Journal of Medical Systems (2022) 46:82
https://doi.org/10.1007/s10916-022-01870-8

IMAGE & SIGNAL PROCESSING

Advances in Deep Learning for Tuberculosis Screening using Chest X-rays: The Last 5 Years Review

KC Santosh[1] · **Siva Allu**[1] · **Sivaramakrishnan Rajaraman**[2] · **Sameer Antani**[2]

Received: 20 January 2022 / Accepted: 19 September 2022 / Published online: 15 October 2022

Abstract

There has been an explosive growth in research over the last decade exploring machine learning techniques for analyzing chest X-ray (CXR) images for screening cardiopulmonary abnormalities. In particular, we have observed a strong interest in screening for tuberculosis (TB). This interest has coincided with the spectacular advances in deep learning (DL) that is primarily based on convolutional neural networks (CNNs). These advances have resulted in significant research contributions in DL techniques for TB screening using CXR images. We review the research studies published over the last five years (2016-2021). We identify data collections, methodical contributions, and highlight promising methods and challenges. Further, we discuss and compare studies and identify those that offer extension beyond binary decisions for TB, such as region-of-interest localization. In total, we systematically review 54 peer-reviewed research articles and perform meta-analysis.

Keywords Tuberculosis · Chest x-rays · Deep learning · Medical imaging · Systematic review

Introduction

The WHO Global Tuberculosis (TB) Report 2020 [1] reports the worldwide disease burden due to pulmonary TB. In spite of falling numbers overall, in 2019, 10 million people fell ill with TB; 465,000 people fell ill with drug-resistant TB; and 1.4 million people died of TB. In other words, the preventable and curable disease still remains the top infectious killer in the world claiming close to 4000 lives per day. Global health efforts aim for early screening and treatment to stem this. Two tests are commonly used to determine TB infection: (i) the Mantoux TB skin test; and, (ii) the interferon-gamma release assays (IGRAs) blood tests. The skin test is sensitive but unable to unilaterally determine of a positive test is due to latent infection or active. The IGRA is highly accurate but expensive and challenging to implement in low and middle income regions (LMIR) of the world. Chest imaging is commonly used in the diagnosis of intrathoracic TB, where chest X-rays (CXR) is the most commonly used modality due to lower cost and relatively easy implementation, particularly for pediatrics [2]. The authors of [2] state that CXRs can be used to "detect abnormalities compatible with pulmonary, pleural, pericardial, and lymph node disease, which can all be caused by TB." However, it is limited by its two-dimensional representations of three-dimensional anatomy and suffers from modest specificity [3–6]. Alternatively, computed tomography (CT), magnetic resonance imaging (MRI), and nuclear medicine techniques, including positron emission tomography/computed tomography (PET/CT) provide greater accuracy in diagnosing both pulmonary and extrapulmonary Tuberculosis [7]. These techniques are expensive to implement in LMIR and all require expertise for interpreting them [8]. In spite of the limitations described earlier, CXR images remain the most widely used imaging modality for screening TB patients [9]. This is primarily

This article is part of the Topical Collection on *Image & Signal Processing*

✉ KC Santosh
santosh.kc@usd.edu

Siva Allu
sivasaivenkata.allu@coyotes.usd.edu

Sivaramakrishnan Rajaraman
sivaramakrishnan.rajaraman@nih.gov

Sameer Antani
sameer.antani@nih.gov

[1] Applied Artificial Intelligence (2AI) Research Lab
Computer Science Department, University of South Dakota, Vermillion, SD 57069, USA

[2] National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

due to their cost effectiveness, low ionizing radiation, and portability for use in remote regions [10]. As some forms of TB show distinctive patterns in CXR images (see Fig. 1), AI tools that analyze CXRs are now a prevalent medical imaging aids for TB screening [9, 11], particularly in LMIRs.

Within the last decade, there has been an explosive growth in automated CXR image analysis techniques toward addressing the expertise and access gap in global effort for early TB screening. The advent of convolutional neural network (CNN)-based deep learning (DL) provides the basis for imaging-based Artificial Intelligence (AI) solutions [12–14]. AI-guided solutions (e.g. DL methods) aim to supplement clinical decision-making [15, 16]. Such medical imaging tools have potential to mitigate the heavy burden on medical experts by triaging cases [14].

Traditional machine learning algorithms use hand-crafted features that are limited to specific problem/dataset as they require expert-based feature extraction [17–20]. In other words, such features are data dependent and therefore they cannot be generally applied. In contrast, deep features are automatically extracted and do not require specific manipulation of parameters [21–23]. DL-enabled methods for TB screening are cited as a promising solution for clinicians' challenges especially when we consider low-resource regions [1, 23]. In [24], authors discussed the use of both, handcrafted features as well as deep features for TB screening using CXR images. Regardless of the method, early and accurate TB detection is essential to achieve global control of the disease [1, 25].

Before we start our review, let us follow a workflow representing different phases of systematic review, where it primarily includes identification, screening, eligibility and included criteria as shown in Fig 2. For identification, we used the following keywords: 'Tuberculosis,' 'chest x-ray,' and 'deep learning,' Using all of them, we search research articles in the following repositories: a) PubMed and b) Web of Science In our screening, duplicate items were removed. Research articles that were published between 2016 - 2021

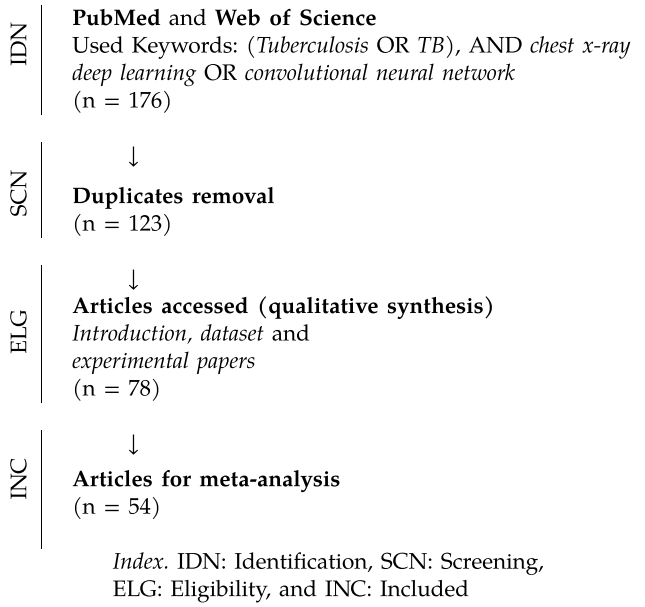

Fig. 2 Workflow representing different phases of the systematic review (source: PRISMA criteria [26])

are eligible for a review. To make our review up-to-date, we considered recently published research articles as well. Pre-print articles (e.g., ArXiv, medRxiv and TechRxiv) were strictly avoided as they are not peer-reviewed. For better meta-analysis, we screened for appropriate dataset (size and source), algorithmic factors (DL models) and corresponding performance scores. For meta-analysis, we included experimental-based research articles. Not to be confused, our aim is not to convey how research articles are compared in terms of performance scores, but to analyze how well DL models has been progressed so far since 2016. Furthermore, with the scope of DL models, we consider transfer learning, data augmentation and visualization (disease localization).

The remainder of the paper is organized as follows. Following our selection criteria and our study scope, we

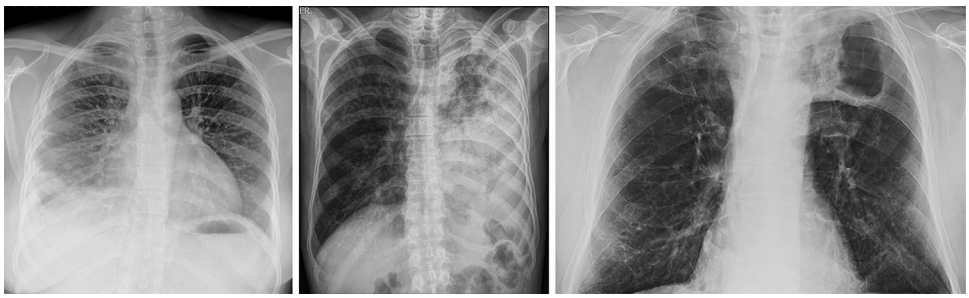

Fig. 1 TB cases in chest X-rays showing **a** scattering bilateral reticular and nodular shadows more evident at the right middle and lower zone (source: radiopaedia.org, rID: 39290); **b** bilateral micronodular insterstitial effusion, left apical bronchiectasis at the level of the pulmonary lingula (granulomatous infectius process), and pleural apical thickening (source: radiopaedia.org, rID: 28037); and **c** secondary TB with apical lung cavities (source: https://www.kumc.edu/)

start with data collections, availability, and their respective sources in "Datasets and Availability". "TB Screening using CXRs (2016-2021)" focuses on the last five-year review on TB screening imaging tools (2016 – 2021) using CXR images. In this section (*ref.* "TB Screening using CXRs (2016-2021)"), we review performance of the DL-based medical imaging tools in accordance with the datasets used. Following the discussion outline in "Discussion Outline", we summarize reviews by taking the following factors into account: a) performance comparison (*ref.* "Performance Comparison)" and b) analysis DL-based medical imaging tools (*ref.* "Based Medical Imaging Tools: Decision-Making to Data Visualization") that mainly includes transfer learning, data augmentation, and disease localization. "Conclusions" concludes this paper.

Datasets and Availability

DL requires large amount of data for training the network. In this section, we discuss various data sets identified in the literature and their availability for research use. They are ordered incrementally according to their data set size. We also summarize them in Table 1.

C1. *Thomas Jefferson University Hospital Dataset (JUHD, USA):* The data set consists of 119 CXRs which are available in DICOM format with 10 cases and 110 controls.

C2. *Montgomery County Dataset (MC, USA):* This data set was collected from the TB control program of the Department of Health and Human Services (HHS) of Montgomery County (MC), Maryland, USA. It contains 138 posterior-anterior X-rays and are available in both PNG and DICOM formats. In the data set, 80 CXRs are controls and 58 CXRs are cases with manifestations of TB. The set covers a wide range of TB-related abnormalities, including effusions and miliary patterns.

C3. *Infectious Disease Institute Dataset (Kampala, Uganda):* The dataset was collected as a part of observational study of patients enrolled at the integrated TB-HIV out-patient clinic of the Infectious Disease Institute in Kampala, Uganda. CXRs were taken in the Radiological Department of Mulago National Referral Hospital. It consists of 138 CXRs of HIV-infected patients diagnosed with TB.

C4. *National Institute of Tuberculosis and Respiratory Diseases Dataset (NITRD, India):* It contains two different sets: DA and DB of CXRs (153 cases, each of which 78 are abnormal and 75 are normal) that were collected from different X-ray machines.

C5. *Japanese Society of Radiological Technology Dataset (JSRT, Japan):* It consists of 247 CXRs with 93 are normal and 154 are abnormal cases, some of which present TB-like manifestations.

C6. *Kasturba Hospitals Dataset (Manipal, India):* The approximate annual number of PTB cases at Kasturba Hospital were estimated between 300–400 cases and 600–800 controls would present over the year. In total, 317 cases and 612 controls were included in the analysis (*ref.* Nash et al. [27]).

C7. *Gugulethu TB Clinic Dataset (Cape Town, South Africa):* A total of 392 patient records are available. Of all, 73 were diagnosed with TB.

C8. *Belarus TB Dataset (Belarus):* The National Institute of Allergy and Infectious Diseases, Ministry of Health, Republic of Belarus, collected the data for a drug resistance study. There are 422 CXR pictures in the dataset, representing 169 people, all of whom have TB. The Kodak Point-of-Care 260 system was used to take chest radiographs with a resolution of 2248 x 2248 pixels.

