# Computer-Aided Assessment of Tuberculosis with Radiological Imaging

## From rule-based methods to Deep Learning

**Pedro Macías Gordaliza**

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in

Multimedia and Communication
Universidad Carlos III de Madrid

Advisors:

Arrate Muñoz Barrutia
Juan José Vaquero López

Tutor:

Arrate Muñoz Barrutia

February 2022

# Published and submitted contents

## Journals Articles

1. **Pedro M. Gordaliza**, Arrate Muñoz-Barrutia, Mónica Abella, Manuel Desco, Sally Sharpe, and Juan José Vaquero. *Unsupervised CT Lung Image Segmentation of a Mycobacterium Tuberculosis Infection Model*. Scientific Reports, 8(1), 12 2018. ISSN 2045-2322. doi:10.1038/s41598-018-28100-x.

   Author contributions: Pedro M. Gordaliza was responsible for conception of the study and the design of the experimental framework, software development, data analysis, writing, editing and reviewing the paper. A.M.B and J.J.V contribute to the conception of the study and interpretation of the results. S.S. provided the CT data and contributed to the revision of the manuscript. All authors were responsible for writing, editing and reviewing the paper.

   Contribution completely included in Chapter 2.

2. **Pedro M. Gordaliza**, Arrate Muñoz-Barrutia, Laura E. Via, Sally Sharpe, Manuel Desco, and Juan José Vaquero. *Computed Tomography-Based Biomarker for Longitudinal Assessment of Disease Burden in Pulmonary Tuberculosis*. Molecular Imaging and Biology, 21(1):19–24, 2 2019. ISSN 1536-1632. doi: 10.1007/s11307-018-1215-x.

   Author contributions: Pedro M. Gordaliza was responsible for conception of the study and the design of the experimental framework, software development, data analysis, writing, editing and reviewing the paper. A.M.B, J.J.V and L.E.V contribute to the conception of the study and interpretation of the results. S.S. and L.E.V provided the data. All authors were responsible for writing, editing and reviewing the paper.

   Contribution completely included in Chapter 2.

3. Paula Martin-Gonzalez, Estibaliz Gomez-de-Mariscal, M. Elena Martino, **Pedro M. Gordaliza**, Isabel Peligros, Jose Luis Carreras, Felipe A. Calvo, Javier Pascau, Manuel Desco, and Arrate Muñoz-Barrutia. *Association of visual and quantitative heterogeneity of 18F-FDG PET images with treatment response in locally advanced rectal cancer: A feasibility study*. Plos One, 15(11):e0242597, 11 2020. ISSN 1932-6203. doi:10.1371/journal.pone.0242597.

Author contributions: Paula M. G. was responsible for conception of the study, the design of the experimental framework and data analysis. Pedro M. Gordaliza was responsible for code implementation and data analysis. All authors were responsible for writing, editing and reviewing the paper.

The code employed is the same developed for Chapter 3.

4. Esperanza Naredo, Javier Pascau, Nemanja Damjanov, Gemma Lepri, **Pedro M Gordaliza**, Iustina Janta, Juan Gabriel Ovalles-Bonilla, Francisco Javier López-Longo, and Marco Matucci-Cerinic. *Performance of ultra-high-frequency ultrasound in the evaluation of skin involvement in systemic sclerosis: a preliminary report*. Rheumatology, 59(7):1671–1678, 10, 2019. ISSN 1462-0324. doi:10.1093/rheumatology/kez439 .

   Author contributions: E.N. was responsible for conception of the study, the design of the experimental framework and data analysis. Pedro M. Gordaliza was responsible for code implementation and data analysis. All authors were responsible for writing, editing and reviewing the paper.

   The code employed is the same developed for Chapter 3.

5. Covadonga M. Diáz-Caneja, Clara Alloza, **Pedro M. Gordaliza**, Alberto Fernández-Pena, Luciá De Hoyos, Javier Santonja, Elizabeth E.L. Buimer, Neeltje E.M. Van Haren, Wiepke Cahn, Celso Arango, René S. Kahn, Hilleke E. Hulshoff Pol, Hugo G. Schnack, and Joost Janssen. *Sex Differences in Lifespan Trajectories and Variability of Human Sulcal and Gyral Morphology*. Cerebral cortex, 31(11):5107–5120, 11 2021. ISSN 1460-2199. doi:10.1093/cercor/bhab145.

   Author contributions: C.M.D was responsible for conception of the study and the design of the experimental framework. Pedro M. Gordaliza was responsible for data analysis. All authors were responsible for writing, editing and reviewing the paper.

6. Verónica Aramendía-Vidaurreta, **Pedro M. Gordaliza**, Rebeca Echeverria-Chasco, Gorka Bastarrika, A Muñoz-Barrutia, and María A. Fernández-Seara. *Reduction of motion effects in myocardial arterial spin labeling.*. Magnetic Resonance in Medicine, vol. 87, n. 3, pp. 1261-1275, March 2022. doi:10.1002/mrm.29038

   Author contributions: V.A.V. was responsible for conception of the study, the design of the experimental framework, code implementation and data analysis. Pedro M. Gordaliza was responsible for code implementation and data analysis. All authors were responsible for writing, editing and reviewing the paper.

7. Joost Janssen, Covadonga M. Díaz-Caneja, Clara Alloza, Anouck Schippers, Lucía de Hoyos, Javier Santonja, **Pedro M. Gordaliza**, Elizabeth E.L. L Buimer, Neeltje E.M. van Haren, Wiepke Cahn, Celso Arango, René S. Kahn, Hilleke E. Hulshoff Pol, Hugo G. Schnack.

*Dissimilarity in sulcal width patterns in the cortex can be used to identify patients with schizophrenia with extreme deficits in cognitive performance*. Schizophrenia Bulletin, page 2020.02.04.932210, 2020. ISSN 0586-7614. doi: 10.1093/schbul/sbaa131

Author contributions: J.J was responsible for conception of the study, the design of the experimental framework and data analysis. Pedro M. Gordaliza was responsible for data analysis. All authors were responsible for writing, editing and reviewing the paper.

8. Joost Janssen, Clara Alloza, Covadonga M. Díaz-Caneja, Javier Santonja, Laura Pina-Camacho, **Pedro M. Gordaliza**, Alberto Fernández-Pena, Noemi Lois, Elizabeth E.L. Buimer, Neeltje E.M. Van Haren, Wiepke Cahn, Eduard Vieta, Josefina Castro-Fornieles, Miquel Bernardo, Celso Arango, René S. Kahn, Hilleke E. Hulshoff Pol, and Hugo G. Schnack. *Longitudinal allometry of sulcal morphology in health and schizophrenia*. Journal of Neuroscience, (In press), 2022. Preprint: https://www.biorxiv.org/content/10.1101/2021.03.17.435797v1

   Author contributions: J.J. was responsible for conception of the study, the design of the experimental framework and data analysis. Pedro M. Gordaliza was responsible for data analysis. All authors were responsible for writing, editing and reviewing the paper.

9. Alberto Fernández-Pena, Daniel Martín-Blas, Luis Marcos-Vidal, **Pedro M. Gordaliza**, Joost Janssen, Susanna Carmona, Manuel Desco, and Yasser Alemán-Gómez. *ABLE: Automated Brain Lines Extraction Based on Laplacian Surface Collapse*. Scientific Reports, (Accepted), 2022. Preprint: https://www.biorxiv.org/content/10.1101/2022.01.18.476370v1

   Author contributions: A.F.P. was responsible for conception of the study and the design of the experimental framework, software development and data analysis. All authors were responsible for writing, editing and reviewing the paper.

## Conference Proceedings:

1. **Pedro M. Gordaliza**, Juan José Vaquero, Sally Sharpe, Manuel Desco, and Arrate Munoz-Barrutia. *Towards an informational model for tuberculosis lesion discrimination on X-ray CT images*. In 15th International Symposium on Biomedical Imaging, Washington, DC, USA, 4-7 April, 2018. doi:10.1109/ISBI.2018.8363570

   Author contributions: Pedro M. Gordaliza was responsible for conception of the study and the design of the experimental framework, software development, data analysis, writing, editing, reviewing the paper and preparing the presentation. A.M.B and J.J.V contribute to the conception of the study and interpretation of the results. S.S. provided the CT data and contributed to the revision of the manuscript. All authors were responsible for writing, editing and reviewing the paper.

Contribution completely included in Chapter 3.

2. **Pedro M. Gordaliza**, Juan José Vaquero, Sally Sharpe, Manuel Desco, and Arrate Munoz-Barrutia. *Radiomics for the Discrimination of Tuberculosis Lesions*. In European Molecular Imaging Meeting, Donostia, Spain, 20-23 March, 2018. http://eventclass.org/contxt_-emim2018/

   Author contributions: Pedro M. Gordaliza was responsible for conception of the study and the design of the experimental framework, software development, data analysis, writing, editing, reviewing the paper and preparing the oral presentation. A.M.B and J.J.V contribute to the conception of the study and interpretation of the results. S.S. provided the CT data and contributed to the revision of the manuscript. All authors were responsible for writing, editing and reviewing the paper.

   Contribution completely included in Chapter 3.

3. **Pedro M. Gordaliza**, Juan José Vaquero, Sally Sharpe, Fergus Gleeson, and Arrate Muñoz-Barrutia. *A Multi-Task Self-Normalizing 3D-CNN to Infer Tuberculosis Radiological Manifestations*. In Medical Imaging with Deep Learning (MIDL), London, UK, 8-10 July, 2019. http://arxiv.org/abs/1907.12331

   Author contributions: Pedro M. Gordaliza was responsible for conception of the study and the design of the experimental framework, software development, data analysis, writing, editing, reviewing the paper and preparing the presentation. A.M.B and J.J.V contribute to the conception of the study and interpretation of the results. S.S. and F.G. provided the data and contributed to the revision of the manuscript. All authors were responsible for writing, editing and reviewing the paper.

   Contribution completely included in Chapter 4.

4. **Pedro M. Gordaliza**, Juan José Vaquero, Sally Sharpe, Fergus Gleeson, and Arrate Muñoz-Barrutia. *Tuberculosis lesions in CT images inferred using 3D-CNN and multi-task learning*. In 16th International Symposium on Biomedical Imaging, Venice, Italy, 8-11 April, 2019. doi: 10.1109/ISBI.2019.8759321

   Author contributions: Pedro M. Gordaliza was responsible for conception of the study and the design of the experimental framework, software development, data analysis, writing, editing, reviewing the paper and preparing the oral presentation. A.M.B perform the oral presentation and together with J.J.V contribute to the conception of the study and interpretation of the results. S.S. and F.G. provided the data and contributed to the revision of the manuscript. All authors were responsible for writing, editing and reviewing the paper.

   Contribution completely included in Chapter 4.

5. **Pedro M. Gordaliza**, Juan José Vaquero, and Arrate Munoz-Barrutia. *Translational Lung Imaging Analysis Through Disentangled Representations*. In Proceedings of Machine Learning Research (Under Review), pages 1–13, 12 2021. openreview.net/forum?id=16efiNAl$_{V}A$

   Author contributions: Pedro M. Gordaliza was responsible for conception of the study and the design of the experimental framework, software development, data analysis, writing, editing, reviewing the paper and preparing the presentation. A.M.B and J.J.V contribute to the conception of the study and interpretation of the results. All authors were responsible for writing, editing and reviewing the paper.

   Contribution completely included in Chapter 5.

6. **Pedro M. Gordaliza**, Verónica Aramendía-Vidaurreta, J.J. Vaquero, Gorka Bastarrika, María A. Fernández-Seara, and A. Muñoz-Barrutia. *Automatic Segmentation Of The Myocardium in Cardiac Arterial Spin Labelling Images Using a Deep Learning Model Facilitates Myocardial Blood Flow*. In 27th International Symposium Magnetic Resonance and Medicine (ISMRM), Montreal, 11-16 May, 2019. index.mirasmart.com/ISMRM2019/PDFfiles/4794.html

   Author contributions: Pedro M. Gordaliza was responsible for conception of the study and the design of the experimental framework, software development, data analysis, writing, editing, reviewing the paper and preparing the presentation. V.A.V. was responsible code implementation and data analysis. All authors were responsible for writing, editing and reviewing the paper.

7. Verónica Aramendía-Vidaurreta, **Pedro M. Gordaliza**, Rebeca Echeverria-Chasco, Gorka Bastarrika, A Muñoz-Barrutia, and María A. Fernández-Seara. *Groupwise Non Rigid Registration For Temporal Myocardial Arterial Spin Labeling Images*. In 27th International Symposium Magnetic Resonance and Medicine (ISMRM), Montreal, 11-16 May, 2019. index.mirasmart.com/ISMRM2019/PDFfiles/4484.html

   Author contributions: V.A.V. was responsible for conception of the study, the design of the experimental framework, code implementation and data analysis. Pedro M. Gordaliza was responsible for code implementation and data analysis. All authors were responsible for writing, editing and reviewing the paper.

8. Verónica Aramendía-Vidaurreta, **Pedro M. Gordaliza**, Marta Vidorreta, Rebeca Echeverria-Chasco, Gorka Bastarrika, Arrate Muñoz-Barrutia, and María A. Fernández-Seara. *Comparison of Myocardial Blood Flow Measurements with Arterial Spin Labeling in Breathhold and Synchronized Breathing Acquisitions*. In International Symposium Magnetic Resonance and Medicine (ISMRM), Virtual Conference, 8-14 Agust, 2020. archive.ismrm.org/2020/2208.html

   Author contributions: V.A.V. was responsible for conception of the study, the design of the experimental framework, code implementation and data analysis. Pedro M. Gordaliza was responsible for code implementation and data analysis. All authors were responsible for writing, editing and reviewing the paper.

9. Leandro A. Hidalgo-Torres, David Pérez-Benito, Rigoberto Chil, **Pedro M. Gordaliza**, and Juan José Vaquero. *Predicting 3D Photon Interaction in a Hexagonal Positron Emission Tomography Detector: A Deep Learning Approach.* Congreso Anual de la Sociedad Española de Ingeniería Biomédica (CASEIB), Virtual Conference, 25-27 November 2020, dialnet.unirioja.es/articulo

   Author contributions: L.H.T was responsible for conception of the study and the design of the experimental framework, software development and data analysis. All authors were responsible for writing, editing and reviewing the paper.

## 1870

Hay una pereza activa
que mientras descansa piensa,
que calla porque se vence,
que duerme pero que sueña.

Es como un leve reflejo
de la majestad suprema,
que eternamente tranquila,
sobre el universo reina.

¡Oh asilo del pensamiento
errante, dulce pereza;
mil veces feliz el hombre
que de ti goza en la tierra

**La pereza. Augusto Ferrán**

# Table of contents

# List of figures

# List of tables

# Nomenclature

## Roman Symbols

| | |
|---|---|
| $A$ | advection function |
| $F$ | complex function |
| f | Function $f$ |
| $I$ | Intensity function |
| $v$ | voxel |
| m | population sample average |
| $N$ | Number of observations |
| $P$ | Probability distribution |
| $P$ | propagation function |
| $p$ | Probability density function of $P$ |
| $r_a$, $r_e$ | radius |
| R | Risk function |
| $S$ | Sigmoid function |
| s | population sample standard deviation |
| $x$ | Draw from a random variable $X$ |
| $y$ | Draw from a random variable $Y$ |
| $X,Y,Z$ | Random Variable |

| | |
|---|---|
| $Z$ | spatial curvature modification function |

**Greek Symbols**

| | |
|---|---|
| $\alpha, \beta, \gamma$ | factor |
| $\Pi$ | prior probability |
| $\Delta$ | Change / Offset |
| $\nabla$ | gradient |
| $\kappa$ | mean contour curvature |
| $\mu$ | population mean |
| $\varepsilon$ | noise |
| $\pi$ | $\simeq 3.14\ldots$ |
| $\prod$ | productory |
| $\Psi$ | contour level set |
| $\sigma$ | population standard deviation |
| $\sum$ | summatory |

**Superscripts**

| | |
|---|---|
| $j$ | superscript index |

**Subscripts**

| | |
|---|---|
| $i$ | subscript index |

**Other Symbols**

| | |
|---|---|
| $\lvert\cdot\rvert$ | absolute value |
| $\mathscr{B}$ | binomial distribution |
| $dx$ | Differential respect to $x$ |
| $\mathbb{E}$ | Expectation function |
| $\mathscr{F}$ | Functions Space |

| | |
|---|---|
| $\mathscr{L}$ | Loss Function |
| $\mathscr{N}$ | Normal distribution |
| $\frac{\partial f}{\partial t}$ | Partial derivative of $f$ with respect to time/step |
| $p(x)$ | Probability density function of $P_X$ at point $x$ |
| $P_X$ | Probability distribution of $X$ |
| $\mathscr{X},\mathscr{Y}$ | Metric Space |

**Acronyms / Abbreviations**

| | |
|---|---|
| AI | Artificial Intelligence |
| BN | Batch Normalization |
| CADe | Computer Aided Detection |
| CADx | Computer Aided Diagnosis |
| CE | Cross Entropy |
| CFU | Colony Forming Unit |
| CNN | Convolutional Neural Network |
| CT | Computed Tomography |
| CV | Cross Validation |
| CXR | Chest X-Ray |
| DAG | Direct Acyclic Graph |
| DL | Deep Learning |
| DNN | Deep Neural Network |
| DSC | Dice Similarity Coefficient |
| EM | Expectation Maximization |
| ERA4TB | European Accelerator of Tuberculosis Regime Project |
| ERM | Empirical Risk Minimization |

| | |
|---|---|
| FCL | Fully Connected Layer |
| FNE | False Negative Error |
| FNN | Feed-forward Neural Network |
| FPE | False Positive Error |
| FTIH | First Time In Humans |
| GAN | Generative Adversarial Network |
| GDL | Geometric Deep Learning |
| GGO | Ground Glass Opacity |
| GLCM | Grey Level Co-occurrence Matrix |
| GLRLM | Grey-Level Run Length Matrix |
| HDA | Hausdorff Distance Averaged |
| HD | Hausdorff Distance |
| HIL | Human-in-the-loop |
| HRCT | High-Resolution Computed Tomography |
| HU | Hounsfield Units |
| ICC | Intra-class Correlation Coefficient |
| ICM | Independent Causal Mechanism |
| IGRA | Interferon $\gamma$ Release Assays |
| i.i.d. | independent and identically distributed |
| KL | Kullback–Leibler divergence |
| LBP | Local Binary Patterns |
| LD | Linear Discriminant |
| LN | Lymphatic Nodule |
| LR | Linear Regression |

| | |
|---|---|
| LTBI | Latent Tuberculosis Infection |
| MALDI-MS | Matrix-Assisted Laser Desorption/Ionization - Mass Spectrometry |
| MDR | Multidrug-Resistant |
| ML | Machine Learning |
| MRI | Magnetic Resonance Imaging |
| *Mtb.* | Mycobacterium tuberculosis |
| NHP | Non-Human Primates |
| NN | Neural Network |
| OOD | Out-Of-Distribution |
| PDE | Partial Differential Equation |
| PD | Pharmacodynamic |
| PD | Presented Dose |
| PET | Positron Emission Tomography |
| PK | Pharmacokinetic |
| PLS | Pathological Lung Segmentation |
| PReLU | Parametric Rectified Linear Unit |
| RF | Random Forest |
| RMSE | Root Mean Square Error |
| ROI | Region Of Interest |
| RR | Rifampicin-Resistant |
| SELU | Scaled Exponential Linear Units |
| SGD | Sustainable Development Goals |
| SNN | Self-Normalizing Neural Network |
| SNR | Signal-to-Noise-Ratio |

| | |
|---|---|
| SOTA | State Of The Art |
| SRM | Statistical Region Merging |
| SSL | Self-Supervised Learning |
| SVM | Support Vector Machines |
| TB | Tuberculosis |
| WHO | World Health Organization |
| TST | Tuberculin Skin Test |
| US | Ultrasound Imaging |
| VAE | Variational Autoencoder |
| VD | Volume Dissimilarity |
| VOI | Volume of Interest |
| WP | Work Package |

# Abstract

Tuberculosis (TB) is an infectious disease caused by *Mycobacterium tuberculosis* (*Mtb.*) that produces pulmonary damage due to its airborne nature. This fact facilitates the disease fast-spreading, which, according to the *World Health Organization* (WHO), in 2021 caused 1.2 million deaths and 9.9 million new cases.

Traditionally, TB has been considered a binary disease (latent/active) due to the limited specificity of the traditional diagnostic tests. Such a simple model causes difficulties in the longitudinal assessment of pulmonary affectation needed for the development of novel drugs and to control the spread of the disease.

Fortunately, *X-Ray Computed Tomography* (CT) images enable capturing specific manifestations of TB that are undetectable using regular diagnostic tests, which suffer from limited specificity. In conventional workflows, expert radiologists inspect the CT images. However, this procedure is unfeasible to process the thousands of volume images belonging to the different TB animal models and humans required for a suitable (pre-)clinical trial.

To achieve suitable results, automatization of different image analysis processes is a must to quantify TB. It is also advisable to measure the uncertainty associated with this process and model causal relationships between the specific mechanisms that characterize each animal model and its level of damage. Thus, in this thesis, we introduce a set of novel methods based on the state of the art *Artificial Intelligence* (AI) and *Computer Vision* (CV).

Initially, we present an algorithm to assess *Pathological Lung Segmentation* (PLS) employing an unsupervised rule-based model which was traditionally considered a needed step before biomarker extraction. This procedure allows robust segmentation in a *Mtb.* infection model (*Dice Similarity Coefficient*, DSC, $94\% \pm 4\%$, *Hausdorff Distance*, HD, $8.64\,\mathrm{mm} \pm 7.36\,\mathrm{mm}$) of damaged lungs with lesions attached to the parenchyma and affected by respiratory movement artefacts.

Next, a Gaussian Mixture Model ruled by an *Expectation-Maximization* (EM) algorithm is employed to automatically quantify the burden of *Mtb.*using biomarkers extracted from the segmented CT images. This approach achieves a strong correlation ($R^2 \approx 0.8$) between our automatic method and manual extraction.

Consequently, Chapter 3 introduces a model to automate the identification of TB lesions and the characterization of disease progression. To this aim, the method employs the *Statistical Region Merging* algorithm to detect lesions subsequently characterized by texture features that feed a *Random Forest* (RF) estimator. The proposed procedure enables a selection of a simple but powerful model able to classify abnormal tissue.

The latest works base their methodology on *Deep Learning* (DL). Chapter 4 extends the classification of TB lesions. Namely, we introduce a computational model to infer TB manifestations present in each lung lobe of CT scans by employing the associated radiologist reports as ground truth. We do so instead of using the classical manually delimited segmentation masks. The model adjusts the three-dimensional architecture, *V-Net*, to a multi-task classification context in which loss function is weighted by homoscedastic uncertainty. Besides, the method employs *Self-Normalizing Neural Networks* (SNNs) for regularization. Our results are promising with a *Root Mean Square Error* of 1.14 in the number of nodules and $F_1$-scores above 0.85 for the most prevalent TB lesions (i.e., conglomerations, cavitations, consolidations, trees in bud) when considering the whole lung.

In Chapter 5, we present a DL model capable of extracting disentangled information from images of different animal models, as well as information of the mechanisms that generate the CT volumes. The method provides the segmentation mask of axial slices from three animal models of different species employing a single trained architecture. It also infers the level of TB damage and generates counterfactual images. So, with this methodology, we offer an alternative to promote generalization and explainable AI models.

To sum up, the thesis presents a collection of valuable tools to automate the quantification of pathological lungs and moreover extend the methodology to provide more explainable results which are vital for drug development purposes. Chapter 6 elaborates on these conclusions.

# Motivation and Objectives

Tuberculosis (TB) is an infectious disease caused by *Mycobacterium tuberculosis* (*Mtb.*) that produces pulmonary damage. TB causes around three thousand deaths per day around the globe; thousand more than the *Coronavirus* disease (COVID-19) at its uprising, a fact even more dramatic considering that such a terrible number maintains steady during the last decades. Thus, numbers in 2021 amount to 9.9 million new cases and 1.2 million deaths, according to the World Health Organization (WHO) [333].

As part of the efforts to control this devastating epidemic, proper modelling of TB as a continuous spectrum between latent and active stages is urgently needed [235]. To this aim, several projects are ongoing worldwide, being this thesis part of one of them, the European Accelerator of Tuberculosis Regime Project, ERA4TB [76].

This thesis presents a set of methods based on the *Artificial Intelligence* (AI) state of the art to enrich the ability of *x-ray Computed Tomography* (CT) images to depict specific manifestations of TB, contrarily, to regular diagnostic tests which suffer from limited specificity. By exposing such specific patterns, CT images allow proper modelling of TB, which is essential for disease prognosis and its longitudinal monitoring, and consequently at the development of more effective drugs.

The common practice is that experts meticulously examine CT images manually for various purposes. Namely, to delimit regions/volumes of interest (ROIs/VOIs) in the images, the lungs in this work context, to enable "post-hoc" studies. Also, to examine abnormal regions within the ROIs to establish whether these are characteristic disease manifestations. With this information, they can subsequently provide longitudinal descriptions to enable hypotheses about the mechanisms that rule the disease and its interactions with in-development drugs.

Nevertheless, this approach is highly time-consuming, prone to human errors and therefore, infeasible at large studies. Namely, at (pre-)clinical trials in which thousands of images belonging to different animal, disease burden and strains models are involved, as in the ERA4TB project.

Therefore, the development of new tools capable of automating expert tasks entails the fundamental goal of this work. It particular, we aim for the methods to capture extra information from CT images that capture new findings of disease inner-workings that could contribute to for TB eradication. To this aim, the following specific objectives are proposed in this thesis:

- Develop methods to automatically segment lungs damaged by Tuberculosis.
- Develop methods for the automatic quantification of TB burden to characterize the disease progression and response to therapy.
- Develop methods for the automatic detection and characterization of the main TB manifestations to assess the development of the disease and the effectivity of the new drugs.
- Perform automatic analysis of translational animal models, namely, the mammal models usually employed in clinical trials (i.e., mouse, macaque, human).
- Provide analysis tools able to yield valuable information about the disease physiopathology under speculative scenarios to leverage interdisciplinary experts hypothesis.

# Thesis Outline

The rest of the document comprehend the following chapters:

1. The Key Role of Artificial Intelligence in Tuberculosis Assessment, consists of a comprehensive review of essential background concepts. Thus, briefly introduces both: a) medico-social aspects of TB (e.g., the impact of Tuberculosis in the world population, the importance of radiological imaging in its eradication) and, b) concepts related to the Artificial Intelligence methodology (e.g., advantages and limitations of the frameworks that enable task automation, the appearance of biases in the algorithms due to data scarcity or dataset shifts) used in the different works presented in the document. Finally, both topics joint in the state-of-the-art portray for the automatic analysis of pathological lung images.

2. Lung Segmentation and Quantification with Rule-based Approach, presents an algorithm for Pathological Lung Segmentation for a macaque model of Tuberculosis and a Gaussian Mixture based method for quantification.

