Evaluation of an artificial intelligence (AI) system to detect tuberculosis on chest X-ray at a pilot active screening project in Guangdong, China in 2019

Qinghua Liao[a,1], Huiying Feng[a,1], Yuan Li[b], Xiaoyu Lai[a], Junping Pan[a], Fangjing Zhou[a], Lin Zhou[a] and Liang Chen[a,*]

[a]*Center for Tuberculosis Control of Guangdong Province, Guangzhou, Guangdong, China*
[b]*Scientific and Technological Project Manager, Beijing Capital Technological Project Management Co., Ltd, Beijing, China*

Abstract.
BACKGROUND: Although computer-aided detection (CAD) software employed with Artificial Intelligence (AI) system has been developed aiming to assist tuberculosis (TB) triage, screening, and diagnosis, its clinical performance for tuberculosis screening remains unknown.
OBJECTIVE: To evaluate performance of an CAD software for detecting TB on chest X-ray images at a pilot active TB screening project.
METHODS: A CAD software scheme employed with AI was used to screen chest X-ray images of participants and produce probability scores of cases being positive for TB. CAD-generated TB detection scores were compared with on-site and senior radiologists via several performance evaluation indices including area under the ROC curves (AUC), specificity, sensitive, and positive predict value. Pycharm CE and SPSS statistics software packages were used for data analysis.
RESULTS: Among 2,543 participants, eight TB patients were identified from this screening pilot program. The AI-based CAD system outperformed the onsite (AUC = 0.740) and senior radiologists (AUC = 0.805) either using thresholds of 30% (AUC = 0.978) and 50% (AUC = 0.859) when taking the final diagnosis as the ground truth.
CONCLUSIONS: The AI-based CAD software successfully detects all TB patients as identified from this study at a threshold of 30%. It demonstrates feasibility and easy accessibility to carry out large scale TB screening using this CAD software equipped in medical vans with chest X-ray imaging machine.

Keywords: Tuberculosis (TB), active TB case screening, artificial intelligence, computer-aided detection, chest X-ray

---

[1]These authors contributed equally to this work.

*Corresponding author: Liang Chen, MD., Center for Tuberculosis Control of Guangdong Province, No. 485 Huangpu Avenue, Guangzhou, Guangdong 510620, China. Tel.: +00186 18928929722; Fax: +00186 38905745; E-mail: davidcl@ 126.com.

1. Introduction

Tuberculosis (TB) is one of the top 10 causes of death and the leading cause from a single infectious disease worldwide [1]. In 2015, World Health Organization (WHO) proposed the End TB Strategy, aiming at a world free of TB [2]. The strategy stressed integrated, patient-centered care and prevention. Early diagnosis of TB, treatment of TB patients and preventive treatment of persons at high risk are key components of patient-centered care and prevention.

Detecting and curing TB patients are currently recognized as the most effective and cost-effective TB control and prevention measures. Around the globe, passive detection strategies have been widely used and recognized as a cost-effective way to identify TB patient, whereas can lead to the failure of some TB patients to receive timely diagnosis and treatment, resulting in further transmission in the community. Therefore, it has been recommended by WHO that countries should carry out active screening and detection of TB in high-risk groups at national and provincial levels, based on their epidemic situations and economic levels [3].

Although bacteriological examinations, including molecular test, sputum smear, and sputum culture, are the main approaches to testing active TB, they are not practical in the large-scale screening of TB in high-risk population. Chest radiography is the most preferred imaging technique for early diagnosis of TB, and in developed countries, chest X-rays (CXRs) have been used for the evaluation of persons with symptoms of active pulmonary TB (PTB), and for TB screening among high-risk groups for several decades [4]. However, CXR has not been extensively used for screening active TB patients for the lack of radiologists in low-resource, high burden areas [5]. Besides, screening large-scale CXRs can lead to the fatigue of radiologists.

