Original Article

Deep learning-based pulmonary tuberculosis automated detection on chest radiography: large-scale independent testing

Wen Zhou[1,2#], Guanxun Cheng[1#], Ziqi Zhang[3], Litong Zhu[4], Stefan Jaeger[5], Fleming Y. M. Lure[6], Lin Guo[6]^

[1]Department of Radiology, Peking University Shenzhen Hospital, Shenzhen, China; [2]Department of Radiology, Peking University First Hospital, Beijing, China; [3]Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen, China; [4]Department of Medicine, Queen Mary Hospital, University of Hong Kong, Hong Kong, China; [5]National Library of Medicine, National Institutes of Health, Bethesda, MD, USA; [6]Shenzhen Smart Imaging Healthcare Co., Ltd., Shenzhen, China

Contributions: (I) Conception and design: L Guo, W Zhou; (II) Administrative support: G Cheng, FYM Lure; (III) Provision of study materials or patients: W Zhou, G Cheng; (IV) Collection and assembly of data: W Zhou, Z Zhang, FYM Lure; (V) Data analysis and interpretation: Z Zhang, L Zhu, S Jaeger; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Lin Guo. No. 1, Qingxiang Road, Longhua District, Shenzhen, China. Email: guolin913@outlook.com.

Background: It is critical to have a deep learning-based system validated on an external dataset before it is used to assist clinical prognoses. The aim of this study was to assess the performance of an artificial intelligence (AI) system to detect tuberculosis (TB) in a large-scale external dataset.

Methods: An artificial, deep convolutional neural network (DCNN) was developed to differentiate TB from other common abnormalities of the lung on large-scale chest X-ray radiographs. An internal dataset with 7,025 images was used to develop the AI system, including images were from five sources in the U.S. and China, after which a 6-year dynamic cohort accumulation dataset with 358,169 images was used to conduct an independent external validation of the trained AI system.

Results: The developed AI system provided a delineation of the boundaries of the lung region with a Dice coefficient of 0.958. It achieved an AUC of 0.99 and an accuracy of 0.948 on the internal data set, and an AUC of 0.95 and an accuracy of 0.931 on the external data set when it was used to detect TB from normal images. The AI system achieved an AUC of more than 0.9 on the internal data set, and an AUC of over 0.8 on the external data set when it was applied to detect TB, non-TB abnormal and normal images.

Conclusions: We conducted a real-world independent validation, which showed that the trained system can be used as a TB screening tool to flag possible cases for rapid radiologic review and guide further examinations for radiologists.

Keywords: Tuberculosis detection; deep learning; chest radiography; external validation; large-scale test

Submitted Jul 05, 2021. Accepted for publication Dec 23, 2021; Published online: 24 Jan 2022.
doi: 10.21037/qims-21-676
View this article at: https://dx.doi.org/10.21037/qims-21-676

^ ORCID: 0000-0002-5910-3464.

Introduction

Tuberculosis (TB) is a pulmonary infectious disease caused by Mycobacterium tuberculosis. The World Health Organization (WHO) estimated that 1.5 million people died of TB in 2018 (1). TB has become the second leading cause of infectious disease death (2). The key methods to prevent and control TB are rapid and accurate diagnoses and timely treatments. Chest X-ray is one of the most commonly used methods to detect TB (3), and it has the potential to allow screening for TB at an early stage.

Artificial intelligence (AI) has played an increasingly important role in medical imaging. LeCun *et al*. (1998) proposed the multilayer network LeNet-5 after Waibel *et al*. (1989) developed the first convolutional neural network (CNN) for speech recognition (4,5). Since then, CNNs have become the subject of much investigation (6,7). With consistent improvements in computer and AI technologies, advanced AI techniques have been integrated into medical big data and used as a way of assisting diagnosis. Meanwhile, the accuracy, specificity and speed of diagnoses have also been improved. Many studies have shown that artificial intelligence can help classify lesions and identify the nature of lesions (8,9).

Research in the application of deep learning to radiology (10,11) is a rapidly growing field as a result of its promising performance in disease detection, such as pleural effusion and cardiomegaly detection on chest radiographs (12,13), and mediastinal lymph node and lung nodule detection on computed tomography (CT) (14,15). Researchers have realized that AI-based chest X-ray (CXR) is a very promising tool for diagnosing TB, especially in resource-limited rural areas (10). In a study where the AlexNet and GoogLeNet networks were used to detect TB on chest radiographs, 1,007 chest radiographs were used for two different deep CNN tests, and the final result showed that the best-performing classifier achieved an AUC of 0.99 (10). Reviews of different computer-aided diagnosis (CAD) methods for TB can be found in the studies of Jaeger *et al*. (16) and Fatima and Shah (17). These CAD systems can automatically score the TB likelihood of each chest image. However, the test datasets from many previous studies had either a small sample size or were acquired from a single source; furthermore, most of the studies conducted model validation only on internal datasets, which might not have comprehensively evaluated the performance of the developed AI models (18). Therefore, the purpose of our study was to develop an AI system for TB detection on chest radiographs and to validate its performance using both internal and external datasets. We present the following data in accordance with the STARD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-21-676/rc).

Methods

Data

This study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). This study was approved by the Peking University Shenzhen Hospital Institutional Review Board with a waiver of informed consent. All patient identification on internal and external datasets were removed. The internal dataset was gathered from five different sources and separated into training, validation, and testing subsets to develop the AI system (*Figure 1*). Among these images, 2,736 images were of TB images, 2,169 images were of non-TB abnormal images, and 2,120 images were normal images. Patients of separated subsets were different and exclusive from each other. Then, 358,169 images of the external dataset were also read by the developed AI system to conduct independent validations (*Figure 1*). The external dataset included posterior-anterior (PA) or anteroposterior (AP) chest X-ray radiographs, among which there were 3,355 pulmonary TB images, 251,495 non-TB abnormal images and 103,319 normal images. In this study, TB cases were bacteriologically confirmed by at least one positive laboratory investigation including specimens of sputum, Xpert MTB/RIF, bronchoalveolar lavage and endotracheal aspirate. Non-TB abnormal cases were verified by medical records from hospitals, which included clinical information along with laboratory and radiological findings, and normal cases were judged by radiological reports and a 2-year follow-up examination. All radiological reports were made by consensus readings by two experienced radiologists with at least 10 years of radiological experience.

