation, DE, USA https://python.org) as programming language were used in deep learning frameworks. The desktop computer was equipped with Intel i9-9820X CPU @3.30 GHz, 64G RAM, and dual NVIDIA GeForce RTX 2080Ti @11G GPU. Figure 1 shows the outline of the DCNN model.

G. Training

All layers except Batch Normalization layers in the base model were set non-trainable as the mean and variance of the training dataset differed from ImageNet[37]. Hyperparameter tuning for the added layers was carried out by Keras Hyperband Tuner[38] with max_epochs=10 and hyperband_iterations=5 to optimize the Dense layer units, Dropout rate, and Learning rate. The training was performed using Categorical cross-entropy as losses function, Adam[39] optimizer. The network was trained with mini-batches of 32 samples. The mini-batches were shuffled on each epoch to randomize the training method and decrease overfitting. We also employed Early Stopping aided by validation loss to reduce overfitting. We did not use any data augmentation technique in this work. The model selection was performed by Keras Tuner. The summary of the optimal Resnet50 and Densenet121 models for various datasets is shown in Supplementary Table S2.

H. Evaluation Metrics

Testing of the model was done using intramural holdout and extramural test sets. The following metrics were used to evaluate the efficacy of the model using the test sets, (1) Sensitivity, (2) Specificity, (3) Area Under the Receiver Operating Characteristic (AUC), (4) Accuracy, (5) Precision and (6) F1-Score. Confidence Interval (CI) for AUC were calculated using Hanley & McNeil test[40], and CI for adjusted sensitivity and specificity were obtained using the Wilson Score method[41]. Statistical analysis was done using Python 3.7 statistical library, and a P-value of 0.05 was considered statistically significant.

Ethics approval and consent

Ethical approval for this study was obtained from the Institutional Ethics committee for Observational studies of Jawaharlal Institute of Postgraduate Medical Education and Research (JIPMER), Puducherry, India (JIPMER ethics committee number JIP/IEC/2019/533). Waiver of written informed consent was approved by JIPMER institutional ethics committee as all data sources used (patient demography, laboratory records, and chest radiography) were previously available, and no patients needed to be contacted. Additionally, all data were collected anonymously and de-identified using study identifier before reading by the radiologist, model development, validation, and testing. All methods were carried out in accordance with Indian Council of Medical Research (ICMR) and International Committee on Harmonization of Good Clinical Practice (ICH-GCP) guidelines and regulations.

Results

The SH collection consists of 662 PA CXR in which 336 CXR having various manifestations of TB, and 326 CXR were healthy control patients with 35 pediatric (<=18 years of age) CXR. In the MC collection, 138 PA CXR with 13% were pediatric, resulting in 58 abnormal CXR and 80 healthy patients. NIAID collections entail 1678 abnormal images with 45 pediatric images. IN dataset entails 4392 PA chest X-rays consisting of 30% abnormal images and 70% normal images, with 3% (172) images in the collections being pediatric. The exclusion criteria for CXR are given in Supplementary Table S1. We excluded patients under 15 years of age from the training of model as per the new WHO guidelines[6]. We also excluded inactive TB and Pneumonia images from SH and MC collections as the objective of this study is to find the diagnostic accuracy of active TB cases.

Demographic characteristics of TB subjects and healthy controls of all datasets are presented in Table 2. Among the active TB case and healthy control in all datasets, the female distribution of CXR was significantly different except for the SH dataset. The mean age (SD) was significantly different among active TB cases and healthy controls in all datasets. Multiple images from the same patients were predominantly seen in TB CXR compared to control CXR. Except for the MC dataset, all three other datasets are balanced with the number of CXR images in each class.