C9. *Shenzhen Hospital Dataset (SH, China):* The X-rays were acquired as part of the routine care at Shenzhen Hospital (Shenzhen No.3 Hospital in Shenzhen, Guangdong province, China. Of 662 frontal CXRs, 326 are normal and 336 are abnormal showing various TB manifestations.

C10. *GF Jooste and Khayelitsha Hospitals Dataset (South Africa):* It combined two datasets that were collected from two different hospitals (GF Jooste Hospitals collected by [28] and Khayelitsha hopsitals collected by [29]). It contains CXRs of 677 HIV-positive patients with suspected TB.

C11. *Sehatmand Zindagi Dataset (2016-2017) (Pakistan):* The dataset was collected from a study conducted between July 2016 to April 2017 in 30 private TB treatment and diagnostic centers called 'Sehatmand Zindagi' (Healthy Life) Centers and in community, mobile X-ray based TB screening camps, located in low middle-income neighborhoods. A total of 694 individuals with a diagnosis of DM (and 31.1% of them were newly diagnosed) were screened with CAD4TB and simultaneously provided sputum for Xpert MTB/RIF testing.

C12. *Easter Asian Hospital Dataset (EAH):* CXRs were collected in cooperation with radison for TB and non-TB diseases, and was followed pathological diagnosis. Of 864 CXRs, 492 are normal and 372 are abnormal cases.

C13. *Medical Surveillance Dataset (Yonsei University, South Korea):* The annual medical surveillance data for workers at Yonsei University, beginning from 2009 was used to create this dataset of 39,677 individuals of which 1202 individuals have TB.

Table 1 Dataset collections and their respective sources

	Data collections	Size	Source
C1.	Thomas Jefferson University Hospital Dataset (JUHD, USA)	119	Source not available
C2.	Montgomery County Dataset (MC, USA)	138	https://data.lhncbc.nlm.nih.gov/public/Tuberculosis-Chest-X-ray-Datasets/Montgomery-County-CXR-Set/MontgomerySet/index.html
C3.	Infectious Disease Institute Dataset (Kampala, Uganda)	138	Availabel upon request [31]
C4.	National Institute of Tuberculosis and Respiratory Diseases Dataset (NITRD, India)	153(DA) 153(DB)	https://sourceforge.net/projects/tbxpredict/files/data/
C5.	Japanese Society of Radiological Technology Dataset (JSRT, Japan)	247	http://db.jsrt.or.jp/eng.php
C6.	Kasturba Hospitals Dataset (Manipal, India)	317	Available upon request [27]
C7.	Gugulethu TB Clinic Dataset (Cape Town-South Africa)	392	Source not available
C8.	Belarus Tuberculosis Dataset (Belarus)	422	https://www.kaggle.com/raddar/drug-resistant-tuberculosis-xrays
C9.	Shenzhen Hospital Dataset (SH, China)	662	https://data.lhncbc.nlm.nih.gov/public/Tuberculosis-Chest-X-ray-Datasets/Shenzhen-Hospital-CXR-Set/Annotations/index.html
C10.	GF Jooste and Khayelitsha Hospitals Dataset (South Africa)	677	Available upon request [28, 29]
C11.	Sehatmand Zindagi Dataset (Pakistan) (2016-2017)	694	Available upon request [32]
C12.	Eastern Asia Hospital Dataset (EAH)	864	Available upon request [33]
C13.	Medical Surveillance Dataset (Yonsei University, South Korea)	39,675	Available upon request [34]
C14.	Nepal and Cameroon Dataset	1,196	Available upon request [35]
C15.	Indiana Dataset (Indiana, USA)	4,104	https://openi.nlm.nih.gov/faq#collection
C16.	Mendeley Dataset (UK)	5,232	https://data.mendeley.com/datasets/rscbjbr9sj/3
C17.	First Affiliated Hospitals of Xi'an JiaoTong University Dataset (FAHXJU, China)	5,344	Available upon request [36]
C18.	Pediatric Pneumonia CXR Dataset (USA)	5,856	https://data.mendeley.com/datasets/rscbjbr9sj/3
C19.	Sehatmand Zindagi Dataset (Pakistan) (2013-2015)	6,090	Available on request [37]
C20.	Korean Institute of Tuberculosis Dataset (KIT, South Korea)	10,848	Source not available
C21.	Tuberculosis CXR Dataset (TBX11K, China)	11,200	http://mmcheng.net/tb/
C22.	Open-i Dataset (NLM, USA)	11,425	https://openi.nlm.nih.gov/faq#collection
C23.	Sitec Medical Dataset (Philippines)	14,094	Source not available
C24.	Radiological Society of North America Dataset (RSNA, USA)	26,684	https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data
C25.	Find and Treat Dataset (UK)	47,510	Source not available
C26.	Chest X-rayS - NIH (MD, USA)	112,120	https://nihcc.app.box.com/v/ChestXray-NIHCC/folder/36938765345
C27.	AbiyeV Dataset (Turkey)	120,120	Available upon request [38]
C28.	CheXpert CXR Dataset (USA)	223,648	https://stanfordmlgroup.github.io/competitions/chexpert

C14. *Nepal and Cameroon Dataset (Nepal & Cameroon):* A total of 1,196 individuals (515 from Nepal and 681 from Cameroon) were included in this dataset with a median age of 46. Each participant received a posterior-anterior CXR using digital X-ray machines (Phillips Digital Diagnost in Nepal and Carestream Direct View Classic CR in Cameroon). In both sites, each person was classified as 'abnormal' if any pulmonary abnormality was detected by human readers, regardless of the abnormality being TB-specific or not.

C15. *Indiana Dataset (Indiana, USA):* It includes 4,104 postero-anterior CXRs (2,378 are abnormal and 1,726 are normal) that were collected from hospitals affiliated with the Indiana University School of Medicine, and archived at the National Library of Medicine (NLM). The images are available through the Open-i multimodal search engine (*ref.* C22).

C16. *Mendeley Dataset (UK):* It contains validated Optical Coherence Tomography (OCT) and CXR images. It

consists of 5,232 children's CXRs, 3883 of them are labeled as abnormal and the rest 2538 are normal.

C17. *First Affiliated Hospitals (Xi'an JiaoTong University, China):* It was collected from First Affiliated Hospitals of Xi'an JiaoTong University (FAHXJU) in Shanxi, China. Of 5,344 CXRs, 2382 are classified as normal and 2962, abnormal with TB manifestations.

C18. *Pediatric Pneumonia CXR (USA):* It includes 1,583 anterior-posterior CXRs showing normal lungs and 4,273 CXRs showing bacterial and viral pneumonia manifestations. They were collected from pediatrics (of 1 to 5 years of age) at the Guangzhou Women and Children's Medical Center, China. The images are acquired as a part of routine clinical care with IRB approvals.

C19. *Sehatmand Zindagi Dataset (2013-2015) (Pakistan):* Samples were collected from TB treatment and diagnostic centers, called 'Sehatmand Zindagi' (Healthy Life) centers, in Karachi, Pakistan, from October 2013 to September 2015. During the period, a total of 6,845 individuals with presumptive TB were enrolled. Of these, 755 individuals, with invalid, error, no result were excluded from the analysis, and final dataset consists of 6,090 CXRs.

C20. *Korean Institute of Tuberculosis Dataset (KIT, South Korea):* This dataset consists of 10,848 DICOM data of which 7,020 are normal and 3,828 abnormal (TB) cases. These were collected from the Korean Institute of Tuberculosis (KIT) under Korean National Tuberculosis Association (KNTA), South Korea.

C21. *Tuberculosis CXR Dataset (TBX11K, China):* TBX11K dataset [30] consists of 11,200 CXRs, including 5000 healthy cases, 5000 sick but non-TB cases, and 1200 cases with TB manifestations. 1,200 TB CXRs include 924 active TB cases, 212 latent TB cases, 54 cases with active and latent TB at the same time, and 10 uncertain cases whose TB types cannot be identified under current medical conditions. The 5000 sick but non-TB cases are gathered in order to cover as many types of radiograph diseases as possible in the clinical scenarios.

C22. *Open-i Dataset (NLM, USA):* The U.S. NLM's Open-i service allows for search and retrieval of abstracts and images (including charts, graphs, clinical images, and so on) from open source literature and biomedical image collections. Over 3.7 million images from approximately 1.2 million PubMed Central® articles are available through Open-i, as are 7,470 CXRs with 3,955 radiology reports, 67,517 images from the NLM History of Medicine collection, and 2,064 orthopedic illustrations.

C23. *Sitec Medical Dataset (Philippines):* In Palawan Island, Philippines, among high risk TB groups, the screening took place in a mobile clinic (Sitech Medical, Gyeonggi-do,Korea). All participants above 15 years underwent screening. Of 14,094 CXRs, 389 of them are abnormal (TB manifestations).

C24. *Radiological Society of North America Dataset (RSNA, USA):* IT was released as a part of the RSNA Kaggle pneumonia detection challenge, jointly organized by the radiologists from RSNA, Society of Thoracic Radiology (STR) and the NIH. It includes 26,684 images of which 17,833 are abnormal and 8,851 are normal.

C25. *Find & Treat Dataset (UK):* Find &Treat Screening Programme, organized in London, UK, screened a high-risk population of homeless people, prisoners and problem drug and alcohol users accessing homeless hostels, day centres, soup kitchens, drug treatment services and detention facilities. Large image database consisting of 47,510 postero-anterior CXRs from 39,328 individuals was created.

C26. *Chest X-rays - NIH (MD, USA):* It is currently the largest public repository of CXRs. It has 112,120 frontal-view (both postero-anterior and antero-posterior) CXRs of 30,805 unique patients. This dataset was collected as a part of routine care at NIH Clinical Centre, Bethesda, MD, USA. Of 112,120 CXRs, 60,360 are normal and 51,760, abnormal.

C27. Abiyev and Ma'aitah [38] *Dataset:* It contains 120,120 CXRs. No further information about the dataset was provided but the data is available upon request from the corresponding author upon request.

C28. *CheXpert CXR - Stanford Hospital Dataset (USA):* It includes 223,648 CXRs from 65,240 patients at Stanford Hospital, California, USA. They are labeled for the presence of 14 observations as positive, negative, or uncertain.

TB Screening using CXRs (2016-2021)

In this section, we summarize advances in Deep Learning (DL) for TB screening using CXRs. In our systematic review, we consider their methodologies and results in accordance with dataset size. Regarding dataset, we study and provide availability (for research purpose) as well as their respective sources (previous section). In our study, we also consider sate-of-the-art articles that used Covid-19 and Pneumonia cases in addition to TB.

In Table 2, starting from 2016, we summarize our systematic review on DL-based imaging tools (using CXR images), where we consider dataset size and performance that are commonly measured in accuracy, area under the curve, specificity and sensitivity. We also briefly describe their respective high-level methodologies, including their objectives/intentions.