Internal data set

The internal dataset was collected from two public datasets [CHNCXR and MCUCXR (19)] and three different Chinese organizations, which were labeled Site 1, Site 2, and Site 3, respectively. Site 1, Site 2 and Site 3 were all in the TB high-prevalence area of China. Site 1 was a specialist hospital for infectious diseases, having the largest proportion of TB cases among its patients (1,102/1,377, 80.0%) when compared with Site 2 (150/1,351, 11.4%)

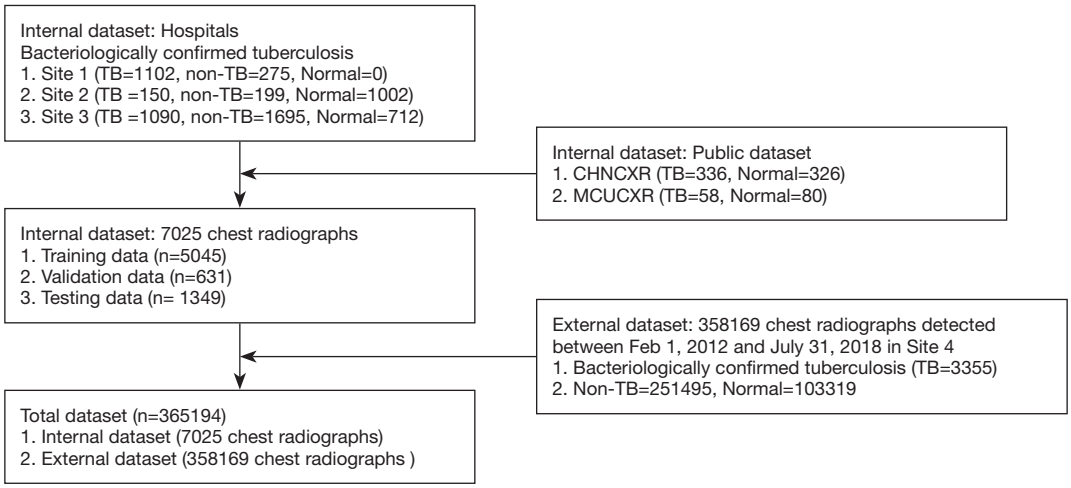


Figure 1 Flowchart for study dataset.

Table 1 Distribution of internal dataset

	Site 1	Site 2	Site 3	CHNCXR	MCUCXR	Subtotal
TB	1,102	150	1,090	336	58	2,736
Non-TB	275	199	1,695	0	0	2,169
Normal	0	1,002	712	326	80	2,120
Total	1,377	1,351	3,497	662	138	7,025

Table 2 Distribution of training, validation, and testing data generated from internal data

	TB	Non-TB	Normal	Subtotal
No. training data	2,186	1,734	1,125	5,045
No. validation data	273	217	141	631
No. testing data	277	218	854	1,349
Total	2,736	2,169	2,120	7,025

and Site 3 (1,090/3,497, 31.2%). A total of 7,025 images were classified in *Table 1* according to data source and type, among which 5,045 images (71.8%), 631 images (9.0%) and 1,349 images (19.2%) were randomly selected for training, validation, and testing, respectively. The detailed information is listed in *Table 2*.

External data set

As shown in *Table 3*, the external dataset of 358,169 images labeled as Site 4 was retrospectively collected from February 1, 2012, to July 31, 2018, in a low-prevalence setting (TB prevalence of 1.0%) to conduct an independent validation of the developed AI system. The gender distribution is shown in *Figure 2*. More men contract TB than women. The ratio of TB images to normal images is 3.24%. The TB distribution, normal and sickness rates relevant to age and gender are displayed in *Figure 2*. As shown in *Figure 2C*, people between 60 and 70 years of age are more likely to contract TB, with over 100 people per 100,000 contracting TB in that age group. The trained AI system in this study was installed at Site 4 to identify TB, non-TB abnormal and

Table 3 Summary of external dataset

Abnormality types	Number
Normal	103,319
Pneumonia	52,140
Aorta (widening) distortion	31,724
Pleural thickening	27,012
Nodules	26,918
Fibrosis	27,367
Bronchitis	21,911
Aortic calcification	18,655
Cardiomegaly	19,536
Effusion	11,670
Calcification	4,177
Primary/secondary tuberculosis	3,355
Pneumothorax	2,613
Emphysema	2,318
Fractures	2,269
Lung cancer	1,042
Atelectasis	1,966
Pulmonary edema	177
Total cases	358,169

normal images.

Artificial intelligence (AI)-based detection system

The AI system was developed using the following procedures. First, the chest radiographic images in DICOM format were converted into Portable Network Graphics format, and then these converted images were de-noised and resized to 500×500 pixels as training input. Augmentation was then performed by random horizontal flip, random crop, and color jittering in sequence to overcome problems caused by insufficient training data or the uneven class balance within the datasets (20,21). Specifically, the crop ratio of random crops ranged from 0.08 to 1, and color jittering was implemented by randomly changing the contrast, brightness, and saturation of images in the range of 0.8 to 1.2. Second, a U-Net-based algorithm was trained to automatically segment the lung area. CXR images worked as input, and then U-Net output the probabilities for each class based on the softmax function. The U-Net we trained consisted of four contraction and expansion processing parts. In each contraction processing, high-level abstract features were extracted by consecutive application of pairs of convolutional and pooling layers. In the expansion part, the low-level abstract features were merged with the features from the contractive part. As a result, the system provided a highly accurate delineation of the lung region boundaries with a Dice coefficient of 0.958 (22) (*Figure 3*). Finally, a transfer learning approach was applied to train a pretrained ResNet model to build the final DCNN to identify and locate TB lesions. ResNet is a network architecture that allows training a large number of layers while still achieving compelling performance (23). The DCNN was first developed by an internal dataset that involved 2,736 TB and 2,120 normal CXRs (including training, validation, and testing set), as described in *Table 1*.