Table 3 shows the estimates of diagnostic accuracy of Resnet50 and Densenet121 tested on the intramural and external test set. The highest sensitivity and specificity across models tested with intramural test set was 0.985 [0.971 – 1.000] for both models trained and tested IN dataset and for specificity, it was 0.988 [0.977 – 0.999] Densenet121 trained and tested NIAID. The range of sensitivity for intramural test set on Resnet50 was 0.627 – 0.985 while Densenet121 ranges from 0.611 – 0.985. The specificity for intramural test set varies 0.730 – 0.986 for Resnet50 and Densenet121 it fluctuates 0.638 – 0.988. However, sensitivity and specificity vary drastically in extramural test sets across models. In Resnet50, sensitivity and specificity for extramural test set varies from 0.152 – 0.994 and 0.055 – 0.974 respectively. Performance of Densenet121 in extramural for sensitivity and specificity was 0.138 – 0.994 and 0.012 – 0.998 correspondingly. Across model intramural test set exabits over 95% sensitivity and specificity for IN and NIAID datasets while SH and MC diverge from 61 % to 91%. In dataset wise, extramural case, MC and IN dataset shows promising results in sensitivity and specificity with both models trained in SH. The AUC across intramural across model Resnet50 was 1.000 and for extramural Densenet121 trained in NIAID shows AUC of 0.976 on IN dataset.

Figure 2 reveals the Receiver Operating Characteristic (ROC) curve for intramural and extramural test sets using Resnet50 and Densenet121. Supplementary figure S3 portrays the comparison of 95% CI of AUC across Resnet50 and Densenet121 models. Its observed intramural test set used in the MC dataset was less than 10 in TB and Healthy normal CXR and thus had wide CI. It is evident from Table 3 the evaluation matric results that the Resnet50 model trained on SH and tested in NIAID showed the worst performance matric compared to all other extramural performance matric. Overall accuracy for Resnet50 was 30.02% – 100%. Precision across all models ranges from 21.55% - 100%, with the

lowest in the Resnet50 model trained in SH and tested in NIAID. The highest was the Resnet50 model trained and tested in the IN dataset, whereas Densenet121 trained model ranges from 26.67% - 100% with lowest were NIAID trained and tested in MC, and the highest was NIAID trained and tested in intramural and IN datasets. In case of F1-score, the results range from 17.78% - 100% in Resnet50 models and 14.81% - 99.62% in Densenet121 models. Details of accuracy, precision, and F1-score are available in Supplementary Table S3.

Discussion

The study aimed to analyze the efficacy of transfer learning and training on one geographical dataset impacted the diagnostic accuracy of different geographical dataset and their performance estimates. Our experiments show that sensitivity, specificity, and AUC for intramural test sets are consistent with previous studies on detecting TB with DCNN[9,24,42]. Our cross-geographical train test experiments exhibit the instability in various performance estimates among models compared with similar published work. Our results show better performance on AUC, sensitivity, and specificity on Densenet121 trained on MC and tested on SH compared with InceptionNet V3 model trained on MC and tested on SH reported by Das, Santosh, & Pal[26], they employed histogram equalization for contrast enhancement. Santosh & Antani[10] reported better results than our experiments but they employed voting ensemble of three different classifiers and used handcrafted feature selection and lung symmetry for detecting the TB. Moreover, both works[10,26] used off-the-shelf MC and SH datasets but our experiments excluded CXR from MC and SH based on the exclusion criteria.

A similar study reported[25] a considerable drop in the performance of AUC on the SH trained InceptionNet V3 model tested on the ChextX-ray8 dataset even though the authors employed several data augmentation methods to improve the generalisability. Cross dataset experiments reported by Hwang et.al[43] used modified AlexNet trained over 14,000 iterations with Korean Institute of Tuberculosis dataset (>10000 data points) reported an AUC of 0.884 for MC and SH was 0.926 better than our Resnet50 trained on IN (1050 data points) with AUC 0.784 for MC and SH was 0.865. It is observed that models trained on NIAID tested on intramural and extramural IN dataset depicts similar performance matric because India (College of Pharmaceutical Sciences, Manipal, India) also contributes CXR to the NIAID data set.

Data distribution shift plays a crucial role in the failure of machine learning systems in terms of generalisability. Although extensive data augmentation can generally improve generalization, in radiological datasets, the use of unfitting data augmentation can lead to an adverse effect on model learning (horizontal flipping of CXR inadvertently creates a medical condition called situs inversus). We also checked the dataset used in commercial CE-marked CAD products on the market for diagnosing Tuberculosis. Genki, reported 1,500,000 CXR from 10 countries used for training and tested over 30 different imaging machines, CAD4TB trained with 1,000,000 CXR from several countries and continents, qXR reported 3,700,000 CXR acquired globally used for training and tested over 20 leading CXR imaging machines, and JLK-JLD-02K reports 1,500,000 training data composed from five countries mainly from southeast Asia region[44]. This indicate use of extremely large dataset can eliminate the data shift in predictive modelling. Tang, Y.-X. *et. al*[12] *also* reports the use of large dataset can diminish the presence of dataset shift.