Table 2 Chest X-ray imaging tools, dataset size and their performance measured in Accuracy (ACC), Area Under the Curve (AUC), Specificity (SPEC), and Sensitivity (SEN)

Authors (year)	Method	Data collection	Performance			
		(size)	ACC in %	AUC	SPEC	SEN
Hwang et al. (2016) [39]	CNN	C2 (138)	67.40	0.88	–	–
		C9 (662)	83.70	0.93	–	–
		C20 (10, 848)	90.30	0.96	–	–
Melendez et al. (2016) [40]	CAD4TBV3.07	C7 (392)	–	0.84	0.49	0.95
Lakhani and Sundaram (2017) [12]	Ensemble: AlexNet	C1, C2, C8 (#88),	–	0.99	1.00	0.97
	GoogLeNet	C9 (1, 007)				
Lopes and Valiati (2017) [41]	CNN: GoogLenet	C2 (138)	82.60	0.93	–	–
	VGGNet, ResNet	C9 (662)	84.70	0.90	–	–
Abiyev and Ma'aitah (2018) [38]	CNN	C27 (120, 120)	92.40	–	–	–
Rajpurkar et al. (2018) [42]	CheXNeXt	C26 (112, 120)	–	0.85	–	–
Melendez et al. (2018) [43]	CAD4TB	C25 (#38, 961)	–	0.90	0.56	0.95
Zaidi et al. (2018) [37]	CAD4TBV3.07	C19 (6, 090)	–	0.84	0.68	0.82
Bekar et al. (2018) [31]	ViDi	C3 (138)	–	0.82	0.98	0.91
Yadav et al. (2018) [46]	ResNet	C2,C9,				
		C26 (112, 920)	94.89	–	–	–
Qin et al. (2019) [35]	CAD4TB(V6)	C14 (1, 196)	92.00	0.92	0.96	0.47
	Lunit	C14 (1, 196)	94.00	0.94	0.97	0.58
	qXR	C14 (1, 196)	94.00	0.94	0.96	0.71
Ge et al. (2019) [47]	D121-BL	C26 (112, 120)	–	0.84	–	–
Hwa et al. (2019) [48]	Ensemble: VGG16 and	C2, C9 (800)	89.77	–	0.88	0.91
	Inception V3					
Pasa et al. (2019) [49]	CNN and SVM	C2, C8 (#304)				
		C9 (1, 104)	86.20	0.93	–	–
Evangelista and Guedes (2019) [50]	Ensemble: Inception	C2, C5 (#93),				
	ResNet and VGG	C9 (893)	93.82	–	0.94	0.93
Ahsan et al. (2019) [51]	VGG16	C2, C9 (800)	81.25	–	–	–
Heo et al. (2019) [34]	DCNN	C13 (39, 675)	–	0.97	0.96	0.81
Philipsen et al. (2019) [52]	CAD4TB	C23 (#13, 255)	–	0.93	0.87	0.82
Kim et al. (2020) [53]	DCNN	C26 (#111, 622)	–	0.87	0.76	0.85
Nash et al. (2020) [27]	qXR	C6 (317)	–	0.81	0.80	0.71
Rajpurkar et al. (2020) [54]	CheXiad	C10 (677)	65.00	–	0.61	0.73
Das et al.(2020) [55]	Truncated Inception Net	C2, C9 (800)	99.92	0.99	1.00	0.93
Sathitratanacheewin et al. (2020) [56]	DCNN	C9 (662)	–	0.98	0.82	0.72
		C26 (112, 120)	–	0.71	–	–
Yoo et al. (2020) [33]	CNN	C9 (#120),	80.00	0.80	0.89	0.72
		C12 (#372)(492)				
Sahlol et al. (2020) [57]	Mobile Net-AEO	C9 (662)	90.20	–	0.90	0.91
		C16 (5, 232)	94.10	–	0.97	0.87
Xie et al. (2020) [36]	Faster RCNN	C2 (138)	92.60	0.98	0.923	0.93
		C9 (662)	90.20	0.94	0.95	0.85
		C17 (5, 344)	97.40	0.99	0.96	0.98
Rajaraman and Antani (2020) [58]	Ensemble:	C9, C15, C18	94.10	0.99	0.96	0.93
	InceptionResNet-V2,	C24 (37, 306)				
	DenseNet-121					
	Inception-V3, VDSNet					
Rahman et al. (2020) [59]	CheXNet	C2, C8, C9	96.40	–	0.96	0.96
		C24 (27, 484)				

Table 2 (continued)

Authors (year)	Method	Data collection (size)	Performance			
			ACC in %	AUC	SPEC	SEN
Guo et al. (2020) [60]	Ensemble: VGG16, VGG19, Inception V3, ResNet34, ResNet50, and ResNet101	C9 (662)	94.50	–	0.95	0.99
		C26 (112, 120)	95.60	–	0.98	0.98
Abideen et al. (2020) [61]	B-CNN	C2 (138)	96.42	–	–	–
		C9 (662)	86.46	–	–	–
Murphy et al. (2020) [62]	CAD4TBV6	C19 (6, 090)	–	0.99	0.98	0.90
Habib et al. (2020) [32]	CAD4TB	C11 (694)	–	0.78	0.42	0.91
Rajaraman et al. (2020) [63]	Ensemble: VGG-16, VGG-19 Xception, NASNet-mobile Inception-V3, MobileNet DenseNet-121	C24, C28 (250, 332)	91.63	0.97	0.88	0.924
Ayaz et al. (2021) [64]	Ensemble: Inceptionv3, InceptionResnetv2 VGG16, VGG19, MobileNet ResNet50, Xception	C2 (138)	93.47	0.97	–	–
		C9(662)	97.59	0.99	–	–
Rajaraman et al. (2021) [65]	ResNet-BS	C2 (138)	92.30	0.96	0.97	0.88
		C9 (662)	88.79	0.95	0.89	0.88

In data collection CX(# YY), # YY refers to YY number of samples used in that study from any specific collection index X

2016 - 17

Hwang et al. [39] designed a Computed Aided Diagnosis (CAD) system based on deep CNN for automatic TB screening, where transfer learning mechanism was used. Computational experiments were conducted using three data collections: C2 (MC, USA), C9 (SH, China) and C20 (KIT, South Korea). On C2, C9 and C20 collections, authors achieved TB screening AUC of 0.88, 0.93 and 0.96, respectively. On the whole, the average accuracy and AUC with transfer learning are 0.903 and 0.964, respectively using train/validation/test (70/15/15) split evaluation protocol. Melendez et al. [40] introduced a novel machine learning framework to exploit additional information by combining automatic CXR scoring (via CAD) with clinical data, including symptoms and HIV status. It employs a data collection, C7 (Cape Town, South Africa) and that includes a CAD score and 12 clinical features. Using 10-fold cross validation, the reported accuracy when CAD scores and clinical information were combined was 0.84, whereas the AUCs for the alternative approaches using either type of information were 0.78 or 0.72, respectively. The combined strategy offers improved accuracy and specificity. In their study, it is observed that the most valuable clinical features reported in the integrated approach are HIV status, auxiliary temperature and lung auscultation findings.

In a similar fashion, Lakhani and Sundaram [12] assessed a deep CNN's (AlexNet and GoogLeNet) effectiveness for detecting TB in CXR images. The study was retrospective and the experiments were performed on four different publicly available data collections, namely C1 (JUHD, USA), C2 (MC, USA), C8 (Belarus) and C9 (SH, China). Using split-based validation approach (train/validation/test: 68/17.1/14.9), the reported accuracy of the best performing classifier is 0.99. The radiologists augmented reported a sensitivity of 0.973 and specificity of 1.0. In their study, authors reported that there are 13 misclassifications out of 150 test cases, and the study can be further enhanced to increase the accuracy in disagreement cases that radiologists correctly interpreted. Lopes and Valiati [41] investigated computational techniques for automating TB diagnosis using CXRs. In their work, pre-trained CNNs (GoogleNet, VGGNet and ResNet) were used to extract features on publicly available data collections, namely C2 (MC, USA), C9 (SH, China) and C4 (NITRD, India). Of all, using 5-fold cross validation, the best test AUC of 0.926 and 0.904; and accuracy of 82.6% and 84.7% were reported on C2 (MC, USA) and C9 (SH, China), respectively.

2018

As before, Abiyev and Ma'aitah [38] used CNN architecture and compared it with backpropagation neural networks (BPNNs) with supervised learning and competitive neural networks (CpNNs). The idea was not just to improve the performance but also to check whether CNN architecture can be generalized. On a data collection, C27 (Turkey) of size 112,120 frontal-view CXRs, CNN achieved the highest

accuracy for training and testing data as compared to BpNN and CpNN. Using split-based evaluation (train/test: 70/30)), the reported accuracy for CNN is 92.4% with MSE of .0013 MSE, and the accuracy of BpNN is 80.04% with MSE of .0025 MSE, the accuracy of CpNN is 89.57% with MSE of .0036. The study concluded that CNN has relatively better generalization power, and authors reported that the proposed network outperformed all the other state-of-the-art works.

Not just limited to binary decision, Rajpurkar et al. [42] proposed a CNN-based model called 'CheXNeXt' to detect fourteen different pathologies. In addition, authors compared the investigating capabilities of the DL algorithm with radiologists. ChexNeXt employed a 121-layer DenseNet architecture. Each layer was directly connected to every other layer within a block, and for each layer, the feature maps of all preceding layers were used as inputs, and its own feature maps were passed on to all following layers as inputs. In their experiment, on a data collection C26 (MD, USA), the suggested model, was at par with radiologists for 10 pathologies and performed less in three pathologies (Cardiomegaly, Emphysema and Hernia) but algorithmperformed better in detecting atelectasis. In their experiments, they employed train/tune/validate approach, and the average AUC for all pathologies for both radiologists and algorithm are 0.871 and 0.849, respectively.

In Melendez et al. [43], authors evaluated TB detection performance in triaging CXRs in a high-throughput digital mobile TB screening program using CAD4TB5 software. The likelihood score calculated was compared for each CXR. The experiment was done on a data collection of size 38,961 posteroanterior (PA) CXRs that are collected between 2005 and 2010 during the Find and Treat screening program organized in London (C25, UK). In their experiments, they reported a specificity of 0.557, sensitivity of 0.95 and AUC of 0.90. However, there validation protocol was not clear. In this work, the generated heatmap helps in highlighting regions-of-interest. Additionally, it provides explanations for assigned TB scores. However, the study claims that there is a preference for software in high burden and resource-constrained areas, and for triage, the analysis is performed on a low percentage of TB active cases.

Similarly, Zaidi et al. [37] evaluated the performance of CAD4TB, which was primarily designed to help (non-expert) readers detect tuberculosis more accurately and cost-effectively. It resulted a score (0–100) that can be interpreted as the probability that the human subject's severity level (from active TB visible in CXR). The speed of digital CXRs combined with DL and remote expertise made CAD4TB a valuable asset in the fight against TB. Their experiments were performed on a different data collection, C19 (Pakistan) of size 6,845 individuals with presumptive TB enrolled in centers of Pakistan. As in [43], the software calculated the likelihood score for each CXR, and a high score indicated a more severe abnormality score for each CXR. CAD4TB with patients demographic (age and gender) and symptoms gave higher AUC of 0.84. However, their validation protocol was not clear. Authors claim that in low resource settings, CAD4TB as a triage tool could minimize use of experts.

Bekar et al. [31] evaluated the feasibility of pathological patterns of TB in CXR using DL-based detection and classification, which they call 'ViDi' (Suite v2.0; ViDi Systems,Villaz-Saint-Pierre, Switzerland)[44, 45]. On a data collection C3 (Kampala, Uganda) of size 138 patients (with previously diagnosed HIV and TB co-infection) and using train/test: 70/30 approach, their reported AUC was 0.82, and specificity and sensitivity were 0.978 and 0.906, respectively.

Yadav et al. [46] conducted a study on coarse-to–fine knowledge transfer learning to fine-tune the model further using multiple data augmentation techniques. Using data collections C2 (MC, USA), C9 (SH, China), and C26 (MD, USA), their study received a significant improvement in performance 94.8% in detecting TB. They followed split-based validation approach (train/test: 8/20 and 80/15).

