UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Transferring Knowledge to improve classification of Tuberculosis in Chest X-rays

Fábio Miguel Simões Neves

Mestrado em Bioinformática e Biologia Computacional

Dissertação orientada por:
Prof.ª Doutora Sara Guilherme Oliveira da Silva
Dr.ª Anet Potgieter

2021

Acknowledgments

I would like to greatly thank Prof. Sara Silva and Dr. Anet Potgieter for their help, guidance and patience over the last year. Most of my academic and professional breakthroughs were largely inspired by Prof. Sara work which has always pleasantly welcomed me in her office to address any existing issue. Dr. Anet Potgieter was only introduced to me in the last year, however I am very thankfull for her accepting to guide this project. I can say that Dr.[a] Anet has never failed to address my issues in either a work or academic environment, and continues to be a very inspiring person.

Additionally, I would like to thank Vincent Meurrens, Thys Potgieter and the rest of the EPCON team that embraced with open arms regardless of my often limited schedule. I am very thankful for the opportunity and will ensure that my work continues to be worthy of your trust.

I also want to thank my parents, grandparents and Mariana for their unconditional love and support over this last year, looking beyond the stress and burnout.

Lastly, I would like to thank a few friends, that regardless of my mood have always sticked by, to either help, or annoy me further: Ricardo Santos, Marta Melissa, André Neves and Pedro Gomes and so many others who I have the pleasure to share a friendship with.

This work was partially supported by FCT through funding of LASIGE Research Unit (UIDB/00408/2020 and UIDP/00408/2020) and project BINDER (PTDC/CCI-INF/29168/2017).

Abstract

Tuberculosis (TB) continues to be one of the main sources of global health concern, with increased incidence in third world countries in Africa and Southwest Asia, which account for 84% of the 1.5 million deaths due to TB during the year of 2017. The interpretation of X-ray is a strong indicator in the diagnosis of TB which, when combined with other indicators such as cough, fever or other suspicious symptoms, can lead to a very accurate diagnosis. The interpretation of an X-ray image requires the expertise of an experienced Radiologist, a limited resource emphasized by the incidence of TB in third world countries. This interpretation can be assisted through the use of Convolutional Neural Networks (CNN) which, when properly trained, can surpass the performance of health professionals. However, the correct training of CNN requires large amounts of classified images, a resource that does not exist in the public domain for TB. The use of Transfer Learning is a very popular solution when implementing CNNs for the interpretation of medical images, bypassing the wide requirements of images. However, its common implementation tends to not use an effective approach, and few studies explore the advantages of using Transfer Learning. This work seeks to explore the use of Transfer Learning for the optimization of CNN training in very limited TB datasets. Exploration involves the use of Random Baselines and Baselines trained on the large dataset ImageNet, exploring the advantages of Transfer Learning. In addition to these, five additional Baselines are trained on two large-scale X-ray sets, the ChestX-ray8 and the CheXpert, in an attempt to optimize the transfer of knowledge for the classification of TB. The training of models for TB uses the "Shenzhen Hospital X-ray Set" dataset for training, validation and testing. The "Montgomery Hospital X-ray Set" dataset is used for testing purposes only. The result of this work is 155 TB classifiers, for which the best results are achieved using a Baseline trained in the complete set of CheXpert, reaching a median value of 0.65 WAF, and 0.77 of AUROC, on the external test set. Additionally, this work verifies more optimistic results for AUROC measures. This difference results from the threshold used to summarize the output of the networks, for which this work suggests an alternative estimate using a limited number of test data that ends up improving the results of WAF, bringing them closer to the AUROC measures.

Keywords: Tuberculosis, X-ray, Convolutional Neural Networks, Transfer Learning

Resumo

A Tuberculose (TB) continua a ser um dos principais problemas de saúde global na actualidade, com especial incidência em países de terceiro mundo pertencentes a África e Sudoeste Asiático, que somam 84% dos 1.5 milhões de óbitos derivados de TB durante o ano de 2017. A interpretação de Raio-X é um indicador forte no diagnóstico de TB que, quando combinado com outros indicadores como tosse, febre ou outros sintomas suspeitos, pode levar a um diagnóstico bastante preciso. A interpretação de uma imagem de Raio-X requer a competência de um Médico Radiologista experiente, um requisito limitado especialmente considerando a incidência de TB em países de terceiro mundo. Esta interpretação pode ser facilitada através do uso de Redes Neuronais Convolucionais (CNN) que, quando treinadas correctamente, conseguem ultrapassar o desempenho de profissionais de saúde. No entanto, o correcto treino de CNN requer largas quantidades de imagens classificadas, um recurso inexistente no domínio público para TB. O uso de Aprendizagem por Transferência, de fácil implementação para CNNs, é uma solução bastante popular na implementação de CNNs para a interpretação de imagens médicas, contornando os largos requisitos de imagens. Contudo, a sua comum implementação tende a não usar uma abordagem eficaz, e poucos trabalhos exploram as vantagens do uso de Aprendizagem por Transferência. Este trabalho procura explorar o uso de Aprendizagem por Transferência para a optimização do treino de CNNs em conjuntos de dados de TB bastante limitados. A exploração passa pelo uso de Bases de Referência Aleatórias e treinadas no grande conjunto de dados ImageNet, de modo a explorar as vantagens do uso de Aprendizagem por Transferência. Além destes, cinco Bases de Referência adicionais são treinadas em dois conjuntos de Raio-X de larga escala, o ChestX-ray8 e o CheXpert, na tentativa de optimizar a transferência de conhecimento para a classificação de TB. O treino de modelos em TB faz uso do conjunto de dados "Shenzhen Hospital X-ray Set", no qual os modelos são treinados, validados e testados. O conjunto de dados "Montgomery Hospital X-ray Set" é usado apenas para teste. O resultado deste trabalho são 155 classificadores de TB, para os quais os melhores resultados são atingidos usando uma Base de Referência treinada no conjunto completo de CheXpert, atingindo um valor mediano de 0.65 de WAF, e 0.77 de AUROC, no conjunto de teste externo. Adicionalmente, este trabalho verifica resultados mais optimistas pelas medidas de AUROC. Esta diferença resulta do limite usado para sumarizar o output das redes, para o qual este trabalho sugere uma estimativa alternativa usando um número limitado de dados de teste que acaba por melhorar os resultados de WAF, aproximando-os das medidas de AUROC.

Palavras-chave: Tuberculose, Raio-X, Redes Neuronais Convolucionais, Transferência de Aprendizagem

Resumo Alargado

A Tuberculose (TB) continua a ser um dos principais problemas de saúde global na actualidade. Esta doença infecciosa, provocada pela bactéria *Mycobacterium tuberculosis*, causou 1.5 milhões de óbitos durante o ano de 2017, com principal incidência em países de terceiro mundo pertencentes a África e Sudoeste Asiático, que somam 84% das mortes ocorridas. Estes números revelam um grave problema de saúde pública que consequentemente leva a problemas económicos associados. O estudo do tratamento e prevenção de TB em países do Sudoeste Asiático, com principal incidência em adultos e crianças, revela estimativas de custo entre os 677.4 milhões e os 1272.7 milhões de dólares internacionais.

A interpretação de um Raio-X é um indicador forte no diagnóstico de TB que, quando combinado com outros indicadores como tosse, febre ou outros sintomas suspeitos, pode levar a um diagnóstico bastante preciso. A interpretação de uma imagem de Raio-X requer a competência de um Médico Radiologista experiente, um requisito relativamente limitado especialmente quando considerada a principal incidência de TB em países de terceiro mundo. O uso de algoritmos baseados em Redes Neuronais Convolucionais (CNN) revela bastante potencial para a interpretação de imagens no âmbito da medicina, com alguns exemplos ultrapassando o desempenho de profissionais de saúde. Desta forma, este tipo de algoritmos revela-se a ferramenta perfeita para o auxílio do diagnóstico para imagens de Raio-X. O potencial de CNNs na determinação de TB em imagens de Raio-X está limitado à disponibilidade de imagens classificadas para TB. Algoritmos baseados em CNNs requerem números enormes de dados classificados para se atingir o seu desempenho óptimo, um número que não existe para imagens de Raio-X classificadas para TB. Este problema de recursos pode ser resolvido através do uso de Aprendizagem por Transferência.

Aprendizagem por Transferência corresponde ao "melhoramento da aprendizagem numa nova tarefa através da transferência do conhecimento reunido para uma tarefa semelhante já aprendida" segundo Lisa Torrey e Jude Shavlik no seu trabalho "Transfer Learning". Corresponde a um processo muito semelhante à aprendizagem encontrada em animais, em que o conhecimento reunido para uma tarefa vai facilitar a aprendizagem de uma futura tarefa semelhante. Para o caso de CNNs, o conhecimento encontra-se armazenado nos parâmetros treináveis da rede. Estes parâmetros, que podem corresponder a qualquer componente da CNN cujo valor seja ajustado durante o treino, podem ser facilmente usados de uma tarefa para outra, preservando o conhecimento pré-estabelecido. O uso de Aprendizagem por Transferência é bastante popular na implementação de algoritmos baseados em CNN no campo da medicina, devido aos largos requisitos de dados classificados para o treino de CNNs de raiz, e pela relativa escassez de imagens de equipamento

especializado médico, devido a questões de privacidade. O senso comum nesta área chama o uso de Bases de Referência pré-treinadas no grande repositório de imagens naturais ImageNet com subsequente treino em conjuntos de imagens classificadas para condições de saúde. No entanto, existem poucos trabalhos que explorem a real eficácia da Aprendizagem por Transferência na optimização do treino de classificadores de imagens médicas.

Este trabalho explora o uso de Aprendizagem por Transferência para a optimização do treino de CNNs em conjuntos de dados bastante limitados classificados para TB. O seu foco procura explorar diferentes Bases de Referência treinadas (ou não) em conjuntos de dados com propriedades diferentes e avaliar o impacto de cada Base de Referência nos resultantes modelos de TB. Existe também um cuidado acrescido para o uso de dados de acesso público, e uma discussão atenta às condições e dinâmica do mundo real.

O procedimento experimental seguido por esta dissertação explora a Aprendizagem por Transferência através de três tipos de Bases de Referência diferentes: Aleatórias; treinadas em ImageNet; e treinadas em Raio-X. Os dois primeiros tipos de Bases de Referência são directamente adquiridos da plataforma Keras que disponibiliza CNNs com parâmetros iniciais aleatórios, ou já treinados em ImageNet. As Bases de Referência pré-treinadas em Raios-X variam em três factores: o tipo de Base de Referência usado em treino, o conjunto de dados, e o tamanho do conjunto de treino usado. Quanto ao tipo de Base de Referência, cada Base de Referência de Raio-X é treinada a partir de uma rede aleatória, identificada no nome pelo sufixo "-R", ou de uma Base de Referência de ImageNet obtida pela mesma plataforma, identificada no nome pelo sufixo "-I". Quanto ao conjunto de dados, o trabalho usa dois conjuntos de dados de larga escala, o ChestX-ray8 com 112 mil imagens, identificado no nome por "Chest_X", e o CheXpert com 224 mil imagens, identificado no nome por "CheX". O tamanho do conjunto de treino apenas varia para Bases de Referência treinadas em CheXpert. A extensão do conjunto de dados CheXpert é aproximadamente três vezes maior que a extensão dos dados de ChestX-ray8, depois de partidos em subconjuntos de treino, validação e teste. Devido à alta sensibilidade da CNN à quantidade de imagens no treino, são gerados dois tipos de conjuntos de treino usando CheXpert: um conjunto pequeno identificado no nome por "Small", reduzido a 33% do tamanho original, e um conjunto maior identificado no nome por "Big", usado na sua totalidade.

Partindo das Bases de Referência referidas anteriormente, cada uma é usada para o treino de um conjunto de modelos de TB, usando a mesma porção de 69% das imagens contidas no "Shenzhen Hospital X-ray Set", um conjunto 656 imagens de Raio-X classificadas para TB. Para teste, o procedimento usa 15% deste conjunto de dados, dando lugar ao conjunto de teste interno, e 100% das imagens contidas no "Montgomery-County Hospital X-ray Set", um conjunto de 138 imagens de Raio-X classificadas para TB, dando lugar ao conjunto de teste externo. A arquitectura de CNN usada para todos os modelos nesta dissertação é a DenseNet121 recolhida da plataforma Keras, sendo a camada de output alterada para a acomodação das diferentes tarefas. O treino prossegue usando lotes de 16 imagens, e transformação das imagens de treino com inversão horizontal e rotação aleatória de modo a prevenir a memorização das imagens de treino. Os modelos treinados em doenças pulmonares gerais usam o optimizador Adadelta com as definições originais, treinando

até os modelos não registarem perdas de Entropia Cruzada Binária em 5 épocas consecutivas. Os modelos treinados para TB usam o optimizador Nadam com as definições originais, treinando até os modelos não registarem perdas de Entropia Cruzada Binária em 10 épocas consecutivas.

Na primeira fase deste trabalho, referente ao treino das Bases de Referência de Raio-X em imagens de Raio-X classificadas para doenças comuns, são usadas duas medidas diferentes, Área Debaixo da Curva ROC (em inglês "Area Under Receiving Characteristic Curve" ou AUROC) e Precisão Média (em inglês "Average Precision" ou AP), e o teste estatístico de Kruskal-Wallis para identificar diferenças significativas entre cada série de modelos. Verifica-se que a AUROC falha na avaliação dos resultados. Embora seja uma medida bastante usada pela comunidade, esta produz resultados demasiado optimistas para classes raras, algo bastante presente tanto no conjunto de dados ChestX-ray8 como no CheXpert. Ponderando o resultado das medidas de AP, não são verificadas diferenças significativas entre modelos treinados com Bases de Referência Aleatórias e Bases de Referência de ImageNet, com a excepção das classes "Nódulo" e "Massa" para modelos treinados em ChestX-ray8.

Na segunda fase deste trabalho, são treinados um total de 155 classificadores de TB, usando Bases de Referência Aleatórias, de ImageNet, e as Bases de Referência de Raio-X preparadas na primeira fase. As medidas usadas são AUROC e a média ponderada de F-scores (em inglês "Weighted Average F-score" ou WAF), para as quais é determinado um acréscimo no poder de generalização para modelos treinados a partir de Bases de Referência de ImageNet, e Bases de Referência de Raio-X treinadas a partir de ImageNet. É discutido que este aumento de generalização seja o resultado de características mais robustas capturadas através do treino nos 14 milhões de imagens contidos em ImageNet. No geral, a série de modelos de TB mais bem-sucedida é a "TB-CheXBig-I", atingindo um valor de 0.65 para WAF e um valor de 0.77 para AUROC, no conjunto de teste externo. Adicionalmente, este trabalho regista uma notória diferença entre os resultados de AUROC e WAF, com os resultados de AUROC mais otimistas em conjuntos de teste externos. Esta diferença está relacionada com o decréscimo generalizado dos valores de saída da rede, no processamento de dados externos. Este decréscimo faz com que o limite medido em treino para o cálculo da WAF não seja apropriado, levando a resultados mais pobres. Este problema é extensível ao mundo real, no qual é necessário a determinação de um limite para distinguir uma classificação positiva de uma negativa. Este trabalho corrige este problema através da estimação de um novo limite, com base num número mínimo de imagens de teste, de modo a tornar o processo compatível com o acesso a dados de teste no mundo real.

Como trabalho futuro é sugerido o uso de conjuntos de dados ainda mais extensos que os usados neste trabalho, como o MIMIC-III, contendo acima de 300 mil imagens. As Bases de Referência que produzem os melhores resultados são treinadas em conjuntos de dados mais extensos, deixando bastante potencial para o treino de Bases de Referência usando o MIMIC-III. Adicionalmente, são sugeridas algumas experiências que não puderam ser realizadas devido a restrições de tempo e recursos, tal como a integração de Metadados usando tecnologias híbridas CNN-BN.

Palavras-chave: Tuberculose, Raio-X, Redes Neuronais Convolucionais, Aprendizagem por Transferência

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Tuberculosis (TB) continues to be one of the main causes of global health concerns. In 2017, it killed 1.5 million people worldwide, 84% of these deaths taking place in Africa (665 thousand deaths) and South-East Asia (666 thousand deaths) (Floyd et al., 2018). Most procedures for diagnosing chest TB do not evaluate Chest X-rays without an assortment of wet lab procedures to confirm the infection. However, Chest X-rays can rule out the presence of chest TB, being commonly used in the triage of potentially infected subjects, and assessing the development of the disease. Many procedures order for a mandatory Chest X-ray for 14-year-old applicants or older, such as those found in the instructions of the Centre for Disease Control of the United States for the Medical Examination for Immigrant or Refugee Applicants (DS-2053) (for Disease Control et al., 2006).

The correct interpretation of Chest X-rays requires trained radiologists' expertise, a scarce resource in poor communities, which hold the largest percentage of people at risk of contracting TB. Algorithms based in Convolutional Neural Networks (CNNs), when properly trained, can outperform field specialists' performance, as the work of (Rajpurkar et al., 2017) shows. However, proper training requires an extensive amount of images labelled for TB, an amount that, to the best of our knowledge, is currently unavailable for public access, with the Shenzhen and Montgomery TB X-ray sets providing a total of 814 images. The lack of TB labelled images for the training of CNN based TB classifiers calls for Transfer Learning.

Transfer Learning promotes the reuse of the knowledge gathered from one task to the other (Torrey & Shavlik, 2010), similar to the biological learning process of animals, where the stored knowledge from one task heightens the learning of another task. The trainable parameters of the CNNs hold their knowledge. Software providers such as Keras (*Keras Documentation for Aplications*, 2020) even distribute pre-trained CNNs, or Baselines on the very large ImageNet dataset (*Towards Fairer Datasets*, 2020), containing 14 million images. The Baseline trainable parameters encode robust features in the higher layers near the input, such as edges, shapes and other simple patterns (Y. H. Liu, 2018), reusable in different tasks for the extraction of visual information. Baselines trained in the vast ImageNet dataset provide features that could not be learned without an extensive training set, improving training procedures on smaller X-ray datasets (Gozes & Greenspan, 2019). These features result from a dataset composed of everyday subjects,

which might not pose an optimal Baseline for handling X-ray images. The ImageNet Baselines trained in large general Chest X-ray datasets can theoretically serve as a more robust Baseline for learning on smaller, more specific datasets. Large datasets such as ChestX-ray8 (Wang et al., 2017) and CheXpert (Irvin et al., 2019) provide thousands of public images access and portray a much closer problem to the classification of TB.

The use of ImageNet Baselines for CNN implementations is very popular in Radiology and is used extensively by previous radiology works. However, to the best of our knowledge, hardly any work fully explores the actual improvements using Transfer Learning from an ImageNet Baseline to X-rays. Even less research is dedicated to Baselines' training on Large Chest X-rays to tackle training on smaller X-ray datasets. Works such as (Raghu et al., 2019) defend that the impact of Transfer Learning for the training of Chest Disease Classifiers is minimal. Their work evaluates Transfer Learning's effectiveness using an ImageNet Baseline to benefit Chest Classifiers' training on the CheXpert X-ray set, an extensive dataset. It fails to show how Transfer Learning, using pre-trained Baselines on Large Chest X-ray datasets, benefits classifiers trained on small datasets, such as those available for TB. The work of (Gozes & Greenspan, 2019) explores Transfer Learning using Baselines pre-trained on the Chest X-ray 8 dataset to benefit the training of TB classifiers trained on small datasets. Their work shows improved results using Chest X-ray Baselines. Still, they focus on an alternate problem, showing only the results of two models, one trained with an ImageNet Baseline, and another trained with a Chest X-ray Baseline, but not focusing enough on this issue.

1.1 Objectives and Contributions

This work explores Transfer Learning for the training of TB models, using a small amount of labelled images. It aims to explore different Baselines and measure effective improvements when used to train TB models. The exploration will involve collecting public access Chest X-ray datasets labelled after non-specific chest diseases, and TB labelled datasets. The collected datasets provide the training data for Chest X-ray Baselines. These Baselines will train with and without Transfer Learning, providing the Chest X-ray Baselines for the training of TB models. Different TB models train using ImageNet and Chest X-ray Baselines. A single series of models train without Transfer Learning, serving as the control for our experiment. The resulting TB models are gathered in the end to determine how the Baseline used affects model performance. The main contributions of this work are the following:

1. Advancement of the state of the art of Computer Automated Diagnosis. This work represents the only thorough Transfer Learning study for Chest X-ray Images, training TB models on a minimal dataset using Random Baselines, ImageNet Baselines and Chest X-ray Baselines. For Chest X-ray Baselines, this work goes on to explore its effectiveness when trained in either the Chest X-ray8 (Wang et al., 2017) or CheXpert dataset (Irvin et al., 2019). This exploration examines the impact of the dataset properties and size for Baselines' training used in Transfer Learning.

2. Exploration of metric related issues for the evaluation of model performance. This work develops an improved approach for threshold estimation to transform the TB models' continuous output into binary labels for the measurement of F-score. The strategy used for threshold estimation tailors a real life setting, where the optimal threshold in the training data may not be optimal for a new task. It uses a minimal amount of labelled data from the target task to estimate an optimized threshold.

3. An extensively trained baseline that can be used by future works to optimize the training of Chest X-ray Classifiers for diseases with limited labelled data. This Baseline consists of DenseNet121 architecture trained in the CheXpert dataset. It will be published in open access in the future, together with the paper (currently in preparation) that documents our findings.

