Rethinking Computer-aided Tuberculosis Diagnosis

Yun Liu[1*] Yu-Huan Wu[1*] Yunfeng Ban[2] Huifang Wang[2] Ming-Ming Cheng[1]✉
[1]TKLNDST, College of Computer Science, Nankai University [2]InferVision
http://mmcheng.net/tb/

Abstract

As a serious infectious disease, tuberculosis (TB) is one of the major threats to human health worldwide, leading to millions of deaths every year. Although early diagnosis and treatment can greatly improve the chances of survival, it remains a major challenge, especially in developing countries. Computer-aided tuberculosis diagnosis (CTD) is a promising choice for TB diagnosis due to the great successes of deep learning. However, when it comes to TB diagnosis, the lack of training data has hampered the progress of CTD. To solve this problem, we establish a large-scale TB dataset, namely Tuberculosis X-ray (TBX11K) dataset. This dataset contains 11200 X-ray images with corresponding bounding box annotations for TB areas, while the existing largest public TB dataset only has 662 X-ray images with corresponding image-level annotations. The proposed dataset enables the training of sophisticated detectors for high-quality CTD. We reform the existing object detectors to adapt them to simultaneous image classification and TB area detection. These reformed detectors are trained and evaluated on the proposed TBX11K dataset and served as the baselines for future research.

1. Introduction

As the second leading cause of death by infectious disease (after HIV), tuberculosis (TB) is one of the major global health threats [33, 34]. Every year, there are about 8,000,000 - 10,000,000 new TB patients, and about 2,000,000 - 3,000,000 people died of TB [34]. TB is induced by Mycobacterium TB, which can be spread by sneezing, coughing or other means of excreting infectious bacteria. Hence TB typically occurs in the lungs through the respiratory tract. Opportunistic infections in immunocompromised people such as HIV patients and malnourished persons in developing countries have exacerbated this problem.

If not treated, the mortality rate of TB is very high, but diagnosing TB in the early stage and imposing treatment with antibiotics greatly improves the chances of survival [6,17,19]. Early diagnosis of TB also helps control the spread of infection [6]. The increase in multidrug-resistant TB also leads to the urgent need for a timely and accurate method of TB diagnosis to track the process of clinical treatment [11]. Unfortunately, TB diagnosis is still a major challenge [1,2,5,6,17,19,31]. The *golden standard* for TB diagnosis is microscopic examination of sputum and culture of bacteria for the identification of Mycobacterium TB [1, 2]. Therefore, biosafety level-3 lab (BSL-3) is needed for the culture of Mycobacterium TB. It requires *several months* for this process [1, 2, 19]. What's worse, hospitals in many developing countries and resource-constrained communities *cannot afford* such conditions.

On the other hand, X-ray is the most common and data-intensive screening method in current medical image examination, and X-ray is also one of most commonly used way for TB screening. Early TB screening through X-ray has great significance for the early detection, treatment, and prevention of TB [5, 19, 21, 35, 40]. However, the results of radiologists' examination of X-rays often go wrong [21,35], because it is often difficult for human eye to distinguish TB areas from X-rays where human eye is not sensitive enough to many details. In our human study, experienced radiologists from top hospitals only have *an accuracy of 68.7%* when compared with the golden standard.

Motivation and Contributions Thanks to the powerful representation ability of deep learning, especially convolutional neural networks (CNNs) [12, 15, 37], machines have outperformed human in many fields, such as face recognition [38], image classification [14], object detection [13], and edge detection [29]. Deep learning can capture details [16, 28, 29] and never feel tired like people. It is a natural idea to adopt deep learning for computer-aided TB diagnosis/screening with X-ray images. However, deep learning is always data-hungry, and it is difficult to collect large-scale TB data because they are very expensive and private. The lack of publicly available X-rays has prevented **computer-aided tuberculosis diagnosis** (**CTD**) from successfully ap-

*Equal contribution.

2575-7075/20/$31.00 ©2020 IEEE
DOI 10.1109/CVPR42600.2020.00272

plying deep learning for improving performance. For example, the existing largest public X-ray dataset for TB diagnosis is Shenzhen chest X-ray set proposed in [18]. The Shenzhen dataset consists of 662 X-ray images, including 336 X-rays with manifestations of TB and 326 normal X-rays, with only binary image-level labels. Just using these several hundreds of images is insufficient to train deep CNNs. Therefore, many state-of-the-art CTD methods only adopt handcrafted features [5, 19, 20] or pretrained CNNs as feature extractors without fine-tuning [31], while ignoring the powerful ability of automatic feature learning of deep CNNs.

In order to actually deploy the CTD system to help TB patients around the world, we must first solve the problem of insufficient data. In this paper, we contribute to the community with a large-scale **Tuberculosis X-ray** (**TBX11K**) dataset, through the long-term cooperation with major hospitals. This new dataset is superior to previous CTD datasets in the following aspects: i) Unlike previous datasets [6, 18] that only contain several tens/hundreds of X-ray images, TBX11K has 11,200 images that are about $17\times$ larger than the existing largest dataset, *i.e*., Shenzhen dataset [18], so that TBX11K makes it possible to train very deep CNNs; ii) Instead of only having image-level annotations as previous datasets, TBX11K annotates TB areas using bounding boxes, so that the future CTD methods can not only recognize the manifestations of TB but also detect the TB areas to help radiologists for the definitive diagnosis; iii) TBX11K includes four categories of healthy, active TB, latent TB, and unhealthy but non-TB, rather than the binary classification for TB or not in previous datasets, so that future CTD systems can adapt to more complex real-world scenarios and provide people with more detailed disease analyses. Each X-ray image in TBX11K is tested **using the golden standard** (*i.e*., diagnostic microbiology) and then annotated by experienced radiologists from major hospitals. TBX11K dataset has been *de-identified* by the data providers and *exempted* by relevant institutions, so it can be made publicly available to promote future CTD research.