3. Radiomics for TB Manifestations Classification, introduces a Radiomics approach for Tuberculosis manifestations classification by extracting texture features and applying a Random Forest classifier.

4. Deep Learning for TB Manifestation Classification, presents the Deep Learning model yielding a multi-task architecture. It is a self-normalized set-up with weights computed from an estimation of uncertainty. It is design to identify lesion from complete three-dimensional Computed Tomography volumes.

5. Translational Lung Imaging Analysis Through Disentangled Representations, presents the model able to disentangle significant factors. We show its application on image synthesis and the automatic and robust delimitation of pathological lungs.

6. Conclusions and Prospective work

# Chapter 1

# The Key Role of Artificial Intelligence in Tuberculosis Assessment

The bulk of this work lies in the study and subsequent proposal of several solutions based on different branches of Artificial Intelligence (AI). Similar to the temporal evolution of AI itself, the first work presented in the manuscript (Chapter 2) initially applies, adapts and develops classical AI methods (based on coding human-defined rules). Then, it presents solutions employing some of the more common Machine Learning (ML) techniques developed in the last 20 years, which are based on statistical dependence within the study dataset but with few or minimal assumptions about the mechanisms causing such dependencies. Finally, it ends with more complex models that aim to add the predictive power and explainability to the previous ones.

These novel methodologies can be adapted to a wide range of problems and hopefully benefit a large part of the scientific community. However, this scientific work has a strong engineering and therefore translational character. Namely, it is oriented to a specific application such as the automatic evaluation of lungs damaged by Tuberculosis as verbalized in its title.

Thus, while in subsequent chapters the particular objective, the mathematical machinery, the performance analysis or the implementation of the different methodologies are presented in a self-contained manner[1], this opening chapter aims to justify the need for their development.

For this reason, initially (see Tuberculosis in numbers) a summary of the updated epidemiological data[2] of Tuberculosis (TB) worldwide is provided showing its pandemic status.

[1] Each chapter consists of adaptations of individual works shown in Published and submitted contents. These extend those sections that help to synthesize the entire manuscript.

[2] Based on 2020 WHO report. [332]

Next, the biomedical framework of TB diagnosis based on microbiological, pathological and immunological studies and their interactions with treatments is introduced. The main biological causes of the pandemic are exposed together with the different ways to dampen them by employing the strategies presented in the Section 1.2 (Eradicating Tuberculosis: The need for continuous assessment).
The role of the different medical imaging modalities depending on the disease markers critical at the several framework components presented in the previous section are introduced subsequently in Section 1.3. Finally, the main approaches existing in the literature to carry out the objective of quantifying damaged lungs, obviously with the focus on those devoted to Tuberculosis damage, and which have served as inspiration and support for the development of the different methodologies introduced in the subsequent chapters, are presented in Section 1.4. Since the approaches rely on the aforementioned AI principles, Section 1.4 is divided in two, namely, a first part which describes AI Learning Principles with Emphasis in Lung Analysis and a second with the specific application's summary.

## 1.1 Tuberculosis in numbers

The data shown below is not intended to simplify the harm caused by Tuberculosis throughout the history of humanity and especially nowadays. The mere expectation is to show the unaware reader the reality about the pandemic that causes the most deaths daily at the time of this writing. We would like also to alert them to the need to tackle it globally, in a similar way to how unfortunately and abruptly the society had done since the outbreak of COVID-19; especially given the neglected status of Tuberculosis.

According to the World Health Organization (WHO) estimations [333], in 2021, there were 9.9 million (range, 8.9–11.0 million) million incident cases and 1.3 million (range, 1.2–1.4 million) deaths caused by tuberculosis (TB). The incidence and mortality rates per 100.000 population per year and country are within the upper part of Fig.1.1 (maps *a* and *b*). More strikingly, latent TB (see Section From a binary perspective to continuous spectrum of diagnosis) is present in about a quarter of the world's population. Within this infected population, *Mycobacterium tuberculosis* (*Mtb.*), the causative agent of TB, becomes active in 10% of the cases and mainly damages the lungs owing to its airborne nature.

As can be seen in Fig.1.1, the TB burden distribution is far from homogeneous around the globe, being low and middle-income countries much more hit by the pandemic; mainly due to the well-known relationship between undernutrition and a depressed immunological system [228, 263]; Fig.1.2 shows this trend clearly. Meanwhile, rich countries understood

TB as a disease of the past, turning it into a neglected disease associated with poverty, marginalization and social exclusion of individuals suffering from it.

This fact does not benefit the fight against the pandemic at all. However, a terrible game-changer has become crucial in the last 25 years. Resistance to TB drugs [19, 86, 67, 341, 355] has increased markedly during this time, especially in those westernized countries that considered the disease eradicated. The data corresponding to new cases in 2020 with MDR-TB (Multidrug-Resistant Tuberculosis) and RR-TB (Rifampicin-Resistant Tuberculosis), as well as the percentage of cases that had already been treated for TB and developed MDR and RR are shown by country in maps *C* and *D* of Fig.1.1. This situation has shifted the projections for TB eradication from around 2030 [19, 341] to 2050 [129].

Thus, it is evident that in the current context of globalization, TB has become a global concern. Due to undesirable reasons and because they are directly involved in the problem, the wealthy countries have finally been forced to intervene much more active in recent years, not least the United Nations (UN) has included the WHO "End TB Strategy" among its SGDs (Sustainable Development Goals) for the period 2015-2035. This strategy is deployed through several multisectoral projects with different objectives, some of which aimed at improving the aforementioned social risk factors, while others, such as the work carried out in this project framework (see Project Framework: ERA4TB), focus on developing new and more effective treatments for TB (e.g., new drugs, regimes, vaccines), as shown in the following sections.

## 1.2 Eradicating Tuberculosis: The need for continuous assessment

In 1882, Robert Koch discovered that TB was caused by the *Mycobacterium tuberculosis* complex, while the symptoms were already well known since earlier dates. Concretely, molecular evidence of ancient TB-related clades has been found in Egyptian mummies ( 1550 – 1080 BC) [222] and recent studies based on the molecular clock of *Mtb.* [207] confirm the hypothesis that all *Mtb.* lineages go back around 2500 years, while the Most Common Recent Ancestors (MRCAs) of the *Mtb.*, such as *Mycobacterium bovis*, *Mycobacterium pinnipedii* or *Mycobacterium canettii* are dated back 11.000 years.

While the antiquity of the disease may be surprising to less familiar readers, the literature point to the fact that the emergence of new clades is not particularly prolific for *Mtb.* when comparing to other diseases of bacteriological or virological origin (*Mtb.* mutation rate is estimated around $1x10^{-8}$ and $5x10^{-}7$ nucleotide changes per site per year) [207]. Thus,

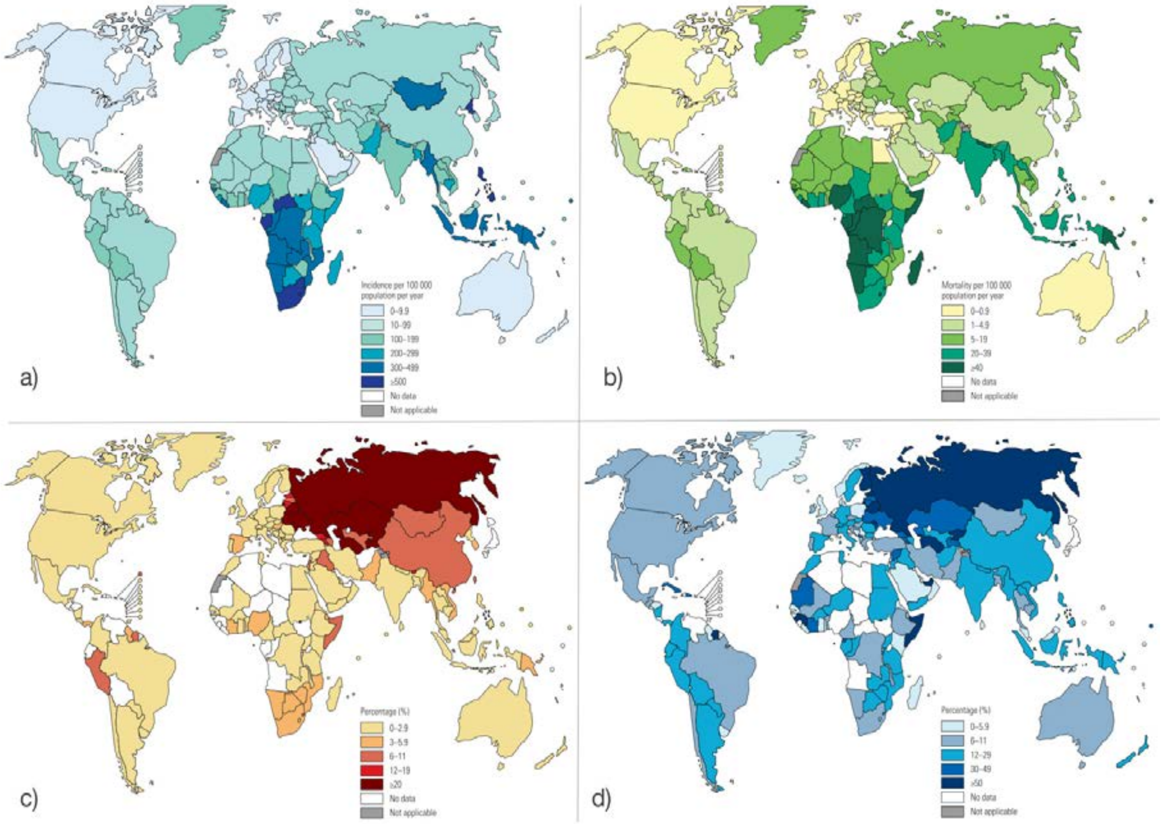


Fig. 1.1 Tuberculosis (TB) epidemiology data per country: **a)** TB prevalence per 100000 population per year; **b)** TB mortality per 100000 population per year; **c)** Multidrug-resistant TB (MTB-TB) percentage; **d)** Rifampicin-Resistant TB (RR-TB). Extracted from the WHO (World Health Organization) TB report 2020 [332].

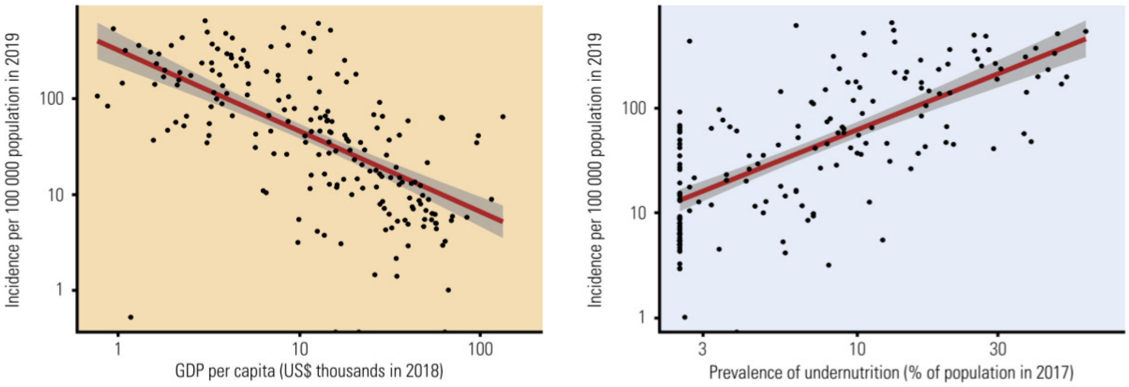


Fig. 1.2 Social cofounders of TB. **Left)** Relationship between TB prevalence and Gross Domestic Product (GDP); **Right)** Relationship between TB prevalence and undernutrition. Extracted from the WHO TB report 2020 [332].

producing efficient drugs in a reasonable time before the appearance of new clades is feasible in this context. However, the new infections and reinfections numbers show that the design of previously used drugs has been insufficient, mainly due to the lack of knowledge of the molecular events of *Mtb.* itself and its interaction with social risk factors which has led to the recent emergence of new, more adapted and therefore resistant clades and lineages of the disease [104, 225].

Partly due to the urgency of the situation given by TB pandemic status and to technical limitations, drug design has been mainly based on a binary interpretation of TB (e.g., active or non-active, infected or non-infected), complicating the understanding of the molecular mechanisms causing the clinical stages traditionally described for TB (see Fig.1.3). As shown in the next section, this fact highlights the need for a paradigm shift in assessing the disease.

### 1.2.1 From a binary perspective to continuous spectrum of diagnosis

Depending on multiple factors (i.e., the viral load, the strain of *Mtb.* complex, the attacked immune system condition after transmission), bacteria can either be eliminated from the organism through an innate or adaptive immune response (T-cell mechanisms) or remain in a latent (LTBI, Latent TB Infection) or active state [19, 153, 234].

LTBI is not transmissible, and infected subjects present no symptoms. Active TB patients suffer from persistent cough, fever, weight loss, haemoptysis, among other maladies, and they can be transmitters of the bacteria. Due to this fact, from the traditional clinical and public health point of view, TB is understood as a binary disease [217, 235]. This binary conceptualization is reflected in the most popular tests for disease assessment, which lack the specificity and sensitivity to provide a correct non-dual result. Illustratively, the main tests are shown below, divided according to whether they are used to find the presence of LTBI or for active TB.

#### Tests to detect LTBI

**Tuberculin Skin Test (TST):** Also known as Mantoux test or Mendel–Mantoux test or Purified Protein Derivative (PPD) [135]. In this test, a protein of *Mtb.* is injected intradermically, usually on the left forearm. The test is read around 72 hours later (48-96 hours). The evaluation consists of measuring the local inflammation caused. If there is no inflammation, the subject diagnostic is negative. In the event of inflammation, the diagnosis is given from the diameter of the inflammation and the patient's risk factors. It is important to note that the test can be positive both when the patient has acquired the corresponding antibodies and has previously eliminated the bacteria (adaptive immune response) and when the TB persists as latent.

The subjectivity test reading leads to false positives and false negatives, making complementary diagnostic tests such as X-rays or a second binary test such as Interferon-$\gamma$ needed [17].

**Interferon-$\gamma$ Release Assays (IGRA):** Employed as an alternative to TST [69]. The test involves exposing a blood sample from a subject to *Mtb.* antigens and measuring the amount of interferon released by T-lymphocytes. As TST, positive results do not directly link to the persistence of latent infection [4]. Those subjects who have overcome the infection and maintain T-cells from the adaptive immune response will also release interferon.

### Tests to detect Active TB

**Sputum smear:** The traditional test analyze the sputum using a conventional microscope to locate the bacteria. The minimum sensitivity is over the threshold of 5000 bacteria per millimeter [61, 217]. Thus, it is impossible to detect active TB at early stages. Besides, the test is not very specific since *Mtb.* appears the same as *non-mycobacterium tuberculosis*, which is added once again to the lack of anatomic location of the TB burden.

**Culture:** The culture substantially improves the sensitivity of the sputum smear, finding the *Mtb.* from 10 mycobacteria per millimetre [61, 125]. In return, with the best and most expensive culture methods, the results are obtained in two weeks, although generally, the duration lasts up to six. The test cannot also characterize the longitudinal evolution of the sources of infection beyond deciding whether they are still active or not.

Besides the mentioned particular features, it is important to remark the tests relatively low cost. However, they cannot provide the continuous longitudinal characterization of the disease (from latent to active) that arises in modern literature [12, 77, 153, 235]. This longitudinal characterization is fundamental to understand the mechanisms of the immunological life cycle of the disease to facilitate the development of new drugs and the characterization of the resistance to them. Namely, it helps to understand under what circumstances do a latent infection reactivate or what markers are significant in remission [19, 235]. This is especially important in the current global scenario in which therapies pretend to shorten their duration to 4 months or less [339].

The Fig. 1.3 illustrates this idea showing TB as a continuous spectrum [235, 341] divided into six main phases (infection, immune response, latent infection, reactivation, active infection and transmission). Such an approach is usual in clinical practice and essential during clinical assays. Subject identification in this scale depends on the symptoms but more on the infection burden for which the tests act as proxies. However, the traditional tests lack enough specificity/sensitivity to characterize the whole spectrum undermining the drug discovery process.

For obvious ethical reasons, the development of new treatments avoids to perform tests in humans until its very last stage. So, the hypothesized molecular mechanisms fighting the disease at each phase are extrapolated from different animal models (see Section 1.2.2 for further details) [339].

When employing binary or categorical tests, the relationship between the drug mechanisms and the outcome could be due to confounded correlations between pairs of outputs (i.e., marker $M$ is employed as the output of treatment $T$ for two animal models $X$ and $Y$. However, output $M$ could be due to the specific T-cell mechanism in model $X$ while in $Y$ is just a product of the risk factors). Causative mechanisms could more easily remain unknown under the binary scenario [120, 186, 318]. So, attributing an effect to a specific molecule in a new drug or to some variation in the composition of a vaccine, would be impossible or very expensive. Contrarily, when continuous TB characterization models are employed, the model translation results less uncertain.

Different techniques indirectly allow for the needed continuous assessment of TB. Blood tests and some culture techniques such as those mentioned above allow obtaining the Colony Forming Units (CFUs) from a sample [19, 235]. CFUs represent the infectious burden of a subject but present some limitations: a) In cases of LTBI, the sensitivity is quite low or nonexistent, b) samples must be taken from several anatomical regions because c) the measurement is not a unilateral marker and lacks anatomical location. For example, the same number of units can represent a subject with a large localized infection or a pattern of a small source of infections in the different pulmonary lobes and even extrapulmonary. Another alternative is given by the extraction of the lungs (or the region of interest) for their subsequent dissection and complete evaluation. Even if the method is very exhaustive and allows to characterize the different manifestations of the disease in relation to their anatomical location. It is obvious that: a) the longitudinal evaluation of an excised subject is infeasible; b) it is costly; c) it implies bacteriological risks and d) it does not allow direct comparison with the disease in humans. The limitations of the different methods highlight the need to use techniques that: 1) leverage a longitudinal analysis of the subjects; 2) provide sufficient specificity and sensitivity for the detection and characterization of the different TB

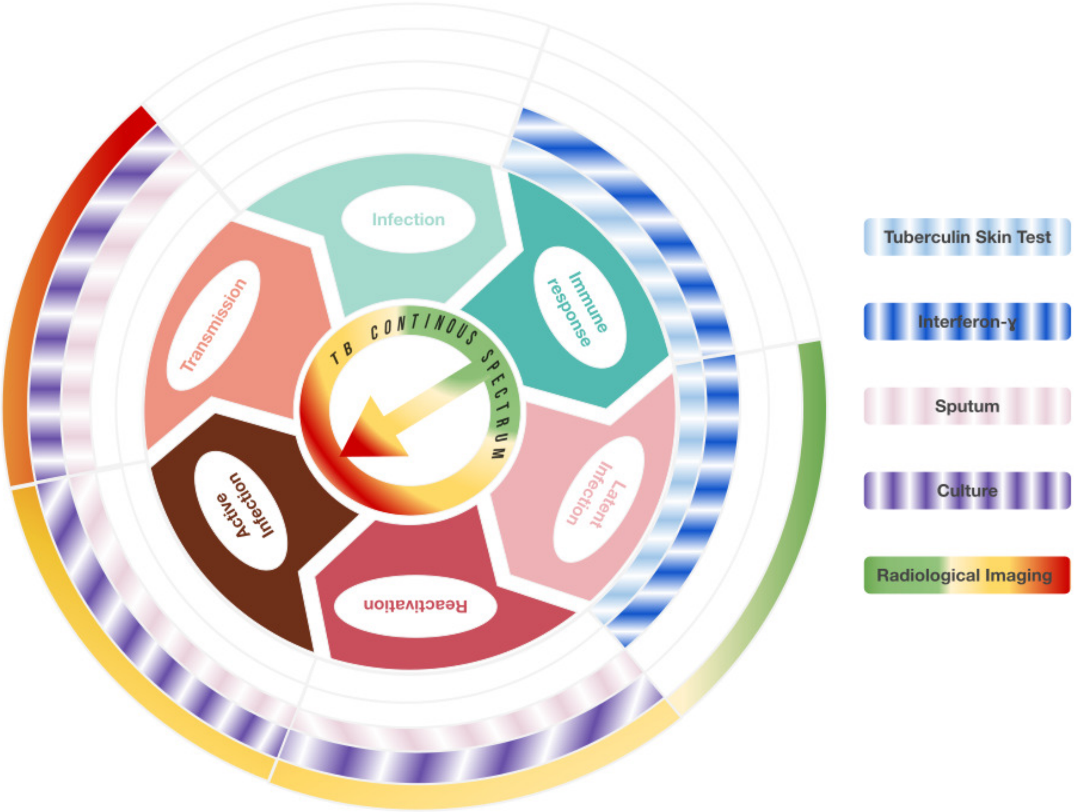


Fig. 1.3 Life cycle of *Mtb.* and main tests to characterise the entire disease spectrum. The inner cycle names the traditional categorical clinical stages of the continuous spectrum of TB immunological life cycle. Each outer circle represent each TB assessment tests capability. Blank spaces for lack of sensibility, bicolour ones represent the binary character of the test, while gradient representation represents the ability to provide a continuous value.

markers and 3) be translational between the different animal models.

Without forgetting that active TB disease requires a microbiological diagnosis, in this context, the use of *in vivo* medical images is an almost perfect fit [48, 217]. Different imaging modalities allow the longitudinal study of the characteristic radiological manifestations. Longitudinal follow ups provide the required continuous character to the evaluation. The screening helps to leverage the most accurate TB infection status that is essential to accomplish the objective of eradicating TB by 2035[3] [19, 341]. Since each medical image technique present advantages and disadvantages and given their relevance to this work, an introduction to the modalities employed for TB assessment are presented in a subsequent section, Medical Imaging for Tuberculosis Assessment.

[3] 2050 attending to the new models [129]

### 1.2.2 Project Framework: ERA4TB

The previous sections clearly show how in addition to evolution of the social factors, the eradication of TB involves characterizing the mechanisms of bacterium propagation and its interaction with different drugs longitudinally for the continuous improvement of the compounds. This process is enabled by clinical trials that study infected subjects using "in-vivo" imaging. The process requires a highly interdisciplinary environment with diverse scientific profiles ranging from chemists to engineers, biologists and physicians. Namely, chemist develop new drugs. Engineers work on automatic extraction of imaging biomarkers. Biologists and physicians come to an understanding of the biological processes involved. Thus, given the magnitude of the task, it is organized into large projects, especially after the inclusion of *"End TB Strategy"* among the SGDs. Specifically, the work presented in this thesis has been developed within the framework of the European Accelerator of Tuberculosis Regime Project (ERA4TB) project. ERA4TB is *"a public-private initiative devoted to accelerate the development of new treatment regimens for tuberculosis"* through a *"platform based on a progression pipeline that can cater for a variety of molecules at different stages of development"* [76]. To this aim, ERA4TB is divided into modules or Work Packages (WPs) briefly summarised below, which articulate the needed interfaces during trials, being this thesis subject framed under WP4:

- **WP1, Data and Pipeline Management:** The activities in WP1 support the development and implementation of a data management (Drug Development Information Management (DDIM)) platform supporting project efforts. A Graphical User Interface (GUI) as a portal to the clinical and preclinical data and an image storage repository. Through the platform, data will be curated, standardized and accessible to researchers. Besides, the platform will provide a plug-and-play infrastructure to employ the software analysis implemented, as presented in this thesis.

- **WP2, *In Vitro* Profiling:** As the first stage for the characterization of the interactions between the drug and the bacteria at cell-level, WP2 aims to provide *in vitro* profiling capacity needed for both: (1) the preclinical profiling of single drugs and (2) the knowledge generation pathway of preclinical combos.

- **WP3, *In Vivo* Profiling:** After *in vitro* validation through the methodology instilled in the platform, WP3 will investigate the efficacy of the identified compounds, alone and in combination, in experimental animal models, focusing initially on relevant mouse models mimicking TB pathogenesis in humans. Subsequently, promising regimens will move to Non-Human Primate (NHP) models for evaluation.

The works presented in this manuscript focus on CT volumes analysis of NHP and mice models to detect similar imaging biomarkers as those described in Medical Imaging for Tuberculosis Assessment. While the mice model is worth it for the initial characterization of the compound effects, it does not completely recapitulate the full range of characteristics of the pulmonary pathology in humans [339]. Experts cannot visually distinguish between different lesions in the mice model, so they cannot inject the knowledge into the automation systems. Consequently, the drug evaluation is approximated by analyzing the lungs as a whole.

Fine evaluation requires NHP models that have been proven to recapitulate relevant clinical characteristics of the human disease. This is due to the high level of gene homology, which underlies anatomical, physiological and immunological similarities [154, 245, 272]. These similarities lead to the development of comparable disease pathology, clinical signs and immune features following *Mtb.* infection. Animal models are fundamental for developing novel treatments, as they provide a platform in which the efficacy of new interventions can be evaluated against infectious challenges. Longitudinal images of the TB macaque model can be acquired from live animals using medical imaging systems [63, 180, 273] – e.g., chest radiographs (CXR), computed tomography (CT) and position emission tomography (PET) – and employed to visualize the evolution of pulmonary disease.

- **WP4, Imaging:** Imaging technologies are instrumental, enabling translational tools for drug development. Different modalities are employed to characterize the disease evolution from single cells to tissue specimens and in vivo subjects; (1) Single-cell imaging using microfluidic systems will be used to quantify responses to dynamic exposure to single molecules and drug combinations and to determine the PK (pharmacokinetic) driver; (2) MALDI-MS (Matrix-Assisted Laser Desorption/Ionization - Mass Spectrometry) imaging in infected tissues will provide quantitative information about drug penetration and distribution at the site of action in different types of TB lesions; (3) PET/CT on infected mice and NHP will provide non-invasive PK/PD (pharmacodynamic) assessments that will be integrated into response prediction models, in which imaging biomarkers will be incorporated, in close cooperation with WP5.

- **WP5, Modelling and Simulation:** WP5 aims to ensure effective translation and extrapolation of experimental findings into clear criteria for selecting candidate molecules for combination therapy. WP5 includes world-leading partners with expertise in mathematical and statistical modelling and simulation in the field of PK/PD of anti-infective drugs, ensuring data integration and translation from WPs 1–4 and 6 with the ultimate goal of ranking suitable candidate compounds for progression into clinical trials. These actions facilitate the implementation of an appropriate database /data management system within WP1 and an efficient workflow for a rational dose selection of single compounds (monotherapy) and combination therapy to be evaluated in Phase I studies.

- **WP6, Preclinical Development:** In charge of developing suitable synthetic route for manufacturing each molecule, and its formulation into Drug Products for later use in WP7 to study their safety and to evaluate their PK in humans ensuring their rapid transfer to in-patien trials (Phase II readiness).

