IDIAP COMMUNICATION REPORT

ACTIVE TUBERCULOSIS DETECTION FROM FRONTAL CHEST X-RAY IMAGES

Geoffrey Raposo

Idiap-Com-01-2021

JULY 2021

Centre du Parc, Centre du Parc, Rue Marconi 19, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Active tuberculosis detection from frontal chest X-ray images

Thesis performed in the

Biosignal Processing Research Group
at the **Idiap Research Institute** in the context of the
UniDistance Artificial Intelligence Programme

to obtain the degree of

Master of Science in Artificial Intelligence

by

Geoffrey Raposo
Student number: 11-877-677

Project supervisors / company supervisor:
Dr André Anjos, Dr Anete Trajman / Flavio Tarsetti

Martigny, July 14, 2021

Acknowledgements

First, I would like to particularly thank Dr. André Anjos for his excellent supervision throughout this project. His expertise, advice, and conviviality made the realization of the present thesis a great experience. It was a real pleasure to work with him.

Next, I would like to thank Dr. Anete Trajman for answering all the questions related to the medical domain and for offering a different and critical look at this work.

I would also like to thank Flavio Tarsetti whose recommendations were of great help to me in writing the initial version of this report.

Finally, thanks to all the teachers and assistants at Idiap for having allowed me to acquire crucial skills for this project.

G. R.

Abstract

Tuberculosis (TB) is one of the leading causes of death from a single infectious agent in the world. In many high-burden regions, which often lack specialized healthcare professionals, Chest X-Ray (CXR) exams continue to be of vital importance in the diagnosis and follow-up of the various presentations of the disease. In this context, automated systems to support diagnosis from CXR images constitute a fundamental cog as the World Health Organization (WHO) confirmed in early 2021 that they can be used in place of human readers for the interpretation of digital CXRs.

In this study, we investigate the benefits of automatic Pulmonary Tuberculosis (PTB) detection methods based on radiological signs found on CXR. Contrary to direct scoring from images, implemented in most related work, indirect detection offers natural interpretability of automated reasoning. We identify generalization difficulties for direct detection models trained exclusively on the modest amount of publicly available CXR images from PTB patients. We subsequently show that a model, pre-trained on tens of thousands of CXR images using automatically annotated radiological signs, offers a more adequate base for development. By relaying radiological signs through a simple linear classifier, one is able to obtain state-of-the-art results on three publicly available datasets (test AUC on Montgomery County-MC: 0.97, Shenzhen-CH: 0.90, and Indian-IN: 0.93). We further discuss limitations imposed by the limited number of PTB-specific radiological signs available on public datasets, and evaluate possible performance gains that could be obtained if more were available (test AUC MC: 0.98, CH: 0.98, IN: 0.93).

We then analyze the relative importance of each of the radiological signs for PTB prediction using two distinct methods and conclude that more than a specific sign, it is their combination that allows a reliable detection of the disease.

Finally, we propose a visual overview of the radiological signs predictions over radiographs using grad-CAMs and highlight the importance of annotating PTB datasets to study the reliability of these visualizations.

Our work is made open-source[1] and fully reproducible in the hopes it becomes useful to further explore the application of Deep Learning to PTB screening.

[1] https://gitlab.idiap.ch/bob/bob.med.tb

Table of contents

List of Figures

List of Tables

List of Abbreviations

AIDS	Acquired Immunodeficiency Syndrome
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating characteristic Curve
BFGS	Broyden–Fletcher–Goldfarb–Shanno algorithm
CAD	Computer-Aided Diagnosis
CH	Shenzhen
CNN	Convolutional Neural Network
CXR	Chest X-Ray
DL	Deep Learning
HIV	Human Immunodeficiency Viruses
IGRA	Interferon-Gamma Release Assay
IN	Indian Collection
LTBI	Latent TB Infection
MC	Montgomery County
ML	Machine Learning
PTB	Pulmonary Tuberculosis
ROC	Receiver Operating Characteristic
TB	Tuberculosis
TST	Tuberculin Skin Test
WHO	World Health Organization

1. Introduction

Despite being preventable and curable, Tuberculosis (TB) is still one of the leading causes of death from a single infectious agent in the world [1]. Indeed, it is estimated that about one quarter of the world population is infected with TB bacteria (*Mycobacterium tuberculosis*), and carry the disease in an asymptomatic latent form [2]. However, only a small percentage of infected people (10-15% [3]) will become sick with active pulmonary TB (PTB). The rest of the infected population will keep it in a latent, asymptomatic, and non-contagious form, typically described as Latent TB Infection or LTBI. Although being a global disease, over 95% of cases and deaths are localized in developing countries (Figure 1.1). In fact, the risk of developing PTB is greater if an LTBI patient is affected by another disease impairing their immune system (co-morbidities), or if people suffer from undernutrition; precisely the kind of situations more frequent in developing countries. For example, chances to develop PTB is 19 times greater for HIV-infected people [4]. Common signs and symptoms of an active TB case include cough with sputum and a small amount of blood at times, chest pains, weakness, weight loss, fever, and night sweats. On the lungs, it is characterized by several combinations of radiological findings including infiltration, pleural effusion, and more.

Although tuberculosis is considered a rare disease in developed countries, it is a major threat worldwide, killing more people than HIV/AIDS. In 2019, 1.4 million people died from TB, and an estimated 10 million people contracted it [1].

Hopefully, TB is curable and preventable. The ambitious aim of the United Nations is, incidentally, to end TB epidemic thanks to the World Health Organization (WHO) "End TB Strategy" [5]. This strategy corresponds to a 95% reduction in the number of deaths and a 90% reduction in TB incidence rate by 2035 compared to the baseline of 2015.

To reach United Nations goal, both active TB and LTBI cases need to be treated (Figure 1.2). While active TB cases can be suspected thanks to patient's symptoms and microbiologically tested, LTBI asymptomatic patients are more difficult to identify. Therefore, WHO's recommendations for LTBI elimination consist of a cascade of steps starting with at-risk populations identification followed by active TB rule out, LTBI testing, and treatment (Figure 1.3). At-risk populations spawn a very large set of potential LTBI treatment candidates including adults and children infected with HIV, HIV-negative contacts of either active TB or HIV patients, and other HIV-negative at-risk groups.

Although active PTB patients usually have symptoms, some are asymptomatic and could potentially be wrongly placed in the LTBI at-risk population. Those patients need to be eliminated from the group which will receive an LTBI treatment (as they require an active

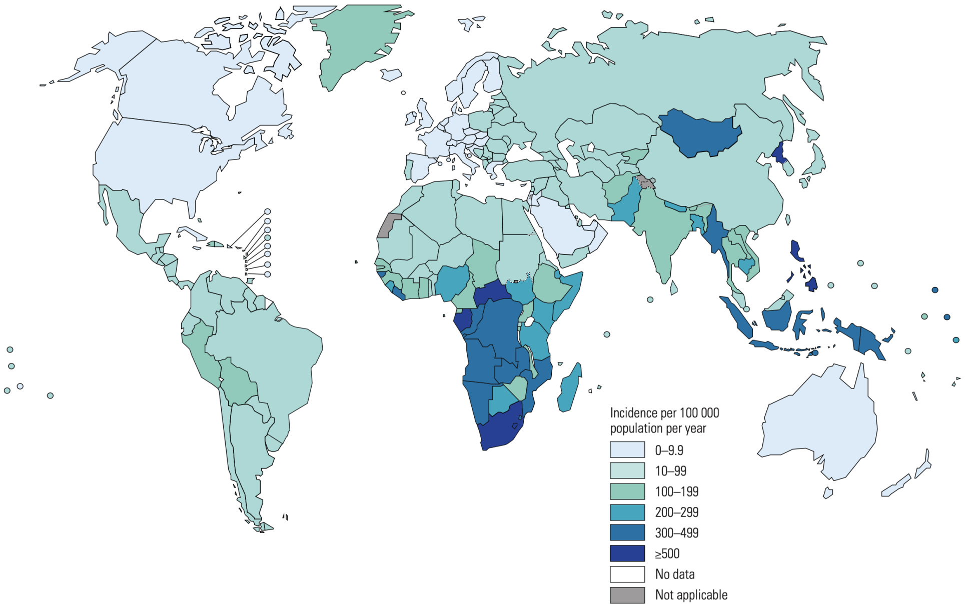


Figure 1.1: Estimated TB incidence rate (percentage of new cases in a population), 2019 [1]. Over 95% of cases and deaths are localized in developing countries where co-morbidities and undernutrition are frequent.

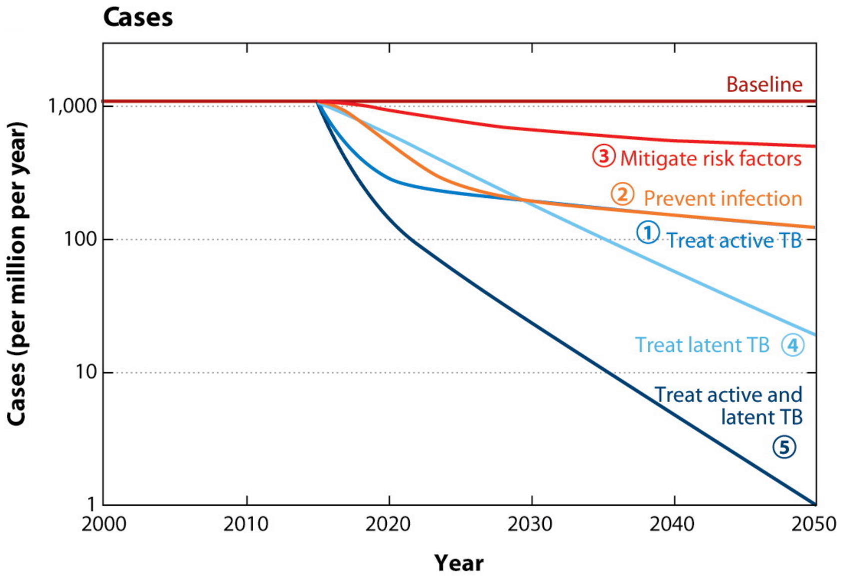


Figure 1.2: TB elimination projections from [6]. Only by treating both active TB and LTBI, United Nations goal can be reached.

TB treatment instead). Since asymptomatic active TB patients can not do a microbiological test in the absence of sputum, chest radiography is required to rule them out. In consequence, every patient having a positive TST (tuberculin skin test) or IGRA (interferon-gamma release assay) will be screened using Chest X-Ray (CXR) imaging before receiving an LTBI treatment [8].

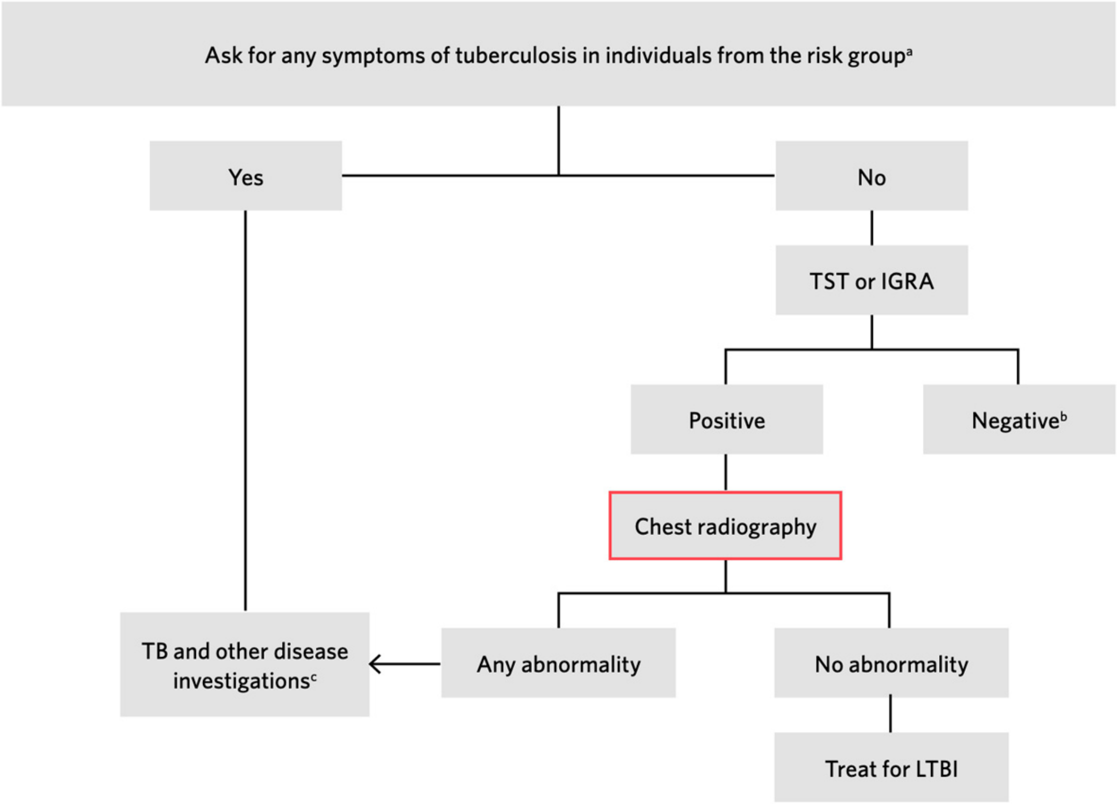

Ask for any symptoms of tuberculosis in individuals from the risk group[a]
Yes
No
TST or IGRA
Positive
Negative[b]
Chest radiography
TB and other disease investigations[c]
Any abnormality
No abnormality
Treat for LTBI

Figure 1.3: Algorithm for targeted diagnosis and treatment of LTBI and exclusion of active TB in HIV-negative household contacts aged >= 5 years and other at-risk populations from [7]. The red box indicates the stage of the process we are going to work on. It is where active TB cases are ruled out of the LTBI at-risk group.

Being the most basic form of radiography, CXR technology is widely available. Chest radiography is also considered the most accurate screening tool for detecting PTB in the general population among standard testing tools protocoled by the WHO [9]. But despite being considered a simple exam (and rather non-sensitive), accurate interpretation of CXR images requires experience as, depending on disease progression and clinical covariates, PTB may lead to different radiological signs (or to their absence). Unfortunately, there is often a lack of specialized professionals at the location of the examination or the few available are overburdened by the sheer volume of images to analyze. In this context, automated systems to support diagnosis from CXR images constitute a fundamental cog. Furthermore, the use of computer-aided detection software is expected to grow in the future as the WHO confirmed in early 2021 that they can be used in place of human readers for the interpretation of digital CXRs, for screening and triage for TB disease [9]. Such systems contain advanced pattern recognition software enabling the identification of radiologic abnormalities corresponding to the manifestations of diseases.

In the current clinical workflow, the patient is referred to CXR imaging and a radiologist analyzes the image to identify radiological signs, confirming or not the presence of PTB.

In both cases, the radiologist will send back a report of the analysis, including supporting evidence (radiological signs) to the original referrer. While discernment of PTB signs from this imaging modality is considered to be a highly specialized task, this relationship between the radiologist and the referring healthcare professional can be characterized by the flow of image annotations (signs) and impressions. From an engineering perspective, radiological signs constitute, therefore, a natural and interpretable basis in this context. However spontaneous, the automated detection of PTB via radiological signs is not a common practice in the literature [10], as a direct and less explainable approach, indicated in Figure 1.4, is usually preferred.

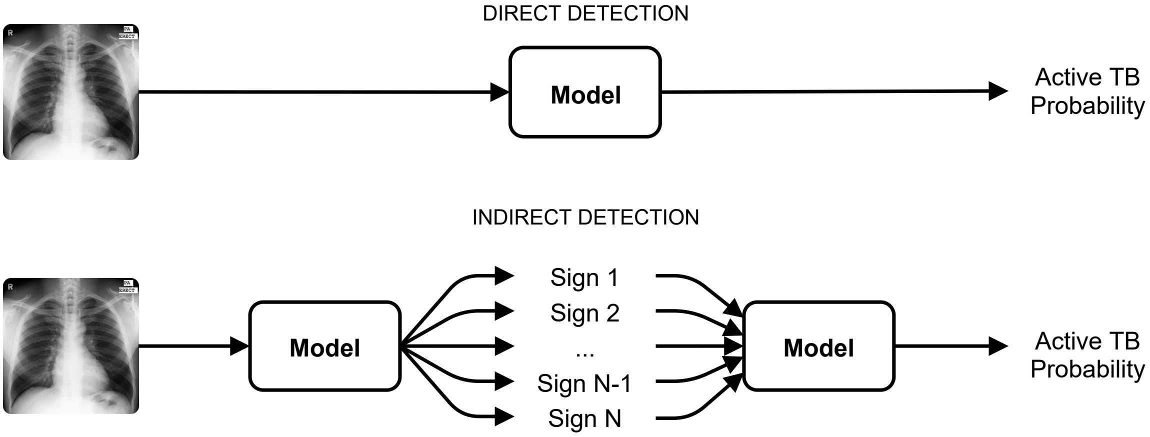


Figure 1.4: Direct and indirect PTB assessment from CXR images. Machine Learning models found in the literature usually implement direct detection while healthcare professionals an indirect approach.

In the present work, we investigate if an indirect detection of PTB, exclusively based on radiological signs extracted from CXR imaging provides better generalization and is capable to reach state-of-the-art results, via Deep Learning (DL) methods, with minimal tuning. Our work, contrary to most publications on this realm, is made fully reproducible (open-source and exclusively built on public datasets).

As it follows, in Section 2, we provide an overview of related work in automatic PTB detection and introduce the Bob framework that we use to make our work reproducible. In Section 3, we present our evaluation process, public datasets used, and direct detection baselines for comparison. Then, the details of our indirect approach are exposed in Section 4, where we present our results, compare them with the baselines, evaluate the importance of each radiological sign, and propose a visualization of the predictions. We finally conclude with the contributions and limitations of our study in Section 5.

2. Related Work

A large amount of studies addressing automatic PTB detection from CXR images exist in the literature [10, 11]. Albeit various Machine Learning (ML) methods were tried in the past [12, 13, 14], current state-of-the-art results stem from Deep Learning (DL) techniques based on convolutional neural networks (CNN) [10], which offer great plasticity and generalization capabilities in image classification [15, 16, 17].

PTB detection is generally posed as a binary classification problem, where one needs to identify PTB patients from healthy subjects [10]. It seems common practice to base ground-truth labels of available datasets [18, 19] on the results of standard skin and sputum tests while avoiding the sometimes prohibitive workload of enumerating and locating various radiological signs on CXR imaging for the classification problem. In practice though, realistic scenarios where computer-aided diagnosis (CAD) from CXR imaging for PTB could be useful are rather different [20]. In high-burden countries, for example, PTB must be screened against the general population, with individuals potentially presenting various other (pulmonary) diseases. Patients with positive skin or sputum tests, in different stages of the disease, or due to other clinical covariates (e.g. HIV-positive), may not present classical PTB symptoms clearly visible on CXR images [21]. In this scenario, we argue that CAD from CXR should limit itself to identify factual and reportable radiological signs that can be identified on the original images. Attempts to perform a direct detection from images that have not gone through rigorous radiological screening, coupled with the relatively small size of available datasets, could lead to unaccounted biases.

In the present section, we will first review studies addressing active PTB detection and then review studies addressing radiological findings detection.