Findings from our study indicate that it was unlikely to construct an accurate deep learning model using transfer learning with Resnet50 and Densenet121 trained and validated on the same dataset to detect the active TB in geographically different population datasets. By far best, even though lung pathology is consistent across the population for active Tuberculosis, the discrepancy in the specification and standard operating procedure for medical imaging machines and the underlying image manipulation techniques to produce Digital Radiography and Computed Radiography modality are inevitably having an essential role in performance shift in cross-population train and test method. Moreover, the technical quality of the CXR, image conversion (DICOM to JPG), and image resolution also play a crucial role in classification accuracy. Different modalities and equipment exist for image acquisition in the dataset used for our experiment, which potentially influenced the learning and classification for active TB.

There are some limitations of the study. First, we didn't use any data augmentation methods and k-fold cross-validation. We used a relatively small dataset (MC,109 instances) and mid-size (IN, 2655 instances) in our experiments. The study's strength was to use multiple datasets across different geographical regions, state-of-the-art deep learning architecture, and hyperparameter tuning using hyperband Keras tuner to optimize model parameters.

Conclusion

This work has addressed the efficacy and fidelity of using transfer learning on geographically different train and test datasets for detecting TB. We demonstrated the cross-population test on two state-of-the-art models without any data augmentation for classification accuracy. The results revealed significant variance in AUC, sensitivity, specificity, accuracy, and other potential measures on where train and test datasets are different. The paucity of Tuberculosis dataset and their annotations in public domain was the main hindrance for many researchers in this field. Further investigation on this area combines CXR from multiple countries or CXR generated from different imaging machines and modalities to use deep learning in real-world TB screening scenarios.

Declarations

Funding statement

This work was supported by an intramural grant from the Jawaharlal Institute of Postgraduate Medical Education and Research for Ph.D. work.

Acknowledgments

The authors wish to acknowledge the Indo US TB projects under the Department of Preventive and Social Medicine, Jawaharlal Institute of Postgraduate Medical Education and Research (JIPMER), for extended support to complete this study.

Disclosure statement

The authors declare no potential conflict of interest.

Additional Information

We used CLAIM checklist to prepare the manuscript. Also, refer to supplementary materials for additional tables and figures.

Data Availability

The dataset used in this study, Montgomery and Shenzhen, are available at https://data.lhncbc.nlm.nih.gov/public/Tuberculosis-Chest-X-ray-Datasets/Montgomery-County-CXR-Set/MontgomerySet/index.html and https://data.lhncbc.nlm.nih.gov/public/Tuberculosis-Chest-X-ray-Datasets/Shenzhen-Hospital-CXR-Set/index.html. TB Portals Program data maintained by the National Institute of Allergy and Infectious Diseases available at https://tbportals.niaid.nih.gov/. Due to specific institutional requirements governing privacy protection, the Indian dataset were not available in public.

Code Availability

Our implementation is available at https://github.com/jamestd-td/dataset_shift

Author contributions statement

J.D, S.L, S.A, and M.R conceived the study. S.A, M.R, and H.G developed the CXR reporting format. J.D collected the CXR with the support of S.A and M.R. H.G read all CXR and S.A and M.R peer validated CXR. J.D developed the model, carried out numerical experiments, analyzed the data, and wrote the main parts of the manuscript. H.G and S.A provided expert radiology advice. M.R provided expert pulmonology/ Tuberculosis advice. A.B provided expert Deep Learning advice and code valuation. S.L verified the analysis and figures. All authors reviewed the manuscript and contributed to the final manuscript.

