2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Identifying Drug-Resistant Tuberculosis in Chest Radiographs: Evaluation of CNN Architectures and Training Strategies

Manohar Karki[1], Karthik Kantipudi[2], Hang Yu[1], Feng Yang[1],
Yasmin M. Kassim[1], Ziv Yaniv[2] and Stefan Jaeger[1]

***Abstract*— Tuberculosis (TB) is a serious infectious disease that mainly affects the lungs. Drug resistance to the disease makes it more challenging to control. Early diagnosis of drug resistance can help with decision making resulting in appropriate and successful treatment. Chest X-rays (CXRs) have been pivotal to identifying tuberculosis and are widely available. In this work, we utilize CXRs to distinguish between drug-resistant and drug-sensitive tuberculosis. We incorporate Convolutional Neural Network (CNN) based models to discriminate the two types of TB, and employ standard and deep learning based data augmentation methods to improve the classification. Using labeled data from NIAID TB Portals and additional non-labeled sources, we were able to achieve an Area Under the ROC Curve (AUC) of up to 85% using a pretrained InceptionV3 network.**

I. INTRODUCTION

Tuberculosis (TB) is a global disease caused by the bacterium *Mycobacterium tuberculosis*, which is spread through the air. According to the World Health Organization, in 2019 an estimated 10 million people were infected with TB and about 1.4 million died from the disease [1]. Efforts to control TB have been hindered by the rise of drug-resistant strains, where in 2019 about half a million people developed rifampicin-resistant TB out of which 78% were multidrug-resistant [1]. Early detection of drug resistance enables more specific drug treatment, reduces the period of infectiousness and disease spread in addition to improving outcomes [2].

Current diagnostic methods for identifying drug-resistant TB (DR-TB) infections include conventional culture growth over several weeks and rapid molecular testing [3]. These procedures are not feasible globally, especially for countries unable to scale up their testing capacities. An automated computational approach that utilizes widely available technology is thus desirable. Chest X-rays (CXRs) are extensively used in detection of tuberculosis, and thus offer a potentially natural avenue for discriminating between DR-TB and drug-sensitive TB (DS-TB).

In this work, we evaluate multiple CNN architectures and training strategies with the aim of differentiating between DR-TB and DS-TB. We evaluate both pre-trained CNNs as simple N-layer custom CNNs. In terms of training strategies, we evaluate the use of different data augmentation approaches, augmenting the data statically beforehand or dynamically during training. Along with that, we generate synthesized images for DR-TB and DS-TB from the original images using Generative Adversarial Networks (GANs). We utilize a unique TB dataset provided by the US National Institute of Allergy and Infectious Diseases [4]. This patient based dataset includes clinical, genomic, and radiological data (CXRs and CT), but most importantly, it includes the results of drug susceptibility testing. Finally, we utilize several publicly available TB image datasets with unknown drug susceptibility to further enhance our classifier training.

II. PREVIOUS WORK

Computational identification and classification of lung diseases in medical images has been greatly facilitated by advancements in deep learning [5]. In the context of TB, usage of CXRs to classify an image as TB/not-TB has been described in multiple publications. Even simpler architectures such as AlexNet and GoogleNet, used with around 1000 training images, have shown good performance, exceeding 95% accuracy on some datasets [6]. The specific task of detecting TB in CXRs has seen great success, with multiple commercial products available, and a recent study reporting an area under the receiver operating characteristic curve of 0.92 or greater, when evaluated on unseen data [7].

Very few works have dealt with identifying the type of TB, DR-TB or DS-TB, from images. As part of the ImageCLEF 2017 and 2018 challenges, this question was posed using CT images. In 2017/2018 participants of the challenge were provided with a training set comprised of 230/259 training CTs and 223/236 testing CTs. After running the challenge for two years, the organizers said that "After two editions we concluded that the MDR (Multi-Drug Resistant) subtask was not possible based only on the image." [1] It should be noted that the size of the training dataset was very small, and likely adversely affected deep learning based approaches.