- **WP7, Phase I. First Time In Humans (FTIH):** For each molecule entering Phase I, WP7 conducts FTIH studies and other Phase I trials on healthy volunteers as needed (i.e., dose-ranging, single ascending dose, multiple ascending doses, combination regimen) for the Phase II dossier. Other Phase I trials (i.e., drug-drug and drug-food interaction studies) could be carried out if prioritised by the development team. FTIH and other Phase I trials will be designed and developed following EMA guidelines and to the highest scientific, quality and ethical standards. The final study protocol design to be implemented for each molecule will consider the recommendations arising from the integrated PK/PD modelling and simulation performed by WP5.

- **WP8, Management, Outreach and Sustainability:** This WP provides scientific guidance and professional project management for solving trade-offs between scope, time, quality and cost to ensure adequate progress and successful project completion. Additionally, this WP aims at developing outreach and sustainability strategies for the long-term maintenance of the ERA4TB platform.

- **WP9, Ethics and Data Privacy:** This WP aims to define and follow up all the Ethical and Data Privacy implications to be considered throughout the project in a cross-sectional way. It also ensures that research integrity is followed across the project. It also takes care that all its activities fulfil ethical and regulatory requirements for preclinical experimentation and clinical trials under the applicable local and EU regulation on ethics and data privacy. A dedicated chapter in the Consortium Agreement ensures that Ethics and Data Privacy is be a matter for discussion included in all the routine meetings and decisions of ERA4TB's governing bodies.

## 1.3 Medical Imaging for Tuberculosis Assessment

The main *in vivo* imaging modalities employed to provide the needed TB longitudinal and continuous assessment under different scenarios are listed below. Their application is limited by the quality of the lung images acquired (i.e., SNR, resolution, artefacts). In the most aggressive cases, TB can attack the majority of the organs (extrapulmonary tuberculosis) [262, 282, 325]. However, the lung parenchyma is often the most and main damaged region. Therefore, most studies attempt to characterize the disease progression (or remission) based on their hypothesis about the evolution of the TB manifestations within the respiratory system.

**Ultrasound (US):** It is usually used in paediatrics or in the extended clinical environment where patients cannot remain immobile during the time that image acquisition with scanners requires. Ultrasound sensitivity is enough for the detection of Pleural Effusion (see Section 1.3) and extrapulmonary manifestations (i.e., hepatosplenomegaly and abdominal lymphadenopathy) but lacks the power to show for other main findings (i.e, differ among granulomas, conglomerations or milary TB, ground glass opacities, cavities walls) [88, 121].

**Magnetic Resonance Imaging (MRI):** It is prescribed for extrapulmonary tuberculosis [325, 341]. In addition, due to its non-ionizing nature, it is usually performed for pregnant women and very young patients. However, the air-like structure of the lung parenchyma yields a low (received) signal from the tissue, which translates into a poor image contrast. Still, it is beneficial in lymph nodes study, abnormalities in the pleura and caseation (see Section 1.3) [262, 283].

**Chest X-Ray (CXR):** Together with the previously mentioned sputum tests and cultures (see Section 1.2.1) form the triumvirate for the initial diagnosis of the disease [24, 217]. Although a wide variety of disease manifestations can be detected and sometimes localized by CXR images [317], the modality does not provide enough information to accurately measure them. So, the longitudinal follow up of the disease using CXR is difficult as the images just provide a planar projection of a 3D volume and are contaminated by noise. Therefore, CXRs just provide a quick, cheap and rough assessment.

**Computed Tomography (CT):** CT allows the acquisition of volumetric images (see Fig.1.4), providing a more reliable representation of the different tissues, unlike the CXR. Current high-resolution CTs (HRCT) leverage the characterization of structures at

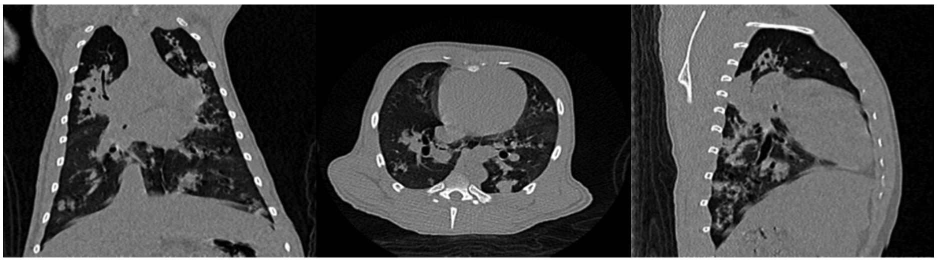


Fig. 1.4 The abdomen of a macaque infected with TB: a) Coronal view; b) Axial view; c) Sagittal view.

submillimetre resolution [84]. This allows the detection and quantification of manifestations that remain hidden under other modalities; for example, Ground Glass Opacities (GCO), miliary and conglomerated nodules, necrosis in Lymphatic Nodules (LNs) or other parenchymal lesions.

**Positron Emission Tomography-CT (PET-CT):** PET addition, normally fluorodeoxyglucose - PET (FDG-PET), to the CT scanner adds to the virtues of the second the possibility of detecting inflammation and infection through the use of Standardized Uptake Values (SUVs). Although the use of this technique in TB is currently under consideration [215, 231], mainly due to the lack of a specific PET radiotracer for TB, the results to date suggest that PET could be of great help for better measurement of the *Mtb.* activity since the additional insights suggest that some manifestations as calcification are not static [232, 314, 339]. This recent discovery could be a breakthrough for PK/PD modelling which would be fundamental to characterize the response to drugs administration [232, 314].

Under the listed principles, it is clear that CXR and especially CT, enable the study of the disease in the most detailed way from its macroscopic lung manifestations (shape, size, texture, localization, rate of change) to obtain suitable biomarkers which able to provide a complete spectrum for TB. Fig.1.5 illustrates this idea. Each manifestation is positioned at the interval of the TB immunological cycle where they typically appear. The detection and quantification power of CXR and CT turn up by the colours of their inner and outer circumferences. A brief description of these radiological manifestations is provided below [38, 217]. Similarly to a clinical radiological description, the standard units of CT images, *Hounsfield Units* (HUs) [165] are referred to describe contrast differences within the TB findings and the rest of the tissues.

- Granuloma or tuberculoma: They are the most characteristic lesions produced by TB and the essential biomarker during the latent stage. Granulomas are spherical and

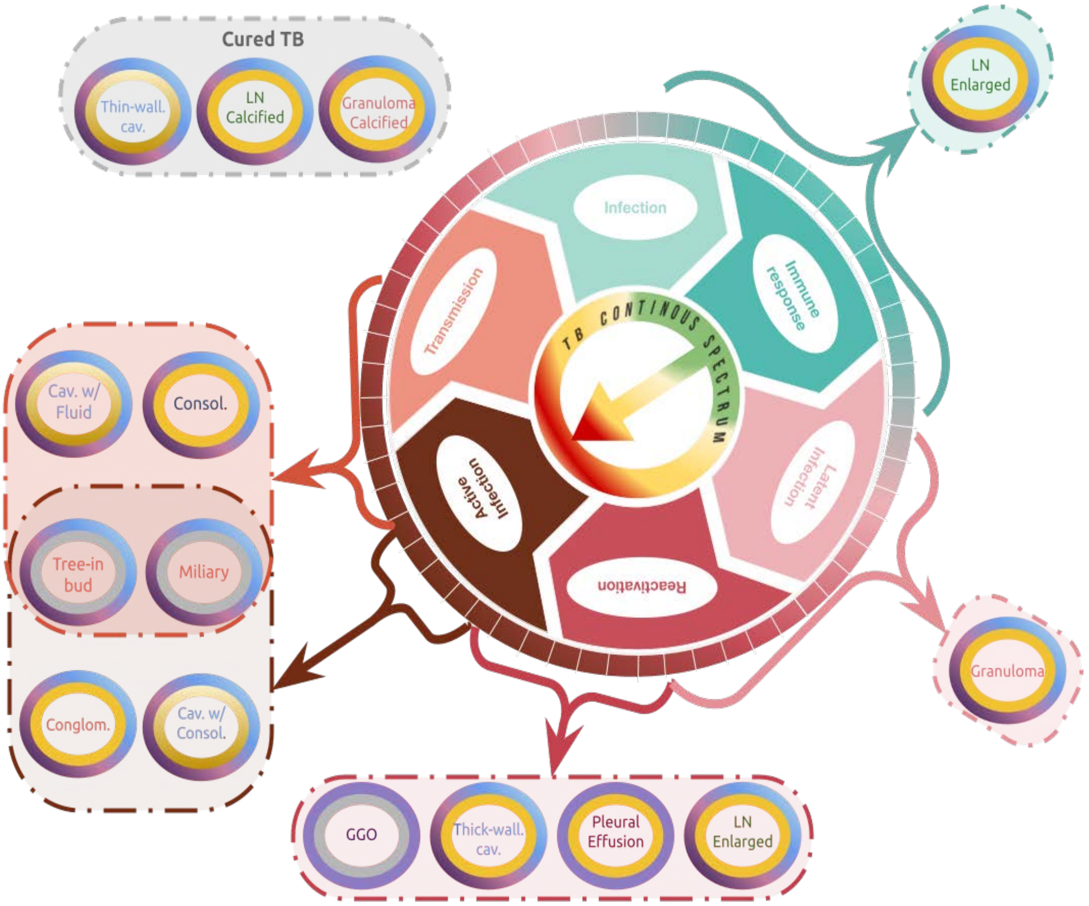


Fig. 1.5 Assessment power for Computed Tomography (CT) and Chest X-Ray (CXR) at main TB manifestations through the continuous epidemiological cycle of TB. Main TB manifestations, namely, Lymph Nodes (LN): Enlarged and Calcified, Pleural Effusion, Ground Glass Opacities (GGO), Consolidation, Cavities: Thin-walled Cavity, Thick-walled Cavity, Cavity with Consolidation and Cavity with Fluid and Nodules: Granuloma, Calcified Granuloma, Conglomeration, Tree-in-bud and Miliary nodules, are displayed across the TB spectrum range where they are usually most common. CT and CXR scan quantification power per manifestation is showed by the outer and inner toroids surrounding them respectively; purplish for CT and yellowish for CXR. Gradient colouring toroids yield for detection, localization and size measurement capacity, solid colour for detection and localization and solid grey for lack of sensibility of the imaging modality.

present high homogeneous values on the HU scale. Due to their similar intensity on this scale to blood vessels and mediastinal tissue, it is easy to classify them incorrectly (Fig.1.6a).

- Nodule Conglomeration: Adhesion of granulomas usually occurs when the disease is advanced and not treated. Its structure can be seen in Fig.1.6a. Alternatively, tuberculomas could appear as randomly distributed micronodules, named miliary TB[4].

- Tree-in-bud: It appears when the infection begins to occupy air-like structures (i.e., alveoli, airways) during medium and high active stages. For this reason, they are identified by observing opacities (very bright areas) within the structures, as can be seen in Fig.1.6b.

- Infiltrate or consolidation: Opacification of air spaces within the lung parenchyma. The consolidation may be dense and may have irregular, poorly defined, or hazy margins (see Fig.1.6c).

- Cavity: They are air-filled areas inside granulomas. It originates at the beginning of the disease in immunosuppressed subjects, in whom the defence mechanisms cannot contain the infection. First, a proliferative lesion is formed. It has a reactive inflammatory component around the infectious focus that tends to evolve to necrosis in the central part. Necrosis is called caseosis because of the whitish appearance reminiscent of cheese. Cavities are formed when the foci of caseosis are emptied (see Fig.1.6e). In the most active periods of the disease, the cavity can fill with liquid.

- Ground Glass Opacity (GGO): Refers to an area of greatest attenuation in the lung with bronchial and vascular damage (see Fig.1.6d).

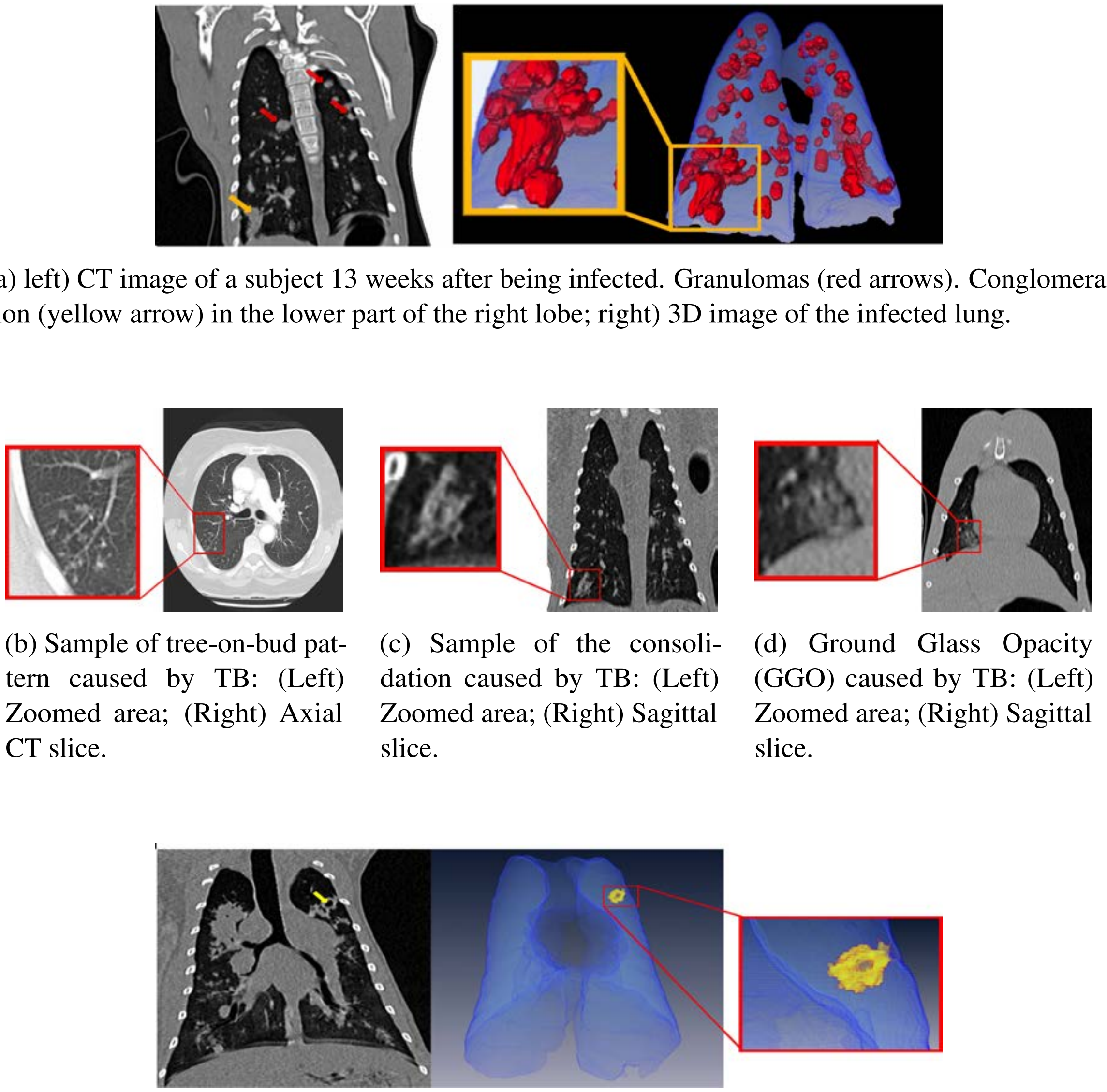
The CT capacity to represent specific radiological TB manifestations (as shown in the Fig. 1.6) makes this modality a workhorse for damaged lung assessment [84, 217].

However, volumetric CT image evaluation is complex. The CT scanners high level of detail comes at the cost of generating large images. Traditionally, the CT volume assessment is done manually by experts radiologists who face a tedious and time-consuming task, prone to errors and with a wide intra- and inter-expert variability being their interpretation subjective [313] (*"With great power comes great responsibility"*[5]).

Aforementioned, the development of new drugs requires massive clinical trials, including different animal models with their corresponding needed lung image analysis, which is infeasible manually.

[4] The Miliary term refers to the random distribution of small nodules which is not exclusive of TB disease

[5] Spider-Man phrase widely attributed to the character Uncle Ben. *Amazing Fantasy. vol.15*, 1962.



(a) left) CT image of a subject 13 weeks after being infected. Granulomas (red arrows). Conglomeration (yellow arrow) in the lower part of the right lobe; right) 3D image of the infected lung.

(b) Sample of tree-on-bud pattern caused by TB: (Left) Zoomed area; (Right) Axial CT slice.

(c) Sample of the consolidation caused by TB: (Left) Zoomed area; (Right) Sagittal slice.

(d) Ground Glass Opacity (GGO) caused by TB: (Left) Zoomed area; (Right) Sagittal slice.

(e) Sample of the cavity caused by TB: (Left) Sagittal slice. The cavity is pointed with a yellow arrow; (Middle) 3D visualization of the segmented lung (blue) and the cavity (yellow); (Right) Zoom showing the segmented cavity in yellow.

Fig. 1.6 Radiological Tuberculosis Manifestations

In this way, the need to develop automatic tools for lung damage assessment, such as those presented in the following chapters, is obvious.

The following section contextualizes the methodological environment which encapsulates such tools, introducing the main AI methodology principles from medical imaging and the most representative approaches.

## 1.4 Computer Aided Diagnosis: The way to automated quantification

Numerous and varied approaches that use Computer-Aided Diagnosis/Computer-Aided Detection (CADx/CADe) for medical imaging analysis can be found in the literature, obviously twinned with the advances and trends of AI methods. Indeed, the papers presented in this thesis redirect AI general-purpose principles to the specific problem of pathological lung CT image analysis, relying on works with a similar goal developed over the last three decades. Therefore, this section presents the relevant AI concepts in different applications that act as scaffolding for the more relevant literature subsequently introduced. The literature described is extended in each particular chapter on specific goals. In this manner, each chapter maintains its self-contained organization.

### 1.4.1 AI Learning Principles with Emphasis in Lung Analysis

From a general perspective, the particular lung imaging analysis question can be treated as an inverse problem[6] that encompasses it within a mathematical framework that provides answers through different AI approaches (i.e., rule-based models, statistical learning, machine learning, causal learning). The solutions provided intend to mathematically skeletonize a physical system from a limited set of $N$ observations:

$$(x_1,y_1),(x_2,y_2)\dots(x_i,y_i)\dots(x_N,y_N), \tag{1.1}$$

where $x_i \in \mathscr{X}$ and $y_i \in \mathscr{Y}$ are inputs and outputs draws of the system. In the most frequent case within this thesis, $x_i$ are CT images and $y_i$ the corresponding segmentation or diagnosis labels. The inputs and outputs are traditionally treated as *independent and identically distributed* (i.i.d) variables, that build up the sample of the random variables $(X_1,Y_1)\dots(X_n,Y_n)$ with unknown probability distribution, $P_{XY}$.

[6] Although many works do not make explicit the resolution of an inverse problem, this framework allows to characterize them

Informally contextualized, the physical systems modelled in the works presented in this thesis correspond to different tasks typically performed manually by radiologists. Thus the inputs to the system, $x_i$, would correspond to chest CT volumes.Concurrently, experts in the specific problem field depict the outputs yielded as labels, $y_i$ (e.g., binary masks, tabulated reports), in the best-case scenario. Namely, experiments in which the experts are previously trained to provide annotations in a specific format for a supervised problem (see below).

Formally, the characterization by the function $f$ must accomplish:

$$f : \mathscr{X} \longrightarrow \mathscr{Y} \tag{1.2}$$

being $f$ usually found over some set of functions $\mathscr{F}$ under the optimization of a *risk* or error:

$$\underset{f \in \mathscr{F}}{\text{minimize}} \quad R(f), \tag{1.3}$$

representing $R(f)$ as [7]

$$R(f) = \int \mathscr{L}(f(x), y) p(x, y) dxdy \tag{1.4}$$

where $p(x, y)$, if $P_{X,Y}$ admits a density, is its probability density function. $\mathscr{L}$ is a loss function (e.g., Mean Square Error, Cross Entropy) to measure the difference between the system prediction, $f(x)$ or $\hat{y}$, and the real observed output, $y$. However, since $P_{XY}$ is unknown, $R$ is approximated in most statistical learning approaches, employing the *Empirical Risk Minimization* (ERM), which converges when $N$ tends $\rightarrow \infty$ (*consistent*) [31, 320]. It is formulated as:

$$ERM(f) = \frac{1}{N} \sum_{i=0}^{N-1} \mathscr{L}(f(x_i), y_i). \tag{1.5}$$

Within this context, from the type of observations and the necessary assumptions to characterize the distribution, $P_{X,Y}$, and the model, $f$, several AI modelling concepts and their limitations, can be distilled.

Shelling the framework components out, the model, $f$, defined in eq. (1.2), is naively identified as the kernel. Usually, in modelling, two stages are distinguished. In the first stage, features are extracted from the input data to obtain robust *representations* [22] of each entity main characteristics. In the second phase, the features are assembled to decide which ones resembles best the system output. Ideally, the function (1.2) must be bijective.

Thus, the polymorphic mathematical machinery that governs $f$ mutates depending on the application, giving rise to different algorithms instantiated through their parameters within

[7] Several derivations of $R$ and $\mathscr{L}$ can be found in the literature depending on the methodological framework (from ML to Bayesian). Indeed, $\mathscr{L}$ arise from the assumptions about $P_{X,Y}$ and $f$

the domain $\mathscr{F}$. Despite being this process consisting in determining the most suitable model, $f^*$ ($f^* \in \mathscr{F}$), referred to as *learning*, highly heterogeneous, some essential principles rule them.

Each learning principle is closely related to primary AI branches [239], which development and transitory predominance are closely linked to the historical evolution of the computational processing power. The processing gain boost models and their optimization processes to obtain better representations. Focusing on the learning representation process and algorithms complexity, two of the three frameworks for the automation of image processing tasks contemplated in this thesis chapters can be distinguished. Namely, the two most commonly used up to date:

1. `Rule-based models:` Tentatively, the model encodes the input data through code statements to obtain the main features (e.g., in image problems: blobs isolation, blobs roundness, edges orientation in objects). Namely, experts define representative characteristics (input representation) automatically extracted through computer programming (rules). Consecutively, the features are combined through a heuristically parameterized model.

   While this modelling is hugely advantageous in terms of a) being an unsupervised process (does not require annotated data for implementation, at least explicitly), b) usual computational simplicity, and c) *explainability* (defined below) since all processes are well-known; the enormous disadvantage is the very low *generalization* capacity since expert knowledge is primarily based on perception. Coding perception based on predetermined rules is highly subjective and data-dependent, which supposes a constant need for model reparametrization even with slight domain changes in the input data. The segmentation method presented in Chapter 2 employs a rule-based algorithm.

2. `Machine Learning Models:` Contrary to rule-based models, the parameters that govern the machine learning models are obtained by explicitly exploiting the statistical dependencies between the features that represent the inputs of the observed sample (*representations*) (e.q. (1.1)), modelled as random variables and their outputs (e.g. $\mathbb{E}[y|f(x)]$), usually under i.i.d. conjecture. Based on different assumptions, many algorithms are available. All of them relate the available features through a large number of parameters which are fitted by exhaustive search computation processes to optimize a cost function (1.4).

   Traditionally, handcrafted features are employed to feed algorithms, such as *linear regression* (LR) [81], *Neural Networks* (NNs) [174], *Support Vector Machines* (SVM) [320] or *Random Forests* (RF) [33] among others in the *supervised* case. Clustering techniques as *k-means* [28], latent variable models as *Expectation-Maximization*, etc., are implemented in unsupervised setups. The quantification approach of Chapter 2 implements an EM algorithm (see Section Quantifying trough correlation in a closed environment) (EM) [220]. These features can be similar to those extracted for rule-based models. However, considering their limitations and the capacity of modern ML algorithms to work in high-dimensional domains, it is frequent to use as many features as can be obtained without prior knowledge about their representation capacities. This approach intersects with radiological imaging in what is referred to as *Radiomics* [91, 172].

*Radiomics* techniques extract numerous quantitative characteristics (e.g., texture features [105], scale-invariant feature transform (SIFT)) leveraged by digitized radiological images. The extracted features are combined to obtain the predictions, $Y$, that best suits the statistical assumptions aforementioned. The work presented in Chapter 3 is an adaptation of radiomics techniques for the detection of the different TB lesions.

These techniques are based on the use of statistical descriptors as features to represent the input entities. They are difficult to interpret by humans but have been proved to yield better results than those based on more human-understandable ones [340]. So much that the classical two-phase modelling approaches (feature extraction + mathematical modelling) has become a minority in the recent literature through the introduction of end-to-end models. In this alternative scenario, features are automatically extracted and combined in a single phase employing Deep Neural Networks (DNNs), which give rise to the now hackneyed term *Deep Learning* (DL) [173]. The results with DL have equalled or overpass human performance for predictive analytics [78, 256]. However, DL models are usually understood as *black boxes* [193], which results in trustability issues, as introduced later in this section.

Following with the dissection of the framework terms presented in equations (1.1)-(1.4), from the role of the outputs (labels) arise the aforementioned concept of *supervised* or *unsupervised* learning (*self-supervised learning*, SSL[8]), depending on whether or not the model employs system outputs in its design (or to learn the model). Although ML literature shows better results for supervised techniques, obtaining enough data is not always possible (e.g., it is not ethical to infect humans with TB to rely on a bigger $N$). Besides, novel SSL techniques could be essential for model generalization purposes [47, 124, 302].

[8] Self-Supervised Learning is more appropriate. As *Le Cun* points out, unsupervised is a "confusing and overloaded" term [175].

Therefore, in the clinical environment with a frequent lack of annotated data, such techniques represent fresh opportunities [42, 136, 155]. Besides, even when is possible to obtain good quality annotations, they require expert work for long periods of time which is an unaffordable cost in many scenarios. Unfortunately, the biomedical environment is paradigmatic in this case since only a few highly qualified people[9] can generate annotations (i.e., radiologists, pathologists) [200, 309].

The lack of data or *data scarcity* turns the inverse problem (eq. (1.2)) into an ill-posed one [249, 320]. Namely, for each realization $(x_i, y_i)$ that is not in the observed sample, $P_{XY}$ is not defined. Therefore, untractable and unknown as defined before. Among others (i.e., Variational Bayesian Methods [29]), a classical way to circumvent this fact fall on building a more tractable distribution as the conditional probability $P(Y|X)$, relying on probability theory techniques. Usual ML terms as *regression* or *classification* problems arise from the assumption of such approach, namely, $f(x) = \mathbb{E}[Y|X = x]$, being $\mathscr{Y} = \mathbb{R}$ or $f(x) = arg\, max_{y \in \mathscr{Y}} P(Y = y|X = x)$, being $\mathscr{Y} = \mathbb{Z}$. Chapter 4 follow this approach for a multi-task classification problem. Employing just the conditional probability is enough in specific environments, where unseen datasets present conditional distributions similar to those employed during model learning.