Diagnostic performances of the trained system

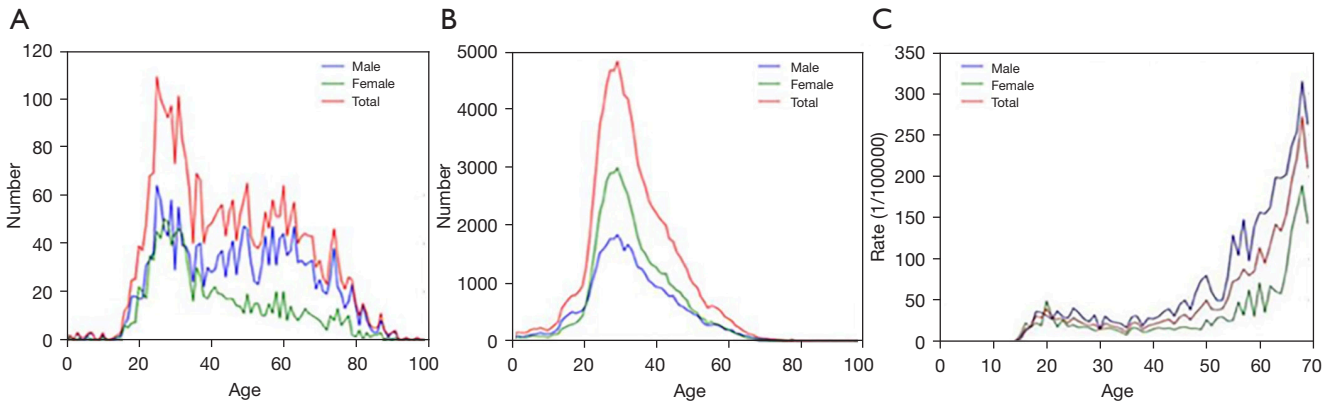
The test of the trained AI detection system was conducted as follows: (I) TB and normal images where TB CXRs were positive and normal CXRs were negative were identified. (II) TB, non-TB abnormal and normal images where TB and non-TB abnormal CXRs were positive and normal CXRs were negative, were identified. These three datasets did not overlap. The first experiment was to test the AI system's capability to detect TB and normal images, and the second experiment was to evaluate its performances to identify TB, non-TB abnormal and normal images.

TB and normal images detection by the trained system

The trained system was applied to detect TB and normal images on internal and external datasets. First, 277 TB and 854 normal CXR images of the internal dataset were involved in the test. Second, the system was validated by an external dataset that included 3,355 TB and 103,319 normal CXR images. Finally, indicators of the system's performances were recorded and analyzed.

TB, non-TB and normal image detection by the trained system

TB, non-TB and normal images from the internal and external datasets were used in this section. Data collected from Site 4, as shown in *Table 3*, were read to conduct independent validation of AI performance, and different



D

Disease	Total images	No. of males	No. of femals	Imaging time	Comorbidity
TB	3,355	2,154	1,201	2012-2018	56.48%
Normal	103,319	44,653	58,666	2012-2018	0%
Ratio #TB/#Normal	3.24%	4.82%	2.05%	NA	NA

Figure 2 Description for study population. (A) Distribution of TB. (B) Normal rates relevant to age and gender for the external dataset. (C) Disease rates relevant to age and gender for the external dataset. (D) Corresponding table displaying the distribution. TB, tuberculosis.

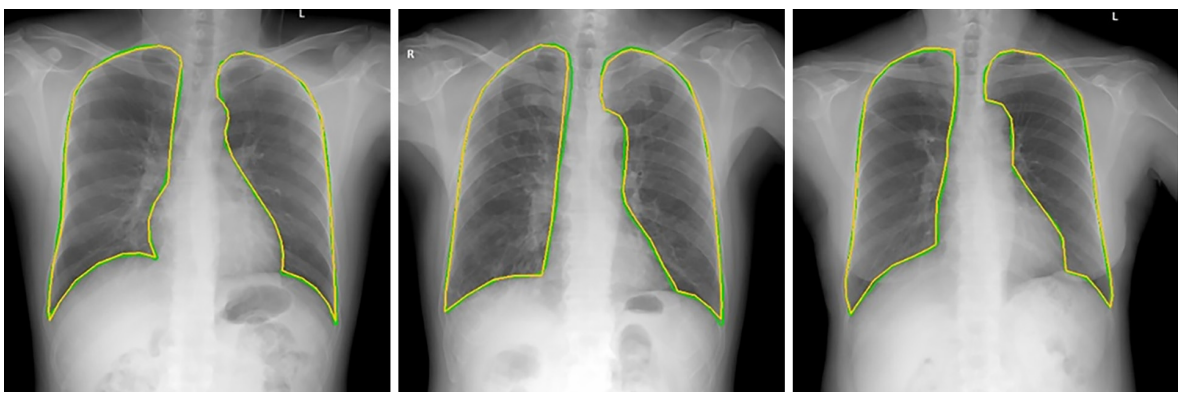


Figure 3 Results of lung segmentation using U-Net (yellow lines represent the ground truth boundaries of the lung area, and green areas denote U-Net results). The Dice coefficient of lung region segmentation is 0.958.

contrast experiments were performed to test the trained AI network. For the internal dataset, contrast experiments consisted of (A) TB images that were positive and non-TB and normal images that were negative; (B) TB images that were positive and non-TB images that were negative; and (C) abnormal (TB and non-TB) images that were positive and normal images that were negative. For the external dataset, contrast experiments were (A) positive images consisting of TB and negative images consisting of nodules, pneumonia and normal cases; (B) positive images consisting of TB and negative images consisting of nodules only; (C) positive images consisting of TB and negative images consisting of pneumonia only; (D) positive images consisting of TB, nodules, as well as pneumonia, and negative images consisting of normal cases only.