In a different study [8], our group had moderate success in differentiating between DR-TB and DS-TB using CXRs, achieving an AUC of 0.66 utilizing hand-crafted shape and texture features. In clinical research, several publications describe using imaging (CXR or CT) to identify clinical findings that potentially differentiate between DR-TB and DS-TB. In [9], the authors found the DR-TB class to have more large lesions whereas the DS-TB class had more medium and small lesions. In [10], the authors found that the DR-TB class was characterized by having thick-walled

[1] Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD 20894 USA
[2] Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases, Bethesda, MD 20894 USA

[1] https://www.imageclef.org/2019/medical/tuberculosis/

cavities. Finally, in [11], the authors found that presence of multiple cavities was a predictor of DR-TB.

Based on our initial results, and the more recent clinical observations, we believe CXRs can potentially be used for differentiating DR-TB from DS-TB using a deep learning approach, which is described in the following sections.

III. Methods

To discriminate between DR-TB and DS-TB, this work collects and processes CXR images from different sources, selects models trained with deep learning based approaches, and uses training strategies to improve classification performance. The CXRs used in this work are from the following sources: TB Portals [4], Montogomery County and Shenzhen chest X-ray sets [12], and the TBX11K large scale tuberculosis dataset [13]. Table I lists the number of samples for each set. The TB Portals dataset is the only one which contains results of drug susceptibility testing, indicating if the image is DR-TB or DS-TB. For all other datasets, we assume the images are DS-TB as that is significantly more common. To ensure that our evaluation is valid, we only use images from the TB Portals dataset in our testing.

A. Data preprocessing

1) Data Selection: The TB portals dataset contains images from hospitals in 16 countries. Because of this, there are variations in the quality of images. We discarded images that are non-pulmonary, from lateral views, or non-grayscale.

Because an early-stage distinction of drug sensitivity or resistance is desirable, only images from a patient's first visit were selected for this analysis. Further, to give equal weight to all patients, a single image was used per patient even when multiple images were acquired on that visit.

2) Cropping of lung regions from CXRs: CXR images often include significantly sized regions that are outside the lungs, such as shoulders and neck. These regions are not relevant for the classification and in fact can be a hindrance in developing accurate models. Cropping a tight region around the lungs and removing unnecessary regions also allows for a more consistent size of the lungs across multiple images once they are rescaled. We therefore use a deep learning approach to crop the original CXRs to the lung region. During the cropping process, the CXRs are blurred by Gaussian smoothing with a standard deviation of 0.5 to reduce the high frequency signal components. Each smoothed image is resampled to a fixed dimension (256x256), before normalizing the intensities to zero mean and unit standard deviation. The CXRs are passed through a U-Net based segmentation model [14]. The resulting lung masks are used to compute a bounding box to crop the lung region from the original CXRs. The segmentation model was trained on two datasets [12], [15] yielding an IoU of 0.971 and 0.956, respectively. Subsequently, the original images are cropped using the bounding box coordinates, downsampled, and renormalized. Figure 1 illustrates this process.

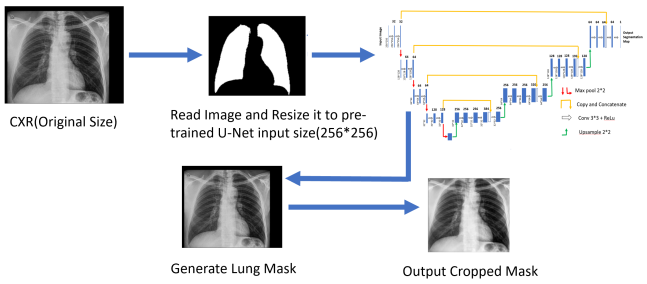

Fig. 1: Preprocessing pipeline for CXRs

B. Network Architectures

For the classification task, we evaluate several standard CNN architectures as well as three custom CNNs. The standard networks include: AlexNet [16], DenseNet [17], InceptionV3 [18], ResNet [19], and Xception [20]. For each of the standard networks, we removed the dense layers after the final convolutional layer and added new dense layers. Table II shows the number of parameters for each of the networks.