This aspect is rarely fulfilled in real-world problems, resulting in a lack of generalization of the proposed model. This concept, which refers to the effectiveness of $f$ to generalize solutions, is named *capacity*. It usually depends on the model *expressiveness*, in the sense of complexity of functions [60]. In addition to *data scarcity*, the model *capacity* is severely affected by *i.i.d.* assumption, generally false, broken due to data mismatches (*distribution shifts* and *selection biases*). They occur naturally in the real world and particularly and very significantly in clinical scenarios, penalizing the *robustness* of the model. To cope with this issue is vitally important to analyze the recognizable mismatches, and for this thesis purpose, how they are extrapolated to clinical datasets. Following the nomenclature given by Castro et al. [42] (in brackets the nomenclature in traditional ML literature), the following are reckon[10]:

*Population shift* (*covariate shift* [140]): It refers to the case when the observed realizations (1.1), employed during model implementation, and the realizations features feeding subsequent inference processes (predictions over new clinical data), are in separated regions of their domain (i.e., different age, smoking status, stays at-risk countries, at samples of subjects included in a TB study).

[9] Experts' knowledge is the physical system to be modelled by $f$.

[10] Further details, especially how the shifts arise from the causal learning framework can be found in the remarkable work [42].

*Annotation shift* (*concept shift* [56, 214]): It arises when observed instances belonging to the same class are labelled differently due to annotators subjectivity (i.e., annotator experience, class definitions).

*Prevalence shift* (*target shift* [343]): It appears in the cases with class balance differences between observed data and new data.

*Acquisition shift* (*domain shift* [247, 271]): Usually, it is due to the use of different scanner and imaging protocols that introduces spurious correlations in the datasets. For example, the same ROI in a CT image, even for the same subject acquired at practically the same time look different under two CT scanners.

*Manifestation shift* (*conditional shift* [310, 343]): Contrarily to the acquisition shift case, manifestation shift occurs when the system outputs present changes between datasets.

*Sample selection bias* [118, 299]: Contrarily to the rest of the shifts, in which the mismatch arises at the data generation process, selection bias occurs at the data collection process. It happens when the dataset subsampling is not uniform due to some selection criteria as image quality control or patients admission criteria.

Given the almost constant need of mitigating distribution shifts, a vast recent literature present methods that favours the model *transfer* to new *Out-Of-Distribution* (OOD) samples (in other words, different datasets). The result is an improved model generalization [150, 247, 347, 348].

Conceptually, this set of methods is coined under the broad term, *Domain Adaptation*, closely linked to hackneyed *transfer learning* [152, 224, 255, 303, 354]. These depend intimately, and their instantiation has produced almost innumerable options captured in the literature. Reviewing each particular technique is outside the scope of this work. Still, among *Domain Adaptation* techniques, it is noteworthy to mention *Data Augmentation* given its widespread use, even before the explosion in the use of DL-based models [75, 136, 246, 289, 306]. *Data Augmentation* techniques could not properly perform *Domain Adaptation* but are included since these augment the joint distribution. Namely, *Data Augmentation* intends to alleviate distributions shifts lessening *data scarcity* by generating new artificial images (experiment draws) transforming available samples to increase the dataset size and its quality. The augmented images depend on *a priori knowledge* about the problem (*domain knowledge* [179]) since they should simulate real-world data. To this aim, colour space modifications, mixing images, projective transformations[11], noise addition, patches deletion,

[11] Generalization framework for DL provided by *Geometric Deep Learning* [36]

among other strategies, are taken into account. Even when the data generation process is mostly unknown, the transformations are inferred using dedicated generative models [156]. This fact and the nature of the *Data Augmentation* algorithms, based on classic transformations or deep generative models (i.e., PixelCNN [230], Variational Autoencoders - VAEs- [160], Generative Adversarial Networks -GANs- [92], Cycle-GANs [353]) [161, 351], make *Data Augmentation* itself an important field inside the AI framework.

The previous paragraphs present several principles for automation of the image analysis task from the different strategies that exist to approximate the members in eq. (1.4). While their development and adaptation to particular problems have recently [123, 188, 351] yielded exceptional and previously inconceivable results, there is still "an elephant in the room". Namely, how trustable are the predictions and therefore AI for the automation tasks? *Trustability*, vital in many complex AI-driven applications, should be a design requirement in healthcare applications [200, 280, 309] to take a sector with die-hard extended ideas to the next level. Trustability is an abstract concept that gives rise to various interpretations [138]. For the sake of clarity, in this work, *trustability* is understood as the addition of two related but commonly studied independently pillars: *uncertainty quantification* and *explainability* [279]. *Uncertainty* is usually inferred as the measure of confidence in the predictions given by a model [83, 167, 301]. Again, there are many different approaches to quantifying *uncertainty* (specific examples appear throughout the thesis), which intend to characterize the following sources of *uncertainty*:

*Epistemic uncertainty:* This type is due to the model structure and parameters chosen to explain the available observed data. Usually, epistemic uncertainty reduces when the number of observed data increases and the complexity of the model decreases.

*Aleatoric uncertainty:* This type is caused by noise unexplained from the data, which mainly arise from *homoscedastic/heteroscedastic* noise and labels overlap. Contrary to *epistemic uncertainty*, *aleatoric uncertainty* is not reducible with extra data. Formally, it is referred as:

1. *Homoscedastic or task-dependent uncertainty*: Occurs when the predictions errors variance have the same distribution independently of the input data values. If a single model is designed to yield more than one prediction/task simultaneously, each task will have its own homoscedastic uncertainty. Chapter 4 presents an approach which estimates Homoscedastic uncertainty to optimize the proposed model loss function.
2. *Heteroscedastic or data-dependent uncertainty*: Contrarily to *Homoscedastic uncertainty*, the variance of the prediction errors depends on the input data.

Finally, *Explainability* qualifies the model by estimating the knowledge humans have about how the model makes a decision [16, 103]. This knowledge allows us to deal with *uncertainty* since it enables the following two control procedures: 1) Interventions in choosing and assembling model mechanisms that better mimic the underlying physical system and 2) direct interventions on these mechanisms values, which have pronounced beneficial effects over healthcare applications [128]. Therefore, it is necessary to understand as best as possible the processes of the DL model. This way, *explainability* comprehend *interpretability*, two terms that are often mistakenly used interchangeably. However, the second one measures humans ability to predict model outputs given different inputs or model parameters variations regardless of "why?". Thus, ML-based models, especially in DL, suffer from the "black box" effect [193, 196, 292] that can be interpretable but hardly explainable. Not surprisingly, techniques such as *disentangling* [213, 344, 346] have recently re-emerged in the field to provide explainability in DNNs. All the concepts above are encapsulated within the novel in-development paradigm, Causal Representation Learning [191, 275], which is the third automation framework used in this thesis. The work in Chapter 5 follows a causal approach for disentangling. Namely,

3. `Causal Representation Learning Models:` Formally, this paradigm arises from the interaction of ML and causal inference [90, 198, 242, 275]. As mentioned, the superior performance of current ML/DL models in predictive tasks for medical imaging is indisputable. Their success lies in the enormous capacity to extract representations of the input data that are strongly related to the output observations. However, these models "are a victim of their success". When modelled naively, ML/DL methods seek relationships based on a mere statistical correlation on the available data [62, 261]. Since correlation usually appears in biased environments (see dataset shifts above), generalization and robustness problems emerge, in addition to explainability ones.

   To alleviate these problems and at the same time exploit the high capacity of DL models to extract features, several approaches appear in the literature giving rise to significant subfields (some of them already mentioned) [100]. We highlight here three archetypal approaches: 1) Incorporating new terms to the lost function, either explicit regularizers[12] or linked to the particular problem (e.g., including overlapping coefficients in a medical image segmentation problem) [93, 170, 223]; 2) Including a *Data Augmentation* or *Domain Adaptation* step; 3) Implementing a new architecture [36, 53, 114, 210, 264].

[12] In Chapter 5 a novel regularizing technique, *Self-Normalizing Neural Networks (SNNs) [164]* is adapted to the medical imaging field

Regardless of the complexity and usefulness of each approach, these techniques allow us to limit the vast solutions space, $\mathscr{F}$, that models with millions of parameters can generated, based on a series of assumptions derived from prior knowledge [27, 321]. For example, from experiments to validate models, it is well-known that parameters with high specific values are a clear symptom of overfitting. This is controlled by inserting regularization terms as $L^1$ or $L^2$[28, 187][13]. Alternatively, the Convolutional Neural Networks (CNNs) assume spatial correlation in the input data [74, 342]. As expected in the neighbourhoods of the images for which, in the automation of their analysis, the CNN present the best results in the literature. This is said considering that the penetration of Transformers literature is still low [45, 72, 111].

This set of assumptions or inductive biases [20, 100] is fundamental to the success of DL models. They delimit the solution space [137]. Namely, they force the model to find more general representations which intend to hold across different datasets. Therefore, high-correlated but spurious signals present in shifted datasets used in the learning phases are not prioritized [10, 42]. However, introducing inductive biases in models guided by statistical learning may not be enough. Even with enriched models able to adapt to more environments or different datasets, solutions are build on the correlations between the representations and the observed outputs of the system under this framework. Since "correlation does not imply causation"[206], both the generalization and the explainability of the models are difficult to prove.

Fortunately, this problem is not new in AI. Causal inference theory pioneered by *Judea Pearl* [240, 243, 244] and developed during the last 40 years allows finding causal-effect relationships from the statistical characterization of the observed variables. However, its integration with DL models is not trivial and is currently a hot topic in the literature, which adopts the mathematical machinery of causation (*calculus of causation*) [241, 312] to enable causal DL models. Thus, we can develop automated algorithms formalized through causal models based on *a priori* knowledge (inductive biases) embedded as causal graphs, structural equations and more [120, 243, 249].

This intersectional framework is still under development and important concepts beyond the scope of the present work such as *identifiability* needs to be integrated [248, 308]. Adopting this framework, it is possible to establish causal models guided by graphical diagrams. Their high-dimensional input and output mechanisms are defined by DL architectures. They allow to establish hierarchical models governed

[13] From a Bayesian perspective, $L^1$ and $L^2$ correspond to adding a Laplacian and Gaussian prior over the parameters distribution, respectively

by meaningful variables in the image generation process, as in the work presented in Chapter 5.

This configuration intends to build the causal knowledge structure defined by Pearl as "Ladder of Causation" [243]. This structure enables not only the statistical characterization of particular observations from the available datasets but also perform interventions on the variables that govern the model. Interventions empower counterfactuals that enable imagined spaces [181, 195]. They are potentially vital in medical applications [99, 238, 258] to answer questions such as, what would be the evolution of a damaged lung if the patient had followed a different treatment? What would happen to a human lung if the treatment only consisted of clinical trials on non-human primates and mice?.

Summarizing, this section briefly presents several basic principles of AI that are fundamental in their intersection with the field of medical imaging and specifically, with the analysis of damaged lungs. The constant evolution of these principles and the adaptation to the problem at hand is presented in the following section, through the fundamental papers published in recent years.

### 1.4.2 AI in service of CT Pathological Lung Assessment Approaches

The successful use of radiological imaging for diseases assessment falls on the reliability of the information yielded by imaging biomarkers.. Imaging biomarkers can be of a very different nature; ranging from the segmentation and measurement of the entire Region of Interest (ROI) to those extracted by a DNN, including the measurement of specific manifestations or the characterization of voxel neighbourhoods using statistical descriptors (*radiomics*).

While it is true that we can attempt to quantify the damage caused by TB in the lungs without first isolating them, prior segmentation of damaged lungs is a convenient initial step to limit the area in which lesions will be located [204]. Following this approach, we drastically avoid extracting spurious correlation from the ROI context, which harms biomarker quantification [70, 176, 203] (see Section 1.4.1). Indeed, the segmented ROI acts as an inductive bias [100].

Moreover, Pathological Lung Segmentation (PLS) is critical for a majority of CADx or CADe applications [84, 278, 337]. While the quantification of disease-specific lesions is more a particular problem. For instance, in the case of lung cancer or Chronic Obstructive Pulmonary Disease (COPD) the characteristic manifestations to quantify are emphysema, fibrosis or specific nodules [109, 119, 147, 277, 304], while for human TB are cavitations, tree in buds, etc. (see Section Medical Imaging for Tuberculosis Assessment).

Besides, the development of new drugs, ERA4TB project motivation (Project Framework: ERA4TB), depends on translational biomarkers. Usually there is not a unique correspondence between human and other animal models TB manifestations or sometimes the radiological manifestations are not even defined. For that, the correct segmentation of the whole pathological lung represents a suitable alternative.

Given the particular importance of the segmentation problem, it is common in the literature to categorize the works into those based solely on segmentation or those based on the extraction of imaging biomarkers for diagnostic, even when the preprocessing includes automatic segmentation. Adding this fact to the significant concepts introduced through the section, the relevant literature is examined with a view in: 1) the model $f$, namely, a) rule-based models or ML models, meaning, b) radiomics (with or without handcrafted features) and c) DL models; 2) whether the application aims PLS or diagnosis, and 3) their capabilities in terms of trustability and generalization as shown in Fig.1.7. The values of generalization and explainability in Fig. 1.7 depends on qualitative criteria, i.e., the number and diversity of the datasets, the analysis of the ablation experiments, the ability to generate realistic synthetic images from random and intervened models. Therefore, the figure provides an approximate representation of the exact capabilities presented in each paper.

As it is depicted in Fig.1.7, rule-based methods generalize poorly, in contrast to their trustability. This fact is explained by both the chance to measure epistemic uncertainty in less complex models and the higher levels of explainability due to the direct injection of expert knowledge to obtain features as inductive biases. Actually, in a counterproductive manner, it is common to find models excessively biased to fit a small dataset that describes a particular pathology in a specific sample as in Abdillah et al. [3].

In general, the complexity of the algorithms grows to reflect the biological heterogeneity within clinical datasets due to both inter/intra-subject and pathological variability. The initial approaches, mainly rule-based, focus on the segmentation of healthy lungs using thresholding algorithms, the simplest inductive bias. These methods extract objects (blobs) from the distribution of grey values in the image. To this aim, the algorithm attempts to find the threshold value that best separates the objects of interest (the lungs, in this case).

This technique works well with CT images since these have grey levels associated with the different tissues of interest. However, the algorithms are very sensitive to noise and abnormal patterns present in the infected tissue. Thus, it is necessary to apply several morphological operations to obtain a still sub-optimal segmentation. Thresholding methods cannot provide proper segmentation of damaged lungs. However, they represent the first step for many algorithms. In the literature, we find paradigmatic thresholding works for the segmentation of healthy lungs, such as the one by Hu et al. [131]. This work solves a simple problem employing an iterative thresholding model that is easily interpretable. The model is able to adapt to the different domains of the healthy lung problem much better than most rules-based algorithms mentioned in Section 1.7, given that they focus on PLS.

To improve the generalization of rule-based models, the algorithms add complexity up by combining thresholding with region growing techniques [5], as in Shen et al. [286]. However, the study presents the aforementioned overfitting issue; parameterization fits a particular dataset (image acquired just from one scanner).

As mentioned, most thresholding techniques focus on the segmentation of the whole lung since inferring the mechanisms to delimit lesions is much more complex. Even so, there are approaches for specific cases where thresholding techniques (i.e., segmenting nodules) are applied for diagnosis. For example, in the work presented by  Roy et al. [267].

Continuing with the description of rule-based techniques, we come across region-based methods such as that of  Hojjatoleslami and Kittler [127]. Their method follows a traditional seeded region growing algorithm. It consists in the evaluation of voxels close to a previously deterministically established voxel, the seed, and setting a criteria to decide whether the evaluated voxels belong to the lung region.

The accuracy of this class of methods depends heavily on: a) the correct identification of significant voxels that can be considered seeds; b) the definition of the neighbourhood, i.e., which voxels are close enough to it to be evaluated, and c) the definition of the neighbourhood resemblance criteria. Thus, human expert intervention controlling all the parameters becomes frequently essential to achieve appropriate segmentation or diagnostic results as in  Farag et al. [79].

Alternatively, employing techniques halfway between rule-based and ML methods to improve model expressiveness and therefore generalization is another explored via, such as Grady [102] where growth is ruled through a *Random Walk* algorithm.

The next class of rule-based methods are shape-based models. These methods segment the structure employing an atlas defined or built by experts as Li et al. [182]. An atlas consists of a lung template for CT images in the present case, containing labels of the anatomical structures of the regions of interest. This template is aligned, or in the image processing jargon, registered, with the image to be segmented by optimizing an indicator of similarity between the two (e.g., the mutual information between the images). Once both images are aligned, the region of interest is segmented. These methods work well enough for cases where lesions have not abruptly modified the expected structure. However, in the case of tuberculosis, they are often ineffective because the lungs of damaged subjects differ significantly from healthy lungs.

The last major group of rule-based methods for PLS commonly defined in the literature are neighbouring anatomy-guided methods. This class of methods uses information about the structures expected to appear around the organ of interest. For example, the rib cage, the heart or the diaphragm as in Artaechevarria et al. [14], in the case of the lungs. This procedure is valuable when the area of interest is so damaged that it is impossible to recognize it. Its main problem is that it requires that the neighbouring organs are not damaged or affected by image acquisition artefacts or pathologies to function correctly.

Although the segmentation or diagnosis obtained is better than with less complex methods, most algorithms do not generalize well. Moreover, it is common to use larger datasets to adapt the parameters of the system and sacrifice quality for an improvement in generalization as in the work Kuhnigk et al. [169], Soliman et al. [294]. Thus, most rule-based methods require the introduction of the Human-in-the-loop (HIL) to refine the results [169].

To avoid this fact, radiomics encompasses both those methods that use a combination of handcrafted features obtained from the *a priori* knowledge of the radiologists and those methods that extract descriptive statistics from the images. In both cases, the features are extracted from the image input through an *ad-hoc* method, oppositely to DL methods which obtain them automatically. Commonly, classifiers use all kinds of features together.

Thus, the different radiomics methodologies presented in the literature are distinguished by the nature of the features that feed the model, $f$, and the algorithm that governs it. The model choice depends on the study of diverse factors related to the nature of the data available and the capabilities of each model to deal with noise, mismatches, etc.

In general, as hypothesized by the blue shaded area (Fig.1.7), radiomics methods present an intermediate trustability and generalization capacity. Since features are known, their importance can be measured through multivariate analysis as in Coroller et al. [59] and Hawkins et al. [112]. The specific statistical dependencies on each model can also be exploited to estimate the importance of the features by leveraging uncertainty measures and explainability. The explainability is conditioned by the features nature too. Thus, the model design must consider the trade-off between trustability and generalization, as a consequence of employing human-interpretable features and those that are not but provide greater expressiveness.

The most up-to-date methodological examples can be found by closely scrutinizing among "all smoke and mirrors" COVID related works [261]. Thus, in Shi et al. [287], Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) correlated features were handcrafted and exploited by a biased version of an RF [33]. It this way, the model was general enough to work with a multicenter sample achieving a good comparison against radiologist scores.

RF was the most successful algorithm, in performance terms, until the advent of DL. Thus, a large part of the literature employs RF models fed by mixed feature sets (texture, wavelet, handcrafted, etc.) [91, 172]. As it is the case of Tang et al. [305] which exploit both texture features and the volumes of the regions to be classified or Wilson and Devaraj [331] which adds wavelets features [260]. In addition, RF is also often used for feature reduction employing the Gini importance metric [208]. Different approaches select the prediction most correlated features with an RF in the first stage of the model. Subsequently, they built a simpler classifier such as an LR, similar to that proposed by Qi et al. [254]. This way, the model is more interpretive, and its uncertainty is easier to estimate. Oppositely, for numerous works, the analysis of the features leading to predictions is not the goal. In this context, Christodoulidis et al. [52] presents an hybrid approach between DL and radiomics to classify lung tissue patterns by extracting features with a CNN.

There are a reduced number of works in the literature, employing radiomics for PLS. Algorithms can recognize texture patterns and classify them as parenchymal tissue or not by setting a threshold for the model metric. However, the fine delimitation of such a region presents a problem. Most algorithms extract features and assign the same class (tissue, lesion, etc.) from regular fixed-dimensional voxel neighbourhoods ($3x3$ for 2D images, $3x3x3$ for image volumes), providing a coarse segmentation. Necessarily, a post-processing algorithm to provide a finer level of detail is applied as in the work of Liu et al. [189].

Although RF is the classifier most used in recent years for radiomics, classifiers of a very different nature, such as SVMs based on kernel methods, are also common: Chen et al. [44], Alam et al. [7] or Singh and Gupta [291]. Since the SVMs performance is similar to other models and the lack of interpretability caused by the use of kernels, SVMs are less common in clinical environments.

However, the reluctance to "black box" models have diminished with the surge of DL models that yield results that far exceed those obtained with previous paradigms. Thus, as already mentioned, the current state-of-the-art (SOTA) is dominated by DL approaches and the incorporation of tools to improve their generalization. As a proof, we have the recent COVID-19 related literature that could serve as a decalogue of the prevailing methodology concerning the automation of image analysis of pathological lung, even despite its pitfalls [261] given by the pandemic emergency and the annoying publication bias [139]. Cao et al.

[39] paper illustrates the capabilities of DL models to detect and delimit specific disease manifestations using a toy dataset of two cases. Several works employ the ubiquitous U-Net for segmentation as in Chen et al. [46] or the U-Net three-dimensional counterpart [210, 264] in the search for abnormalities in lung structure [281]. The U-nets and other architectures, as the *ResNet* [114] in the work of Li et al. [183] and Song et al. [296], are sometimes used as the baseline architecture for detection and classification. Work examples that extend naive DL models such as the above, incorporating new tools that mitigate the negative effect of strongly biased datasets (see 1.4.1), include, among others, the work of Zheng et al. [350] using *weak supervision* [352] or those of Gozes et al. [101] and Wu et al. [334] adding explainability to the models by the use of *saliency maps* [290].

Beyond COVID-related literature, we found essential contributions to the intersection between automation of pathological lung analysis and DL. Thus, it is very remarkable the research of Van Tulder and De Bruijne [319]. One of the first works adapting DL-based representation learning theory to lung CT image processing. To do so, it combines generative and discriminative models, named *hybrid models*. This work is before the explosion of deep generative model's [92, 160] but still proves the effectiveness of such approaches.

Alternatively, there exists numerous works pioneering PLS tasks using DL. It is worth mentioning the work of Gao et al. [87] that already employs multiresolution analysis to address pattern recognition in *Interstitial lung diseases* (ILD). Also, the paper of Alakwaa et al. [6] among so many appeared to tackle the cancer nodule detection problem presented at the *Kaggle Data Science Bowl 2017* [149][14]. Regarding the generalization problems already mentioned, it is necessary to point out that the *Kaggle* challenge was subsequently won by Liao et al. [185]. They adopted for the medical imaging field, a 3-D Deep Leaky Noisy-OR Network. Shortly after, Google AI researchers [9] surpassed these results employing a 3-D Mask R-CNN [116]. However, both applications turn out to be hardly integrable in the clinical workflow [142].

Continuing with significant work for PLS, the paper by Harrison et al. [108], published in 2017, presents a remarkable alternative using "progressive and multi-path holistically nested neural networks" to the U-net architecture. U-Net is still the baseline in the field, probably because the chance of obtaining similar results to those reported depends more on the diversity of the data available during learning than the model choice, as pointed out by Hofmanninger et al. [126].

For this reason, this summary includes innovative work on models that encourage this generalization, even employing limited datasets. Thus, Gerard et al. [89] paper stands out, which shows that training a model in several steps allows the use of images of multiple mam-

[14] https://www.kaggle.com/c/data-science-bowl-2017

malian species for PLS. A similar strategy to segment human lung lobes in high-resolution CT images follows Lee et al. [176] work. Likewise, Xie et al. [335] also gives an alternative to segment lobes on images acquired with a high-resolution protocol. They add structural relationships that act as inductive biases for the DL model. Finally, the work of Amyar et al. [8] enriches the model by proposing multi-task learning. Namely, learning the lung segmentation masks together with the severity degree caused by the pathology. The Chapter 4 follows a multi-task approach for the segmentation of TB radiological manifestations.

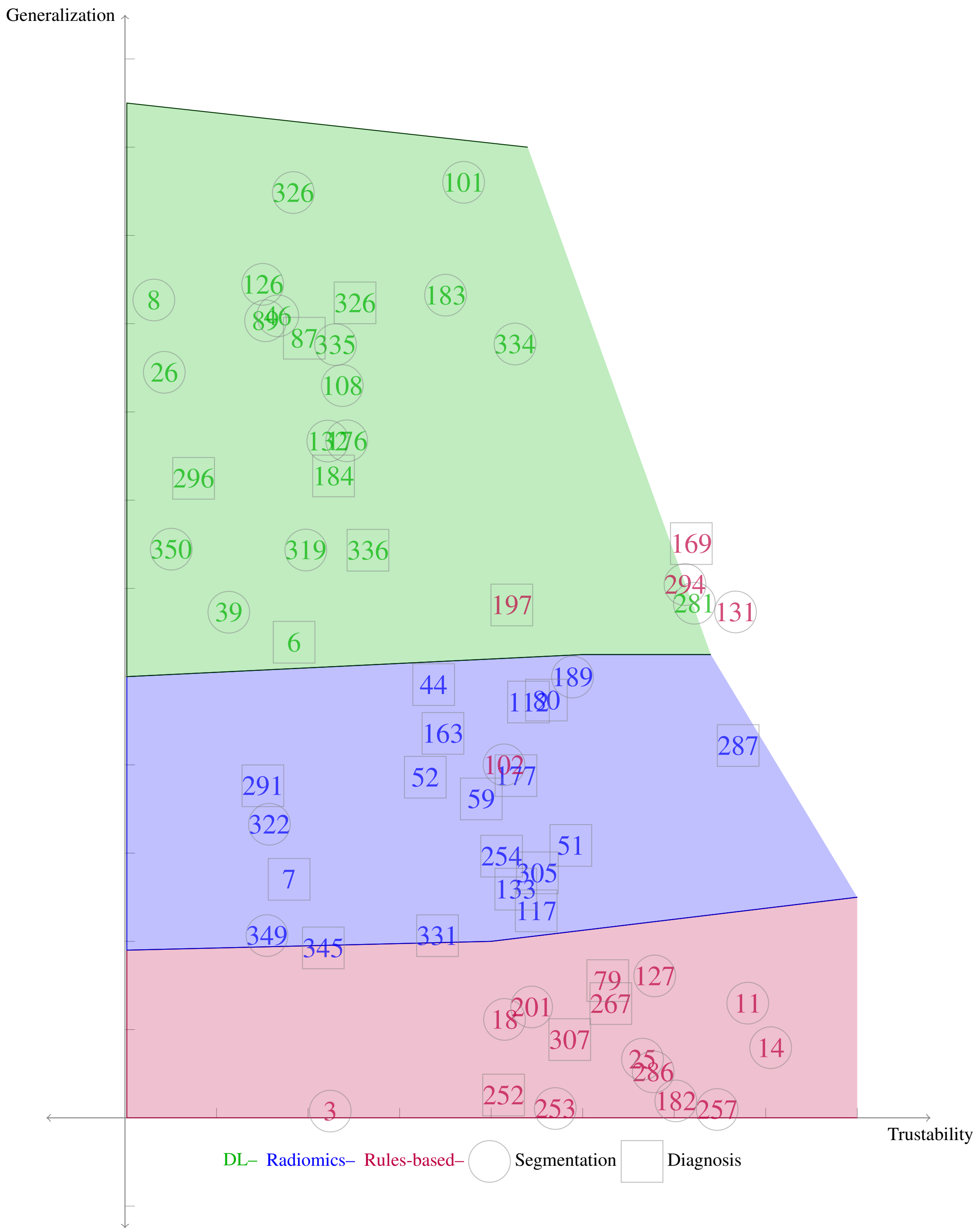


Fig. 1.7 Qualitative categorization of representative pathological lung analysis works (bibliography reference number) depending on their generalization and trustability (uncertainty measure and explainability) characteristics. Red, blue and green depict Rules-based, radiomics and DL models, respectively, for the dominant AI approach of each work. ○ or □ represent the main focus of each entry, namely, Segmentation or Diagnosis (which depends on some previous segmentation algorithm). Shadowed areas conceptualize each AI branch capacity for our purpose.
Note that [131] consider only healthy subjects and [169, 294] are Human-in-the-loop (HIL) approaches.