2019

Qin et al. [35] conducted a retrospective evaluation of three different CAD systems: CAD4TB, Lunit INSIGHT and qXR for detecting TB-associated abnormalities in CXR images from outpatients in Nepal and Cameroon. Their experiment analysis was not clear about validation approach. For CAD4TB, we refer to [37]. Lunit INSIGHT [15] was designed with augmented accuracy in the detection of chest abnormalities and prioritized work list so radiologists' workflow can be enhanced. qXR software (Qure.ai, India) received CE-certification that integrated with radiology workflow, and the whole screening process was done in 10 milliseconds. Further, they used a heat map or bounding box to point out abnormalities (to the clinician, facilitating rapid confirmation, proving a valuable supplement to the existing healthcare systems). All 1,196 individuals (data collection, C14) received a expert MTB/RIF assay and a CXR read by two groups of radiologists and the DL systems (*ref* C14 in Table 1). Expert was used as the reference standard. In their experiments, AUCs are similar; 0.94 from Lunit, 0.94 from qXR and 0.92 from CAD4TB. Using DL models to read CXRs could cut the number of expert MTB/RIF tests required by 66% while keeping sensitivity at 96% or greater.

Ge et al. [47] proposed two novel error functions: multi-label Softmax loss and correlation loss, for any DL model for better medical image classification. This study's primary motivation is to resolve the two critical issues of medical image classification: multiple labels and visually similar

data. Using train/validate/test (70/10/20) approach, on a data collection C26 (MD, USA), the proposed functions claimed to perform better with the backbone network and with other models for improved performance. The DenseNet bottleneck layer (D121-BL-proposed) gave an AUC of 0.843, W-AUC of 0.589, D-AUC of 0.799 and N-AUC of 0.874.

Hwa et al. [48] developed an ensemble DL for TB detection using CXR as well as their corresponding edge map to handle the diagnosis challenges. The experiment used two public data collections: C2 (MC, USA) and C9 (SH, China). The study reported an accuracy of 89.77%, a sensitivity of 0.909%, and a specificity of 0.886% using 10-fold cross validation. The results showed that features extracted from different images could improve the detection rate, further focusing on types of features and TB classification based on severity level.

More often, DL-based algorithms suffer from computational burden/complexity. It does not necessarily due to they need large amount of data for training; it could also potentially be the size of the architectures. In Pasa et al. [49], authors proposed a simple CNN to automatically diagnosis TB using CXRs, where their primary motivation is to reduce the computational and memory burden. Their CNN architecture was light. It consisted of 5 convolutional blocks, followed by a global average pooling layer (which compressed each feature map to its mean value) and a fully connected Softmax layer with two outputs. Each convolutional block had two 3×3 convolutions with ReLUs, followed by a max-pooling operation. The pooling size was 3×3 with stride 2, similarly to AlexNet. The convolutions were all zero-padded to preserve the input resolution and each convolutional layer also made use of batch normalization to speed up the training procedure and reduce overfitting. Interestingly, authors took an additional step on data visualization. It used saliency maps and gradient weighted class activation method (gradCAMs) for an excellent visual explanation, and helped clinical officers better review and interpret. The experiment used benchmark data collections: C2 (MC, USA), C8 (Belarus), C9 (SH, China), and combined datasets (*ref* in Table 1). They reported accuracies of 79% (C2) , 84.4% (C9), 86.2% (combined) and AUCs of 0.811 (C2), 0.90 (C9), 0.925 (combined). Their experiments followed 5-fold cross validation approach. Authors did not report separate test results on C8 (Belarus) data collection.

In Evangelista Guedes [50], authors developed an ensemble CNN (Inception, ResNet and VGG) to detect TB using CXRs. On publicly available data collections: C2 (MC, USA) and C9 (SH, China), their reported accuracy was higher than 93%, where specificity was 0.94 and sensitivity was 0.93. In their experiments, split-based approach - train/validate/test (70/10/20) was followed.

Ahsan et al. [51] employed VGG16 to perform TB screening with and without data augmentation. On two data collections C2 (MC, USA) and C9 (SH, China) were used in their study, and their model achieved the highest possible accuracy of 81.25% (80%) with (without) data augmentation. In their experiments, split-based approach - train/test (75/25) was followed.

Integrating demographic factors with CXR data provides better and consistent decision-making. In Heo et al. [34], authors examined shift workers' health data at Yonsei University to detect TB using CXRs. The study's motivation was to check the impact of demographic factors, and they compared the performance of image CNN with demographic CNN. To extract features, CNN architectures: VGG19, InceptionV3, ResNet50, and DenseNet121 and InspectionResNetV2 were used. Data collection C13 (Yonei University, South Korea) was used for the study. Both I-CNN (image based CNN) and D-CNN (demography based CNN) models were trained with 1000 chest X-ray images both positive and negative for TB. The main demographic factors were age, weight, height, and gender. The AUC values of the D-CNN models were greater than those from I-CNN . Also, D-CNN (demography based CNN) models resulted in higher sensitivity than I-CNN models (0.815 versus 0.775) and specificity of 0.962. They followed split-based approach (train/test: 80/20) to evaluate their method. Authors claimed that the demographic factors can help build better/consistent TB detection in addition to the use of CXRs.

It is always interesting to compare the performance of AI-guided tools in reading CXRs against radiologist readings. Philipsen et al. [52] evaluated TB detection performance based on computerized CXR readings. Both physician and software (CAD4TB (v5)) read CXRs. The experiment used a data collection C23 (Philippines) of size 14,094 CXRs. Using software, the reported AUC of the software is 0.93; the physician had a sensitivity and specificity of 0.82 and 0.87. In this study, it was not clear about their validation protocol. Even though the reported accuracy shows that the software had a slightly higher sensitivity than the physician, the study lacks measures of statistical significance. The performance of automated CXR is at par with physicians.

2020-21

Without a surprise, data collection is the primary task. In addition, examining them by radiologist another expensive task in the process. In Kim et al. [53], authors collected a total of 111,622 CXRs (*ref.* C26 (MD, USA), in Table 1), where a cardiothoracic radiologist examined a subset of 11,000 CXRs and classified them as positive or negative for possible TB signs. The best-performing algorithm from TBNet (in phase II) was subsequently tested against CXRs from three different sites (2 in the United States, 1 in China) with clinically confirmed TB cases. The algorithm generalized well to CXRs obtained from a tertiary care hospital,

achieving an AUC of 0.87; TBNet's sensitivity, specificity, positive predictive value, and negative predictive value were 0.85, 0.76, 0.64 and 0.90, respectively. Their study followed split-based approach (train/test: 80/20) for evaluation.

Nash et al. [27] carried out a case-control study to evaluate the diagnostic accuracy of CAD software (qXR, Qure.ai, Mumbai, India) using microbiologically-confirmed PTB as the reference standard. On a data collection C6 (Manipal, India) of size 317 CXRs, an AUC for qXR for microbiologically confirmed PTB detection was 0.81. Radiologists had a sensitivity and specificity of 0.56 (95% confidence interval (CI)%) and 0.80 for detecting microbiologically confirmed PTB, respectively. Their algorithm includes a model in assisting clinicians in diagnosing HIV +ve TB cases, but it is not clear about how they evaluated their experiment.

Rajpurkar et al. [54] proposed 'cheXiad' (a DL algorithm) to diagnose TB using CXRs in addition to clinical information. It used a 121-layer DenseNet architecture to extract image features of size 1024. The network forked into two modules, one for TB diagnosis using the image features and the clinical covariates, and the other for predicting the occurrence of six clinical findings that were diagnosed by radiologists. The TB module first used a linear layer to learn 20 image features (from original feature vector of size 1024), and then combined them with the 8 couples to feed the resulting 28-dimensional patient representation into a two layer neural network to predict TB. During inference, each of the five algorithms in the ensemble produced a probability of TB and these probabilities were averaged to get a final, ensemble probability. Using CheXaid, authors performed a diagnostic accuracy study comparing physicians with and without algorithm assistance at the task of diagnosing active pulmonary TB for HIV positive patients (Source: C10 (South Africa)) Physicians' mean accuracy was 0.60 (95% CI 0.57, 0.63) without the algorithm and 0.65 (95% CI 0.60, 0.70) with assistance. Sensitivity was 0.70 (95% CI 0.64, 0.77) without assistance and 0.73 (95% CI 0.66, 0.80) with assistance; specificity was 0.52 (95% CI 0.45, 0.59) without assistance and 0.61 (95% CI 0.52, 0.70) with assistance. However, their validation protocol is not clear.

Rather than investing on TB versus normal CXR classification, creating a large dataset by considering other pulmonary abnormalities is an open problem. In addition, reducing DNN architectures without degrading performance has been one of open challenges. Das et al. [55] proposed a 'Truncated Inception Net' to detect/classify not only TB but also Pneumonia, Covid-19 in addition to normal cases. Authors reduced the model complexity and eventually the number of trainable parameters, which they called 'truncated.' The model was truncated at a point, where it retained 3 Inception modules and 1 grid size reduction block from the beginning. The point of truncation was chosen experimentally, that yielded the best classification results. In their experiments, when using two publicly available TB data collections: C2 (MC, USA) and C9 (SH, China), they reported an accuracy of 99.92% (AUC of 0.99) in categorizing Covid-19 positive cases from combined Pneumonia, TB and normal/healthy cases. To avoid possible bias, they followed 10-fold cross validation. The study claims that they outperformed state-of-the-art results. State-of-the-art techniques are not rich in generalized DL models. In Sathitratanacheewin et al. [56], authors developed a supervised Deep CNN (DCNN) model aiming to check whether machine learning models can be generalized. Using a data collection C9 (SH, China) of size 662 CXRs, an AUC of 0.9845 was reported. Similarly, they reported an AUC of 0.7054 on a different dataset C26 (MD, USA) of size 112,120 CXRs. Their experiments followed train/validate/test (75/15/10) approach. Their results suggested that considering the training dataset from a population may have different diagnostic performances for a diverse population. The technical specifications, severity distribution and dataset distribution (socio-demographics) are vital factors that need examination.

Decision tree has been widely used classifier in the literature. Yoo et al. [33] proposed a DL-based decision tree to predict Covid-19, where they employed three different decisions (binary) in a series. The first decision was made between normal and abnormal CXRs; the second was made between TB and non-TB; and the third one, Covid-19 (with/without TB) and non-Covid-19. Their experiments reported an accuracy of 98%, and 80% for the first two decisions on publicly available data collections C9 (SH, China) and C12 (EAH). The second decision between between TB and non-TB resulted an AUC of 0.80, a senstivity of 0.72 and specificity of 0.89. Like others, their experiments followed train/test (85/15) approach. The study did not use pathologically confirmed data for all disease types (TB and Covid-19) during training process. As a result, results are limited to binary decisions. In our study, we do not consider Covid-19 screening performance.

Feature selection is the must in order not just to remove (redundant features) but also to select distinguished/releveant attributes. Sahlol et al. [57] proposed a novel hybrid method, 'MobileNet-AEO' to classify CXRs, and their motivation was to use Artificial Ecosystem based Optimization (AEO) as a feature selector from redundant deep features that are generated from CNN. The algorithm employed MobileNet to extract parts from CXR that was previously trained on the ImageNet dataset. AEO algorithm was used as a feature selector to check the most relevant features. In their experiments, using split-based approach (train/test:80/20), they reported an accuracy of 90.2%, a specificity of 0.901 and a sensitivity of 0.914 on a data collection C9 (SH, China). Using the exact same validation framework, on another data collection C16 (UK), they reported an accuracy of 94.1%, a specificity of 0.97 and a sensitivity of 0.872.

The algorithm claimed to successfully reduce the number of features from 50K to only 25 and 19 for C9 (SH, China) and C16 (UK), respectively. They further worked by combining the transfer learning model with a meta-heuristic swarm optimisation algorithm.

Instead of performing binary decisions: TB versus non-TB, few works focused on the regions-of-interest with TB lesions. Xie et al. [36] introduced a scalable pyramid structure called faster RCNN (region-based CNN) to detect different categories of TB lesions in CXRs. In brief, their system integrated learning scalable pyramid structure into Faster RCNN and employed reinforcement learning in the trained lesion detection model to detect categories of TB lesions. They used public data collections: C5 (JSRT, Japan), C2 (MC, USA), C9 (SH, China) and an additional local data collection C17 (FAHXJU, China). In their experiments, using split-based approach (train/test:80/20), they reported an accuracy of 92.6% (AUC of 0.977) on C2, and an accuracy of 90.2% (AUC of 0.941) on C9. Authors claimed that their method supported clinical applications.