Statistical analysis

To evaluate the performances of the trained system, we illustrated test results by receiver operating characteristic (ROC) curve and the value of the area under the curve (AUC), accuracy, sensitivity (SS), specificity (SP), false-

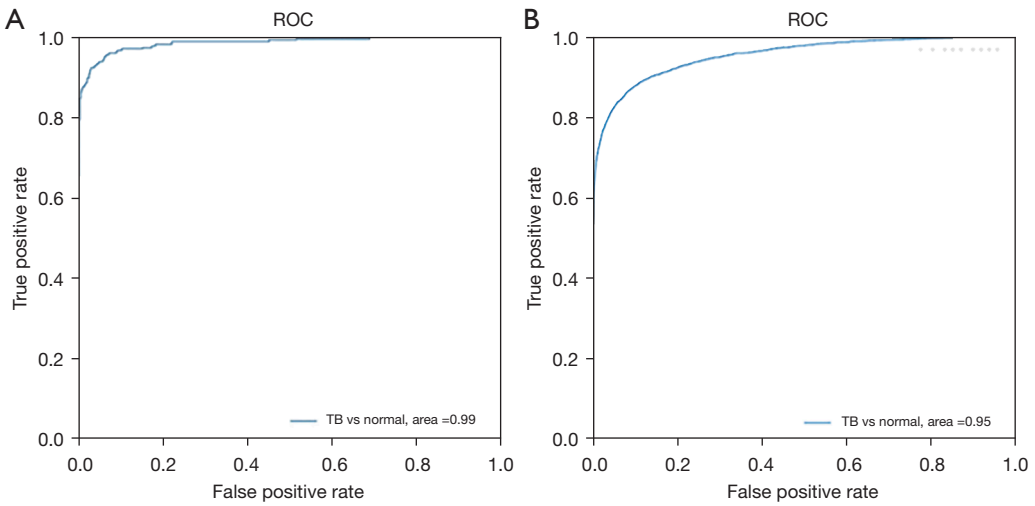


Figure 4 Testing the performance of the trained AI system. (A) The ROC curve for the internal testing dataset. The internal testing dataset contains 277 TB-positive and 854 normal cases, and the area under the ROC curve (AUC) is 0.99. (B) The ROC curve for the independent external dataset. The independent external dataset contains 3,355 TB-positive and 103,319 normal cases, and the area under the ROC curve (AUC) is 0.95. AI, artificial intelligence; TB, tuberculosis; ROC, receiver operating characteristic curve; AUC, area under the curve.

Table 4 Performance of the AI system to distinguish TB and normal cases on internal and external datasets

Independent testing	AUC (95% CI)	Accuracy (95% CI)	SS (95% CI)	SP (95% CI)	FPR (95% CI)	PPV (95% CI)	NPV (95% CI)
Internal validation	0.998 (0.981–1.004)	0.948 (0.933–0.960)	0.798 (0.746–0.841)	0.996 (0.989–0.999)	0.004 (0.001–0.011)	0.986 (0.960–0.997)	0.938 (0.921–0.952)
External validation	0.956 (0.937–0.975)	0.931 (0.914–0.945)	0.844 (0.802–0.880)	0.934 (0.917–0.948)	0.066 (0.053–0.084)	0.292 (0.265–0.322)	0.995 (0.988–0.998)

AI, artificial intelligence; TB, tuberculosis; AUC, area under the curve; SS, sensitivity; SP, specificity; FPR, false-positive rate; PPV, positive predictive value; NPV, negative predictive value.

positive rate (FPR), positive predictive value (PPV) and negative predictive value (NPV).

Results

The first experiment test

Internal data set

As shown in *Table 2*, the detection system was evaluated by the ROC curve and the AUC, as shown in *Figure 4A*. There are more performance measures listed in *Table 4*, which show that the system yielded high performance with an AUC of 0.998 (95% CI: 0.981–1.004) on the internal dataset, and its accuracy, SS, SP, FPR, PPV, and NPV were 0.948, 0.798, 0.996, 0.004, 0.986 and 0.938, respectively.

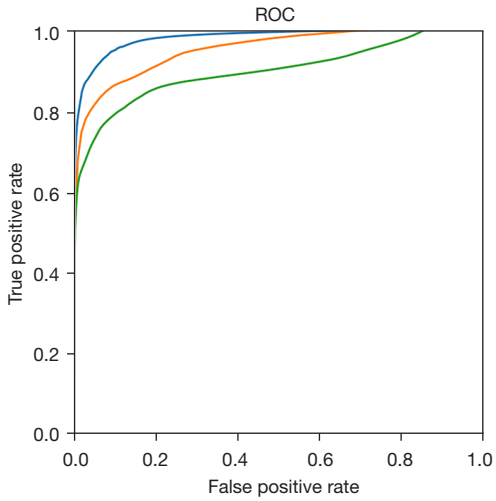
External data set

The system's performance on the external dataset is shown in *Figure 4B*, and our algorithm was again evaluated by accuracy, SS, SP, FPR, PPV, and NPV (*Table 4*), which were 0.931, 0.844, 0.934, 0.066, 0.292 and 0.995, respectively. The system's performance based on the external dataset yielded a high performance with an AUC of 0.956 (95% CI: 0.937–0.975). The AUC value of the internal validation was larger than that of the external validation, indicating better performance of the AI model on internal validation. However, overall, our trained network achieved promising accuracy when it was applied to a large and multi-disease external dataset.