The development of artificial intelligence especially the recent break-through of deep learning technologies, has shed light in using computer-aided detection (CAD) software for automatic classification of TB from normal and other abnormal CXRs to assist TB triage, screening, and diagnosis [6–8]. The Stop TB Partnership has evaluated multiple CAD products from different countries with retrospective datasets from TB triage settings, and all these products performed better than local human readers [9, 10]. In a recent systematic review, Harris et al. concluded that CAD products are promising, but also pointed out that majority of work in the literature had been on development rather than clinical evaluation. All the 13 studies of clinical evaluation included in Harris's review are about CAD4TB (Delft Imaging, Netherlands). In addition to CAD4TB, nine more CAD products are under WHO evaluation [11], but evidence about their clinical performance are limited. A recent prospective study evaluated qXR (version 2.0) and CAD4TB (version 6.0) among symptomatic adults in a triage setting with a reference of mycobacterial culture and reported that both software achieved non-inferior accuracy (qXR: sensitivity = 0.93, specificity = 0.75; CAD4TB: sensitivity = 0.93; specificity = 0.69) to WHO-recommended minimum values [12], further approve the value of CADs.

The clinical performance of CAD in TB screening settings, however, is still largely unknown. In this study, one of the products under WHO evaluation, JF CXR-1, was employed in an active screening pilot project in Guangdong Province (TB incidence 65/100 thousand compared to 56/100 thousands nationwide in China, 2019) [13]. We evaluated its performance compared with human readers and clinical diagnosis.

2. Materials and methods

2.1. Pilot project location and study population

We implemented this active TB screening project in Yingde, a county-level city with a total population of 1.18 million in Northern Guangdong Province. This pilot study of TB screening aimed to recruit local
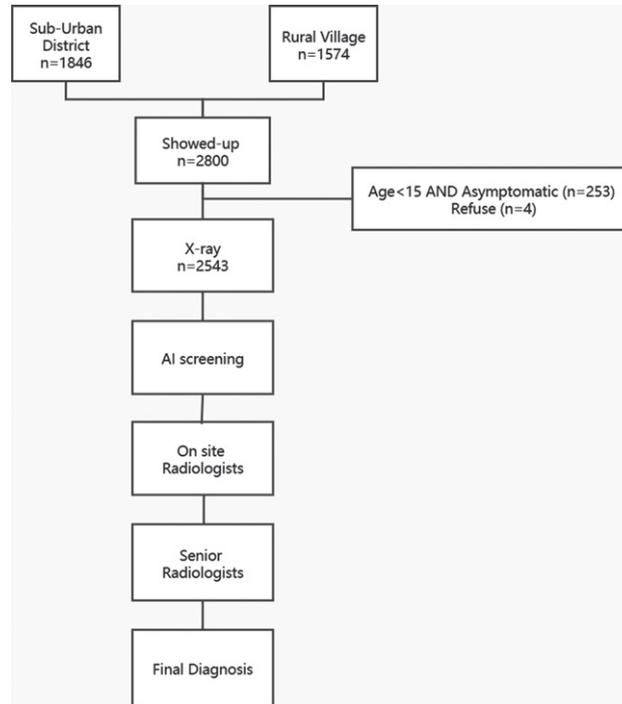
Fig. 1. A flow-chart of inclusion and exclusion criteria.

residents who lived in Yingde for more than 6 months. A sub-urban district (target population 1846) and a rural village (target population 1574) of Yingde were selected to recruit the participants. The local Health Departments organized the recruitment and mobilization door-by-door to notify residents about the screening program with propaganda and TB-related health education materials.

2.2. Screening procedures

All participants (including known TB patients) who showed up at the screening sites were verbally screened for TB symptoms; TB infection was tested by interferon-gamma release assay (IGRA, QIAGEN, QuantiFERON®-TB Gold), and those who aged 15 years and above received a posterior-anterior CXR (PA-CXR). Subjects aged 5–14 years were examined by CXR only if any of the following conditions occurred, (a) with TB symptoms, (b) close contacts of TB patients (family members, classmates and neighbors), or (c) IGRA positive.