References

1. WHO, G. Global tuberculosis report 2020. *Glob Tuberc Rep* **2020**, (2020).
2. Assembly, U. G. *Political declaration of the high-level meeting of the General Assembly on the fight against tuberculosis*. https://www.un.org/en/ga/search/view_doc.asp?symbol=A/RES/73/3 (2018).
3. Nachiappan, A. C. *et al.* Pulmonary Tuberculosis: Role of Radiology in Diagnosis and Management. *RadioGraphics* **37**, 52–72 (2017).
4. World Health Organization. *Chest radiography in tuberculosis detection: summary of current WHO recommendations and guidance on programmatic approaches*. (World Health Organization, 2016).
5. van't Hoog, A. H. *et al.* Screening Strategies for Tuberculosis Prevalence Surveys: The Value of Chest Radiography and Symptoms. *PLOS ONE* **7**, e38691 (2012).
6. World Health Organization. WHO consolidated guidelines on tuberculosis: module 2: screening: systematic screening for tuberculosis disease. (2021).
7. Hoog, A. H. van't *et al.* High sensitivity of chest radiograph reading by clinical officers in a tuberculosis prevalence survey. *Int. J. Tuberc. Lung Dis. Off. J. Int. Union Tuberc. Lung Dis.* **15 10**, 1308–14 (2011).
8. Melendez, J. *et al.* An automated tuberculosis screening strategy combining X-ray-based computer-aided detection and clinical information. *Sci. Rep.* **6**, 25265 (2016).
9. Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D. & Pfeiffer, D. Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. *Sci. Rep.* **9**, 6268 (2019).

10. K. C. Santosh & S. Antani. Automated Chest X-Ray Screening: Can Lung Region Symmetry Help Detect Pulmonary Abnormalities? *IEEE Trans. Med. Imaging* **37**, 1168–1177 (2018).
11. Rajpurkar, P. *et al.* CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *Npj Digit. Med.* **3**, 115 (2020).
12. Tang, Y.-X. *et al.* Automated abnormality classification of chest radiographs using deep convolutional neural networks. *Npj Digit. Med.* **3**, 70 (2020).
13. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems* (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) vol. 25 (Curran Associates, Inc., 2012).
14. S. Jaeger *et al.* Automatic Tuberculosis Screening Using Chest Radiographs. *IEEE Trans. Med. Imaging* **33**, 233–245 (2014).
15. Vajda, S. *et al.* Feature Selection for Automatic Tuberculosis Screening in Frontal Chest Radiographs. *J. Med. Syst.* **42**, 146 (2018).
16. S. Rajaraman *et al.* A novel stacked generalization of models for improved TB detection in chest radiographs. in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* 718–721 (2018). doi:10.1109/EMBC.2018.8512337.
17. Jaeger, S. *et al.* Detecting drug-resistant tuberculosis in chest radiographs. *Int. J. Comput. Assist. Radiol. Surg.* **13**, 1915–1925 (2018).
18. Reamaroon, N. *et al.* Robust segmentation of lung in chest x-ray: applications in analysis of acute respiratory distress syndrome. *BMC Med. Imaging* **20**, 116 (2020).
19. Schultheiss, M. *et al.* A robust convolutional neural network for lung nodule detection in the presence of foreign bodies. *Sci. Rep.* **10**, 12987 (2020).
20. X. Wang *et al.* ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 3462–3471 (2017). doi:10.1109/CVPR.2017.369.
21. Moreno-Torres, J. G., Raeder, T., Alaiz-Rodríguez, R., Chawla, N. V. & Herrera, F. A unifying view on dataset shift in classification. *Pattern Recognit.* **45**, 521–530 (2012).
22. Qui??onero-Candela, J., ,. *Dataset shift in machine learning*. (MIT Press, 2009).
23. Becker, A. & Becker, J. Dataset shift assessment measures in monitoring predictive models. *Knowl.-Based Intell. Inf. Eng. Syst. Proc. 25th Int. Conf. KES2021* **192**, 3391–3402 (2021).
24. Lakhani, P. & Sundaram, B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology* **284**, 574–582 (2017).
25. Sathitratanacheewin, S., Sunanta, P. & Pongpirul, K. Deep learning for automated classification of tuberculosis-related chest X-Ray: dataset distribution shift limits diagnostic performance generalizability. *Heliyon* **6**, e04614 (2020).
26. D. Das, K. C. Santosh, & U. Pal. Cross-Population Train/Test Deep Learning Model: Abnormality Screening in Chest X-Rays. in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)* 514–519 (2020). doi:10.1109/CBMS49503.2020.00103.
27. Dockès, J., Varoquaux, G. & Poline, J.-B. Preventing dataset shift from breaking machine-learning biomarkers. *GigaScience* **10**, (2021).
28. Guo, L. L. *et al.* Evaluation of Domain Generalization and Adaptation on Improving Model Robustness to Temporal Dataset Shift in Clinical Medicine. *medRxiv* 2021.06.17.21259092 (2021) doi:10.1101/2021.06.17.21259092.
29. Glavatskikh, M., Leguy, J., Hunault, G., Cauchy, T. & Da Mota, B. Dataset's chemical diversity limits the generalizability of machine learning predictions. *J. Cheminformatics* **11**, 69 (2019).
30. Stacke, K., Eilertsen, G., Unger, J. & Lundström, C. *A closer look at domain shift for deep learning in histopathology*. (2019).
31. Jaeger, S. *et al.* Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quant. Imaging Med. Surg. Vol 4 No 6 Dec. 2014 Quant. Imaging Med. Surg.* (2014).
32. Rosenthal Alex *et al.* The TB Portals: an Open-Access, Web-Based Platform for Global Drug-Resistant-Tuberculosis Data Sharing and Analysis. *J. Clin. Microbiol.* **55**, 3267–3282 (2017).
33. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. in 234–241 (Springer, 2015).
34. K. He, X. Zhang, S. Ren, & J. Sun. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016). doi:10.1109/CVPR.2016.90.
35. G. Huang, Z. Liu, L. Van Der Maaten, & K. Q. Weinberger. Densely Connected Convolutional Networks. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2261–2269 (2017). doi:10.1109/CVPR.2017.243.
36. Landis, J. R. & Koch, G. G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **33**, 159–174 (1977).
37. A. S. B. Reddy & D. S. Juliet. Transfer Learning with ResNet-50 for Malaria Cell-Image Classification. in *2019 International Conference on Communication and Signal Processing (ICCSP)* 0945–0949 (2019). doi:10.1109/ICCSP.2019.8697909.