2.1 Detecting Tuberculosis

A systematic review assessing the diagnostic accuracy of studies discriminating PTB and healthy cases has been published by Harris et al. [10] in 2019. Looking at the top 7 best performing deep learning models of this study (Table 2.1), one takes note of the very high detection accuracy of published work, identifying a strong relationship between candidate PTB images and available ground-truth labels. The area under the specificity versus sensitivity "ROC" curve (AUC[1], range $[0.0, 1.0]$) is used here as a figure of merit for performance

[1] AUC systematically refers to the area under the receiver operating characteristic curve (AUROC) throughout this report.

reporting. It is possible to observe AUC results ranging from 0.82 to an impressive 0.99 with relatively tight confidence intervals in some of the studies [22, 23], for which estimation is not described. Hwang et al. [22] encouragingly indicate that these results would be comparable to the accuracy of trained thoracic radiologists in detecting PTB simply using CXR images. Most of the entries report results on private datasets, and none provide source code, which makes reproducibility impossible to achieve.

Table 2.1: Results of the top 7 models, from Harris et al. [10] systematic review, classifying PTB vs healthy cases solely using CXR. We notice impressive AUC with relatively tight confidence intervals indicating a strong relationship between PTB images and ground-truth labels. Datasets references are available in the original study but most are private.

Authors	**Datasets used for training**	**Number of CXRs used for training**	**Datasets used for testing**	**Number of CXRs used for testing**	**AUC (95% CI)**
Heo et al. [24]	YU AWHE	2000	YU AWHE	37475	0.91, 0.92
Hwang et al. [22]	SNUH	60989	SNUH, BMC, KUHG, DEMC, MC, CH	NR	0.988 (0.976-0.999)
Lakhani et al. [23]	MC, CH, TJH, Belarus	857	MC, CH, TJ, Belarus	150	0.99 (0.96-1.00)
Santosh et al. [25]	MC, CH, IN	976	MC, CH, IN	976	0.92 (MC), 0.82 (CH), 0.96 (IN)
Lopes et al. [26]	NR	NR	CHMC, CI, NR	1031	0.834 (CH), 0.926 (MC)
Santosh et al. [27]	NR	NR	CHMC, CI	878	0.93 (CH), 0.88 (MC)
Hwang et al. [28]	KIT	9221	KIT, MC, CH	2427	0.96

To the best of our knowledge, the only fully reproducible PTB-specific CAD study has been published by Pasa et al. [29]. The authors propose a simple CNN optimized for deployment in mobile settings, evaluating its performance on the publicly available Montgomery County (from now on referred to as "MC") and Shenzhen ("CH") datasets [18]. In this work, the test AUC reaches 0.811 for MC, 0.9 for CH, and 0.925 when both datasets are combined (Figure 2.1). The higher AUC on the combined dataset is likely due to a better generalization of the model when trained on more images. Given the small and efficient DL architecture proposed in that work, results are notable. Yet, considering the limited number of model parameters and the small amount of CXR images used for training, it is not possible to exclude a poor generalization on other data. Indeed, no cross-dataset analysis has been performed in this study.

The algorithm used by Pasa et al. is also able to generate either grad-CAMs or saliency maps to visualize where TB has been identified on the image, as illustrated in Figure 2.2. However, it is not clear whether the authors were able to verify the validity of these results as the datasets used do not include any ground-truth.

More recently, another DL model has been applied to the special case of PTB detection

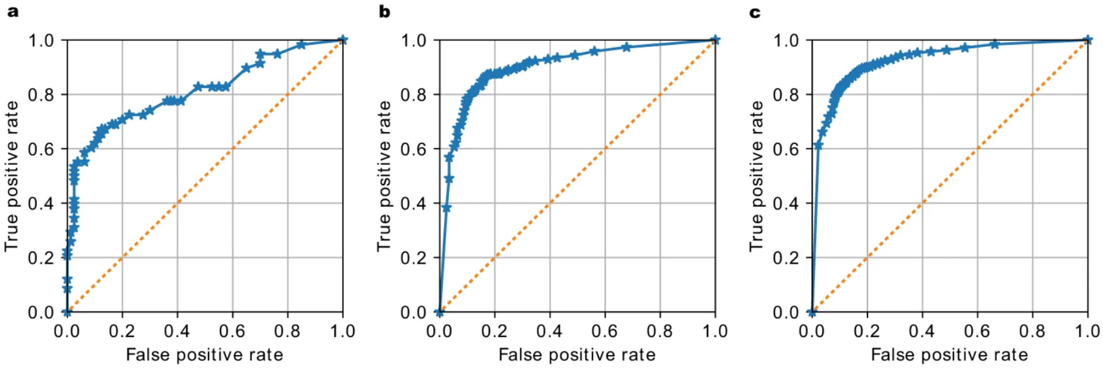


Figure 2.1: Receiver Operating Characteristic (ROC) curves from Pasa et al. [29]. Corresponding AUC: (a) 0.811 for the Montgomery dataset, (b) 0.9 for the Shenzhen dataset, and (c) 0.925 for the combined dataset. It is remarkable that the AUC is higher on the combined dataset.

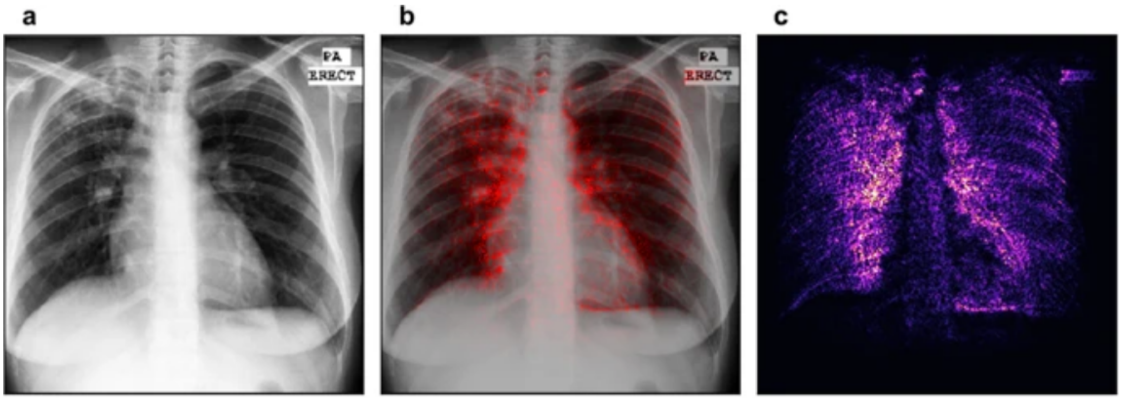


Figure 2.2: Saliency map with overlay for one correctly classified case. Panel (a) shows the chest image of the patient, panel (c) shows the saliency map, while panel (b) shows the saliency map overlaid on the chest image for comparison. From Pasa et al. [29].

in HIV-positive patients by Rajpurkar et al. [30]. The development of TB in HIV-positive patients is indeed specific as TB injury is a result of the immune response of the host, which is likely impaired in this group of patients [31]. Radiological signs may be thus atypical, with many patients presenting no CXR alterations. This work makes use of six radiological signs and clinical covariates among which the age, oxygen saturation, and the patient's prior TB history. Their encouraging results suggest that the use of a DL assistant improves the diagnostic capabilities of radiologists and confirm that radiological signs can provide an effective basis for TB identification. Datasets used in this study are private, making this work difficult to reproduce.

2.2 Detecting Radiological Signs

In a more general context, other studies [32, 33, 34] highlight the ability of DL models to extract various radiological signs from images. In particular, the CheXNeXt study by Rajpurkar et al. [35] presents a model concurrently detecting 14 clinically important radiological findings, using an ensemble of dense DL models. Trained on the NIH CXR14 dataset [32], this model is able to predict if each of the 14 signs is present on input images, with a performance similar to trained radiologists (Table 2.2). Although this is not the first study where DenseNet models are used to predict radiological signs [36, 37, 38], it is the first where the performance of the model is directly compared with that of radiologists. Moreover, the code as well as the training and validation datasets used are publicly available. Unfortunately, this is not the case for the test data on which AUCs were calculated.

Table 2.2: AUC results from CheXNeXt [35]. For ten radiological findings, the model is able to give equivalent performances while giving better results for one finding and worse results for three others respectively. We observe that some radiological findings seem easier to detect for both radiologists and CheXNeXt.

Radiological sign	**Radiologists (95% CI)**	**Algorithm (95% CI)**	**Algorithm - Radiologists Difference (99.6% CI)**	**Advantage**
Atelectasis	0.808 (0.777 to 0.838)	0.862 (0.825 to 0.895)	0.053 (0.003 to 0.101)	Algorithm
Cardiomegaly	0.888 (0.863 to 0.910)	0.831 (0.790 to 0.870)	-0.057 (-0.113 to -0.007)	Radiologists
Consolidation	0.841 (0.815 to 0.870)	0.893 (0.859 to 0.924)	0.052 (-0.001 to 0.101)	No difference
Edema	0.910 (0.886 to 0.930)	0.924 (0.886 to 0.955)	0.015 (-0.038 to 0.60)	No difference
Effusion	0.900 (0.876 to 0.921)	0.901 (0.868 to 0.930)	0.000 (-0.042 to 0.040)	No difference
Emphysema	0.911 (0.866 to 0.947)	0.704 (0.567 to 0.833)	-0.208 (-0.508 to -0.003)	Radiologists
Fibrosis	0.897 (0.840 to 0.936)	0.806 (0.719 to 0.884)	-0.091 (-0.198 to 0.016)	No difference
Hernia	0.985 (0.974 to 0.991)	0.851 (0.785 to 0.909)	-0.133 (-0.236 to -0.055)	Radiologists
Infiltration	0.734 (0.688 to 0.779)	0.721 (0.651 to 0.786)	-0.013 (-0.107 to 0.067)	No difference
Mass	0.886 (0.856 to 0.913)	0.909 (0.864 to 0.948)	0.024 (-0.041 to 0.080)	No difference
Nodule	0.899 (0.869 to 0.924)	0.894 (0.853 to 0.930)	-0.005 (-0.058 to 0.044)	No difference
Pleural thickening	0.779 (0.740 to 0.809)	0.798 (0.744 to 0.849)	0.019 (-0.056 to 0.094)	No difference
Pneumonia	0.823 (0.779 to 0.856)	0.851 (0.781 to 0.911)	0.028 (-0.087 to 0.125)	No difference
Pneumothorax	0.940 (0.912 to 0.962)	0.944 (0.915 to 0.969)	0.004 (-0.040 to 0.051)	No difference

In order to take into account the labeling errors of the dataset, the authors first trained multiple 121-layer DenseNet models [36] on the training subset to predict the 14 radiological signs. A selection of those models, based on their performance on the validation subset, was then used to build an ensemble producing predictions by computing the mean over the predictions of each individual network. This ensemble was subsequently used to relabel the training and validation subsets in the following way: the label for each radiological sign was defined as positive if the prediction of the ensemble or the original label was positive. Next, they trained new networks on the relabeled training subset and selected the 10 best ones according to their average error on the relabeled validation subset to generate predictions on the test subset.

While multiple variables, among which patient's history, age, and environment, are usually taken into account to establish a diagnosis, in the CheXNeXt study only the image was used by the model and the radiologists. In a real context, with access to this information, performances are expected to improve.

CheXNeXt also includes the ability to generate an overlay, through the use of class activation mappings (grad-CAMs), highlighting parts of the image most indicative of each predicted finding (Figure 2.3).

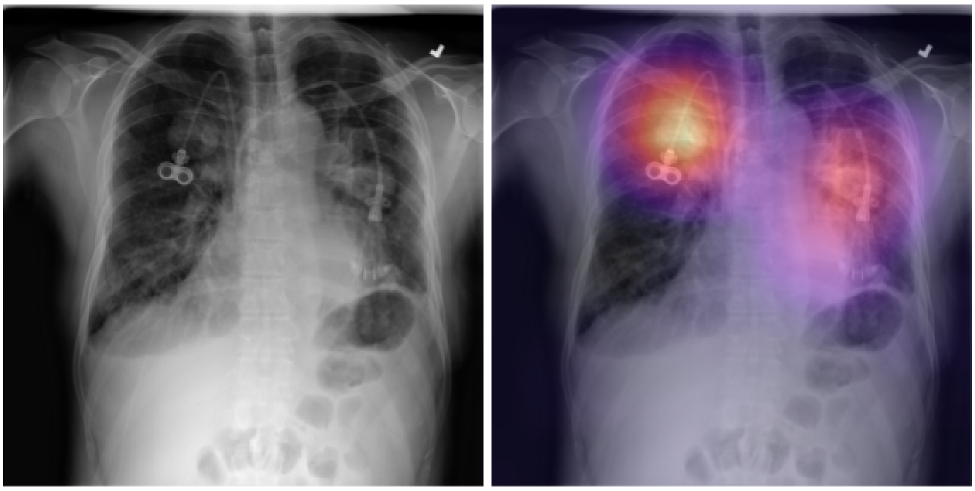


Figure 2.3: Example of a CheXNeXT prediction. Original patient's chest radiograph image (left) and 2 upper-lobe pulmonary masses with both right and left-sided central venous catheter highlighted on the same image (right) [35].

2.3 Detecting Tuberculosis Using Radiological Signs

As we have seen, active PTB is typically detected on CXR images using binary labels (direct CAD) [10, 29] although one study has used radiological signs along with clinical covariates in the special case of TB detection in HIV-positive patients [30]. To the best of our knowledge, this last study is the only existing one proposing to use radiological signs as a basis for indirect disease detection. Other models for CXR computer-aided diagnosis (i.e. CheXNeXt) currently focus on radiological findings identification [35].

While clinical covariates are important for final diagnosis, we hypothesize that a modular indirect ML model efficiently detecting and localizing PTB-related radiological signs can be equivalently efficient in detecting PTB, solely considering CXR images. We notice that such a detector would be naturally interpretable and immediately relatable to healthcare workers, given the workflow similarities. In what follows, we conduct a series of experiments to validate this hypothesis.

2.4 Reproducibility With The Bob Framework

As we have seen in our literature review, most PTB detection research is not reproducible as no source code or precise algorithm description is usually made available. To solve this problem, we implemented our experiments using the Bob framework methodology [39, 40]. Bob is a free signal-processing and machine learning toolbox designed to facilitate continuous reproducibility of data science related projects. It is designed to be efficient and to reduce development time. A typical workflow in machine learning and pattern recognition is illustrated in Figure 2.4.

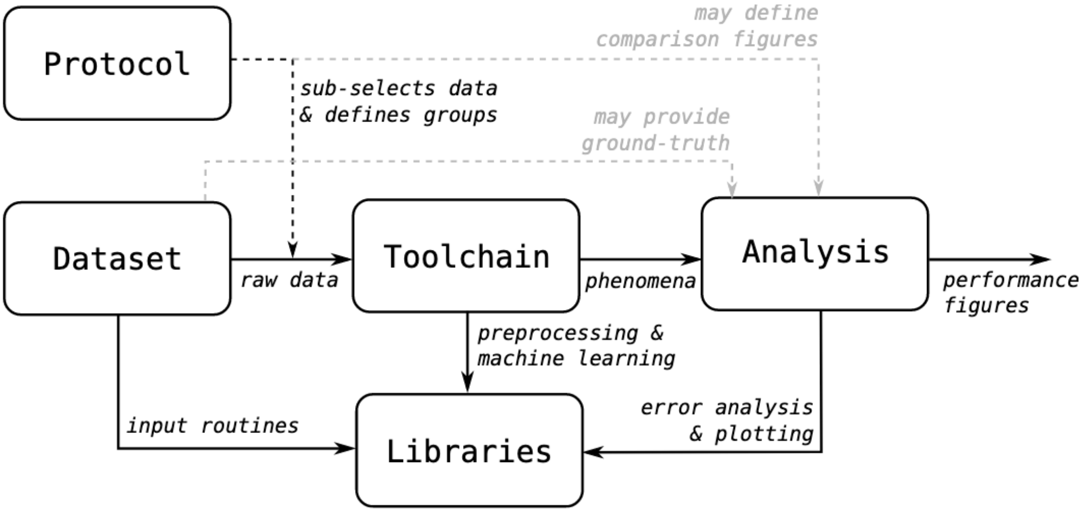


Figure 2.4: Typical workflow in machine learning and pattern recognition. Data protocols, pre-processing steps, and all experiments and analysis are standardized in order to be reproducible. From Bob framework [40].

Thanks to this toolkit and to the use of publicly available datasets, we have standardized data treatments as well as our experiments so that they can be reproduced by anyone.

We also took special care in implementing both unit and integration tests. Unit tests verify the functionality of particular sections of code, such as specific functions. While integration tests allow us to check command lines as if a user was executing them. Examples of this second class of tests are the execution of a training epoch of one of our models or the prediction of PTB on a dataset. All tests cover more than 93% of the code and their execution is automated with Gitlab's continuous integration so that they are systematically executed whenever the code is modified.

In terms of documentation, we have taken care to describe the use of each command available in our package, to list the main results as well as to describe precisely the process of model optimization and the required resources. As we do not provide the datasets themselves, links to their download locations are provided alongside their descriptions. The list of possible commands and of all their parameters is automatically created, as is the API of the package. When the code is modified on Gitlab, all the documentation is automatically generated to keep it up to date.

The package allowing the execution of all the experiments described in this report is available on the GitLab of the Idiap Research Institute[2], along with the documentation.

[2] https://gitlab.idiap.ch/bob/bob.med.tb

3. Direct Detection Baselines

To compare results of the proposed indirect detection algorithm, we introduce three public datasets, our evaluation protocol and metrics, as well as two baselines for the direct detection method, that in our understanding, are representative of the accessible state-of-the-art.

3.1 Active Pulmonary Tuberculosis Datasets

At the time of writing, four frontal CXR datasets containing postero-anterior (PA) images and labels for active PTB classification as well as a larger additional dataset featuring a subset of PTB cases are publicly available: MC, CH, Indian Collection ("IN"), NIAID TB and PadChest [18, 19, 41, 42]. Given the limited amount time available for this study, our work is focused on the first three datasets (MC, CH, and IN) for which a compact overview is provided in Table 3.1.

Table 3.1: Overview of the three publicly available PTB datasets we will use in the present study. We notice that all datasets contain information about the final diagnosis but no radiological signs annotations.

	TB cases	**Normal cases**	**Resolution**	**Annotations**
Montgomery County [18]	58	80	4020×4892 or 4892×4020	Final diagnosis, Region segmentation, and lung masks
Shenzhen [18]	336	326	Varying dimensions from 948 to 3001 pixels	Final diagnosis and Region segmentation
Indian collection (New Delhi) [19]	78	77	1024×1024 or more	Final diagnosis

Within these datasets, CXR images are divided into normal (i.e healthy lungs) and active PTB cases depending on the final diagnosis (samples in Figure 3.1), which is based on the results of standard skin and sputum tests for the MC and CH datasets (no information is available for IN). Active PTB cases are also commonly referred to as "positive cases" in the medical literature, whereas normal cases are known as "negative cases".

