Tuberculosis Classification Survey using Computer Vision models

Sofia Blinova **Ekaterina Trimbach** **Saqib Javed**

Abstract

Tuberculosis (TB) is a serious infectious disease that mainly affects the lungs caused by Mycobacterium tuberculosis (MTB) bacteria. Recently, deep learning based methods have shown promising results to detect TB using patients' digital Chest X-rays(CXR). However, digital CXR is unavailable in many parts of Asia or Africa, where TB is most commonly found. In this paper, we investigate the possibility of predicting the presence of TB in HIV and diabetic patients using the images of their CXR taken by a camera. First, we investigate the potential of already known models for predicting tuberculosis, then, having one fixed model, we apply various techniques for changing the sampling algorithm and loss function to improve the final quality. As a result, we managed to achieve accuracy **95.7%** and F1 score **88.4%** for ResNet-18 model.

1. Introduction

The article (WHO, 2018) provides statistics that as of 2018, occurred more than 10 million tuberculosis cases all over the world. TB can be considered as a very dangerous disease as it is the 13th leading cause of death and the second leading infectious killer after COVID-19 in 2020 according to Global tuberculosis report 2021(WHO, 2021).

Furthermore, diagnosing TB is not an easy task by any means and has certain complications. The golden standard test for TB is still the technique for isolating bacteria which is a very time-consuming process. Despite the fact that Polymerase Chain Reaction (PCR) based technique solves this problem by providing timely results, unfortunately, they cannot be performed by all medical institutions due to their high cost(Duong et al., 2021). To this end, many researchers are trying to use medical imaging techniques based on deep learning to assist doctors.

Problem statement. The main objective of this research is to analyse whether deep learning can be used in the detection of TB in HIV and diabetic patients using the images of their Chest X-rays. We study our results specially for the HIV and diabetic patients as they are more likely to have TB.

2. Related Work

Generally, CXR of the patients are studied by the doctors to detect if the patient is infected with TB or some other pulmonary disease where the advanced facilities are not available. Various studies have been done in the past that can aid doctors in the detection of tuberculosis and other pulmonary diseases. As deep learning based methods normally work very well with problems that involve image data, many researchers have been using deep learning to detect these diseases from the CXR images of the patients. Some of these methods got very accurate detection of TB or other pulmonary diseases [(Wong et al., 2021),(Rahman et al., 2020),(Ajas-tb, 2021)]. However, digitized CXR images[(Pasa et al., 2019),(Jaeger et al., 2014)] were used in these studies which are not available everywhere.

In this regard, our work deals with a private dataset consisting of images of the patient's CXR taken by the camera as shown in figure 3 instead of digitized CXR images. In addition, we also incorporate the data from HIV and diabetic patients as they are more likely to have TB. One such study that uses deep learning has been done on HIV patients to detect TB using patients' CXR and many other clinical covariates (i.e. CD4 count, WBC count, etc.)(Rajpurkar et al., 2020). They also crop CXR images to exclude regions outside of the lungs using human annotations of the lung regions and get very good results. Nevertheless, human annotation of the lung regions in an image is a very time consuming process and was not studied during the course of this project. In our research, we explore the fact of whether we can rely only on the camera images of these CXR of the patients.

3. Dataset

3.1. Data Analysis

Our dataset contains 2879 CXR images for normal cases and 598 TB cases. A distinctive feature of our dataset is that it consists of real photographs of CXR taken very close to reality. The data collection involved patients from Benin and Guinea. People with HIV and diabetes participated in the study and you can see the number of cases from each class and country mentioned in Table 1. Moreover, some additional samples of confirmed TB cases were added as we

Table 1. Dataset Overview

	Benin		Guinea		Additional Samples
	Diabetic	HIV	Diabetic	HIV	
TB	10	7	7	27	547
Normal	1718	384	424	359	0

had more normal cases.

The second distinguishing feature of our dataset is the presence of labels on the CXR images. Due to the fact that our data are photos of real CXR, they may contain the doctor's handwritten comments or stickers indicating some of the medic's assumptions. This fact does not prevent testing models trained on third-party datasets, but it is necessary to take this into account during training models on our dataset.

3.2. Data Preprocessing

Dataset's images have some soft labels(colored rectangle) and it could cause label leakage. To this end, some preprocessing was required and we started by using OpenCV functions. The method that worked the best, finds the rectangular shapes in the image after applying a threshold to make the image black and white. An example of a good and a bad image after label removal using our best preprocessing technique can be seen in the figure 3. Initially, both pictures contained labels, but in the first picture the label was successfully removed (in the center of the picture can be noticed a gray square on which there used to be a sticker), but one picture 2 our approach did not detect that rectangle. We tried several other methods such as finding rectangles of a specific color, finding rectangles after applying canny edge detection, and template matching. Moreover, some morphological operations like dilation and erosion were used along with these strategies to get better results. Nevertheless, we were not able to remove the labels perfectly from all the images.