# Chapter 2

# Lung Segmentation and Quantification with Rule-based Approach

## 2.1 Coding human perception and expert knowledge

From a general perspective, the specific goal of segmenting and quantifying a pathological lung image could be seen as the attempt to capture heuristic knowledge from the human experts to encode it as methods, processes or algorithms to perform the task.

Regardless of whether the knowledge comes as a *rule of thumb*, intuitive judgments or educated guesses [239] as in the problem case (from experienced radiologists), the most traditional way to employ Artificial Intelligence (AI) needed for task automation is through a set of encoding rules (`if A: do B...`) [54]; in contrast with Machine Learning (ML), where the rules are learned from labelled data and/or features prescribed (supervised or unsupervised learning) or even just the data (e.g., Deep Learning (DL)) employing methods in the intersection between statistics and computer science (see Section 1.4.1).

This primary approach to AI for CADx/CADe presents some limitations. The encoded models do not just intend to represent a human-defined procedure (for example, decision tree protocols) or a well-known shape (e.g., ellipse detection through the Hough transform [73, 130]), but also need to encode the complexity of human perception, which is loosely defined for this purpose. DL models are based on NNs , which are strongly inspired in the neurological basis [134, 173]. Taking this into account, it is maybe less surprising that DL methods, instead of rule-based ones, have recently achieved the most successful results for perception related tasks [188, 309].

Besides, the expert knowledge cannot be adequately encoded using a set of classic programming control structures. Usually, users apart from the author are unable to set up the proper parameters for a new environment (e.g., new animal model, new acquisition scanner) (see Section 2.4).

However, we can gain much valuable information employing a rule-based approach. We can obtain easily informative models for domain-specific problems that, even with limitations, could be enough in some scenarios.

The use of a traditional methodology at this thesis stage serves a double purpose: 1) to promote a solution to the problems mentioned above in datasets composed of TB-infected lungs for different animal models (see Section 1.4.2) and 2) to gain further insight into the segmentation problem. In this way, we avoid the usual black box effect of the more optimized but less informative ML/DL most employed models [188, 216, 309], to subsequently enrich them, by injecting in different ways, as far as possible, the knowledge acquired during this first approach.

## 2.2 Rules-based Lung Segmentation Method for a Specific Domain

As was mentioned in Section AI in service of CT Pathological Lung Assessment Approaches, segmentation of TB-infected lungs is complex in clinical and preclinical studies. The expected variability of the pulmonary inflation caused by the respiratory cycle is increased and less predictable than the healthy subjects. This is due to the changes in lung compliance caused by the disease and the breathing difficulties experienced by anaesthetized infected animals. Moreover, CT image acquisition in TB animal models is usually performed on free-breathing animals to avoid the additional level of complexity added by the intubation. It results in the presence of significant respiratory motion artefacts. This effect produces fuzzy boundaries, especially in the diaphragm area (Fig. 2.1). Thus, it implies an uncertain delimitation of the lungs beyond the segmentation technique used.

Manual segmentation of the lungs is subject to wide intra-, and inter-expert variability in the presence of those fuzzy boundaries [313]. Most of the state-of-the-art methods for automatic lung segmentation are not designed to deal with the specific problems present in *Mtb*-infected lungs under the presence of strong respiratory motion artefacts [204], as was mentioned before, even under the DL scope (see PLS approaches). They generally are not able to differentiate between the neighbouring soft tissue and the lesions attached to the pleura since their density (Hounsfield Units, HU) is similar [209]. Well-known thresholding methods [131] perform well when extracting healthy tissue but cannot cope with HU variability. Region-based methods [102, 127] fail in the presence of abnormalities

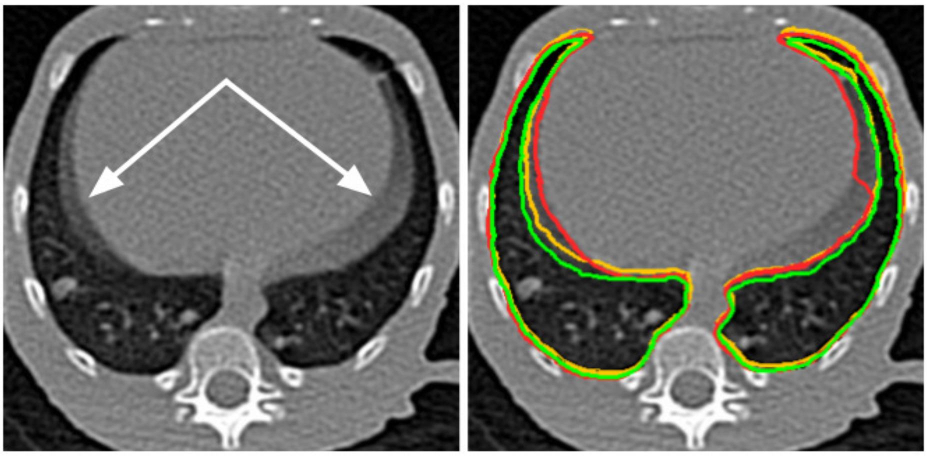


Fig. 2.1 (Left) Sample slice from a chest CT volume of a subject infected with *Mycobacterium tuberculosis*. The presence of fuzzy boundaries (white arrow) caused by respiratory movement artifacts makes it difficult to delimit the lung boundary; (Right) The annotations performed by the experts are combined to explicitly illustrate the differences and shown with a red, yellow and green outline, respectively.

and are highly user-dependent. Atlas-based methods [182] fail to obtain a suitable general model able to capture the singularity of the disease. The more recent approaches are primarily based on supervised learning methods [340]. They require a large dataset labelled by an expert to ensure appropriate training and even then, are not free from bias.

In the remainder of this chapter, an automatic pipeline able to segment lungs infected with *Mtb.* placing considerable importance on the robust and consistent identification of fuzzy boundaries is presented. This is our first approach to the segmentation of infected lungs and follows a traditional rule-based methodology. This method was already published in the paper, *Unsupervised CT Lung Image Segmentation of a Mycobacterium Tuberculosis Infection Model*, [95]. Therefore, most of the content of the following paragraphs were already presented within it.

## 2.2.1 Materials

### 2.2.1.1 Experimental Animals

Male cynomolgus macaques, aged 3 to 4 years, were obtained from an established UK breeding colony for these studies. Genetic analysis of this colony has previously confirmed the cynomolgus macaques to be of Indonesian genotype [211]. The absence of previous exposure to mycobacterial antigens was confirmed. All animal procedures and study designs were approved by the Public Health England Animal Welfare and Ethical Review Body, Porton Down, UK, and authorized under an appropriate UK Home Office project license. All animal procedures were performed on a facility with biosafety level 3 laboratories.

#### 2.2.1.2 Aerosol Exposure

Macaques were challenged by exposure to aerosols of *Mtb* as previously described [284, 285]. Mono-dispersed bacteria in particles were generated using a 3-jet Collison nebuliser (BGI, Waltham, MA, USA) and, in conjunction with a modified Henderson apparatus, delivered to the nares of each sedated primate via a modified veterinary anaesthetic mask. The challenge was performed on sedated animals. They were placed within a head out plethysmography chamber (Buxco, Wilmington, North Carolina, USA) to enable the aerosol to be delivered simultaneously to measure respired volume. The calculations to derive the presented dose (PD) (the number of organisms that the animals inhale) and the retained dose (the number of organisms assumed to be retained in the lung) have been described previously [107, 284, 285].

#### 2.2.1.3 CT Imaging

Our dataset comprises 63 CT scans of the chest acquired from 9 different subjects at 7 time points (0, 3, 12, 16, 20, 24 and 28 weeks after aerosol exposure to *Mtb*). The subjects were treated with different combinations of antibiotics (see Table 2.3) [285]. The chest CT scans were acquired with a 16-slice Lightspeed CT scanner (General Electric Healthcare, Milwaukee, WI, USA) with voxel spacing of 0.23 mm x 0.23 mm x 0.625 mm and in-plane resolution of 512 pixels x 512 pixels.

### 2.2.2 Methods[1]

#### 2.2.2.1 Automatic Lung Segmentation

The automatic lung segmentation pipeline is composed of three main steps, as depicted in Fig. 2.2 and explained in the following sections.

#### 2.2.2.2 Preliminary Lung and Airway Tree Segmentation

*Automatic Adaptive Thresholding:* The first step goal is to obtain a rough segmentation of the lungs, including the airway tree, similarly as was introduced by *Hu et al.* [131] by separating air-filled structures (i.e., healthy parenchyma, stomach, airways, image background)

[1] The `C++/ITK` [146] code implementation for the methodology description can be found under the Tuberculosis Lung Segmentation (TLS) GitHub repository

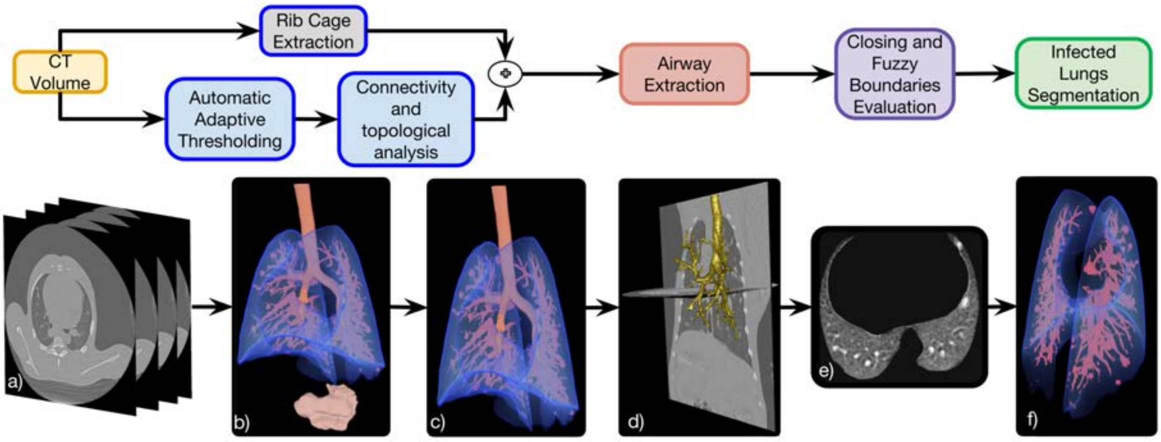

CT Volume
Rib Cage Extraction
Automatic Adaptive Thresholding
Connectivity and topological analysis
Airway Extraction
Closing and Fuzzy Boundaries Evaluation
Infected Lungs Segmentation
a)
b)
c)
d)
e)
f)

Fig. 2.2 Automatic lung segmentation pipeline: (a) Source chest CT volume; (b) 3D rendering of the air-like structures detected in the image using automatic adaptive thresholding; (c) 3D rendering of the preliminary lung and connected airways segmentation obtained using a set of topological operations based on the position of all pre-segmented structures; (d) Isolated airways tree extracted with a propagating wavefront approach; (e) Axial slice of the final lung segmentation in which the lesions caused by *Mtb* and attached to the pleura have been included and the motion artefacts discarded; (f) 3D rendering of the final lung segmentation including healthy parenchyma, the damaged parenchyma and the blood vessels.

from more dense tissues in the whole image volume (Fig. 2.2 (a) and (b)) in which a bi-modal histogram distribution is expected, especially for healthy lungs, by finding a threshold $T$ given by the iterative equation 2.1.

$$T_{i+1} = \frac{\mu_a + \mu_{na}}{2}, \tag{2.1}$$

where $\mu_a$ y $\mu_{na}$ are the average intensity of voxels below and above $T$ for the iteration $i$. The find ends once $T_{i+1} = T_i$. While this method have been widely employed in several studies is quite biased as is based in a unreal assumption (inductive prior) for this work environment: volumetric images to segment present big and isotropic Signal-to-Noise ratio (SNR).
Since the study images present movement artefacts as was mentioned before (see Fig. 2.1), *Hu et al.* method was discarded in favour of *Otsu* method [233, 316]. *Otsu*'s method adapt the classic Fisher Linear Discriminant (LD) [28] for image thresholding. To this aim, the algorithm assumes the existence of a bimodal (two Gaussians) distribution (air-like and non air-like structures), two tissue classes ($c = [1,2]$), and minimize the within-class variance, e.g: $s_1^2 + s_2^2$, while maximize the separation between the class means ($S_B$), e.g: $(m_2 - m_1)^2$,

minimizing the class overlap under an optimized $T$ as follows:

$$\arg\max_{T} \frac{(m_2(T) - m_1(T))^2}{s_1(T)^2 + s_2(T)^2}, \tag{2.2}$$

being

$$m_k(T) = \frac{1}{N_k} \sum_{v \in k(T)} I(v) \qquad s_k^2(T) = \sum_{v \in k(T)} (I(v) - m_k)^2, \tag{2.3}$$

where $I(v)$ is the image intensity value (HUs) at voxel $v$ and $N_k$ the number of voxels belonging to class $k$.

*Rib Cage Extraction:* Although the literature contains robust approaches to rib cage and sternum segmentation [151, 190, 298], it was not necessary for our purpose— and beyond the scope of the present study— to implement a highly accurate and time-consuming segmentation. Instead, we used a simple technique, which, although unable to capture each bone's specific shape, was good enough to establish a convex hull for the ribcage. First, we defined voxels with a value similar to the rib cage bones (over 900 Hounsfield units (HU)) as seeds. Then, we perform region-growing segmentation using the criteria given by the confidence connected segmentation method [250].

*Connectivity and Topological Analysis:* In order to isolate the lungs from the rest of the segmented air-filled structures, as described in [15, 178], we utilized the differences in size and anatomical location of the hidden objects as follows: a) excluding the objects located outside the convex hull formed by the partially extracted ribcage (Fig. 2.2, as those which a volume less than $10\,mm^3$ (c)) and b) selecting as lung tissue, the structures at the minimal Euclidean distance to the ribcage centroid.

#### 2.2.2.3 Airway Tree Extraction

Due to the intricate morphology of the airway tree, a specific algorithm was needed to extract it from the overall lung volume (Fig. 2.2 (d)). Our approach adapted a method based on modelling a propagating wavefront through the trachea, as introduced by *Schlathoelter et al.* [274] and extended by *Bulöw et al.* [37]. The implementation described in the following sections focuses on leakage detection, which is usually a significant problem for the mentioned approaches.

*Trachea detection and initialization:* The origin of the trachea is detected using a slice-by-slice search for the first isolated, air-filled area with a diameter from 5.5 *mm* to 8.5 *mm*

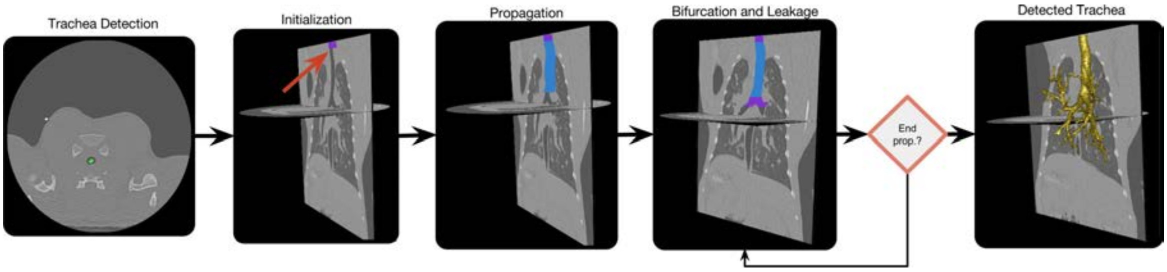


Fig. 2.3 Airway tree extraction workflow. Step 1: Trachea seed detected by morphological analysis of the *HRCT* slices; Step 2: The trachea section is initialized by adding the neighboring voxels to the seed, thus creating a dome; Step 3: Spherical wavefront propagation ruled by the algorithm [37, 274]; Step 4: Check for bifurcations and leakages of the wavefront into the lungs; Step 5: The resultant isolated trachea after propagation.

(depending on the animal's weight [251]) and a roundness above 0.9. The centre of mass (highlighted in green in Fig. 2.3-Step 1) is chosen to be the seed to form a dome, including the surrounding voxels and emulating a spherical wave.

*Wavefront Propagation:* The wavefront propagates, and the decision on whether to add voxels from the neighborhood (segments) is based on a 3D fast marching level set algorithm, which is ruled by the $time\ step$ and two thresholds as defined in *Artaecheverria et al.* [13, 65, 82]:

- $T_i = \mu_s + \alpha \cdot \max(\sigma_{s-1}, \sigma_{s-2})$ for the similarity between a voxel and its neighborhood, where $\mu_s$ is the mean intensity of the neighborhood voxel, $\sigma_{s-1}$ y $\sigma_{s-2}$ the standard deviation of the voxels added HUs at the two previous iterations and $\alpha$ the propagation factor.

- and $T_s$ for the intensity gradient given by the evaluated voxel and its neighborhood computed using a three-dimensional Sobel filter.

After each propagation step, the dome shape is checked to detect possible bifurcations and the presence of leakages. If these are detected, the propagation of the current wavefront ends, thus defining a segment. If a bifurcation is found, two new wavefronts are initialized.

*Bifurcation Detection:* The algorithm exploits the expected wavefront shape to determine when a bifurcation occurs. The following rule is used: if $r_a > \beta \cdot r_e$, then mark that a bifurcation has occurred. The parameter $r_a$ is the actual radius of the dome, $r_e$ the expected radius and $\beta$ a scalar factor (commonly chosen between one and two).

*Leakage Detection:* Owing to the presence of partial volume effects, beam hardening, motion artefacts or low-radiation induced noise, the contrast between the airway lumen and the walls might become insufficient to guide the segmentation. Consequently, the wavefront could leak into the lungs. Two control mechanisms are implemented [13]: (a) The number of newly generated wavefronts after bifurcation is restricted to two. As a larger number generally indicates that several small segments are growing next to each other; this is a common indicator of leakage; (b) To be accepted, a fully grown segment needs to comply with three restrictions as measured by the growth rate, the compactness and the differences between wavefront sizes.

The *Growth Rate, GR*, indicator evaluates whether the waveform has propagated uniformly. It is defined as:

$$GR = \frac{1}{N}\sum_{i=1}^{N}\frac{|W_i|}{|W_{i-1}|} \quad < \quad T_{GR}, \tag{2.4}$$

where $|W_i|$ is the number of wavefront voxels at propagation step $i$, $N$ the number of propagation steps and $T_{GR}$ a threshold. Commonly, $T_{GR}$ is chosen slightly larger than one.

The *Discrete Compactness*, $C$ [35], is computed as:

$$C = \frac{n - \frac{A}{6}}{n - (\sqrt[3]{n})^2} > T_C \tag{2.5}$$

where $n$ is the number of voxels of the solid volume, $A$ is the segment surface area and $T_C$ is a threshold defined to separate correct from incorrect segments. The typical range for $T_C$ is $[0,1]$.

Finally, the difference between the sizes of the last $(W_{Last})$ and the first $(W_{First})$ wavefronts is also computed and compared with a threshold $T_W$ because a large difference (over 10%) is a typical sign of leakage:

$$|W_{last} - W_{first}| < T_W \tag{2.6}$$

#### 2.2.2.4 Morphological Closing and Fuzzy Boundaries Evaluation

The last step of the automatic lung segmentation procedure is a refinement process to include missing lesions attached to the pleura and remove the fuzzy boundaries produced by the respiratory motion artefact.

*Morphological 3D Hole Filling:* Holes, defined as black voxels of the mask that are not connected to the boundaries of the lung segmentation, are removed with an iterative hole-filling filter using the approach described in *Janaszewski et al.* [144] (see Fig. 2.4 (b)). At each iteration, a hole neighbourhood ($1mm$ x $1mm$ x $1mm$) was evaluated to add new voxels to the mask. It is important to remark that the parameters driving the morphological

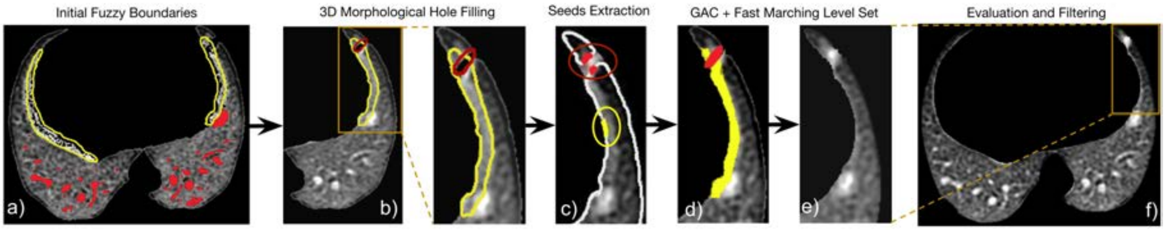


Fig. 2.4 Lung segmentation evaluation workflow illustrated using a sample sagittal CT slice multiplied by its lung mask: (a) Axial slice of the segmented lung obtained after the Lung and Airway Segmentation and Airway Extraction processes showing holes (black areas inside the parenchyma) and fuzzy boundaries (in yellow); (b) Segmentation after the 3D morphological hole filling process including the holes enclosed by the lung parenchyma; (c) Seeds extracted on the eroded lung surface both in fuzzy boundaries (in yellow) and in TB lesions attached to the pleura (in red); (d) Respiratory motion artefact in the diaphragm area (in yellow) and TB lesion mask (in red) extracted by the combined level set and active contour approach; (e) Final segmentation in which the lesion attached to the pleura has been included and the fuzzy boundaries excluded.

operations are fixed based on the prior knowledge about the subject's anatomy (see Experimental Animals), and its value is kept the same for all CT volumes.

*Fuzzy Lung Border Segmentation and Evaluation:* We specifically propose excluding movement artefacts and including lesions attached to the pleura in our lung segmentation using level sets and active geodesic contours [41], which have proven successful in similar tasks [79, 300]. First, the lung surface was extracted from the mask obtained after the morphological 3D hole-filling process: the lung surface was computed as the subtraction of the mask, and an eroded version was computed using a kernel of 1 *mm* radius. Then, to obtain the seeds automatically for the level-sets algorithm, we assumed that the fuzzy regions (lesions or respiratory movement artefacts) had the highest values at the lung boundary (see Fig. 2.4 (b)). Therefore, the seeds are chosen to be the outliers (or less probable values) of the intensity distribution at the previously delimited lung boundary (see Fig. 2.4 (c)). We set a voxel, $v_i$, as the seed based on the following criteria:

$$v_i \in seeds \iff I(v_i) \geq \mu_{sp} + 2.5\sigma_{sp} \quad \forall v_i \in sp_{border} \tag{2.7}$$

where $I(\cdot)$ is the voxel intensity, $sp$ represents the segmented lung parenchyma obtained as the output from the morphological hole-filling routine, $sp_{border}$ corresponds to the boundary voxels, and $\mu_{sp}$ and $\sigma_{sp}$ are the mean and standard deviation of the intensities of the voxels within $sp_{border}$, respectively. Assuming a Gaussian distribution of the intensities and setting the number of standard deviations from the mean to 2.5, we retain 0.65% of the voxels (one-tail, highest values), to capture just a few reliable outliers.

These seeds were used to create the initial contours for the fast marching level sets. Several seeds could be placed at a given fuzzy boundary, but the level sets will expand, evolving into complex shapes and merging since the intensity gradient is smooth. However, the level sets placed on the fuzzy boundaries do not merge with those placed on the lesion areas, as can be observed in Fig. 2.4 (c), where the intensity gradient was too large.

Several coarse level sets were obtained as output. These were used as initial contours ($x_0$) for the geodesic active contour algorithm [41]. Namely, a contour was fitted to the region ruled by the following partial differential equation (PDE):

$$\frac{\partial \Psi}{\partial t} = -\alpha \mathbf{A}(\mathbf{x}) \cdot \nabla \Psi - \beta P(\mathbf{x}) |\nabla \Psi| + \gamma Z(\mathbf{x}) \kappa |\nabla \Psi|, \tag{2.8}$$

where $\Psi$ is the level set, $\mathbf{x}$ is a point of the contour, $\mathbf{A}(\mathbf{x})$ controls the advection, $P(\mathbf{x})$ is the propagation and $Z(\mathbf{x})$ is the spatial modification of the mean curvature $\kappa$; $\alpha$,$\beta$ and $\gamma$ are scalars which module each term of the contour evolution. Their value was heuristically set to $\alpha = 1.0$, $\beta = 0.25$, $\gamma = 2.0$. The outputs were refined level-set contours for both the lesions and the fuzzy boundaries. Once the contours were determined, lesions were discriminated from artifacts based on the prior morphological information: contours with a sphericity over 0.85 were selected as lesions and included within the segmented lung (see Fig. 2.4 (d)).

### 2.2.3 Lung Segmentation Evaluation

The quality of the automatic segmentation for medical imaging applications is commonly estimated with respect to a manually or semi-automatically generated ground truth. The most commonly used evaluation measures are computed as an average of the intersected volumes between both segmentations (i.e., Dice Similarity Coefficient (DSC)) [203, 227]. For our application, suitable values of the measures could be misleading [259] if relatively small volumes at the fuzzy boundaries (i.e., lesions, respiratory motion artefacts) are incorrectly segmented. In those cases, the perceived decrease in quality given by the measure will be minor, but these errors in lung segmentation would generate considerable bias in the subsequent quantification of disease burden.