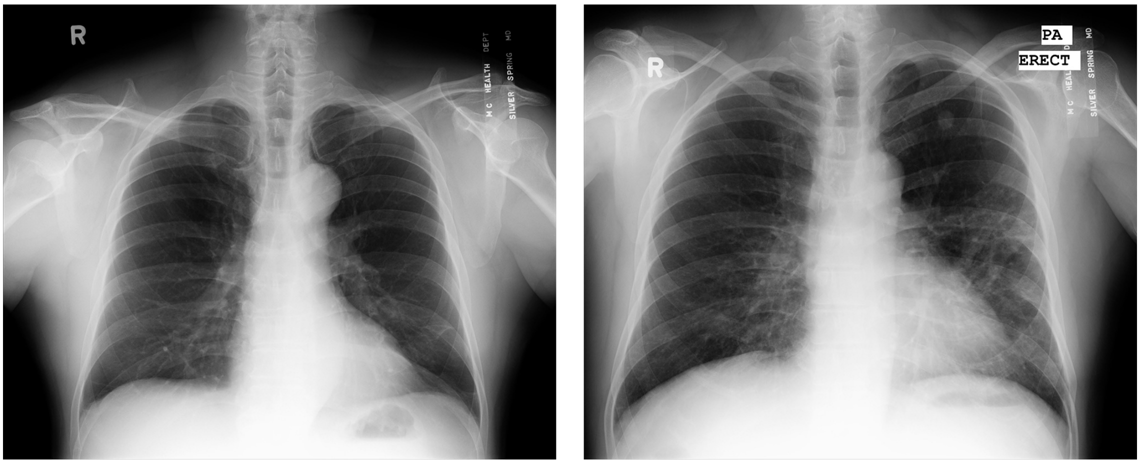


Figure 3.1: Publicly available frontal CXR datasets typically pose the diagnosis of PTB as a binary classification problem between healthy and PTB patients. Two samples from the Montgomery County dataset [18] are displayed here (left: healthy patient, right: active PTB case).

We briefly introduce each of these three datasets in the present section. Each of them was first separated into three subsets (training/validation/testing), the details of which are provided in this section, so that the best model hyperparameters could be easily identified using a grid search. We then applied a stratified k-fold cross-validation, presented in detail in Section 3.1.2, on each dataset to improve the reliability of our final results.

Montgomery County [18] The Montgomery County (MC) set is made available by the U.S. National Library of Medicine and contains radiographs from Montgomery County's Tuberculosis screening program. It is composed of 138 frontal CXR images, including 58 TB cases. They are provided as 8-bits greyscale PNG files and resolution of images is either 4020×4892 or 4892×4020 pixels.

The group to which each CXR image belongs is indicated by the end of the file name: 1 for TB, 0 for healthy. Moreover, text files with gender and age of the patients and scan's diagnosis information are supplied. A typical reading has the following form:

Patient's Gender: F
Patient's Age: 031Y
cavitary nodular infiltrate in RUL; active TB

Some clinical readings include details about radiological findings and their position on the lungs, but not all.

The average patient's age of this dataset is 40.11 years old and CXR images come from 73 women, 61 men, and 1 unspecified. Each radiograph corresponds to a single patient, except 4 which come from 2 patients.

Additionally, the Montgomery dataset contains manually segmented lung masks (sample in Figure 3.2) usable for the evaluation of automatic segmentation methods.

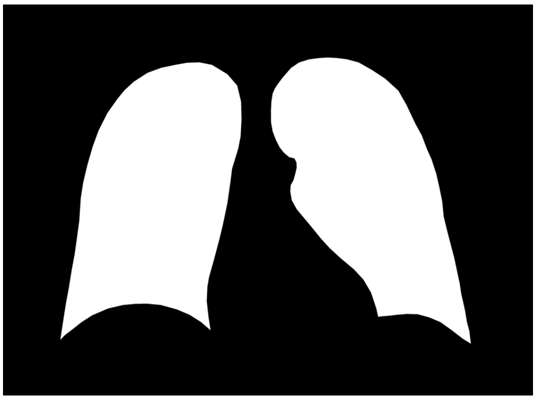


Figure 3.2: Lung masks from the Montgomery County dataset [18].

To identify the best model hyperparameters, we have randomly separated the dataset in a training, a validation, and a testing set as detailed in Table 3.2. The list of images included in each subset can be found in the code.

Table 3.2: Overview of the number of samples in the different subsets randomly created from the Montgomery dataset for the hyperparameters optimization.

	Training n (%)	**Validation n (%)**	**Testing n (%)**
Number of Positive Cases	37 (42%)	9 (41%)	12 (43%)
Number of Negative Cases	51 (58%)	13 (59%)	16 (57%)
Total Cases	88	22	28

Shenzhen [18] The Shenzhen (CH) set is also made available by the U.S. National Library of Medicine and contains radiographs collected at Shenzhen No.3 People's Hospital (China). Of the 662 frontal CXR images, 326 represent healthy patients and 336 are cases with manifestations of TB. Scans are provided as 8-bits RGB PNG files with varying resolution (width and height from 948 to 3001 pixels).

The group (healthy or TB, samples in Figure 3.3) to which each CXR image belongs is indicated by the end of the file name, similarly to the MC dataset. Patient's information and diagnosis are also available in a similar form.

The average patient's age is 35.43 years old and CXR images come from 449 women and 213 men (one radiograph per patient).

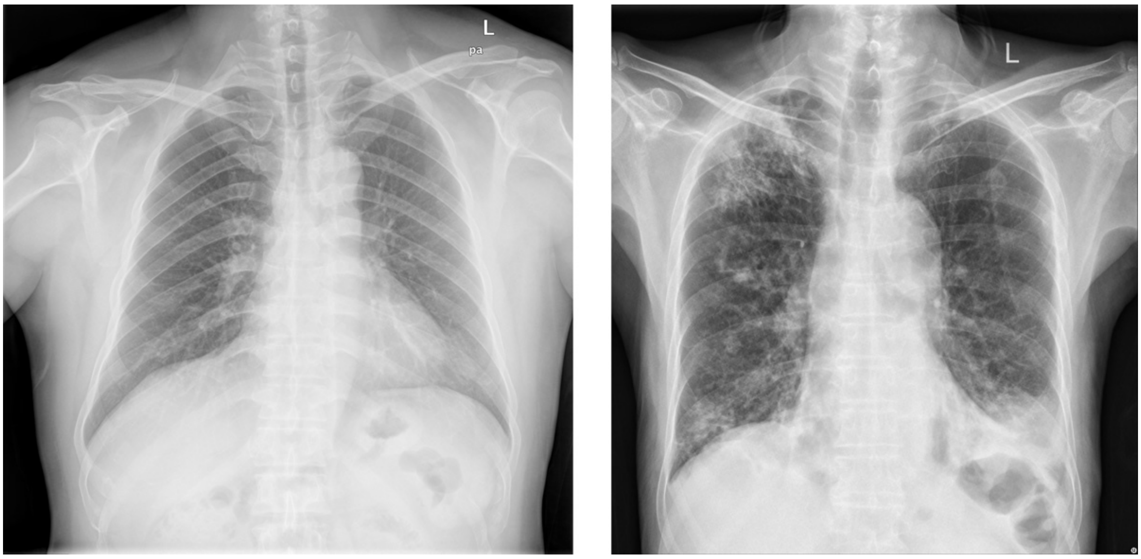


Figure 3.3: Two samples from the Shenzhen dataset [18] (left: healthy patient, right: active PTB case).

To identify the best model hyperparameters, we have randomly separated the dataset in a training, a validation, and a testing set as detailed in Table 3.3.

Table 3.3: Overview of the number of samples in the different subsets randomly created from the Shenzhen dataset for the hyperparameters optimization.

	Training n (%)	**Validation n (%)**	**Testing n (%)**
Number of Positive Cases	215 (51%)	54 (50%)	67 (50%)
Number of Negative Cases	207 (49%)	53 (50%)	66 (50%)
Total Cases	422	107	133

Indian collection [19] The Indian Collection (IN) set has been obtained from two different CXR machines at the National Institute of Tuberculosis and Respiratory Diseases of New Delhi (India). It contains 155 CXR images of which 78 are healthy patients and 77 are TB cases. Radiographs are provided as 8-bits greyscale JPEG files of at least 1024×1024 pixels resolution.

Like previously cited datasets, files are binary classified between normal (healthy patients) and TB cases. The category to which an image belongs is indicated in the file name: "nx" for normal, "px" for TB. Two samples are displayed in Figure 3.4.

Unfortunately, no additional patients metadata is provided with this set.

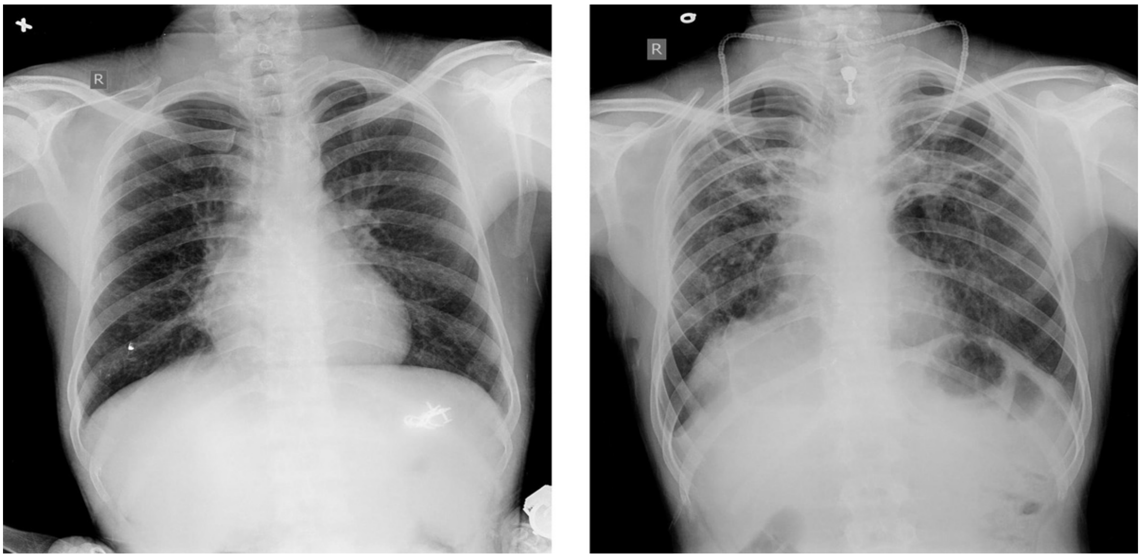


Figure 3.4: Two samples from the Indian Collection dataset [19] (left: healthy patient, right: active PTB case).

This dataset is already divided into training and testing subsets of 103 and 52 CXR respectively. To identify the best model hyperparameters, we kept the provided train-test split, but further separated 20% of the training subset to form a validation one as indicated in Table 3.4.

Table 3.4: Overview of the number of samples in the different subsets of the Indian Collection dataset for the hyperparameters optimization.

	Training n (%)	**Validation n (%)**	**Testing n (%)**
Number of Positive Cases	42 (50%)	10 (50%)	26 (50%)
Number of Negative Cases	41 (50%)	10 (50%)	26 (50%)
Total Cases	83	20	52

3.1.1 Aggregated Datasets

To perform a cross-dataset analysis, we decided to train the models progressively on increasingly more data to see the evolution of prediction performance on the test subsets of the MC, CH, and IN datasets, as explained in Section 3.3. To do so, we created two new datasets (MC-CH and MC-CH-IN) by aggregating the subsets of our three original datasets as shown in Figure 3.5. The training subset of the MC-CH dataset is, for instance, created by aggregating the training subsets of the MC and CH datasets together.

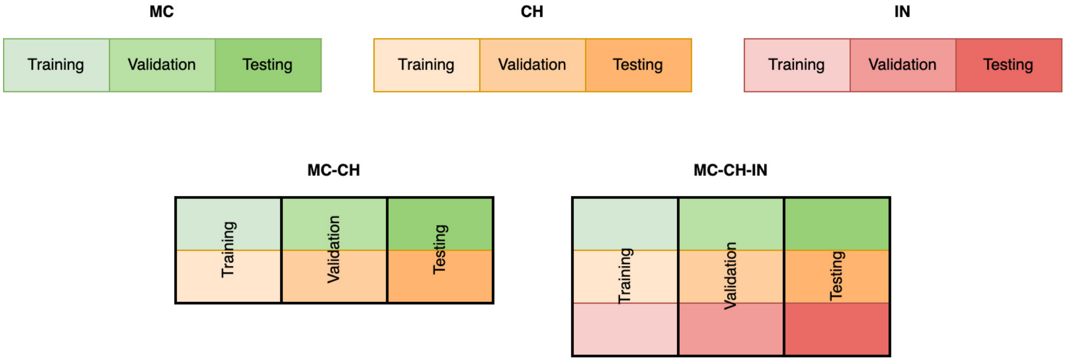


Figure 3.5: The MC-CH and MC-CH-IN datasets were created by aggregating the subsets of the MC, CH, and IN datasets together.

3.1.2 Cross-Validation

Cross-validation is a statistical method commonly used to assess the performance of ML models in the presence of a limited amount of data [29].

For instance, the Montgomery dataset is composed of 138 images and its testing subset of only 28. Since these 28 images were randomly selected, they may not be representative of the whole dataset and this could induce a bias in the performance evaluation of our models. Furthermore, the correct or incorrect PTB identification on a single additional image will greatly influence the percentage of success of the model given the limited number of images in the testing subset.

To address this problem, we applied stratified k-fold cross-validation with a k equal to 10. This means that from our initial dataset we created 10 new sets with a different training/validation/testing separation, as illustrated in Figure 3.6. Thus, each of the 10 testing subsets contains one tenth of the total number of images and together they represent the entire initial dataset. Instead of training a single model on our dataset, we trained 10 times each model on the 10 training subsets separately and predicted the PTB on the 10 testing subsets. In this way, we evaluated the performance of the model on all images in the dataset while avoiding a selection bias.

It is important to note that the "stratified" adjective indicates that we systematically used the same proportion of positive and negative cases as in the original dataset within each of the subsets created for every 10 new sets during this procedure.

We thus used cross-validation for all PTB datasets when conducting the experiments in this work. The list of samples included in each of the folds is available in the code.

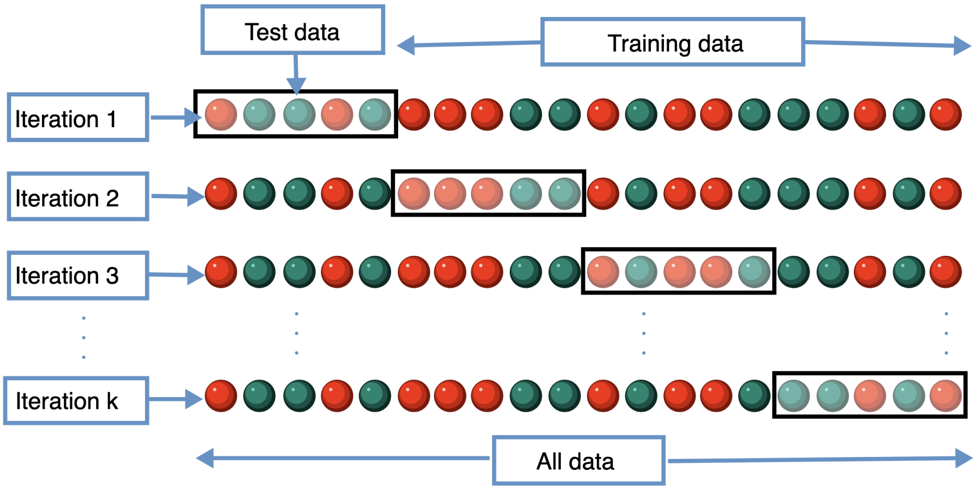


Figure 3.6: Illustration of a k-fold cross-validation from [43].

3.1.3 Data Augmentation

No specific image pre-processing is required to use radiographs as they are typically treated identically to conventional images. However, to improve resilience with variable quality CXR images and to prevent overfitting, data augmentation is frequently implemented [29, 35, 24, 22].

Data augmentation is a powerful technique used to generate artificial variations in existing datasets. The range of possible data augmentations includes transformation techniques like rotation, cropping, random contrast degradation, addition of noise, and more advanced techniques like elastic deformation (illustrated in Figure 3.7).

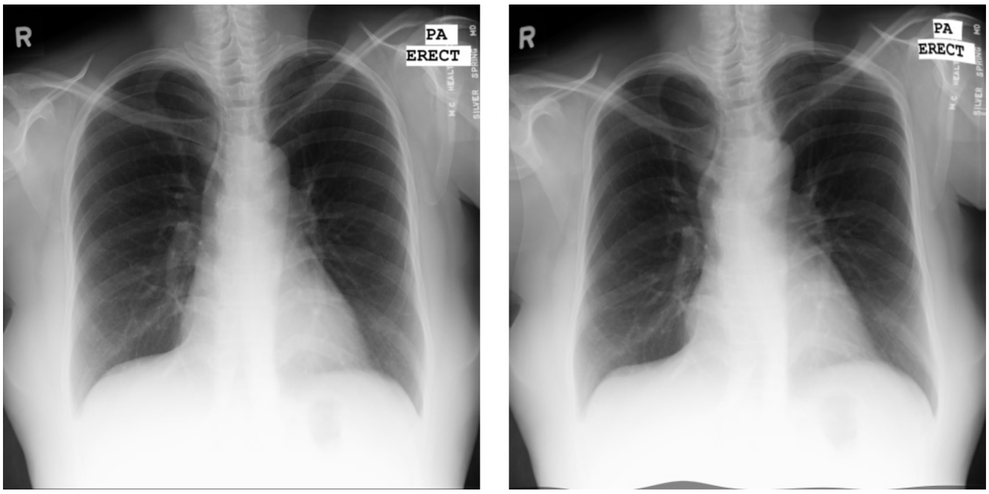


Figure 3.7: An original CXR image alongside the same image that has undergone elastic deformation as part of a data augmentation process. The use of data augmentation techniques improves the resilience of machine learning models.

In our case, data augmentation techniques need to take into account probable variations of CXR images. Indeed, radiographs are generated on a wide range of differently configured machines and could look slightly different on each of them. We choose to rely on the elastic deformation technique used in related work [29].

All images from PTB datasets will thus be pre-processed to remove black contours, resized to a resolution of 512 x 512 pixels, and be subject to data augmentation via elastic deformation with a probability of 80% before going through the ML models, like in [29], as illustrated in Figure 3.8.

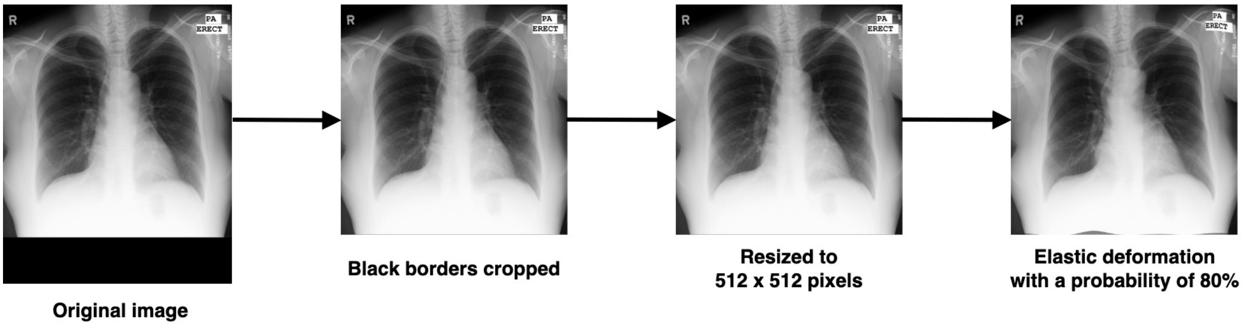


Figure 3.8: Illustration of the pre-processing pipeline applied to all images from PTB datasets.

An image standardisation step has also been implemented in order to facilitate models optimisation:

$$x_{stand} = \frac{x - mean(x)}{standard\ deviation(x)} \tag{3.1}$$

The latter has been directly incorporated into all models so that the mean and standard deviation are automatically calculated and stored as parameters when the training subset is fed in.

3.2 Models

To establish a performance baseline for the direct detection method, we assessed the model proposed by Pasa et al. [29] as well as a DenseNet-121 network [36] on the aforementioned datasets and splits.