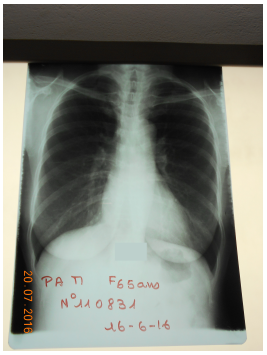

Figure 1. Good case

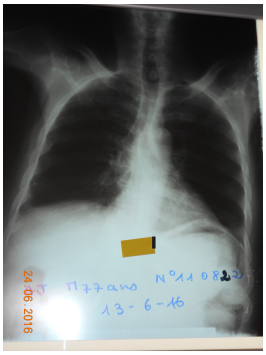

Figure 2. Bad case

Figure 3. Example of a good and a bad image after label removal.

After trying a lot of different strategies, we decided to use one of the simplest approaches i.e converting RGB images to grayscale to reduce the impact of these different color soft labels. Despite its simplicity, the method of translating images into grayscale format is reliable and has met us many times in various articles and implementations.

Furthermore, we also tried to crop the CXR out of the images and remove the background using OpenCV. We could not test many different methods here due to time limitations. One of our best approaches here uses edge detection techniques such as Canny edge and hough lines along with morphological operations mentioned before on our images. The results obtained are fed into the OpenCV function *goodFeaturesToTrack*. This function tries to find the best four points from the edges detected in the image and should correspond to the corners of the CXR in the image. Notwithstanding, we were not able to get any good results here but this approach can be fine-tuned and better results could be obtained.

4. Existing Models and Methods

4.1. Tuberculosis-TB-Analyzer

Tuberculosis-TB-Analyzer (TTBA) is a convolutional neural network consisting of 20 layers[1]. That model was trained on an open-source dataset published at the kaggle tuberculosis competition containing 336 tuberculosis and 325 normal cases. Also, the model's authors provided a web application for detecting tuberculosis online[2]. Testing this model on our dataset shows poor results with F1_score 0% and 4% without preprocessing and grayscale conversion of images respectively. We assume that this happened because TTBA trained on clear X-ray images, while we are dealing with images of images.

After getting low accuracy, we decided to train the model on our dataset using grayscale conversion. Nevertheless, we were not able to get any better results and got 0% F1_score.

4.2. Tuberculosis ChestXray Classifier

The article (Duong et al., 2021) proposed Vision Transformer model for predicting tuberculosis [3]. This model uses transfer learning with various data augmentations such as RandomRotation, RandomResizedCrop, RandomHorizontalFlip, ColorJitter, RandomErasing. These transformations should improve the accuracy of the model and make it more resistant to various image distortions. Unfortunately, after training it on our dataset the Tuberculosis ChestXray Classifier (TCXC) predicted all images as normal cases. We

[1]https://github.com/vbookshelf/Tuberculosis-TB-Analyzer
[2]http://tb.test.woza.work/
[3]https://github.com/linhduongtuan/Tuberculosis_ChestXray_Classifier

assume that this model has too many parameters and the size of our dataset is too small for it, as a result of which we are more likely to face under-training.

4.3. AJAS-TB

AJAS-TB is a convolutional neural network consisting of 14 layers[4]. The original model was trained on an open-source kaggle dataset containing 3500 normal and 700 tuberculosis cases [5]. During testing on our data AJAS-TB also showed bad results with metrics only 25% as accuracy and 21% F1_score. We assume that the reason for such results is similar to the TTBA model. As a next step, as well as for the TTBA model, we trained this AJAS-TB on our data using grayscale preprocessing. As a result, we achieved 18% accuracy and 30% F1_score.

5. Our method

It seems that the existing models failed to show good performance for our data. Thus, we decided to use ResNet-18 model pretrained on the ImageNet dataset. However, it also resulted in the same bad accuracy and F1_score as the existing models. Such results mean that models cannot distinguish differences between classes due to imbalance which can be addressed with the right training strategy. We identify three ways how we can deal with this problem and improve model performance: re-weighting, loss function adaptation and re-sampling.

5.1. Model setup

We changed the classification head of the original ResNet-18 to apply it for binary classification. Furthermore, we resized input images to 224 x 224 and transformed them to grayscale. For the training set, we additionally apply random cropping, flipping and rotation. Apart from this, we normalise the pixel values based on the per-channel mean and standard deviation, precomputed on the training set.