CXR was taken in two mobile vans equipped with a digital X-ray machine (HEDY, DXR-530) which was connected with an AI-based CAD system (JF CXR-1, v2). The AI algorithm was deployed on Ali-Cloud, and DICOM images and AI results were transferred via 4G connection. On-site certified radiologists read the CXR image simultaneously and took into consideration of participant's demographic and other clinical information to issue a report.

Sputum tests were performed (sputum smear, sputum culture and a TB molecular test) for subjects meeting any of the following conditions, (a) with TB symptoms, or (b) with abnormal CXR (whether suggestive to TB or not), (c) with known active TB disease, or (d) with positive IGRA results.

After the on-site data collection, all images were re-read by senior radiologists, and participants with CXR suggestive to TB but missed by the on-site radiologists were followed up to have sputum tests. (Fig. 1).

2.3. AI-based CAD system

The AI-based CAD system (JF CXR-1, v2) used in this pilot study was developed by a domestic Hi-tech enterprise (JF Healthcare, Nanchang, China). It used cutting-edge deep learning technology to detect multiple major thorax diseases simultaneously on PA-CXR image, including TB, pneumonia, lung nodule/mass, and pleural diseases etc. This system was trained on approximately 150,000 CXR from township level hospitals in multiple provinces across China. Each TB image was annotated by one of three board-certified radiologists to localize TB lesions through bounding boxes or masks. The full training dataset was randomly split into 90% for training and 10% for validation. A ResNet-34 backbone Convolutional Neural Networks (CNN) with a customized mask layer was trained to recognize TB, which achieved area under the curve (AUC)0.94 with 0.91 sensitivity and 0.81 specificity at the cutoff value of 0.20 for the TB probability score. Heatmaps were generated to locate the lesion on the image [13, 14]. The probability scores of TB generated by this AI-based CAD system were recorded in this screening study for data analysis.

2.4. Diagnosis and treatment

TB diagnosis in this study followed China's National TB Diagnosis Guideline (WS288-2017) [15]. Newly detected TB patients were registered in the National TB Program (NTP) and given standard treatment in local TB-designated healthcare facilities.

2.5. Data analysis

The performance of the AI-based CAD system, including AUC, as well as the sensitivity and specificity at different thresholds of the probability score, was calculated comparing with the results of on-site radiologists (group 1), senior radiologists (group 2) when taking final diagnosis (group 3) as the ground truth, respectively. All abnormal CXR reports were reviewed and those with "TB" or "suspect TB" in the diagnosis suggestion session were considered positive for TB. "Normal" and "other abnormal" CXR were considered negative for TB. Pycharm CE (JetBrains, PyCharm 2020.1.2) and SPSS 22.0 (IBM) software packages were used for data analysis.

2.6. Ethics

The study was approved by the Institutional Review Board of Center for Tuberculosis of Guangdong Province. Each adult participant who received the TB screening provided a written informed consent. Parental consent was obtained for participants less than 18 years of age.
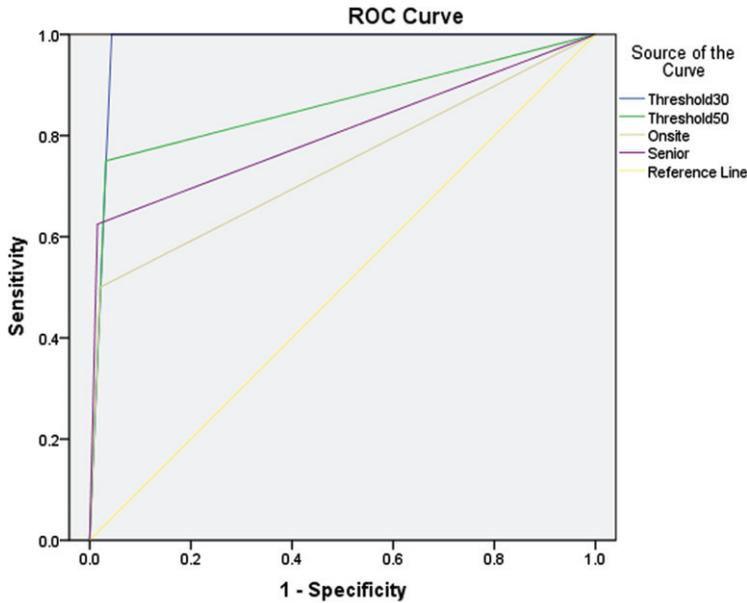
2.7. Role of the AI developers

The AI developers had no role in study design, data collection, and analysis. The AI developers participated in the AI-based CAD system manuscript writing, revision, and approved the final version.