3.2.1 The Pasa Model

For the first baseline, we selected the model proposed by Pasa et al. [29] because it is the only open source model which is both specifically dedicated to PTB detection and evaluated on public datasets. It is a small model (201'905 parameters) built to be trained on CXR images annotated with binary labels and deployed in mobile settings. It is made of 5 convolutional-blocks followed by a global average pooling and a fully-connected layer with two outputs. Each convolutional-block is composed of two 3×3 convolutions (stride of 2) with ReLUs, one 1×1 parallel convolution and a 3×3 max-pooling operation (stride of 2). Additionally, batch normalization is used after each convolution to improve training performance and prevent overfitting. Following the original code, we implemented the model in our own framework. While the original implementation has two output nodes, we decided to use only one for having a code compatible with multi-class classifiers. As a consequence, we used a Binary Cross-Entropy loss in place of the original Cross-Entropy loss.

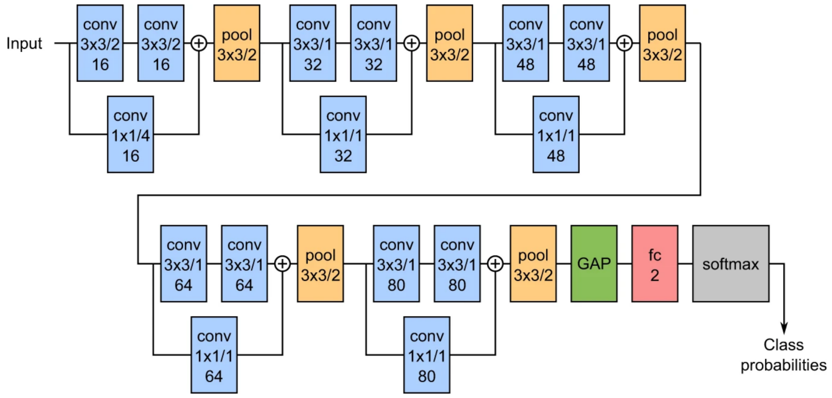


Figure 3.9: Schematic representation of the network architecture proposed by Pasa et al. [29].

We retained the hyperparameters for training the model proposed by Pasa et al. [29]: batch size of 4, learning rate of 8×10^{-5}. The evolution of losses when the model is trained on the MC-CH-IN dataset with these parameters is illustrated in Figure 3.10. We notice a high variability in the validation loss which is probably due to variations of images coming from three distinct datasets. Training has been conducted with a Binary Cross-Entropy loss for 500 epochs via an Adam optimizer with the default parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$. The checkpoints with the lowest validation losses have been kept for evaluation.

3.2.2 The Densenet-121 Model

For our second baseline, we chose a Densenet-121 model (7'216'513 parameters) [36]. Our motivation to use this network stem from the fact that various studies on our review used relatively dense models in such a task with promising results [23, 26]. Moreover, DenseNets encompass various regularization mechanisms (skip connections and batch normalization) that are important in image classification tasks.

A grid search was conducted on the aggregated MC-CH-IN dataset to determine the best hyperparameters for training this model, as shown in Table 3.5: learning rate of 5×10^{-5}, batch size of 8, and the default Adam optimizer parameters. We used a maximum batch size

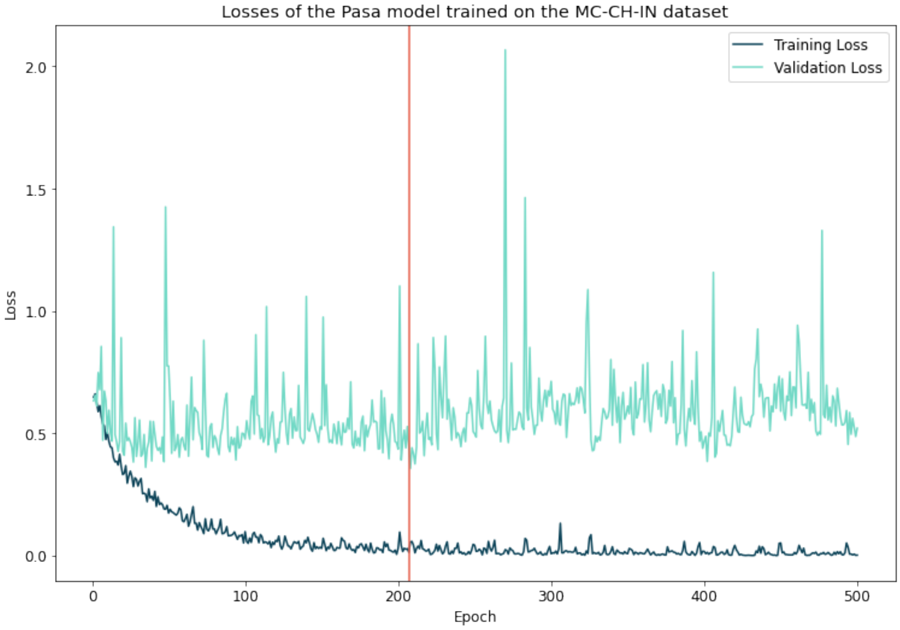


Figure 3.10: Losses evolution of the Pasa model trained with hyperparameters proposed by Pasa et al. [29] on the MC-CH-IN dataset. The vertical red line indicates the location of the lowest validation loss.

of 8 as the available memory did not allow loading more images at once. The evolution of losses when the model is trained on the MC-CH-IN dataset with the best hyperparameters is illustrated in Figure 3.11 alongside another losses evolution with a smaller learning rate. In both cases, the lowest validation loss is quickly reached and we notice the same high variability as with the Pasa model. That a lower learning rate does not reduce the loss and that we reach the lowest point so quickly indicates that the variability of the data in the three datasets and their small quantity leads to difficulties for the model optimization. To generate the final results, we trained the model from random initialization for 2000 epochs and the checkpoints with the lowest validation Binary Cross-Entropy losses have been retained for the evaluation.

Table 3.5: Results of the grid search conducted on the aggregated MC-CH-IN dataset to identify the best hyperparameters for the Densenet-121 model. This table indicates the minimum validation loss obtained for each combination of learning rate and batch size.

Learning rate	**Batch size: 4**	**Batch size: 8**
1×10^{-4} (training for 600 epochs)	0.3658	0.3676
5×10^{-5} (training for 150 epochs)	0.3490	**0.3168**
1×10^{-5} (training for 1000 epochs)	0.3791	0.3831

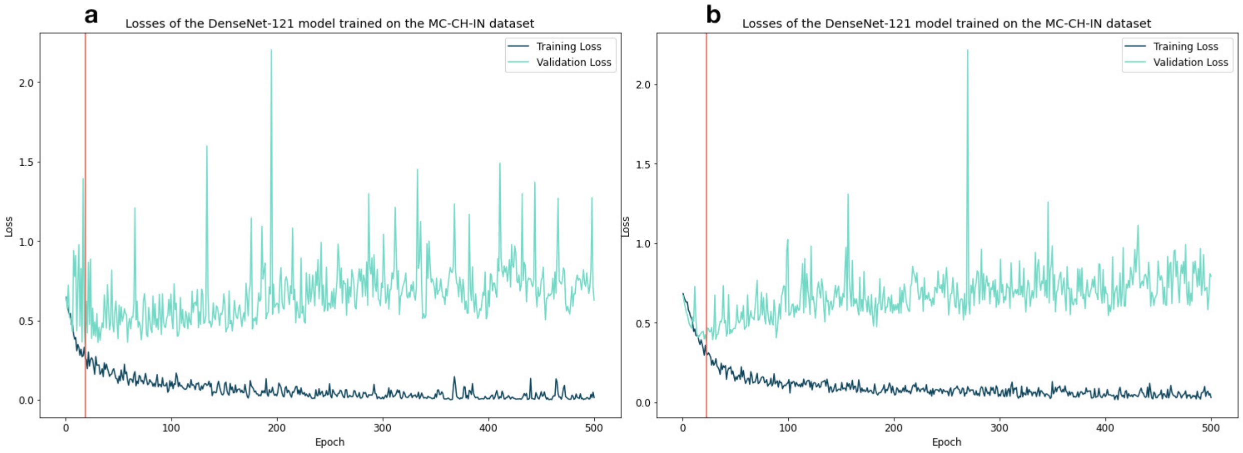


Figure 3.11: Losses evolution of the DenseNet-121 model trained with a batch size of 8 and a learning rate of a) $5e^{-5}$ (leading to the minimum validation loss) or b) $1e^{-5}$ on the MC-CH-IN dataset. The vertical red line indicates the location of the lowest validation loss.

3.3 Evaluation Protocol

Each model has been trained in three different scenarios, using an increasing amount of datasets in the training mix: MC only, MC+CH, and MC+CH+IN. This setup allowed the evaluation of generalization as more data is gradually used to train model parameters. The threshold used to classify a CXR image as PTB-positive or healthy in the test subset has been systematically selected as being the threshold giving the best F1-score on the validation subset of the dataset on which the model has been trained. Since we made use of cross-validation, our thresholds correspond to the average of the 10 thresholds identified on the validation subsets, as illustrated in Figure 3.12. Table 3.6 shows the respective thresholds used.

To avoid biases towards any of the datasets, a data balancing mechanism has been implemented to ensure feedback loss is not dominated by datasets with more samples. More precisely, this mechanism consists of sampling with replacement an equal number of samples in each of the datasets whatever their size. The same mechanism is applied to ensure that an equal number of PTB and healthy cases is being fed to the models.

We subsequently evaluated each of the three models on the individual test subsets of MC, CH, and IN. Since we used cross-validation, we employed each of the 10 models to predict the presence of PTB on the corresponding 10 testing subsets of each dataset. We then aggregated these predictions, for each dataset, to perform the evaluation.

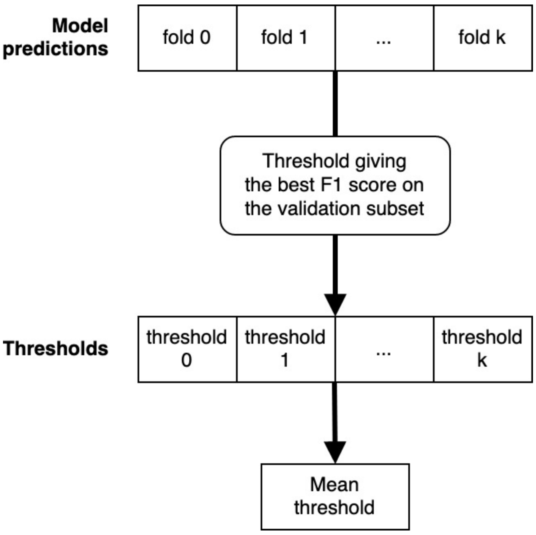


Figure 3.12: Calculation process of the optimal threshold to evaluate predictions on tests subsets during cross-validation.

Table 3.6: Thresholds used to classify a CXR image as PTB-positive or healthy for our two baselines. The threshold has been identified by maximizing the F1-score on the validation subset of the dataset on which the model has been trained. The threshold indicated here is an average of 10 thresholds as we made use of cross-validation.

Model	**Threshold**
Pasa (train: MC)	0.5057
Pasa (train MC+CH)	0.4966
Pasa (train: MC+CH+IN)	0.4135
DenseNet-121 (train: MC)	0.5183
DenseNet-121 (train: MC+CH)	0.2555
DenseNet-121 (train: MC+CH+IN)	0.4037

3.3.1 Measures

To be able to compare our results to state-of-the-art research, we mainly use the Area Under the Receiver Operating Characteristic Curve (AUROC or AUC) to evaluate the performance of our models on the various test subsets. The AUC is therefore introduced in this section alongside the score distribution, which allows for a better understanding of it. Additionally, we present the F1 score which we used in the threshold optimization process, as well as the precision and the recall measures.

Score Distribution

For binary labels, the score distribution plots the probability distribution of positive (PTB cases) and negative (healthy cases) classes with the model output on the x-axis and normalized counts on the y-axis. The score distribution in Figure 3.13 illustrates a situation in which the model does not perfectly separate positive (green curve) and negative (red curve) cases. We can observe the presence of an overlap containing both false positives and false negatives. Indeed, each model prediction can be classified into one of the four categories of the confusion matrix (Figure 3.14): true positive, true negative, false positive, or false negative. Depending on the chosen threshold, more items will be classified as positive and the number of both true and false positives will increase simultaneously. The score distribution of a perfect model will contain two non-overlapping curves while that of a bad model will display totally superimposed ones.

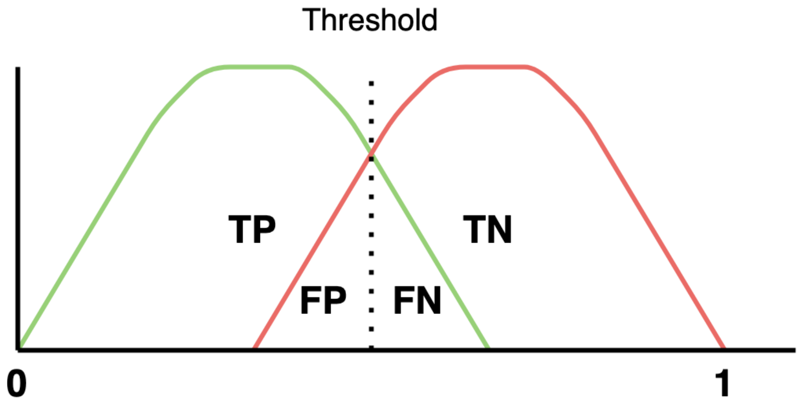


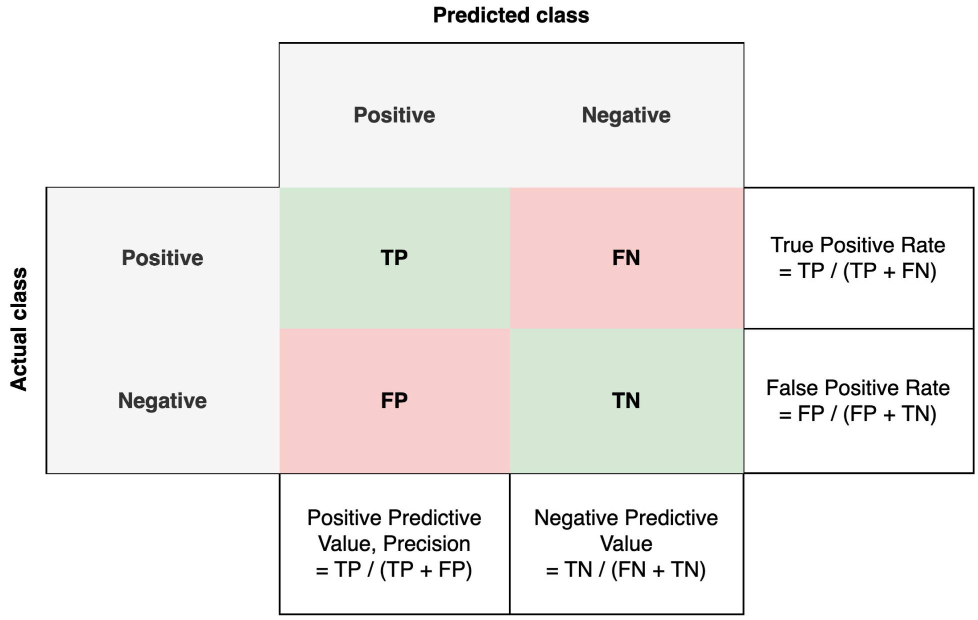
Figure 3.13: The score distribution of an imperfect model: the positive (green curve) and negative (red curve) cases are not perfectly separated, introducing false positives and false negatives. The green and red curves of a perfect model would not overlap.

Area Under the Curve

One of the most widely used measures is the Area Under the Receiver Operating Characteristic Curve (AUROC or AUC) [10, 11, 29]. It measures the percentage of area underneath the entire receiver operating characteristic curve (ROC). The latter is simply the summary of the score distribution, plotting true positive rate against false positive rate. Therefore, the AUC shows us the classification model performance across all classification thresholds (Figure 3.15).

$$TPR\ (True\ Positive\ Rate) \;=\; \frac{TP}{TP+FN}$$

$$FPR\ (False\ Positive\ Rate) \;=\; \frac{FP}{FP+TN}$$



		Predicted class		
		Positive	Negative	
Actual class	Positive	TP	FN	True Positive Rate = TP / (TP + FN)
	Negative	FP	TN	False Positive Rate = FP / (FP + TN)
		Positive Predictive Value, Precision = TP / (TP + FP)	Negative Predictive Value = TN / (FN + TN)	

Figure 3.14: Each model prediction can be classified into one of the four categories of the confusion matrix: true positive (TP), true negative (TN), false positive (FP), or false negative (FN). With a successful model, most of the predictions will be true positives or true negatives. Multiple rates can be computed from those four categories.

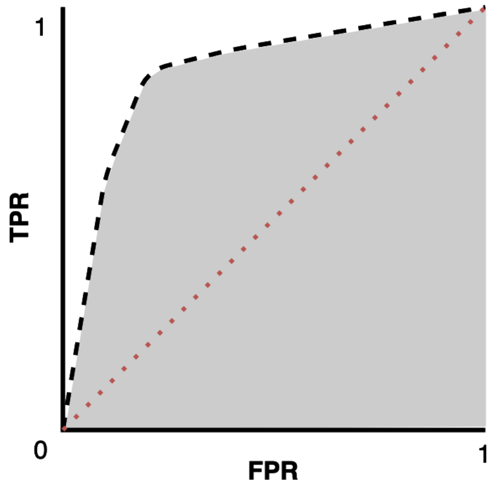


Figure 3.15: Area under the ROC curve. The AUC of a model having good prediction capabilities will be close to 1.0 (100% correct) while systematic misprediction will lead to 0.0 (100% wrong). A model predicting at random would give a diagonal (red dotted line).

Precision and Recall

Although it is not the most published evaluation diagram, the precision-recall curve is helpful when classes are imbalanced. In this case, the ROC curve is not perfectly representative of the performances [44]. However, since the state-of-the-art of PTB detection uses AUC, we choose not to use precision and recall as primary measures because our results would not be comparable. Nevertheless, we define precision and recall here so that we can introduce the F1 score below.

Precision tells us what proportion of the positive predictions were actually correct and is defined as follows:

$$Precision \;=\; \frac{TP}{TP + FP}$$

On the other hand, recall indicates what proportion of actual positive classes were identified correctly. It is the TPR from previous section.

$$Recall \;=\; \frac{TP}{TP + FN}$$

The precision-recall curve is the plot of those two parameters for each threshold. Again, improving one parameter degrades the other and vice versa.

F1 Score

To define a threshold classifying a sample as positive or negative according to the score predicted by the models, we optimized the F1 score on the validation subset of the dataset on which the model was trained.

The F1 score is defined as the harmonic mean of precision and recall. The maximum value it can take is 1.0, indicating perfect precision and recall, and the minimum value 0.0 when either precision or recall is 0.

$$F1\ score \;=\; 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

3.4 Results

Table 3.7 summarizes the cross-validated results of our baselines. With both models, a lack of generalization is observable and can be confirmed by looking at the corresponding ROCs presented in Figures 3.16 and 3.17. As a general rule, performances are consistent with reported values in the literature only after the dataset being tested was also seen during training. We also notice that the results of the DenseNet-121 model are slightly superior in most cases, probably due to the higher number of parameters.

It is important to keep in mind that even when the AUC is around 0.9, neither model indicates which radiological signs are present on the CXR image.