4. A poster in the 2019 LASIGE workshop, under the name "XrayAme – Combining Knowledge to Improve Classification of Tuberculosis in Chest X-Rays", briefly introducing the ideas behind this work.

1.2 Document Structure

The document is organized as follows. Chapter 2 covers some major concepts related to this work, providing some background to the subject. Chapter 3 explores the related work of the last five years surrounding this theme. Chapter 4 describes the general approach used for the exploration of Transfer Learning. Chapter 5 describes the outcomes of the experiments. Chapter 6 explores and discusses the results obtained in the previous section. Finally, Chapter 7 concludes the work and presents some ideas for future work.

Chapter 2

Concepts

2.1 Tuberculosis

TB is an old epidemiological problem, often forgotten due to its eradication from most modern communities in recent years. However, it still takes the lives of millions of people every year, especially in poorer communities that lack the infrastructure to deal with TB outbreaks. This section provides the facets surrounding TB in the modern world and ultimately entices this work into fruition. Section 2.1.1 describes the economical problem of TB in the real world, section 2.1.2 portrays the epidemiological dynamic of TB and finally section 2.1.3 provides the current approach to the diagnosis of TB.

2.1.1 The cost of TB

TB is predominantly a third world problem, with 84% of its casualties taking place in Africa and South-East Asia (Floyd et al., 2018). In 2005, a study checked the rough estimates of the average treatment cost of TB in South-East Asia, with high adult and infant TB incidence. The estimates come to a minimum of 677.4 million and a maximum of 1272.7 million international dollars[1] (Rob et al., 2005). However, modern countries still tackle with TB, with another study from 2013 measuring an average cost per person infected with TB in the old EU-15 states, Cyprus, Malta and Slovenia. This study determines a cost of about 10 thousand euros for normal TB, 57 thousand euros for multidrug-resistant TB (MDR-TB), and 170 thousand euros for extensively drug-resistant TB (XDR-TB). Values remain 2 to 3 times lower for the remaining European Union, with 3 thousand euros for susceptible TB and 24 thousand euros for MDR- TB/XDR-TB[2].

The economic burden of TB in the EU, regarding a range of related health expenses, achieved five thousand million euros in 2012 (Diel et al., 2014). The large difference in incidence between Africa and South-East Asia compared to the rest of the world does not translate very well into monetary values. Therefore one should reason over the regional differences in 3rd World Countries with lower incomes and limited medication access.

[1] also known as the Geary-Khamis dollar, a hypothetical currency unit with the chronological value of the dollar

[2] The paper stated the same numbers for both MDR and XDR TB

Table 2.1: Even though TB is not a predominant concern in modern Europe, the costs related to the treatment of this disease are still very high, leaving room for modern, more cost-effective methods.

Accounted EU Countries	TB(€)	MDR-TB(€)	XDR-TB(€)
Old EU and Cyprus, Malta, Slovenia	10 282	57 213	170 744
Other	3,427	24,166	24,166

2.1.2 TB contagious patterns

Mycobacterium tuberculosis, the bacteria responsible for the infection with TB, is carried in aerosols whenever an infected individual exhales or coughs, with the bacterial agent remaining active for up to 12 hours (Schwartzman & Menzies, 2000). It takes prolonged exposure for one individual to infect another, placing healthy individuals interacting with a carrier of TB bacilli in a social setting at exponentially higher risk, such as schools (Sacks et al., 1985; Rogers, 1962), hospitals (Haley et al., 1989; George et al., 1986), or familiar settings (Hahn, 1943; Spector, 1939; PATERSON et al., 1940). Environments with limited air circulation such as hospitals without proper air quality management represent a risk of infection of uninfected patients from infected patients placed in different wards, compromising the containment of the disease (Sultan et al., 1960; Riley et al., 1959).

The infection only progresses to the active form of the disease in about 10% of infected individuals, because of the immune system response that leads to the formation of granulomas. The eradication of the infection carries out in about 10% of cases, with the 90% left developing Latent Infection (LTBI), an asymptomatic state of the condition (Ilievska-Poposka et al., 2018). LTBI possesses a cumulative reactivation risk of 9.5 years (Sloot et al., 2014), with half of such cases registered within the first five years (Styblo, 1985). The reactivation risk will depend on external factors to the disease such as age (Marais et al., 2004) or infection by Human Immunodeficiency Virus (HIV) (Akolo et al., 2010).

TB achieved a reproducible rate of 4.3 in China in the year 2012 (Zhang et al., 2015), meaning that on average, each person infected with TB will infect at least four other people. It links these rates with the total yield of infected people in each country, population density, cultural patterns, and social education. Recent estimates on the global burden of LTBI cases shows 1.7 thousand million people infected in the year 2014, representing a quarter of the world population at risk of developing and spreading TB in their lifetime (Houben & Dodd, 2016). The silent presence of latent TB in modern communities can slow the fight set in place by large initiatives like Stop TB Partnership, which among other milestones, plans to treat 29 million people with TB and prevent 49 million people from contracting the disease (*Stop TB Partnership — The Global Plan to End TB — The Global Plan to Stop TB 2016 - 2020*, 2019).

2.1.3 Diagnosing TB from X-ray

Early diagnosis of LTBI is crucial in the combat against TB spread, to comply with the Stop TB targets for prevention and treatment of people from 2016 to 2020 (*Stop TB Partnership — The Global Plan to End TB — The Global Plan to Stop TB 2016 - 2020*, 2019). Common guidelines often require a combination of Tuberculin Skin Test (TST), Interferon-gamma Release Assay (IGRA), and chest radiography (X-ray) to screen for TB (Bothamley et al., 2008). Unlike TST or IGRA, the results from a Chest X-ray are not always clear. Studies show that the image's quality may have harmful effects in the radiologist's assessment, leaving visible conditions out of the completed report (Siewert et al., 2008). It is important to consider the limitations of human reading and interpretation, since factors such as previous experience (Renfrew et al., 1992), communication between doctors and radiologists (Brady et al., 2012), and the workload of the radiologist in a given setting (FitzGerald, 2013), all can contribute to wrong or incomplete assessments of the image at hand, with X-ray imaging serving as a complementary step in the diagnosis. Nonetheless, it is an important tool to determine false negatives from bacteriological tests, which is a running issue in children were these tests often deliver false negatives (Organization, 2014).

The analysis of a Chest X-ray for TB diagnosis will depend on the local guidelines. A patient assigned for examination may undergo a series of three sputum smear tests. If the three tests' outcome turns positive, the patient does not require any further assessment by chest radiology. However, the medical practitioner in charge may order it if the subject:

- **Shows breathlessness** , because of pneumothorax, pericardial effusion or pleural effusion that requires specific treatment.
- **Coughs blood** , also known as haemoptysis, to exclude other conditions.
- **Tests positive for only one smear test** , requiring the follow-up inspection of the Chest X-ray.

Upon closer examination a radiology expert can identify abnormalities in the Chest X-ray, such as cavitation, upper lobe infiltration, bilateral infiltrates, and pulmonary fibrosis and shrinkage, among other potential phenomena associated with the development of TB. These abnormalities in the Chest X-rays are not a guaranteed factor. Depending on the condition's development and the radiologist's experience, the diagnosis from a chest radiology image may vary considering the resulting accuracy (Burrill et al., 2007). These factors further stress the usefulness of unbiased and accurate computer algorithms, capable of identifying abnormalities in Chest X-rays to address the practitioner's attention and reduce the number of lost features.

2.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs), introduced in 1980 by Kunihiko Fukushima through the "neocognitron", takes inspiration from the visual cortex's biological design. This inspiration leads to the establishment of the architecture whose convolutional and downsampling layers allow for

correct image recognition regardless of shifting relative to a visual scene (Fukushima, 1980). Large improvements have been made to CNN designs to achieve maximum performance in the classification of images with the least overfitting. However, their architecture derives from a combination of stacked segments, Normalisation, Pooling, Convolution, and Fully Connected Layers, also known as Dense Layers. This section describes some basic concepts for CNNs. Section 2.2.1 explains the base components of CNNs, and section 2.2.2 explains its training process.

2.2.1 Properties of CNNs

CNN show a very modular nature to them using a collection of stacked building blocks named CNN layers, each performing basic tasks namely **normalization**, **pooling**, **convolution**, and **fully connected layers**.

- **Normalisation** layers are commonly used in the pre-processing of the input from one set of layers to the other to reduce the presence of outlier values and convert each value into a unit (x"). The process results from the division of each pixel's value (x') by the mean standard deviation of all the pixels in the image.

$$x^{''} = \frac{x^{'}}{\sqrt{\frac{\sum_{i=1}^{N}(x_i - \hat{x})^2}{N-1}}} \tag{2.1}$$

- **Pooling** layers use a hovering matrix of a given size that iterates over an input feature map while using simple functions to extract the covered portion by a single value. An effortless pooling operation is max Pooling, where the maximum value held by the part is selected, as shown in fig. 2.1. Another simple pooling operation is average Pooling, which calculates the mean value of the enveloped collection. The stride length configuration fine-tunes of the pooling layers, i.e. the number of pixels transitioned by the kernel in length and height, and the matrix's size. This type of task will proceed on the downsampling of the input to a more compact representation, which will have a minimal effect when faced with moderate changes from the view of the object (Goodfellow et al., 2016). The final feature map will have its height (h') and length (l') dependent on the height (h) and length (l) of the original feature map, as well as the stride length (s) and the size of the hovering matrix (f).

$$h' = \frac{h - f + s}{s}, l' = \frac{l - f + s}{s} \tag{2.2}$$

- **Convolution** layers, a little bit like the pooling layers use a hovering matrix, called a convolutional filter/kernel, although unlike the pooling layer, the filter contains values. The filter is multiplied by the covered patch in the input feature map, for which resulting values are then summed to a total and retrieved to a single value in the output feature map. Depending on the filter's size and the feature map, this convolutional filter can downscale or maintain the

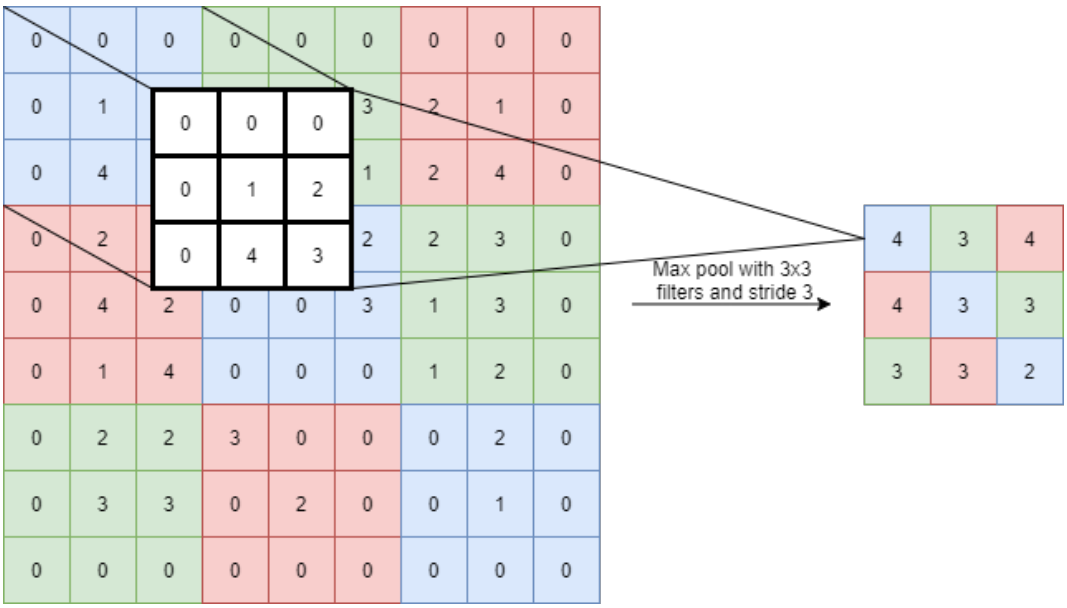

Figure 2.1: Hovering 3x3 matrix with a stride of 3 units performing max pooling operations. This allows the down sampling of the original input, promoting lower processing demand with low impact in performance.

original image resolution, with downscale being useful due to the cut down on parameters and the cost on processing power. As in the pooling layers, the modification of the filter size, and stride length fine-tunes the convolution layers' action. However, zero-padding is also essential, since zero paddings' different values may result in the loss or maintenance of the original image size. To maintain the initial image size, the user needs to use full Convolution. For fxf sized filter, full Convolution requires a zero-padding with length equal to f-1. The height (h') and length (l') with padding depend on the height (h), length(l), filter size(f) and stride length (s), similar to the pooling layers with the addition of a length of the zero-padding around the border (p).

$$h' = \frac{h - f + s + p}{s} l' = \frac{l - f + s + p}{s} \tag{2.3}$$

Convolutional layers convert the normal original images into meaningful features like edges, shapes, gradients while scaling down the matrix. It reduces the number of parameters and computational processing power, converting the information shared by the network as a set of features, that should reduce the network's likeliness to memorise the image, and provide a more meaningful interpretation (Lawrence et al., 1997). These features are hierarchical, with higher-order (simple) features like edges captured near the input, and lower order features (complex) features like patterns captured next to the Fully Connected layers towards the end of the network (Y. H. Liu, 2018).

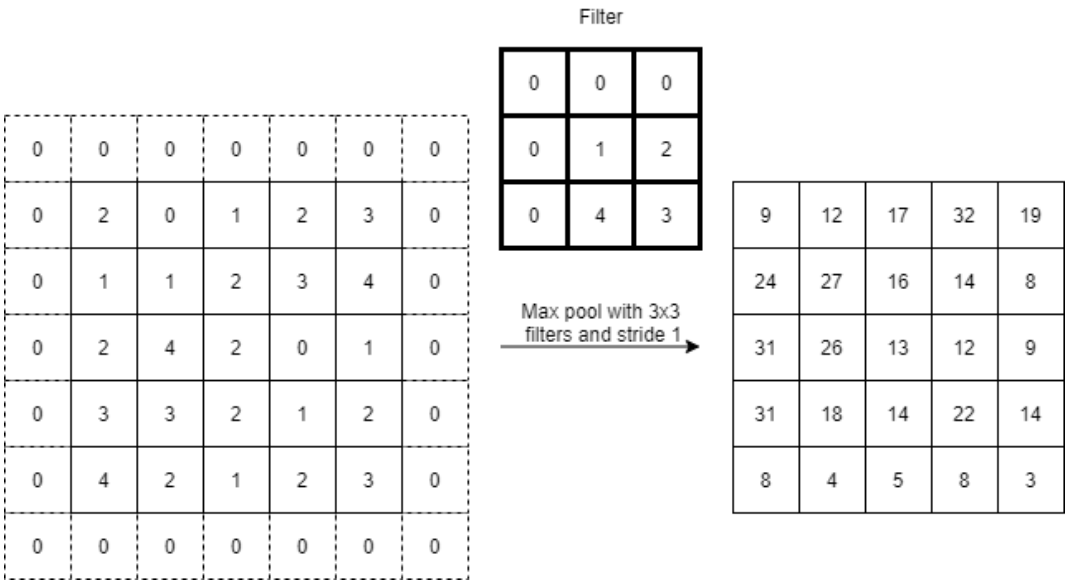

Figure 2.2: Although convolutional layers are many times used for downsampling, in tasks such as image denoising and hyper resolution, maintaining the same image size is useful. Without zero paddings, a stride of 1x1 presented in the figure would not return a feature map with the same dimensions as the input.

- **Fully Connected** layers follow the flattening the input from the previous convolutional and Pooling layer, allowing the network to learn a solution based on the features extracted from the image. The final layer is responsible for the input classification, where its length is directly related to the number of unique labels assigned to the images fed through the network.

2.2.2 Training CNNs

CNNs were a novel topic in 1980. Still, it wasn't until the millennium turn that faster implementations on Graphical Processing Units (GPU) stirred some interest on CNNs, which were competing with other algorithms with lower resource cost-effectiveness (Chellapilla et al., 2006; Steinkraus et al., 2005; Oh & Jung, 2004; Hinton et al., 2006). The use of back-propagation (Hirose et al., 1991) alongside GPU processing has established itself as the popular standard among the community, with its results verified by the state of the art (He et al., 2016; Schmidhuber, 2015; Cireşan et al., 2010; Ciregan et al., 2012; Ciresan et al., 2011).

When using a feed-forward CNN, the network's input is propagated through the network in a process called "forward propagation". The output is held in the final layer, summarising the information passed through each of the hidden units contained in the network, from which the Loss can be calculated (Hirose et al., 1991, pp. 200–220). The Loss, also mentioned as cost function or error, is used to measure the difference between the training truth values, and the network output. The Loss Function outcome can then compute the gradient, also known as derivative, of the function contained within each node in the network. This computation accomplishes by passing

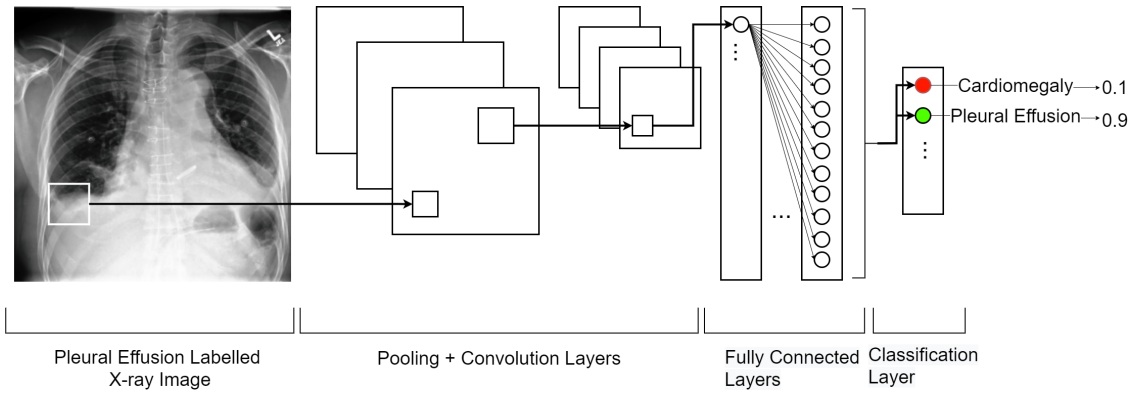

Figure 2.3: X-ray Image from the CheXpert dataset. Fully connected layers placed near the end of the architecture learn the correlation of the truth labels and the features extracted from previous layers. This setup allows a highly accurate classification of the images regardless of rotation, scale, or contrast in the original picture.

information from the Loss Function to each unit placed before a given starting unit, in a backwards fashion, embodying the term back-propagation.

Although being often misunderstood as a learning algorithm, back-propagation represents only the process of flowing information backwards through the network to calculate the gradient. Iterative algorithms, also known as optimisers, such as the stochastic gradient descent, achieve the actual learning, aiming to minimise an objective function, for a given parameter and a set of observations gathered from the dataset (Taddy, 2019, pp. 303–307). Stochastic Gradient Descent (SGD) is a single representative of a whole range of optimisers available over the years. SGD, RMSprop, Adam, Adadelta, Adagrad, Adamax, Nadam, and Ftrl are examples of optimisers provided by services such as Keras (*Keras Documentation for Optimizers*, 2020).

The use of back-propagation and optimiser algorithms limits the training to labelled datasets for supervised learning. However, other research shows quite interesting workarounds on this limitation. The work in (Xie et al., 2019) shows that by using the best performing models in a given stage for the expansion of the training data with unlabelled data, it is possible to train regular CNNs with back-propagation in a semi-supervised manner. They achieve this using the trained models for labels' assignment to the unlabelled data to expand the training datasets with incorrectly labelled data. Noise used directly in the labelled images avoids bias problems, and the label quality improves along with the knowledge of the resulting models.

2.3 Transfer Learning

Transfer learning is "the improvement of learning in a new task through the Transfer of knowledge from a related task that has already been learned" (Torrey & Shavlik, 2010). It is exciting on machine learning algorithms with very high data requirements for the achievement of optimal performance. Gathering high amounts of data is especially challenging in supervised machine

learning. Not only is labelled data harder to obtain, but special care is required when considering the available labels, since wrongfully labelled data may be present in either manual or autonomous labelling processes.

This section explains how to perform Transfer Learning using CNNs in section 2.3.1 and the used metrics to evaluate its effect in section 2.3.2.

2.3.1 Transfer Learning with CNNs

Transfer Learning for use on CNN training is quite simple. The trainable parameters stores the knowledge gathered during training within the network. The Convolutional Layer filters, Dense Layers and any other trainable network component which value adjusts through back-propagation make up the network's trainable parameters. The Transfer of this knowledge only requires the modification of the last layers to accommodate the new task at hand, easily performed using the Keras Functional API (*Keras Documentation for Functional API*, 2020).

Keras and other software providers already provide trained networks in the ImageNet dataset. These networks have very well developed features trained for the classification of general everyday subjects, knowing how to detect edges, shapes and other high-level features useful for extracting visual information. Transfer Learning used for training on smaller datasets contributes effectively to the network's higher-order features, found in the convolutional layers near the input. On very large CNNs, these initial convolutional layers require extensive training procedures for Back-propagation to make any changes to it, whereas shorter training procedures leave these layers mostly unchanged (Raghu et al., 2019). Transfer Learning from the ImageNet dataset can be especially useful in faster training procedures, providing high-level features that the network would otherwise not learn during training.

2.3.2 Metrics

To measure the improvement or degradation of the models when attempting to use transfer learning, three measures are used to discuss its appropriateness for a given couple of source-target tasks (Torrey & Shavlik, 2010).