The second experiment test

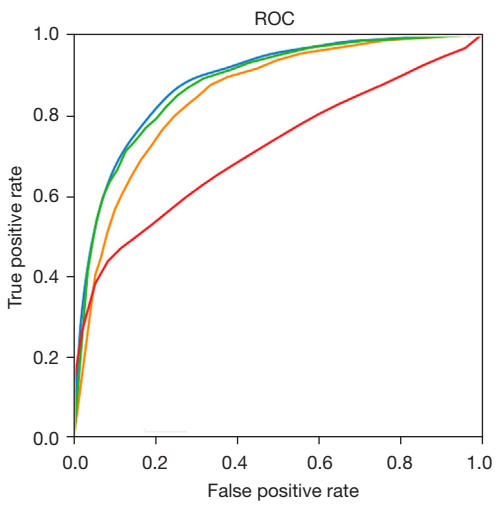
Internal data set

The ROC curves of three different contrast experiments based on the internal dataset are presented in *Figure 5*,



	Disease	Blue: Positive: TB Negative: NonTB & Normal	Orange: Positive: TB Negative: NonTB	Green: Positive: Abnormal (TB & NonTB) Negative: Normal
Test Performance	AUC (95%CI)	0.985 (0.968-1.002)	0.943 (0.924-0.962)	0.868 (0.846-0.889)
	Accuracy (95%CI)	0.946 (0.932-0.957)	0.859 (0.825-0.887)	0.805 (0.783-0.825)
	SS (95%CI)	0.798 (0.764-0.841)	0.798 (0.746-0.841)	0.475 (0.431-0.519)
	SP (95%CI)	0.984 (0.975-0.990)	0.936 (0.894-0.962)	0.997 (0.989-0.999)
	PPV (95%CI)	0.928 (0.888-0.956)	0.940 (0.902-0.965)	0.987 (0.962-0.997)
	NPV (95%CI)	0.950 (0.935-0.961)	0.785 (0.731-0.830)	0.766 (0.740-0.790)

Figure 5 Classification performance on the internal dataset for three different contrast experiments. Performance on the trained system was measured in terms of SS, SP, PPV, NPV and AUC. SS, sensitivity; SP, specificity; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve.



	Disease	Blue: Positive: TB Negative: Nodule & Pneumonia & Normal	Orange: Positive: TB Negative: Nodule	Green: Positive: TB Negative: Pneumonia	Red: Positive: TB & Nodule & Pneumonia Negative: Normal
Test Performance	AUC (95%CI)	0.883 (0.861-0.904)	0.856 (0.839-0.876)	0.880 (0.863-0.897)	0.728 (0.704-0.751)
	Accuracy (95%CI)	0.791 (0.759-0.820)	0.773 (0.724-0.815)	0.790 (0.748-0.827)	0.519 (0.482-0.556)
	SS (95%CI)	0.839 (0.821-0.856)	0.839 (0.821-0.856)	0.839 (0.821-0.856)	0.436 (0.423-0.449)
	SP (95%CI)	0.775 (0.737-0.809)	0.704 (0.681-0.726)	0.755 (0.738-0.773)	0.910 (0.892-0.925)
	PPV (95%CI)	0.549 (0.487-0.608)	0.750 (0.730-0.769)	0.707 (0.687-0.727)	0.958 (0.949-0.965)
	NPV (95%CI)	0.937 (0.909-0.957)	0.805(0.783-0.825)	0.870 (0.854-0.884)	0.255 (0.243-0.269)

Figure 6 Classification performance on external dataset. Performance on the trained system was measured in terms of SS, SP, PPV, NPV and AUC. SS, sensitivity; SP, specificity; PPV, positive predictive value; NPV, negative predictive value; AUC, area under the curve.

which shows that the AI system achieved an AUC of more than 0.9 to identify TB images. Specifically, the blue line with an AUC of 0.985 (95% CI: 0.968–1.002) represented the test to identify TB, non-TB and normal images. Its accuracy, SS, SP, PPV, and NPV were 0.946, 0.798, 0.984, 0.928 and 0.950, respectively. The orange line with an AUC of 0.943 (95% CI: 0.924–0.962) represented the test to identify TB and normal images. Its accuracy, SS, SP, PPV, and NPV were 0.859, 0.798, 0.936, 0.940 and 0.785, respectively. The green line with an AUC of 0.868 (95% CI: 0.846–0.889) represented the test to identify TB, non-TB and normal images. Its accuracy, SS, SP, PPV, and NPV were 0.805, 0.475, 0.997, 0.987 and 0.766, respectively.

External data set

The testing results are displayed in *Figure 6*. The performances on contrast groups (A) and (C) were highest with an AUC of 0.883 (95% CI: 0.861–0.904); contrast