Moreover, we reform the existing object detectors, including SSD [27], RetinaNet [25], Faster R-CNN [36], and FCOS [39], for simultaneous image classification and TB area detection. Specifically, we introduce a classification branch onto these detectors and propose an alternative training strategy. These reformed methods can be viewed as baselines for future CTD research. We also adapt the metrics for image classification and object detection to TB diagnosis on the proposed TBX11K dataset. The baselines are evaluated in terms of these metrics to build the initial benchmarks.

In summary, our contributions are twofold:

- We build a large-scale CTD dataset that is much larger, better annotated, and more realistic than existing TB datasets, enabling the training of deep CNNs.
- We establish the CTD benchmark by i) reforming the existing object detectors for CTD; ii) adapting classification and detection metrics to CTD, which is expected to set a good start for future CTD research.

Datasets	Year	Class	Label	Sample
MC [18]	2014	2	Image-level	138
Shenzhen [18]	2014	2	Image-level	662
DA [6]	2014	2	Image-level	156
DB [6]	2014	2	Image-level	150
TBX11K	2020	4	Bounding box	11200

Table 1. Summary of publicly available TB datasets.

2. Related Work

2.1. Tuberculosis Datasets

Since TB data is very private and it is difficult to diagnose TB with golden standard, the publicly available TB datasets are very limited. We provide a summary for the publicly available TB datasets in Tab. 1. Jaeger *et al*. [18] proposed two chest X-ray datasets for TB diagnosis. The Montgomery County chest X-ray set (MC) [18] is collected through the cooperation with Department of Health and Human Services, Montgomery County, Maryland, USA. MC dataset consists of 138 X-ray images, 80 of which are healthy cases and 58 are cases with manifestations of TB. Shenzhen chest X-ray set (Shenzhen) [18] is collected through the cooperation with Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China. Shenzhen dataset is composed of 326 norm cases and 336 cases with manifestations of TB, leading to 662 X-ray images in total. Chauhan *et al*. [6] proposed two datasets, namely DA and DB, which are obtained from two different X-ray machines at the National Institute of Tuberculosis and Respiratory Diseases, New Delhi. DA is composed of training set (52 TB and 52 non-TB X-rays) and the independent test set (26 TB and 26 non-TB X-rays). DB contains 100 training X-rays (50 TB and 50 non-TB) and 50 test X-rays (25 TB and 25 non-TB). Note that all these four datasets are annotated with image-level labels for binary image classification.

These datasets are too small to train deep neural networks, so recent research on CTD has been hindered although CNNs have achieved numerous successful stories in the computer vision community. On the other hand, the existing datasets only have image-level annotations, and thus we cannot train TB detectors with previous data. To help radiologists make accurate judgements, we are expected to detect the TB areas, not only an image-level classification. Therefore, the lack of TB data has prevented deep learning from bringing success to practical CTD systems that have potential to save millions of TB patients every year. In this

paper, we build a large-scale dataset with bounding box annotations for training TB detectors. The presentation of this new dataset is expected to promote the future research for CTD and promote more practical CTD systems.

2.2. Computer-aided Tuberculosis Diagnosis

Owing to the lack of data, traditional CTD methods cannot train deep CNNs. Most traditional methods mainly use hand-crafted features and train binary classifiers. Jaeger *et al*. [19] first segmented the lung region using a graph cut segmentation method [4]. Then, they extracted hand-crafted texture and shape features from this lung region. Finally, they apply a binary classifier, *i.e*., support vector machine (SVM), to classify X-rays as normal and abnormal. Candemir *et al*. [5] adopted image retrieval-based patient specific adaptive lung models to a nonrigid registration-driven robust lung segmentation method, which would be helpful for traditional lung feature extraction [19]. Chauhan *et al*. [6] implemented a MATLAB toolbox, TB-Xpredict, which adopted Gist [32] and PHOG [3] features for the discrimination between TB and non-TB X-rays without requiring segmentation [8, 30]. Karargyris *et al*. [20] extracted shape features to describe the overall geometrical characteristics of lungs and texture features to represent image characteristics.

Instead of using hand-crafted features, Lopes *et al*. [31] adopted the fixed CNNs pretrained on ImageNet [10] as the feature extractors to compute deep features for X-ray images. Then, they train SVM to classify these deep features. Hwang *et al*. [17] trained an AlexNet [22] for binary classification (TB and non-TB) using a private dataset. Other private datasets are also used in [23] for image classification networks. However, our proposed dataset, *i.e*., TBX11K, will be made publicly available to promote research in this field.

3. The Tuberculosis X-ray (TBX11K) Dataset

3.1. Data Collection and Annotation

For the data collection and annotation, we follow three steps: i) taxonomy establishment, ii) X-ray Collection, iii) professional data annotation, which are introduced below.

3.1.1 Taxonomy Establishment

The existing TB datasets only contain two categories: TB and non-TB. The non-TB refers to healthy cases. In practice, the abnormalities in chest X-rays, such as TB, atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, *etc*., share similar abnormal patterns (*e.g*., blurry and irregular lesions), while being significantly different from healthy X-rays that have almost the same clear patterns. Therefore, only using healthy X-rays as the negative category has the bias to lead to large false positives in the model prediction for clinical scenarios where there are many sick but non-TB patients. To promote CTD to practical applications, we consider a new category, sick but non-TB, in our dataset. Moreover, besides the recognition of TB, it is also very important to differentiate the active TB and latent TB. Active TB is caused by Mycobacterium TB infection or as a reactivation of latent TB, while people with latent TB are neither sick nor contagious. The differentiation between active TB and latent TB can help doctors provide patients with proper treatment. Considering this, we divide the TB into two categories of active TB and latent TB in our dataset. With the above analyses, we include four categories in the proposed TBX11K dataset: healthy, sick but non-TB, active TB, and latent TB.