C. Data Augmentation

Most deep CNNs require a large amount of good quality data for the models to generalize well. As the number of available samples for each class is relatively small, we use two augmentation approaches:

1) Image transformations: The following transformations are applied to the original images: rotation ($\pm 10^o$), translation (± 5 pixels), blurring ($\mathcal{N}(0.0, 1.0)$), and histogram equalization. We intentionally apply small parameter values for these methods as they replicate the relatively small variations in X-ray images compared to images from other domains. We evaluate the usage of **static augmentation**, one time application of the transformations to the entire dataset before training starts, and **dynamic augmentation**, where original images are modified on the fly during batch training.

2) Synthetic Image Generation: Aside from image transformations, we synthesize images from both categories to increase the number of samples. We use the progressive growing of generative adversarial networks (PG-GANs) [21]. PG-GANs were chosen as they have been shown to generate relatively stable, quality, and variant images. For generating synthetic images for each category during each growth phase of $[4{\times}4, 8{\times}8, \ldots, 128{\times}128]$, batch sizes and epochs

Sources	DR-TB	DS-TB
TB Portals	1821	878
Montgomery County [12]*	0	58
Shenzhen [12]*	0	336
TBX11K [13]*	0	549
Synthetic (using GAN)	1000	1000
Total	2821	2821

TABLE I: Number of images from each source. * Patients from [12] and [13] are assumed to be drug sensitive.

of $[128, 64, 64, 32, 32, 16]$ and $[100, 250, 250, 250, 250]$ were used respectively. The final outputs are up-sampled to the input size of the classifying network.

IV. Experiments

We evaluate model capabilities to distinguish between DR-TB and DS-TB on a patient-level basis. In all experiments, we use 10-fold cross validation.

We start by evaluating multiple models on the TB portals dataset using non-augmented training. We then evaluate the effects of various augmentation strategies on the best models. Finally, we add TB images from external sources, labeling all of them as DS-TB, to the best model from the last step.

A. Model Selection

The pretrained network architectures were designed to address multi-class classification tasks. While we only deal with two classes (DR-TB and DS-TB), the size of the available dataset is much smaller in comparison. We therefore initially evaluate multiple standard architectures and several custom CNNs using the TB Portals dataset.

B. Effects of Augmentation

Once we identify the more promising architectures, we explore the effects of dynamic and static augmentation strategies as well as utilizing synthetically generated images in the training stage. For this experiment, we only select the best performing pretrained-networks (InceptionV3 and Xception) and the best performing custom network. We also evaluate the effect of increasing the amount of statically augmented data on the balanced dataset created in the previous experiment.

C. Including Additional Data

As shown in Table I, the number of DR-TB images in the TB portals dataset is significantly higher than the number of DS-TB images. In an effort to utilize images from all available patients in the imbalanced TB portals dataset, additional TB images from other sources were also included. We label these images as DS-TB, as this is the prevalent type of TB. The previous augmentation strategies were combined with the additional data to see if they influence the overall AUC performance. Note that these images are only used for training purposes as there is no drug susceptibility testing associated with them.

V. Results

In our network architecture comparison, without any augmentation, pretrained InceptionV3 and Xception networks had the best performance, as can be seen in Table II. These two networks, and several custom CNNs (6-layer, 10-layer, 12-layer), were also trained with random initialization. Among the custom networks, the 6-layer CNN had the best area under the ROC curve (AUC) with $0.74 \pm .04$ compared to the rest of the custom networks. When randomly initialized, the performance of InceptionV3 and Xception deteriorated.

Different augmentation methods and addition of synthetic images did not yield better performance for these networks, as shown in Table III. Performance did not scale with the increase in augmented data. When number of samples was increased to 3X and 4X original samples size by static augmentation, performance decreased. Interestingly, the performance of the custom 6-layer network improved with the same training strategy. We chose to continue our evaluation using the InceptionV3 network as its performance remained the most consistent with these augmentation strategies.