38. O'Malley, T. *et al.* Keras Tuner. *Githubaccessed 31 January 2021* (2019).
39. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *ArXiv Prepr. ArXiv14126980* (2014).
40. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36 (1982).
41. Wilson, E. B. Probable Inference, the Law of Succession, and Statistical Inference. *J. Am. Stat. Assoc.* **22**, 209–212 (1927).
42. Xie, Y. *et al.* Computer-Aided System for the Detection of Multicategory Pulmonary Tuberculosis in Radiographs. *J. Healthc. Eng.* **2020**, 9205082–9205082 (2020).
43. Hwang, S., Kim, H.-E., Jeong, J. & Kim, H.-J. A novel approach for tuberculosis screening based on deep convolutional neural networks. in vol. 9785 97852W (International Society for Optics and Photonics, 2016).
44. Naheyan, T. AI Products for Tuberculosis Healthcare. *AI4HLTH* https://www.ai4hlth.org.

Tables

Table 1 Overview of Chest X-Ray Datasets used for training and testing of active tuberculosis and its characteristics. IN - JIPMER Hospital, India, MC - Montgomery County Hospital, USA, SH- Shenzhen hospital, China, NIAID - National Institute of Allergy and Infectious Diseases, USA. TIFF - Tag Image File Format, PNG - Portable Network Graphics, DICOM - Digital Imaging and Communications in Medicine. BPP - Bits Per Pixel. DR – Digital Radiography, CR – Computed Radiography. * Please refer Supplementary Table S1 for detailed exclusion criteria. ** Net images used for model training, validation, intramural test and extramural test.