3.1.2 X-ray Collection

The collection of TB X-rays faces two difficulties: i) The chest X-rays, especially TB X-rays, are of high privacy and leaking these data will expose people to risk of breaking the law, so that it is almost impossible for individuals to access the raw data; ii) Although there are millions of TB patients worldwide, the TB X-rays that are definitively tested by the golden standard are scarce, due to the complex and lengthy (*i.e*., several months [1, 2]) process of examination of Mycobacterium TB. In order to overcome these difficulties, we cooperate with top hospitals to collect X-rays. The resulting TBX11K dataset consists of 11200 X-rays, including 5000 healthy cases, 5000 sick but non-TB cases, and 1200 cases with manifestations of TB. Here, each X-ray belongs to a unique person. The 1200 TB X-rays are composed of 924 active TB cases, 212 latent TB cases, 54 cases that contain active and latent TB simultaneously, and 10 uncertain cases whose TB types cannot be recognized under today's medical conditions. The 5000 sick but non-TB cases are collected to cover as many of the types of radiograph diseases as possible in the clinical scenarios. All X-rays are in the resolution of about 3000×3000. We also include the corresponding sex and age for each X-ray to provide more comprehensive clinical information for TB diagnosis. The data have been de-identified by data providers and exempted by relevant government institutions, so we can make this dataset publicly available legally.

3.1.3 Professional Data Annotation

Every X-ray image in our dataset has been definitively tested using the golden standard, but the golden standard can only provide the image-level labels. For example, if the sputum of one patient has manifestations of TB, we would know the corresponding X-ray falls into the category of TB, but we do not know the exact location and area of TB in this

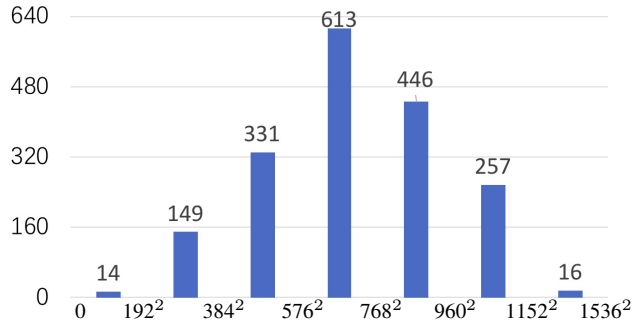

Figure 1. Distribution of the areas of TB bounding boxes. The left and right values of each bin define its corresponding area range, and the height of each bin denotes the number of TB bounding boxes with an area within this range. Note that X-rays are in the resolution of about 3000×3000.

X-ray. On the other hand, detecting TB areas is of vital importance to help radiologists make the final decision. With only image-level predictions, it is still hard for human eye to find the TB areas, which can be proved by the low accuracy of radiologists in the clinical examination as shown in Sec. 3.3. If CTD systems could simultaneously provide image classification and TB localization results, radiologists will make decisions more accurately and efficiently by watching the detected TB areas.

In order to achieve the above objectives, we provide bounding box annotations for TB X-rays in TBX11K dataset. To the best of our knowledge, this is the first dataset for TB detection. The bounding box annotations are performed by experienced radiologists from top hospitals. Specifically, each TB X-ray is first labeled by a radiologist who has 5-10 years of experience in TB diagnosis. Then, his box annotations are further checked by another radiologist who has $>$10 years of experience in TB diagnosis. They not only label bounding boxes for TB areas but also recognize the TB type (active or latent TB) for each box. The labeled TB types are double-checked to make sure that they are consistent with the image-level labels produced by the golden standard. If a mismatch happens, this X-ray will be put into the unlabeled data for re-annotation, and the annotators do not know which X-ray was labeled wrong before. If an X-ray is incorrectly labeled twice, we will tell the annotators the gold standard of this X-ray and ask them to discuss how to re-annotate it. This double-checked process makes the annotated bounding boxes are highly reliable for TB area detection. Moreover, non-TB X-rays are only with image-level labels produced by the golden standard. We show some examples of the proposed TBX11K dataset in Fig. 3. We display the distribution of the areas of TB bounding boxes in Fig. 1. The areas of most TB bounding boxes are in the range of $(384^2, 960^2]$.

	Class	Train	Val	Test	Total
Non-TB	Healthy	3000	800	1200	5000
	Sick & Non-TB	3000	800	1200	5000
TB	Active TB	473	157	294	924
	Latent TB	104	36	72	212
	Active & Latent TB	23	7	24	54
	Uncertain TB	0	0	10	10
Total		6600	1800	2800	11200

Table 2. Split for the proposed TBX11K dataset. "Active & Latent TB" refers to X-rays that contain active and latent TB simultaneously. "Active TB" and "Latent TB" refers to X-rays that only contain active TB or latent TB, respectively. "Uncertain TB" refers to TB X-rays whose TB types cannot be recognized under today's medical conditions.