To mitigate this issue in evaluating the goodness of the proposed lung segmentation method, we use the procedure described below to select the slices that most probably have fuzzy boundaries. Rough segmentations of the lungs were semi-automatically computed in 63 subjects using an in-house platform [236] created explicitly for the interactive segmentation of TB-infected lungs. To segment the lungs using the platform, the user specifies at least 1 seed in the centre of the left lung and right lung. The segmentation then propagates employing a region-growing algorithm. The user can manually specify frontier surfaces to prevent the segmentation from reaching adjacent air-filled regions. The platform has added functionalities to enable manual correction of the results. Once the lungs are interactively segmented, the Hausdorff distances between the automatic lung segmentation obtained before and after the refinement step with respect to the semi-automatic segmentations are computed. The differences in the Hausdorff distances are due to the corrections performed by the refinement routine. The differences point out to those slices in which the segmentation is more uncertain due to the variability introduced by each subject and the disease course. We then choose the 156 slices with the most considerable differences in the Hausdorff distance to build a surrogate ground truth, as described in detail in Appendix A.1.

Three experts interactively segmented the selected slices, paying particular attention to the boundary delimitation. The very accurate segmentations obtained were then combined by consensus to provide a surrogate ground truth [328]. Characterization of the agreement, computing the intra-class correlation coefficient (ICC), between the lung segmentation performed by the experts showed excellent consistency (details can be found in Appendix A.2).

The individual expert segmentations and the surrogate ground truth are compared with the proposed method (refined -Ref-) and two other approaches intended for healthy or slightly damaged lung segmentation. Namely, the aforementioned manual segmentation (referred to as semi-auto -Semi-) and the traditional fuzzy connectedness–based lung segmentation (referred to as FC), which has a publicly available open-source software lung segmentation tool (http://www.nitrc.org/projects/nihlungseg/) [202]. For the latter, we used the best performing manual seeding mode, as recommended by the authors, for refining segmented region maps, namely, filling holes with a 0.44 mm-diameter binary filter and checking fuzzy connectedness.

The similarity is measured as both volume overlap and distance between surfaces with the following metrics: Dice similarity coefficient ($DSC$), Hausdorff distance ($HD$), Hausdorff distance averaged ($HDA$), false-positive error ($FPE$), false-negative error ($FNE$) and volume dissimilarity ($VD$). The $HD$ and $HDA$ measures are indicators of a given method's ability to delineate the tissue boundaries. The $FPE$, $FNE$ and $VD$ indexes provide additional information for the volume overlap measured by the $DSC$. In particular, $FPE$ is related to over-segmentation, $FNE$ to under-segmentation and $VD$, evidently, to volume differences.

To better understand the measures dispersion, box plot charts for each similarity index are also obtained. The dispersion characterization of the similarity indexes is particularly interesting in our case, owing to the complexity of the dataset used. We refer to each comparison between a method and the surrogate ground truth for a given similarity index specifying the method as sub-index (e.g., $DSC_{Ref}$. refers to the median DSC of the comparison between the refined segmentation and the surrogate ground truth).

Finally, we studied the statistical significance of our results to assure the objectivity of our conclusions. For each evaluation metric and each reference segmentation, the outputs of the three segmentation methods were compared using a paired t-test. A $p$ value below 0.05 was considered statistically significant.

## 2.2.4 Results

### 2.2.4.1 Qualitative Results

Fig. 2.5 illustrates the computed lung segmentation on a representative slice from those retained (i.e., those in which the segmentation is most uncertain). The segmentations corresponding to the semi-automatic approach (panel c) are subject to over-segmentation: the delimitation of the lungs goes beyond the lung parenchyma, including respiratory movement artefacts. As per the FC approach (panel d), we observed that several lesions, independently of their localization, were not included in the segmentation due to the method's lack of sensitivity to those areas. The amount of over- and under-segmentation (highlighted in red and yellow, respectively) caused by the proposed method was reduced with respect to the other two approaches.

### 2.2.4.2 Quantitative Results

Fig. 2.6 shows the box plot charts for each similarity index of the refined (Ref), the semi-automatic (Semi) and the fuzzy connectedness lung segmentation (FC) against the manual annotations performed by each expert (Exp. #) and the consensus surrogate ground truth (Maj.). The numerical results are provided in Table 2.1. The refined segmentation provides the most similar results with respect to the experts' delimitation and, thus, with respect to the surrogate ground truth. In this sense, the proposed method achieves the largest volume overlap, as reflected by the $DSC$ (mean $DSC_{Ref} = 0.933$; median $DSC_{Ref} = 0.943$). The second best-performing method, the FC, which was intended for the segmentation of slightly infected lungs, presents a close mean $DSC$ (mean $DSC_{FC} = 0.926$) but more distant median $DSC$ (median $DSC_{FC} = 0.922$). Our method achieves much lower distances ($HD$ and $HDA$) with respect to the surfaces of the surrogate ground truth than the others (between

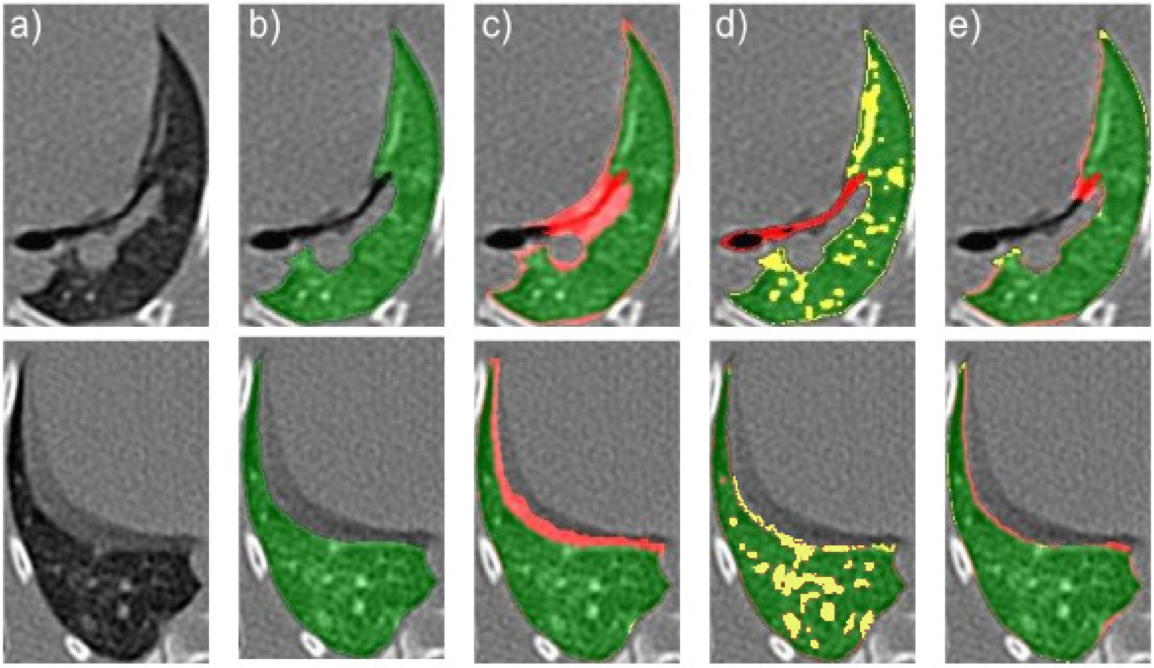


Fig. 2.5 Sample lung segmentations on a representative slice (a) corresponding with the surrogate ground truth (b), the semi-automatic segmentation (c), the fuzzy connectedness segmentation (d), and our proposed method (e). The regions in which there is overlap with the surrogate ground truth are colored in green, the false-positive errors in red and the false-negative errors in yellow.

1.2 and 5.1 mm with respect to the median ($HD_{Ref} = 5.537$ mm) and between 2.8 and 11 mm with respect to the average value ($HD_{Ref} = 8.642$ mm)). The method presents similar rates of under- and over-segmentation, around 6%. In contrast, the semi-auto approach achieve a larger over-segmentation rate (median $FPE_{Semi} = 15\%$, mean $FPE_{Semi}$ =16%) but a much smaller under-segmentation rate (median $FNE_{Semi} = 0.2\%$, mean = 0.6%) while the FC method provide the opposite results (mean $FPE_{FC} = 2.4\%$, median $FPE_{FC} = 2.2\%$, mean $FNE_{FC} = 11\%$ and median $FPE_{FC} = 10.4\%$). These imbalances make the differences between the volumes obtained by the experts (consensus) and those obtained with the semi-automatic and the fuzzy connectedness methods much higher than those measured for our approach. The volume dissimilarity index for the latter is close to zero in all cases (mean $VD_{Ref} = 0.026$, median $VD_{Ref} = -0.0009$). All the differences as illustrated in Fig. 2.6, are statistically significant except for the *HDA* index on the Refined and FC segmentations when Expert 2 is used as reference.

Fig. 2.7 displays *DSC*, *HD* and *HDA* plots over the slices arranged in ascending order as given by the *DSC* of the semi-automatic segmentation with respect to the surrogate ground truth. The data have been filtered following the locally weighted scatterplot smoothing

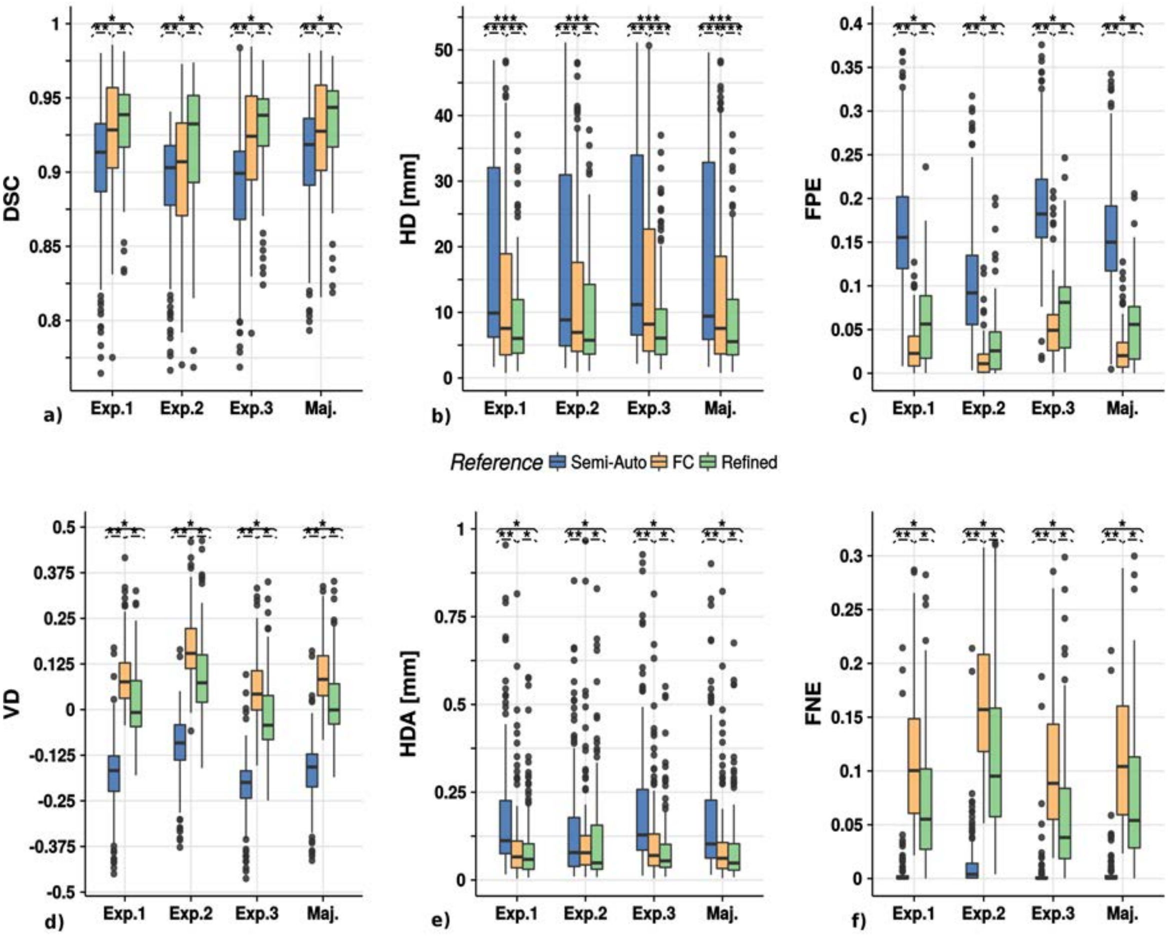


Fig. 2.6 Boxplot charts for the similarity indexes: (a) Dice Similarity Coefficient (*DSC*); (b) Hausdorff Distance (*HD*); (c) False Positive Error (*FPE*); (d) Volume Dissimilarity (VD); (e) Hausdorff Distance Averaged (*HDA*); (f) False Negative Error (*FNE*). The lung segmentation obtained with the proposed method (refined) is compared with the semi-automatic (semi-auto) and the fuzzy connectedness approaches in the individual expert annotations (Exp. 1, Exp. 2 and Exp. 3) and the surrogate ground truth obtained by the expert consensus as explained in the A.1 (Selecting CT Slices With The More Uncertain Boundaries). The asterisks over each group of boxes indicate statistically significant differences between the lung segmentation methods compared: $p < 0.05 \equiv *$, $p < 0.01 \equiv **$ and $p < 0.001 \equiv ***$.

(LOESS) [55] model in order to achieve a better appreciation of the patterns and the differences between the approaches. The *DSC* plot shows that the gap between the proposed and the semi-automatic method (about 10% for the first slice) decreases as we move towards higher *DSC* slice values, while the difference with the *FC* method remains relatively stable. The *HD* index corresponding to the proposed method is smaller than the other methods for

| Expert | Comparison | $\overline{DSC} \pm \sigma_{DSC}$ | $\overline{HD} \pm \sigma_{HD}$ | $\overline{HDA} \pm \sigma_{HDA}$ | $\overline{FPE} \pm \sigma_{FPE}$ | $\overline{FNE} \pm \sigma_{FNE}$ | $\overline{VD} \pm \sigma_{Vol.Dis}$ |
|---|---|---|---|---|---|---|---|
| Exp. 1 | Semi-Auto | 0.904 ± 0.04 | 18.741 ± 14.78 | 0.206 ± 0.25 | 0.167 ± 0.07 | 0.007 ± 0.03 | -0.179 ± 0.10 |
| | FC | 0.926 ± 0.04 | 12.768 ± 12.79 | 0.105 ± 0.14 | 0.028 ± 0.03 | 0.112 ± 0.07 | 0.092 ± 0.08 |
| | Refined | 0.931 ± 0.03 | 8.801 ± 7.37 | 0.093 ± 0.11 | 0.059 ± 0.04 | 0.074 ± 0.06 | 0.017 ± 0.10 |
| Exp. 2 | Semi-Auto | 0.891 ± 0.04 | 17.448 ± 14.66 | 0.159 ± 0.23 | 0.106 ± 0.07 | 0.013 ± 0.03 | -0.101 ± 0.09 |
| | FC | 0.901 ± 0.04 | 12.346 ± 12.24 | 0.111 ± 0.13 | 0.015 ± 0.02 | 0.167 ± 0.07 | 0.170 ± 0.09 |
| | Refined | 0.920 ± 0.04 | 9.576 ± 7.95 | 0.117 ± 0.15 | 0.033 ± 0.04 | 0.118 ± 0.08 | 0.095 ± 0.11 |
| Exp. 3 | Semi-Auto | 0.889 ± 0.04 | 19.835 ± 15.19 | 0.235 ± 0.27 | 0.193 ± 0.07 | 0.005 ± 0.02 | -0.212 ± 0.09 |
| | FC | 0.919 ± 0.04 | 13.993 ± 13.44 | 0.116 ± 0.15 | 0.052 ± 0.04 | 0.105 ± 0.06 | 0.059 ± 0.09 |
| | Refined | 0.931 ± 0.03 | 8.825 ± 7.43 | 0.089 ± 0.09 | 0.059 ± 0.06 | 0.075 ± 0.05 | -0.016 ± 0.10 |
| Maj. | Semi-Auto | 0.909 ± 0.04 | 18.674 ± 14.81 | 0.199 ± 0.25 | 0.16 ± 0.07 | **0.006 ± 0.02** | -0.171 ± 0.09 |
| | FC | 0.926 ± 0.04 | 12.786 ± 12.85 | 0.103 ± 0.14 | **0.024 ± 0.02** | 0.116 ± 0.06 | 0.101 ± 0.08 |
| | Refined | **0.933 ± 0.03** | **8.642 ± 7.36** | **0.091 ± 0.11** | 0.054 ± 0.04 | 0.077 ± 0.06 | **0.026 ± 0.09** |

Table 2.1 Overall performance of the refined, the semi-automatic and the fuzzy connectedness (FC) lung segmentation against the manual annotations made by each expert (Exp. 1, Exp. 2 and Exp. 3) and the consensus surrogate ground truth (Maj.). Mean, and standard deviation are provided for each index. For the surrogate ground truth, the best performing method is highlighted in bold for each index. Note: Dice similarity coefficient (DSC), Hausdorff distance (HD), Hausdorff distance averaged (HDA), false-positive error (FPE), false-negative error (FNE) and volume dissimilarity (VD).

all the slices. The improvement is $5-10$ mm with respect to the Semi-Automatic approach and $0.5-7.5$ mm with respect to the FC approach. Finally, the $HDA$ index exhibits an exponential decay for all the methods.

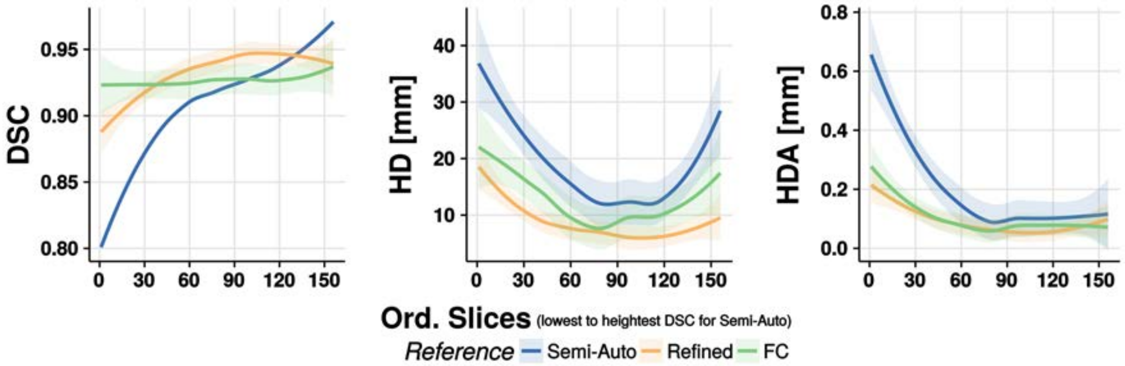


Fig. 2.7 Dice similarity coefficient (DSC), Hausdorff distance (HD), and Hausdorff distance averaged (HDA) plots along the slices sorted in ascending order based on the DSC of the semi-automatic segmentation with respect to the surrogate ground truth. Data have been filtered with the locally weighted scatterplot smoothing (LOESS) model. The 95% confidence interval is drawn as a shadow of the same colour as the corresponding line.

### 2.2.5 Discussion

The experiments performed reveal substantial improvements when an input volume is processed through our pipeline. As expected from a method focused on improving boundary detection, the Hausdorff distance is significantly smaller than other methods while presenting reasonably good results for the volume overlap measures. This behaviour is explained by the ability to reject fuzzy boundary artefacts while retaining most of the damaged tissue (especially the lesions attached to the pleura). Since the Hausdorff distance computes the maximum among the minimal distances for all points in the two surfaces compared, small changes when delimiting a complex shape (such as those generated by the diseased lung) result in large Hausdorff distance values. Fortunately, the boundaries created by our method are consistent and stable, and inaccuracies in the boundary delimitation are less frequent. Moreover, improved delimitation enables the target volume to be filled more accurately, as reflected in the DSC values.

For the macaque model context, where the lung segmentation is a preparatory step for quantifying the TB lesions burden during the disease, these small differences are vital. Especially, for relapse models in which small lesions are regularly the most often ones. High-quality segmentation is critical in the early stages. The sensitivity given by the radiological images is crucial when assessing latent tuberculosis due to the small parenchymal damage associated with this stage of the disease. Therefore, the fact that the advanced method achieves the lowest Hausdorff distance measured by far in almost all the slices (Fig. 2.7) is a major step towards the proper quantification of disease burden, even with the current dispersion of the measure. This dispersion is mostly due to the intrinsic noise inherent in the delineation of complex slices. Thus, it is likely to appear in any segmentation method, including manual delineations [313]. In the Section A.2, the inter-agreement differences between the experts' delimitation are presented. They show a good intra-class correlation coefficient (ICC) for the overall surface delimitation ($HD = 0.88$, $HDA = 0.85$) and lower values for the volume indicators of performance ($DSC = 0.74$, $FPE = 0.71$ and $FNE = 0.6$). The fact that small variations in delineation produce large dissimilarity values is even more obvious for the Hausdorff distance averaged. Although, as observed in Fig. 2.6, the values of this measure are much smaller than the Hausdorff distance. Many outliers are present due to the relatively large distance between the surfaces corresponding to pairs of compared segmentations at several slices within the data set.

The more conservative segmentations are those provided by the fuzzy connectedness–based method and our proposal. They perform better in terms of $HD$ and $HDA$ (Fig. 2.5). Hence, the segmentations they provided are more suitable for subsequent quantification of the TB lesion burden.

It is important to emphasize that our method achieves a good balance between false positive and false negative errors, in contrast to the semi-auto segmentation results, which show, on average, 15% over-segmentation. The lung segmentation includes fuzzy regions, which will contaminate the subsequent analysis. In contrast, the FC segmentation is excessively conservative. It presents a tiny percentage of over-segmentation and 15% false-negative errors on average for the most uncertain slices in the dataset. Thus, it potentially generates a misleading evaluation of TB infection. Although the advanced method balances out possible errors, it still exhibits 5% false negative errors on average, which could still influence the quantification of disease burden, although less severely than the FC method.

The information from the error types makes it possible to explain the volume dissimilarities shown in Fig. 2.6. The semi-auto method presents the previously mentioned problems of over-segmentation, which account for the almost parabolic shape of the HD when the DSC increases in Fig. 2.7. The method presents a limit (at the parabola vertex), from where the segmentation is unable to fill the region of interest without growing beyond. Thus, the method presents a few slices with better overlap (DSC) than the proposed approach at the expense of losing sensitivity at the boundaries. Consequently, the HD remains flat, between the 90th and the 120th slice, only to increase dramatically afterwards, while a suitable segmentation should decrease or, at least, keep a constant low distance. The HD plot for the FC method in Fig. 2.7 presents similar behaviour to the semi-auto method, albeit for different reasons. As illustrated with the examples in Fig. 2.5, the FC method misses an important part of the volume-of-interest, resulting in considerable volume dissimilarity (see Fig. 2.6). Although the DSC trend in Fig. 2.7 is flatter than the one corresponding to the semi-auto method, it also presents a parabola vertex, which indicates an inability to capture the intricate shape of the selected surrogate ground truth. In contrast, the refined method shows a negligible value of volume dissimilarity (see Fig. 2.6) and a much less marked parabola shape (see Fig. 2.7). To further improve the accuracy of the lung segmentation, it could be much more appropriate to use novel indicators of segmentation performance more closely associated with the ulterior quantification than the overlap and surface indicators. They are clearly of limited validity owing to the variability of human criteria during the segmentation process [259]. To this aim, we have introduced a quantification method, presented in the subsequent section Quantifying trough correlation in a closed environment, which makes use of the proposed pipeline for lung segmentation and that presents satisfactory results [94]

The framework allows re-parametrization to other models (e.g., mice, humans) by fine-tuning of the parameters as shown in Table 2.2. As was mentioned at the beginning of the chapter, this hyper-tuning process would not probably achieve the best possible results given

Table 2.2 Pipeline process (first column), algorithm (second column), parameters (third column) and their values (fourth column).

| Section | | Parameter | Value |
|---|---|---|---|
| Preliminary Lung Segmentation | Adaptive Thresholding | Otsu Threshold | Auto. |
| | Rib Cage Extraction | Seeds | $> 900$ HU |
| | Connectivity and Topological Analysis | Min. Object size | 10 $mm^3$ |
| Airway Tree Segmentation | Trachea detection | Expected Perimeter | $5.5 - 8.5$ $mm$ |
| | | Roundness | $> 0.9$ |
| | Wavefront Propagation | *Time Step* | 0.8 |
| | | $T_i$ | $-625$ HU |
| | | $T_s$ | 2.5 |
| | | $\alpha$ | 1.4 |
| | Bifurcation Detection | $\beta$ | 2 |
| | Leakage Detection | $T_{GR}$ | 1 |
| | | $T_C$ | 0.72 |
| | | $T_W$ | 10% |
| Closing and Fuzzy Boundaries | Morphological 3D Hole Filling | *Kernel Radius* | 1 $mm$ |
| | Fuzzy Lung Border Segmentation | $\alpha$ | 1.0 |
| | | $\beta$ | 0.25 |
| | | $\gamma$ | 2 |
| | | sphericity | $> 0.85$ |

the SOTA DL methods available (see Rule-Based Methods Under Unseen Domains & Lack of Generalization). Nevertheless, the proper understanding of each parameter function provides instrumental knowledge for prospective segmentation models. Thus, to improve the results and extend the framework to the segmentation of extremely damaged lungs, within this thesis (see chapters Deep Learning for TB Manifestation Classification and Translational Lung Imaging Analysis Through Disentangled Representations), AI/DL techniques are examined and developed. These implementations shown promising results for segmentation, besides further developments are introduced to cope with common limitations for DL models like the loss of resolution [108]. These limitations impede the proper identification of boundaries, biases (Section 1.4.1), and results in the need for a large refined ground truth [266], which results quite challenging to obtain. To get the best labels (segmentations) to train the DL models introduced in the thesis, the unsupervised segmentation volumes obtained with the tool presented in this chapter are reviewed and corrected when necessary by experts. Sustantially minimizing in this way the time invested by them to create a good Ground Truth, especially in comparison with the use of the tools mentioned above (see Section 2.2.3).

## 2.3 Quantifying trough correlation in a closed environment

As mentioned several times before, automatic segmentation is a big step [84]. However, the final goal is the implementation of techniques for the characterization of TB as a continuous spectrum employing radiological images given its sensitivity for findings TB manifestations [217, 235] (see Fig. 1.5). Therefore, identifying biomarkers to ease the radiologists' evaluation of TB in extensive studies needs to be automatized. However, the few methods dealing with TB damaged lungs do not contemplate quantification or are not automatic [48, 323–325].

To ease this task, within this section, our first approximation of a complete methodology to automatically extract biomarkers from CT images is presented (see Section 1.4.2). Although with limitations, it is able to estimate the evolution of TB burden and could be used to assess the response to treatment of infected subjects when there is a causal relationship between the damaged lung tissue and TB burden [48]. This scenario could be customary in several animal models, as was already pointed out in the our published work *Computed Tomography-Based Biomarker for Longitudinal Assessment of Disease Burden in Pulmonary Tuberculosis* [94]. Most of the following sections were extracted from that manuscript and which `R` and `Python` based code can be found in https://github.com/BIIG-UC3M/TLS-Piped.