3. Results

3.1. Participant characteristics and TB case finding

Ten days of the on-site screening from December 16–26, 2019 tested 2543 residents who showed up at the TB screening site out of the total 3420 local residents from the two selected area (one sub-urban

Taking the final diagnosis (including bacteriologically confirmed and clinical diagnosed TB) as the ground truth.

Areas Under ROC Curves

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Threshold 30	0.978	0.005	0.000	0.968	0.988
Threshold 50	0.859	0.091	0.000	0.680	1.000
Onsite	0.740	0.113	0.019	0.518	0.962
Senior	0.805	0.105	0.003	0.598	1.000

The test result variable(s): Threshold30, Threshold50, Onsite, Senior has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

Fig. 2. Performance of the AI-based CAD algorithm.

district and one rural village). Female to male ratio was 1.37 : 1, and more than half of the participants (63.43%) were 25–60 years old adults (Table 1).

About 24.18% participants had positive IGRA results (615 positive, which is consistent with recent findings of LTBI prevalence in China [16]. In addition, 10.66% (271/2543) of participants showed TB symptoms, 52 of which also had abnormal CXRs. On-site radiologists detected 56 CXRs "suggestive to TB (including old TB)", 211 "other abnormal", 2276 "normal". The senior radiologists confirmed 2110 participants' CXRs as "normal," while 389 "other abnormal," and 44 "suggestive to TB (including old TB)" (Table 1).

Eight TB patients were identified from this screening pilot (Table 1), four of which as bacteriologically confirmed (three by molecular test and one by culture), and the other four had clinical diagnosis

ROC Curve

Diagonal segments are produced by ties.
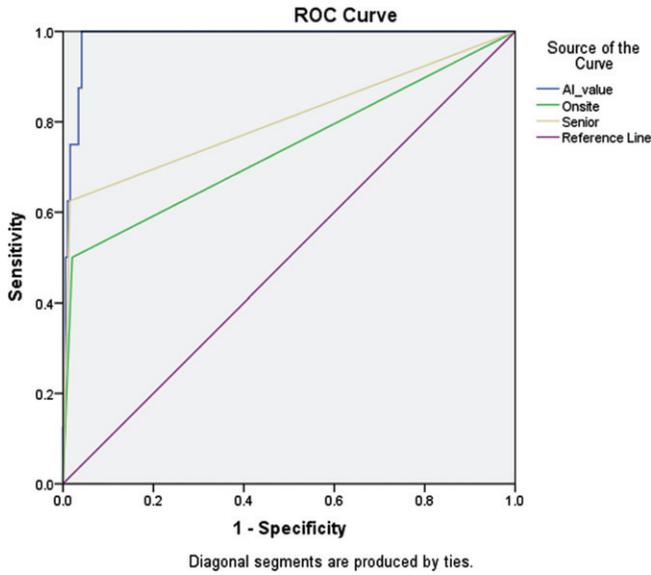
The AI-CAD score was generated as a continuous variable in our test.