3.2. Dataset Split

We divide the data into three subsets for training, validation, and testing, respectively. The details of our split are summarized in Tab. 2. For more representative, we consider four different cases of TB: i) only active TB appears; ii) only latent TB appears; iii) both active and latent TB appear in an X-ray; iv) uncertain TB whose TB type cannot be recognized. For various TB cases, the ratio of the number of TB X-rays for training, validation, and test sets is $3 : 1 : 2$. Note that the uncertain TB X-rays are put into the test set, and researchers can perform the evaluation for class-agnostic TB detection using these 10 uncertain X-rays. Consistent with the scientific experiment settings, we suggest researchers training their models on the training set and evaluating on the validation set when they tune hyper-parameters. Once their models are fixed, they can retrain using the union of the training set and validation set, and then report the performance of their models on the test set.

3.3. Human Study by Radiologists

The human study of radiologists is important for us to understand the role of CTD in clinical TB diagnosis. We first randomly choose 400 X-rays from the test set of the proposed TBX11K dataset, including 140 healthy X-rays, 140 sick but non-TB X-rays, and 120 TB X-rays. The 120 TB X-rays consist of 63 active TB, 41 latent TB, 15 active & latent TB, and 1 uncertain TB. Then, we invite an experienced radiologist with $>$10 years of work experience from a major hospital to label these X-rays with image-level labels from four categories, *i.e.*, health, sick but non-TB, active TB, and latent TB. He assigned both active TB and latent TB to an X-ray if both active TB and latent TB appear. Note that this radiologist is different from radiologists who labeled our dataset.

Compared with the ground truth produced by the golden standard, the radiologist only achieves an accuracy of 68.7%. If ignoring the differentiation between the active TB

and latent TB, the accuracy is 84.8%, but the recognition of TB types is important for clinical treatment. This low performance is one of the major challenges in TB diagnosis, treatment, and prevention. Different from natural color images, chest X-rays are in grayscale and usually have fuzzy and blurry patterns, which causes significant difficulty for recognition. However, the TB diagnosis with golden standard takes several months [1, 2] and there is no such condition in many parts of the world. The challenge in TB diagnosis is one of the main reasons why TB becomes the second leading infectious disease worldwide (after HIV). In the following study, we will show that deep learning CTD methods trained on the proposed TBX11K dataset can significantly outperform the experienced radiologists.

3.4. Potential Research Topics

With the proposed TBX11K dataset, we can conduct research on X-ray image classification and TB area detection. With many health and sick but non-TB data, our test set can simulate the clinical data distribution to evaluate the CTD systems. We think the simultaneous X-ray image classification and TB area detection systems would be a challenging and interesting research topic. It is convenient to deploy such systems to help radiologists for TB diagnosis.

Besides the simultaneous detection and image classification, another challenge of our dataset is the imbalanced data distribution across different categories. However, this data imbalance is in line with actual clinical scenarios. Intuitively, when a person goes to the hospital for a chest examination, he is likely to feel uncomfortable, so the probability of getting sick is higher than usual, but TB is only one of many chest diseases. In the proposed TBX11K dataset, we assume only 44.6% examination takers are healthy, 44.6% are sick but non-TB, and only 10.7% takers are infected by TB. The latent TB can be caused by two ways: i) exposure to active TB and ii) conversion from active TB after treatment. Most people with latent TB are caused by the first way. The people with latent TB in the hospital are usually the above second case, because people with latent TB are neither sick nor contagious and they are unlikely to go to the hospital for examination. Therefore, our dataset has much more active TB cases than latent TB. Therefore, future CTD methods should be designed to overcome the data imbalance problem in practice, *e.g.*, how to train models on our imbalanced TBX11K training set.

4. Experimental Setup

In this section, we first build some baselines for the simultaneous X-ray image classification and TB area detection. Then, we elaborate on the evaluation metrics.

4.1. Baselines

Existing object detectors do not consider the background images. More specifically, they usually ignore these images that have no bounding-box objects [9, 25, 27, 36, 39, 41]. Directly applying existing object detectors into CTD task will lead to many false positives, because of the large number of non-TB X-rays in practice. To solve this problem, we propose to conduct simultaneous X-ray image classification and TB area detection, so that the image classification results can filter out the false positives of detection.

We reform several well-known object detectors, including SSD [27], RetinaNet [25], Faster R-CNN [36], and FCOS [39], for simultaneous X-ray classification and TB area detection. The image classification branch learns to classify X-rays into **three categories**, *i.e.*, healthy, sick but non-TB, and TB using a *Softmax* function. The TB detection branch learns to detect TB with **two categories**, *i.e.*, active TB and latent TB. In the clinical diagnosis, the image classification results can help radiologists judge whether TB appears in an X-ray. Then, the TB detection results provide radiologists with TB areas that help radiologists make the final decision. With the above definitions, we add an image classification branch to the existing object detectors after the final convolution layer of their backbone networks, *i.e.*, *conv5_3* and *res5c* for VGG16 [37] and ResNet-50 [15], respectively. For the classification branch, we use five sequential convolution layers, each of which has 512 output channels and ReLU activation. The first convolution layer has a stride of 2 only for SSD and 1 for other methods. A max pooling layer with a stride of 2 is connected after the third convolution layer. After these convolutions, we use a global average pooling layer and a fully connected layer with 3 output neurons for classification into 3 classes. Since we focus on providing some workable baselines for the analyses of the proposed dataset, we do not make careful parameter tuning.

We introduce a two-stage strategy to train such networks. First, we omit the image classification branch and train the object detector with default settings. Then, we freeze the backbone network and object detection branch, and only train the image classification branch to adapt the detection features for image classification. The first-stage training only uses the TB X-rays in the TBX11K *trainval* (train + validation) set. The second-stage training not only uses all (*i.e.*, TB and Non-TB) TBX11K *trainval* X-rays but also the random half of the MC [18] and Shenzhen [18] datasets as well as the training sets of the DA [6] and DB [6] datasets. The other half of the MC [18] and Shenzhen [18] datasets as well as the test sets of the TBX11K, DA [6] and DB [6] datasets are used to evaluate the performance of image classification. TB detection evaluation has two modes: i) using all (*i.e.*, TB and Non-TB) TBX11K *test* X-rays and ii) using only the TB X-rays in the TBX11K *test* set.