How well knowledge transfer can be made in DL models has been an open problem in computer vision. Rajaraman and Antani [58] evaluated the effectiveness of knowledge transfer gain through ensemble DL models to detect TB using a large public CXR dataset. The motivation is to improve TB classification algorithms. In their study, they used C15 (Indiana, USA), C18 (USA) and C24 (RSNA, USA) data collections. The knowledge acquired through transfer gain applied to CXRs in C9 data collection. They reported the highest possible accuracy of 94.1% (a 95% CI of 0.899, 0.985) and AUC of 0.995 (95% with CI 0.945, 1.00) when considering the top-3 pre-trained models, using train/test (80/20) split-based validation framework. Even though the study shows the knowledge transfer method has helped improve classification, the study lacks to generalize it due to data availability. The study could potentially be enhanced with larger dataset(s).

Rahman et al. [59] used nine different deep CNNs (ResNet18, ResNet50, ResNet101, ChexNet, InceptionV3, VGG19, DenseNet201, SqueezeNet, and MobileNet) with transfer learning from their pre-trained initial weights and were trained, validated as well as tested for classifying TB and non-TB (normal) cases. Using data collections C2 (MC, USA), C9 (SH, China), C8 (Belarus), and C24 (RSNA, USA), they reported an accuracy, a sensitivity, and a specificity of best performing model (ChexNet) are 96.40%, 0.964 and 0.965, respectively. Their experiments followed 5-fold cross validation approach. When considering lung segmentation, their performance was better using DenseNet201 (accuracy of 98.60, sensitivity of 0.985 and specificity of 0.985).

Guo et al. [60] employed an integrated process to improve TB diagnostics via CNNs and localization in CXRs via deep-learning models. The first step of TB diagnostics process includes modifying CNN model structures, the second step includes model fine-tuning via artificial bee colony algorithm and the final step includes the implementation of linear average- based ensemble method. Comparisons were made across all three steps using deep CNN models on two publicly available data collects namely C9 (SH, China) and C26 (MD, USA). The accuracy, specificity and AUC of ensemble (VGG16, VGG19, Inception V3, ResNet34, ResNet50, and ResNet101) model for standard train/valid ratio 8:2 are 94.50% (95.60%), 0.955 (0.985) and 0.986 (0.976), respectively on C9 (C26) collection. Their experiments followed 10-fold cross-validation approach.

When traditional CNN models do not take uncertainty into account, integrating Bayesian concept is a wise idea. In Abideen et al. [61], authors proposed a TB identification system, which is primarily based on Bayesian-based CNN (B-CNN) to classify between TB and non-TB in uncertain cases. In other words, they focused on the problem with the traditional CNN models, and they did not consider uncertainty to classify CXRs (for TB cases). The experiments used two publicly available data collections: C2 (MC, USA) and C9 (SH, China), and their results are summarized as follows. Using split-based (train/test: 80/20) approach, the highest possible accuracies of 96.42% and 86.46% were achieved on C2 and C9, respectively. In their comparison study, B-CNN showed its superiority over other state-of-the-art methods. Results prove the supremacy of B-CNN for the identification of TB and non-TB sample CXRs as compared to counterparts in terms of accuracy, variance in the predicted probabilities and model uncertainty.

Murphy et al. [62] evaluated the latest version of CAD4TB V6, a commercial software platform (DL-based model) for automatic and inexpensive TB screening, with the motivation to use it in high burdened regions. The experiment used 5,565 CXRs with GeneXpert (Xpert), which was C19 (Pakistan). Using split-based (train/test: 80/20) approach, Version 6 of CAD4TB achieved an AUC of 0.986, a specificity of 0.98 and a sensitivity of 0.90. Authors claimed that CAD4TB v6 is cost-effective and was aimed to be used for in resource-constrained and/or high-task burdened regions.

Habib et al. [32] evaluated the performance of CAD4TB software for people with diabetes (PWD). The experiment used 694 participants enrolled as a part of a comprehensive bi-directional screening program for TB and Diabetes in Pakistan (C11 data collection). They screened with CAD4TB and concurrently provided sputum for Xpert MTB/RIF testing. The experiment reported the AUC was 0.78 (95% CI: 0.77–0.80). The software offers good diagnostic accuracy as a triage test for TB screening among PWD using Xpert MTB/RIF as the reference standard.For a CAD cut-off score of 50 the specificity and sensitivity are 0.424 (42.4%) and 0.905 (90.5%) respectively. The study did not

tell us the version of this software. Also, it did not include a patient history for PWD identification, an association of glycemic control with CAD4TB scores, and MTB positivity needs evaluation. Their train/test evaluation protocol is not clear either.

Different models produce different results. Combining their decisions attracts researchers' interests, where one requires to understand whether models complement each other. In Rajaraman and Antani [63], authors proposed methodology explicit collective learning toward improving abnormality detection in CXRs. Model predictions are combined using different ensemble strategies toward reducing prediction variance and sensitivity to the training data while improving overall performance and generalization. The publicly available data collections: C24 (RSNA, USA) and C28 (USA) were used in this retrospective study. Using hold-out validation approach: 80% for training and 20% for testing, it was observed that the model groups show prevalent confinement execution as far as Intersection of Union (IoU) and mean Average Precision (mAP) measurements as compared to any individual constituent model. Ensemble 5: VGG-16, Xception, NASNet-mobile, Inception-V3 and MobileNet showed prevalent execution for IoU and mAP metrics (0.433 and 0.447 separately). Ensemble of top 7 CNNs: VGG-16, VGG-19, Xception, NASNet-mobile, Inception-V3, MobileNet and DenseNet-121 were used for detection, where the weighted average of these models gave an accuracy of 91.63%, an AUC of 0.974, a specificity of 0.884 and a sensitivity of 0.924. Further, they used them for visualizing abnormalities in CXRs.

In Ayaz et al. [64], through ensemble learning, a unique TB detection methodology was suggested that combines hand-crafted features with deep features (convolutional neural network-based). Gabor Filter was used to extract hand-crafted features, while pre-trained DL models were used to extract deep features. The suggested approach was evaluated using two publicly available data collections: C2 (MC, USA) and C9 (SH, China). Using k-fold cross-validation, for C2 (k=6) and C9 (k=10) data collections, AUCs of 0.97 and 0.99 were achieved, respectively, demonstrating the superiority of the suggested approach.

In CXRs, textures (within the lung sections) change in accordance with abnormalities. As a result, changes in bone/rib structure happen. As such changes vary with the severity level of pulmonary abnormality, they complicate abnormality decision-making. In Rajaraman et al. [65], authors developed a DL-based bone suppression model that detected and removed occluding bony structures in frontal CXRs. With 6-fold cross validation, the best-performing model (ResNet-BS) (PSNR = 34.0678; MS-SSIM = 0.9828) is used to suppress bones in the publicly available and TB CXR collections C2 (MC, USA) and C9 (SH, China). The primary aim was to help minimize radiological interpretation errors. In addition, DL algorithm was used to detect TB manifestations on two different data collections: C2 and C9. In their experiments, they proved that models that were trained on bone-suppressed CXRs (C2: AUC = 0.9635 ± 0.0106; C9: AUC = 0.9535 ± 0.0186) significantly outperformed ($p < 0.05$) the models trained on non-bone-suppressed CXRs (C2: AUC = 0.8567 ± 0.0870; C9: AUC = 0.8991 ± 0.0268). Models trained on bone-suppressed CXRs enhanced TB detection and resulted in compact clustering of data points in the feature space, indicating that bone suppression increased model sensitivity for TB classification.

Table 3 Comparative study on C2 (MC, USA) data collection

Authors (year)	ACC (in %)	AUC	SPEC	SEN
Hwang et al. (2016) [39]	67.40	0.88	–	–
Lakhani and Sundaram (2017) [12]*	–	0.99	1.00	0.97
Lopes and Valiati (2017) [41]	82.60	0.93	–	–
Yadav et al. (2018) [46]*	94.89	–	–	–
Hwa et al. (2019) [48]*	89.77	–	0.89	0.91
Pasa et al. (2019) [49]*	86.20	0.93	–	–
Evangelista and Guedes (2019) [50]*	93.82	–	0.94	0.93
Ahsan et al. (2019) [51]*	81.25	–	–	–
Das et al. (2020) [55]*	99.92	0.99	1.00	0.93
Xie et al. (2020) [36]	92.60	0.98	0.92	0.93
Rahman et al. (2020) [59]*	96.40	–	0.97	0.96
Abideen et al. (2020) [61]	96.42	–	–	–
Ayaz et al. (2021) [64]	93.47	0.97	–	–
Rajaraman et al. (2021) [65]	92.30	0.96	0.97	0.88

*Other data collections were employed in addition to C2

Discussion Outline

In this section, we elaborate our study by taking into account the following factors: a) performance comparison and b) generic analysis of DL-based medical imaging tools. For a fair performance comparison, we are required to study whether exact same datasets and evaluation protocols were employed. Whereas, for DL-based medical imaging tools, we are not limited to binary decision-making (TB versus normal) but also to check whether such tools can be used for region-of-interest (or pathology) localization.

Performance Comparison

We repeat, comparison may not be fair and comprehensive among different DL-based methods if we consider one specific year as we need to check whether their methods used exact same datasets and evaluation protocol as well as

Table 4 Comparative study on C8 (Belarus) data collection

Authors (year)	ACC (in %)	AUC	SPEC	SEN
Lakhani and Sundaram (2017) [12]⋆	–	0.99	1.00	0.97
Pasa et al. (2019) [49]⋆	86.20	0.93	–	–
Rahman et al. (2020) [59]⋆	96.40	–	0.97	0.96

⋆Other data collections were employed in addition to C8

Table 6 Comparative study on C24 (RSNA, USA) data collection

Authors (year)	ACC (in %)	AUC	SPEC	SEN
Rajaraman and Antani (2020) [58]⋆	94.10	0.99	0.96	0.93
Rahman et al. (2020) [59]⋆	96.40	–	0.97	0.96
Rajaraman et al. (2020) [63]⋆	91.63	0.97	0.88	0.92

⋆Other data collections were employed in addition to C24

performance metrics. Since we found that different systems used different datasets as well as validation protocols, primarily hold-out or split-based, k-fold cross validation, it is not fair to compare them. However, we aim at providing an idea of how far they have been advancing over years (since 2016) using specific data collections such as C2 and C8.

Following datasets (*ref.* Table 1) and Table 2, in Tables 3, 4, 5, 6 and 7, we compare different DL-based methods. In what follows, we discuss on their performance scores (Tables 3, 4, 5, 6 and 7).