Areas Under ROC Curves

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
AI_value	0.985	0.005	0.000	0.975	0.996
Onsite	0.740	0.113	0.019	0.518	0.962
Senior	0.805	0.105	0.003	0.598	1.000

The test result variable(s): Onsite, Senior has at least one tie between the positive actual
state group and the negative actual state group. Statistics may be biased.
a. Under the nonparametric assumption
b. Null hypothesis: true area = 0.5

Fig. 3. Comparison of TB detection performance among the onsite radiologists, senior radiologists and AI-based CAD algorithm.
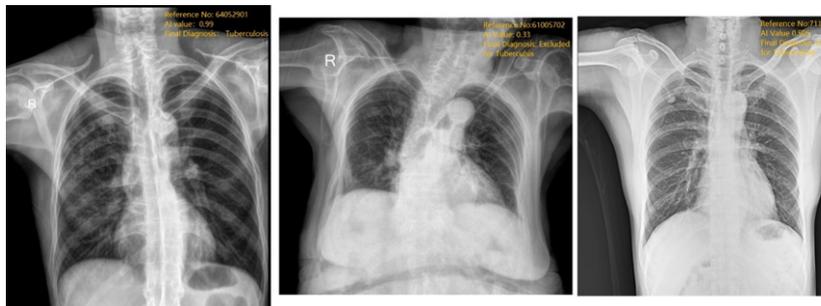

Fig. 4. Images of normal tuberculosis under threshold of AI-CAD Score = 0.3 and threshold of AI-CAD Score = 0.5.

Table 1
Demographic characteristics of participants in the active TB screening pilot project, Yingde, Guangdong, December 16–26, 2019

	Overall	On site Human Reader Categories				Senior Human Reader Categories			Final Diagnosis	
	2543	Normal	Abnormal but not suggestive of TB	TB	Not sure	Normal	TB	Abnormal but not suggestive of TB	Not TB	TB
< 15 years	35	5	1	0	29	17	0	18	35	0
Young age (15–25years)	43	14	1	0	28	36	0	7	43	0
Middle age (25–60 years)	1613	540	86	23	964	1436	19	158	1608	5
Elder age (>60 years)	852	191	123	33	505	621	25	206	849	3
Male	1073	314	90	36	633	851	28	194	1065	8
Female	1470	436	121	20	893	1259	16	195	1470	0

according to China's National TB Diagnosis Guideline (WS288-2017). Seven were new, and one more was on TB treatment for six months with abnormal CXR suggestive to TB as read by senior radiologists, positive molecular test result but negative IGRA result at the screening, and AI TB probability score of 46%. It should be noted that the eighth patient was diagnosed four months later via follow-up management due to suspicious CXR (AI score = 33.1%) and strong positive IGRA result at the screening.

3.2. Performance of AI-based CAD system

In this 2543 participants dataset, the AI-based CAD system achieved AUCs of both above 0.85 when compared with the on-site radiologists (0.740) and senior radiologists (0.805). The highest AUC (0.985) was achieved when comparing AI results to the final diagnosis (including bacteriologically confirmed and clinical diagnosed TB patients as defined in China's National TB Diagnosis Guideline WS288-2017). Sensitivity, specificity, and PPV at the thresholds of 30% and 50% are shown in Table 4. On sensitivity, AI-based CAD system shown better performance than human reader. At threshold of 50% (which was recommended before the start of the pilot), the AI-based CAD system missed two TB patients, one of which was very early on the disease stage at screening and was followed-up to be diagnosed three months later, and the other was on TB treatment for 6 months with positive molecular test but negative IGRA at the screening. All eight TB patients could have been detected by the AI-based CAD system at threshold 30%, but 97 participants would have to receive the subsequent sputum tests compared to 71 at threshold of 50%.

4. Discussion

This active screening pilot project in Yingde identified seven new TB patients; one on-going patient under TB treatment. The two screening sites had a total population of 3420, and therefore, the rough TB prevalence was 234/100,000, with rough incidence 205/100,000,which is not only higher than the estimated incidence rate of 58/100,000 in China, but also higher than the notification rate 65/100,000 in Guangdong Province in 2019 [13]. Unlike other retrospective studies [9, 17, 18], our study is one of the few studies perform prospectively screening on healthy population.