All experiments are based on the open-source mmdetection toolbox [7] with 4 RTX 2080Ti GPUs. The batch size is always 16. The first-stage training runs 38400 and 76800 iterations for training with ImageNet pretraining [10] and training from scratch, respectively. The initial learning rate is 0.005 except that SSD [27] uses 0.0005. The learning rate is divided by 10 after 25600 and 32000 iterations when with ImageNet pretraining or after 51200 and 64000 iterations when training from scratch. We train the second stage for 24 epochs with the initial learning of $1e$-3 that is divided by 10 after 12 and 18 epochs. The X-ray images are resized to 512×512 when inputting to networks.

4.2. Evaluation Metrics

In this section, we introduce the metrics for the evaluation of CTD task. For X-ray classification, CTD aims at classifying each X-ray into three categories, which is evaluated using six metrics:

- Accuracy that measures the percentage of X-rays that are correctly classified as one of the three classes;
- Area Under Curve (AUC) that computes the area under the Receiver Operating Characteristic (ROC) curve that plots the true positive rate against the false positive rate for TB class;
- Sensitivity that measures the percentage of TB cases that are correctly identified as TB, *i.e.*, the recall for TB class;
- Specificity that measures the percentage of non-TB cases that are correctly identified as non-TB, *i.e.*, the recall for non-TB class, where non-TB includes healthy and sick but non-TB classes;
- Average Precision (AP) that computes the precision of each class and takes the average across all classes;
- Average Recall (AR) that computes the recall of each class and averages over all classes.

For the evaluation of TB detection, we adopt the average precision of bounding box (AP^{bb}) proposed by the COCO dataset [26]. The default AP^{bb} refers to the AP^{bb} averaged over IoU (intersection-over-union) thresholds of $[0.5 : 0.05 : 0.95]$. AP^{bb}_{50} refers to AP^{bb} at the IoU thresholds of 0.5. In order to facilitate the observation of the detection of each TB type, we report the evaluation results for active TB and latent TB separately. Here, the uncertain TB X-rays are ignored. We also report category-agnostic TB detection results, where the TB categories are ignored, to describe the detection for all TB areas. Here, the uncertain TB X-rays are included. Moreover, we introduce two evaluation modes by using i) all test X-rays or ii) only TB X-rays in the test set. With these metrics, we can analyze the performance of CTD systems in more useful aspects.

5. Experiments and Analyses

5.1. Image Classification

We summarize the evaluation results for image classification in Tab. 3. When ImageNet pretraining [10] is enabled, Faster R-CNN [36] achieves the best performance. When there is no ImageNet pretraining [10], SSD [27] performs best. We also observe that ImageNet pretraining can significantly improve performance except for SSD that achieves better performance without pretraining. Maybe this is because SSD adopts a shallower backbone, VGGNet-16 [37], which is easier to be trained than ResNet-50 [15] with FPN [24] used by other methods. Note that FCOS [39] crashes if without ImageNet pretraining, so we do not include the results of this setting in Tab. 3.

Faster R-CNN [36] achieves a high sensitivity rate of 91.2%, which suggests deep learning can recognize most of the TB X-rays. The specificity is 89.9%, which means 10.1 of 100 non-TB X-rays will be classified as TB. Moreover, in terms of accuracy, all methods (*i.e.* SSD without pretraining and ResNet-50 based methods with pretraining) can outperform radiologists as shown in Sec. 3.3, so deep learning based CTD is a promising research field. The future progress in this direction has the potential to promote practical CTD systems to help millions of TB patients.

5.2. TB Area Detection

Here, we report not only the performance on the whole TBX11K test set but also the performance evaluated using only TB X-rays in the test set. Since there is no target TB areas in non-TB X-rays, evaluating using only TB X-rays could provide precise detection analyses, while evaluating using all X-rays includes the influence of false positives in non-TB X-rays. When evaluating using all X-rays, we abandon all predicted boxes in those X-rays that are predicted as non-TB by the image classification branch. This filtering process is useless for evaluating using only TB X-rays. Since the training of image classification for FCOS [39] without ImageNet pretraining crashes, we do not report the evaluation results using all X-rays in this case.

Except for SSD [27], ImageNet pretraining [10] can improve the detection performance. SSD achieves similar performance with or without pretraining. SSD achieves the overall best performances in most cases except for the evaluation using all X-rays with ImageNet pretraining where Faster R-CNN [36] achieves the best performance. All methods fail to accurately detect latent TB areas, but the evaluation results of category-agnostic TB are better than that of active TB, which means many latent TB targets are correctly located but wrongly classified as active TB. We guess that this is caused by the limited number of latent TB X-rays in TBX11K where there are only 212 latent TB X-rays but 924 active TB X-rays. Therefore, future research

Method	Pretrained	Backbone	Accuracy	AUC (TB)	Sensitivity	Specificity	Ave. Prec.	Ave. Rec.
SSD [27]	Yes	VGGNet-16	84.7	93.0	78.1	89.4	82.1	83.8
RetinaNet [25]		ResNet-50 w/ FPN	87.4	91.8	81.6	89.8	84.8	86.8
Faster R-CNN [36]		ResNet-50 w/ FPN	**89.7**	**93.6**	**91.2**	**89.9**	**87.7**	**90.5**
FCOS [39]		ResNet-50 w/ FPN	88.9	92.4	87.3	**89.9**	86.6	89.2
SSD [27]	No	VGGNet-16	**88.2**	**93.8**	**88.4**	89.5	**86.0**	**88.6**
RetinaNet [25]		ResNet-50 w/ FPN	79.0	87.4	60.0	**90.7**	75.9	75.8
Faster R-CNN [36]		ResNet-50 w/ FPN	81.3	89.7	72.5	87.3	78.5	79.9

Table 3. X-ray image classification results on the proposed TBX11K test data. "Pretrained" indicates whether to pretrain the backbone networks on ImageNet [10]. "Backbone" refers to the backbone networks of each baseline, where FPN denotes the feature pyramid network [24] for object detection.