- **Initial performance of the transferred** , knowledge without training on the target task. Using the example of a DCNN, this would be equivalent to loading the trained model from disk, with the trained weights adjusted for the source task, fine-tuning the final classification layers for the new task, and measuring its performance in the target task. It allows the user to acknowledge how the ignorant target model performs with just the transferred knowledge.

- **difference in learning time towards full performance** , between the model trained from scratch, and the model using transferred knowledge. As was explained before,

- **Final performance** , which compares the model trained from scratch and the one with transferred knowledge to guarantee performance improvement. In some cases, this may not be true, resulting in a loss of final performance.

Depending on the amount of training data available for a given target task, transfer learning is usually an exciting proposition to evaluate. However, even for small datasets, if the source-target task pair are not sufficiently correlated, transfer learning can aggravate performance already achieved in default training with a limited dataset. This Loss of performance compared to the lack of knowledge transfer is known as Negative Transfer and highlights the essential roles of transfer learning metrics.

Chapter 3

State of the Art

In the early 2000s, CNNs were computationally expensive and very prone to overfitting. Other classical machine learning methods were more popular, requiring much less data than current Deep Learning techniques. In the last few years, CNNs have already proved themselves powerful machine learning algorithms capable of high generalization on suitable amounts of data. The ImageNet dataset, released in 2010, follows the popularization of CNNs in the latter half of the 2000s, with processing power rising exponentially at a lower cost for the consumer market and other breakthroughs in training optimization, such as the use of GPU. This dataset comprises 14 million images over 20,000 categories (Deng et al., 2009).

The ImageNet challenge hosts constant breakthroughs in machine learning and image classification. These breakthroughs often show innovative solutions that fit with the needs of the current technologies. An example of this is the model under the name "Noisy Student". Currently taking the fourth place at the time of writing this paragraph (28-12-2020), it uses a custom EfficientNet L2 architecture (Tan & Le, 2019), achieving 88.4% top-1 accuracy on ImageNet (Xie et al., 2019). This work proposes a self-training workflow to improve the performance of existing architectures. It initially trains a model on the existing labelled data in a fully supervised manner until said model plateaus' performance. Following this initial training stage, the trained model is used to classify unlabelled data to increase the training dataset's size, with a new model training until it plateaus again. Given that the new model performs better than the original model that provided the virtual labels, it replaces the previous model. It provides its virtual labels for another round of training. This process repeats indefinitely to improve the quality of the virtual labels while expanding the training dataset. It is, by itself, an exciting workflow for the problem discussed in this thesis, since there is a lack of adequately labelled images, especially in the medical field.

Open Access medical X-rays are very limited in nature. Such images must be anonymous before distribution, ensuring patient privacy, provided by the hospital entities that own the images and reports obtained in a medical environment's daily routine. Once published, the pictures either come with pre-assigned classes or with a textual description processed through Natural Language Processing (NLP) techniques, for the training of supervised machine learning algorithms (such as CNNs). Another difficulty for developing Deep Learning algorithms for medical purposes is the

severe bias in clinical data. On the one hand, different healthcare environments may capture images using various equipment at different stages. The protocols used for identifying a given disease may also change along with the error from the human practitioner. On the other hand, diseases are rare, limiting the number of samples available for a given problem. This limitation calls for optimizing the training procedures to make the most out of the limited data available.

The rising trend in the release of Open Access data provides room for more diverse and improved algorithms. It is beneficial for algorithms with increased data demands such as CNN. On the remainder of this chapter, section 3.1 portrays a timeline of the most relevant Chest X-ray datasets, used directly or indirectly in the preparation of TB classifiers, and section 3.2 describes the most pertinent works released in the last few years for the classification of TB. And finally, section 3.3 inspects the existing works that explore Transfer Learning for medical applications.

3.1 Chest X-ray Datasets

The PLCO (Prostate, Lung, Colorectal and Ovarian) Lung Dataset (Team et al., 2000), released in the year 2000, contains 185,421 X-rays and their respective textual reports. The data does not ship with classes; however, the sheer amount of images has made it the chosen dataset for Abnormality classifiers' training such as in (Guendel et al., 2018). The images for this dataset result from oncology screenings, and therefore neither the reports nor images show findings directly related to the classification of TB. However, the number of images makes it a great candidate for the training of X-ray Baselines.

The year 2014 marks the release of the Shenzhen Hospital TB X-ray Set, and the Montgomery-County TB X-ray Set (Jaeger et al., 2014), essential for the training of the TB classifiers, although hosting a small number of images. These datasets provide Chest X-ray images classified for TB, along with the textual reports containing information on the position of the image, age and gender of the patient. The Shenzhen X-ray Set provides lung masks (i.e., images delineating the lungs' boundaries), very useful for the creation of machine learning algorithms capable of extracting the lung fields from the image. When used in conjunction with TB classifiers, it allows the models to learn without the noise generated by the area outside of the lung boundaries. The majority of the TB classifiers mentioned further along (Lakhani & Sundaram, 2017; Huang et al., 2017; Islam et al., 2017; Gozes & Greenspan, 2019) use these two TB datasets for training and testing.

The Indiana Open I dataset, originating from the Indiana University and released in 2016, provides 7,470 Chest X-rays. This dataset contains images automatically annotated for multiple conditions, making it a good contender for Chest X-ray Baselines' training. However, they are limited, especially compared with newer open-access datasets containing upwards of 100 thousand images, like the ChestX-ray8 and the CheXpert datasets released in the subsequent years.

The ChestX-ray8 dataset is released in 2017 by the National Institute of Health, an agency of the U.S. Department of Health (Wang et al., 2017). On release, the dataset contained only eight classes assigned to each image, with six more common thorax diseases added with a follow-up update to the dataset. The mining of the disease labels makes use of NLP, that extracts mentions

and negations for each of the diseases, creating the necessary tags for supervised training. However, the original paper does not provide any hand labelled images that can guarantee a certain level of truth, since the mining procedures are prone to mistakes that a human experts committee is much less susceptible to make.

The CheXpert dataset, released in 2019 by the Stanford Machine Learning Group, tries to consolidate the truthfulness of benchmarks in the classification of Lung Diseases. The team also uses NLP to extract labels from radiology reports, generating different classes than those found for the ChestX-ray8 dataset. The dataset provides a separate validation set of 420 images examined and labelled by a radiology expert team to guarantee that the models have a knowledgeable ground-truth. The authors do not publish the testing set publicly to maintain the published results' integrity on this dataset. The testing results are only available through the submission of the trained model as an executable file to the CheXpert sponsored Contest.[1]

The latest X-ray dataset released is the MIMIC-CXR. Another bigger and richer source of Chest X-ray images, released in 2020 with the newest version (at the time being v2.0.0) containing 377,110 images from radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston (Johnson et al., 2016). These X-rays include the radiological reports attached to them, processed through the use of Neg Bio, and the CheXpert tool, provided as part of the CheXpert dataset, meaning that both CheXpert and MIMIC-CXR share the same classes, useful for the generation of external tests.

3.2 CNNs for Chest X-ray Images

In 2016, a team of researchers used a custom version of the AlexNet architecture to tackle the classification of TB (Hwang et al., 2016)[2]. The work used KIT (Korean Institute of Health), a relatively large private dataset with 10,848 DICOM images[3] for the training procedure, comprising 7020 normal and 3828 abnormal (TB), divided into 70%, 15% and 15% for the Training, Validation, and Testing sets, respectively. The best performing models scored 0.877 AUC score in the Montgomery County Dataset and 0.919 AUC score in the Shenzhen Dataset (Hwang et al., 2016).

On the same year, a team conducted similar research (Cao et al., 2016), training a variation of GoogLeNet pre-trained on the ImageNet dataset, for the classification of Lung TB. However, their results are dubious, since they take a very unbalanced private dataset provided by Peruvian partners at "Socios en Salud". This dataset comprises 453 images with no findings, and 4248 abnormal images with indications of Lung TB, without mention to a class-wise balancing of the dataset, or data augmentation of any sort. The paper also does not state any benchmark in an external dataset that could rule out overfitting. The article reports 89.6% accuracy, which is an inappropriate measure to use for such an unbalanced dataset since the stated value could

[1] accessible at https://stanfordmlgroup.github.io/competitions/chexpert/

[2] AlexNet gathered much attention winning the ImageNet challenge in 2012 and introducing dropout layers that help reduce overfitting by randomly "forgetting" connections between layers.

[3] DICOM or Digital Imaging and Communications in Medicine is the standard format used in the communication and management of medical images. It holds information used to ease the interoperability of medical image systems.

represent the dataset class ratio when the trivial model determines the same class for every image. Average Precision or F1 score are much better metrics since they consider both precision and recall. Regardless of the high reported accuracy, their average class-based accuracy is ultimately very low, scoring 62.07% (Cao et al., 2016), which validates previous concerns.

By the following year, a team composed by most of the primary authors of the previous work released a new model named "TX-CNN", as an attempt to test other architectures, and improve the management of the poorly balanced dataset provided by "Socios en Salud" (C. Liu et al., 2017). Faced with the low representation of some of the classes, visible in Table 3.1, the team attempted to balance their data during training, by assigning a maximum number of images for each category, followed by random sampling with replacement. The GoogLeNet model achieved an average class-based accuracy of 91.72%. This work used the F1 score to test the model performance, achieving up to 0.95 F1 for some classes in their test dataset. However, this is also a very doubtful report, since the text states that the over-sampling procedure used for balancing data carried out repeated data from the training set to the testing set. Therefore, the possibility of overfitting cannot be discarded.

Table 3.1: Data distribution of the dataset provided by "Socios en Salud". Abnormal images such as Miliary Disease and Ghon Focus are severely under represented against the other classes, which can lead the model to underperform on the classification of these images.

Category	Number of Images
Miliary Disease	25
Cavitation	1182
Lympahadenopathy	202
Ghon Focus	27
Alveolar Infiltrates	2252
Other	560

One 2017 work (Lakhani & Sundaram, 2017) that shows outstanding results, with minimal room for criticism, uses AlexNet and GoogLeNet, trained on a collection of data comprising 1007 patients. From this data, 492 images show signs of TB manifestation. The X-rays are the product of multiple combined datasets, including the Shenzhen, Montgomery, and the privately provided Belarus TB public Health Program and Thomas Jefferson University Hospital TB datasets. In the results, deep image augmentation and the pre-training of the models in the ImageNet dataset show the best performance on models trained for a single class, with image augmentation showing promising results in improving model results seen in Table 3.2. The GoogleNet and AlexNet models achieve an AUC value of 0.98%, and a staggering 0.99% AUC when using models' ensembles. None of the stated results involves an external dataset, with the test dataset corresponding to a subset of 14.9% from the original combined dataset. An external dataset would portray how the model performs when faced with unfamiliar sources of X-ray images. Testing on an internal dataset may

lead to optimal results that are not observed when testing occurs in a dataset that does not share the original training data's properties, such as an external testing set.

Table 3.2: Data representing the performance of the AlexNet, GoogLeNet and Ensemble models for classification of TB. Augmented stands for the additional use of deeper augmentation of the images during training.

Architecture	Untrained	Pre-trained	Augmented Untrained	Augmented Pre-trained
AlexNet	0.90	0.98	0.96	0.98
GoogLeNet	0.88	0.97	0.94	0.98
Ensemble	-	-	-	0.99

In 2018, a team from the University of Stanford addressed the automatic diagnosis of Pneumonia in radiology images (Rajpurkar et al., 2017). When trained, the CheXnet CNN model diagnosed Pneumonia with a higher F1 score than the combined average of four radiology practitioners. DenseNet121 was the chosen CNN architecture (Huang et al., 2017), pre-trained in ImageNet and provided by Keras. Benchmarking carries out in the CheXpert testing set, and with the aid of a team of four radiology practitioners to classify the same testing set blindly. The model achieved an F1 score of 0.435, overcoming the combined average of the four invited radiologists, that attained an F1 score of 0.387. Radiologist 4, the most seasoned practitioner with 28 years of experience, was the only one able to slightly surpass the algorithm, achieving an F1 score of 0.442.

Another work (Islam et al., 2017) explores the AlexNet, VGG and ResNet CNN architectures for the classification of chest abnormalities. The datasets used include the publicly available Indiana Dataset (Demner-Fushman et al., 2016), comprising 7284 frontal and lateral Chest X-rays, with annotations for Cardiomegaly, Pleural Effusion, Pulmonary Edema, and Opacity, gathered from the Indiana University School of Medicine; the private Japanese Society of Radiological Technology (JSRT) dataset, with 247 images annotated for nodule abnormalities; and the previously mentioned Shenzhen Tuberculosis with 662 images labelled for TB. This work sets out to explore the extraction of features from different layers, followed by fully connected layers for the disease classification, to determine which depth maximizes the performance of the models. The authors report better overall results for models trained with features gathered from upper convolution blocks, specifically the ones retrieved from the RELU activation layer in the block Res4B from a ResNet152 model architecture, which complies with the conclusions of (Lakhani & Sundaram, 2017). The paper also tests ensembles of models, containing 1 to 24 different CNNs. The results show improvements in classification robustness when using larger ensembles. The model achieves an AUC score of 0.94, although it is unclear where the team gathers this result.

In 2019, a team of researchers (Gozes & Greenspan, 2019), inspired by the outstanding results achieved in (Huang et al., 2017), trained the same architecture for the classification of TB. Their work follows the steps taken in the CheXnet paper, training a DenseNet121 in the CheXpert

dataset to classify the 14 labelled chest conditions, while following the same procedures for both image normalization and augmentation. The determination of TB is possible with the replacement of the classification layer by a single channel dedicated to TB classification, followed by training on the publicly available Shenzhen TB dataset. Since the Shenzhen TB dataset provides the images in DICOM format, containing additional information like gender, age and position, the authors create a second different custom model named MetaCheXnet. The authors customize the classification layer for the classification of TB, Gender, Age and Position. Benchmarking carries out for both the models in the Montgomery County TB dataset. The MetaCheXnet models improve over the default model for Lung TB classification, achieving an AUC score of 0.93.

3.3 Transfer Learning

Transfer Learning is prevalent in the field of deep learning applied to medical imaging. Most works use the same formula, which breaks down to using a network pre-trained in the ImageNet dataset, adjusted for the extraction of information from everyday objects, followed by the training on the target task. The solutions presented before for the automated classification of lung diseases in Chest X-rays use this formula step-by-step (Hwang et al., 2016; Cao et al., 2016; C. Liu et al., 2017; Lakhani & Sundaram, 2017; Rajpurkar et al., 2017).

To our knowledge, only two works explore the use of Transfer Learning in CNNs for Medical Images. The first one is (Gozes & Greenspan, 2019), inspired in the models trained for pneumonia in (Rajpurkar et al., 2017), deviates from the traditional formula by performing Transfer Learning in two steps. In the first step, Transfer Learning performs through training on the ChestX-ray8 dataset (Wang et al., 2017), using an ImageNet Baseline. In the second step, Transfer Learning performs through training on the Shenzhen TB dataset, using the ChestX-ray8 trained model in the previous step as a Baseline. The authors very briefly compare the results against a traditional Transfer Learning procedure to find a lower performance. The other work that explores Transfer Learning in Chest X-rays is (Raghu et al., 2019), released in 2019, it compares the training of Random and ImageNet trained networks in Chest X-ray Images and a smaller retinal dataset. The Chest X-ray dataset is the mentioned CheXpert dataset (Irvin et al., 2019), where the authors do not find any difference between the use of Transfer Learning or lack thereof. The authors go ahead and show that the training on small datasets such as the retinal fundus dataset leads to minimal modifications on the higher convolutional layers closer to the input (Raghu et al., 2019).

Chapter 4

Experimental Setup

This section describes the experimental setup used to create algorithms specialised in the visual classification of TB in Chest X-rays.

Section 4.1 describes the datasets used as part of this work, covering each dataset's technical details. Section 4.2 describes the general methodology used for this work. It approaches the technical information surrounding the training and testing of the models. And finally, section 4.3 describes the actual training of different CNNs. This section explores the details of the training framework, and the measurements used to gather the results.

4.1 Datasets

The Chest X-ray image datasets used in this work can be divided into two main categories: those labelled for multiple general lung diseases, and those labelled after the infection of pulmonary TB.

The X-ray datasets labelled for general lung diseases contain labels for manifestations of a given disease, identified by a field specialist in an X-ray image. These patterns help determine or rule out a possible underlying condition. They are much more numerous due to the broad spectrum of cases that these labels fit in, not being limited by the disease's cause. On the other hand, TB screenings make up the labelled X-ray sets, for which the specific subject limits the number of images available. The training procedures split these datasets into three different subsets, training, validation and testing, where:

- Training subsets provide the labelled data over each epoch. Each image contained in the training set is provided once during one epoch. Loss measures the difference between the CNN's output value when processing the training image and the truth label provided for the same image. The Optimizer uses the Loss value to perform the necessary adjustments to the network components to minimise loss.

- Validation is used at the end of each epoch of training. The CNN processes every image contained in the validation set and determines the Average Loss. Validation Loss determines

the early stopping of training. The adjustment of the network topology cannot use the Validation data.

- Testing is used after model training. All the results reported in this work originate from the Testing set. The training procedure does not use this subset. Overfitted CNNs show inferior results in the Testing set. Testing sets separate into two different subclasses:
  - Internal testing sets sourced from the same dataset as the training set. They share the same properties of the training set and therefore, often report optimal results.
  - External testing sets sourced from external datasets. They contain different properties that strain the CNN performance, often resulting in lower performance. However, an external testing set portrays a very realistic approximation to the actual CNN performance when deploying a model in the field.

What follows is a thorough description of the open-source datasets used for the training of General Lung Disease Classifiers, as well as the ones used in the training of TB Classifiers.

4.1.1 General Lung Disease Datasets

From the datasets mentioned in the state-of-the-art, this work uses the **ChestX-ray8** and the **CheXpert** dataset. This work does not use the PLCO, the OpenI or the MIMIC-CXR dataset. The PLCO sources oncology screening images and therefore, does not portray a problem relevant for TB classification. The OpenI dataset has reduced size, containing only 7,470 images. And finally, the MIMIC-CXR dataset does not comply with storage requirements. It contains over 300 thousand high-quality images, occupying 4.6 TB of internal storage, an amount that surpasses the storage assigned to this project.

The remainder of this section describes in detail the ChestX-ray8 and CheXpert datasets.

ChestX-ray8

The ChestX-ray8 dataset was released by the National Institute of Health, an agency of the U.S. Department of Health, in 2017 (Wang et al., 2017). Upon release, the dataset contained only eight classes assigned to each image, including Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, and Pneumothorax. Six other common lung diseases come after the original publication, namely Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening and Hernia.

The origin of the images in the dataset is not exact. In the paper, the authors determine that the images are sourced from their institute's PACS system, querying what were initially eight common thoracic diseases commonly observed by practitioners, and later extending the scope to 14 common lung diseases as part of an update to the original dataset. The final product is a dataset composed of 112,120 X-ray images, from 32,717 different patients.

Due to the significant number of images, manual labelling procedures would take a very long time to complete. NLP techniques provide the required labels, extracting them from the radiological reports attached to each X-ray. The mining procedures assign lung diseases with the help

of **DNorm** (Leaman et al., 2015), a machine learning method used for the detection of disease concepts, and **MetaMap** (Aronson & Lang, 2010), an ontology-based method used for the detection of bio concepts. The authors report improved results on a set of reports obtained from the OpenI API search engine (NIH, 2020), against the results obtained with the sole use of MetaMap. The result is an array of positive and negative disease labels assigned for each image, from which the CNNs learn. The co-occurrence statistics of different lung diseases are visible in fig. 4.1, and according to to (Wang et al., 2017), agree with the empirical knowledge of domain experts, such as the diagnosis of Infiltration in images diagnosed for Atelectasis and Effusion.

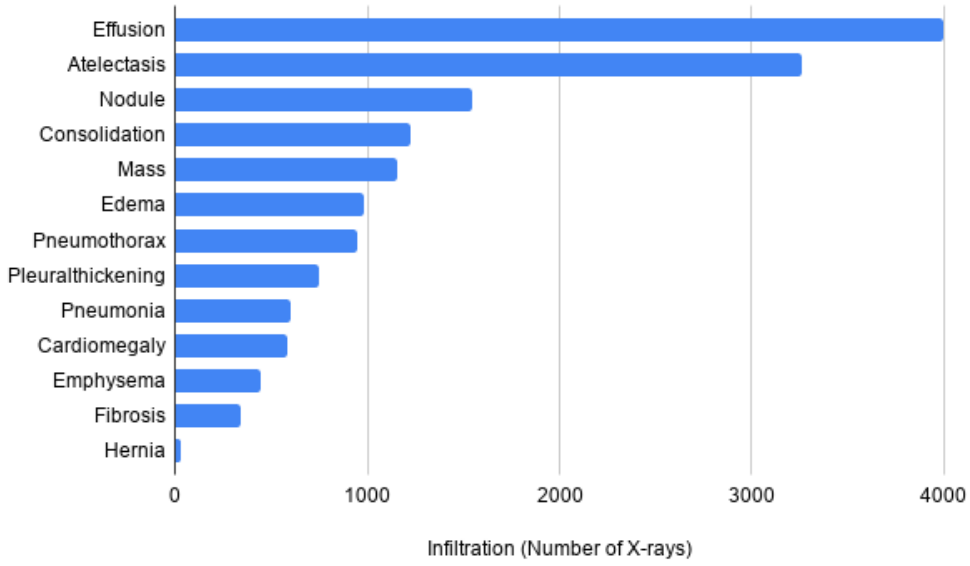

Figure 4.1: The bar chart represents the co-ocurrence of Infiltration with each of the other 13 labels in the ChestX-ray8 dataset.