Table 2
AI-CAD results compared with different groups

	Overall	AI Result Threshold Score: 30		AI Result Threshold Score: 50	
		Negative	Positive	Negative	Positive
	2543	2425	118	2456	87
On-site radiologists	Negative	2408	79	2435	52
	Positive	17	39	21	35
Senior radiologists	Negative	2414	85	2443	56
	Positive	11	33	13	31
Final diagnosis	Negative	2425	110	2454	81
	Positive	0	8	2	6

The AI-based CAD system employed in this active screening pilot achieved better AUCs than on-site and senior radiologists when taking the results of final clinical and bacteriological result as the ground truth. This result is consistence with results reported by Cao XF, et al., [17]. In our study, with sensitivity of 90%, the AI-based CAD system shows specificities larger than 70%, which meets the WHO-recommended minimum values (90% for sensitivity, 70% for specificity). Nevertheless, our study result indicates a threshold score of 30% (with sensitivity 100% and specificity 96%) is suggested in future screening project. Furthermore, our study, conducted in 2019, is an innovative study to test the performance of a commercially available AI technique on real world on-site screening, rather than those study focusing on potential patients showed up on hospital triage [12]. Therefore, even though our study evaluates one available AI-based CAD system, it showed a promising performance of the AI-based CAD system, which indicates more and more concerned should be put on future AI screening. Further analysis of AI's false positive CXR at threshold of 30%, more were abnormal but not suggestive to TB as determined by on-site radiologists (77.7%) and senior radiologists (93.5%) (Table 2). Less than 74.7% were normal CXR as determined by human readers. Most likely, AI tends to mistakenly group other abnormal CXR as TB. Differential diagnosis is challenging in CXR interpretation for AI and for human readers. To flag other abnormal CXR as TB to raise the attention of clinicians and patients may waste the resource of TB testing, but could benefit patient care.

There are a few limitations in this study. Not every participant received sputum test due to resource restriction and difficulties to collect sputum from apparent healthy subjects. As recommended in Harris's review, microbiologic reference standard of culture or nucleic acid amplification tests (NAAT) are preferred to assess CAD accuracy [18]. This study mainly assessed CAD performance against human readers (on-site and senior). In addition, CAD score were not considered as one of the criteria for subsequence sputum test, and therefore, the false positivity of CAD may be over-estimated. It has been reported that 19.5% of bacteriological positive patients who had CXR not suggestive to TB as by human readers [19]. The CAD threshold of 30% as recommended based on findings from this study may not be the ideal threshold in other settings; additional evaluations from sites with different demographic and disease prevalence baseline other than Yingde are needed to better determine the threshold of this CAD for TB screening purpose. Last but not least, we failed to include a cost-effective analysis, which plays an important role in decision makings of public health program.

In conclusion, this is one of few studies so far to evaluate CAD in an active TB screening program. This commercially available CAD successfully detected all TB patients as identified from this study at a threshold of 30%. It is quite feasible and accessible to carry out large scale TB screening using

Table 3
AI results and demographic distribution

Overall	AI Result Threshold Score: 30		AI Result Threshold Score: 50	
	≥30	< 30	≥50	< 50
2543	118	2425	87	2456
Age group				
< 15 years	1	34	0	34
Young age (15–25 years)	0	43	0	43
Middle age (25–60 years)	44	1569	32	1581
Old age (>60 years)	73	779	55	797
Gender				
Male	80	993	64	1009
Female	38	1432	23	1447

Table 4
Comparison of Sensitivity, Specificity and Positive predict value (PPV) between
AI results and radiologists

	Sensitivity	Specificity	PPV
On-site Radiologists	50%	97.90%	7.10%
Senior Radiologists	62.50%	98.50%	11.40%
AI (Threshold 30)	100%	95.70%	6.80%
AI (Threshold 50)	75%	96.80%	6.90%

The gold standard was defined as bacteriologically confirmed and clinical diagnosed TB.

CADs equipped in medical vans with DR. In future scaled-up studies, CAD could be considered to assist the on-site radiologists to further increase the efficiency of TB screening.