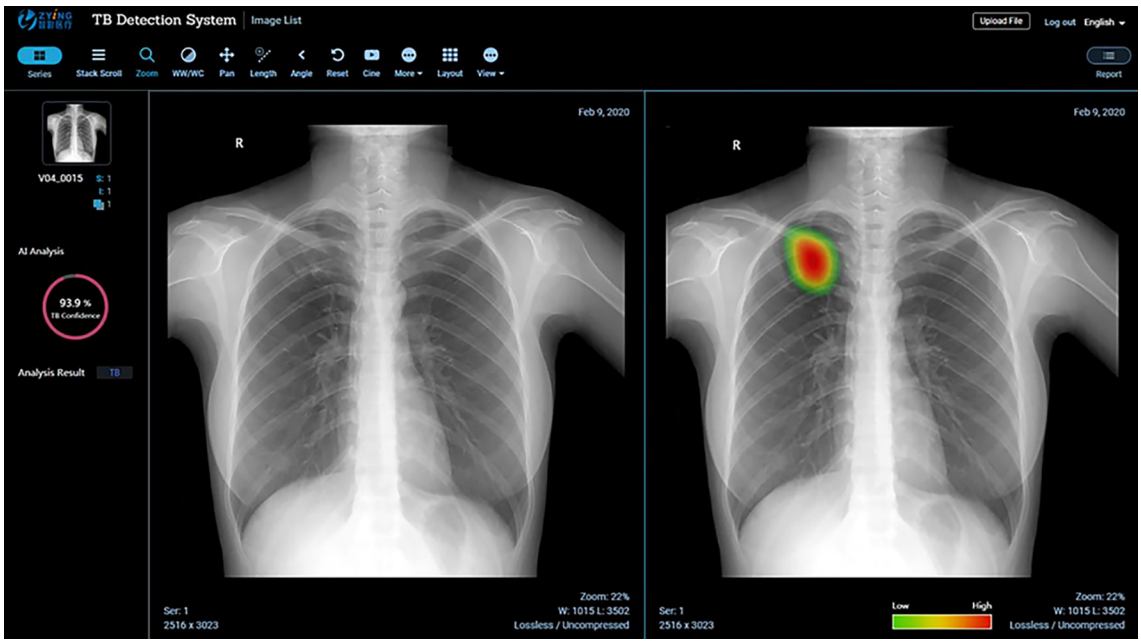


Figure 7 Detection results generated by the AI system. Both the likelihood score (between 0 and 100%) of the presence of TB and a heatmap are provided. The left panel shows the original X-ray image of the patient. The right panel shows the abnormal area with a heatmap, where red indicates the highest probability of TB and green shows the less probable TB regions. AI, artificial intelligence; TB, tuberculosis.

group (B) was slightly lower at AUC =0.856 (95% CI: 0.839–0.876); and contrast group (D) was the lowest with AUC =0.728 (95% CI: 0.704–0.751), which showed that the developed system can detect TB at very high accuracy (SS =0.839) for group (A), (B), and (C). This TB detection can also detect other non-TB abnormalities but with less accuracy. The specificity for group (D), where only normal images were used as negative, was highest at 0.910, which indicated that the system can identify normal images at high accuracy and confidence. As shown in *Figure 6*, the developed system was also able to (I) differentiate TB from other abnormalities and normal tissues; (II) differentiate between TB and nodules; (III) differentiate between TB and pneumonia; and (IV) identify TB and abnormalities in the lung with AUCs of 0.883, 0.856, 0.880, and 0.728, respectively.

Discussion

In this study, we developed a DCNN and evaluated its ability to detect TB, non-TB and normal CXRs. The AI system applied U-Net to automatically segment the lung area and used ResNet to make classifications. The detection results are presented in the form of a heatmap (*Figure 7*), and representative examples of two false-positive cases and two false negative cases detected by the AI system are shown in *Figure 8*.

In the current study, to improve the generalizability of the AI model, only those datasets that met the image quality criteria of each hospital were included, and then multicenter datasets and additional external independent validation sets were used as internal and external datasets, respectively. Moreover, we conducted image preprocessing for all the included datasets before AI modeling. The AI system was first "trained" by an internal dataset, and the results showed that it is possible to achieve a high accuracy (>0.931) on this five-source dataset. Tests based on the internal dataset yielded a high accuracy of 0.948 when the AI system was used to identify TB from normal images. Hwang also reported that their CNN model obtained an accuracy of 0.90 to identify TB on a dataset of 10,848 CXRs (24). The internal dataset in this study was composed of five different sources, which was different from previous reports where they mainly relied on publicly available CXR datasets, as creating a large annotated medical dataset is not easy (25). An accuracy of 0.931 was obtained for the external independent validation, which showed that using multisource datasets to train the DCNN may be effective in creating a robust