The inability to separate TB cases from healthy ones depending on the dataset used during training can also be seen thanks to the score distribution. We compare on Figure 3.18 the score distribution of the worst performing direct detection model (Pasa trained on MC, tested on CH) with that of the best one (DenseNet-121 trained on MC-CH-IN, tested on CH). We observe that the positive (TB) and negative (healthy) cases are well separated with the second model, whereas this is not the case at all for the first one.

Table 3.7: Baseline benchmark results: models are trained on the train subsets in parenthesis and tested on the three tests subsets using cross-validation. The test AUC is reported and the best result in each column is highlighted.

AUC	**MC test**	**CH test**	**IN test**
Pasa (train: MC)	**0.890**	0.576	0.642
Pasa (train MC+CH)	0.870	0.893	0.669
Pasa (train: MC+CH+IN)	0.881	0.898	0.848
DenseNet-121 (train: MC)	0.822	0.607	0.625
DenseNet-121 (train: MC+CH)	0.883	0.905	0.672
DenseNet-121 (train: MC+CH+IN)	0.860	**0.917**	**0.850**

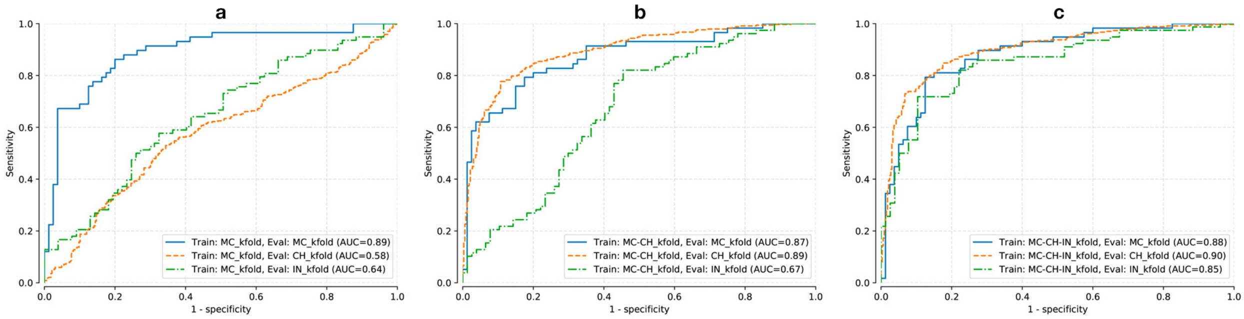


Figure 3.16: ROCs of our implementation of the Pasa model. The cross-validated predictions were generated by the model after training it on the (a) MC, (b) MC-CH, and (c) MC-CH-IN datasets. Performances are consistent with reported values in the literature only after the dataset being tested was also seen during training.

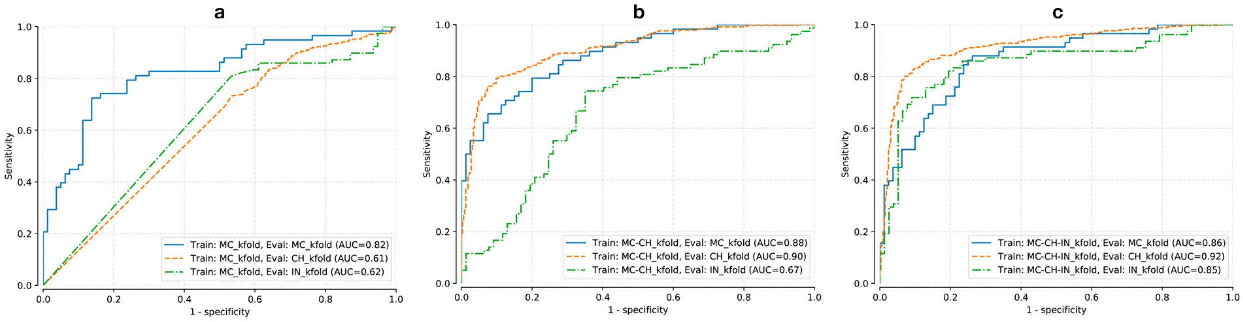


Figure 3.17: Cross-validated ROCs from the DenseNet-121 model trained on the (a) MC, (b) MC-CH, and (c) MC-CH-IN datasets. A similar lack of generalization is observable.

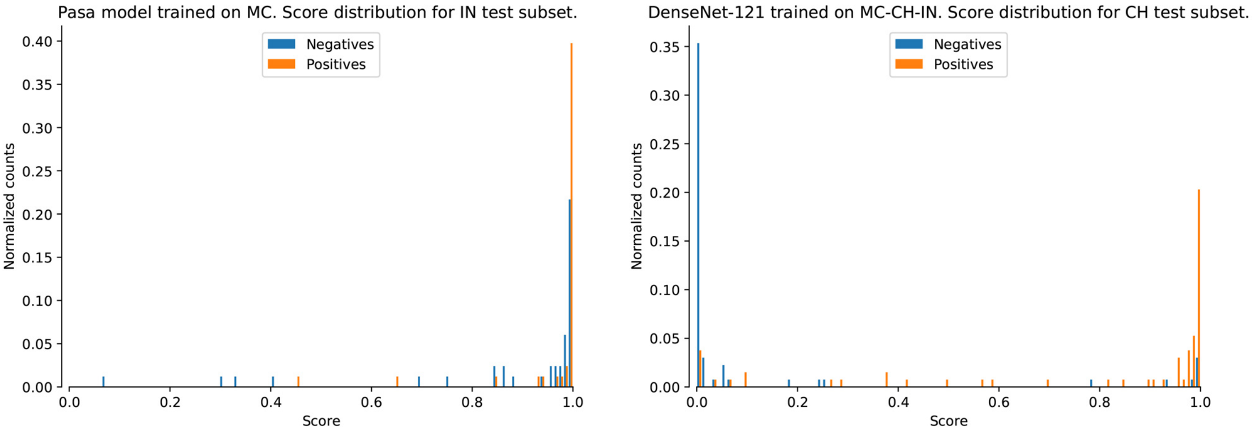


Figure 3.18: Score distribution of the worse and best direct detection models. We observe an inability to separate TB cases from healthy ones for the left model.

4. Indirect Detection

We will now introduce and evaluate the proposed indirect detection method illustrated in Figure 4.1. A first model will classify radiological signs present on the CXR image which will then be fed to a second model taking care of PTB prediction.

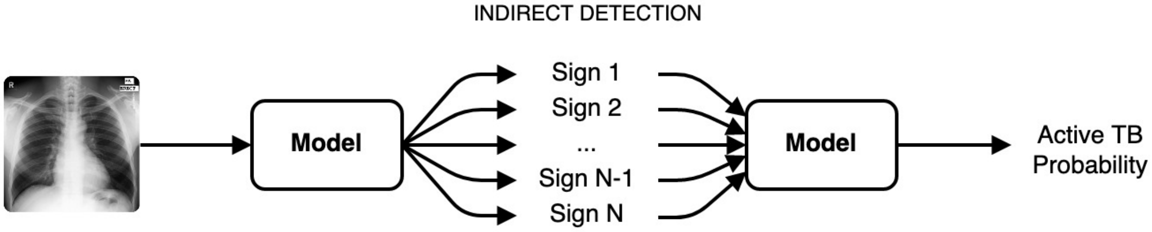


Figure 4.1: The proposed indirect detection method is composed of two models.

4.1 General Radiological Signs Dataset

To train the first of the two models, a dataset of CXR images annotated with radiological findings is required. Several large datasets of this kind are publicly available: NIH CXR14, CheXpert, MIMIC CXR [32, 45, 46, 47]. We decided to work with the NIH CXR14 dataset as it is composed of good quality CXR images annotated with multiple radiological signs (illustrated in Figure 4.2), some of which are TB-related. The main characteristics of this dataset, which will be used to train a classifier to detect the underlying radiological abnormalities of TB, are summarized in Table 4.1.

Table 4.1: Summary of the NIH CXR14 dataset annotated with radiological findings. This dataset does not contain information about patient's final diagnosis.

	Nb of images	**Resolution**	**Findings**
NIH CXR14 [32, 45]	112'120	1024×1024	Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural thickening, Cardiomegaly, Nodule, Mass and Hernia

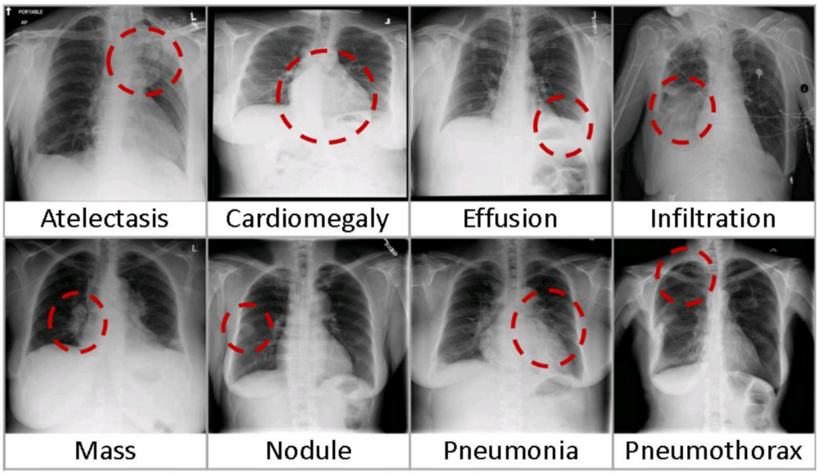


Figure 4.2: Eight visual examples of common thorax diseases from the NIH CXR14 dataset [32].

The NIH CXR14 dataset has been extracted from the clinical PACS database at the National Institutes of Health Clinical Center (USA) and represents 60% of all their radiographs. It is composed of 112'120 images of 30'805 unique patients with labels for fourteen common radiological signs including atelectasis, consolidation, infiltration, pneumothorax, edema, emphysema, fibrosis, effusion, pneumonia, pleural thickening, cardiomegaly, nodule, mass, and hernia. The greyscale CXRs are provided in PNG format with 8-bits depth in a standardized resolution of 1024×1024 pixels.

Metadata are provided in a CSV file reporting the following information for each CXR: image index, finding labels, follow-up number, patient id, patient age, patient gender, view position, original image size, and original image pixel spacing.

On top of that, hand-labeled bounding-boxes in CSV format are supplied for approximately 1'000 images of the set. For each of these boxes, the visible radiological finding is indicated.

The authors of the CheXNeXt study [35], following the identification of partially incorrect original labels, have relabelled the training and validation subsets. We will therefore use their version of the annotations and their split, detailed in Table 4.2. The authors have been careful to put images related to the same patient in only one of the subsets. Regarding the test subset, as it has not been relabelled, we will use annotations provided in the original dataset.

The average patient's age is 46.6 years old with radiographs from 63'340 men and 48'780 women.

The fourteen radiological signs annotated on the NIH CXR14 dataset can be separated into three categories according to their link with PTB. Some indicate a probable presence of the disease while others do not, as summarized in Table 4.3. In the context of active

Table 4.2: Overview of the number of samples in the different subsets of the NIH CXR14 dataset with the split proposed by the CheXNeXt study authors [35].

	Training n (%)	**Validation n (%)**	**Testing n (%)**
Number of Cases	98'637 (90%)	6'350 (6%)	4'054 (4%)

PTB rule-out, signs indicating a probable presence of the disease imply that TB preventive therapy should not be started. When the signs are possibly related to PTB, a medical doctor should interpret the radiograph to take a decision. And when radiological findings are not TB-related, the TB preventive therapy could start.

Table 4.3: Summary of the relationship between the different radiological signs annotated on the NIH CXR14 dataset and PTB. This table has been produced by a team of medical experts on PTB.

Radiological sign	**Link with PTB**
Pleural Effusion	Likely PTB
Infiltration	Likely PTB
Pneumonia	Likely PTB
Mass	Could be PTB
Nodule	Could be PTB
Atelectasis	Could be PTB
Fibrosis	Could be PTB
Consolidation	Could be PTB
Cardiomegaly	Unlikely PTB
Emphysema	Unlikely PTB
Hernia	Unlikely PTB
Pneumothorax	Unlikely PTB
Pleural thickening	Unlikely PTB
Edema	Unlikely PTB

4.2 Models

4.2.1 Radiological Signs Classification Model

Inspired by the CheXNeXt study [35], we decided to use a DenseNet-121 model for radiological sign classification. On this model, we replaced the final layer with a new fully-connected one, producing a 14-dimensional output (one dimension per radiological finding). But this time, rather than starting from a random initialization of the weights, we selected the model provided by PyTorch, pre-trained on the ImageNet dataset [48]. While the original work from Rajpurkar et al. [35] uses model ensembles to maximize performance, we satisfy ourselves of the simplification and use a single model. Although we did not compare them with an ensemble of 10 models, we consider the performance of a single model to be sufficient for the purpose of our study. Indeed, the tenfold increase in computational cost required by the ensemble would not alter our conclusion.

Like in [35], we use a Multi-Class Cross-Entropy loss, a batch size of 8 samples, a learning rate of 1×10^{-4}, and the default Adam optimizer parameters. We train the model for 10 epochs on the NIH CXR14 [45] dataset, as in [35], and the checkpoint with the lowest validation loss, illustrated in Figure 4.3, is retained for our pipeline.

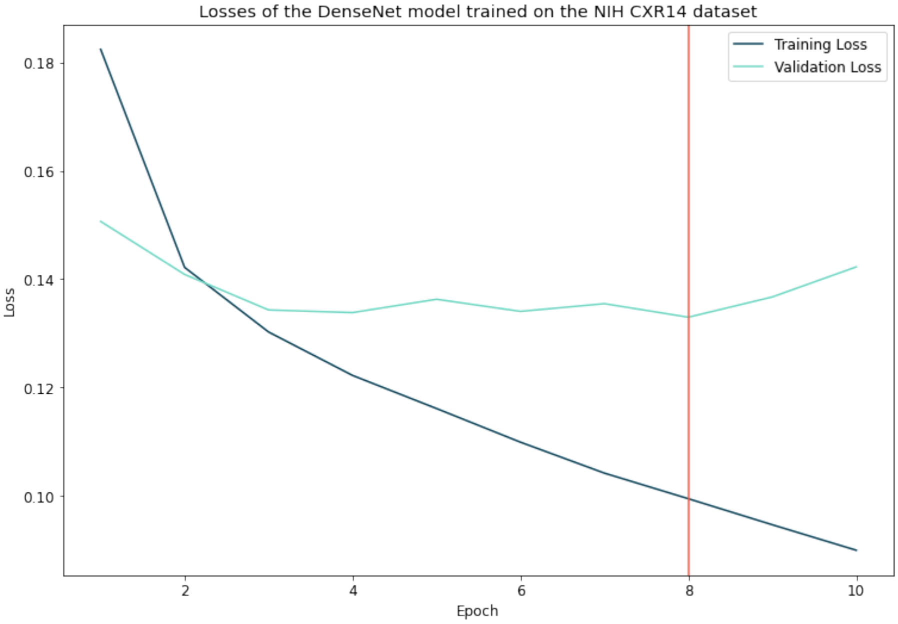


Figure 4.3: Losses evolution of the DenseNet model trained with hyperparameters from Rajpurkar et al. [35] on the NIH CXR14 dataset [45]. The vertical red line indicates the location of the lowest validation loss.

Table 4.4 presents the AUC for each radiological sign classified by our model alongside the corresponding one from the CheXNeXt study [35]. We notice that our results are higher for three radiological signs and lower for the eleven others. This difference is primarily explained by the fact that we used the full original test subset of the NIH CXR14 dataset [32] while CheXNeXt used a selection of 420 images newly annotated by three radiologists. Hence, our test subset may contain labelling errors as this was the case in the training and validation subsets. The second reason is the use of an ensemble of 10 models by CheXNeXt when we use only one model on our side. Note that our implementation is strictly similar to the CheXNeXt one as consistent predictions are generated when we input the same radiographs to our model configured with identical parameters.

Table 4.4: AUC results from our implementation of a DenseNet-121 model on the original test subset of the NIH CXR14 dataset [32] compared with those from CheXNeXt [35]. We observe that CheXNeXt achieves better results for 11 radiological signs.

Radiological sign	**Our model**	**CheXNeXt (95% CI)**
Atelectasis	0.667	0.862 (0.825 to 0.895)
Cardiomegaly	0.855	0.831 (0.790 to 0.870)
Consolidation	0.680	0.893 (0.859 to 0.924)
Edema	0.819	0.924 (0.886 to 0.955)
Effusion	0.737	0.901 (0.868 to 0.930)
Emphysema	0.858	0.704 (0.567 to 0.833)
Fibrosis	0.755	0.806 (0.719 to 0.884)
Hernia	0.854	0.851 (0.785 to 0.909)
Infiltration	0.636	0.721 (0.651 to 0.786)
Mass	0.782	0.909 (0.864 to 0.948)
Nodule	0.664	0.894 (0.853 to 0.930)
Pleural thickening	0.759	0.798 (0.744 to 0.849)
Pneumonia	0.693	0.851 (0.781 to 0.911)
Pneumothorax	0.780	0.944 (0.915 to 0.969)

4.2.2 Pulmonary Tuberculosis Prediction Model

To predict PTB from the fourteen radiological signs classified by the DenseNet-121 model described above, we tested two different models: a simple logistic regression classifier and a shallow network with a single layer of hidden neurons. Since we obtained superior performance with the former, we will focus on it here.

The classifier has been trained via backpropagation like all other DL models so far and has been subject to hyperparameter tuning: batch size of 4, learning rate of 10^{-2}, and the default Adam optimizer parameters. Although we could have used a closed-form solution such as L-BFGS (a quasi-Newton method approximating the Broyden-Fletcher-Goldfarb-Shanno algorithm) instead of Adam, we chose to use the latter because it allowed us to standardize the training code for all models. The details of the performed grid search are presented in Table 4.5 and the evolution of the losses with the optimal hyperparameters in Figure 4.4. We observe a much lower validation loss variability here, probably related to the small amount of information (14 scalars) that the model receives as input, than with direct detection models. However, we once again reach the lowest validation loss very quickly because our sample number is still limited. We trained the linear models for 100 epochs and the checkpoints with the lowest validation Binary Cross-Entropy losses have been retained for the evaluation.

Table 4.5: Results of the grid search conducted on the aggregated MC-CH-IN dataset to identify the best hyperparameters for the logistic regression classifier. This table indicates the minimum validation loss obtained for each combination of learning rate and batch size.