Architecture	**Parameters (in millions)**	**AUC**
Pretrained Networks		
AlexNet [16]	5.7	0.79 (± .02)
DenseNet121 [17]	7.2	0.79 (± .02)
DenseNet201 [17]	18.6	0.80 (± .02)
InceptionV3 [18]	22.3	**0.81 (± .03)**
InceptionResNetV2 [18]	54.7	0.77 (± .05)
ResNet50 [19]	24.1	0.80 (± .03)
ResNet152 [19]	58.7	0.77 (± .03)
Xception [20]	21.3	**0.81 (± .02)**
Random initialization		
6-layer CNN	3.0	**0.74 (± .04)**
10-layer CNN	8.6	0.70 (± .03)
12-layer CNN	9.1	0.65 (± .04)
InceptionV3 [18]	22.3	0.76 (± .03)
Xception [20]	21.3	0.76 (± .03)

TABLE II: Mean AUC (Area Under ROC Curve) of 10-fold cross validation results when various pretrained and custom networks are tested on TB Portals dataset

Figure 2 summarizes the performance evaluation of InceptionV3, using various datasets and augmentation strategies. We see that static augmentation has an overall positive effect on model performance compared to the dynamic augmentation strategy. We also see that the addition of images from other sources to the training set combined with static augmentation lead to the best AUC overall performance of 85%.

Although the inclusion of *both* the *synthetic* data and data from *additional sources* improves performance when using dynamic augmentation, it did not have an effect when using static augmentation. The variance in performance is slightly better when *both* synthetic and additional data are included.

Finally, to inspire confidence in the predictions of our network, we utilize GradCAM heatmaps [22] to visualize its focus. Figure 3 shows two heatmaps for correctly predicted .

Network Architecture	**Dynamic**	**Static**			**Synthetic**
		2X	*3X*	*4X*	
InceptionV3 (pretrained)	0.80 (±.03)	**0.81** (±.03)	0.80 (±.02)	0.79 (±.02)	**0.81** (±.02)
Xception (pretrained)	0.80 (±.03)	0.80 (±.03)	0.77 (±.04)	0.79 (±.03)	**0.81** (±.03)
6-layer CNN	0.76 (±.03)	0.76 (±.02)	0.74 (±.02)	0.76 (±.03)	0.75 (±.03)

TABLE III: AUC with *dynamic* and *static* augmentation and with augmentation using GAN generated images.

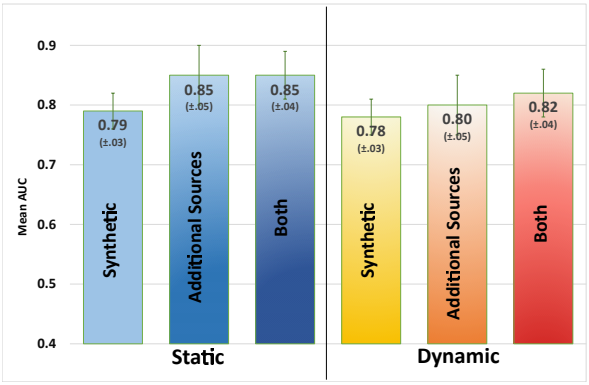

Fig. 2: Mean AUC performances of the InceptionV3 network with *static* or *dynamic* augmentation and including a) *synthetic* images, b) images from [12] and [13] (referred in figure as *additional sources*) c) *both*. These additional images with static augmentation provided the best performance.

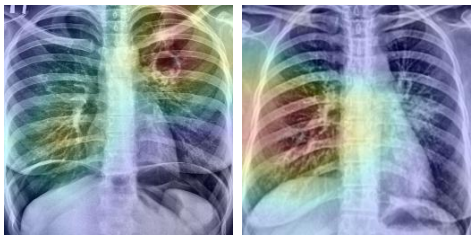

Fig. 3: GradCAM heatmaps superimposed on the original images, Classified as DR-TB (left image) and DS-TB (right image) due to likelihood values of .99 and .05 respectively.