Dataset	No of abnormal images			No of healthy control images			File type	BPP	Resolution (approx.)	Modality	Medical Imaging Machine
	Original	Excluded*	Net images**	Original	Excluded*	Net images**					
IN	1350	38	1312	3042	27	3015	TIFF	24	1530x1896 / 4280x3520	DR	Portable Samsung Retrofit DR System
MC	58	19	39	80	10	70	PNG	12	4020x4892 / 4892x4020	CR	Eureka stationary X-Ray systems
SH	336	11	325	326	23	303	PNG	8	3000x3000	DR	Philips DR Digital Diagnose system
NIAID	1678	6	1672	0	0	0	DICOM	16	2248x2724 / 2724x2248	CR	Not available

Table 2 Demographic Characteristics of Subjects in Chest-X-Ray Datasets. CXR – Chest Radiography. SD – Standard Deviation. * Healthy controls taken from IN collection. Bold letters denote statistical significance at 95% Confidence Level (P<0.05)

Data set	Characteristic	Chest X-Rays		P Value
		No. of case positive for TB	No. of healthy control patients	
JIPMER Hospital, India (IN) (n= 2655)	No of patients	837	1335	
	No of CXR	1312	1343	
	Mean age in years (SD)	46.33 (14.49)	52.75 (16.62)	**<0.0001**
	Gender, Female n(%)	244 (18.5)	640(47.65)	**<0.0001**
Shenzhen hospital, China (SH)(n=628)	No of patients	323	303	
	No of CXR	325	303	
	Mean age in years (SD)	37.93 (14.86)	35.32 (12.15)	**0.0150**
	Gender, Female n(%)	102 (31.38)	97 (32.01)	0.8650
Montgomery County Hospital, USA (MC) (n=109)	No of patients	32	66	
	No of CXR	39	70	
	Mean age in years (SD)	47.58(20.58)	37.92(12.81)	**0.0100**
	Gender, Female n(%)	13(33.33)	48(68.57)	**0.0003**
NIAID TB Portals Program (NIAID) (n=1672)	No of patients	786	1655*	
	No of CXR	1672	1672	
	Mean age in years (SD)	40.16 (14.25)	51.62 (15.61)	**<0.0001**
	Gender, Female n(%)	607 (36.3)	674 (40.3)	**0.0171**

Table 3 Estimates of diagnostic accuracy of intramural and extramural test sets in Resnet50 and DenseNet121 models. IN - JIPMER Hospital, India, MC - Montgomery County Hospital, USA, SH- Shenzhen hospital, China, NIAID - National Institute of Allergy and Infectious Diseases, USA. AUC – Area Under the Receiver Operating Characteristic. CI – Confidence Interval. R50 – Resnet50 Model. D121 – DenseNet121 Model. * Intramural Test Set. # Extramural Test Set. 95% CI for AUC calculated using Hanley & McNeil method. CI for Sensitivity and Specificity calculated using Wilson Score method.