5.2. Imbalance learning methods

In medical imaging datasets, class imbalance is a very natural problem. Different techniques that we used to solve this problem are discussed in this section.

Due to huge dataset imbalance, accuracy alone is not the best measure for assessing performance. Thus, we also take into account F1 score.

Re-weighting. For these experiments we use Cross Entropy loss with different weights. Firstly, we apply empirical frequency re-weighting, i.e.

$$w_i = \frac{1}{N_i}, \quad i \in \{0, 1\}. \tag{1}$$

Then, we use re-weighting function that is usually applied to medical tasks (Paszke et al., 2016). The formula for this re-weighting technique is:

$$w_i = \frac{1}{1.02 + \log\left(\frac{N_i}{N}\right)}, \quad i \in \{0, 1\}. \tag{2}$$

Finally, we apply weights that consider empirical effective number of samples per class (Cui et al., 2019). The formula for this technique includes the effective number of samples per class $N_i^{eff} = \frac{1-\beta}{1-\beta^{N_i}}$ that is used for calculating final class weights

$$w_i = \frac{N_i^{eff}}{\sum_{i=1}^{n} N_i^{eff}}, \quad i \in \{0, 1\}. \tag{3}$$

In a sequel, the best result we obtain with the logarithmic re-weighting. Therefore, we used this method in the further work. Comparison of different re-weighting approaches is shown in Table 2.

Table 2. Re-weighting comparison

	Freq. RW	Log. RW	Eff. RW
Acc, %	17.7	**51.7**	82.3
F1_score, %	30.1	**34.2**	0.0

Loss functions. In this section, we compare two loss functions for imbalanced learning with the simple cross entropy loss. The results are presented in Table 3. Below, we also provide a brief explanation to these loss functions.

Focal loss. We use Focal loss (Lin et al., 2017) which gives lower weights to easily classified samples. It adds a modulating factor to the cross entropy $\text{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log p_t$, where $p_t = \mathbb{1}[y = 1]p + \mathbb{1}[y = 0](1 - p)$, α_t are loss weights, e.g. from Equation (2), γ is a tunable.

Super loss. The SuperLoss (Castells et al., 2020) takes the conventional loss (in our case, Cross Entropy) of a given sample ℓ_i and its confidence σ_i. It is composed of a loss-amplifying term and a regularization term controlled by the hyper-parameter $\lambda > 0$:

$$L_\lambda(\ell_i, \sigma_i) = (\ell_i - \tau)\sigma_i + \lambda(\log \sigma_i)^2,$$

where τ is a threshold that ideally separates easy samples from hard samples based on their respective loss. If the optimisation of confidence parameters σ_i is complicated due to task-specific overheads, one can leverage the optimal confidence:

$$\sigma_\lambda^*(\ell_i) = e^{-W\left(\frac{1}{2}\max\{\frac{2}{e}, \beta\}\right)}, \text{ with } \beta = \frac{\ell_i - \tau}{\lambda},$$

where W stands for the Lambert W function.

Re-sampling. The most obvious approach to avoid imbalance in data is re-sampling, e.g. under-sampling of the head classes and over-sampling the tail ones. The idea of

[4]https://github.com/ajas-tb/tuberculosis

[5]https://www.kaggle.com/tawsifurrahman/tuberculosis-tb-chest-xray-dataset

Table 3. Loss function comparison

	Cross Entropy	Focal	Super
Acc, %	**84.0**	17.7	82.3
F1_score, %	**17.6**	30.1	0.0

Table 4. Re-sampling comparison

	IB sampling	Eff. sampling
Acc, %	**95.7**	93.7
F1_score, %	**88.4**	83.3

re-sampling is to sample classes into a batch with the probability different from uniform. We list strategies that we tested below. The first method is re-sampling with probabilities proportional to the relative number of samples in a give class:

$$p_j = \frac{N_j}{\sum_{i=1}^{c} N_i}, \quad j \in 0, 1,$$

where $N_j, j \in 0, 1$ is number of samples in the jth class. In the literature, this method is usually referred to *Instance-Based sampling (IB)*. Another approach utilises the same formula but replace the usual number of samples with the effective one as introduced in Equation (3). We call it *Effective sampling*. Results for different re-sampling approaches described in Table 4.

6. Results

In table 5, we show that our model outperforms existing methods.