The diagnosis of different lung diseases depends on external factors, such as the rarity of the finding and the sociological and environmental conditions. Figure 4.2 shows that in the ChestX-ray8 dataset, Infiltration has the most substantial representation in data, assigned to 17.71% of the images, with lower numbers for other diseases such as Pneumonia, assigned to 1.28% or even Hernia, assigned to 0.20% of the entire dataset.

CheXpert

The CheXpert dataset was released in 2019 by a Stanford University team (Irvin et al., 2019). This dataset provides two different sets for training/validation and testing, with the training dataset providing 224,316 Chest X-ray images from 65,240 patients. The testing contains 200 X-rays labelled by certified board specialists. The images and associated radiology reports originate from the Stanford Hospital PACS system, involving studies between October 2002 and July 2017.

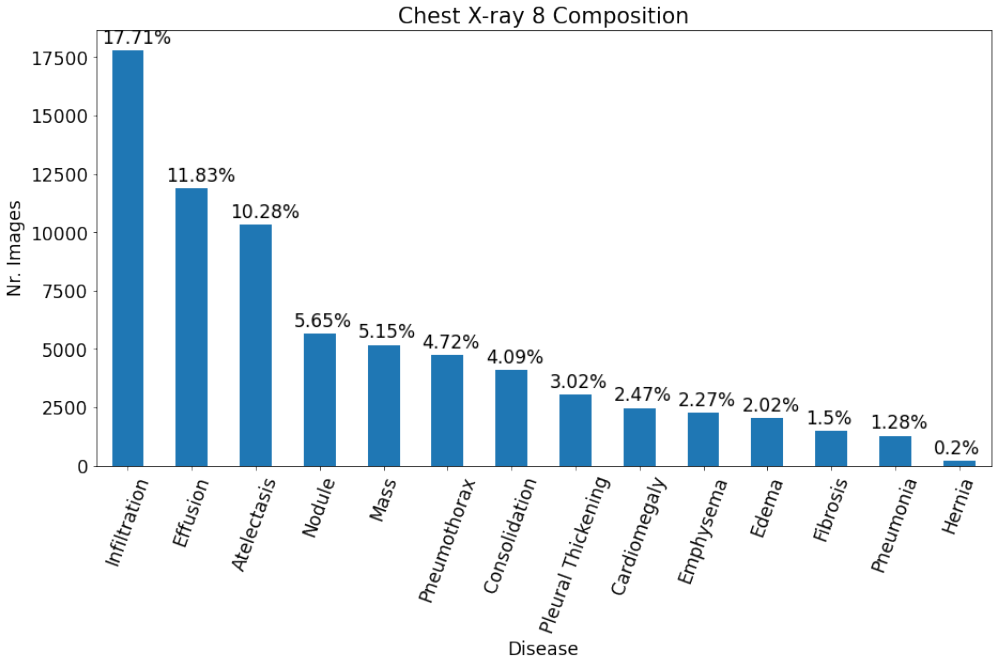

Figure 4.2: Total number of images for each disease. The value displayed on top of the bars is the relative size to the complete ChestX-ray8 dataset.

A total of 14 chest diseases, mined from the textual reports, classify each of the images, including No Finding for the absence of chest diseases visible/reported in the text, Enlarged Cardiomegaly, Cardiomegaly, Lung Lesion, Lung Opacity, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, Support Devices. These labels result from text mining methods, except for the validation and test set, which are labelled by three specialists to determine the ground truth. The team reinforces the use of rule-based extraction, resulting in an improved performance when compared with the methods used in the Chest-X-ray8 dataset. The used procedure captures the Mention, where the text corpus states the presence of the disease, Negation, where the text corpus discards the presence of a disease, and Uncertainty, where the mention of the condition is detected but the method cannot determine if it approves or discards its presence (see (Irvin et al., 2019)) for more information). A set of rules then aggregates these three tasks. The improvement is visible in fig. 4.3, with improved F1 score over Mention, Negation, and Uncertainty syntax captured in the labelling processes. The CheXpert and ChestX-ray8 labellers only extract a total of 7 shared classes, complicating testing procedures between datasets.

Similar to what happens in the ChestX-ray8 dataset, the product of the mining procedures are images labelled for multiple lung diseases. This dataset does not escape from the same issue related to the biased distribution of positive cases. This problematic distribution is visible in fig. 4.4, with labels such as Pleural Effusion, Support Devices and Lung Opacity being over-represented compared with classes such as Pneumonia or Lung Lesion.

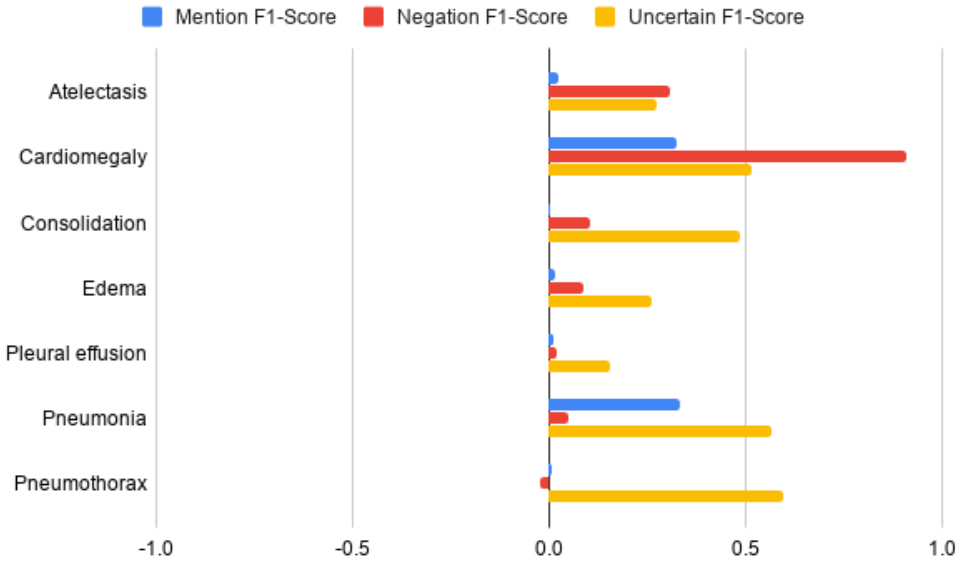

Figure 4.3: The F1 Score difference for each disease in the CheXpert dataset (CheXpert F1 - Chest X-ray). The labellers perform using a set of rules that aggregate the Mention, Negation or Uncertainty of the given label in the text corpus. The CheXpert labeller performs better in every process, for diseases shared between the two datasets.

4.1.2 Tuberculosis Datasets

This work uses the Shenzhen Chest X-ray set to train, validate and test the TB models, and holds the Montgomery-County set only for testing, following the steps of other works such as (Lakhani & Sundaram, 2017) and (Gozes & Greenspan, 2019). The remainder of this section provides the current information for these TB datasets.

Shenzhen Hospital X-ray Set

The Shenzhen Hospital X-ray Set (Jaeger et al., 2014) provides 662 images, with 326 pictures labelled positive for TB. These are hand labelled pictures for TB and have been previously used effectively in the training of Deep CNNs (Gozes & Greenspan, 2019; Lakhani & Sundaram, 2017; Islam et al., 2017). The Shenzhen dataset was collected in Guandong Medical College, Shenzhen China, from routine operations during September 2012. A Philips DR Digital Diagnostic system captures the images, provided in PNG format with an approximate size of 3000x3000 pixels. The authors of the dataset also provide a TXT file with the case report, with info on age, gender, and TB type.

Montgomery-County Chest X-ray Set

The Montgomery-County Chest X-ray Set labelled for TB provides 138 Chest X-rays, with 58 images labelled positive for TB. The images originate from the Montgomery County's Tuberculosis

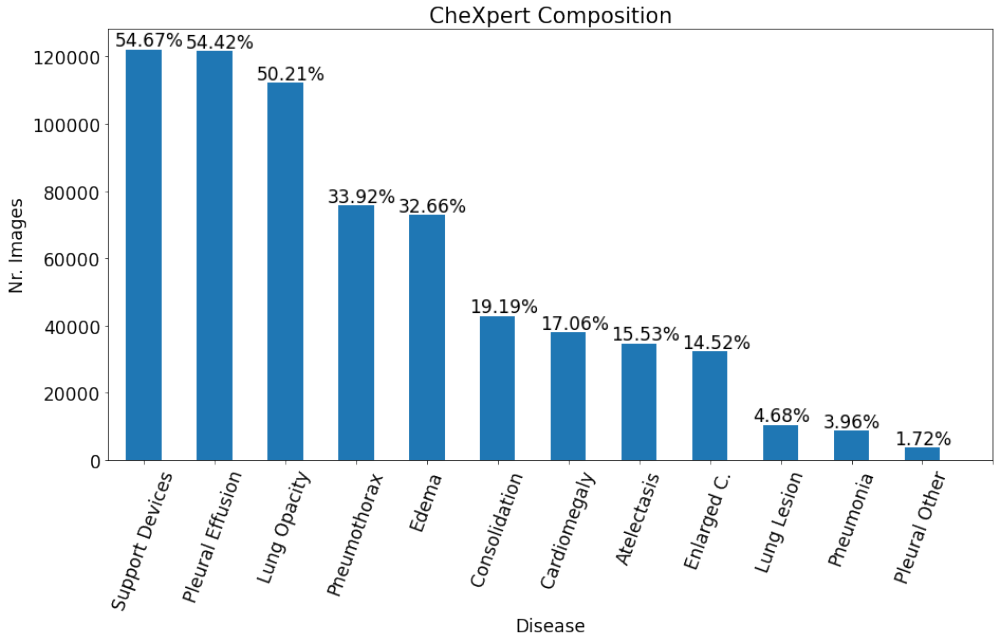

Figure 4.4: Total number of images for each disease. The value displayed on top of the bars is the relative size to the complete CheXpert dataset.

screening program, captured with a Eureka stationary X-ray machine (CR), and provided in PNG format with an image size 4020x4892 pixels. Similar to what happens in the Shenzhen Chest X-ray set, a TXT file linked to the X-ray images provide the clinical reading, with information regarding the patient's age, gender and abnormality seen in the lung (if any abnormal finding is present).

4.2 Methodology

This work explores the effects of Transfer Learning and Data in TB classifiers using small datasets. The Baselines used in this work to perform Transfer Learning are CNNs adjusted for a different problem, in a separate dataset, loaded at the start of each training procedure. This work uses three main types of Baselines:

- **Random Baselines**, networks with randomly generated weights. These networks are "blank", in other words, they do not contain any learned features and require large amounts of images for the development of robust features;

- **ImageNet Baselines**, networks trained in the ImageNet dataset containing 14 million RGB images labelled for general everyday subjects. These networks can segment the input into robust features, such as edges, patterns or shapes, useful for interpreting visual data.

- **Chest X-ray Baselines**, networks trained in either one of the Chest X-ray datasets used in this work. They result from the training procedures for the classification of Lung Diseases in

X-rays, training from either a Random or ImageNet Baseline. The TB Classifiers use these Baselines for Transfer Learning.

Each Baseline trains a set of CNNs for each series. Each series contains a set of CNNs trained with the same Baseline and training set, named after the task, dataset, and portion of data used during training [1]. The figure in fig. 4.5 displays the experiments performed in this work. The figure depicts the whole experimental design, providing the number of CNNs prepared for each series, its name, and the relationship between the series. The following subsections describe the methodology used for the training of Lung Disease and TB classifiers.

4.2.1 Lung Disease

For the ChestX-ray8 dataset, this work follows previous research steps that train CNNs on the same dataset (Yao et al., 2017; Wang et al., 2017; Rajpurkar et al., 2017; Gozes & Greenspan, 2019), using 20% of the original dataset for testing. Neither of the previous works determines precisely the portion used for testing, randomly sampling the dataset to fulfil each subset's desired length. This work attempts to discard any undesirable sampling effects by splitting the ChestX-ray8 dataset into 20% folds. Therefore, single replicates use one fold for testing and the other four folds for training and validation. With this procedure, the replicates train on all the data available in the original dataset.

CheXpert provides an extensive amount of data processed by a better performing labeller than the ChestX-ray8 dataset (Irvin et al., 2019). Both factors should contribute positively to model training. This work divides the CheXpert dataset into three 33% training subsets, training one replicate for each, producing 3 CNN replicates. This step approximates the CheXpert and ChestX-ray8 training subset size, discarding any performance differences related to dataset size. An additional model trains on 100% of the CheXpert dataset, capturing any improvements associated with the training set's size. Model training uses the hand labelled validation set provided by CheXpert, following the steps of previous research (Irvin et al., 2019). Testing uses one of the 33% subsets not used during training. Testing for the CheXBig-I CNN does not carry out, since this model uses 100% of the original dataset for training, no data exists for testing.

Figure 4.5 visually portrays this first section under the horizontal bar named "General Chest Abnormality Classifiers". Table 4.1 provides the absolute amount of images provided in each subset for the training, validation and testing of ChestX-ray8, and CheXpert trained models. To summarise the training process, a total of 17 CNNs train for the classification of Lung Disease where:

- Five use a Random Baseline and another five use an ImageNet Baseline, both obtained from a 64% training subset of the ChestX-ray8 dataset, producing the **Chest_XSmall-R** and **Chest_XSmall-I** models;

[1] For example, the series "Chest_XSmall-R" trains on the **ChestX**-ray8 dataset, with a **Random** Baseline and about **72 thousand** images.

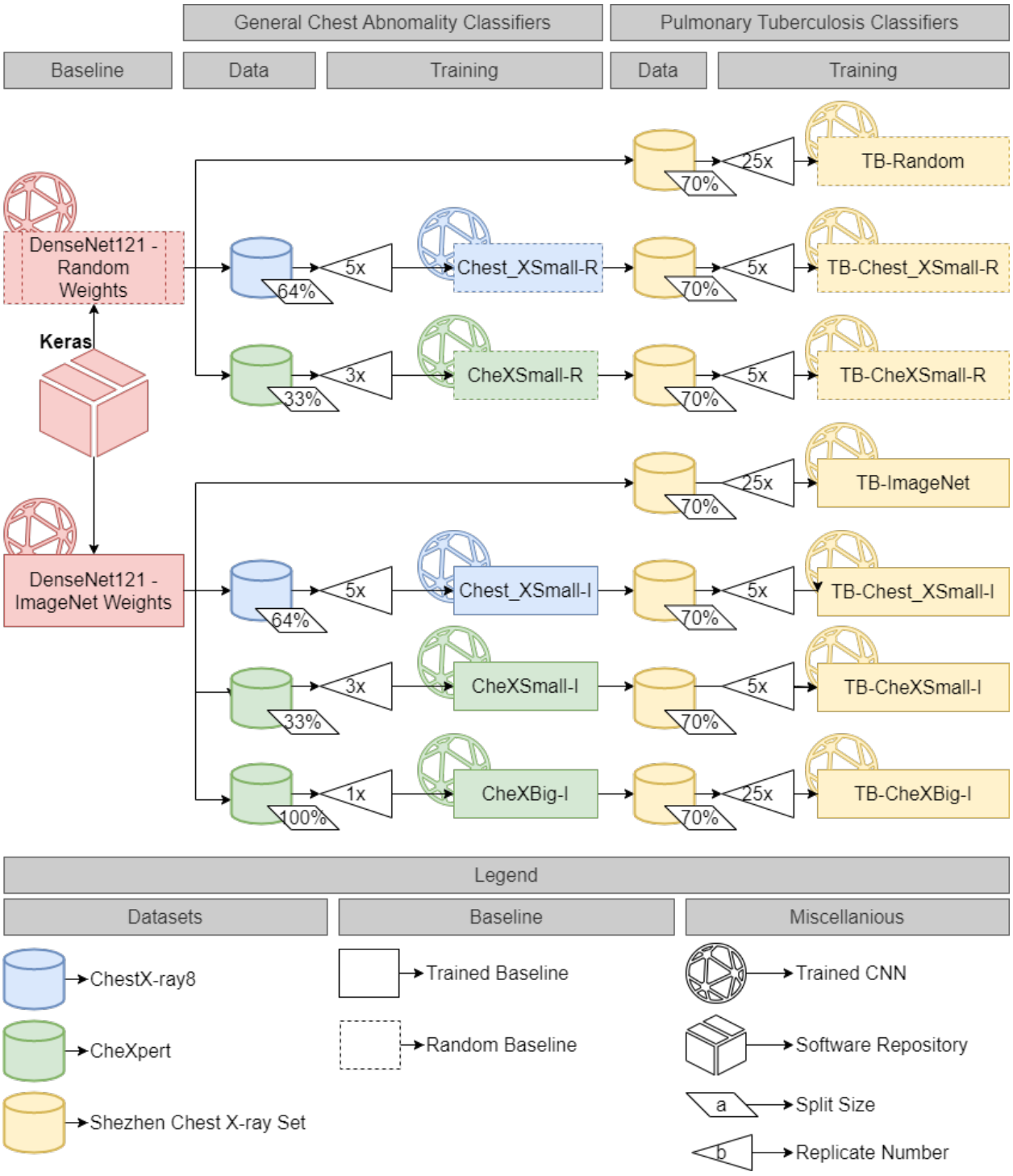

Figure 4.5: Diagram representative of the experimental procedures depicted in the folowing sections. The upstream CNNs represents the Baseline used for the CNN connected by an arrow ($Baseline \longrightarrow TrainingCNN$).

- Three use a Random Baseline and another three use an Imagenet Baseline, both obtained from a 33% training subset of the CheXpert dataset, producing the **CheXSmall-R** models and **CheXSmall-I**;
- One uses an ImageNet Baseline and 100% training set, producing the **CheXBig-I** model.

Table 4.1: The different data partitions prepared for CheXpert and ChestX-ray8.

Dataset	Training	Validation	Testing
ChestX-ray8	71756 ($\sim$ 64%)	17939 (= 16%)	22424 (= 20%)
CheXpert (33%)	74471 ($\sim$ 33%)	234 ($\sim$ 3%)	74471 ($\sim$ 33%)
CheXpert (100%)	223414 ($\sim$ 99%)	234 ($\sim$ 1%)	-

4.2.2 Tuberculosis

For TB datasets, this work follows the data layout used in (Gozes & Greenspan, 2019), visible in table 4.2. Since it produces well-performing models with $>$ 90% AUROC, with internal testing set useful for comparison with the results in external testing sets, such as the Montgomery County Chest X-ray set used in this work. Validation and testing both retrieve 100 images from the Shenzhen dataset, with an equal number of images labelled positive and negative for TB. The training subset contains the rest of the images not used in the validation or testing set, for a total of 462 images, 236 labelled positive for TB. This work reserves the Montgomery dataset as an external testing set.

Table 4.2: The different splits assigned to the Shenzhen and Montgomery dataset.

	Training	Validation	Testing
Shenzhen	456 ($\sim$ 69%)	100 (= 15%)	100 (= 15%)
Montgomery	-	-	138 (= 100%)

Training for all the TB models uses the same 70% subset of the Shenzhen TB X-ray Set. Figure 4.5 displays the training processes involving Tuberculosis datasets under the horizontal bar "Pulmonary Tuberculosis Classifiers". The bars named "Random Baseline" and "ImageNet Baseline" refer to the procedures using an initial Random Baseline and an initial ImageNet Baseline, respectively.

The experiments use seven different Baselines, producing a total of 155 other TB models, where:

- 65 CNNs train with an initial Random Baseline:
  - 25 replicates use a Random Baseline, off-the-shelf, provided by Keras, producing the **TB-Random** series;

  - Five replicates train from each of the five Chest_XSmall-R models, generating 25 models obtained from a 64% subset of the ChestX-ray8 dataset with a Random Baseline, producing the **TB-Chest-X67-R** series;
  - Five replicates train from each of the three CheXSmall-R models, generating 15 models obtained from a 33% subset of the CheXpert dataset with a Random Baseline, producing the **TB-CheXSmall-R** series;

- 90 CNNs train with an initial ImageNet Baseline:
  - 25 replicates use an ImageNet Baseline, off-the-shelf, provided by Keras, producing the **TB-ImageNet** series;
  - Five replicates train from each of the five Chest_XSmall-I models, generating 25 models obtained from a 64% subset of the ChestX-ray8 dataset with an ImageNet Baseline, producing the **TB-Chest-XSmall-I** series;
  - Five replicates train from each of the three CheXSmall-I models, generating 15 models obtained from a 33% subset of the CheXpert dataset with an ImageNet Baseline, producing the **TB-CheXSmall-I** series;
  - 25 replicates use the CheX_I model, obtained from a 100% subset of the CheXpert dataset with an ImageNet Baseline, producing the **TB-CheXBig-I** series.

4.3 CNN Architecture and Training

This section summarises the architecture and general configuration used for the training of the CNN models. Section 4.3.1 determines the details surrounding the used architecture for all the training procedures of this work, and section 2.2.2 explains the used training procedures.