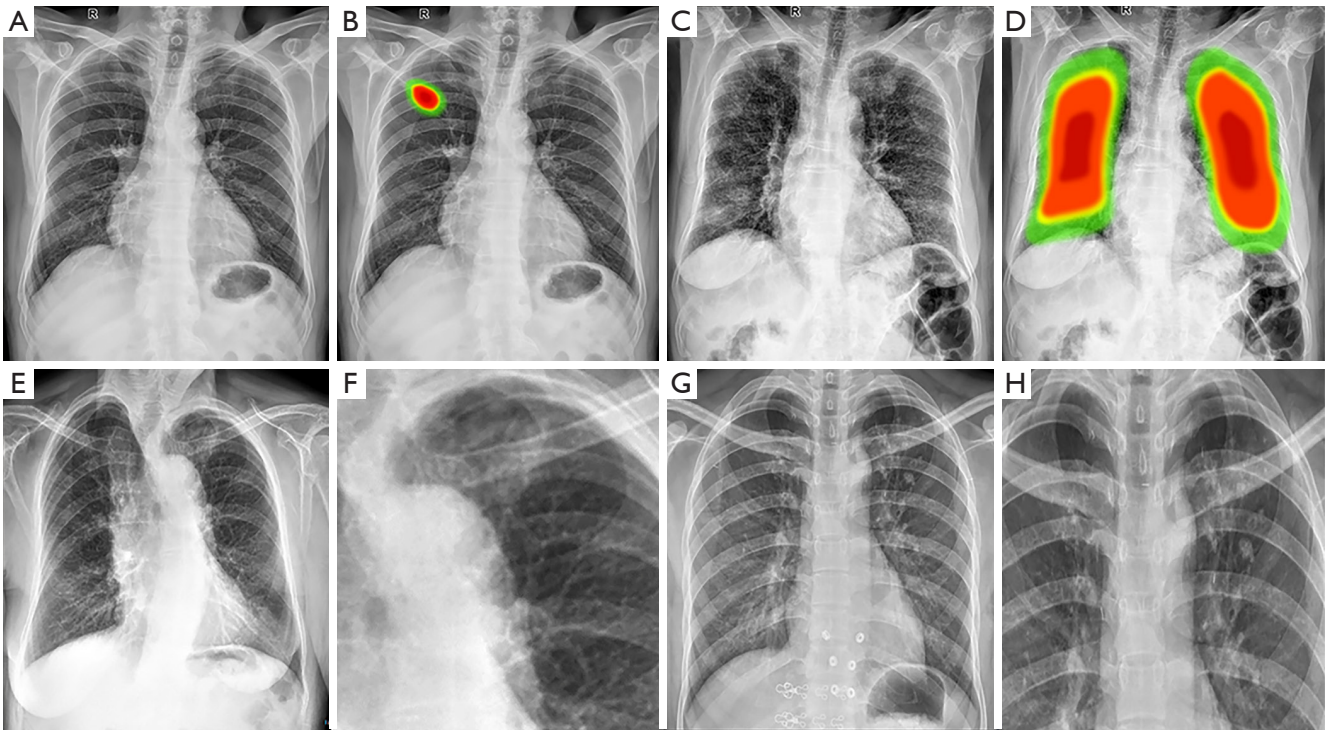


Figure 8 Examples of two false-positive cases and two false negative cases of the AI system for detecting TB from chest radiographs. (A) A normal case. (B) A false-positive case: the AI system misdiagnosed this normal case as TB with a TB lesion heatmap in the right upper lung field. (C) A chronic bronchitis case. (D) A false-positive case: the AI system misdiagnosed this chronic bronchitis as TB with lesion heatmaps in bilateral lung fields. (E) A TB case. (F) A false-negative case: the AI system misdiagnosed this TB case as normal with missed TB lesions in the left upper lung field. (G) A TB case. (H) A false-negative case: the AI system misdiagnosed this TB cases as normal with bilateral upper lung fields. AI, artificial intelligence; TB, tuberculosis.

detection system that is able to maintain its accuracy across datasets (26).

It is noteworthy that the developed AI system had better performance at the detection of TB *vs*. non-TB abnormalities on both the internal and external dataset when compared to any abnormality *vs*. normality test. Given that our algorithm was trained with a larger number of TB cases than non-TB abnormalities (2,186 *vs*. 1,734, *Table 2*), the system extracted more TB features (grayscale, shape, texture, etc.) to correctly identify TB. In addition, 1,734 cases of non-TB abnormalities were divided into nodules, pneumonia and other abnormalities, which led to each kind of abnormality having a smaller dataset to train the system. Therefore, when detecting all types of abnormalities (TB and non-TB abnormality) from normal images, the system could identify TB well but had poor performance in detecting non-TB abnormalities, resulting in the lowest AUC value among all test situations (*Figures 6*,*7*). Future studies dealing with these issues are likely to provide better performance in detecting radiologic abnormalities other than pulmonary tuberculosis, which is an important issue in the area of pulmonary diseases (26).

Clinically, this research is important in two ways. First, to our knowledge, this is the first study to use a large-scale, long-term dynamic cohort accumulation dataset to conduct an additional validation, where this external dataset went through more than 6 years of analysis and included 358,169 images. Unlike the internal testing, the external dataset in this study possessed a small number of TB cases with a TB prevalence of 1.0% because these cases were experimentally designed to be representative rather than balanced. Therefore, they reflected a real systematic screening population. We found that in such situations, the system showed good performance, achieving an accuracy of 0.931 to identify TB from normal images, and the sensitivity and specificity were 0.844 and 0.934, respectively. The latest WHO consolidated guidelines on tuberculosis (27) indicated the sensitivity and specificity ranges of computer-

aided detection software, referring to a minimum sensitivity of 0.90 when the specificity reaches 0.7. In the current study, the sensitivity was 0.94 for a specificity of 0.70 when classifying TB from normal cases on external independent validation. This is larger than the values reported by WHO, indicating the promising potential of our AI model in clinical application. Second, few previous studies conducted such a strict external test to evaluate the performance of AI models in detecting TB images. In our study, we had no access to the 358,169 images from the external dataset; since the developed AI system was connected to the hospital intranet, and doctors sent back results directly after the test was completed. This differed from the 138 images from the USA and 662 images from China, which were used to validate all internally downloaded and stored images (24). It should be noted that external validation is strongly recommended for all prediction models. Kang Zhang *et al.* (2020) evaluated AI performance on external datasets to detect COVID-19 after the internal test (28), and the lack of appropriate external validation for AI algorithms is a growing concern for its clinical application (18). However, a strict external validation such as ours would be preferable. Therefore, we believe that our results can provide supporting evidence for the usefulness of AI systems in real-world applications.

Differential diagnosis of TB is important, as it is essential to guide clinical patient management. However, sometimes it may be difficult to distinguish TB from other pulmonary diseases, such as pneumonia, by limited radiologic features (29). In some images, TB manifests as dense and homogeneous parenchymal consolidation in the lower and middle lobes of the lung, which is often indistinguishable from the appearance of bacterial pneumonia. However, it can be differentiated on the basis of radiographic evidence of lymphadenopathy (30). Therefore, our system may be a convenient approach to differentiate TB and other common pulmonary diseases as a result of its ability to identify a large number of radiologic features, which is an improvement over other reported studies focusing on classifying TB and normal images only.

There are some limitations to this study. Although a large number of chest radiographs from multiple institutions and external, independent evaluations were used to evaluate the performance of the AI system, this was still a retrospective study where datasets were available at the time of the study. For the next step of our study in the future, we aim to conduct a prospective clinical evaluation with human participants by comparing the performance of manual methods with or without AI assistance in an external, independent validation to further explore the clinical benefit of the AI method.

Conclusions

The AI system developed in our study was validated on both internal and external sets and demonstrated excellent detection performances in real-world independent validation. In addition, it should be noted that this AI system can only be used for TB routine screening to flag possible cases for rapid radiologic review and guide further examinations by radiologists, at this stage of development.

Acknowledgments

Funding: This work was supported by the National Key Research and Development Program of China (No. 2019YFE0121400), the Shenzhen Science and Technology Program (No. KQTD2017033110081833; JSGG20201102162802008), and the Shenzhen Fundamental Research Program (No. JCYJ20190813153413160). This research was also supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

Footnote

Reporting Checklist: The authors have completed the STARD checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-21-676/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-21-676/coif). FYML has the stock of Shenzhen Smart Imaging Healthcare Co., Ltd., and Lin Guo is the Medical Director of Shenzhen Smart Imaging Healthcare Co., Ltd. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the Peking University Shenzhen Hospital Institutional Review Board with a waiver of informed consent.