Author contributions

LC was in charge and designed this project. QH-L, FJ-Z, HY-F and YL performed the onsite project, data-analysis and quality control of AI programing. HY-F and YL prepared the first draft of the manuscript. All authors made significant contribution to this study, and took part in revising and/or critically reviewing this article as well. All authors gave final approval of the version to be published and agreed on the journal to which the article has been submitted.

Funding

Acknowledgment

References

[1]   World Health Organization. *Global tuberculosis report 2018.* Geneva: 2018. Licence: CC BY-NC-SA *WHO/HTM/ TB/2018.2*. 2018, Switzerland.

[2]   World Health Organization, *The end TB strategy: global strategy and targets for tuberculosis prevention, care and control after 2015*. World Health Organization 2015. Available online: https://www.who.int/tb/post2015_TBstrategy.pdf

[3]   World Health Organization, *Systematic screening for active tuberculosis: an operational guide*.2015, Geneva, Switzerland WHO Document Production Services Available online: https://apps.who.int/iris/bitstream/handle/10665/181164/ 9789241549172_eng.pdf?sequence=1

[4]   Williams and H. Francis, The use of x-ray examinations in pulmonary tuberculosis, *The Boston Medical and Surgical Journal* **157**(26) (1907), 850–853.

[5]   T. Pande, et al., Use of chest radiography in the 22 highest tuberculosis burden countries, *European Respiratory Journal* **46**(6) (2015), 1816–1819.

[6]   P. Lakhani and B. Sundaram, Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks, *Radiology* **284**(2) (2017), 574–582.

[7]   L. Ma, Y. Wang, L. Guo, et al., Developing and verifying automatic detection of active pulmonary tuberculosis from multi-slice spiral CT images based on deep learning, *J Xray Sci Technol* **28**(5) (2020), 939–951.

[8]   M. Nijiati, Z. Zhang, A. Abulizi, et al., Deep learning assistance for tuberculosis diagnosis with chest radiography in low-resource settings, *J Xray Sci Technol* **29**(5) (2021), 785–796.

[9]   Z.Z. Qin, S. Ahmed, M.S. Sarker, et al., Can artificial intelligence (AI) be used to accurately detect tuberculosis (TB) from chest x-ray? A multiplatform evaluation of five AI products used for TB screening in a high TB-burden setting. arXiv.2006.05509, 2020.

[10]   Z.Z. Qin, et al., Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems, *Sci Rep* **9**(1) (2019), 15000.

[11]   World Health Organization, *Global tuberculosis report 2020.* Geneva, Licence CC BY-NC-SA 3.0 IGO 2020.

[12]   F.A. Khan, et al., Chest x-ray analysis with deep learning-based software as a triage test for pulmonary tuberculosis: a prospective study of diagnostic accuracy for culture-confirmed disease, *The Lancet Digital Health* **2**(11) (2020), e573–e581.

[13]   Q. Wang, T. Li, X. Du, et al., The analysis of national tuberculosis reported incidence and mortality, 2015–2019, *Chin J Antituberc* **43**(2) (2021), 107–112.

[14]   Y. Li, et al., Deep learning algorithm classifies active TB, normal, and other abnormal chest x-rays with high accuracy on large scale dataset. 51st World Conference on Lung Health of the International Union Against Tuberculosis and Lung Disease, *IJTLD* **24**(10) (2020), s406.

[15]   World Health Organization, W.H., *Treatment of tuberculosis: guidelines.* 2010.

[16]   X.W. Gao and Y. Qian, Prediction of multidrug-resistant TB from CT pulmonary images based on deep learning techniques, *Mol Pharm* **15**(10) (2018), 4326–4335.

[17]   X.F. Cao, et al., Application of artificial intelligence in digital chest radiography reading for pulmonary tuberculosis screening, *Chronic Dis Transl Med* **7**(1) (2021), 35–40.

[18]   M. Harris, et al., A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest X-rays for pulmonary tuberculosis, *PLoS One* **14**(9) (2019), e0221339.

[19]   K. Ito, [Limits of chest X-ray investigation in the diagnosis of recurrent pulmonary tuberculosis], *Kekkaku* **80**(7) (2005), 521–6.