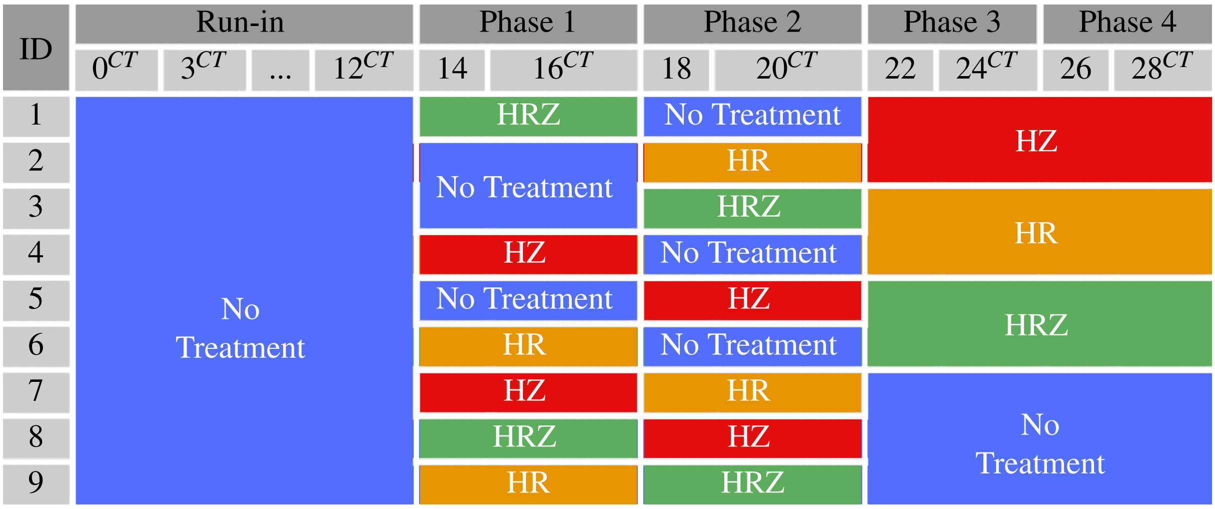
### 2.3.1 Materials

#### 2.3.1.1 Computer Tomography Images

The CT scans are already described in CT Imaging. It is important to remember that the macaques were treated with a different antibiotic cocktail of Isoniazid (H), Rifampicin (R), and Pyrazinamide (Z) [285] in four phases, as is shown in the Table 2.3.
All animal procedures and study designs were approved by the Public Health England Animal Welfare and Ethical Review Body, Porton Down, UK, and authorized under an appropriate UK Home Office project license.

### 2.3.2 Lungs Segmentation

The entire procedure is detailed in Section 2.2.2, therefore, if the reader is familiar with it can skip to Section 2.3.3 at this point. For those readers primarily interested in quantification, the previous segmentation step is illustrated in the left part of Fig. 2.8 and summarized in the following paragraph.



| ID | Run-in | | | | Phase 1 | | Phase 2 | | Phase 3 | | Phase 4 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $0^{CT}$ | $3^{CT}$ | ... | $12^{CT}$ | 14 | $16^{CT}$ | 18 | $20^{CT}$ | 22 | $24^{CT}$ | 26 | $28^{CT}$ |
| 1 | No Treatment | | | | HRZ | | No Treatment | | HZ | | | |
| 2 | | | | | No Treatment | | HR | | | | | |
| 3 | | | | | | | HRZ | | HR | | | |
| 4 | | | | | HZ | | No Treatment | | | | | |
| 5 | | | | | No Treatment | | HZ | | HRZ | | | |
| 6 | | | | | HR | | No Treatment | | | | | |
| 7 | | | | | HZ | | HR | | No Treatment | | | |
| 8 | | | | | HRZ | | HZ | | | | | |
| 9 | | | | | HR | | HRZ | | | | | |

Table 2.3 Antibiotic cocktail per week and subject. Each color represents the treatment (No Treatment, HR (*Isoniazid* + *Rifampicin*), HRZ (*Isoniazid* + *Rifampicin* + *Pyrazinamide*), HZ (*Isoniazid* + *Pyrazinamide*)) taken by a subject during each treatment phase. Weeks with the CT superindex indicate the acquisition of a computed tomography volume at that week.

Initially, air-like organs (e.g., healthy lungs, airways tree, stomach) presented in the chest CT scans (Fig. 2.8.a) are identified employing an adaptive thresholding method (Fig. 2.8.b) to subsequently isolate the object formed by the lungs and airways studying the topology and connectivity of the organs (Fig. 2.8.c). Next, the intricate airways tree structure is computed employing a region growing algorithm which propagates simulating a spherical wavefront ruled by active contours [43] (Fig. 2.8.d) and is removed from the segmented lungs. Finally, unsegmented pulmonary regions, corresponding to damaged parenchyma and TB lesion are included by a morphological hole filling process [144] which is refined using *Geodesic Active Contours* [41] to segment the most uncertain regions in the lungs boundary to include discarded lesions and expel previously included artefacts (Fig. 2.8.e).

### 2.3.3 Computer Tomography Biomarker Extraction

In order to automatically retrieve quantifiable information as a CT biomarker, the proposed method is inspired in  Chen et al. [48] work. Within *Chen's* work, the tissue belonging to the lungs is divided into three disease-associated volumes manually. This division depends on the grey level intensity of the voxels, measured employing Hounsfield Units (HU), and two thresholds selected by experts, which establish three regions in the lungs histogram corresponding with the following kind of tissues.

- *Healthy Tissue*, which corresponds to the voxels with the lower intensities in the lungs and free of TB
- *Soft Tissue*, which match with voxels found in lower density of abnormal tissue, corresponding with forming or healing lesions.
- *Hard Tissue*, corresponding to intensity values in high density abnormal tissue.

In other words, *Chen's* approach assigns a discrete class (*healthy*, *soft* or *hard*) to a range of values distributed around an expected intensity given a variability that captures the subtle differences in the composition of each kind of tissue. The expected intensity value and variability of each class are intrinsically determined via the selection of thresholds by the experts. Fortunately, for the problem domain, this empiric approach can be computationally modelled employing the well known *Gaussian Mixture Model* (GMM) and the more likelihood volumes separation obtained through the Expectation-Maximization (EM) algorithm [28]. The GMM model allows us to represent a known histogram, like the one belonging to segmented lungs (see Fig. 2.8 right part), as a probability distribution composed of several overlapped Gaussian variables, in our case, the distribution of each kind of tissue.
Generally speaking is formulated as:

$$p(\mathbf{x}) = \sum_{i=1}^{K} \pi_k \mathscr{N}(\mu_k, \Sigma_k), \tag{2.9}$$

where $\mathbf{x}$ is a vector of observed features (the intensity values of each voxel represented in the histogram), $K$ is the number of expected Gaussians ($K = 3$ corresponding to *healthy*, *soft* and *hard* tissues), and $\mathscr{N}(\cdot)$ represent each one of the overlapped normal distributions ($k$) of the voxels grey level, being: $\pi_k$, the *a priori* probability; $\mu_k$, the mean; and $\Sigma_k$ the covariance, respectively, for each distribution. These parameters are computed employing the EM algorithm, selecting those that set the Gaussians which overlapping is most similar to the known histogram (Eq. 2.9). This way, each voxel is assigned to a lung tissue depending on which of the fitted Gaussians provides the biggest probability for a given voxel intensity.

#### 2.3.3.1 Gold Standard Computer Tomography Biomarker

To measure the performance of the proposed automatic biomarker extraction method, the *soft* and *hard* volumes aforementioned were manually extracted by an expert from the original 63 CT scans comprised within the dataset.

### 2.3.4 Evaluation Methods

In order to provide a more general biomarker of the *Mtb* burden, besides the volumes defined in the previous section, we include the total volume of diseased tissue for the comparison between the proposed method and the gold standard, the volume is defined as:

$$Diseased\ \ Vol. = Soft\ \ Vol. + Hard\ \ Vol. \tag{2.10}$$

Additionally, to avoid the effects of the changes in the whole lung volume due to the subjects' growth during the 28 weeks, the diseased volume is normalized to study the longitudinal disease behaviour. This volume is easily defined as follows:

$$Relative\ Diseased\ \ Vol. = \frac{Diseased\ \ Vol.}{Healthy\ \ Vol.} \tag{2.11}$$

To evaluate the longitudinal change, we employ a multirow bar plot, known as waterfall (i.e., Fig. 2.9), in which the first row shows a relative volume at each subject in baseline time point and the rest of them the change in the volume at a concrete time point in a $log_2$ scale, therefore, the subject change is computed as:

$$change\ in\ vol. = \log_2\left(\frac{Vol.\ at\ week\ of\ change}{Vol.\ at\ baseline}\right) \tag{2.12}$$

### 2.3.5 Results

Fig. 2.9 depicts the longitudinal change of the *Relative diseased volume* through a waterfall plot for the nine subjects (horizontal axis) in four of the seven-time points for the shake of clarity. Concretely, the first row contains the relative volume of diseased tissue at week three after infection, while the rest represents the $log_2$ change (see Section 2.3.4) at weeks $16, 20$ and 28 with respect to the first row, the baseline. Beyond meaningful quantitative differences between equal treatments, it can be observed how subjects under the same drug cocktail (each treatment is shown with a different colour) at the end of the study (week 28) present a similar response to treatment to the baseline.

Fig. 2.10 shows the diseased volume obtained employing the manual delimitation of regions against the volume obtained by the proposed automatic extraction method at each one of the 63 segmented lungs in the dataset (see Section 2.3.1.1) together with the corresponding Bland-Altman plot in order to show the agreement between methods. The similarity between measures results is primarily independent of the subject, treatment and study time point

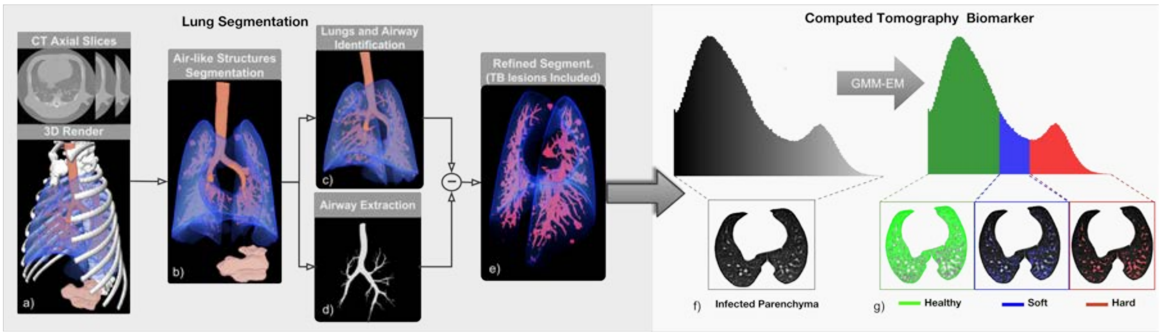

Lung Segmentation
CT Axial Slices
3D Render
a)
Air-like Structures Segmentation
b)
Lungs and Airway Identification
c)
Airway Extraction
d)
Refined Segment. (TB lesions Included)
e)
Computed Tomography Biomarker
GMM-EM
f)
Infected Parenchyma
g)
Healthy
Soft
Hard

Fig. 2.8 *Lung Segmentation:* From a chest CT scan (a) all aerated regions (b) are identified using an specific adapted algorithm [131]. Successively, lungs are isolated, including the airways, by exploiting the topological information (c). The intricate structure formed by the airways (d) is extracted and eliminated from the lungs. Finally, to include the non-segmented damaged tissue, the lung segmentation is refined employing a hole-filling method based on active contours [41](e). *CT Biomarkers:* The segmented lungs are separated into three TB-associated volumes as proposed by *Chen et al.* [48]. Each region comprises a grey-level range in the segmented lungs histogram (f) representing: Healthy Tissue: Parenchyma free of infection, Soft Tissue: Forming or healing lesions, Hard Tissue: Abnormal lung parenchyma (g).

(none of these factors shows a remarkable bias). The correlation coefficient was R $\approx$ 0.8 ($p < 1 \times 10^{-4}$), with a tendency to obtain higher values for the volumes obtained automatically. The Bland-Altman plot depicts all the values within the 95 % limits of agreement.

## 2.3.6 Discussion

The results exhibit how the automatic biomarker extraction can provide good results by statistical modelling of the decision-making process carried out by an expert. Concretely, as an analogy between the experts' work and the proposed method, we can assert that our approach automatically assigns the thresholds established deterministically by the specialists. The method can fairly assess the longitudinal evolution of TB by showing significant similarities in the treatment response, as shown in Fig. 2.9. As per the experimental design with a combination of antibiotics and as expected, differences are noticeable at the end-point (week 28). Besides, there is a correspondence between the results obtained automatically and the manual ones, as depicted by the $R^2 = 0.8$. This relation is biased by a factor of 0.47 favouring the volumes obtained with the proposed method. Two causes can mainly explain these differences: a) The difficulty presented in the manual delimitation of complex three-dimensional structures, many times intricate in the healthy tissue (see Fig. 2.1), prevents from segmenting

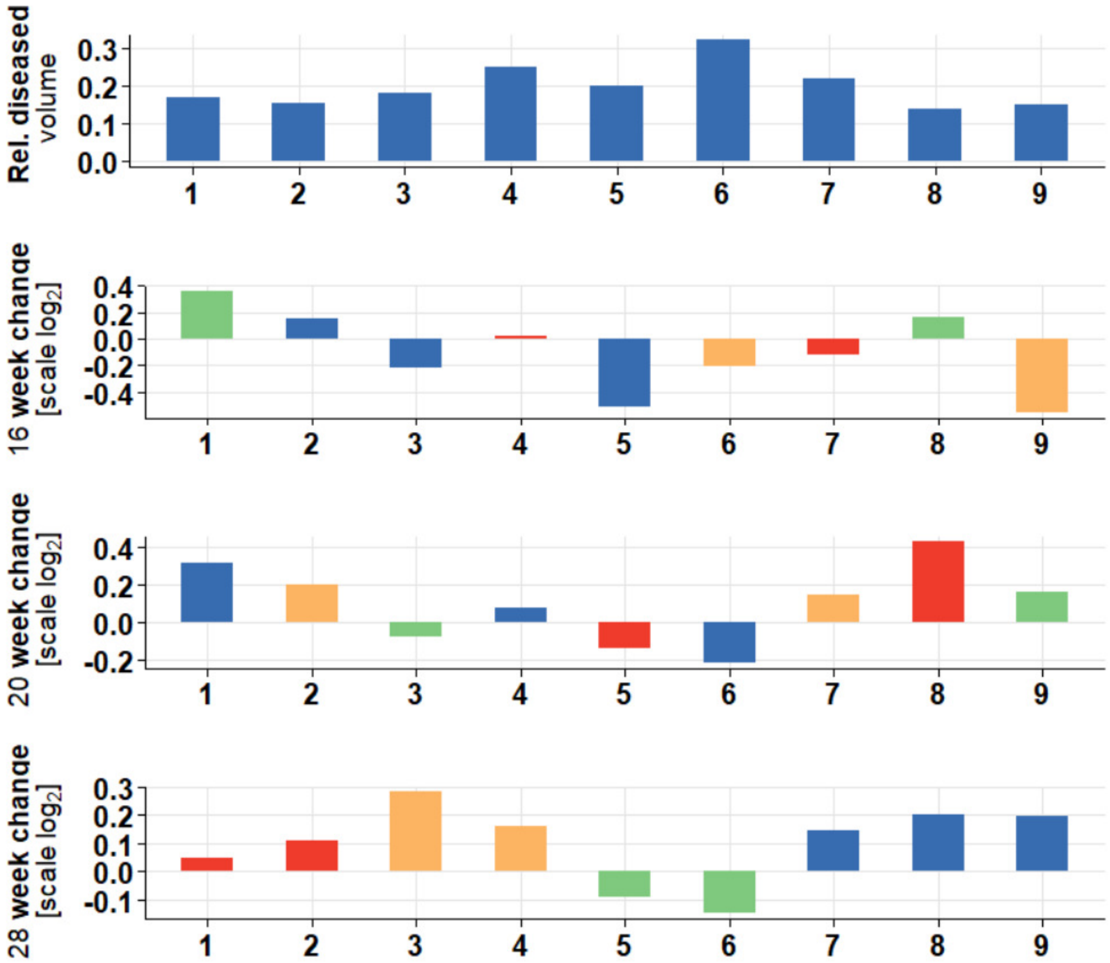


Fig. 2.9 Longitudinal evolution of TB infection: The waterfall plot, depicts the longitudinal change for diseased volume $(soft + hardvols.)$ between the baseline week and the representative weeks as the $log_2$ fold change $(log_2(week\ change/baseline))$ and the correlation with treatments outcomes: 🟦 No-D (No Drugs), 🟧 HR $(Isoniazid + Rifampicin)$, 🟩 HRZ $(Isoniazid + Rifampicin + Pyrazinamide)$, 🟥 HZ $(Isoniazid + Pyrazinamide)$. Subjects with the same treatment in the final phase (week 28) present similar response $(diseased = hard + soft)$ with respect to the baseline

the whole region of interest which results in smaller volumes; b) The automatic extraction of the biomarker tend to include the unsegmented (in the lungs segmentation step) small vessels as diseased tissue, and therefore increase the obtained volumes. The inclusion of vessels as damaged lung tissue is undesirable. However, this side-effect is reduced by employing the normalized relative volume and assuming that the extra volume produced by the inclusion of vessels remains constant over time. Thus, the effect is not meaningful evaluating changes, as the trends presented in Fig. 2.9 seem to indicate. It is also essential to note that the proposed

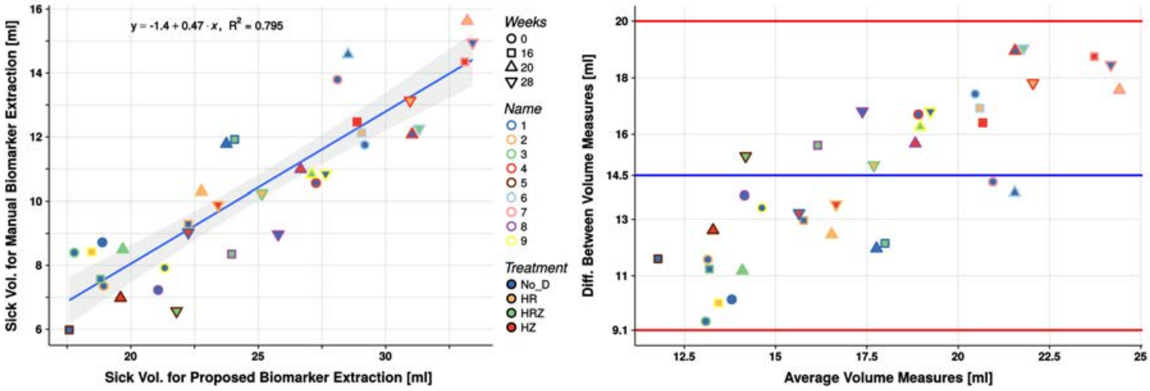


Fig. 2.10 (Left) Biomarker evaluation: Correlation between the manual biomarker(relative diseased volume) and the proposed method ($R^2 \approx 0.8$, $p < 10^{-4}$). The 95% confidence interval is drawn as the shadow of the regression line. (Right) The Bland- Altman plot presents a good agreement between measures with the regression bias of 0.47

method is mainly intended to capture differences in the infection burden for animal models in which the subjects are not at the final stages of the disease. Namely, to establish a continuous spectrum between latent and active disease (Fig. 1.5). Because of this, extra-large cavities (the manifestations proper of tissue destruction where the *Mtb* is not present anymore and the drugs cannot be effective) (see Fig. 1.6e) are not included as damaged tissue, which can provoke small drifts in the data correlation corresponding with very particular high infected subjects. Some of the commented limitations are addressed in the next chapter, Radiomics for TB Manifestations Classification, taking advantage of the Radiomics techniques [172], or in subsequent chapters by injecting part of qualitative information usually employed by the experts, and showed in this first approach, in DL models.

## 2.3.7 Conclusion

In this section, we introduce a complete methodology for the extraction of a biomarker to characterize the gradual change of *Mtb.* infection. The proposed technique yields similar results to the ones obtained manually by a trained specialist. These facts highlight the capability of the method as a quantification tool in clinical assays devoted to design effective drugs against TB. Limitations of the framework, mainly caused by the lack of generalization, are shown in the following Section 2.4.

## 2.4 Rule-Based Methods Under Unseen Domains & Lack of Generalization

We have referred before to the performance decay of traditional automation methods, especially when these receive inputs out-of-domain to those used during their design but whose similarity, however, allows experts to analyze them similarly. This section gives examples of the limited capability of rule-based algorithms to transfer knowledge to new domains and obtain acceptable lung mask delimitations. Thus, the first column of the Fig. 2.11 shows chest CT axial slices corresponding to different mammals and diseases. Namely, the slice in the first row corresponds to the macaque model infected with mild TB as described in Section 2.2.1 (same domain, $P^{PHE_1}$). The slice at the second row also corresponds to the same macaque model but a different cohort ($P^{PHE_2}$). Such cohort models a much active TB infection (see Section 3.2.1) [285]. As can be seen, the lesions, in this case, differ (lung vanishes), which supposes a domain shift. The third image belongs to a dataset of a mouse infected by TB, $M^{GSK}$, (courtesy of *GlaxoSmithKline* (GSK) a ERA4TB partner [76], Section Project Framework: ERA4TB). This example shows a domain shift due to the use of a different animal (i.e., different TB manifestations, change of CT scanner). The last two slices belong to images of human lungs extracted from publicly available clinical datasets [57, 68]. TB is the pathogen for $H^{CLE}$, while $H^{RAD}$ belongs to COVID-infected lungs, thus illustrating the limitations of classical methods in clinical practice [142, 309].

Ideally, automation mechanisms should perform segmentation on domain-shifted lung images exploiting information extracted from the dataset employed during design/learning. This fact holds while such information is similar for all datasets, in the same way, as experts can segment images of different animal models and diseases learning from particular datasets (Ground truth, GT).

However, as illustrated in the third and fourth columns of Fig. 2.11, this only occurs when the input data have the same distribution as the training data. Specifically, the *R-macaq.* column shows the segmentation yield by the algorithm presented in this chapter with the parameters set for the macaque model dataset. The *R-tuned* corresponds to the segmentation obtained when the parameters (see Table 2.2) are tuned to best fit other datasets. As shown, the algorithm performs excellently with data similar to those considered during the design. This approach is of enormous help in automating the analysis of hundreds or thousands of images. However, it is insufficient with severe infection models such as those in the second, third and fourth row of the figure corresponding to other animal and disease models. This fact motivates the creation and implementation of methods, such as those studied in subsequent chapters, with the ability to perform the transfer of meaningful information between datasets. As a teaser, the fifth and sixth columns show the segmentation results obtained with two DL-based approaches; nnU-net [141] (*DL-nnunet*), SOTA method and, an own approximation (*DL-our*), that will be presented in Chapter Translational Lung Imaging Analysis Through Disentangled Representations.

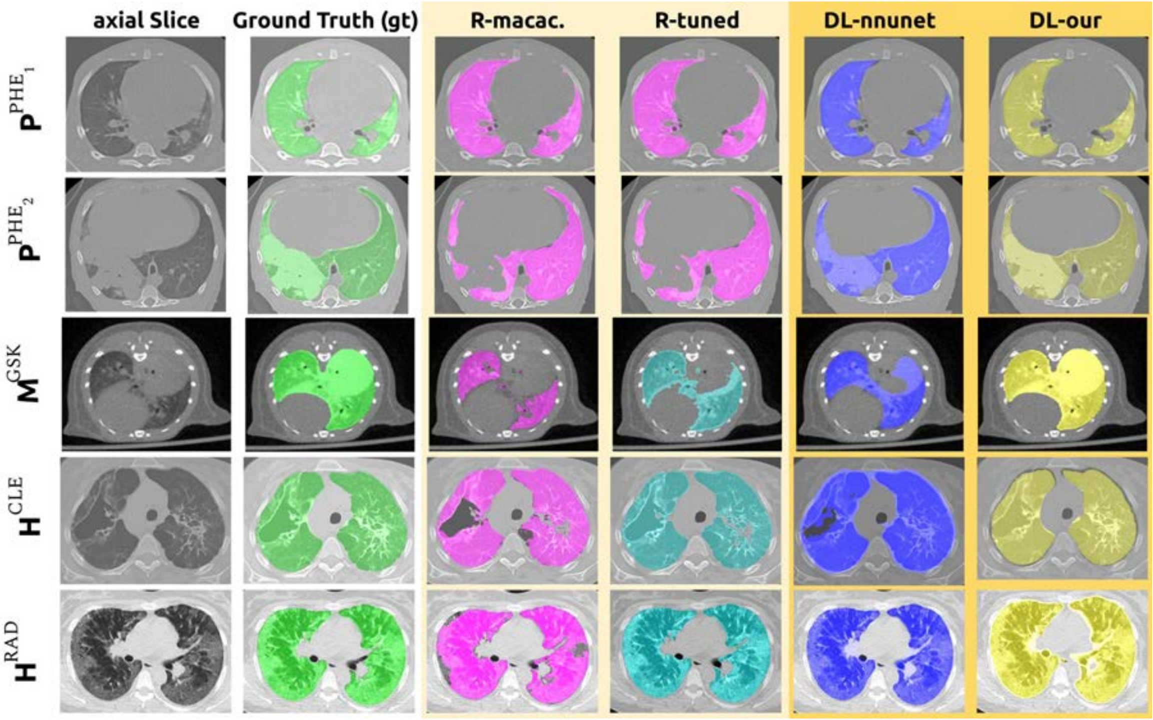


Fig. 2.11 Pathological Lung Segmentation (PLS) masks examples obtained applying the method presented in this chapter and DL-based methods that will be in Chapter 5, employing different animal and disease models to those considered during the initial design. Each row shows an axial slice as an example of, respectively: 1) Initial dataset for which the presented model was designed, namely, mild TB macaque model ($P^{PHE_1}$, see Section 2.2.1), 2) severe TB macaque model ($P^{PHE_2}$, see Section 3.2.1), 3) severe TB mouse model ($M^{GSK}$) [76], 4) human TB ($H^{CLE}$) [68] and 5) human COVID ($H^{RAD}$) [57] (description in Table 5.1).
The columns correspond to a) the original chest CT axial slice, b) the ground truth mask delimited by experts (see details in Sections 2.2.1 and 5.1), c) the mask obtained with the parameters given at the Table 2.2 (*R-macac.*), d) the mask obtained after tuning the parameters for each specific model (*R-tuned.*), e) mask obtained employing the SOTA DL-based method, nnU-Net [141] (*DL-nnunet*) and f) the hybrid (discriminative + generative) DL-based method proposed at Chapter 5 of this work [99] (*DL-our*).