4.3.1 Architecture

The base CNN model is a 121-layer network, the **DenseNet121** (Huang et al., 2017) downloaded from Keras (*Keras Documentation for Aplications*, 2020). Previous works such as (Yao et al., 2017; Rajpurkar et al., 2017; Irvin et al., 2019; Gozes & Greenspan, 2019), using DenseNet shows that it outperforms other established networks in biomedical applications such as ResNet (He et al., 2016), GoogLeNet (Szegedy et al., 2015) and AlexNet (Krizhevsky et al., 2012). Deep Residual Networks, also known as ResNet (He et al., 2016), and Highway Networks (Srivastava et al., 2015) are the main inspiration behind the design of the DenseNet121 architecture. Simple feed-forward CNNs communicate through the output of adjacent convolutional blocks. DenseNets connects blocks at different depths with shortcut connections, meaning that an nth block will receive the output of the nth - 1 block and the output of all the previous blocks.

The images are resized to the input layer's size before use, with the outermost densely connected layers requiring replacement for the new task. The Keras Functional API (*Keras Documentation for Functional API*, 2020) captures the relu layer at the end of the DenseNet121 network. It connects

a GlobalAveragePooling2D Layer, followed by the output layer, a Densely Connected Layer with Sigmoid activation, the best-suited activation function, as explained in section 4.3.2. The Densely Connected Layer's length is equal to the number of classes available in the training dataset.

4.3.2 Loss Function

The labelling method results in an independent classification problem that needs to be taken into account when determining the offset between the truth labels in the dataset and the predicted labels provided by the output after processing. This offset is measured using the appropriate loss function, which for this type of classification is Binary Cross-Entropy Loss. Binary Cross-Entropy loss computes the loss for every CNN output class independently from each other, using the Sigmoid Activation Function to modulate a vector of values in a range between 0 and 1 with the following function:

$$f(s_i) = \frac{1}{1 + e^{-s_i}}, \tag{4.1}$$

where s_i represents the raw value of the CNN output. Cross-Entropy is then measured singularly for each element of the vector to determine the loss value that defines the strength of back-propagation for weight adjustment, depicted by the following function:

$$H(p_i, q_i) = -\sum_i p_i \log q_i, \tag{4.2}$$

where p_i is the observed value (either 0 or 1) and q_i the outcome value of the network for the condition.

TB labelled datasets represent a similar problem to the ones found in ChestX-ray8 and CheXpert, with a single binary problem. The truth values for each image are either positive or negative for TB. Therefore, the training procedures for TB classification also use Binary Cross-Entropy.

4.3.3 Training

The chosen batch size for training depends on the hardware that the models run on; however, it also affects the final models' performance. Recent studies regarding the effect of batch size in the fine-tuning of CNNs in medical image classification determine 16 as an acceptable batch size for similar training procedures (Kandel & Castelli, 2020). The training of the Lung Disease and TB classifiers proceeds with an equal batch size of 16.

During training, before the assignment to batches, the resolution of the images is scaled down to 224x224 with OpenCV, matching the input layer dimensions of the default DenseNet121, as supplied by Keras. Random image manipulation leverages the CNN models' generalisation capacity, introducing random artefacts and preventing the algorithms from memorising the original input. OpenCV performs horizontal axis flipping and rotation with a random value between -360° and

360º, displacing the input channel's pixel values. All the stated manipulations have a 50% chance of taking place. Image augmentation does not apply to the validation or testing set images.

Training proceeds in a slightly different manner depending on the nature of the classifier.

- **Lung Disease Classifiers** proceed with training until the validation loss does not improve after five epochs. When it stops, the program stores the model with the lowest recorded performance for testing. The update rule used to tune the network weights is Adadelta with default settings as provided by Keras.
- **TB Classifiers** train until the validation loss does not improve after ten epochs. The increased number of epochs accommodates the much smaller size of the training set used to train the TB classifiers, providing more room for improvement. However, not so much that would degrade generalisation capacity. The update rule used is Nadam with default parameters, the same update rule successfully used to train TB Classifiers (Gozes & Greenspan, 2019).

4.4 Software and Hardware

The methodology described before is all achieved using Python 3.6 (*Python 3.6*, 2020) for scripting, and Docker 4.3.1 (*Docker Download*, 2020) for version control downloaded from docker. The models' training performs with Keras 2.3.1 (*keras*, 2020) and Tensorflow 2.3.1 (*Tensorflow Download*, 2020) for Python, downloaded from PIP. The image augmentation processes are all performed using OpenCV (*OpenCV*, 2020) for Python, version 4.4.0.44, a distribution of CV2. Scikit-learn 0.23.2 (*Scikit-learn*, 2020) provides the metrics used to evaluate the models' performance, WAF, AP and AUROC Score.

All the models are trained on the GPU from a remote server, with a Geforce GTX 1080 ti.

Chapter 5

Results

The measurement of model performance uses three different metrics, Weighted Average F-Measure (WAF), Area Under the Receiving Operating Characteristic (AUROC), and Average Precision (AP).

WAF is the Average Weighted F-Score of the Positive and Negative classes, where the relative proportion of each class[1] averages its F-score(It requires the assignment of a threshold for the binarisation of the CNN output[2]. WAF is determined as follows:

$$WAF = \sum_{n} F(C_n) \times P_n, n = \{0, 1\} \tag{5.1}$$

For the WAF threshold, a simple script searches the cutoff value that maximises the WAF score in the training data[3]. Testing proceeds, using the optimal training threshold to calculate the WAF score.

AUROC, used in (Wang et al., 2017), (Yao et al., 2017) and (Rajpurkar et al., 2017), determines performance using a floating threshold point. This metric effectively portrays model performance without the need for a fixed optimal threshold, such as in the case for F-score or WAF. However, AUROC is too optimistic for unbalanced data. With a minimal representation of some classes in the training datasets under highly skewed data, the AUROC metric provides unreliable results (Fernández et al., 2018).

AP summarizes the weighted mean of Precisions at multiple thresholds, using the Recall of the previous threshold point as the weight. According to the documentation of Scikit-Learn, this implementation prevents the overly optimistic results resulting from the simple measurement of the Area Under the Precision-Recall Curve (Pedregosa et al., 2011). The Precision-Recall Plot provides a more informative view of model performance in highly skewed data (Davis & Goadrich, 2006), and AP delivers a single value for easier evaluation. For AP, the weighted mean of Precision

[1]The diagnosis of each disease is binary, where False = 0 and True = 1.

[2]The output is a float between 0 and 1, binarisation transforms the float into a boolean regarding whether the float is above the threshold or not.

[3]Each threshold is model and disease specific

(P) measures using the difference of Recall (R) between each immediate step in threshold (n), according to the following equation:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n, n = \{0, 1\} \tag{5.2}$$

The comparison of the different models uses the **Kruskal-Wallis H test** (Kruskal & Wallis, 1952). This test determines significant statistical differences between two or more groups of an independent and continuous variable. For general lung diseases, the statistical analysis compares models trained from Random and ImageNet Baselines, for ChestX-ray8 and CheXpert. This study does not compare models trained in different datasets because their labels do not fully match. For TB, the statistical analysis compares the seven different series between each other. The implementation of the test uses the Scipy method for Python, downloaded from (*Scipy Kruskall-Willis H test*, 2020). When provided with the distributions of values for each group, this method returns the ***p*-value**, and **Kruskal-Wallis H Statistic**. The statistical test evaluates each group against each other, in one-to-one comparisons. It returns the p-values in a matrix and determines statistical significance for groups showing a p-value lower than 0.01. Given two groups, x and y, a Kruskal-Wallis H test performed on x and y (in this exact order), with p-value < 0.01, has its font changed to green if the median value of x has a higher value than y. The font changes to red, if the median p-value of x has a lower value than y.

The following section describes the results gathered in the pursuit of optimal TB classifiers using transfer learning and a limited set of images. Section 4.1 covers CNNs trained for the visual classification of general lung diseases using open-access Chest X-ray datasets. As portrayed in fig. 4.5, Chest X-ray Baselines' generation uses Random and ImageNet Baselines as the starting point. The ChestX-ray8 and CheXpert X-ray datasets provide a plentiful amount of images to tune the original Baseline for a Chest X-ray specific task. AUROC and AP measure the performance for the classification of each disease in internal testing sets. The models trained for general lung disease classification provides the X-ray trained Baselines used in 5.2 for Transfer Learning. This latter section covers the training of CNNs for the classification of TB. It compares models trained from Random, ImageNet, and Chest X-ray Baselines. Chest X-ray Baselines leverage model performance on small datasets by training on a similar problem, such as lung disease classification in Chest X-rays. Here AUROC and WAF measure the performance for the classification of TB in internal and external testing sets.

5.1 Lung Disease Classifiers

Lung abnormality datasets are quite large when compared with open-access datasets labelled for TB. They are the result of the combined effort between the medical institutions that hold X-rays of multiple case studies, human X-ray specialists, and the teams that gather and explore the data with the help of computer-automated mining procedures. Such is the case of the ChestX-ray8

dataset provided by (Wang et al., 2017), and Chexpert provided by (Irvin et al., 2019). These large datasets allow the development of X-ray specific Baselines.

The following section depicts the outcome of model training on Chest Disease labelled Chest X-rays, with Random and ImageNet Baselines. This step empirically determines the impact of Baseline on large datasets and prepares the Chest X-ray Baselines sought after in section 5.2 for TB training procedures. section 5.1.1 reports the performance of models trained on a ChestX-ray8 dataset and section 5.1.2 reports the performance of the models trained on the CheXpert dataset.

5.1.1 ChestX-ray8

The boxplots in fig. 5.1 shows the AP results for the classification of each disease. The number of images with positive labels for each class sorts the X-axis in descending order [4] (see section 4.1). The inspection of fig. 5.1, portraying AP values, reveals a clear pattern showing higher model performance for diseases assigned to a higher number of images. It is the case for Effusion, assigned to 11.43% of ChestX-ray8 dataset and scoring the best median AP value of 0.42 on Random and ImageNet Baselines. Diseases with a lower number of positive labels achieve lower scores, such as Hernia, present in 0.2% of the dataset images, with an AP value of 0.01 for Random Baselines, and 0.05 for ImageNet Baselines.

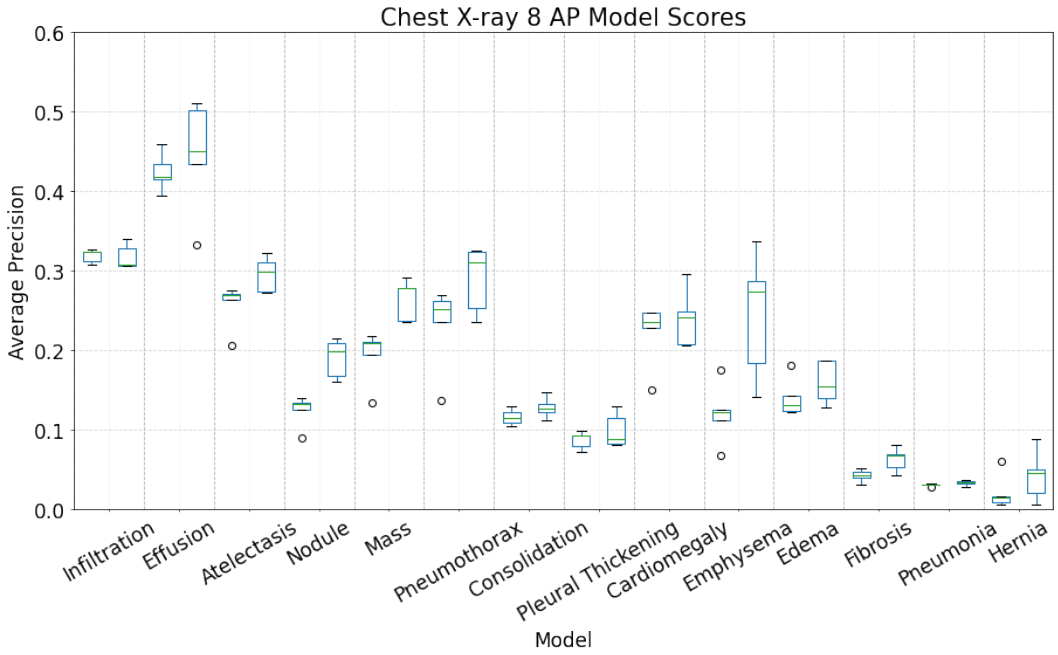

Figure 5.1: Distribution of AP values from the Chest_XSmall-R (Random Baselines, left) and Chest_XSmall-I (ImageNet Baselines, right) in the Chest X-ray testing set.

The comparison of AP values for each Disease and Baseline shows a slight improvement with the use of ImageNet Baselines. The median of ImageNet models surpasses Random Baselines'

[4] more images to the left - fewer images to the right

results for every disease, although the difference varies for each disease. In total, the ImageNet Baselines achieve a median AP value of 0.21 [5], and the Random Baselines achieve a median AP value close to 0.13.

The Kruskal-Wallis H test, shown in 5.1, compares the ImageNet and Random Baselines distribution for each of the 14 classes present in the ChestX-ray8 dataset. The tests determine that only two classes show statistically significant differences for AP metrics, with ImageNet Baselines outperforming Random Baselines for Mass and Nodule classification. [6]. Other results do not offer any significant differences, including the total distribution of each Baseline's outcomes, regardless of disease.

On the other hand, the AUROC metrics, depicted in fig. 5.2 do not share the clear pattern suggested by AP metrics. There is no clear relationship between model performance and the number of images assigned to each disease.

As described at the beginning of this chapter, the AUROC metric is highly inappropriate for poorly weighted datasets, providing very optimistic results for classes with low representation. Diseases like Hernia support this hypothesis, performing very poorly on AP metrics while showing very high median values for AUROC metrics. The results show Hernia classification performing with a median AUROC score of 0.76 for Random Baselines, and 0.84 for ImageNet Baselines.

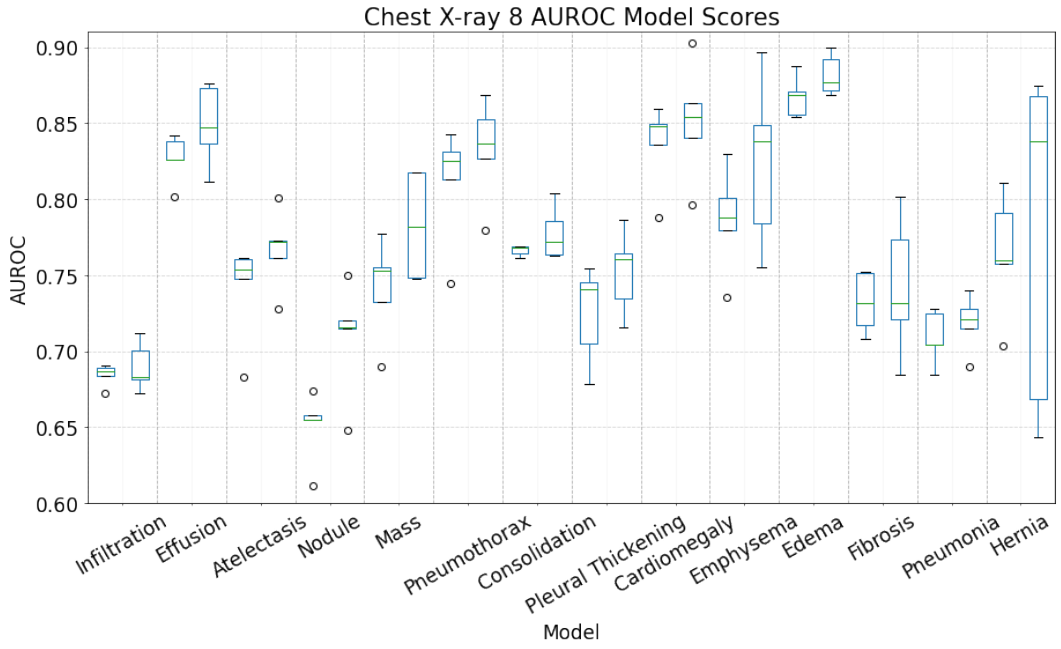

Figure 5.2: Distribution of AUROC values from the Chest_XSmall-R (Random Baselines, left) and Chest_XSmall-I (ImageNet Baselines, right) in the Chest X-ray testing set.

[5] concerning the scores obtained for every disease

[6] E.g. Mass performs better in ImageNet Baselines, according to AP measurements, because $p - value(Mass, AP) < 0.01$ and $ImageNet(Mass, AP)Median > Random(Mass, AP)$.

The statistical tests in 5.1 do not determine any significant differences between the Random and ImageNet Baselines contradicting the results found for AP metrics which determine significant improvements for Mass and Nodule in ImageNet Baselines.

All the statistical test p-values comparing the Random and ImageNet distributions for each disease are available in 5.1 for further review. The diseases sort according to the number of images assigned to them, with diseases covering a larger number of images in the dataset placed at the top. The tests do not show any relationship between the p-value and number of images assigned to each disease.

Table 5.1: P-values from the Kruskal-Wallis H test for models trained in the ChestX-ray8 dataset, comparing Random and ImageNet Baseline models.

	AP	AUC
Infiltration	0.754	0.917
Effusion	0.251	0.175
Atelectasis	0.028	0.117
Nodule	0.009	0.076
Mass	0.009	0.175
Pneumothorax	0.175	0.251
Consolidation	0.175	0.347
Pleural Thickening	0.465	0.175
Cardiomegaly	0.602	0.347
Emphysema	0.016	0.251
Edema	0.175	0.117
Fibrosis	0.047	0.602
Pneumonia	0.117	0.465
Hernia	0.347	0.602

5.1.2 CheXpert

The boxplots in 5.3 show the AP results for the classification of each disease. Similar to the boxplots observed in the previous section, the X-axis sorts in descending order according to the number of images with positive labels for each class. The inspection of 5.3 portraying AP values, reveals the same pattern noticed for models trained in the Chest-Xray8 dataset, with models performing better for diseases assigned to a higher number of images. For example, Pleural Effusion, covering about 54.42% of the images in the CheXpert dataset shows a median AP value of 0.75 for Random Baselines and 0.74 for ImageNet Baselines. On the other hand, Pneumonia, covering 3.96% of the images achieves an AP value of 0.08 for Random and ImageNet Baselines.

The difference between models trained with Random or ImageNet Baselines is almost negligible. The Random and ImageNet Baselines achieve a median AP value of 0.38.

The Kruskal-Wallis H test available in table 5.2, further supports the previous results, not returning p-values lower than 0.01 for any disease or metric used. Therefore the performance

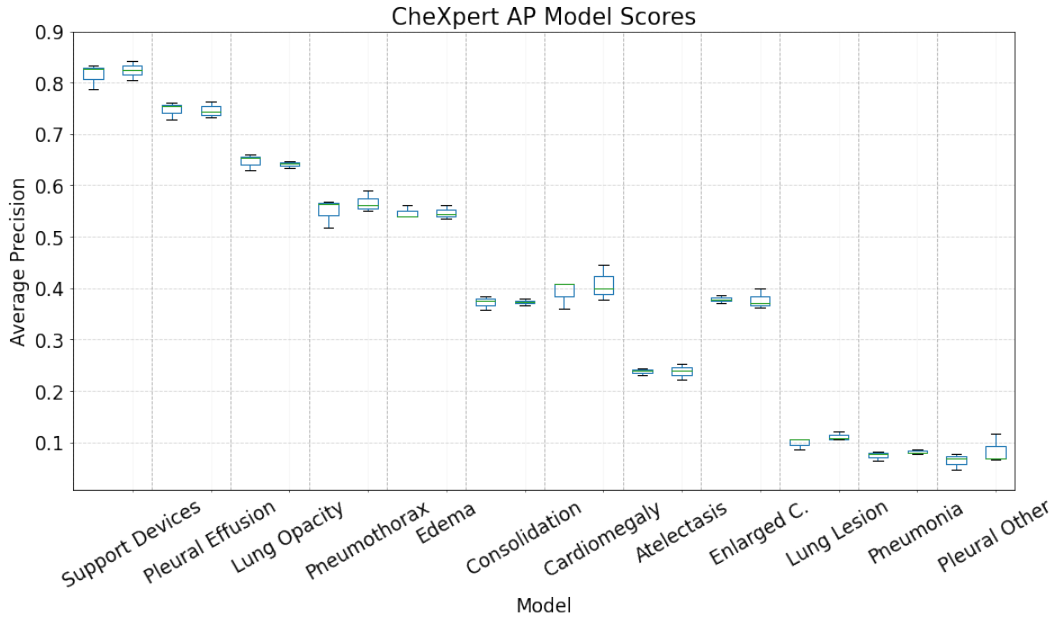

Figure 5.3: Distribution of AP values from the CheXSmall-R (Random Baselines, left) and CheXSmall-I (ImageNet Baselines, right) in the CheXpert testing set.

of ImageNet Baselines, show no statistically significant difference to the distributions of Random Baselines. These results are somewhat similar to the results found in the AP statistical tests between Random and ImageNet Baselines for ChestX-ray8 trained models which determined significant differences for only two diseases, not shared by the CheXpert dataset.

Similar to the ChestX-ray8 trained models findings, the AUROC metric produces suspicious results, as seen in fig. 5.4. They assign similar performance to all the diseases except for Support Devices. It achieves an AUROC value of 0.81 for both Random and ImageNet Baselines, the highest value registered out of all the diseases. Diseases such as Pneumonia, covered in 3.96% of the CheXpert dataset, achieve an AUROC value of 0.69 for Random Baselines and 0.70 for ImageNet Baselines.

The differences between Random and ImageNet Baselines for AUROC comply with the findings for AUROC in the ChestX-ray8 trained models, with both baselines showing very similar median AUROC values of 0.71 and 0.72 for Random and ImageNet Baselines respectively. The statistical tests for AUROC, available in table 5.2, comply with the previous finding and the AP results, showing no improvements between Random and ImageNet Baselines.