References

1. MacNeil A, Glaziou P, Sismanidis C, Date A, Maloney S, Floyd K. Global Epidemiology of Tuberculosis and Progress Toward Meeting Global Targets - Worldwide, 2018. MMWR Morb Mortal Wkly Rep 2020;69:281-5.
2. Adane AA, Alene KA, Koye DN, Zeleke BM. Non-adherence to anti-tuberculosis treatment and determinant factors among patients with tuberculosis in northwest Ethiopia. PLoS One 2013;8:e78791.
3. Cudahy P, Shenoi SV. Diagnostics for pulmonary tuberculosis. Postgrad Med J 2016;92:187-93.
4. Waibel A, Hanazawa T, Hinton G, Shikano K, Lang KJ. Phoneme recognition using time-delay neural networks. IEEE Trans Acoust, Speech, Signal Process 1989;37:328-39.
5. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE 1998;86:2278-324.
6. Li Q, Cai W, Wang X, Zhou Y, Feng DD, Chen M, editors. Medical image classification with convolutional neural network. 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV); 2014 10-12 Dec. 2014.
7. Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Jianming Liang. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? IEEE Trans Med Imaging 2016;35:1299-312.
8. Olveres J, González G, Torres F, Moreno-Tagle JC, Carbajal-Degante E, Valencia-Rodríguez A, Méndez-Sánchez N, Escalante-Ramírez B. What is new in computer vision and artificial intelligence in medical image analysis applications. Quant Imaging Med Surg 2021;11:3830-53.
9. Deepa SN, Devi BA. A Survey on Artificial Intelligence Approaches for Medical Image Classification. Indian J Sci Technol 2011;4:1583-95.
10. Lakhani P, Sundaram B. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology 2017;284:574-82.
11. McBee MP, Awan OA, Colucci AT, Ghobadi CW, Kadom N, Kansagra AP, Tridandapani S, Auffermann WF. Deep Learning in Radiology. Acad Radiol 2018;25:1472-80.
12. Dimopoulos K, Giannakoulas G, Bendayan I, Liodakis E, Petraco R, Diller GP, Piepoli MF, Swan L, Mullen M, Best N, Poole-Wilson PA, Francis DP, Rubens MB, Gatzoulis MA. Cardiothoracic ratio from postero-anterior chest radiographs: a simple, reproducible and independent marker of disease severity and outcome in adults with congenital heart disease. Int J Cardiol 2013;166:453-7.
13. Rai DK, Kirti R, Kumar S, Karmakar S, Thakur S. Radiological difference between new sputum-positive and sputum-negative pulmonary tuberculosis. J Family Med Prim Care 2019;8:2810-3.
14. Li K, Liu K, Zhong Y, Liang M, Qin P, Li H, Zhang R, Li S, Liu X. Assessing the predictive accuracy of lung cancer, metastases, and benign lesions using an artificial intelligence-driven computer aided diagnosis system. Quant Imaging Med Surg 2021;11:3629-42.
15. Cheng JZ, Ni D, Chou YH, Qin J, Tiu CM, Chang YC, Huang CS, Shen D, Chen CM. Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans. Sci Rep 2016;6:24454.
16. Jaeger S, Karargyris A, Candemir S, Siegelman J, Folio L, Antani S, Thoma G. Automatic screening for tuberculosis in chest radiographs: a survey. Quant Imaging Med Surg 2013;3:89-99.
17. Fatima S, Shah SIA. A review of Automated Screening for Tuberculosis of Chest Xray and Microscopy Images. Int J Sci Eng Res 2017;8:405-18.
18. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. Korean J Radiol 2019;20:405-10.
19. Jaeger S, Candemir S, Antani S, Wáng YX, Lu PX, Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant Imaging Med Surg 2014;4:475-7.
20. Mikołajczyk A, Grochowski M, editors. Data augmentation for improving deep learning in image classification problem. 2018 International Interdisciplinary PhD Workshop (IIPhDW); 2018 9-12 May. 2018.
21. Srivastava N, Hinton G, Krizhevsky A, Sutskever I,

Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 2014;15:1929-58.
22. Dai S, Lu K, Dong J, Zhang Y, Chen Y. A novel approach of lung segmentation on chest CT images using graph cuts. Neurocomputing 2015;168:799-807.
23. Günther S, Ruthotto L, Schroder J, Cyr E, Gauger N. Layer-Parallel Training of Deep Residual Neural Networks. SIAM J Math Data Sci 2020;2:1-23.
24. Hwang S, Kim HE, Jihoon JMD, Kim HJ, editors. A novel approach for tuberculosis screening based on deep convolutional neural networks. Medical Imaging: Computer-aided Diagnosis; 2016. 2016.
25. Qin C, Yao D, Shi Y, Song Z. Computer-aided detection in chest radiography based on artificial intelligence: a survey. Biomed Eng Online 2018;17:113.
26. Hwang EJ, Park S, Jin KN, Kim JI, Choi SY, Lee JH, Goo JM, Aum J, Yim JJ, Park CM; Deep Learning-Based Automatic Detection Algorithm Development and Evaluation Group. Development and Validation of a Deep Learning-based Automatic Detection Algorithm for Active Pulmonary Tuberculosis on Chest Radiographs. Clin Infect Dis 2019;69:739-47.
27. Organization WH. WHO consolidated guidelines on tuberculosis: module 2: screening: systematic screening for tuberculosis disease. Geneva: World Health Organization; 2021. Available online: https://www.who.int/publications/i/item/9789240022676.
28. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography. Cell 2020;181:1423-1433.e11.
29. Bhalla AS, Goyal A, Guleria R, Gupta AK. Chest tuberculosis: Radiological review and imaging recommendations. Indian J Radiol Imaging 2015;25:213-25.
30. Burrill J, Williams CJ, Bain G, Conder G, Hine AL, Misra RR. Tuberculosis: a radiologic review. Radiographics 2007;27:1255-73.

Cite this article as: Zhou W, Cheng G, Zhang Z, Zhu L, Jaeger S, Lure FYM, Guo L. Deep learning-based pulmonary tuberculosis automated detection on chest radiography: large-scale independent testing. Quant Imaging Med Surg 2022;12(4):2344-2355. doi: 10.21037/qims-21-676