Table 5. Models comparison

Model name	Training	Acc, %	F1-score, %
TTBA 4.1	kaggle dataset	78.0	4.0
TTBA 4.1	our dataset	82.3	0.0
TCXC 4.2	our dataset	82.3	0.0
AJAS-TB 4.3	kaggle dataset	25.0	21.0
AJAS-TB 4.3	our dataset	18.0	30.0
ResNet-TB 5	our data	**95.7**	**88.4**

7. Discussion

After testing all the approaches, the best accuracy and F1_score for ResNet-18 is 95.7% and 88.4% respectively. These metrics undoubtedly stand out from all others methods that we tested. In this section, we delve into the analysis of our best model and investigate on which instances our model classify images wrong. As mentioned in section 3.1, our dataset contains images from HIV and diabetic patients. On the figures 6 and 5, ResNet-18 results are shown for patients classified as diabetic and HIV. Figure 7 presents results for subdataset marked as Additional Samples. From the confusion matrix, it can be concluded that our model does not perform so well for predicting TB in diabetic and HIV patients.

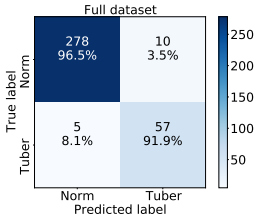

Figure 4. All test data

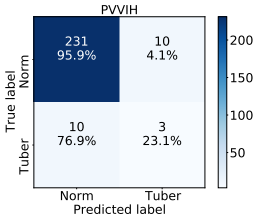

Figure 5. HIV patients

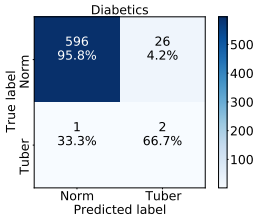

Figure 6. Diabetic patients

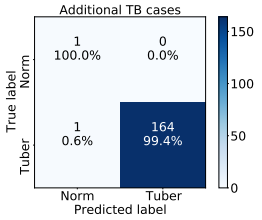

Figure 7. Additional TB cases

Figure 8. Confusion matrices for different data

7.1. Limitations

We certainly have many limitations in this work. However, the biggest of them all is the non-standardized way of taking images of the patients' CXR that causes a huge challenge in interpretability. It could be the case that our model is learning certain lighting factors or white balance instead of trying to find good features in the lung region. This fact makes it harder for model to detect TB using these images. Label leakage is another issue that is caused by the soft label present in the images.

7.2. Future Work

Even after this extensive analysis of real images of CXR, there is still room for a lot of improvement. Preprocessing of the images can be improved in various ways. Firstly, label leakage can be countered by either manually annotating the images or improving the label removal strategy. Secondly, extracting the CXR out of these images by removing the background can also improve the results. However, the best technique would be to perform the segmentation of the lung region as it can be used to improve the results of the TB detection and also solve the problems mentioned above. Moreover, a standard environment should be used to capture the images of these CXR to make them have uniform exposure, contrast, white balance, etc.

References

Ajas-tb. Lung x-ray image classification using deep neural network, 2021. URL https://github.com/ajas-tb/tuberculosis.

Castells, T., Weinzaepfel, P., and Revaud, J. Superloss: A generic loss for robust curriculum learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples, 2019.

Duong, L. T., Le, N. H., Tran, T. B., Ngo, V. M., and Nguyen, P. T. Detection of tuberculosis from chest x-ray images: Boosting the performance with vision transformer and transfer learning. *Expert Systems with Applications*, pp. 115519, 2021.

Jaeger, S., Candemir, S., Antani, S., Wáng, Y.-X. J., Lu, P.-X., and Thoma, G. R. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4 6:475–7, 2014.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D., and Pfeiffer, D. Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization. *Scientific Reports*, 9, 04 2019. doi: 10.1038/s41598-019-42557-4.

Paszke, A., Chaurasia, A., Kim, S., and Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation, 2016.

Rahman, T., Khandakar, A., Kadir, M. A., Islam, K. R., Islam, K. F., Mazhar, R., Hamid, T., Islam, M. T., Kashem, S., Mahbub, Z. B., Ayari, M. A., and Chowdhury, M. E. H. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601, 2020. doi: 10.1109/ACCESS.2020.3031384.

Rajpurkar, P., O'Connell, C., Schechter, A., Asnani, N., Li, J., Kiani, A., Ball, R. L., Mendelson, M., Maartens, G., van Hoving, D. J., Griesel, R., Ng, A. Y., Boyles, T. H., and Lungren, M. P. Chexaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with hiv. *npj Digital Medicine*, 3(1), 2020. doi: 10.1038/s41746-020-00322-2.

WHO. *Global tuberculosis report 2018*. World Health Organization, 2018.

WHO. *Global tuberculosis report 2021*. World Health Organization, 2021.

Wong, A., Lee, J. R. H., Rahmat-Khah, H., Sabri, A., and Alaref, A. Tb-net: A tailored, self-attention deep convolutional neural network design for detection of tuberculosis cases from chest x-ray images, 2021.