Method	Data	Pretrained	Backbone	CA TB		Active TB		Latent TB	
				AP^{bb}_{50}	AP^{bb}	AP^{bb}_{50}	AP^{bb}	AP^{bb}_{50}	AP^{bb}
SSD [27]	ALL	Yes	VGGNet-16	52.3	22.6	50.5	**22.8**	8.1	**3.2**
RetinaNet [25]			ResNet-50 w/ FPN	52.1	22.2	45.4	19.6	6.2	2.4
Faster R-CNN [36]			ResNet-50 w/ FPN	**57.3**	**22.7**	**53.3**	21.9	**9.6**	2.9
FCOS [39]			ResNet-50 w/ FPN	46.6	18.9	40.3	16.8	6.2	2.1
SSD [27]		No	VGGNet-16	**61.5**	**26.1**	**60.0**	**26.2**	**8.2**	**2.9**
RetinaNet [25]			ResNet-50 w/ FPN	20.7	7.2	19.1	6.4	1.6	0.6
Faster R-CNN [36]			ResNet-50 w/ FPN	21.9	7.4	21.2	7.1	2.7	0.8
SSD [27]	TB	Yes	VGGNet-16	68.3	**28.7**	**63.7**	**28.0**	**10.7**	4.0
RetinaNet [25]			ResNet-50 w/ FPN	**69.4**	28.3	61.5	25.3	10.2	**4.1**
Faster R-CNN [36]			ResNet-50 w/ FPN	63.4	24.6	58.7	23.7	9.6	2.8
FCOS [39]			ResNet-50 w/ FPN	56.3	22.5	47.9	19.8	7.4	2.4
SSD [27]		No	VGGNet-16	**69.6**	**29.1**	**67.0**	**29.0**	**9.9**	**3.5**
RetinaNet [25]			ResNet-50 w/ FPN	40.5	13.8	37.8	12.7	3.2	1.1
Faster R-CNN [36]			ResNet-50 w/ FPN	37.4	11.8	35.3	11.3	3.9	1.1
FCOS [39]			ResNet-50 w/ FPN	42.1	14.4	38.5	13.6	4.3	1.1

Table 4. TB area detection results on the proposed TBX11K test set. "Data" indicates whether to use all test X-rays for evaluation or only TB X-rays in the test set. "Pretrained" and "Backbone" refer to the same meaning as in Tab. 3. "CA TB" denotes class-agnostic TB.

should pay more attention to this data imbalance problem. We also find that the performance in terms of AP^{bb}_{50} is usually much better than that in terms of AP^{bb}. This means that although the detection could find the target region, the localization is usually not very accurate. We argue that locating TB bounding box regions is quite different from locating nature object regions. Even experienced radiologists cannot easily locate the precise TB regions. Therefore, AP^{bb}_{50} is more important than AP^{bb} because the predicted boxes having an IoU of 0.5 with TB targets are enough for helping radiologists to find TB areas.

In Fig. 2, we plot PR curves for detection error analyses. All methods are tested with ImageNet pretraining for category-agnostic TB detection. We can clearly see that all methods have very large improvement from IoU threshold 0.75 to 0.5. This shows that these methods struggle at high IoU thresholds due to its poor object localization. The region of "FN" when using all X-rays is larger than that when using only TB X-rays, indicating that the filtering of image classification ignores many correctly detected TB areas, but we claim that this filtering is useful for improving overall detection performance. When evaluating using all X-rays, Faster R-CNN [36] achieves the state-of-the-art performance. When evaluating using only TB X-rays, RetinaNet [25] seems to achieve better performance. Combined image classification and TB area detection, we can conclude that these baselines show their strengths in different aspects.

5.3. Visualization

To understand what CNNs have learned from different X-rays, we visualize the feature map at the $1/32$ scale of the backbone of RetinaNet [25]. Specifically, we use principal component analysis (PCA) to reduce the channels of the feature map into a single channel. This single-channel map is converted to heat map for visualization. We display the results in Fig. 3. The visualization of healthy cases is irregular, while the visualization of sick but non-TB cases has some highlights, maybe the lesion. For TB cases, the