All the statistical test p-values comparing the Random and ImageNet distributions for each disease are available in 5.2 for further review. Similar to table 5.1, the diseases sort according to the number of images assigned to them, with diseases covering a larger number of images in the dataset placed at the top. Similar to the ChestX-ray8 findings, the tests do not show any relationship between the p-value and number of images assigned to each disease.

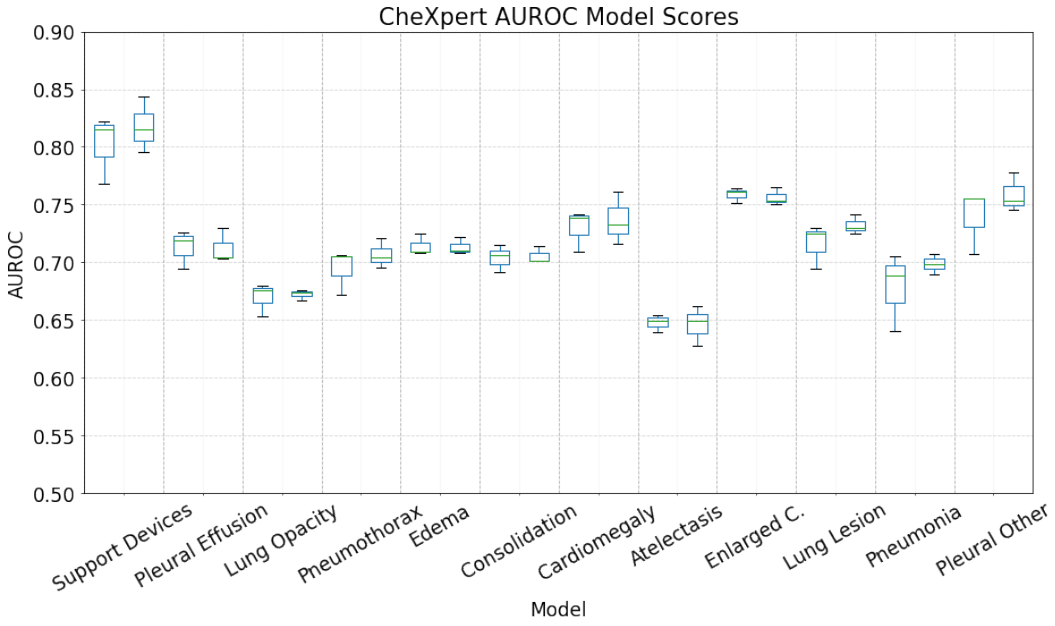

Figure 5.4: Distribution of AUROC values from the CheXSmall-R (Random Baselines, left) and CheXSmall-I (ImageNet Baselines, right) in the CheXpert testing set.

Table 5.2: *P*-values from the Kruskal-Wallis H test for models trained in the CheXpert dataset comparing Random and ImageNet Baselines.

	AP	AUC
Support Devices	0.827	0.827
Pleural Effusion	0.827	0.827
Lung Opacity	0.513	0.513
Pneumothorax	0.827	0.827
Edema	0.827	0.827
Consolidation	0.827	0.827
Cardiomegaly	0.827	0.827
Atelectasis	0.827	0.827
Enlarged Cardiomediastinum	0.513	0.827
Lung Lesion	0.127	0.127
Pneumonia	0.275	0.275
Pleural Other	0.513	0.827

5.2 TB classifiers

The ImageNet Baselines, trained with 14 million images, are capable of robust feature extraction, useful in any problem related to retrieving visual information, such as the detection of shapes and forms linked with pulmonary disease. However, this Baseline does not have its features tuned for the classification of X-rays, possibly struggling to capture optimal visual information in Chest X-rays classification. On the other hand, Chest X-ray Baselines represent a similar problem to the classification of TB. With much larger datasets, the training for identifying general diseases may seed the network with rich features useful for the extraction of visual information from X-ray images. The following results explore different Baselines and how Transfer Learning carries over in the training of TB classifiers.

AUROC and WAF (Weighted Average F-measure) provide the necessary metrics for interpreting model performance. For reference, fig. 5.5 provides the WAF scores and fig. 5.6 provides AUROC measurements for each set of TB models. table 5.3 provides the Kruskal-Wallis H-test results for Shenzhen test set (the internal test set) and table 5.4 provides the Kruskal-Wallis H-test results for Montgomery test set (the external test set). The results separate into two different subsections. The first, section 5.2.1, reports the results for models trained with a single Random Baseline (TB-Random), and Chest X-ray Baselines trained from a Random Baseline (TB-Chest-XSmall-R and TB-CheXSmall-R). The other section, section 5.2.2, reports the results obtained from models trained with a single Imagenet Baseline (TB-ImageNet) and Chest X-ray Baselines trained from an ImageNet Baseline (TB-Chest-XSmall-I, TB-CheXSmall-I and TB-CheXBig-I).

5.2.1 Random Baseline TB models

These CNNs show a WAF score on the Shenzhen test for the three series of models, with a lower median score of 0.69 for TB-Random, and a higher median score of 0.76 TB-Chest-XSmall-R. AUROC measurements determine slightly lower median AUROC values for TB-Random of 0.74. Similarly to WAF, the TB-Chest-XSmall-R also achieves higher values, with an AUROC value of 0.82.

Results from the Montgomery test set show a lower median WAF score of 0.43 for TB-CheXSmall-R, with TB-Random outperforming the other Random Baseline series with a WAF score of 0.49. Contrary to the WAF results, the AUROC results determine TB-Random as the worst performer out of the three model series, achieving an AUROC value of 0.54. TB-Chest-XSmall-R on AUROC shows better median values out of the three, with a median value of 0.57.

The statistical analysis for measurements in the **Shenzhen** test set, in table 5.3, determines significant differences between the TB-Random series and TB-Chest-XSmall-R, with TB-Chest-XSmall-R showing better results for both WAF and AUROC measurements. No other significant differences exist between the series trained with Random Baselines. The statistical analysis of the distributions in the Montgomery Dataset determines better AUROC performance for TB-CheXSmall-R over TB-Random. The three model series perform consistently worse than the TB-CheXBig-I models, for both WAF and AUROC metrics.

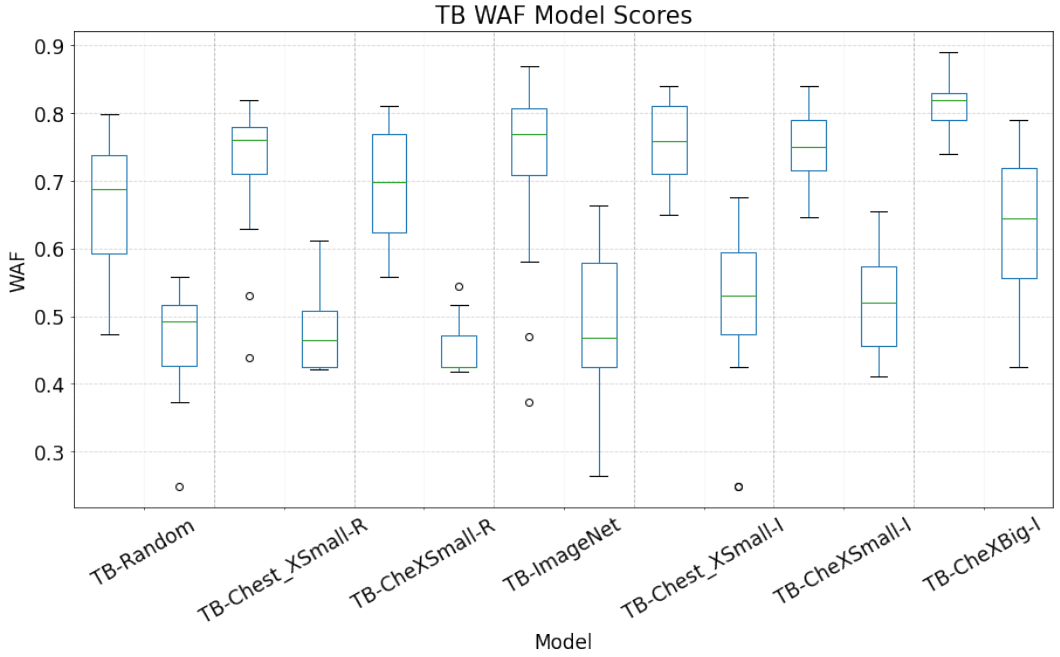

Figure 5.5: WAF-Test score measurements with train threshold for the classification of TB. For each model denoted in x axis, the left box represents the test results on Shenzhen, and the right box determines the test results on the Montgomery County Dataset.

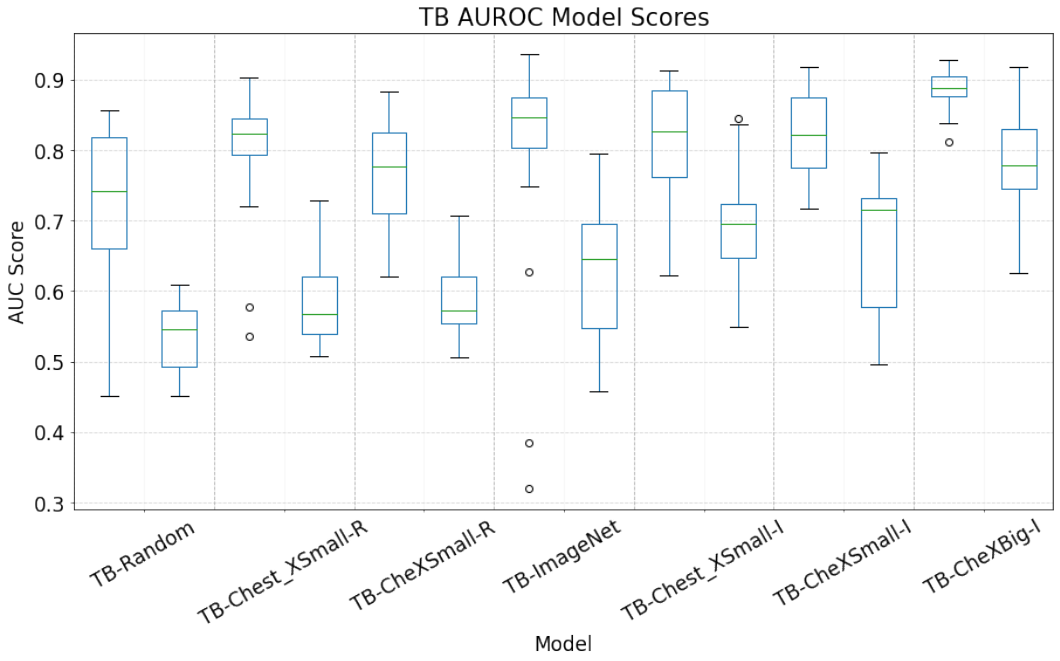

Figure 5.6: AUC measurements for the classification of TB (Shenzhen test measurements on the left, Montgomery County on the right).

5.2.2 ImageNet Baseline TB models

This subsection first approaches the TB-ImageNet, TB-Chest-XSmall-I and TB-CheXSmall-I models, since they show lesser differences between each other, and the Chest X-ray Baselines use a similar amount of data for training.

WAF scores show similar performance between the three in the Shenzhen testing set, ranging from the worst-performing series, the TB-CheXSmall-I, with a median score of 0.75 and the best performing series, TB-ImageNet with a median score of 0.77. AUROC complies with the WAF score for the Shenzhen dataset, revealing a lower AUROC median value of 0.82 for the TB-CheXSmall-I series, and a higher median value of 0.85 for the TB-ImageNet series.

For the Montgomery test set, the models from TB-ImageNet achieve a lower median WAF score of 0.47, and the TB-Chest-XSmall-I achieves a higher median score of 0.53. AUROC exhibits a lower median score of 0.65 obtained for TB-ImageNet, and a higher value of 0.72 for TB-CheXSmall-I.

The statistical tests for values obtained in the Shenzhen test set are available in table 5.3. The tests carried out in the Shenzhen dataset determine that TB-Imagenet, TB-Chest-XSmall-I and TB-CheXSmall-I all reach better scores than the TB-Random model series while performing worse than the TB-CheXBig-I models.These three series do not offer any other significantly different distributions. The Montgomery test set statistical tests, available in table 5.4, determine better AUROC scores for each of the three ImageNet trained models, when compared with the AUROC distribution of TB-Random, but not showing any significant difference for the WAF distributions. Additionally, TB-Chest_XSmall-I achieves better AUROC distribution than TB-ChexSmall-R. TB-Chest-XSmall-I also shows better WAF and AUROC distributions than the TB-CheXSmall-R.

The TB-CheXBig-I series show an overall improved performance for WAF and AUROC metrics. It achieves a median WAF score of 0.82, and an AUROC value of 0.88 in the Shenzhen testing set. On the Montgomery testing set, these models earn a median WAF score of 0.65 and a median AUROC value of 0.77.

The Kruskal-Wallis H-test complies with the previous findings for the Shenzhen test set, determining improved AUROC and WAF distributions over every previously trained series. The statistical tests determine a similar case for the Montgomery test results with TB-CheXBig-I achieving improved AUROC and WAF distributions over every previous series.

Table 5.3: *P*-values from the Kruskal-Wallis H test for models trained in TB data and tested in the Shenzhen TB X-ray Set. The test compares the series in the columns with the series in the lines, pairwise. If the *p*-value shows statistical significance (*p*-value < 0.01), the font colour is changes to green if the column performs better than the row, or red if otherwise.

	TB-Random	TB-Chest_XSmall-R	TB-CheXSmall-R	TB-ImageNet	TB-Chest_XSmall-I	TB-CheXSmall-I	TB-CheXBig-I
TB-Random							
WAF	1.000	0.003	0.426	0.003	0.001	0.002	0.000
AUC	1.000	0.009	0.258	0.001	0.001	0.002	0.000
TB-Chest_XSmall-R							
WAF	0.003	1.000	0.091	0.491	0.594	0.615	0.000
AUC	0.009	1.000	0.154	0.107	0.362	0.442	0.000
TB-ChestXSmall-R							
WAF	0.426	0.091	1.000	0.050	0.022	0.044	0.000
AUC	0.258	0.154	1.000	0.024	0.022	0.044	0.000
TB-ImageNet							
WAF	0.003	0.491	0.050	1.000	0.954	0.944	0.000
AUC	0.001	0.107	0.024	1.000	0.823	0.900	0.000
TB-Chest_XSmall-I							
WAF	0.001	0.594	0.022	0.954	1.000	0.955	0.000
AUC	0.001	0.362	0.022	0.823	1.000	0.967	0.000
TB-CheXSmall-I							
WAF	0.002	0.615	0.044	0.944	0.955	1.000	0.001
AUC	0.002	0.442	0.044	0.900	0.967	1.000	0.001
TB-CheXBig-I							
WAF	0.000	0.000	0.000	0.000	0.000	0.001	1.000
AUC	0.000	0.000	0.000	0.000	0.000	0.001	1.000

Table 5.4: *P*-values from the Kruskal-Wallis H test for models trained in TB data, similar to table 5.3 but portraying the testing results on Montgomery-County TB X-ray dataset.

	TB-Random	TB-Chest_XSmall-R	TB-CheXSmall-R	TB-ImageNet	TB-Chest_XSmall-I	TB-CheXSmall-I	TB-CheXBig-I
TB-Random							
WAF	1.000	0.698	0.084	0.771	0.021	0.162	0.000
AUC	1.000	0.012	0.006	0.001	0.000	0.000	0.000
TB-Chest_XSmall-R							
WAF	0.698	1.000	0.129	0.727	0.014	0.117	0.000
AUC	0.012	1.000	0.548	0.059	0.000	0.021	0.000
TB-ChestXSmall-R							
WAF	0.084	0.129	1.000	0.112	0.002	0.026	0.000
AUC	0.006	0.548	1.000	0.235	0.000	0.059	0.000
TB-ImageNet							
WAF	0.771	0.727	0.112	1.000	0.130	0.301	0.000
AUC	0.001	0.059	0.235	1.000	0.023	0.276	0.000
TB-Chest_XSmall-I							
WAF	0.021	0.014	0.002	0.130	1.000	0.625	0.002
AUC	0.000	0.000	0.000	0.023	1.000	0.548	0.000
TB-CheXSmall-I							
WAF	0.162	0.117	0.026	0.301	0.625	1.000	0.003
AUC	0.000	0.021	0.059	0.276	0.548	1.000	0.001
TB-CheXBig-I							
WAF	0.000	0.000	0.000	0.000	0.002	0.003	1.000
AUC	0.000	0.000	0.000	0.000	0.000	0.001	1.000

Chapter 6

Discussion

The development of fast and precise diagnostic tools is a requirement for the future, being cheaper and easier to distribute than trained professionals in impoverished regions. Improved automated tools help in the early detection and containment of infectious diseases, saving economic resources in struggling communities. Although this work focuses on TB, the improvement and exploration of diagnostic tools are interchangeable with other infectious diseases.

Other works have successfully deployed CNNs capable of the classification of TB (see section 3.2). This work explores CNNs in a constrained setting, using public datasets with a minimum amount of X-ray images for the training of TB models, and Transfer Learning. Transfer learning should help augment the performance of CNNs on limited data, a hypothesis supported by other works (Lakhani & Sundaram, 2017; Rajpurkar et al., 2017; Gozes & Greenspan, 2019) that use a similar approach for Chest X-rays. Bioinformatics consensus determines that Transfer Learning through ImageNet Baselines improves model performance on CNNs trained for medical applications. However, few works explore to what extent an ImageNet Baseline helps model performance. Models trained from a Baseline tuned for a similar problem[1] should perform better when compared with models trained with an ImageNet Baseline.

Before proceeding, it is essential to establish a couple of concepts:

- **Overfitting**, occurs when the model fits too well to the training data, showing high accuracy in the training subset, and poor accuracy in the validation subset. This phenomenon might happen with extended sessions of training but is preventable with a proper validation subset. A validation subset is used at the end of each epoch to measure the model loss and determine early stopping. If a model overfits, it won't perform well on the validation subset, unless there are shared images between the training and validation subset. If there are shared images between the training and validation subset, the loss will lower regardless of Overfitting. The implementation should set extra care in avoiding shared images between the training and validation subset.

[1] For example a Baseline adjusted in Chest X-rays and used in the training of Chest X-ray based TB classifier

- **Specialization** describes models with poor generalization capacity. A model that generalizes well shows minimal difference between the performance obtained in an Internal and an External Testing Set. A Specialized model performs well on the Internal Testing Set, which shares the same properties as the training data. However, performance suffers major losses when tested on an External Testing Set.

The following section provides the discussion of the results gathered in this work. Section 6.1 discusses the results obtained for the training of Chest X-ray Baselines, comparing the metrics used, and the differences between Random and ImageNet Baselines. Section 6.1 discusses the results obtained for the training of TB Classifiers, discussing in detail the effects of Transfer Learning in model performance, and the dynamics of the metrics used.

6.1 Lung Disease Classifiers

A general observation of the Lung Disease models shows values with minimal differences between models using Random and ImageNet Baselines. According to the statistical tests, ImageNet Baselines only perform better than Random Baselines in Mass and Nodule's classification on the Chest_-X-small series. However, when used as Baselines for TB models, models trained with ImageNet Baselines perform much better than those trained with Random Baselines. The extensive training on the ImageNet dataset should provide much more robust features, resulting in the improved Transfer Learning observed in TB models.

The Random and ImageNet Baselines might not show significant differences due to the nature of the testing subset used. The test subsets are internal testing sets, which means they originate from the same dataset as the training data (see section 4.2). Random Baselines have no developed features before training, the features obtained during training might be sub-optimal, specialized for the classification of the training data. Specialization produces favourable test results on the internal testing set, concealing possible generalization issues linked with Random Baselines. TB classifiers have the same issue. TB classifiers trained with Random Chest X-ray Baselines do not show significant differences to the ImageNet Chest X-ray Baselines on the internal testing set. The external testing set reveals improved results for ImageNet models, that were not verified for the internal testing set (see table 5.3 for internal testing results and table 5.4 for external testing results).

Of course the small differences verified for Random and ImageNet Baselines can also be the result of training on large datasets. (Raghu et al., 2019) supports this hypothesis, where the authors find minimal performance improvements while performing Transfer Learning for the training of large CNNs on the entire CheXpert dataset, using an ImageNet Baseline.

Both the ChestXray8 and CheXpert models portray a clear relationship between the number of images labelled positive for a given class and the specific performance on that class, with higher performance for standard classes and lower performance for rare classes. The number of positive cases in a rare class can be exceedingly low when compared to the negative cases, for which the CNN learns to negate the positive class, lowering Recall. Since the AP metric uses Recall to

summarise the Precision-Recall curve at multiple thresholds, and Recall is very low for rare classes, the AP metric registers a lower score by extension (see eq. (5.2)). Given the extensive amount of negative cases, the loss is partially minimized by outputting negative outputs. When a model learns to produce negative outputs for rare classes, the loss is already very low, leading to fewer modifications by the optimizer that could lead to the correct classification of the rare cases.

For example, consider that in a hypothetical total of 1 thousand images, a single image is labelled positive for a given class. The Network solves perfectly 99.9% of the cases when the outcome is always negative, regardless of input. For example, a network that outputs 0 (negative)[2] for every image, achieves a null loss during 99,9% of the training procedure. Due to Gradient Descent Optimizers' inner workings such as Adadelta, a repeated loss equal to zero lowers the learning rate, leading to minimal changes to the network parameters(Ruder, 2016) and preventing the correct classification of the rare classes.