Learning rate	**Batch size: 4**	**Batch size: 8**	**Batch size: 16**
1×10^{-1} (training for 50 epochs)	0.3932	0.4013	0.4229
1×10^{-2} (training for 100 epochs)	**0.3835**	0.3998	0.4126
1×10^{-3} (training for 200 epochs)	0.3875	0.4075	0.4188
1×10^{-4} (training for 800 epochs)	0.3942	0.4059	0.4123

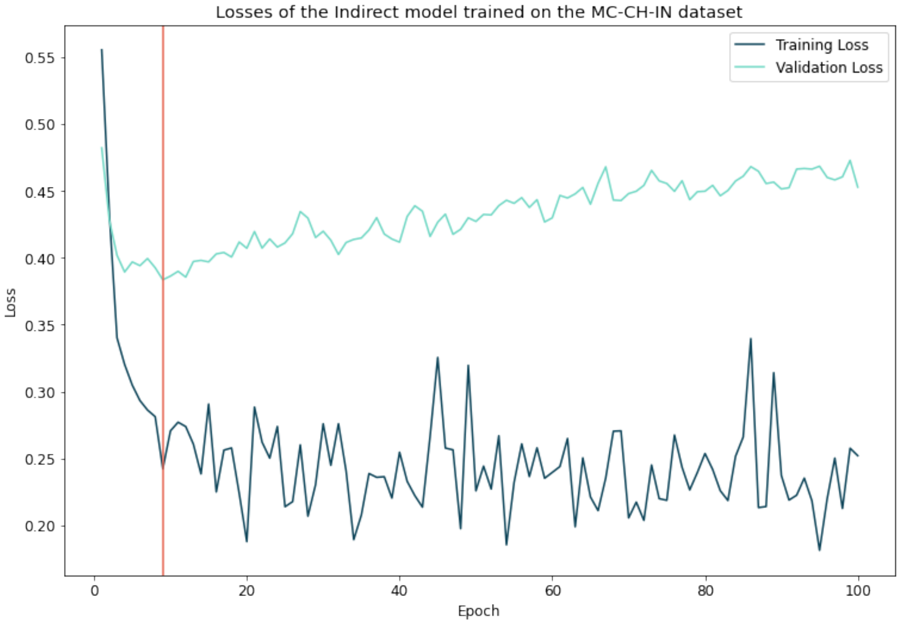


Figure 4.4: Losses evolution of the logistic regression classifier trained with optimized hyperparameters on the radiological signs annotations generated for the MC-CH-IN dataset. The vertical red line indicates the location of the lowest validation loss.

4.3 Prediction Procedure

Once the radiological sign detection model was trained, we used it to predict the presence of the fourteen signs on the three PTB-specific datasets: MC, CH, and IN. We subsequently trained the logistic regression classifier on these three datasets following the same process as for direct detection, detailed in Section 3.3, but feeding the radiological signs to the model rather than CXR images. Again, we made use of cross-validation by using the same 10 folds of each dataset as before. Similarly, the thresholds used for the indirect prediction correspond to the average of the 10 thresholds identified on the validation subsets and are shown in Table 4.6.

Table 4.6: Thresholds used to classify a CXR image as PTB-positive or healthy for our indirect model. The threshold is identified by maximizing the F1-score on the validation subset of the dataset on which the model has been trained. The threshold indicated here is an average of 10 thresholds as we made use of cross-validation.

Model	**Threshold**
Indirect (train: MC)	0.5340
Indirect (train MC+CH)	0.2838
Indirect (train: MC+CH+IN)	0.2371

4.4 Results

Table 4.7 presents AUCs of our indirect detection method subject to the same evaluation protocol as in Table 3.7. We reported the best results from that table here, for ease of comparison. We first observe that, contrary to baseline results for direct detection, cross-database generalization in this scenario becomes consistently independent of the PTB training subset. This conclusion is corroborated by the similarity of the three corresponding ROC plots presented in Figure 4.5. Another visual illustration of this aspect is provided in Figure 4.6 with a comparison of score distributions when predicting on the IN dataset using the Pasa model and the indirect model, both trained on MC. While neither model has seen the CH dataset during training, only the indirect model is able to separate PTB cases (positives) from healthy cases (negatives). Secondly, we observe that the test AUC is now more consistent with state-of-the-art values reported in other less reproducible work (top 3 AUC from the systematic review [10]: 0.99, 0.98, 0.96). We further hypothesize that if a dataset with a larger number of PTB-specific signs would be available, classification performance could be boosted to optimal levels.

To test this hypothesis, we return to a direct classification scheme using the DenseNet-121 model. In place of starting with random initialization of the weights, as we did in our first attempt, we use the model pre-trained on ImageNet provided by PyTorch. We pre-train the model a second time on NIH CXR14, remove its multi-class output and adapt a new output layer (random initialization) for binary classification. We train the resulting model for a further 300 epochs on the TB datasets, using hyperparameters identified using a grid search: learning rate of 10^{-4} and batch size of 8. The details of the performed grid search are

Table 4.7: Indirect detection via a Logistic Regressor from Radiological Signs shows better generalization capabilities over unseen data. The test AUC is reported.

AUC	**MC test**	**CH test**	**IN test**
Direct (best overall)	0.890	**0.917**	0.850
Indirect (train: MC)	**0.966**	0.867	0.926
Indirect (train: MC+CH)	0.961	0.901	**0.928**
Indirect (train: MC+CH+IN)	0.951	0.895	0.920

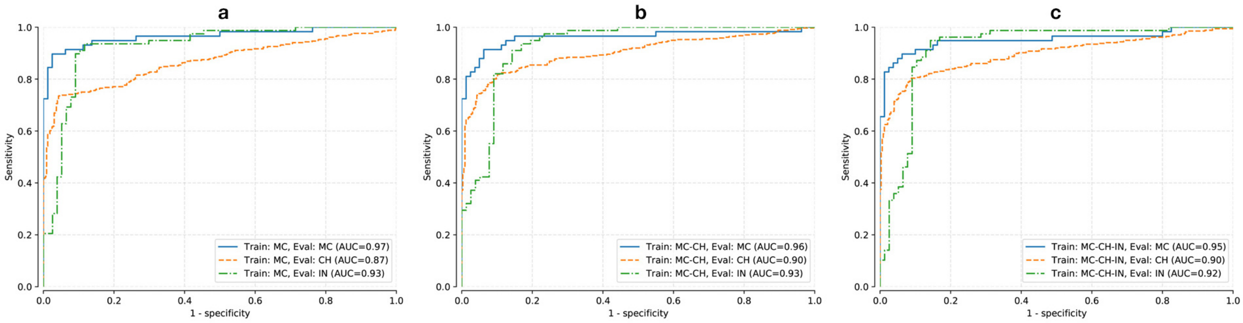


Figure 4.5: ROCs of our indirect detection model. These cross-validated curves were generated after training the model on the (a) MC, (b) MC-CH, and (c) MC-CH-IN datasets. Cross-database generalization is consistently independent of the PTB training subset.

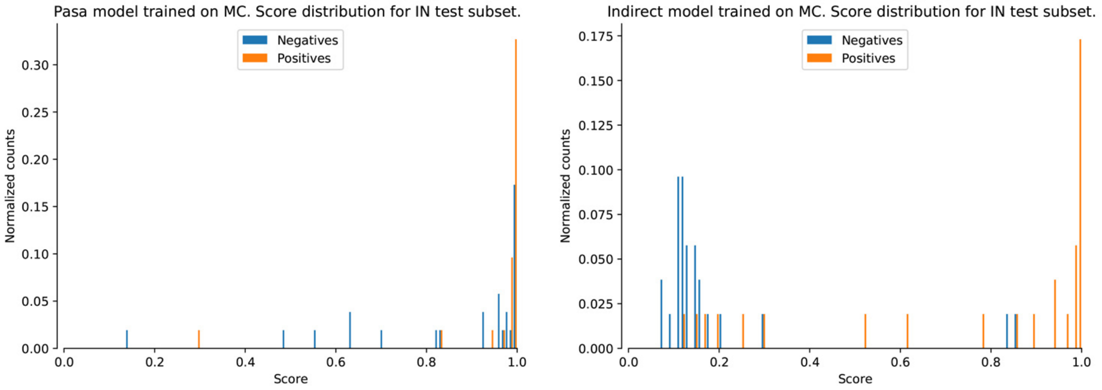


Figure 4.6: Comparison of score distributions when predicting on the IN test subset by the Pasa model and the Indirect model both trained on MC. We observe that only the indirect model is able to separate PTB cases (positives) from healthy cases (negatives).

presented in Table 4.8 and the thresholds used in Table 4.9. The results of the final models are summarized in Table 4.10.

Table 4.8: Results of the grid search conducted on the aggregated MC-CH-IN dataset to identify the best hyperparameters for the DenseNet-121 model pre-trained on ImageNet and NIH CXR14. This table indicates the minimum validation loss obtained for each combination of learning rate and batch size.

Learning rate	**Batch size: 4**	**Batch size: 8**	**Batch size: 16**
1×10^{-4} (training for 300 epochs)	0.2053	**0.1511**	0.2372
1×10^{-5} (training for 500 epochs)	0.1832	0.1931	0.2326
1×10^{-6} (training for 600 epochs)	0.2086	0.2139	0.2138

Table 4.9: Thresholds used to classify a CXR image as PTB-positive or healthy for the DenseNet-121 network pre-trained on ImageNet and NIH CXR14. The threshold is identified by maximizing the F1-score on the validation subset of the dataset on which the model has been trained. The threshold indicated here is an average of 10 thresholds as we made use of cross-validation.

Model	**Threshold**
DenseNet-121@CXR14 (train: MC)	0.4126
DenseNet-121@CXR14 (train MC+CH)	0.3711
DenseNet-121@CXR14 (train: MC+CH+IN)	0.4255

Table 4.10: A Densenet-121 model successively pre-trained on ImageNet and NIH CXR14 may provide insight into possible performance gains using datasets annotated with more TB-specific radiological signs. The test AUC is reported.

AUC	**MC test**	**CH test**	**IN test**
Best overall	0.966	0.917	0.928
DenseNet-121@CXR14 (train: MC)	0.966	0.917	0.901
DenseNet-121@CXR14 (train: MC+CH)	**0.984**	**0.979**	0.869
DenseNet-121@CXR14 (train: MC+CH+IN)	0.965	0.978	**0.931**

We can see that the results are even better, at the cost of a loss of interpretability. Indeed, radiological signs cannot be identified using direct detection. However, we notice that the corresponding ROCs, presented in Figure 4.7, fluctuate more than with indirect detection when the model is trained on different datasets. Thus, it appears that it is the use of radiological signs as an intermediate basis that allows us to obtain more stable results when the training subset varies.

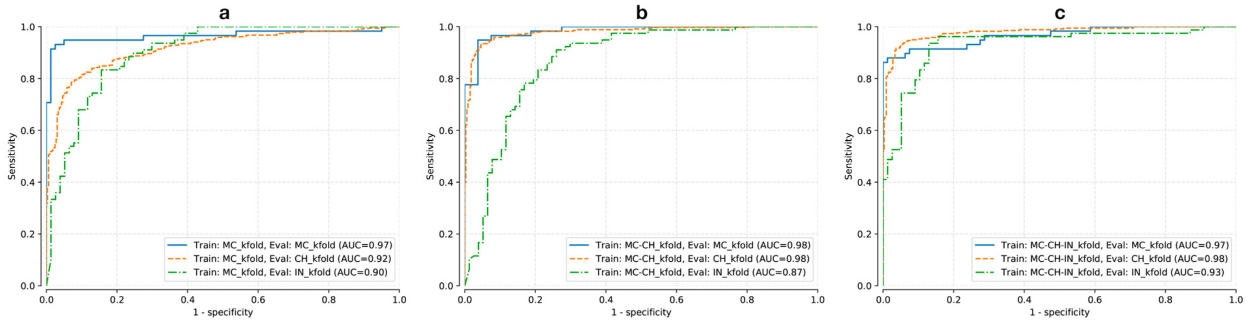


Figure 4.7: The corresponding ROCs of the DenseNet-121 model. We notice less stable curves than with the indirect detection method.

Since the annotations of the NIH CXR14 dataset are not specifically adapted to PTB detection and we still get state-of-the-art results, we believe that PTB-specific annotations would allow our indirect detection model to achieve similar performance to that presented here while having a more interpretable diagnosis and more stable predictions regardless of the PTB training subset. The best of both worlds in other words.

4.5 Radiological Signs Importance Analysis

To assess the relative impact, in PTB prediction, of each of the fourteen radiological signs used within our indirect detection method, we used two different approaches, presented in this section.

4.5.1 Random Permutation

First, we randomly permuted the values of one of the radiological signs within the subsets of a selected dataset and computed the mean squared error over all samples between the new outputs and the original ones:

$$Impact_i = \frac{1}{N}\sum_{j=1}^{N}[output(\vec{x_j}) - output(\vec{x_j} \mid x_{j,i} = x_{k,i})]^2 \tag{4.1}$$

with $k \neq j$ and such that we use each value only once.

This manipulation, proposed in [49, 50], has been performed for each radiological sign, one after the other, and allowed us to measure their relative impact on the output. If our model is particularly sensitive to one variable, the error that its random permutation will generate will be large. Conversely, permuting a variable that is of little use for PTB prediction will generate little or no error. The model used for this process is the Logistic Regression model already trained on the first non-modified fold of the aggregated MC-CH-IN dataset for 100 epochs. We present in Figure 4.8 a plot of the mean squared error generated with the multiple subsets of the MC, CH, and IN datasets.

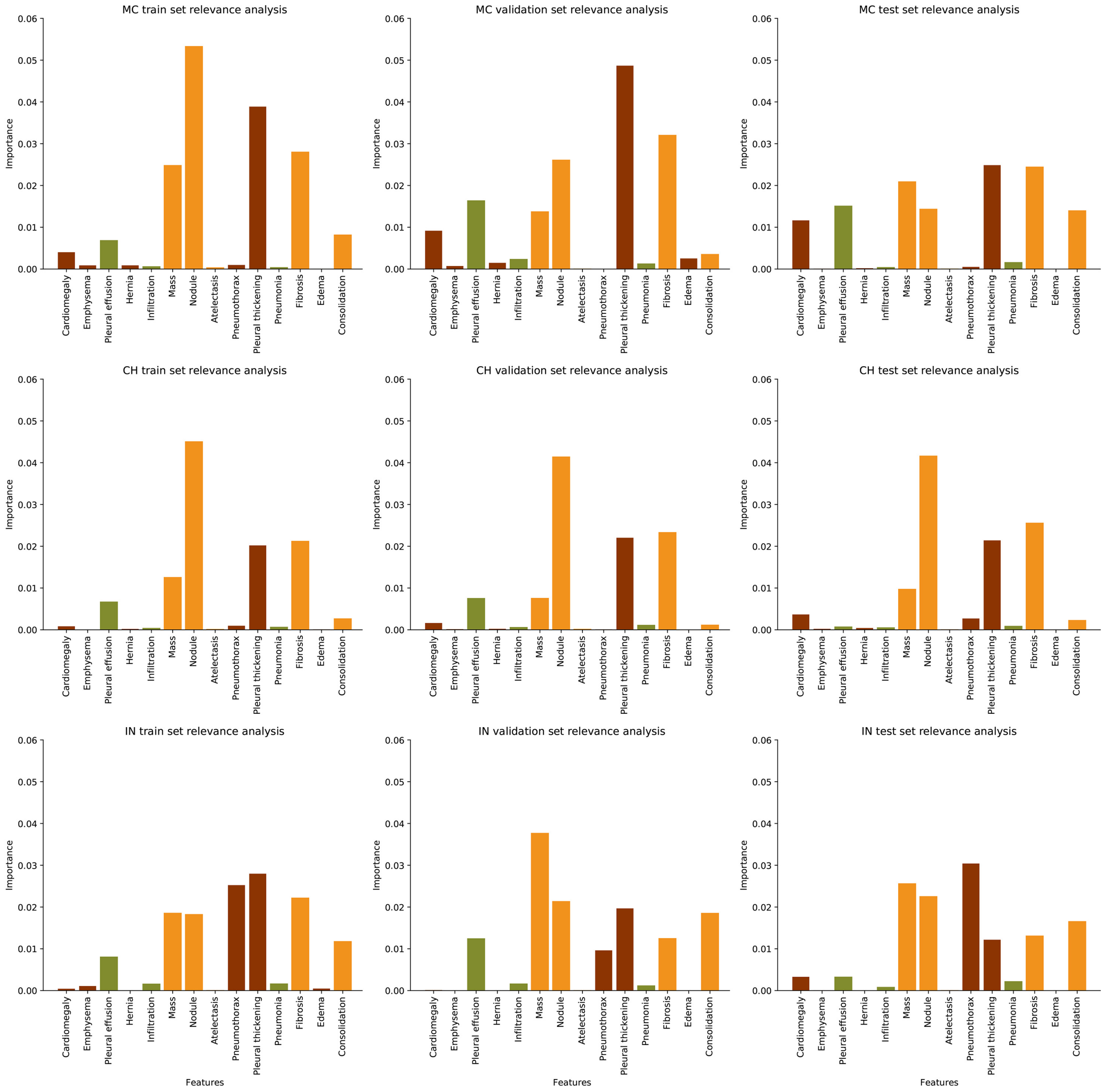


Figure 4.8: Relative importance of individual radiological signs for PTB prediction with our logistic regression model, calculated using random permutations. The color of the bars indicates the relationship between the radiological sign and PTB (green: likely TB, orange: could be TB, red: unlikely TB).

We find that the importance of each radiological sign in isolation is very low when predicting PTB. Indeed, the highest value is below 0.06 in the case of the nodule on the MC dataset. Furthermore, although 4 radiological findings (mass, nodule, pleural thickening, and fibrosis) are represented on all the plots, they are present in varying proportions and are accompanied by different other signs depending on the dataset. These results indicate that more than one sign in particular, it is an ensemble of radiological signs that would allow the detection of PTB. We also note that signs that are relatively more important during diagnosis are not necessarily signs classified as "likely TB" by physicians. Further analysis in collaboration with healthcare specialists is required to better understand this element.

4.5.2 Radiological Sign Dropping

Whereas we did not re-train the model with previous method, a more direct but more computationally costly technique for measuring the importance of each sign involves removing the data for that sign from the dataset for both model training and prediction. Given the unavailability of information about the discarded sign in this process, biases potentially present in the first technique are eliminated, as explained by Parr et al. [50].

This time we started by training and evaluating our logistic regression model on the original MC-CH-IN aggregated dataset. Then we repeated this training and evaluation by systematically discarding the information of one radiological sign from the dataset at a time. The result, shown in Figure 4.9, is a ROC curve and its corresponding AUC for each version of the dataset from which one of the signs has been dropped. We can consequently see to what extent each radiological finding influences the predictive performance of the model. We expect a much lower AUC than the original (0.87) when a sign of major importance in PTB prediction is no longer available. We find, however, that the decrease in AUC is at most 2% in the case of the nodule (AUC of 0.85), which was one of the most influential signs with the random permutation method. These results confirm the fact that no sign taken in isolation has a strong impact on the predictive capacity of the model. Indeed, no significant decrease in performance can be observed when a radiological sign is discarded.

Further analysis of the interactions between the different signs is thus required to draw more accurate conclusions about their role in PTB prediction.