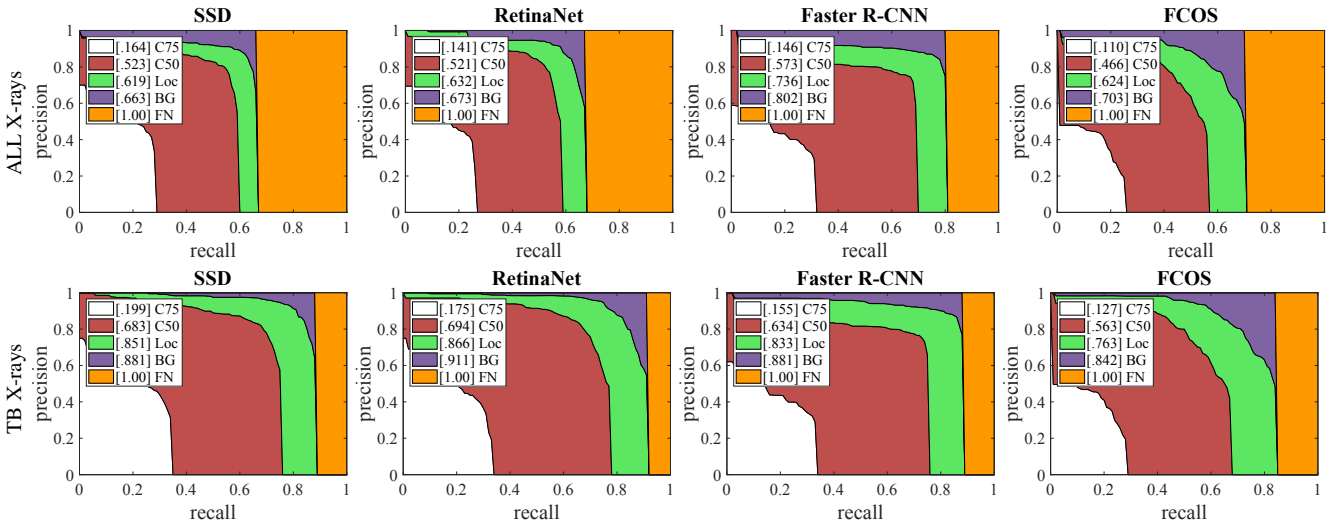

Figure 2. Error analyses of category-agnostic TB detection with ImageNet pretraining [10]. The first row is evaluated using all X-rays, while the second row only uses TB X-rays. C50/C75: PR curves under IoU thresholds of 0.5/0.75. Loc: the PR curve under the IoU of 0.1. BG: removing background false positives (FP). FN: removing other errors caused by undetected targets.

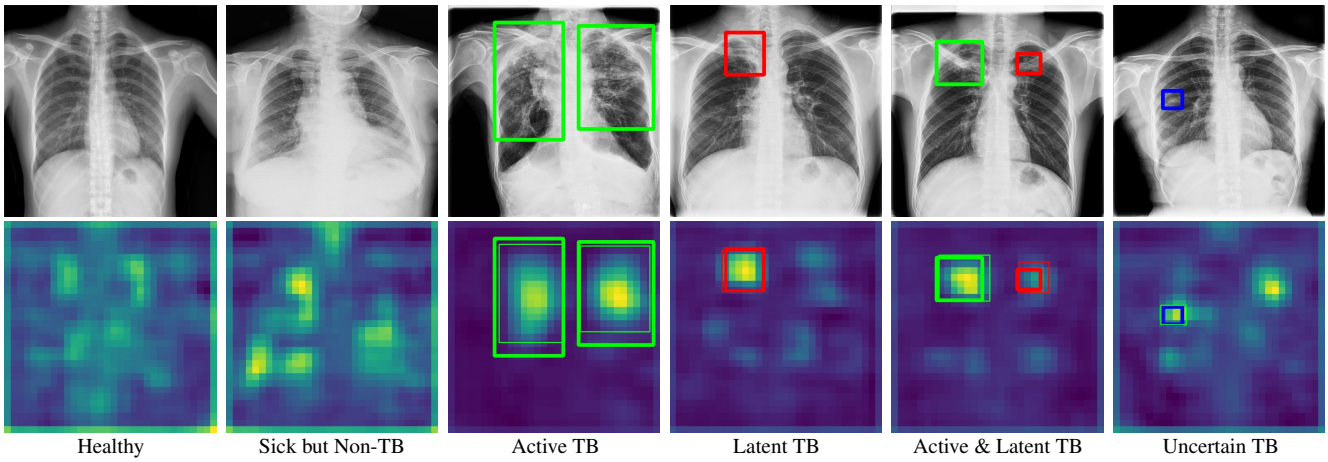

Figure 3. Visualization of the learned deep features from X-rays. All X-rays are randomly selected from the TBX11K test set. For each class listed in Tab. 2, we present one example. The green, red, and blue boxes cover the regions of active TB, latent TB, and uncertain TB, respectively. Boxes with thick and thin lines denote ground-truth boxes and detected boxes, respectively.

highlights in the visualization map is consistent with the annotated TB region.

6. Conclusion

Early diagnosis is important for the treatment and prevention of TB, a leading infectious disease. Unfortunately, TB diagnosis remains a major challenge. The definitive test of TB using the golden standard takes several months and is impossible in many developing countries and resource-constrained communities. Inspired by the successes of deep learning, deep learning based CTD is a promising research direction. However, the lack of data has prevented deep learning from bringing progress for CTD. In this paper, we build a large-scale TB dataset TBX11K with bounding-box annotations, enabling the training of deep CNNs for TB diagnosis. TBX11K is also the first dataset for TB detection. We build an initial benchmark for CTD by further proposing some baselines and evaluation metrics. This new TBX11K dataset and benchmark are expected to promote the research in CTD, and better CTD systems are expected to be designed with new powerful deep networks [12].

Acknowledgement. This research was supported by Major Project for New Generation of AI under Grant No. 2018AAA0100400, NSFC (61922046), the national youth talent support program, and Tianjin Natural Science Foundation (17JCJQJC43700, 18ZXZNGX00110).