Smaller amounts of positive labels in a skewed dataset are challenging to learn. For example, models trained in the ChestX-ray8 data achieve a relatively low performance in the classification of **Hernia**, with an average of 141 images per training subset. Note that while Hernia has 141 positive labels, the rest of the ChestX-ray8 dataset contains 7018 negative labels for Hernia. Since the Loss Function treats positive/negative labels for Hernia in an equal manner, not taking into account its biased distribution, Hernia classification fails because the model adjusts for the negation of Hernia. When provided with testing images, the model tends to negate Hernia. The testing sets' low AP measurements result from the model returning very few true positives and many false negatives.

6.2 TB Classifiers

The following section splits into different subsections to provide a structured discussion of the training of TB models. Section 6.2.1 approaches the performance of models trained with the multiple baselines gathered in the previous stage. Section 6.2.2 addresses the different results given by AUROC and WAF metrics, exploring the nature behind this phenomenon, which gives place to explore the thresholds that provide optimal model performance approached in 6.2.3. And finally, 6.2.4 wraps the discussion regarding the optimal model performance by providing a simple method for threshold estimation.

6.2.1 General performance

The AUROC results, when subject to the Kruskal-Wallis tests, show very nuanced results for models trained with either Random or ImageNet Baselines especially for CNNs tuned in Chest X-ray data. All the ImageNet models perform better than TB-Random, a series of models trained from a Random Baseline, for which features develop in the limited Shenzhen TB dataset.

This is not surprising, since the number of images provided in the Shenzhen training set is not appropriate for CNNs training with such a high number of parameters. The default ImageNet

[2]For a label setting where 0=negative, and 1=positive.

Baseline, used without modification in the TB-ImageNet series, employs 14 million images for training. CNNs are known to require extensive amounts of data to train, and the 462 images provided in the Shenzhen training set are not enough to promote robust features. Although the ImageNet Baselines do not train in Chest X-rays, the higher-order features found near the input might help improve the learning process overall, containing simple features such as edges or shapes critical in the extraction of visual data. These high order features are not present in Random Baselines, resulting in the reduced performance of TB classifiers trained from either Random Baselines, or Random Chest X-ray Baselines.

When looking at the Montgomery testing set results, the metrics do not seem to agree. There are multiple cases where AUROC reports improved results, where WAF does not. AUROC provides a better insight into the model "actual" performance without the degrading effect from the threshold measured in the training set (more on this in 6.2.2).

The AUROC results reported on the Shenzhen and Montgomery testing dataset also show some interesting differences. For example, the results for TB-Chest_XSmall-R in the Shenzhen testing set shows improved outcomes compared with TB-Random, which disappear when testing on the Montgomery testing set. Another Example is the TB-CheXSmall-R, which does not show any significant differences in one testing set compared with the TB-Random series, but does so in the other. Or also the TB_Chest_XSmall-I, not reporting substantial differences when compared with TB_Chest_XSmall-R and TB-CheXSmall-R, but does so in the other. These differences occur when transitioning from the Shenzhen testing set, the internal testing set, to the Montgomery testing set, the external testing set. These differences reflect the improved generalization capacity of some series over the others. Models with poor generalization manage to perform well in the internal testing set, not showing significant differences with another model with better generalization capacity. When testing carries over to the external testing set, the specialized models' results drop, raising substantial differences that the internal testing set does not capture.

The internal testing set shares the same properties as the training data. This type of testing procedure might favour an undesirable model, a product of the adaptation of fragile features to the training data's properties. When faced with testing data with properties different from those found in the training data, model performance quickly drops since the features of specialized models do not extract sufficient information. On the other hand, the improved features found in the ImageNet Baselines, a result of the extensive training in the ImageNet dataset, provide robust features that allow for a higher degree of generalization in external datasets. The internal testing sets cannot capture the lack of generalization portrayed by models trained with Random X-ray Baselines, heightening the importance of testing on external testing sets to guarantee the robustness of the CNN features.

TB-CheXBig-I models improve their AUROC performance over every previous model trained before. The TB-CheXBig-I series Baseline trains almost three times the amount of data provided to TB-CheXSmall or TB-Chest-XSmall series. It gives a much stronger Baseline for the training on small datasets such as the Shenzhen Dataset. This difference is not related to the CheXpert

dataset's properties, since the results show that TB-CheXSmall-I performs better than TB-Chest-XSmall-I, both using a similar amount of training data.

The WAF measurements show similar findings in the Shenzhen testing set but hold back on some of the improvements reported by AUROC metrics in the Montgomery testing set. TB-ImageNet, TB-Chest-XSmall-I and TB-CheXSmall-I portray these improvements in the Montgomery testing set. While AUROC reports progress over TB-Random in the Montgomery testing set, WAF does not show this improvement. This difference might result from the sub-optimal threshold measured in the training set. The determination of an optimal threshold for the testing set is crucial to maximizing the model's performance. The mismatch between the two scores possibly results from AUROC using a range of thresholds that best portray the optimal performance in the testing set (more on this in 6.2.2).

For TB-CheXBig-I, the results for WAF determine the same differences in the Kruskal-Wallis tests for the internal and external tests. These results further suggest the massive impact of dataset size in Baselines' training for Transfer Learning. Baselines trained on 300 thousand X-rays show improved performance overall, regardless of generalization or metric.

6.2.2 Disagreeing WAF and AUROC metrics

WAF and AUROC represent two different problems in our study. WAF provides the test WAF score for a threshold established from the training data for whom the truth values provide the fundamental basis for CNN learning procedures. AUROC finds a range of thresholds for the testing data which best portrays the ratio of TPR (True Positive Rate) and FPR (False Positive Rate). AUROC captures the models' absolute optimal performance in the testing data while requiring prior knowledge of the truth labels.

The mismatch between AUROC and WAF derives from the threshold used to label the output of the CNNs to zero and one. To better understand this occurrence, fig. 6.1 plots the distribution of WAF for a range of thresholds between zero and one. Each plot portrays the results of a single model that achieves median WAF in the corresponding series. The green vertical line represents the threshold used for the measurement of WAF. As stated before, at the beginning of chapter 5, this specific threshold maximizes WAF in the training dataset. Each plot shows an additional two lines, one blue, the estimated threshold in the Shenzhen testing dataset, and another red, indicating the threshold calculated in the Montgomery testing set. The colour of each threshold and WAF distribution is specific to dataset[3] to facilitate the comprehension of each plot. While the rest of this section focuses on the patterns of the curves obtained, the estimated thresholds are further down approached in more detail in subsection 6.2.3.

The distribution of the WAF values follows a similar pattern in the Shenzhen training, represented by the green curve, and Shenzhen testing set, shown by the blue curve. Every model series shows this pattern, and the equal distribution means that similar threshold values determine the maximum WAF in both cases. The proximity between these optimal training and testing thresholds

[3]Green - Shenzhen Training data; Blue - Shenzhen Testing data; Red - Montgomery Testing data

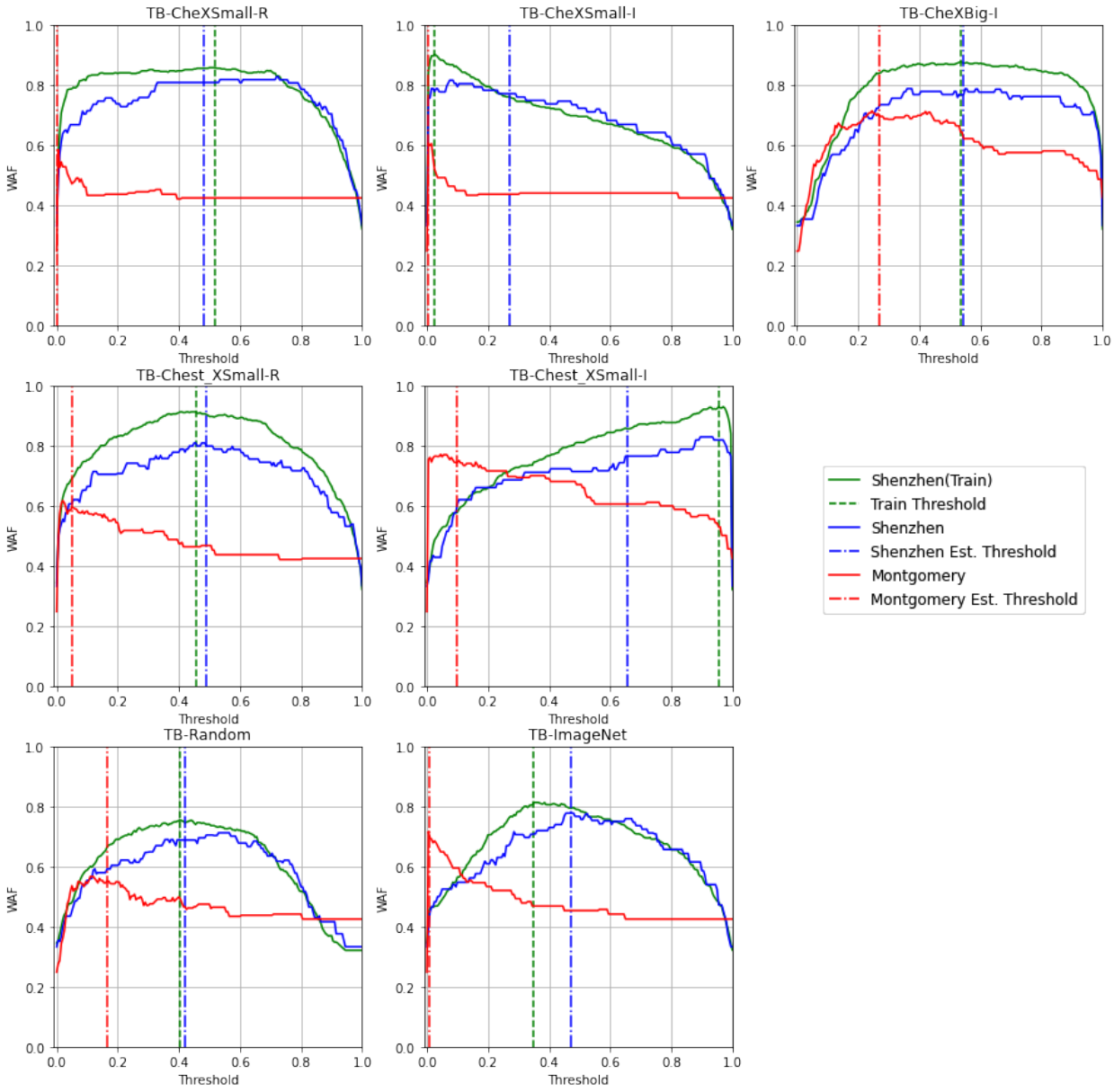

Figure 6.1: Plots of the WAF scores obtained with thresholds ranging 0 to 1. Each plot draws one curve for each source of data. The green vertical line corresponds to the threshold that maximizes WAF in the training dataset. The blue and red vertical lines correspond to estimated test based thresholds.

on the Shenzhen test set justifies why AUROC and WAF agree with each other in the Kruskal-Wallis tests in table 5.3. AUROC summarises the performance of a model with multiple thresholds, emphasizing the optimal performance of the model. The threshold that maximizes WAF in the training set achieves values close to optimal in the Shenzhen testing set, providing similar results to AUROC in the statistical analysis.

Regarding the WAF curve of the Montgomery testing set, drawn with the colour red, and the WAF curve of the Shenzhen training set, marked with the colour green, these two curves show very different shapes. The distinct shape suggests that the optimal threshold measured in training data does not produce an optimal WAF value for the Montgomery testing set. The Montgomery distributions characterize optimal WAF values for lower thresholds, noticeable by a large increase of performance near values closer to zero. Shenzhen distributions achieve optimal WAF values for higher thresholds that return very poor Montgomery testing set values. Such a finding supports the different results portrayed by WAF and AUROC metrics in the Kruskal-Wallis test for the Montgomery testing set in table 5.4.

6.2.3 Optimal WAF for lower thresholds

CNNs trained for TB classification are trained to output the value one (1) for a positive image for TB and zero (0) for negative images. The threshold that optimally separates the two classes shifts from one training session to the other. Table 6.1 shows the median thresholds measured in the training set, used for maximizing WAF, and the estimated thresholds, measured in the testing set. Estimated thresholds are assessed in section 6.2.4. It is visible that the series with odd distributions for the median model in fig. 6.1, such as TB-CheXSmall-I, portray a standard median threshold measured in the training set, as table 6.1. This table does not show noteworthy differences for the median threshold value measured in the training set and the estimated threshold measured in the Shenzhen test set. However, in the Montgomery test set, where the CNNs perform the worst, the optimal threshold achieves much lower values than the training data threshold.

Between the filters that process the original image from one convolution layer to the other, and the final Dense Network that processes the collective information, it is difficult to tell with certainty why these values are so low. The high WAF score for smaller thresholds in the Montgomery testing set determines that the model separates the negative and positive TB images in a non-random manner.

Training on small amounts of data might lead to the generation of features specific to the training data's properties, in the lower Convolutional Blocks, closer to the output. When provided with the Montgomery data, whose properties diverge substantially from the training dataset, these features have reduced sensitivity, propagating smaller values throughout the Network, ultimately resulting in reduced performance and an output value closer to 0. Since the Shenzhen testing set has similar properties to the training data, the Network portrays higher sensibility to the images, outputting larger values, and higher performance. When estimating the Shenzhen testing dataset threshold, the threshold is much larger and closer to 0.5, which supports this hypothesis.

Table 6.1: Median thresholds obtained in training, and estimated thresholds obtained from a small sample of the testing sets.

Model	Testing Dataset	Measured Threshold	Estimated Threshold
TB-Random	Shenzhen	0.255	0.315
	Montgomery	0.255	0.0869
TB-Chest_XSmall-R	Shenzhen	0.455	0.496
	Montgomery	0.455	0.035
TB-CheXSmall-R	Shenzhen	0.355	0.480
	Montgomery	0.355	0.006
TB-ImageNet	Shenzhen	0.400	0.500
	Montgomery	0.400	0.100
TB-Chest_XSmall-I	Shenzhen	0.520	0.506
	Montgomery	0.520	0.123
TB-CheXSmall-I	Shenzhen	0.585	0.517
	Montgomery	0.585	0.046
TB-CheXBig-I	Shenzhen	0.360	0.500
	Montgomery	0.360	0.126

6.2.4 Estimating a better threshold

To estimate a "better" threshold, this work makes sure that the metrics adopted provide a realistic view into the model performance, and that the estimation of the new threshold remains feasible in a real-life scenario.

This work initially used F1 Score in place of WAF, being a commonly used metric in the field. The collected results displayed improved F1 Score for lower thresholds, visible in fig. 6.2 which portrays the F1 Score according to the threshold in a similar approach fig. 6.1. The F1 and AUROC scores disagreed in the same manner that WAF does, using a training measured threshold. Since AUROC uses a set of thresholds that best describe the testing data's performance, the used threshold was the most obvious suspect contributing to the lack of coherence between the two metrics, which led to the discovery of the suspiciously low optimal thresholds.

Previous research determines that the use of very low thresholds (close to 0) for F1 Score does not produce good results on biased testing sets. It promotes completely uninformative models that classify every case as positive[4], undesirable when the prevalence of positive cases is low (Lipton et al., 2014). Both the Shenzhen and Montgomery testing sets have a balanced amount of positive and negative cases, which leaves F1 Score less problematic. However, the prevalence of low thresholds was still a significant cause of concern. To further explore how a "completely uninformative" model would behave, this work generates an array with random numbers between zero and one, in a range of one thousand equidistant units. This array is used in the same manner as fig. 6.1, measuring

[4]The way this work uses a threshold determines that anything above its value is positive. Very low thresholds determine everything as positive.

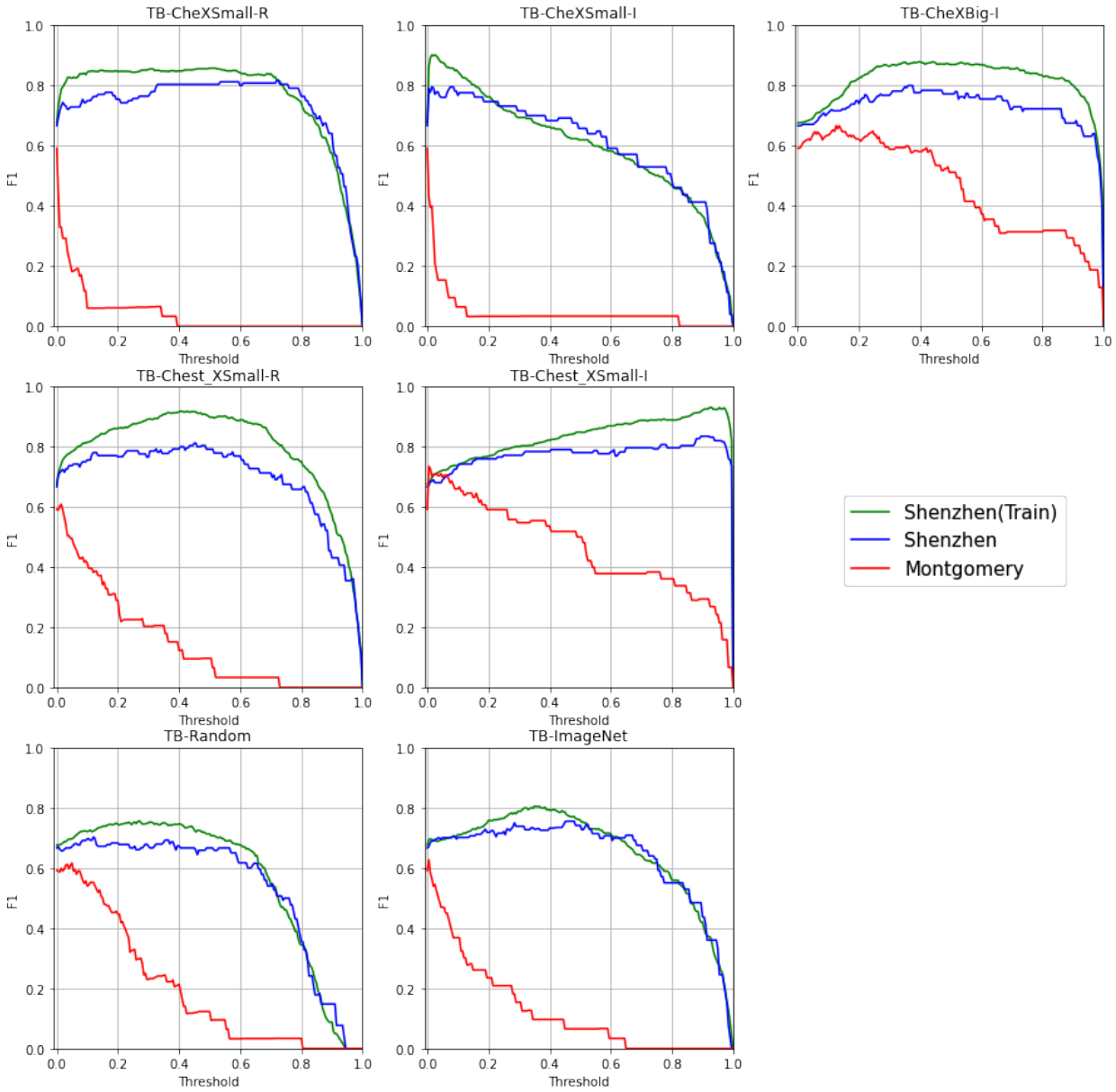

Figure 6.2: Plots of the F1 scores obtained with thresholds ranging 0 to 1. Each plot draws one curve for each source of data.

the Montgomery testing set WAF and F1 at different thresholds. Figure 6.3 shows the result, with F1 Score showing increasing values towards lower thresholds.

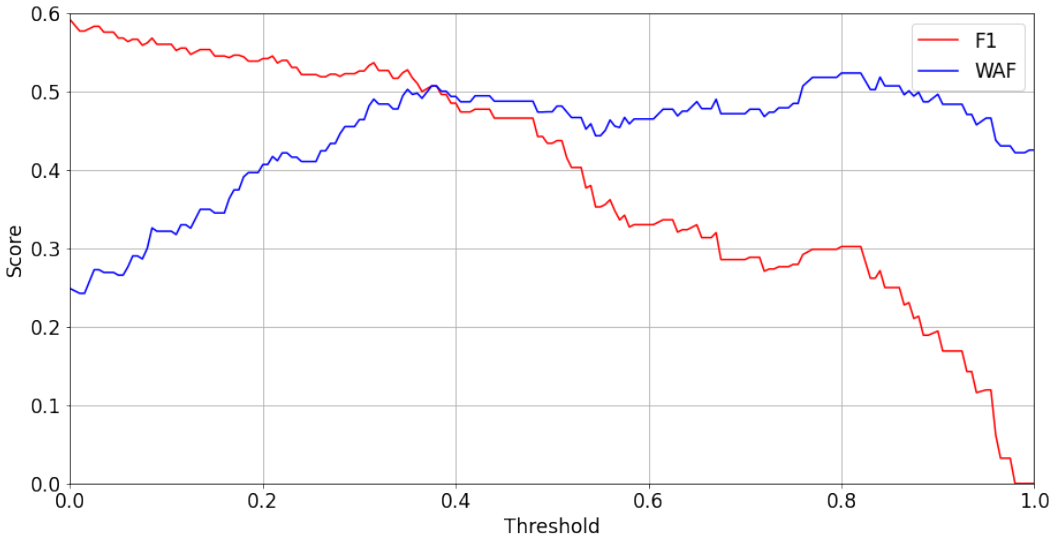

Figure 6.3: F1 and WAF measured in a limited range of thresholds. The curves represent a completely uninformative model that outputs random values between 0 and 1 for the Montgomery testing set.