4.6 Radiological Signs Visualisation

Although we do not possess radiological signs annotations for PTB datasets, we have implemented a grad-CAM prediction functionality to produce a visual overview of the predictions. Among the various possible implementations of grad-CAMs [51, 52, 53, 54], we chose to use the method proposed by Ramprasaath R. Selvaraju et al. in [51]. This method consists of using the gradient of a prediction (a radiological sign in the present case) to produce a

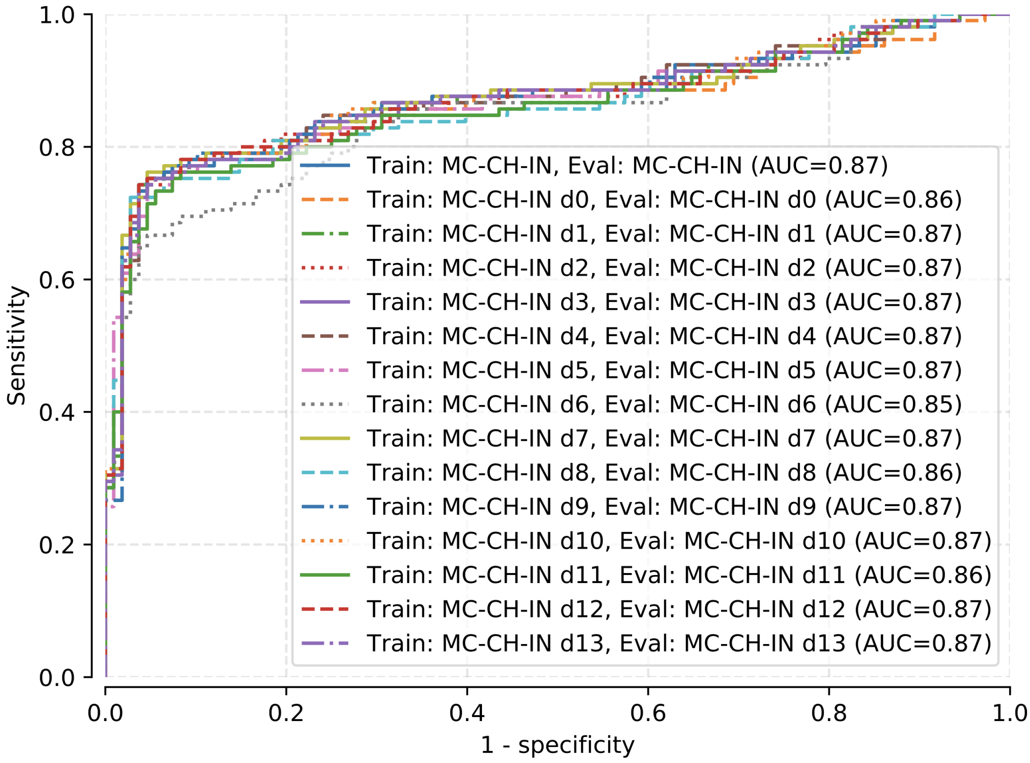


Figure 4.9: Predictive capabilities of our logistic regression model after removing the data for each radiological sign (d0-d13 correspond, in this order, to cardiomegaly, emphysema, effusion, hernia, infiltration, mass, nodule, atelectasis, pneumothorax, pleural thickening, pneumonia, fibrosis, edema, and consolidation).

coarse localization map highlighting the important regions of the chest radiograph for this prediction. To discard less confident predictions, we decided to calculate the grad-CAM only for signs with a score higher than 0.5. We illustrate this work with the four predictions shown in Figure 4.10.

We find that most areas are located within the chest cavity, which is a good sign, but that there are some misses, such as the case of fibrosis prediction using a presumably useless part of the radiograph in the top-left of the second image.

If in the future these PTB datasets are annotated by radiologists, we could further evaluate the precision of these grad-CAMs.

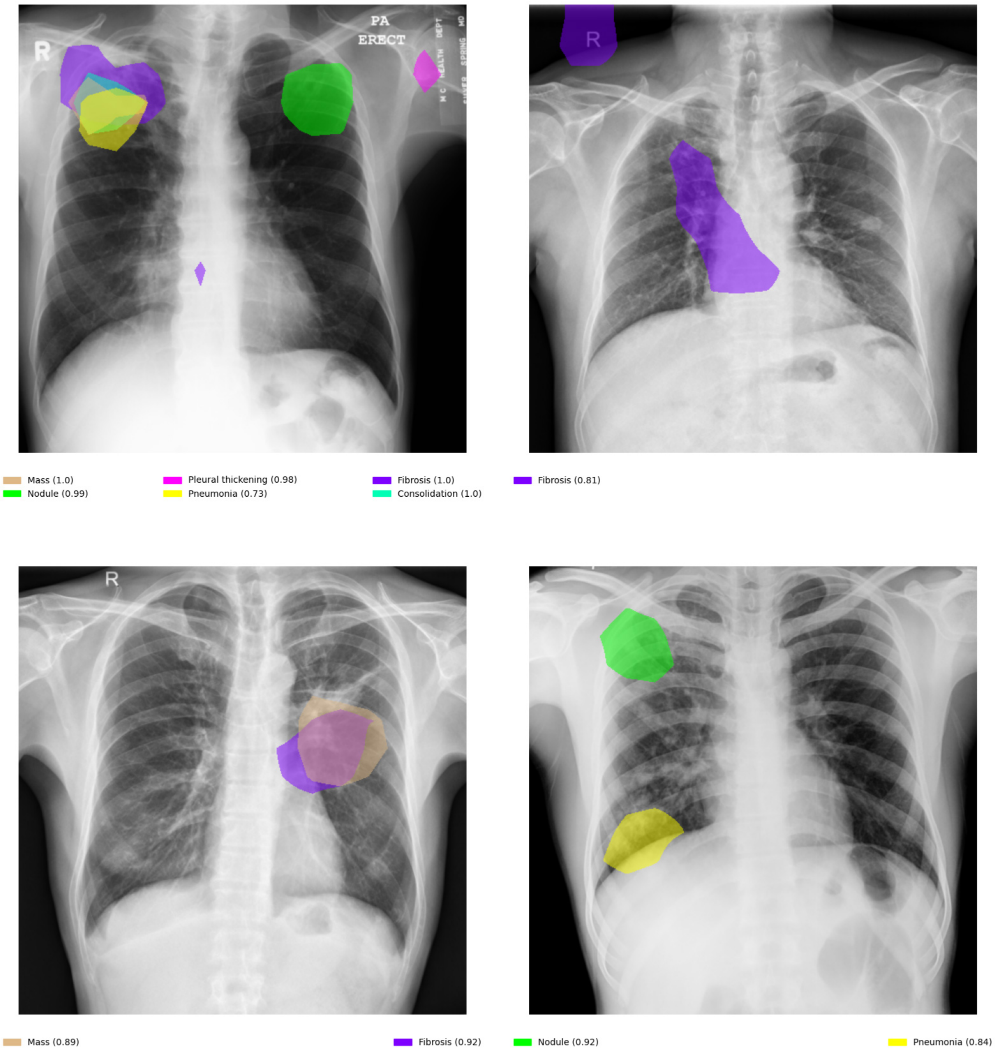


Figure 4.10: Sample grad-CAMs generated on the PTB datasets (no ground-truth available) using the method proposed by Ramprasaath R. Selvaraju et al. in [51].

5. Conclusion

TB is still one of the leading causes of death from a single infectious agent in the world, second only to COVID-19. To mitigate this, the United Nations propose to treat active TB and LTBI simultaneously. In the case of LTBI, an important step in triage, prior to treatment, involves ruling out patients who have developed an active TB. CXR imaging is commonly part of this process as it is a widely available form of radiography, and is considered the most accurate screening tool for detecting PTB in the general population among standard testing tools protocoled by the WHO. Moreover, this technology allows for the identification of asymptomatic patients. Unfortunately, there is often a lack of trained radiologists to analyze images at the location of the examination. In this context, CAD software may constitute a fundamental cog of the solution. More recently, the WHO confirmed that they can be used in place of human readers for the interpretation of digital CXRs, for screening and triage.

A direct detection method is generally preferred for such systems in state-of-the-art research. With this method, a radiograph is fed into a model producing a score reflecting the presence or absence of the disease. Although results from studies using this method are remarkable (top 3 AUC of [10]: 0.99, 0.98, 0.96), they are not interpretable by healthcare professionals who commonly use radiological signs as a basis for their work. Most of these studies are also not reproducible as code or datasets are usually kept private. A second category of studies has successfully investigated the classification of radiological signs using CXRs. We further hypothesized that an indirect detection of PTB, based on the radiological findings identified on a radiograph, could offer naturally interpretable results, better cross-database generalization, and provide state-of-the-art results. The suggested indirect detection method is illustrated in Figure 5.1.

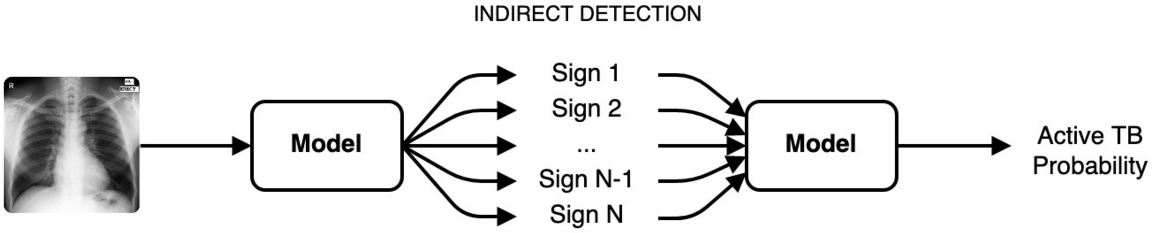


Figure 5.1: The indirect detection method is composed of two models. The first one takes CXR images as input and outputs radiological signs while the second one predicts PTB from these signs.

The baseline from Chapter 3, composed of a model proposed in the only reproducible study by Pasa et al. [29] and of a DenseNet-121, provided a base for comparison and allowed us to confirm the generalization difficulties of these models. Indeed, performances are consistent with reported values in the literature only after the dataset being tested was also seen during training. We visually observed this difficulty on ROCs of the Pasa model, trained on different datasets, presented alongside their corresponding AUCs in Figure 5.2. We also report in Table 5.1 the best results that the direct detection models allowed us to obtain for each test subset.

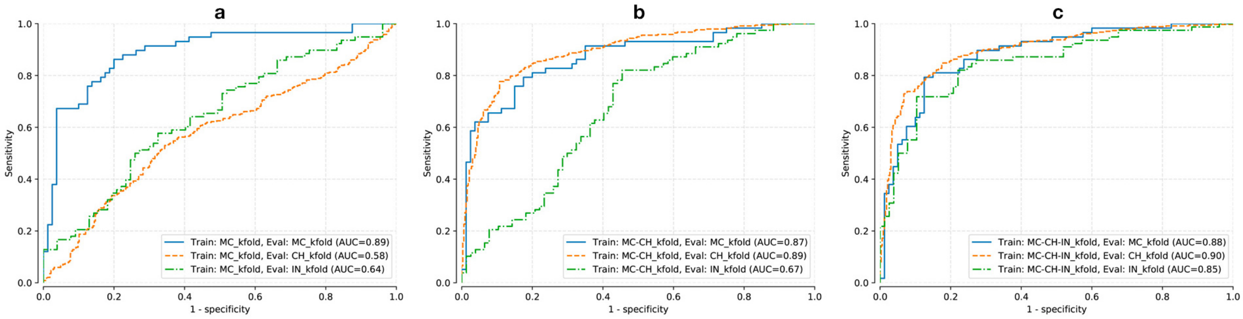


Figure 5.2: ROCs of our implementation of the Pasa model. The cross-validated predictions were generated by the model after training it on the (a) MC, (b) MC-CH, and (c) MC-CH-IN datasets. Performances are consistent with reported values in the literature only after the dataset being tested was also seen during training.

Table 5.1: Best results generated by the direct detection models in our baseline. The test AUC is reported.

AUC	**MC test**	**CH test**	**IN test**
Direct (best overall)	0.890	0.917	0.850

While direct detection is generally preferred in the literature, our study of the indirect detection method presented in Chapter 4 suggests that radiological signs extracted from CXR images constitute a sufficient canvas, close to clinical requirements, to build more interpretable and generalizable CAD for active PTB detection. We obtained state-of-the-art results, presented in Table 5.2, by simply plugging a linear classifier into a DL-based framework detecting radiological signs on CXR images. Our indirect detection algorithm provides better generalization as can be seen in Figure 5.3, more interpretable diagnosis, and state-of-the-art performance while using a training set containing only 8 TB-related radiological signs. These results confirm that our hypothesis should not be rejected at this point and that further research should be carried out in this direction.

We further hypothesized that if a dataset with a larger number of PTB-specific signs would be available, classification performance could be boosted to optimal levels. To test this hypothesis, we returned to a direct classification scheme and fine-tuned a DenseNet-121 model pre-trained on thousands of CXR images. By doing so, we obtained even better results, shown in Table 5.3, in exchange for interpretability. However, we notice that the

Table 5.2: Best results generated by the proposed indirect detection method. The test AUC is reported.

AUC	MC test	CH test	IN test
Indirect (best overall)	0.966	0.901	0.928

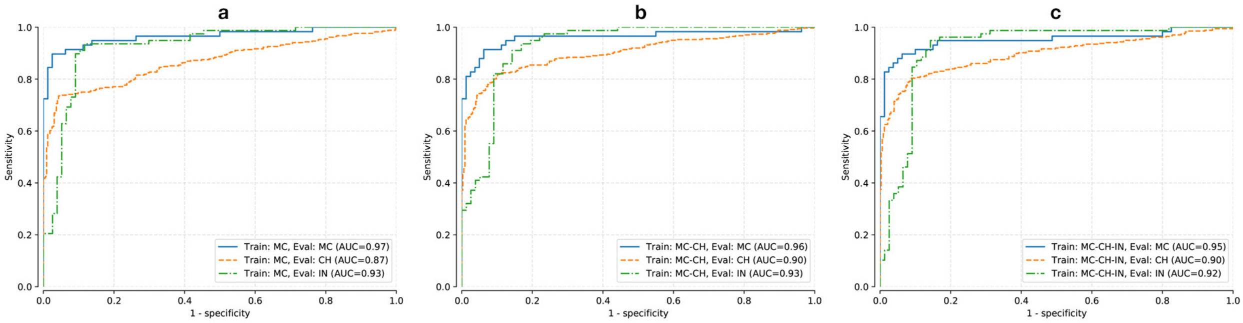


Figure 5.3: ROCs of our indirect detection model. These cross-validated curves were generated after training the model on the (a) MC, (b) MC-CH, and (c) MC-CH-IN datasets. Cross-database generalization is consistently independent of the PTB training subset.

corresponding ROCs, presented in Figure 5.4, fluctuate more than with the indirect detection method. Thus, it appears that it is the use of radiological signs as an intermediate basis that allows us to obtain more stable results in presence of different training subsets. Nevertheless, these results offer a glimpse of the possible performance gains that an adapted PTB dataset with more specific radiological signs annotations could bring.

Table 5.3: A Densenet-121 model successively pre-trained on ImageNet and NIH CXR14 may provide insight into possible performance gains using datasets annotated with more TB-specific radiological signs. The test AUC is reported.

AUC	MC test	CH test	IN test
Direct pre-trained (best overall)	0.984	0.979	0.931

Finally, we measured the impact of individual radiological signs in predicting the disease. Our results suggest that more than one radiological sign in isolation, it is the combination of several signs that allows our model to predict PTB. This is a conclusion that is consistent with the methodology applied by healthcare workers today.

While state-of-the-art results could be extracted in the proposed indirect workflow, it is adequate to highlight the limitations of this work. First and foremost, public PTB datasets are relatively small in size and may not be representative of realistic deployment conditions. A study considering confidence intervals may throw some light on this matter. Secondly, the use of known markers for a disease may limit the discovery of new ones. Thirdly, as

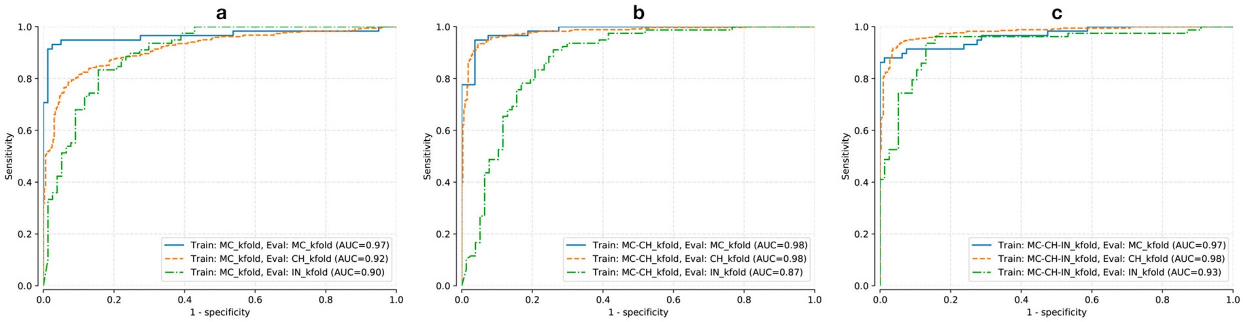


Figure 5.4: ROCs of the pre-trained DenseNet-121 model. We notice less stable curves than with the indirect detection method.

the results of the indirect detection methodology are promising, it would make sense to annotate PTB datasets with radiological signs information to check the consistency of the numerical and visual predictions of our models. Considering all of the above, we believe that a combination of both direct and indirect techniques into a single CAD solution could offer both interpretability and the required robustness in realistic deployments. On the technical side, possible improvements are the implementation of a scheduler reducing the learning rate progressively, the evaluation of different resolution methods for the logistic regressor or the calculation and usage of more precise thresholds for grad-CAMs generation.

Finally, we point out that the proposed workflow could be applicable to other diseases and medical imaging techniques, but this remains untested at the moment. To bridge this gap, we make our findings fully reproducible, distributing code and documentation[1] so these limitations may be eventually addressed.