References

[1] P Andersen, ME Munk, JM Pollock, and TM Doherty. Specific immune-based diagnosis of tuberculosis. *The Lancet*, 356(9235):1099–1104, 2000. 1, 3, 5

[2] Aliya Bekmurzayeva, Marzhan Sypabekova, and Damira Kanayeva. Tuberculosis diagnosis using immunodominant, secreted antigens of mycobacterium tuberculosis. *Tuberculosis*, 93(4):381–388, 2013. 1, 3, 5

[3] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *ACM International Conference on Image and Video Retrieval*, pages 401–408. ACM, 2007. 3

[4] Yuri Boykov and Gareth Funka-Lea. Graph cuts and efficient N-D image segmentation. *International Journal of Computer Vision*, 70(2):109–131, 2006. 3

[5] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE Transactions on Medical Imaging*, 33(2):577–590, 2013. 1, 2, 3

[6] Arun Chauhan, Devesh Chauhan, and Chittaranjan Rout. Role of gist and phog features in computer-aided diagnosis of tuberculosis without segmentation. *PloS One*, 9(11):e112980, 2014. 1, 2, 3, 5

[7] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6

[8] Ming-Ming Cheng, Yun Liu, Qibin Hou, Jiawang Bian, Philip Torr, Shi-Min Hu, and Zhuowen Tu. HFS: Hierarchical feature selection for efficient image segmentation. In *European Conference on Computer Vision*, pages 867–882, 2016. 3

[9] Ming-Ming Cheng, Yun Liu, Wen-Yan Lin, Ziming Zhang, Paul L Rosin, and Philip HS Torr. BING: Binarized normed gradients for objectness estimation at 300fps. *Computational Visual Media*, 5(1):3–20, 2019. 5

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 3, 6, 7, 8

[11] Neel R Gandhi, Paul Nunn, Keertan Dheda, H Simon Schaaf, Matteo Zignol, Dick Van Soolingen, Paul Jensen, and Jaime Bayona. Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. *The Lancet*, 375(9728):1830–1843, 2010. 1

[12] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2Net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 8

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision*, pages 2961–2969, 2017. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *IEEE International Conference on Computer Vision*, pages 1026–1034, 2015. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1, 5, 6

[16] Xiaowei Hu, Yun Liu, Kai Wang, and Bo Ren. Learning hybrid convolutional features for edge detection. *Neurocomputing*, 313:377–385, 2018. 1

[17] Sangheum Hwang, Hyo-Eun Kim, Jihoon Jeong, and Hee-Jin Kim. A novel approach for tuberculosis screening based on deep convolutional neural networks. In *Medical Imaging 2016: Computer-Aided Diagnosis*, volume 9785, page 97852W. International Society for Optics and Photonics, 2016. 1, 3

[18] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative Imaging in Medicine and Surgery*, 4(6):475–477, 2014. 2, 5

[19] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al. Automatic tuberculosis screening using chest radiographs. *IEEE Transactions on Medical Imaging*, 33(2):233–245, 2013. 1, 2, 3

[20] Alexandros Karargyris, Jenifer Siegelman, Dimitris Tzortzis, Stefan Jaeger, Sema Candemir, Zhiyun Xue, KC Santosh, Szilárd Vajda, Sameer Antani, Les 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*, 11(1):99–106, 2016. 2, 3

[21] Anastasios Konstantinos. Testing for tuberculosis. *Australian Prescriber*, 33(1):12–18, 2010. 1

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012. 3

[23] Paras Lakhani and Baskaran Sundaram. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582, 2017. 3

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 6, 7

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, pages 2980–2988, 2017. 2, 5, 7

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014. 6

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37, 2016. 2, 5, 6, 7

[28] Yun Liu, Ming-Ming Cheng, Deng-Ping Fan, Le Zhang, JiaWang Bian, and Dacheng Tao. Semantic edge detection with diverse deep supervision. *arXiv preprint arXiv:1804.02864*, 2018. 1

[29] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Jia-Wang Bian, Le Zhang, Xiang Bai, and Jinhui Tang. Richer convolutional features for edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1939–1946, 2019. 1

[30] Yun Liu, Peng-Tao Jiang, Vahan Petrosyan, Shi-Jie Li, Jiawang Bian, Le Zhang, and Ming-Ming Cheng. DEL: Deep embedding learning for efficient image segmentation. In *International Joint Conference on Artificial Intelligence*, pages 864–870, 2018. 3

[31] UK Lopes and João Francisco Valiati. Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. *Computers in Biology and Medicine*, 89:135–143, 2017. 1, 2, 3

[32] Aude Oliva and Antonio Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, 155:23–36, 2006. 3

[33] World Health Organization. Global tuberculosis report 2015. http://apps.who.int/iris/bitstream/10665/191102/1/9789241565059_eng.pdf, 2015. 1

[34] World Health Organization. Global tuberculosis report 2017. https://www.who.int/tb/publications/global_report/gtbr2017_main_text.pdf, 2017. 1

[35] World Health Organization et al. Chest radiography in tuberculosis detection: summary of current who recommendations and guidance on programmatic approaches. Technical report, World Health Organization, 2016. 1

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 2, 5, 6, 7

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 1, 5, 6

[38] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *Advances in Neural Information Processing Systems*, pages 1988–1996, 2014. 1

[39] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *IEEE International Conference on Computer Vision*, pages 9627–9636, 2019. 2, 5, 6, 7

[40] MRA Van Cleeff, LE Kivihya-Ndugga, H Meme, JA Odhiambo, and PR Klatser. The role and performance of chest x-ray for the diagnosis of tuberculosis: A cost-effectiveness analysis in nairobi, kenya. *BMC Infectious Diseases*, 5(1):111, 2005. 1

[41] Ziming Zhang, Yun Liu, Xi Chen, Yanjun Zhu, Ming-Ming Cheng, Venkatesh Saligrama, and Philip HS Torr. Sequential optimization for efficient high-quality object proposal generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1209–1223, 2018. 5