This means that when using F1 Score and very low thresholds an utterly random model achieves an F1 Score of 0.6. To negate this factor out of the results, WAF takes the place of F1 Score. It portrays lower values for lower thresholds, ensuring that the results reported are not the product of the metric's misuse.

With all the metric related issues sorted the actual estimation of optimal threshold proceeds. A test-based threshold assessment measures the average Score of two labelled images, one positive and one negative, repeating this process for each positive-negative pair. The median value from the resulting array of averages provides the test based threshold. This approach ensures a minimal requirement of available labelled data to optimize a TB model's deployment in a real-life setting. Figure 6.4 shows the results, showing the WAF score achieved in the Montgomery testing set with a trained, measured threshold, and with a test estimated threshold, for each of the model series. The median WAF values achieved by estimated thresholds are always superior to the values achieved with thresholds measured in the training data. This shows the ill-suited nature of the threshold measured in the training data, further supporting the importance of an appropriate threshold during the deployment of a CNN based tool.

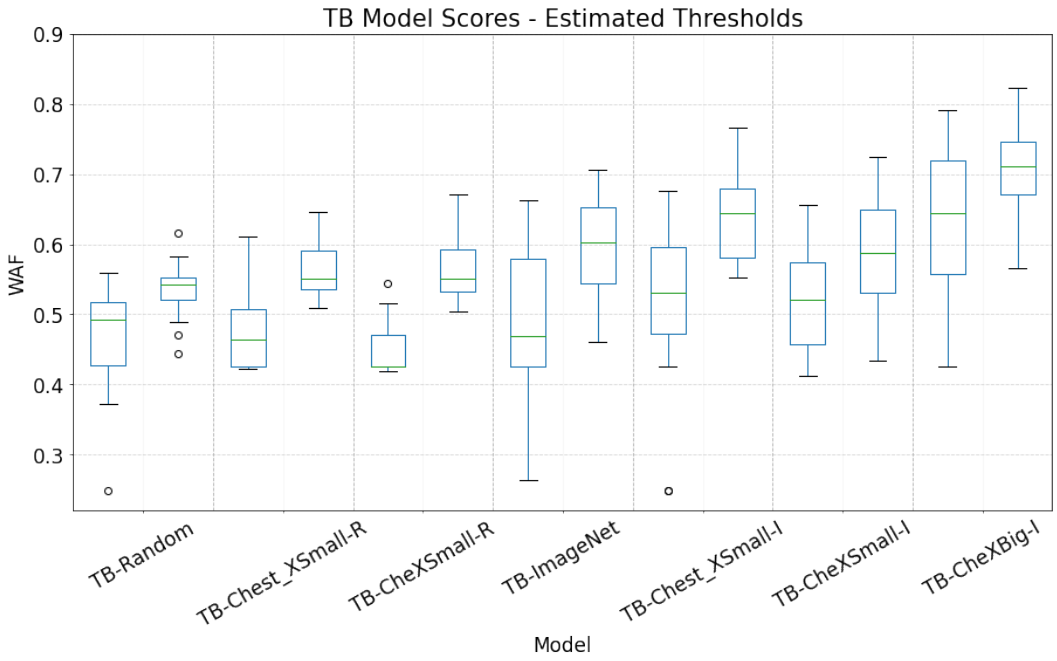

Figure 6.4: Montgomery WAF results measured with a train measured threshold (on the right) and with a test estimated threshold (on the left).

Chapter 7

Conclusion and Future Work

The following sections briefly describe the general work and results obtained as part of this work in 7.1, and 7.2 discusses some ideas on how to follow up according to our initial aim.

7.1 Conclusion

The amount of open-access labelled TB X-Ray images is fairly low, with the Shenzhen Hospital and the Montgomery-County X-ray dataset providing a total of 814 images for the training, validation and testing of the CNNs (Jaeger et al., 2014). The limited amount of resources encourages the use of novel techniques to improve the training of CNNs for the classification of TB. Transfer Learning in CNNs provides lower level features, such as shapes and edges, usable in new tasks. These lower level features benefit from vast training datasets, such as the ImageNet dataset, promoting robust and effective information capture. Transfer Learning bypasses the need for very high amounts of data, reusing the learned features for a new task. This work explores Transfer Learning's use, using multiple Baselines (some generated as part of this work) to evaluate the improvement of TB models using different Baselines.

On a first stage this work trains a total of 17 Chest X-ray Baselines. A total of 4 series of models train using Random and ImageNet Baselines, each using the ChestX-ray8 dataset, and a shortened version of the CheXpert dataset, providing a similar amount of images. An additional series of models trains on the complete CheXpert dataset. All the models use a DenseNet121 CNN architecture, sharing the same image augmentation procedure, training until the validation loss does not improve after five epochs. The statistical tests do not determine any significant differences between models trained using Random and ImageNet Baselines, except for AP scores for Nodule and Mass for models trained in ChestX-ray8 dataset. AP metrics provide the most reliable results showing higher performance for diseases with a higher number of positive cases in the training dataset. AUROC metrics do not portray a realistic outcome, providing very optimistic values for rare diseases. These results further stress the importance of carefully chosen metrics when evaluating model performance.

The second stage addresses the classification of TB models. A total of 155 models train using Random, ImageNet, and the Chest X-ray Baselines prepared in the previous step. The training uses the Shenzhen Hospital TB X-ray set and shares the same image augmentation procedure, stopping after the validation loss does not improve after ten epochs. Here we determine an increased generalization for models trained with a primary ImageNet Baseline, producing competent models that lose less performance when classifying the Montgomery TB X-ray set, an external dataset. The improved generalization is possibly the result of the enhanced features captured during training on the ImageNet dataset containing 14 million images. The statistical tests capture this increasing trend in the Montgomery testing set results of TB-Chest_XSmall-I series using the 70 thousand ChestXray8 Baseline, and the TB-CheXBig-I using the 200 thousand CheXpert Baseline. TB-CheXBig-I provides the best improvement in performance when used in the training of TB classifiers, achieving a median WAF value of 0.65, and a median AUROC value of 0.77 in the Montgomery Testing Set. Additionally, this work reports a significant difference between the results achieved by AUROC and WAF metrics, with AUROC portraying promising results in external testing sets. This difference is related to the general decrease in output values when the CNNs process external datasets. This general decrease leaves the threshold measured in the training data unsuitable for external testing sets. This work corrects this by estimating the threshold with the average of a negative and positive value, successfully achieving more appropriate thresholds that increase every model's performance regardless of Baseline.

In conclusion, this work effectively uses Transfer Learning to improve the training of CNNs on a concise amount of images. Our findings indicate that the current public access X-ray datasets provide the necessary data for the generation of strong baselines. These strong baselines can be shared and reused freely without additional computational costs, lowering the need for extensive amounts of labelled data to implement a competent CNN model. The high requirement of data is essential when tackling rare diseases such as TB where labelled data availability is scarce. CNN models trained with a substantial baseline generalize better to external data, therefore are better suited to perform when assigned with images captured from different X-ray equipment, in different settings and with a multitude of varying noise not present in the training data.

7.2 Future Work

The auspicious results obtained for Baselines trained in extensive datasets encourages the training of Baselines on the MIMIC-III dataset. It is the largest public Chest X-ray dataset at the time of writing. However, it calls for resources that were not available in this work. It would be interesting to explore how the 300 thousand high-quality images could further improve the results obtained by our TB models. Another approach not explored here could combine the ChestX-ray8 and CheXpert dataset for the training of a single Baseline that benefited from the increased training dataset size and higher diversity of images.

Additionally there are a couple of caveats that our work did not correct. One of them is the general Chest X-ray Baselines training, for which training does not achieve its full potential due to

the biased distribution of some diseases. Our work uses all the labels available in the large Chest X-ray datasets, following previous authors' work such as (Gozes & Greenspan, 2019). However, it would be interesting to see how much a well-trained model for a single disease, balanced, would improve TB models' training when used as a Baseline. Another point that our work fails to achieve due to time limitations is the training of a complete CheXpert model with Random Baselines. Our findings suggest that this deliverable was not crucial since the ImageNet Baselines tend to deliver better models overall. The complete CheXpert model's considerable size could either give the worst results than the ImageNet Baseline or similar results due to the extensive training procedure.

The information on gender, age and position of capture is also available in the original TB datasets. The use of this information provides a generalized improvement in performance, as suggested by (Gozes & Greenspan, 2019). However, a CNN trained with this type of information would require the input of metadata for all further evaluations, limiting the deployment scope. It would be interesting for future works to combine the high reasoning power of CNNs with the flexibility of other traditional algorithms such as Bayesian Networks. Such a procedure could allow the integration of metadata without requiring its input for a full diagnosis. Such an implementation could increase the amount of metadata used, taking into account the information provided for each subject in the Shenzhen testing set and any other interesting information related to the topic. This implementation could further increase the correctness of the matter's full evaluation with limited amounts of X-ray images.

Chapter 8

References

References

Akolo, C., Adetifa, I., Shepperd, S., & Volmink, J. (2010). Treatment of latent tuberculosis infection in hiv infected persons. *Cochrane database of systematic reviews*(1).

Aronson, A. R., & Lang, F.-M. (2010). An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, *17*(3), 229–236.

Bothamley, G., Ditiu, L., Migliori, G., & Lange, C. (2008). Active case finding of tuberculosis in europe: a tuberculosis network european trials group (tbnet) survey. *European Respiratory Journal*, *32*(4), 1023–1030.

Brady, A., Laoide, R. Ó., McCarthy, P., & McDermott, R. (2012). Discrepancy and error in radiology: concepts, causes and consequences. *The Ulster medical journal*, *81*(1), 3.

Burrill, J., Williams, C. J., Bain, G., Conder, G., Hine, A. L., & Misra, R. R. (2007). Tuberculosis: a radiologic review. *Radiographics*, *27*(5), 1255–1273.

Cao, Y., Liu, C., Liu, B., Brunette, M. J., Zhang, N., Sun, T., ... Garcia, L. L. (2016). Improving tuberculosis diagnostics using deep learning and mobile health technologies among resource-poor and marginalized communities. In *2016 ieee first international conference on connected health: applications, systems and engineering technologies (chase)* (pp. 274–281).

Chellapilla, K., Puri, S., & Simard, P. (2006). High performance convolutional neural networks for document processing..

Ciregan, D., Meier, U., & Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *2012 ieee conference on computer vision and pattern recognition* (pp. 3642–3649).

Cireşan, D. C., Meier, U., Gambardella, L. M., & Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. *Neural computation*, *22*(12), 3207–3220.

Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J. (2011). Flexible, high performance convolutional neural networks for image classification. In *Twenty-second international joint conference on artificial intelligence.*

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on machine learning* (pp. 233–240).

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., ... McDonald, C. J. (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, *23*(2), 304–310.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Diel, R., Vandeputte, J., de Vries, G., Stillo, J., Wanlin, M., & Nienhaus, A. (2014). Costs of tuberculosis disease in the european union: a systematic analysis and cost calculation. *European Respiratory Journal*, *43*(2), 554–565.

Docker download. (2020). `https://pypi.org/project/docker/`. (Accessed: 2020-10-05)

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, *61*, 863–905.

FitzGerald, R. (2013). Commentary on: workload of consultant radiologists in a large dgh and how it compares to international benchmarks. *Clinical radiology*, *68*(5), e237–e238.

Floyd, K., Glaziou, P., Zumla, A., & Raviglione, M. (2018). The global tuberculosis epidemic and progress in care, prevention, and research: an overview in year 3 of the end tb era. *The Lancet Respiratory Medicine*, *6*(4), 299–314.

for Disease Control, C., Prevention, et al. (2006). *Instructions to panel physicians for completing new us department of state medical examination for immigrant or refugee applicant.*

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, *36*(4), 193–202.

George, R., Gully, P., Gill, O., Innes, J., Bakhshi, S., & Connolly, M. (1986). An outbreak of tuberculosis in a children's hospital. *Journal of Hospital Infection*, *8*(2), 129–142.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press. (`http://www.deeplearningbook.org`)

Gozes, O., & Greenspan, H. (2019). Deep feature learning from a hospital-scale chest x-ray dataset with application to tb detection on a small-scale dataset. In *2019 41st annual international conference of the ieee engineering in medicine and biology society (embc)* (pp. 4076–4079).

Guendel, S., Grbic, S., Georgescu, B., Liu, S., Maier, A., & Comaniciu, D. (2018). Learning to recognize abnormalities in chest x-rays with location-aware dense networks. In *Iberoamerican congress on pattern recognition* (pp. 757–765).

Hahn, R. G. (1943). Tuberculosis in the household, its occurrence in marital partners and other members of the household when the primary case was a parent or another member of the family. *American Review of Tuberculosis*, *47*(3), 316–324.

Haley, C. E., McDonald, R. C., Rossi, L., Jones, W. D., Haley, R. W., & Luby, J. P. (1989). Tuberculosis epidemic among hospital personnel. *Infection Control & Hospital Epidemiology*, *10*(5), 204–210.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527-1554. Retrieved from `https://doi.org/10.1162/neco.2006.18.7.1527` (PMID: 16764513) doi: 10.1162/neco.2006.18.7.1527

Hirose, Y., Yamashita, K., & Hijiya, S. (1991). Back-propagation algorithm which varies the number of hidden units. *Neural networks*, *4*(1), 61–66.

Houben, R. M., & Dodd, P. J. (2016). The global burden of latent tuberculosis infection: a re-estimation using mathematical modelling. *PLoS medicine*, *13*(10), e1002152.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4700–4708).

Hwang, S., Kim, H.-E., Jeong, J., & Kim, H.-J. (2016). A novel approach for tuberculosis screening based on deep convolutional neural networks. In *Medical imaging 2016: computer-aided diagnosis* (Vol. 9785, p. 97852W).

Ilievska-Poposka, B., Metodieva, M., Zakoska, M., Vragoterova, C., & Trajkov, D. (2018). Latent tuberculosis infection-diagnosis and treatment. *Open access Macedonian journal of medical sciences*, *6*(4), 651.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... Shpanskaya, K. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 590–597).

Islam, M. T., Aowal, M. A., Minhaz, A. T., & Ashraf, K. (2017). Abnormality detection and localization in chest x-rays using deep convolutional neural networks. *arXiv preprint arXiv:1705.09850*.

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

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., ... Mark, R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, *3*, 160035.

Kandel, I., & Castelli, M. (2020). The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express*.

keras. (2020). `https://keras.io/getting_started/`. (Accessed: 2020-10-05)

Keras documentation for aplications. (2020). `https://keras.io/applications/`. (Accessed: 2020-04-29)

Keras documentation for functional api. (2020). `https://keras.io/guides/functional_api/`. (Accessed: 2020-10-05)

Keras documentation for optimizers. (2020). `https://keras.io/api/optimizers/`. (Accessed: 2020-05-11)

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–

1105).

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, *47*(260), 583–621.

Lakhani, P., & Sundaram, B. (2017). Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, *284*(2), 574–582.

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE transactions on neural networks*, *8*(1), 98–113.

Leaman, R., Khare, R., & Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, *57*, 28–37.

Lipton, Z. C., Elkan, C., & Naryanaswamy, B. (2014). Optimal thresholding of classifiers to maximize f1 measure. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 225–239).

Liu, C., Cao, Y., Alcantara, M., Liu, B., Brunette, M., Peinado, J., & Curioso, W. (2017). Tx-cnn: Detecting tuberculosis in chest x-ray images using convolutional neural network. In *2017 ieee international conference on image processing (icip)* (pp. 2314–2318).

Liu, Y. H. (2018). Feature extraction and image recognition with convolutional neural networks. In *Journal of physics: Conference series* (Vol. 1087, p. 062032).

Marais, B., Gie, R., Schaaf, H., Hesseling, A., Obihara, C., Starke, J., ... Beyers, N. (2004). The natural history of childhood intra-thoracic tuberculosis: a critical review of literature from the pre-chemotherapy era [state of the art]. *The International Journal of Tuberculosis and Lung Disease*, *8*(4), 392–402.

NIH, U. (2020). *Open-i: An open access biomedical search engine.*

Oh, K.-S., & Jung, K. (2004). Gpu implementation of neural networks. *Pattern Recognition*, *37*(6), 1311–1314.

Opencv. (2020). `https://pypi.org/project/opencv-python/`. (Accessed: 2020-10-05)

Organization, W. H. (2014). *Guidance for national tuberculosis programmes on the management of tuberculosis in children* (Tech. Rep.). Author. Retrieved from `https://www.who.int/tb/publications/childtb_guidelines/en/`

PATERSON, J. F., et al. (1940). Tuberculosis in married couples. *American Journal of Hygiene*, *32*(3), 67–78.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Python 3.6. (2020). `https://www.python.org/downloads/`. (Accessed: 2020-10-05)

Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019). Transfusion: Understanding transfer learning for medical imaging. In *Advances in neural information processing systems* (pp. 3347–3357).

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... Shpanskaya, K. (2017). Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv*

preprint arXiv:1711.05225.

Renfrew, D., Franken Jr, E., Berbaum, K., Weigelt, F., & Abu-Yousef, M. (1992). Error in radiology: classification and lessons in 182 cases presented at a problem case conference. *Radiology*, *183*(1), 145–150.

Riley, R., Mills, C., Nyka, W., Weinstock, N., Storey, P., Sultan, L., ... others (1959). Aerial dissemination of pulmonary tuberculosis. a two-year study of contagion in a tuberculosis ward. *American Journal of Hygiene*, *70*(2), 185–96.

Rob, B., Katherine, F., & Christopher, D. R. (2005). Achieving the millennium development goals for health cost effectiveness analysis of strategies for tuberculosis control in developing countries. *BMJ*.

Rogers, E. F. (1962). Epidemiology of an outbreak of tuberculosis among school children. *Public Health Reports*, *77*(5), 401.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Sacks, J. J., Brenner, E. R., Breeden, D. C., Anders, H. M., & Parker, R. L. (1985). Epidemiology of a tuberculosis outbreak in a south carolina junior high school. *American journal of public health*, *75*(4), 361–365.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85–117.

Schwartzman, K., & Menzies, D. (2000). How long are tb patients infectious? *CMAJ: Canadian Medical Association Journal*, *163*(2), 157.

Scikit-learn. (2020). `https://scikit-learn.org/stable/install.html`. (Accessed: 2020-10-05)

Scipy kruskall-willis h test. (2020). `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.kruskal.html`. (Accessed: 2020-11-03)

Siewert, B., Sosna, J., McNamara, A., Raptopoulos, V., & Kruskal, J. B. (2008). Missed lesions at abdominal oncologic ct: lessons learned from quality assurance. *Radiographics*, *28*(3), 623–638.

Sloot, R., Schim van der Loeff, M. F., Kouw, P. M., & Borgdorff, M. W. (2014). Risk of tuberculosis after recent exposure. a 10-year follow-up study of contacts in amsterdam. *American journal of respiratory and critical care medicine*, *190*(9), 1044–1052.

Spector, H. (1939). Marital tuberculosis, a study of 210 couples in which both husband and wife have clinical tuberculosis. *American Review of Tuberculosis*, *40*(2), 147–156.

Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Highway networks. *arXiv preprint arXiv:1505.00387*.

Steinkraus, D., Buck, I., & Simard, P. (2005). Using gpus for machine learning algorithms. In *Eighth international conference on document analysis and recognition (icdar'05)* (pp. 1115–1120).

Stop tb partnership — the global plan to end tb — the global plan to stop tb 2016 - 2020. (2019). `http://www.stoptb.org/global/plan/plan2/`. (Accessed: 2019-10-31)

Styblo, K. (1985). The relationship between the risk of tuberculous infection and the risk of developing infectious tuberculosis. *Bull IUAT*, *60*(3), 117–119.

Sultan, L., Nyka, W., Mills, C., O'grady, F., Wells, W., & Riley, R. (1960). Tuberculosis disseminators: a study of the variability of aerial infectivity of tuberculous patients. *American Review of Respiratory Disease*, *82*(3), 358–369.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 1–9).

Taddy, M. (2019). *Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions*. McGraw Hill Professional.

Tan, M., & Le, Q. V. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*.

Team, P. P., Gohagan, J. K., Prorok, P. C., Hayes, R. B., & Kramer, B.-S. (2000). The prostate, lung, colorectal and ovarian (plco) cancer screening trial of the national cancer institute: history, organization, and status. *Controlled clinical trials*, *21*(6), 251S–272S.

Tensorflow download. (2020). `https://pypi.org/project/tensorflow/`. (Accessed: 2020-10-05)

Torrey, L., & Shavlik, J. (2010). Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques* (pp. 242–264). IGI Global.

Towards fairer datasets. (2020). `http://image-net.org/update-sep-17-2019`. (Accessed: 2020-01-06)

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2097–2106).

Xie, Q., Hovy, E., Luong, M.-T., & Le, Q. V. (2019). Self-training with noisy student improves imagenet classification. *arXiv preprint arXiv:1911.04252*.

Yao, L., Poblenz, E., Dagunts, D., Covington, B., Bernard, D., & Lyman, K. (2017). Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501*.

Zhang, J., Li, Y., & Zhang, X. (2015). Mathematical modeling of tuberculosis data of china. *Journal of theoretical biology*, *365*, 159–163.