[1] https://gitlab.idiap.ch/bob/bob.med.tb

A. Results of the Experiments

Trained on /tested on	Model	F1 (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUROC
MC training subset / MC testing subset	Pasa	0.79 (0.71, 0.85)	0.83 (0.75, 0.88)	0.78 (0.65, 0.86)	0.86 (0.77, 0.92)	0.890
	Densenet121	0.74 (0.65, 0.81)	0.78 (0.71, 0.84)	0.74 (0.62, 0.84)	0.81 (0.71, 0.88)	0.822
	Logistic Regression	0.91 (0.85, 0.95)	0.93 (0.87, 0.96)	0.90 (0.79, 0.95)	0.95 (0.88, 0.98)	0.966
	Pre-trained Densenet121	0.94 (0.88, 0.97)	0.95 (0.90, 0.97)	0.90 (0.79, 0.95)	0.99 (0.93, 1.00)	0.966

Trained on /tested on	Model	F1 (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUROC
MC training subset / CH testing subset	Pasa	0.67 (0.64, 0.70)	0.51 (0.47, 0.55)	0.99 (0.98, 1.00)	0.01 (0.00, 0.03)	0.576
	Densenet121	0.67 (0.64, 0.70)	0.51 (0.47, 0.55)	0.99 (0.97, 1.00)	0.01 (0.00, 0.03)	0.607
	Logistic Regression	0.71 (0.67, 0.75)	0.77 (0.74, 0.80)	0.56 (0.51, 0.61)	0.99 (0.97, 1.00)	0.867
	Pre-trained Densenet121	0.82 (0.79, 0.85)	0.80 (0.77, 0.83)	0.90 (0.87, 0.93)	0.70 (0.64, 0.74)	0.917

Trained on /tested on	Model	F1 (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUROC
MC training subset / IN testing subset	Pasa	0.68 (0.61, 0.73)	0.52 (0.44, 0.59)	1.00 (0.95, 1.00)	0.03 (0.01, 0.09)	0.642
	Densenet121	0.67 (0.61, 0.73)	0.50 (0.43, 0.58)	1.00 (0.95, 1.00)	0.00 (0.00, 0.05)	0.625
	Logistic Regression	0.90 (0.84, 0.93)	0.90 (0.84, 0.94)	0.88 (0.79, 0.94)	0.91 (0.82, 0.95)	0.926
	Pre-trained Densenet121	0.73 (0.67, 0.78)	0.63 (0.55, 0.70)	1.00 (0.95, 1.00)	0.25 (0.16, 0.35)	0.901

Trained on /tested on	Model	F1 (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUROC
MC-CH training subset / MC testing subset	Pasa	0.72 (0.63, 0.80)	0.78 (0.71, 0.84)	0.67 (0.54, 0.78)	0.86 (0.77, 0.92)	0.870
	Densenet121	0.77 (0.68, 0.83)	0.80 (0.72, 0.86)	0.79 (0.67, 0.88)	0.80 (0.70, 0.87)	0.883
	Logistic Regression	0.88 (0.81, 0.92)	0.88 (0.82, 0.93)	0.97 (0.88, 0.99)	0.82 (0.73, 0.89)	0.961
	Pre-trained Densenet121	0.94 (0.88, 0.97)	0.95 (0.90, 0.97)	0.93 (0.84, 0.97)	0.96 (0.90, 0.99)	0.984

Trained on /tested on	Model	F1 (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUROC
MC-CH training subset / CH testing subset	Pasa	0.83 (0.80, 0.86)	0.83 (0.80, 0.85)	0.83 (0.79, 0.87)	0.82 (0.78, 0.86)	0.893
	Densenet121	0.83 (0.80, 0.85)	0.81 (0.78, 0.84)	0.88 (0.84, 0.91)	0.75 (0.70, 0.79)	0.905
	Logistic Regression	0.84 (0.81, 0.87)	0.85 (0.82, 0.88)	0.79 (0.75, 0.83)	0.91 (0.88, 0.94)	0.901
	Pre-trained Densenet121	0.94 (0.92, 0.95)	0.94 (0.92, 0.95)	0.94 (0.91, 0.96)	0.94 (0.90, 0.96)	0.979

Trained on /tested on	Model	F1 (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUROC
MC-CH training subset / IN testing subset	Pasa	0.70 (0.63, 0.75)	0.59 (0.51, 0.66)	0.94 (0.86, 0.97)	0.23 (0.15, 0.34)	0.669
	Densenet121	0.68 (0.61, 0.74)	0.57 (0.49, 0.64)	0.90 (0.81, 0.95)	0.23 (0.15, 0.34)	0.672
	Logistic Regression	0.86 (0.80, 0.91)	0.85 (0.78, 0.89)	0.97 (0.91, 0.99)	0.71 (0.60, 0.80)	0.928
	Pre-trained Densenet121	0.80 (0.74, 0.85)	0.75 (0.68, 0.82)	0.97 (0.91, 0.99)	0.53 (0.42, 0.64)	0.869

Trained on /tested on	Model	F1 (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUROC
MC-CH-IN training subset / MC testing subset	Pasa	0.80 (0.71, 0.86)	0.83 (0.75, 0.88)	0.81 (0.69, 0.89)	0.84 (0.74, 0.90)	0.881
	Densenet121	0.72 (0.64, 0.80)	0.77 (0.69, 0.83)	0.72 (0.60, 0.82)	0.80 (0.70, 0.87)	0.860
	Logistic Regression	0.87 (0.80, 0.91)	0.88 (0.81, 0.92)	0.95 (0.86, 0.98)	0.82 (0.73, 0.89)	0.951
	Pre-trained Densenet121	0.92 (0.85, 0.96)	0.93 (0.88, 0.96)	0.88 (0.77, 0.94)	0.97 (0.91, 0.99)	0.965

Trained on /tested on	Model	F1 (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUROC
MC-CH-IN training subset / CH testing subset	Pasa	0.83 (0.80, 0.86)	0.82 (0.79, 0.85)	0.87 (0.83, 0.90)	0.78 (0.73, 0.82)	0.898
	Densenet121	0.86 (0.83, 0.88)	0.85 (0.82, 0.88)	0.87 (0.83, 0.90)	0.83 (0.79, 0.87)	0.917
	Logistic Regression	0.85 (0.82, 0.87)	0.85 (0.82, 0.88)	0.80 (0.75, 0.84)	0.91 (0.88, 0.94)	0.895
	Pre-trained Densenet121	0.94 (0.92, 0.95)	0.94 (0.92, 0.95)	0.93 (0.90, 0.95)	0.95 (0.92, 0.97)	0.978

Trained on /tested on	Model	F1 (95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	AUROC
MC-CH-IN training subset / IN testing subset	Pasa	0.79 (0.72, 0.84)	0.77 (0.70, 0.83)	0.86 (0.76, 0.92)	0.68 (0.56, 0.77)	0.848
	Densenet121	0.81 (0.74, 0.86)	0.80 (0.73, 0.86)	0.83 (0.74, 0.90)	0.77 (0.66, 0.85)	0.850
	Logistic Regression	0.86 (0.80, 0.91)	0.85 (0.78, 0.89)	0.97 (0.91, 0.99)	0.71 (0.60, 0.80)	0.920
	Pre-trained Densenet121	0.90 (0.85, 0.94)	0.90 (0.84, 0.94)	0.95 (0.88, 0.98)	0.84 (0.75, 0.91)	0.931

Bibliography

[1] World Health Organization. *Global Tuberculosis Report*. 2020. URL: https://www.who.int/teams/global-tuberculosis-programme/tb-reports (visited on 07/09/2021).

[2] Rein M. G. J. Houben and Peter J. Dodd. "The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling". In: *PLOS Medicine* 13.10 (2016). Publisher: Public Library of Science, pp. 1–13. DOI: 10.1371/journal.pmed.1002152. URL: https://doi.org/10.1371/journal.pmed.1002152.

[3] WHO. *Percentage of active TB cases*. 2020. URL: https://www.paho.org/en/documents/world-tuberculosis-day-2020-infograph%20ic-jpg-treatment-tb-infection-latent-tb (visited on 06/18/2020).

[4] WHO. *Tuberculosis, Key Facts*. URL: https://www.who.int/en/news-room/fact-sheets/detail/tuberculosis (visited on 03/13/2020).

[5] WHO. *The End TB Strategy*. URL: https://www.who.int/tb/strategy/end-tb/en/ (visited on 03/13/2020).

[6] Christopher Dye et al. "Prospects for Tuberculosis Elimination". In: *Annual Review of Public Health* 34.1 (Mar. 18, 2013), pp. 271–286. ISSN: 0163-7525, 1545-2093. DOI: 10.1146/annurev-publhealth-031912-114431. URL: http://www.annualreviews.org/doi/10.1146/annurev-publhealth-031912-114431 (visited on 06/18/2020).

[7] WHO. *Latent TB Infection : Updated and consolidated guidelines for programmatic management*. 2018. URL: https://www.who.int/tb/publications/2018/latent-tuberculosis-infection/en/ (visited on 06/25/2020).

[8] World Health Organization. *WHO consolidated guidelines on tuberculosis. Module 1, Module 1,* 2020. URL: http://www.ncbi.nlm.nih.gov/books/NBK554956/ (visited on 09/16/2020).

[9] World Health Organization. *WHO consolidated guidelines on tuberculosis Module 2: Screening – Systematic screening for tuberculosis disease*. Mar. 22, 2021. URL: https://www.who.int/publications/i/item/9789240022676.

[10] Miriam Harris et al. "A systematic review of the diagnostic accuracy of artificial intelligence-based computer programs to analyze chest x-rays for pulmonary tuberculosis". In: *PLOS ONE* 14.9 (Sept. 3, 2019). Ed. by Pascal A. T. Baltzer, e0221339. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0221339. URL: http://dx.plos.org/10.1371/journal.pone.0221339 (visited on 03/23/2020).

[11] F. Madhani et al. "Automated chest radiography and mass systematic screening for tuberculosis". In: *The International Journal of Tuberculosis and Lung Disease* 24.7 (2020), pp. 665–673. ISSN: 1027-3719. DOI: doi:10.5588/ijtld.19.0501. URL: https://www.ingentaconnect.com/content/iuatld/ijtld/2020/00000024/00000007/art00004.

[12] Stefan Jaeger et al. "Automatic screening for tuberculosis in chest radiographs: a survey". In: *Quantitative Imaging in Medicine and Surgery* 3.2 (2013). URL: http://qims.amegroups.com/article/view/1813.

[13] Kunio Doi. "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential". In: *Computerized Medical Imaging and Graphics* 31.4 (June 2007), pp. 198–211. ISSN: 08956111. DOI: 10.1016/j.compmedimag.2007.02.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0895611107000262 (visited on 07/13/2020).

[14] R. Shen, I. Cheng, and A. Basu. "A Hybrid Knowledge-Guided Detection Technique for Screening of Infectious Pulmonary Tuberculosis From Chest Radiographs". In: *IEEE Transactions on Biomedical Engineering* 57.11 (2010), pp. 2646–2656.

[15] Wenlu Zhang et al. "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation". In: *NeuroImage* 108 (Mar. 2015), pp. 214–224. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2014.12.061. URL: https://linkinghub.elsevier.com/retrieve/pii/S1053811914010660 (visited on 07/13/2020).

[16] Yu-Jen Yu-Jen Chen et al. "Computer-aided classification of lung nodules on computed tomography images via deep learning technique". In: *OncoTargets and Therapy* (Aug. 2015), p. 2015. ISSN: 1178-6930. DOI: 10.2147/OTT.S80733. URL: http://www.dovepress.com/computer-aided-classification-of-lung-nodules-on-computed-tomography-i-peer-reviewed-article-OTT (visited on 07/13/2020).

[17] Hoo-Chang Shin et al. "Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning". In: *IEEE Transactions on Medical Imaging* 35.5 (May 2016), pp. 1285–1298. ISSN: 0278-0062, 1558-254X. DOI: 10.1109/TMI.2016.2528162. URL: https://ieeexplore.ieee.org/document/7404017/ (visited on 07/13/2020).

[18] Stefan Jaeger et al. "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases". In: *Quantitative Imaging in Medicine and Surgery* 4.6 (Dec. 2014), pp. 475–477. ISSN: 2223-4292. DOI: 10.3978/j.issn.2223-4292.2014.11.20. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4256233.

[19] Arun Chauhan, Devesh Chauhan, and Chittaranjan Rout. "Role of Gist and PHOG Features in Computer-Aided Diagnosis of Tuberculosis without Segmentation". In: *PLoS ONE* 9.11 (Nov. 12, 2014). Ed. by Hans A. Kestler, e112980. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0112980. URL: https://dx.plos.org/10.1371/journal.pone.0112980 (visited on 03/21/2020).

[20] World Health Organization. *Chest radiography in tuberculosis detection.* 2016. URL: https://www.who.int/tb/publications/chest-radiography/en/.

[21] Maria de Fátima Militão de Albuquerque et al. “Radiographic features of pulmonary tuberculosis in patients infected by HIV: is there an objective indicator of co-infection?” In: *Revista da Sociedade Brasileira de Medicina Tropical* 34.4 (Aug. 1, 2001), pp. 369–372. ISSN: 0037-8682. DOI: 10.1590/S0037-86822001000400010. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0037-86822001000400010&lng=en&tlng=en (visited on 03/03/2021).

[22] Eui Jin Hwang et al. “Development and Validation of a Deep Learning–based Automatic Detection Algorithm for Active Pulmonary Tuberculosis on Chest Radiographs”. In: *Clinical Infectious Diseases* 69.5 (Aug. 16, 2019), pp. 739–747. ISSN: 1058-4838, 1537-6591. DOI: 10.1093/cid/ciy967. URL: https://academic.oup.com/cid/article/69/5/739/5174137 (visited on 07/13/2020).

[23] Paras Lakhani and Baskaran Sundaram. “Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks”. In: *Radiology* 284.2 (Aug. 2017), pp. 574–582. ISSN: 0033-8419, 1527-1315. DOI: 10.1148/radiol.2017162326. URL: http://pubs.rsna.org/doi/10.1148/radiol.2017162326 (visited on 02/22/2021).

[24] Seok-Jae Heo et al. “Deep Learning Algorithms with Demographic Information Help to Detect Tuberculosis in Chest Radiographs in Annual Workers’ Health Examination Data”. In: *International Journal of Environmental Research and Public Health* 16.2 (Jan. 16, 2019), p. 250. ISSN: 1660-4601. DOI: 10.3390/ijerph16020250. URL: http://www.mdpi.com/1660-4601/16/2/250 (visited on 07/14/2020).

[25] K. C. Santosh and Sameer Antani. “Automated Chest X-Ray Screening: Can Lung Region Symmetry Help Detect Pulmonary Abnormalities?” In: *IEEE Transactions on Medical Imaging* 37.5 (2018), pp. 1168–1177. DOI: 10.1109/TMI.2017.2775636.

[26] U.K. Lopes and J.F. Valiati. “Pretrained convolutional neural networks as feature extractors for tuberculosis detection”. In: *Computers in Biology and Medicine* 89 (Oct. 2017), pp. 135–143. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2017.08.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S0010482517302548 (visited on 02/22/2021).

[27] K. C. Santosh et al. “Edge map analysis in chest X-rays for automatic pulmonary abnormality screening”. In: *International Journal of Computer Assisted Radiology and Surgery* 11.9 (Sept. 2016), pp. 1637–1646. ISSN: 1861-6410, 1861-6429. DOI: 10.1007/s11548-016-1359-6. URL: http://link.springer.com/10.1007/s11548-016-1359-6 (visited on 07/14/2020).

[28] Sangheum Hwang et al. “A novel approach for tuberculosis screening based on deep convolutional neural networks”. In: *Medical Imaging 2016: Computer-Aided Diagnosis*. Ed. by Georgia D. Tourassi and Samuel G. Armato III. Vol. 9785. Backup Publisher: International Society for Optics and Photonics. SPIE, 2016, pp. 750–757. DOI: 10.1117/12.2216198. URL: https://doi.org/10.1117/12.2216198.

[29] F. Pasa et al. “Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization”. In: *Scientific Reports* 9.1 (Dec. 2019), p. 6268. ISSN: 2045-2322. DOI: 10.1038/s41598-019-42557-4. URL: http://www.nature.com/articles/s41598-019-42557-4 (visited on 04/01/2020).

[30] Pranav Rajpurkar et al. “CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV”. In: *npj Digital Medicine* 3.1 (Dec. 2020), p. 115. ISSN: 2398-6352. DOI: 10.1038/s41746-020-00322-2. URL: http://www.nature.com/articles/s41746-020-00322-2 (visited on 02/17/2021).

[31] Hanif Esmail et al. “The Immune Response to *Mycobacterium tuberculosis* in HIV-1-Coinfected Persons”. In: *Annual Review of Immunology* 36.1 (Apr. 26, 2018), pp. 603–638. ISSN: 0732-0582, 1545-3278. DOI: 10.1146/annurev-immunol-042617-053420. URL: http://www.annualreviews.org/doi/10.1146/annurev-immunol-042617-053420 (visited on 03/03/2021).

[32] Xiaosong Wang et al. “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)*. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, July 2017, pp. 3462–3471. DOI: 10.1109/CVPR.2017.369. URL: http://ieeexplore.ieee.org/document/8099852/ (visited on 03/23/2020).

[33] Mark Cicero et al. “Training and Validating a Deep Convolutional Neural Network for Computer-Aided Detection and Classification of Abnormalities on Frontal Chest Radiographs:” in: *Investigative Radiology* 52.5 (May 2017), pp. 281–287. ISSN: 0020-9996. DOI: 10.1097/RLI.0000000000000341. URL: http://journals.lww.com/00004424-201705000-00004 (visited on 07/13/2020).

[34] Yaniv Bar et al. “Chest pathology detection using deep learning with non-medical training”. In: *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*. 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI 2015). Brooklyn, NY, USA: IEEE, Apr. 2015, pp. 294–297. ISBN: 978-1-4799-2374-8. DOI: 10.1109/ISBI.2015.7163871. URL: http://ieeexplore.ieee.org/document/7163871/ (visited on 07/13/2020).

[35] Pranav Rajpurkar et al. “Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists”. In: *PLOS Medicine* 15.11 (Nov. 20, 2018). Ed. by Aziz Sheikh, e1002686. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002686. URL: http://dx.plos.org/10.1371/journal.pmed.1002686 (visited on 04/01/2020).

[36] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. “Densely Connected Convolutional Networks”. In: *CoRR* abs/1608.06993 (2016). _eprint: 1608.06993. URL: http://arxiv.org/abs/1608.06993.

[37] Qingji Guan et al. *Diagnose like a Radiologist: Attention Guided Convolutional Neural Network for Thorax Disease Classification.* _eprint: 1801.09927. 2018.

[38] Li Yao et al. *Learning to diagnose from scratch by exploiting dependencies among labels.* _eprint: 1710.10501. 2018.

[39] A. Anjos et al. “Bob: a free signal processing and machine learning toolbox for researchers”. In: *20th ACM Conference on Multimedia Systems (ACMMM), Nara, Japan.* Oct. 2012. URL: https://publications.idiap.ch/downloads/papers/2012/Anjos_Bob_ACMMM12.pdf.

[40] A. Anjos et al. “Continuously Reproducing Toolchains in Pattern Recognition and Machine Learning Experiments”. In: *International Conference on Machine Learning (ICML).* Aug. 2017. URL: http://publications.idiap.ch/downloads/papers/2017/Anjos_ICML2017-2_2017.pdf.

[41] NIAID. *NIAID TB Portals.* 2020. URL: https://tbportals.niaid.nih.gov/ (visited on 06/19/2020).

[42] Aurelia Bustos et al. “PadChest: A large chest x-ray image dataset with multi-label annotated reports”. In: *Medical Image Analysis* 66 (Dec. 2020), p. 101797. ISSN: 13618415. DOI: 10.1016/j.media.2020.101797. URL: https://linkinghub.elsevier.com/retrieve/pii/S1361841520301614 (visited on 01/29/2021).

[43] Gufosowa. *Illustration of a k-fold cross-validation.* URL: https://commons.wikimedia.org/wiki/User:Gufosowa.

[44] Takaya Saito and Marc Rehmsmeier. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets”. In: *PLOS ONE* 10.3 (Mar. 4, 2015). Ed. by Guy Brock, e0118432. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0118432. URL: https://dx.plos.org/10.1371/journal.pone.0118432 (visited on 07/29/2020).

[45] Xiaosong Wang and Yifan Peng. *ChestXray14-NIHCC Dataset.* URL: https://nihcc.app.box.com/v/ChestXray-NIHCC (visited on 03/23/2020).

[46] Jeremy Irvin et al. “CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison”. In: *arXiv:1901.07031 [cs, eess]* (Jan. 21, 2019). arXiv: 1901.07031. URL: http://arxiv.org/abs/1901.07031 (visited on 03/23/2020).

[47] Alistair E. W. Johnson et al. “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports”. In: *Scientific Data* 6.1 (Dec. 2019), p. 317. ISSN: 2052-4463. DOI: 10.1038/s41597-019-0322-0. URL: http://www.nature.com/articles/s41597-019-0322-0 (visited on 03/23/2020).

[48] *ImageNet database.* URL: http://www.image-net.org/ (visited on 04/01/2020).

[49] Terence Parr, James D. Wilson, and Jeff Hamrick. “Nonparametric Feature Impact and Importance”. In: *CoRR* abs/2006.04750 (2020). _eprint: 2006.04750. URL: https://arxiv.org/abs/2006.04750.

[50] Terence Parr et al. *Beware Default Random Forest Importances.* explained.ai. URL: https://explained.ai/rf-importance/ (visited on 06/17/2021).

[51] Ramprasaath R. Selvaraju et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”. In: *International Journal of Computer Vision* 128.2 (Feb. 2020), pp. 336–359. ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-019-01228-7. arXiv: 1610.02391. URL: http://arxiv.org/abs/1610.02391 (visited on 06/29/2020).

[52] Suraj Srinivas and Francois Fleuret. *Full-Gradient Representation for Neural Network Visualization.* _eprint: 1905.00780. 2019.

[53] Matthew D. Zeiler and Rob Fergus. *Visualizing and Understanding Convolutional Networks.* _eprint: 1311.2901. 2013.

[54] Jost Tobias Springenberg et al. *Striving for Simplicity: The All Convolutional Net.* _eprint: 1412.6806. 2015.