Train	Test	AUC [95% CI]		Sensitivity [95% CI]		Specificity [95% CI]	
		R50	D121	R50	D121	R50	D121
SH	SH*	0.932 [0.866 – 0.997]	0.914 [0.840 – 0.987]	0.695 [0.546 – 0.844]	0.611 [0.452 – 0.770]	0.854 [0.743 – 0.965]	0.913 [0.833 – 0.994]
	MC#	0.724 [0.620 – 0.827]	0.792 [0.698 – 0.885]	0.675 [0.535 – 0.814]	0.791 [0.673 – 0.910]	0.608 [0.497 – 0.719]	0.648 [0.540 – 0.757]
	IN#	0.848 [0.833 – 0.862]	0.844 [0.828 – 0.859]	0.737 [0.713 – 0.760]	0.648 [0.623 – 0.674]	0.810 [0.789 – 0.831]	0.882 [0.864 – 0.899]
	NIAID#	0.254 [0.237 – 0.270]	0.556 [0.536 – 0.575]	0.152 [0.134 – 0.169]	0.251 [0.230 – 0.271]	0.449 [0.425 – 0.473]	0.772 [0.752 – 0.792]
MC	MC*	0.821 [0.531 – 1.000]	0.750 [0.422 – 1.000]	0.627 [0.300 – 0.954]	0.755 [0.510 – 1.000]	0.730 [0.486 – 0.974]	0.638 [0.358 – 0.917]
	SH#	0.775 [0.738 – 0.811]	0.719 [0.679 – 0.758]	0.221 [0.176 – 0.266]	0.547 [0.493 – 0.600]	0.974 [0.957 – 0.990]	0.746 [0.697 – 0.794]
	IN#	0.626 [0.604 – 0.647]	0.858 [0.843 – 0.872]	0.374 [0.348 – 0.400]	0.409 [0.382 – 0.436]	0.808 [0.787 – 0.829]	0.962 [0.952 – 0.972]
	NIAID#	0.457 [0.437 – 0.476]	0.863 [0.850 – 0.875]	0.178 [0.160 – 0.197]	0.477 [0.454 – 0.501]	0.800 [0.781 – 0.819]	0.940 [0.929 – 0.951]
IN	IN*	1.000 [1.000 – 1.000]	0.999 [0.995 – 1.000]	0.985 [0.971 – 1.000]	0.985 [0.971 – 1.000]	0.986 [0.972 – 1.000]	0.978 [0.958 – 0.998]
	SH#	0.865 [0.836 – 0.893]	0.849 [0.818 – 0.879]	0.994 [0.988 – 1.000]	0.994 [0.988 – 1.000]	0.055 [0.030 – 0.080]	0.012 [0.001 – 0.023]
	MC#	0.784 [0.688 – 0.879]	0.794 [0.700 – 0.887]	0.394 [0.248 – 0.541]	0.488 [0.338 – 0.638]	0.865 [0.790 – 0.940]	0.825 [0.740 – 0.909]
	NIAID#	0.527 [0.507 – 0.546]	0.548 [0.528 – 0.567]	0.769 [0.748 – 0.789]	0.815 [0.797 – 0.834]	0.443 [0.420 – 0.467]	0.447 [0.423 – 0.471]
NIAID	NIAID*	0.999 [0.995 – 1.000]	0.999 [0.995 – 1.000]	0.977 [0.957 – 0.996]	0.977 [0.957 – 0.996]	0.982 [0.966 – 0.998]	0.988 [0.977 – 0.999]
	SH#	0.386 [0.342 – 0.429]	0.390 [0.346 – 0.433]	0.705 [0.656 – 0.754]	0.854 [0.816 – 0.892]	0.156 [0.115 – 0.196]	0.084 [0.053 – 0.115]
	MC#	0.361 [0.255 – 0.466]	0.477 [0.364 – 0.589]	0.348 [0.206 – 0.490]	0.138 [0.040 – 0.235]	0.554 [0.440 – 0.667]	0.825 [0.740 – 0.909]
	IN#	0.903 [0.891 – 0.914]	0.976 [0.970 – 0.981]	0.607 [0.580 – 0.633]	0.534 [0.508 – 0.561]	0.972 [0.963 – 0.981]	0.998 [0.997 – 0.999]