1. C2 collection (MC, USA): Of 14 articles, Das et al. [55] reported the highest accuracy of 99.91% (and their corresponding AUC of 0.99, SPEC of 1 and SEN of 0.93). The alarming issue is their sensitivity is lower than other average methods. Also, Das et al. [55] did not work solely on TB screening, they worked on Covid-19 versus TB. Other works provided marginal differences. These results are similar to those from Lakhani et al. [12] (AUC =0.99, SPEC = 1.0 and SEN = 0.973).
2. C8 collection (Belarus): Rahman et al. [59] have obtained the highest accuracy of 96.40 among other two authors who used the same data collection. Lakhani and Sundaram [12] reported higher AUC than Pasa et al. [49] (0.99 versus 0.925). It is observed that even though all three authors results are from combinations of different datasets, numbers of samples in their experiments were different: Lakhani et al. (2017) used 88 CXRs and Pasa et al. (2019), Rahman et al. (2020) used 304 and 306 positive CXRs respectively (from C8 dataset).
3. C9 collection (SH, China): Like C2 (MC, USA), C9 (SH, China) is another popular dataset. Of 19 articles, Das et al. [55] reported the highest accuracy of 99.91% (and their corresponding AUC of 0.99, SPEC of 1 and SEN of 0.93), which is followed by Ayaz et al. [64] with an accuracy and AUC of 97.59 and 0.99, respectively. Noe that Das et al. [55] results were not solely based on C9 collection as compared to Ayaz et al. [64], where results were reported from C9. The lowest performance scores were reported by Hwang et al. [39] (accuracy of 83.70 and AUC of 0.93) - which was the beginning of DL-model based algorithm for TB detection.
4. C24 collection (RSNA, USA): Apart from C24 collection, authors Rajaraman and Antani [58, 63] also used other data collections. Rahman et al. [59] used C2, C8, C9 and C24 (27,484); Rajaraman and Antani [58], used

Table 5 Comparative study on C9 (SH, CHina) data collection

Authors (year)	ACC (in %)	AUC	SPEC	SEN
Hwang et al. (2016) [39]	83.70	0.93	–	–
Lakhani and Sundaram (2017) [12]⋆	–	0.99	1.00	0.97
Lopes and Valiati (2017) [41]	84.70	0.90	–	–
Yadav et al. (2018) [46]⋆	94.89	–	–	–
Hwa et al. (2019) [48]⋆	89.77	–	0.89	0.91
Pasa et al. (2019) [49]⋆	86.20	0.93	–	–
Evangelista and Guedes (2019) [50]⋆	93.82	–	0.94	0.93
Ahsan et al. (2019) [51]⋆	81.25	–	–	–
Das et al.(2020) [55]⋆	99.92	0.99	1.00	0.93
Sathitratanacheewin et al. (2020) [56]	–	0.98	0.82	0.72
Yoo et al. (2020) [33] ⋆	80.00	0.80	0.89	0.72
Sahlol et al. (2020) [57]	90.20	–	0.90	0.91
Xie et al. (2020) [36]	90.20	0.94	0.95	0.85
Rajaraman and Antani (2020) [58]⋆	94.10	0.995	0.96	0.93
Rahman et al. (2020) [59]⋆	96.40	–	0.97	0.96
Guo et al. (2020) [60]	94.5	–	0.96	0.99
Abideen et al. (2020) [61]	86.46	–	–	–
Ayaz et al. (2021) [64]	97.59	0.99	–	–
Rajaraman et al. (2021) [65]	88.79	0.95	0.89	0.88

⋆Other data collections were employed in addition to C9

Table 7 Comparative study on C26 (MD, USA) data collection

Authors (year)	ACC (in %)	AUC	SPEC	SEN
Rajpurkar et al. (2018) [42]	–	0.85	–	–
Yadav et al. (2018) [46]⋆	94.89	–	–	–
Ge et al. (2019) [47]	–	0.84	–	–
Kim et al. (2020) [53]	–	0.87	0.76	0.85
Sathitratanacheewin et al. (2020) [56]	–	0.71	–	–
Guo et al. (2020) [60]	95.60	–	0.99	0.98

⋆Other data collections were employed in addition to C26

C9, C15, C18 and C24 (37,306); and Rajaraman and Antani [63] used C24 and C28 (250,332). The later work used more parameters in their ensemble DNNs. Rahman et al. [59] did not report AUC but specificity and sensitivity are 0.965 and 0.964 respectively, which are highest among the other two. Authors stated that their objective was not just to improve TB detection performance [58] or reliable TB detection method [59] but also to check whether abnormality visualization was possible [59, 63].

5. C26 collection (USA): Of 6 research articles, ensemble architecture proposed by Guo et al. [60] reported the highest accuracy compared to Yadav et al. (2018) [46] (95.60 vs 94.89) and other authors did not report the accuracy. Among all the authors with reported AUC, the highest difference is 0.20. Guo et al. (2020) and Kim et al. (2020) reported the specificity and sensitivity and the earlier reports the highest in both parameters(0.985,0.976 vs 0.76,0.85).

Among all articles, we have observed that a) data collections: C2 (MC, USA) and C9 (SH, China) are mostly used (publicly available), and b) performance increasing trend (starting from 2016) either by increasing datasets or by introducing ensemble DNNs and/or DL-based algorithms. Regarding dataset size, authors combined multiple sources of data as well as applied data augmentation techniques. Unlike in computer vision (generic problem), data augmentation techniques may not guarantee whether their pixel re-arrangements can produce possible clinical TB manifestations.

DL-Based Medical Imaging Tools: Decision-making to Data Visualization

In this section, we discuss on DL-based medical imaging tools (TB screening purpose), where transfer learning, data augmentation and data visualization (disease localization) are considered.

Transfer Learning

Unlike conventional machine learning, huge amounts of annotated data are required for DL models to deliver desirable performance. However, medical image datasets are usually relatively small due to the overall lower incidence of disease, and few large datasets exist that are also appropriately balanced. As a result transfer learning (TL) methods are more commonly used to transfer the learned knowledge on a large collection of natural images [66]. Subsequently, the model is fine-tuned for a target task of interest. TL approaches are employed by several researchers toward the task of TB detection on CXRs. Some of these approaches used a combination of conventional hand-crafted feature descriptors/classifiers and DL models toward the TB detection task [12, 67]. It was observed that the DL models delivered superior results compared to the conventional hand-crafted feature-based classification. Lakhani and Sundaram [12] used ImageNet-pretrained CNN models and fine-tuned them to classify the CXRs from Shenzhen [67], Montgomery [67], and Belarus TB CXR collections as showing normal lungs or those with TB-consistent manifestations. The authors constructed an averaging ensemble of these models to obtain an AUC of 0.99 toward this classification task compared to an AUC of 0.94 obtained by the models trained with random weight initializations. Hwang et al. [39] employed TL methods to fine-tune the ImageNet-pretrained DL models on a private CXR dataset for detecting TB-consistent findings. The DL models pretrained on the ImageNet collection were observed to deliver markedly improved performance with accuracy and AUC of 83% and 0.926 respectively, compared to those models that are trained with random weight initializations. Pasa et al. [49] constructed a custom CNN model to classify Shenzhen, Montgomery, and Belarus TB CXR collections as showing normal lungs or TB-consistent findings. The custom CNN model was found to be 86.2% accurate toward this classification task and obtained 0.925 AUC. Qin et al. [68] used TL methods to transfer the learned knowledge from DL models and classified the 10,848 CXRs in a private collection as showing normal lungs or TB-like manifestations. The fine-tuned models were found to be 90.3% accurate and obtained 0.964 AUC toward this classification task.

CXRs have distinct visual characteristics such as edges, colors, contours, and orientations compared to natural images. Hence, the knowledge transfer from a model pretrained on natural images may not be relevant for a medical visual recognition task. In this regard, researchers retrained the ImageNet-pretrained DL models on a large-scale collection of CXRs to convert the weight layers specific to the CXR modality and help the models to broadly learn the characteristics of normal and abnormal lungs. Rajaraman and Antani [58] performed modality-specific model retraining and transferred the modality-specific learned knowledge to fine-tune for a relevant task of classifying the CXRs as showing TB-consistent findings or normal lungs. It was observed that a stacked ensemble of CXR modality-specific pretrained and fine-tuned models delivered a superior performance with 94.1% accuracy and 0.995 AUC toward classifying the CXRs in the Shenzhen TB collections as showing TB-consistent findings. Yadav et al. [46] performed an ensemble of CXR modality-specific models that were fine-tuned to classify the images in the Shenzhen and Montgomery collections as showing TB-consistent manifestations with 94.89% accuracy.

Data Augmentation

Data augmentation plays a pivotal role in reducing overfitting and improving the generalization performance of DL models. This is particularly true under circumstances

of sparse data availability such as for medical computer vision applications. Augmenting the training data is done by applying traditional random transformations. These include image resizing, unsharp masking, zooming, shearing, channel splitting, rotation, cropping, pixel-shifting, Gaussian smoothing, etc. These and other perturbations are shown to introduce diversity by generating new training instances, thereby helping the model learn robust feature representations and increasing generalization [59, 60]. In this regard, data augmentation helps in regularizing the model by training on diversified data and also alleviating class imbalance issues that may lead to model overfitting. Data augmentation methods are widely used in CXR classification tasks. Ganesan et al. compared the performance of traditional and generative adversarial networks (GAN)-based augmentation techniques toward classifying CXRs as showing normal or abnormal lungs [69]. A progressive-growing GAN (PG-GAN) was trained to synthesize CXRs of the minority (abnormal) class. Traditional augmentations such as Gaussian smoothing, unsharp masking, and minimum filtering were performed to augment the abnormal class. An ImageNet-pretrained VGG-16 model was fine-tuned and evaluated separately on the data augmented with traditional and GAN-based methods. It was observed that the performance obtained with traditional augmentation (84.25% accuracy) was superior compared to GAN-based methods (83.9% accuracy). The authors also observed that the classifier performance improved with increased numbers of augmented images through traditional and GAN-based strategies. In another study, Rajaraman and Antani studied the effects of data augmentation toward the task of classifying the CXRs as showing normal lungs or Covid-19-related manifestations [70]. The authors augmented the training data with CXRs pooled from publicly available datasets showing pneumonia-related manifestations and compared those with models trained on non-augmented data toward this classification task. It was observed that such augmentation improved classification performance (65.36% accuracy) compared to non-augmented model training (50.28% accuracy). However, augmenting the training data using CXRs with Covid-19 labels acquired from cross-institutional collections improved the classification performance further (88.89% accuracy). These empirical observations underscore the fact that Covid-19 manifests with distinct visual characteristics in CXRs compared with those manifested through community-acquired bacterial and viral pneumonia infections. Applied to detecting TB manifestations in CXRs, Ahsan et al. [51] obtained superior performance with data augmentation (81.25% accuracy) compared to non-augmented model training (80% accuracy) while using an ImageNet-pretrained DL model to classify CXRs as showing TB-like manifestations.

Disease ROI Localization

Although DL models demonstrate astounding success in natural and medical visual recognition tasks, their predictions lack explanations due to their black-box behavior. This lack of explainability, or interpretability, is considered a serious bottleneck in medical screening and diagnosis and may restrict their use in clinical practice. In this regard, there is an ever-increasing ask for interpreting the learned behavior of these models, particularly toward supplementing medical decision-making. Currently, researchers have published several studies that attempt to interpret model predictions in natural and medical visual recognition tasks. The study published by Zeiler, and Fergus used a model with deconvolutional blocks and performed ablation studies to explain the features learned at the model layers [71]. Another study by Mahendran and Vedaldi used invert representations to interpret the learned features at the intermediate model layers [72]. The authors observed that the layers learned photometric and geometric features specific to the class samples. Zou et al. proposed a visualization method called class activation mapping (CAM) to localize ROIs that were responsible to categorize the images to their respective classes [73]. The principal limitation of this method is that it can be used to visualize the learned activations in DL models with a fixed architecture. A generalized version of CAM, known as gradient-weighted class activation mapping (Grad-CAM) was proposed by Selvaraju et al. [74] that can be used to visualize the learned ROIs in CNN models with diversified architecture. Another model agnostic visualization approach is the Local Interpretable Model-Agnostic Explanations (LIME) [75]. Rajaraman et al. used it for visualize and explain the predictions of DL models trained to detect pneumonia in CXRs [76].