VI. CONCLUSIONS

This paper presents an evaluation of models for discriminating between drug-resistant and drug-sensitive TB in the TB portals dataset, using augmentation strategies and other publicly available data. With a 10-fold cross validation, we achieve the best AUC performance of 85%. Even without augmentation and additional data, but with pretrained weights, we achieve a 81% AUC performance with InceptionV3 and Xception networks. GradCAM heatmaps affirm that the models learn from relevant areas from the CXRs during the training process. Despite discouraging earlier work in the literature, our work has shown that discriminating between DR-TB and DS-TB can be possible in CXRs for a sufficiently large training set.

ACKNOWLEDGMENT

This work was supported by the Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF), under Interagency Agreement #750119PE080057, and by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. This project has also been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases under BCBB Support Services Contract HHSN316201300006W/HHSN27200002.

REFERENCES

[1] World Health Organization, *Global tuberculosis report*, 2020.
[2] P. O'Riordan, U. Schwab *et al.*, "Rapid molecular detection of rifampicin resistance facilitates early diagnosis and treatment of multidrug resistant tuberculosis: case control study," *PLoS One*, vol. 3, no. 9, p. e3173, 2008.
[3] C. Lange, K. Dheda *et al.*, "Management of drug-resistant tuberculosis," *The Lancet*, vol. 394, no. 10202, pp. 953–966, 2019.
[4] A. Rosenthal, A. Gabrielian *et al.*, "The TB portals: an open-access, web-based platform for global drug-resistant-tuberculosis data sharing and analysis," *Journal of Clinical Microbiology*, vol. 55, no. 11, pp. 3267–3282, 2017.
[5] S. T. H. Kieu, A. Bade, M. H. A. Hijazi, and H. Kolivand, "A survey of deep learning for lung disease detection on medical images: state of the art, taxonomy, issues and future directions," *Journal of Imaging*, vol. 6, no. 12, 2020.
[6] 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.
[7] Z. Z. Qin, M. S. Sander *et al.*, "Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems," *Sci Rep.*, vol. 9, no. 1, 2019.
[8] S. Jaeger, O. H. Juarez-Espinosa *et al.*, "Detecting drug-resistant tuberculosis in chest radiographs," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 12, pp. 1915–1925, 2018.
[9] A. G. Icksan, M. R. S. Napitupulu, M. A. Nawas, and F. Nurwidya, "Chest X-ray findings comparison between multi-drug-resistant tuberculosis and drug-sensitive tuberculosis," *Journal of Natural Science, Biology, and Medicine*, vol. 9, no. 1, p. 42, 2018.
[10] Y. X. J. Wang, M. J. Chung *et al.*, "Radiological signs associated with pulmonary multi-drug resistant tuberculosis: an analysis of published evidences," *Quant Imaging Med Surg.*, vol. 8, no. 2, pp. 161–173, 2018.
[11] S. Flores-Trevino, E. Rodriguez-Noriega *et al.*, "Clinical predictors of drug-resistant tuberculosis in Mexico," *PLoS One*, vol. 14, no. 8, 2019.
[12] S. Jaeger, S. Candemir *et al.*, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative Imaging in Medicine and Surgery*, vol. 4, no. 6, p. 475, 2014.
[13] Y. Liu, Y.-H. Wu, Y. Ban, H. Wang, and M.-M. Cheng, "Rethinking computer-aided tuberculosis diagnosis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
[14] P. Ivan, P. Vitali, and B. Uladzislau, "Lung segmentation (2D)," https://github.com/imlab-uiip/lung-segmentation-2d, 2020.
[15] J. Shiraishi, S. Katsuragawa *et al.*, "Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *AJR. American Journal of Roentgenology*, vol. 174, pp. 71–4, 02 2000.
[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.
[17] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*, 2017.
[18] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*, 2016.
[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.
[21] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.
[22] R. R. Selvaraju, M. Cogswell *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.