Figures

Figure 1

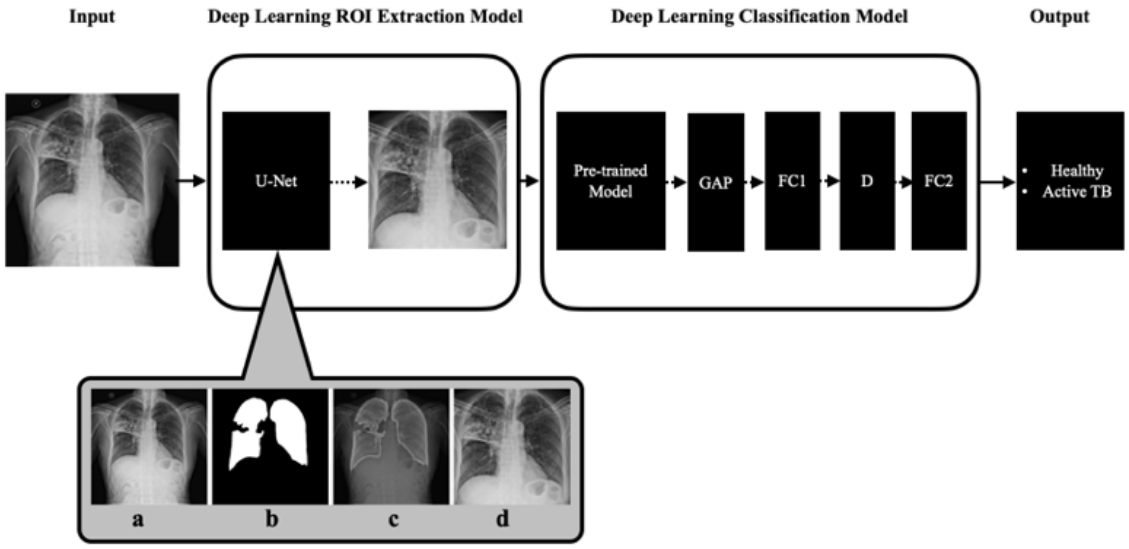

Figure 1

Basic outline of the network architecture. Input - image of chest radiograph. ROI – Region of Interest. In the ROI extraction Model: a. Original Chest X-Ray, b. Mask of lung generated by U-Net, c. Mask overlay to original chest x-ray, and d. Final segmented image of the lungs based on mask in 224x224 size. Pre-trained Model – Resnet50 and DenseNet121trained on ImageNet without the final classifier layer. GAP – Global Average Pooling. FC – Fully Connected Layer. D – Dropout Layer. Due to complexity we omitted layer and classifier activation functions.

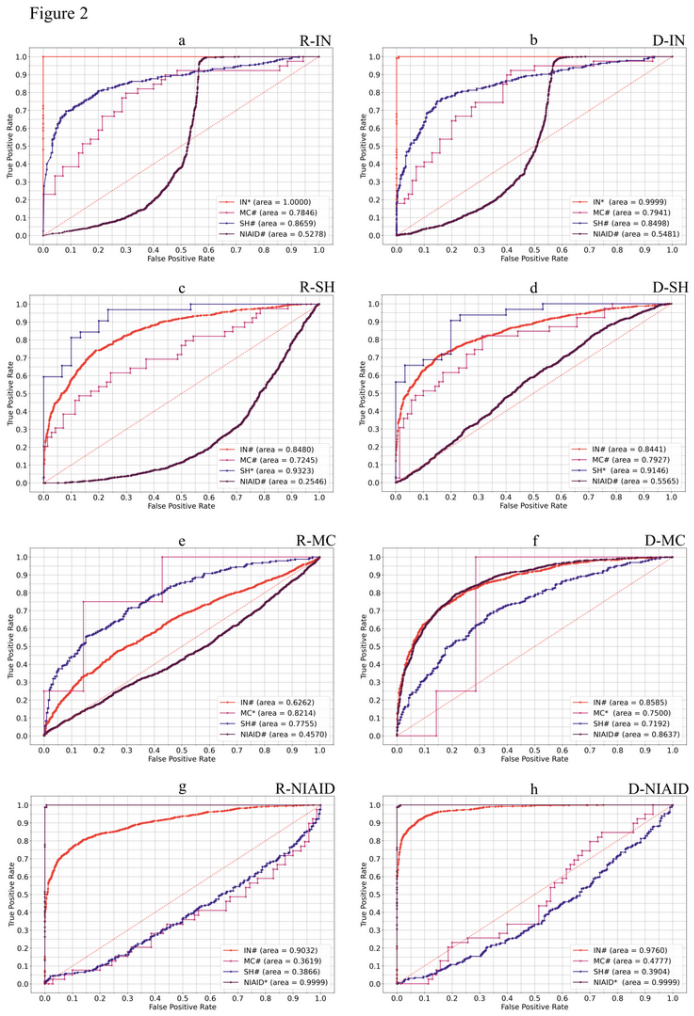

Figure 2

Receiver Operating Characteristic of intramural and extramural test sets. IN - JIPMER Hospital, India, MC - Montgomery County Hospital, USA, SH- Shenzhen hospital, China, NIAID - National Institute of Allergy and Infectious Diseases, USA. a. Resnet50 trained on IN dataset (R-IN). b. DenseNet121 trained on IN dataset (D-IN). c. Resnet50 trained on SH dataset (R-SH). d. DenseNet121 trained on SH dataset (D-SH). e. Resnet50 trained on MC dataset (R-MC). f. DenseNet121 trained on MC dataset (D-MC). g. Resnet50 trained on NIAID dataset (R-NIAID). h. DenseNet121 trained on NIAID dataset (D-NIAID). * Intramural test set. # Extramural test set.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplimentarymaterial.pdf