Applied to disease detection using CXRs, researchers have used these visualization methods to interpret the learned features in their respective classification tasks. Wang et al. [77] used Grad-CAM-based visualization to localize ROIs that stand indicative of pneumonia-like manifestations in CXRs. In another study, the authors used Grad-CAM-based visualization tools to detect ROIs that indicate pneumonia-like manifestations and further categorize these as caused by bacterial and viral infections in CXRs [78]. Pasa et al. explained the learned activations of a customized CNN model [49] using saliency maps and Grad-CAM activations toward the task of TB detection in CXRs. The authors observed salient activations in the right upper lobe of a CXR instance that was consistent with TB-like manifestations. It was observed that the Grad-CAM-based activations demonstrated superior disease-specific ROI localization in deeper model layers compared to earlier layers. However, Grad-CAM activations were reported to be of sub-optimal resolution compared to saliency maps and not found to be

useful for disease diagnosis. Another study [79] proposed a visualization method called class-selective relevance mapping (CRM) that measures both positive and negative spatial element contributions to categorize the images to their respective classes. The proposed visualization method demonstrated improved localization performance compared to Grad-CAM-based visualization in localizing the ROIs in a medical modality classification task. The authors of [63] proposed a modality-specific model ensemble to improve the detection of abnormalities in CXRs. CRM-based visualization was used to interpret the learned behavior of the model ensemble and its constituent models. Increased activations were observed in discriminative ROIs that indicated abnormal manifestations that explained model predictions. Improved disease ROI localization performance measured in terms of mean average precision (MAP) and the intersection of union (IOU) was demonstrated by the model ensemble compared to the individual constituent models.

Conclusions

In this paper, we have systematically reviewed 54 research articles that were reported since 2016. We have focused on Deep Learning (DL) algorithms for Tuberculosis (TB) screening using chest X-ray (CXR) images. The study was not limited to data collections (and their corresponding sources) but also comprehensively reviewed DL's potential, promises and pitfalls in analyzing CXR images, where authors have discussed on performance comparison, transfer learning, data augmentation, decision-making to pathology visualization. Authors observed that, in 2016, research studies were mostly based on decision-making (binary) by taking DL-based algorithms (plug and play), whereas after 2017, extended tools were implemented.

Since the beginning of 2020, we observed that several studies been extended in proposing model pruning protocols [80], analyzing the variability in the annotations [81], and suppressing bony structures to improve soft tissue visibility [82] for improving TB classification and localization performance.

Acknowledgements This research was supported in part by the Intramural Research Program of the National Library of Medicine (NLM), National Institutes of Health (NIH).

Author Contributions KC Santosh conceptualized the work, analyzed results, and wrote/revised the manuscript. S Allu collected review contents. S Rajaraman helped revise the manuscript. S Antani helped conceptualize the work, edited, and proofread the manuscript.

Funding Not applicable.

Data Availibility Statement Not applicable.

Declarations

Ethical Approval and Consent to Participate This article does not include any human participant studies conducted by any of the authors.

Human and Animal Ethics This study did not include any human subjects or animals.

Consent for Publication This article contains no identifying information, so it is inapplicable.

Competing Interests There are no potential conflicts of interest reported by any of the authors.

References

1. World Health Organization (WHO), "Global tuberculosis report 2020: executive summary," 2020.
2. S. Jain, S. Andronikou, P. Goussard, S. Antani, D. Gomez-Pastrana, C. Delacourt, J. Starke, A. Ordonez, P. Jean-Philippe, R. Browning, and C. Perez-Velez, "Advanced imaging tools for childhood tuberculosis: potential applications and research needs," *Lancet Infect Dis.*, vol. 20, no. 11, pp. e289–e297, 2020.
3. R. Piccazzo, F. Paparo, and G. Garlaschi, "Diagnostic accuracy of chest radiography for the diagnosis of tuberculosis (tb) and its role in the detection of latent tb infection: a systematic review," *The Journal of Rheumatology Supplement*, vol. 91, pp. 32–40, 2014.
4. A. Van't Hoog, M. Langendam, E. Mitchell, F. Cobelens, D. Sinclair, M. Leeflang, and K. Lonnroth, "A systematic review of the sensitivity and specificity of symptom-and chest-radiography screening for active pulmonary tuberculosis in hiv-negative persons and persons with unknown hiv status," *Systematic screening for active tuberculosis: principles and recommendations: World Health Organization*, 2013.
5. L. M. Pinto, M. Pai, K. Dheda, K. Schwartzman, D. Menzies, and K. R. Steingart, "Scoring systems using chest radiographic features for the diagnosis of pulmonary tuberculosis in adults: a systematic review," *European Respiratory Journal*, vol. 42, no. 2, pp. 480–494, 2013.
6. J. B. Bomanji, N. Gupta, P. Gulati, and C. J. Das, "Imaging in tuberculosis," *Cold Spring Harbor perspectives in medicine*, vol. 5, no. 6, p. a017814, 2015.
7. E. Skoura, A. Zumla, and J. Bomanji, "Imaging in tuberculosis," *International Journal of Infectious Diseases*, vol. 32, pp. 87–93, 2015. Special Issue: Commemorating World Tuberculosis Day 2015.
8. R. Arora, "The training and practice of radiology in india: current trends," *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, 2014.
9. D. Braun, M. Singhof, M. Tatusch, and S. Conrad, "Convolutional neural networks for multidrug-resistant and drug-sensitive tuberculosis distinction.," in *CLEF (Working Notes)*, 2017.
10. R. Dhoot, J. M. Humphrey, P. O'Meara, A. Gardner, C. J. McDonald, K. Ogot, S. Antani, J. Abuya, and M. Kohli, "Implementing a mobile diagnostic unit to increase access to imaging and laboratory services in western kenya," *BMJ Global Health*, vol. 3, no. 5, 2018.
11. S. Latif, M. Usman, S. Manzoor, W. Iqbal, J. Qadir, G. Tyson, I. Castro, A. Razi, M. N. K. Boulos, A. Weller, *et al.*, "Leveraging data science to combat covid-19: A comprehensive review," *IEEE Transactions on Artificial Intelligence*, 2020.
12. P. Lakhani and B. Sundaram, "Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks," *Radiology*, vol. 284, no. 2, pp. 574–582, 2017.
13. H. Y. Paul, T. K. Kim, J. Wei, J. Shin, F. K. Hui, H. I. Sair, G. D. Hager, and J. Fritz, "Automated semantic labeling of pediatric musculoskeletal radiographs using deep learning," *Pediatric radiology*, vol. 49, no. 8, pp. 1066–1070, 2019.

14. T. K. Kim, H. Y. Paul, J. Wei, J. W. Shin, G. Hager, F. K. Hui, H. I. Sair, and C. T. Lin, "Deep learning method for automated classification of anteroposterior and posteroanterior chest radiographs," *Journal of digital imaging*, vol. 32, no. 6, pp. 925–930, 2019.
15. E. J. Hwang, S. Park, K.-N. Jin, J. I. Kim, S. Y. Choi, J. H. Lee, J. M. Goo, J. Aum, J.-J. Yim, C. M. Park, *et al.*, "Development and validation of a deep learning–based automatic detection algorithm for active pulmonary tuberculosis on chest radiographs," *Clinical Infectious Diseases*, vol. 69, no. 5, pp. 739–747, 2019.
16. C. Tataru, D. Yi, A. Shenoyas, and A. Ma, "Deep learning for abnormality detection in chest x-ray images," in *IEEE Conference on Deep Learning*, 2017.
17. S. Vajda, A. Karargyris, S. Jaeger, K. Santosh, S. Candemir, Z. Xue, S. Antani, and G. Thoma, "Feature selection for automatic tuberculosis screening in frontal chest radiographs," *Journal of medical systems*, vol. 42, no. 8, pp. 1–11, 2018.
18. K. Santosh and S. Antani, "Automated chest x-ray screening: Can lung region symmetry help detect pulmonary abnormalities?," *IEEE transactions on medical imaging*, vol. 37, no. 5, pp. 1168–1177, 2017.
19. K. Santosh, S. Vajda, S. Antani, and G. R. Thoma, "Edge map analysis in chest x-rays for automatic pulmonary abnormality screening," *International journal of computer assisted radiology and surgery*, vol. 11, no. 9, pp. 1637–1646, 2016.
20. S. Govindarajan and R. Swaminathan, "Analysis of tuberculosis in chest radiographs for computerized diagnosis using bag of keypoint features," *Journal of medical systems*, vol. 43, no. 4, pp. 1–9, 2019.
21. A. Karargyris, J. Siegelman, D. Tzortzis, S. Jaeger, S. Candemir, Z. Xue, K. Santosh, S. Vajda, S. Antani, L. Folio, *et al.*, "Combination of texture and shape features to detect pulmonary abnormalities in digital chest x-rays," *International journal of computer assisted radiology and surgery*, vol. 11, no. 1, pp. 99–106, 2016.
22. C. Wang, A. Elazab, J. Wu, and Q. Hu, "Lung nodule classification using deep feature fusion in chest radiography," *Computerized Medical Imaging and Graphics*, vol. 57, pp. 10–18, 2017.
23. M. E. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. Al Emadi, *et al.*, "Can ai help in screening viral and covid-19 pneumonia?," *IEEE Access*, vol. 8, pp. 132665–132676, 2020.
24. M. Owais, M. Arsalan, T. Mahmood, Y. H. Kim, and K. R. Park, "Comprehensive computer-aided decision support framework to diagnose tuberculosis from chest x-ray images: Data mining study," *JMIR medical informatics*, vol. 8, no. 12, p. e21790, 2020.
25. S. M. A. Zaidi, S. S. Habib, B. Van Ginneken, R. A. Ferrand, J. Creswell, S. Khowaja, and A. Khan, "Evaluation of the diagnostic accuracy of computer-aided detection of tuberculosis on chest radiography among private sector patients in pakistan," *Scientific reports*, vol. 8, no. 1, pp. 1–9, 2018a.
26. A. Liberati, D. G. Altman, J. Tetzlaff, C. Mulrow, P. C. Gøtzsche, J. P. A. Ioannidis, M. Clarke, P. J. Devereaux, J. Kleijnen, and D. Moher, "The prisma statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration," *BMJ*, vol. 339, 2009.
27. M. Nash, R. Kadavigere, J. Andrade, C. A. Sukumar, K. Chawla, V. P. Shenoy, T. Pande, S. Huddart, M. Pai, and K. Saravu, "Deep learning, computer-aided radiography reading for tuberculosis: a diagnostic accuracy study from a tertiary hospital in india," *Scientific reports*, vol. 10, no. 1, pp. 1–10, 2020.
28. R. Griesel, A. Stewart, H. van der Plas, W. Sikhondze, M. X. Rangaka, M. P. Nicol, A. P. Kengne, M. Mendelson, and G. Maartens, "Optimizing Tuberculosis Diagnosis in Human Immunodeficiency Virus-Infected Inpatients Meeting the Criteria of Seriously Ill in the World Health Organization Algorithm," *Clinical Infectious Diseases*, vol. 66, pp. 1419–1426, 11 2017.
29. D. J. Van Hoving, H. J. L. Sa'ad Lahri, M. P. Nicol, G. Maartens, and G. Meintjes, "The real-world performance and inter-observer agreement of urine lipoarabinomannan in diagnosing hiv-associated tuberculosis in an emergency center," *Journal of acquired immune deficiency syndromes (1999)*, vol. 81, no. 1, p. e10, 2019.
30. Y. Liu, Y.-H. Wu, Y. Ban, H. Wang, and M.-M. Cheng, "Rethinking computer-aided tuberculosis diagnosis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
31. A. Becker, C. Blüthgen, C. Sekaggya-Wiltshire, B. Castelnuovo, A. Kambugu, J. Fehr, T. Frauenfelder, *et al.*, "Detection of tuberculosis patterns in digital photographs of chest x-ray images using deep learning: feasibility study," *The International Journal of Tuberculosis and Lung Disease*, vol. 22, no. 3, pp. 328–335, 2018.
32. S. S. Habib, S. Rafiq, S. M. A. Zaidi, R. A. Ferrand, J. Creswell, B. Van Ginneken, W. Z. Jamal, K. S. Azeemi, S. Khowaja, and A. Khan, "evaluation of computer aided detection of tuberculosis on chest radiography among people with diabetes in karachi pakistan," *Scientific reports*, vol. 10, no. 1, pp. 1–5, 2020.
33. S. H. Yoo, H. Geng, T. L. Chiu, S. K. Yu, D. C. Cho, J. Heo, M. S. Choi, I. H. Choi, C. Cung Van, N. V. Nhung, *et al.*, "Deep learning-based decision-tree classifier for covid-19 diagnosis from chest x-ray imaging," *Frontiers in medicine*, vol. 7, p. 427, 2020.
34. S.-J. Heo, Y. Kim, S. Yun, S.-S. Lim, J. Kim, C.-M. Nam, E.-C. Park, I. Jung, and J.-H. Yoon, "Deep learning algorithms with demographic information help to detect tuberculosis in chest radiographs in annual workers' health examination data," *International journal of environmental research and public health*, vol. 16, no. 2, p. 250, 2019.
35. Z. Z. Qin, M. S. Sander, B. Rai, C. N. Titahong, S. Sudrungrot, S. N. Laah, L. M. Adhikari, E. J. Carter, L. Puri, A. J. Codlin, *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," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.
36. Y. Xie, Z. Wu, X. Han, H. Wang, Y. Wu, L. Cui, J. Feng, Z. Zhu, and Z. Chen, "Computer-aided system for the detection of multicategory pulmonary tuberculosis in radiographs," *Journal of Healthcare Engineering*, vol. 2020, 2020.
37. S. M. A. Zaidi, S. S. Habib, B. Van Ginneken, R. A. Ferrand, J. Creswell, S. Khowaja, and A. Khan, "Evaluation of the diagnostic accuracy of computer-aided detection of tuberculosis on chest radiography among private sector patients in pakistan," *Scientific reports*, vol. 8, no. 1, pp. 1–9, 2018b.
38. R. H. Abiyev and M. K. S. Ma'aitah, "Deep convolutional neural networks for chest diseases detection," *Journal of healthcare engineering*, vol. 2018, 2018.
39. S. Hwang, H. Kim, J. Jeong, and H. Kim, "A novel approach for tuberculosis screening based on deep convolutional neural networks," in *Medical Imaging 2016: Computer-Aided Diagnosis, San Diego, California, United States, 27 February - 3 March 2016* (G. D. Tourassi and S. G. A. III, eds.), vol. 9785 of *SPIE Proceedings*, p. 97852W, SPIE, 2016.
40. J. Melendez, C. I. Sánchez, R. H. Philipsen, P. Maduskar, R. Dawson, G. Theron, K. Dheda, and B. Van Ginneken, "An automated tuberculosis screening strategy combining x-ray-based computer-aided detection and clinical information," *Scientific reports*, vol. 6, no. 1, pp. 1–8, 2016.
41. U. Lopes and J. F. Valiati, "Pre-trained convolutional neural networks as feature extractors for tuberculosis detection," *Computers in biology and medicine*, vol. 89, pp. 135–143, 2017.
42. P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists," *PLoS medicine*, vol. 15, no. 11, p. e1002686, 2018.
43. J. Melendez, L. Hogeweg, C. Sánchez, R. Philipsen, R. Aldridge, A. Hayward, I. Abubakar, B. van Ginneken, and A. Story,

"Accuracy of an automated system for tuberculosis detection on chest radiographs in high-risk screening," *The International Journal of Tuberculosis and Lung Disease*, vol. 22, no. 5, pp. 567–571, 2018.
44. Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, *Efficient BackProp*, pp. 9–48. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.
45. G. E. Hinton, "To recognize shapes, first learn to generate images," in *Computational Neuroscience: Theoretical Insights into Brain Function* (P. Cisek, T. Drew, and J. F. Kalaska, eds.), vol. 165 of *Progress in Brain Research*, pp. 535–547, Elsevier, 2007.
46. O. Yadav, K. Passi, and C. K. Jain, "Using deep learning to classify x-ray images of potential tuberculosis patients," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2368–2375, 2018.
47. Z. Ge, D. Mahapatra, X. Chang, Z. Chen, L. Chi, and H. Lu, "Improving multi-label chest x-ray disease diagnosis by exploiting disease and health labels dependencies," *Multimedia Tools and Applications*, pp. 1–14, 2019.
48. M. H. A. Hijazi, S. K. T. Hwa, A. Bade, R. Yaakob, and M. S. Jeffree, "Ensemble deep learning for tuberculosis detection using chest x-ray and canny edge detected images," *IAES International Journal of Artificial Intelligence*, vol. 8, no. 4, p. 429, 2019.
49. F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, and D. Pfeiffer, "Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization," *Scientific reports*, vol. 9, no. 1, pp. 1–9, 2019.
50. L. G. C. Evangelista and E. B. Guedes, "Ensembles of convolutional neural networks on computer-aided pulmonary tuberculosis detection," *IEEE Latin America Transactions*, vol. 17, no. 12, pp. 1954–1963, 2019.
51. M. Ahsan, R. Gomes, and A. Denton, "Application of a convolutional neural network using transfer learning for tuberculosis detection," in *2019 IEEE International Conference on Electro Information Technology (EIT)*, pp. 427–433, 2019.
52. R. Philipsen, C. Sánchez, J. Melendez, W. Lew, and B. van Ginneken, "Automated chest x-ray reading for tuberculosis in the philippines to improve case detection: a cohort study," *The International Journal of Tuberculosis and Lung Disease*, vol. 23, no. 7, pp. 805–810, 2019.
53. T. K. Kim, H. Y. Paul, G. D. Hager, and C. T. Lin, "Refining dataset curation methods for deep learning-based automated tuberculosis screening," *Journal of Thoracic Disease*, vol. 12, no. 9, p. 5078, 2020.
54. P. Rajpurkar, C. O'Connell, A. Schechter, N. Asnani, J. Li, A. Kiani, R. L. Ball, M. Mendelson, G. Maartens, D. J. van Hoving, *et al.*, "Chexaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with hiv," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–8, 2020.
55. D. Das, K. Santosh, and U. Pal, "Truncated inception net: Covid-19 outbreak screening using chest x-rays," *Physical and engineering sciences in medicine*, vol. 43, no. 3, pp. 915–925, 2020.
56. S. Sathitratanacheewin, P. Sunanta, and K. Pongpirul, "Deep learning for automated classification of tuberculosis-related chest x-ray: dataset distribution shift limits diagnostic performance generalizability," *Heliyon*, vol. 6, no. 8, p. e04614, 2020.
57. A. T. Sahlol, M. Abd Elaziz, A. Tariq Jamal, R. Damaševičius, and O. Farouk Hassan, "A novel method for detection of tuberculosis in chest radiographs using artificial ecosystem-based optimisation of deep neural network features," *Symmetry*, vol. 12, no. 7, p. 1146, 2020.
58. S. Rajaraman and S. K. Antani, "Modality-specific deep learning model ensembles toward improving tb detection in chest radiographs," *IEEE Access*, vol. 8, pp. 27318–27326, 2020.
59. T. Rahman, A. Khandakar, M. A. Kadir, K. R. Islam, K. F. Islam, R. Mazhar, T. Hamid, M. T. Islam, S. Kashem, Z. B. Mahbub, M. A. Ayari, and M. E. H. Chowdhury, "Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization," *IEEE Access*, vol. 8, pp. 191586–191601, 2020.
60. R. Guo, K. Passi, and C. K. Jain, "Tuberculosis diagnostics and localization in chest x-rays via deep learning models," *Frontiers in Artificial Intelligence*, vol. 3, p. 74, 2020.
61. Z. U. Abideen, M. Ghafoor, K. Munir, M. Saqib, A. Ullah, T. Zia, S. A. Tariq, G. Ahmed, and A. Zahra, "Uncertainty assisted robust tuberculosis identification with bayesian convolutional neural networks," *Ieee Access*, vol. 8, pp. 22812–22825, 2020.
62. K. Murphy, S. S. Habib, S. M. A. Zaidi, S. Khowaja, A. Khan, J. Melendez, E. T. Scholten, F. Amad, S. Schalekamp, M. Verhagen, *et al.*, "Computer aided detection of tuberculosis on chest radiographs: An evaluation of the cad4tb v6 system," *Scientific reports*, vol. 10, no. 1, pp. 1–11, 2020.
63. S. Rajaraman, I. Kim, and S. K. Antani, "Detection and visualization of abnormality in chest radiographs using modality-specific convolutional neural network ensembles," *PeerJ*, vol. 8, p. e8693, 2020.
64. M. Ayaz, F. Shaukat, and G. Raja, "Ensemble learning based automatic detection of tuberculosis in chest x-ray images using hybrid feature descriptors," *Physical and Engineering Sciences in Medicine*, vol. 44, no. 1, pp. 183–194, 2021.
65. S. Rajaraman, G. Zamzmi, L. Folio, P. Alderson, and S. Antani, "Chest x-ray bone suppression for improving classification of tuberculosis-consistent findings," *Diagnostics*, vol. 11, no. 5, p. 840, 2021.
66. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.
67. S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest x-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, 2014.
68. C. Qin, D. Yao, Y. Shi, and Z. Song, "Computer-aided detection in chest radiography based on artificial intelligence: a survey," *Biomedical engineering online*, vol. 17, no. 1, pp. 1–23, 2018.
69. P. Ganesan, S. Rajaraman, R. Long, B. Ghoraani, and S. Antani, "Assessment of data augmentation strategies toward performance improvement of abnormality classification in chest radiographs," in *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 841–844, IEEE, 2019.
70. S. Rajaraman and S. Antani, "Weakly labeled data augmentation for deep learning: a study on covid-19 detection in chest x-rays," *Diagnostics*, vol. 10, no. 6, p. 358, 2020.
71. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), (Cham), pp. 818–833, Springer International Publishing, 2014.
72. A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5188–5196, 2015.
73. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929, 2016.
74. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.
75. M. T. Ribeiro, S. Singh, and C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier," in

Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, (New York, NY, USA), p. 1135–1144, Association for Computing Machinery, 2016.
76. S. Rajaraman, S. Candemir, G. Thoma, and S. Antani, "Visualizing and explaining deep learning predictions for pneumonia detection in pediatric chest radiographs," in *Medical Imaging 2019: Computer-Aided Diagnosis*, vol. 10950, p. 109500S, International Society for Optics and Photonics, 2019.
77. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "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)*, pp. 3462–3471, 2017.
78. S. Rajaraman, S. Candemir, I. Kim, G. Thoma, and S. Antani, "Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs," *Applied Sciences*, vol. 8, no. 10, p. 1715, 2018.
79. I. Kim, S. Rajaraman, and S. Antani, "Visual interpretation of convolutional neural network predictions in classifying medical image modalities," *Diagnostics*, vol. 9, no. 2, p. 38, 2019.
80. S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, L. R. Folio, and S. K. Antani, "Iteratively pruned deep learning ensembles for covid-19 detection in chest x-rays," *IEEE Access*, vol. 8, pp. 115041–115050, 2020.
81. S. Rajaraman, S. Sornapudi, P. O. Alderson, L. R. Folio, and S. K. Antani, "Analyzing inter-reader variability affecting deep ensemble learning for covid-19 detection in chest radiographs," *PLOS ONE*, vol. 15, pp. 1–32, 11 2020.
82. S. Rajaraman, G. Cohen, L. Spear, L. Folio, and S. Antani, "Debonet: A deep bone suppression model ensemble to improve disease detection in chest radiographs," *PLOS ONE*, vol. 17, pp. 1–22, 03 2022.