# Chapter 3

# Radiomics for TB Manifestations Classification

## 3.1 Introduction

The previous chapter, Lung Segmentation and Quantification with Rule-based Approach, illustrates how the more traditional methods for automating Pathological Lung Segmentation (PLS) [95] combined with a classic estimation algorithm such as Expectation-Maximization (EM) [94, 220] allows quantifying infected lungs. As was commented, such an approach works as long as the input images belong to a specific domain, namely, a model of mild TB in macaques (see Sections 1.2.2). Even when this approach works as ideally expected, it may be insufficient to our goal of achieving a characterization of the continuous spectrum of the disease (see Section 1.2.1) that requires the discrimination between the different types of lesions.

Expert radiologists have claimed that TB lesions appear in high-resolution CT images at all disease stages, which radiological manifestations could be used as imaging biomarkers to provide information about the biological course of the disease. Thus, this chapter, presents a complete pipeline to detect TB lesions on thorax CT scans and extract informative features from them. In particular, the method infers the TB lesions after feeding with the texture features a *Random Forest* classifier (see Section AI Learning Principles with Emphasis in Lung Analysis). The model can provide an adequate classification for a complex multi-label problem, distinguishing between five different TB lesions types: granulomas, conglomerations, trees in bud, consolidations and ground-glass opacities (see Section 1.3). The work as previously published and presented orally in the *International Symposium on Biomedical imaging* (ISBI) as *Towards an informational model for tuberculosis lesion discrimination on*

*X-ray CT images* [97] and in the *European Molecular Imaging Meeting* (EMIC) as *Radiomics for the Discrimination of Tuberculosis Lesions* [96].

## 3.2 Materials and Methods

The proposed methodology is summarized in the workflow shown in Figure 3.1 and it is described in detail below:

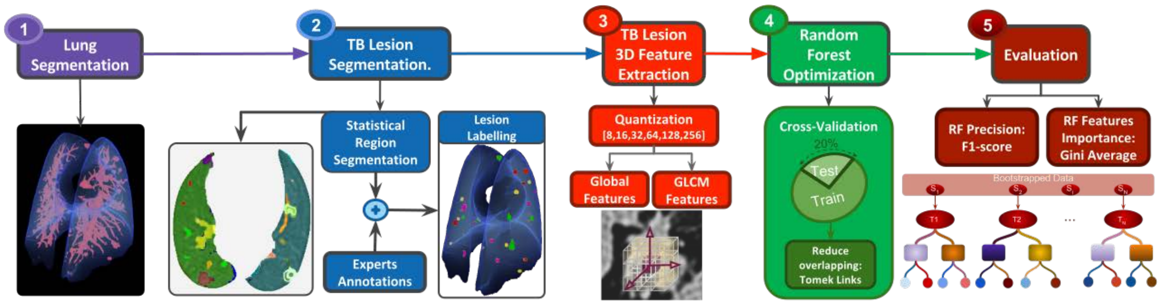


Fig. 3.1 Fully-automatic radiomics workflow for the extraction of informative features on the lung parenchyma: 1) Lung segmentation and airway tree extraction; 2) Selection of relevant volumes employing the Statistical Region merging method [226] matched with the expert annotations of lesions; 3) Extraction of texture features from each volume at 8, 16, 32, 64, 128 and 256 levels of quantization; 22 features are extracted from the Grey Level Co-Occurrence Matrix (GLCM) and 4 are global descriptor of the volume (Mean, Median, Maximum and Minimum); 4) Optimization of the *Random Forest* (RF) hyperparameters (number of trees, minimum number of samples for split and the maximum number of features to evaluate per node); The optimal RF classifier is computed per quantization level and number of features employed. The optimization employ a grid search process with 100-fold cross validation where the training data (80% of the total) in each fold is filtered employing Tomek Links [113] to handle class imbalance; 5) Two-fold evaluation: a) The weighted $F_1 - score$ is employed as a measure of the classification quality of the most frequent TB lesion types; b) The importance of each feature is evaluated using as merit figure the Gini importance.

### 3.2.1 Materials: Computer Tomography Images

In this chapter, forty-two thorax CT scans acquired on a medium size animal model of Tuberculosis were employed. Their voxel size is 0.26 mm × 0.26 mm × 0.63 mm. In order to build a predictor, the identified lesions (Regions-of-Interest, ROIs) were labeled by an expert distinguishing five types of lesions: 2140 granulomas, 350 conglomerations, 82 trees in bud, 80 consolidations and 53 Ground Glass Opacities (see Medical Imaging for Tuberculosis

Assessment).
All animal procedures and study designs were approved by the Public Health England Animal Welfare and Ethical Review Body, Porton Down, UK, and authorized under an appropriate UK Home Office project license.

The reader should note that this dataset differs from the one presented in Section 2.2.1. The former does not have the annotated lesions necessary for this chapter. However, the axial slice in the second row of Fig. 2.11, employed to illustrate a domain shift due to the present manifestations, belongs to the employed dataset.

### 3.2.2 Lungs Segmentation

For completeness a small description of the process is included in this section, for further details see Section 2.2.2.

Air-like organs (e.g., healthy lungs, airways tree, stomach) are identified on a thorax CT scan. Next, the intricate airways tree structure is computed and removed from the segmented lungs. Finally, unsegmented pulmonary regions corresponding to damaged parenchyma and TB lesion are included by morphological hole filling. The whole procedure is summarised in Fig. 3.2.

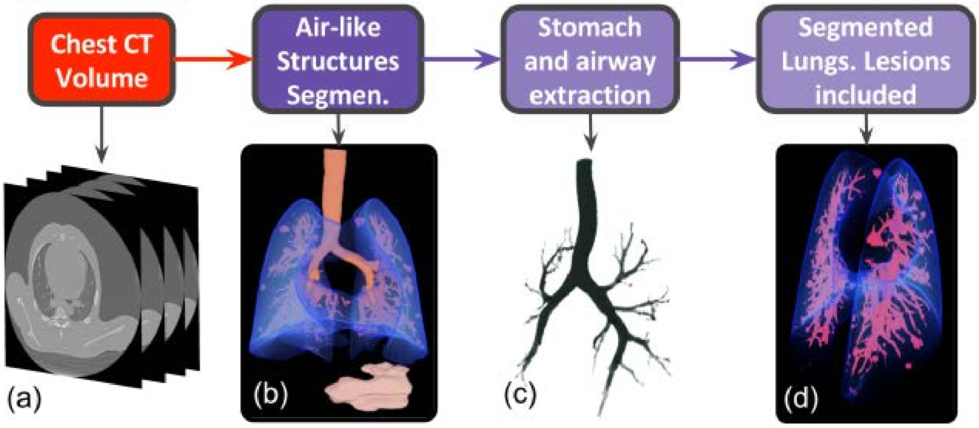

Chest CT Volume
Air-like Structures Segmen.
Stomach and airway extraction
Segmented Lungs. Lesions included
(a)
(b)
(c)
(d)

Fig. 3.2 Summary of lungs segmentation methodology: From a CT scan a) aerated regions are identified via automatic thresholding. b) Then, lungs and airways together are isolated, based on topological priors. Next, the intricate airways are extracted (c). Finally, non-segmented damaged tissue is included by a hole filling method based on active contours (d).
For further details check Automatic Lung Segmentation.

### 3.2.3 Lesions Segmentation

The lung volume is divided into regions of similar appearance (i.e., lesions, vessels, healthy parenchyma) applying the statistical region merging (SRM) approach [226].
The SRM works in two steps: sort and merge. Firstly, the voxels in each 26-connected neighbourhood are sorted based on their similitude. Subsequently, voxels are compared in the earlier order and merged into regions when a given condition is accomplished. The merging predicative is based on intensity similarity and region size (further details can be found in [226]). The process achieves an automatic segmentation of the lesions on the extracted CT lung volumes as illustrated in Fig. 3.1.2.

### 3.2.4 Lesion Characterization [1]

Texture features intend to formally define the spatial distribution of the pixel intensities perceived by experts during manual image evaluation. Numerous types of texture features (e.g., based on *Grey-Level Run Length Matrix* (GLRLM) [85, 329], based on *Local Binary Patterns* (LBP) [229, 295]) can be found in a vast related literature [64]. The nature of the problem to be solved designates the most appropriate set of characteristics to use. Namely, it depends on the imaging technique used, the region to be analyzed, and the specific manifestations of the disease to characterize.
Thus, among the variety of texture features, the ones based on the grey level co-occurrence matrix (GLCM) [105, 106] are proved to be specially useful in our context [64, 203]. GLCM represents the joint frequency over all possible grey levels combinations in every pair of voxels separated by a predefined *offset*, and it is computed as:

$$C_{\Delta x,\Delta y,\Delta z}(l_i,l_j) = \sum_{x=0}^{N_x-1}\sum_{y=0}^{N_y-1}\sum_{z=0}^{N_z-1}\begin{cases}1, & \text{if } I(x,y,z)=l_i \text{ and } I(x+\Delta x,y+\Delta y,z+\Delta z)=l_j\\ 0, & \text{otherwise}\end{cases}$$

where $I(x,y,z)$ is the intensity at $(x,y,z)$, $(\Delta x,\Delta y,\Delta z)$ is the *offset*, $N_x \times N_y \times N_z$ is the image size in voxels and $l_{[i,j]} \in L$ the grey level at the pair of voxels $i$ and $j$.

[1] As in the previous chapter the `C++/ITK` [146] code implementation for the Tuberculosis Lung Segmentation (TLS) https://github.com/BIIG-UC3M/TLS-Piped while the `Python` code for texture features extraction and analysis is in the following GitHub Repository. This code was developed from scratch to enable feature extraction from three-dimensional neighbourhoods, reducing the time complexity of similar approaches. Indeed, this software is the choice for the radiomics application works to the clinical practice: *Performance of ultra-high-frequency ultrasound in the evaluation of skin involvement in systemic sclerosis: a preliminary report* [219] and *Association of visual and quantitative heterogeneity of 18F-FDG PET images with treatment response in locally advanced rectal cancer: A feasibility study* [205] thanks to its versatility and performance.

In this work, we restrict to unitary *offsets* and compute GLCM matrices in the 13 possible directions in $\mathbb{Z}^3$. We further average them obtaining a unique GLCM matrix per ROI from which 26 descriptors are extracted[2] as described in [23, 64]. These are added to the global ROI histogram descriptors: *mean*, *standard deviation*, *minimum* and *maximum* grey value.

Commonly, the GLCM is not computed on images with $2^{16}-1$ grey levels ($L$) [3] that can be found in a CT image, but a quantization is performed, that is, a reduction of the number of grey levels, using rounding and truncation techniques. In this way, it is possible to reduce the computational cost of the GLCM calculation while minimizing the noise [293]. However, it could mean a loss of detail if it is influenced by the artefacts present in the image (for example, the movement of the lungs). Therefore, it could not faithfully represent the tissue that *a priori* characterizes. For this reason, this work studies the effect of quantization to determine whether it is preferable to work with more smoothed images due to the noise or whether it is better to use all the available information. With this objective, the classification study is performed using $8, 16, 32, 64, 128$ and 256 grey levels independently.

### 3.2.5 Random Forest Optimization and Evaluation

It is well-known from the literature [33, 208, 270] that the (RF) classifiers provide a high precision due to their ability to discard irrelevant features. More formally, modelling with RF allows us to obtain a complex model estimator able to define flexible manifold in high dimensional input spaces [33, 110] without completely losing the interpretability of more classic statistical models (see AI Learning Principles with Emphasis in Lung Analysis). Specifically, to estimate the importance of each feature, $\phi$, (the 26 GLCM descriptors in our case), we use the *Gini importance* [199, 208, 216], $I_G$, which is computed as the averaged over all the trees $T$ and all the nodes $N$ of the *Gini impurity*, $G(n) = \sum_{k \in C} p_k(1-p_k)$, change:

$$I_G(\phi) = \sum_{t \in T} \sum_{n \in \mathrm{N}} \Delta G(n), \tag{3.1}$$

where $\Delta G(N) = G(N) - \sum_{s=0}^{S} G(S)$ ($s \in S$ set of nodes splited from $N$), $p_k$ is the probability of having an instance labeled with class $k$ in the set $C$. The Gini importance of each feature varies between 0 and 1 and the sum over the whole feature set adds to 1.

[2] For the full description of the 26 features see the Appendix B.

[3] Note that, typically, not all the procured levels are used when using 16 bits, since the HU are usually set between -1024 and 3024

| | 8 | 16 | 32 | 64 | 128 | 256 |
|---|---|---|---|---|---|---|
| **6** | 0.788 ±0.01 | 0.792 ±0.01 | 0.803 ± 0.01 | 0.818 ± 0.01 | 0.841 ± 0.01 | 0.844 ± 0.01 |
| **26** | 0.806 ±0.01 | 0.808 ± 0.01 | 0.815 ± 0.01 | 0.829 ± 0.01 | 0.838 ± 0.01 | 0.844 ± 0.01 |

Table 3.1 Optimization results employing 6 and 26 most significant features

The RF hyper-parameters are the number of trees $T$ in each RF, the minimum number of samples for split and the maximum number of features to evaluate per node. The optimal RF hyper-parameters per quantization level are found by grid search employing a 100-fold cross-validation (CV) per RF candidate.

The weighted $F_1$-score was used as the quality measure, which is defined as follows:

$$\frac{1}{\sum_{k\in C}|\hat{y}_k|}\sum_{k\in C}|\hat{y}_k|\,F_1(y_k,\hat{y}_k), \tag{3.2}$$

where $y_k$ and $\hat{y}_k$ represent the predicted and the true labels. Namely, it is the weighted average of each class's $F_1$-score (harmonic mean of the precision and the recall).

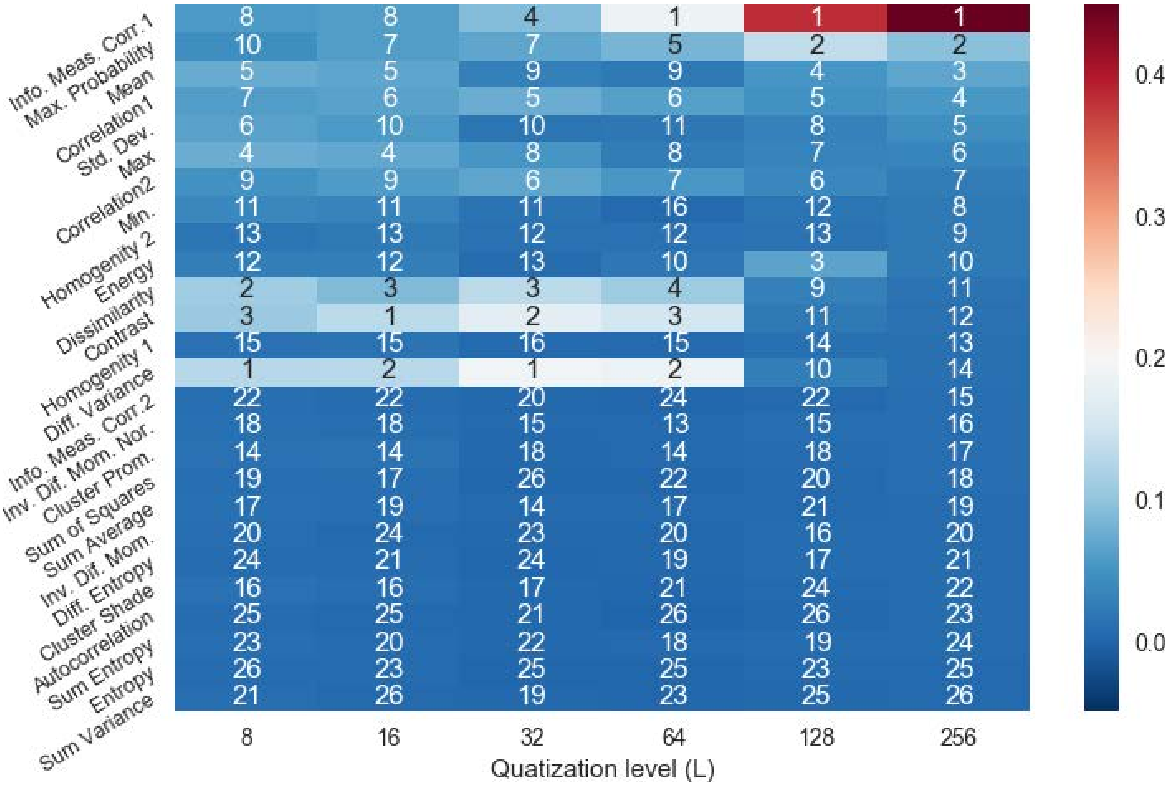
Besides, due to the high imbalance of the annotated dataset, the *Tomek Links* technique [113] is applied to the training data to reduce the overlap in the feature space.

## 3.3 Results

Figure 3.3 presents the Gini importance $I_G$ of each feature for a given quantization level $L$.

It can be observed that the *difference variance, contrast* and *information measure of correlation 1* rank first for at least one quantization level. The Gini importance of the most informative feature increases with $L$ (*difference variance*, $L = 8$, $I_G = 0.14$; *information measure of correlation 1*, $L = 256$, $I_G = 0.45$).

Figure 3.4 shows the weighted $F_1$ score obtained for the optimal estimator at each $L$ in function of the number of features. The observed results confirm the ones presented above. At large quantization levels ($L = 128, 256$), the precision reaches good values using just two or three features, while at small $L$, the increment in the weighted $F_1$-score with the number of features grows more slowly. However, all the estimators show symptoms of convergence when using 6 features (red line in Figure 3.4). The weighted $F_1$-score at convergence (i.e., using 26 features) increases as the quantization level does ($0.844 \pm 0.01$, $L = 256$) as shown in Table 3.1.



| | 8 | 16 | 32 | 64 | 128 | 256 |
|---|---|---|---|---|---|---|
| Info. Meas. Corr.1 | 8 | 8 | 4 | 1 | 1 | 1 |
| Max. Probability | 10 | 7 | 7 | 5 | 2 | 2 |
| Mean | 5 | 5 | 9 | 9 | 4 | 3 |
| Correlation1 | 7 | 6 | 5 | 6 | 5 | 4 |
| Std. Dev. | 6 | 10 | 10 | 11 | 8 | 5 |
| Max | 4 | 4 | 8 | 8 | 7 | 6 |
| Correlation2 | 9 | 9 | 6 | 7 | 6 | 7 |
| Min. | 11 | 11 | 11 | 16 | 12 | 8 |
| Homogenity 2 | 13 | 13 | 12 | 12 | 13 | 9 |
| Energy | 12 | 12 | 13 | 10 | 3 | 10 |
| Dissimilarity | 2 | 3 | 3 | 4 | 9 | 11 |
| Contrast | 3 | 1 | 2 | 3 | 11 | 12 |
| Homogenity 1 | 15 | 15 | 16 | 15 | 14 | 13 |
| Diff. Variance | 1 | 2 | 1 | 2 | 10 | 14 |
| Info. Meas. Corr.2 | 22 | 22 | 20 | 24 | 22 | 15 |
| Inv. Dif. Mom. Nor. | 18 | 18 | 15 | 13 | 15 | 16 |
| Cluster Prom. | 14 | 14 | 18 | 14 | 18 | 17 |
| Sum of Squares | 19 | 17 | 26 | 22 | 20 | 18 |
| Sum Average | 17 | 19 | 14 | 17 | 21 | 19 |
| Inv. Dif. Mom. | 20 | 24 | 23 | 20 | 16 | 20 |
| Diff. Entropy | 24 | 21 | 24 | 19 | 17 | 21 |
| Cluster Shade | 16 | 16 | 17 | 21 | 24 | 22 |
| Autocorrelation | 25 | 25 | 21 | 26 | 26 | 23 |
| Sum Entropy | 23 | 20 | 22 | 18 | 19 | 24 |
| Entropy | 26 | 23 | 25 | 25 | 23 | 25 |
| Sum Variance | 21 | 26 | 19 | 23 | 25 | 26 |

Fig. 3.3 Matrix of the Gini importances estimating the importance of each feature for the optimal RF classifier. The number on each row corresponds to the feature ranking position and the colour, the averaged Gini index. Each column gives the results for a given quantization level.

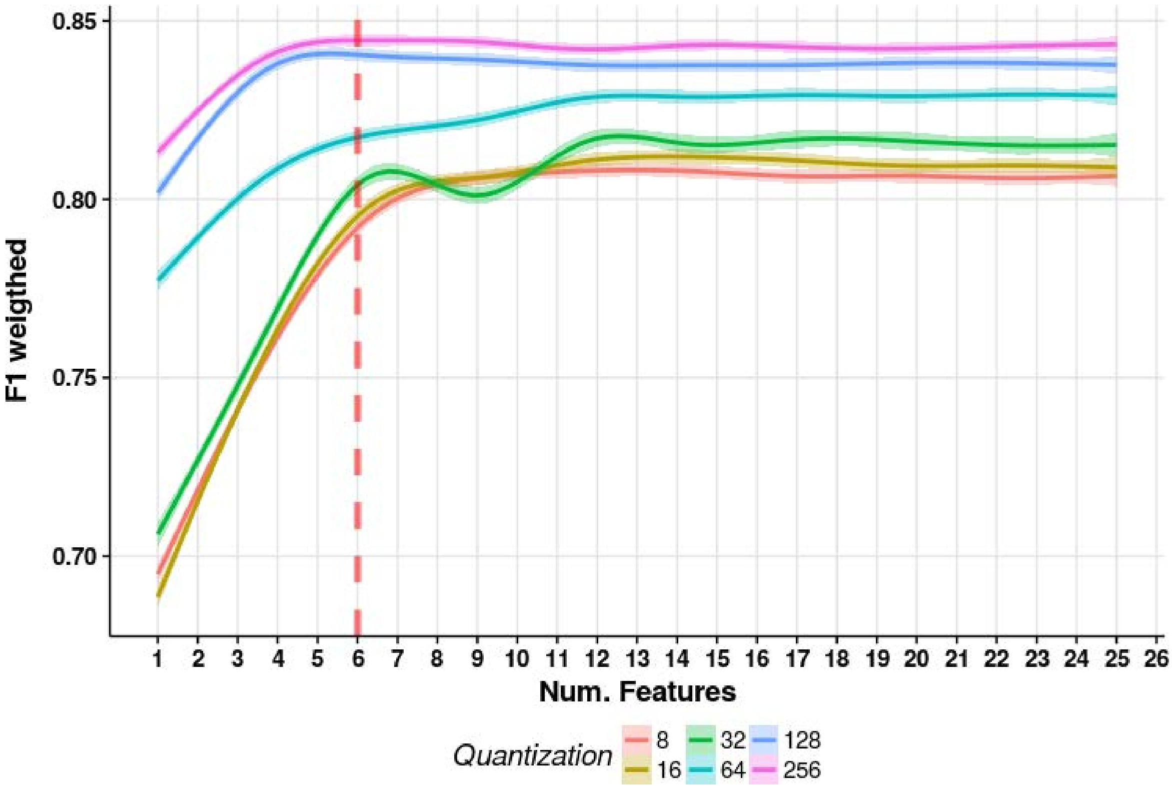


Fig. 3.4 Per quantization level, mean weighted $F_1$ score obtained for groups of sorted features (size 1 to 26) over the 100 cross-validation folds of the best estimator. The features are sorted as per the ranking in Figure 3.3. The 95% confidence interval is drawn as the line shadow.

In Figure 3.5, we present the box plots of the weighted $F_1$-scores obtained for the best RF estimator at each quantization level for two interesting cases (6 and 26 features). The differences between factors were assessed with the one-way analysis of variance test (ANOVA). The differences between quantization levels are statistically significant ($p < 2 \times 10^{-7}$) while those between the number of features are not ($p > 0.05$).

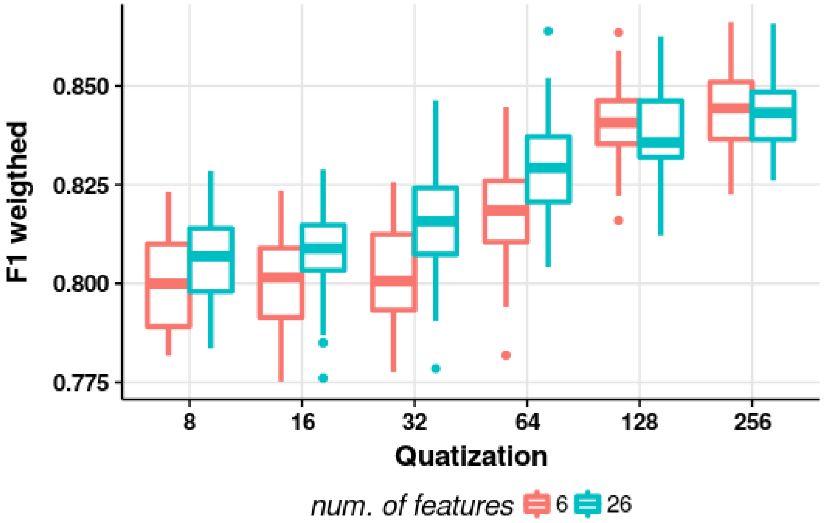


Fig. 3.5 Box plot representation of the weighted $F_1$ scores obtained on the 100 folds of the cross validation process for the best RF estimator. The results are shown for each quantization level when employing the 6 most relevant and all the 26 features.

## 3.4 Discussion

The results presented in this manuscript show that it is possible to obtain a good classification performance for a complex multi-label classification problem by training simple models that are still informative of the disease development.

It is crucial to notice that by using a unique *offset*, averaging the GLCM directional descriptors and performing an independent treatment for each quantization level, we can keep the set of features small ($n = 26$). This is critical to facilitate the model comprehension and departs from the tendency to use models focused on achieving maximal precision at the cost of employing thousands of features.

As previously commented, identifying the lesions improves with the quantization level in parallel to the growth of the Gini importance of the most informative features. In particular, when employing 128 and 256 grey levels, the Gini importance of the first ranked feature, *Information Measure Correlation 1* or $r_0$, is particularly high, 0.38 and 0.45, respectively. This descriptor gives a general measure of the correlation between the intensity of adjacent voxels and can be interpreted as an information gain. Our results point out that when employing the proper level of detail, this descriptor can capture complex relations between the intensities of neighbour voxels characteristic of a particular type of lesion.

In general, we can confirm that our model can quantify changes in the lung parenchyma that are specific to a given type of lesion (e.g., density, heterogeneity). This affirmation gets strongly supported when analyzing the importance of the features at lower quantization levels. Namely, the most important features (*contrast, difference of variance*) are strongly related to *Information Measure Correlation 1*. This fact was expected as the RF classifiers discard features contributing with redundant information. For example, the descriptor, *Information Measure Correlation 2*, defined as $r_1 = (1 - e^{2r_0})^{1/2}$, it is closely related to $r_0$ and consequently, provides similar information. So, it has a low Gini importance in our model for all the quantization levels appearing at the bottom of the ranking shown in Figure 3.3.

## 3.5 Conclusion

The proposed framework gives promising results in its ability to extract informative biomarkers of tuberculosis development. Namely, we demonstrate that we can achieve a reasonable good classification of the most frequent TB lesion types.

This work will be the basis for the studies presented in the next chapters on the characterization of the biological changes induced by TB infection. We have selected the state-of-the-art (SOTA) methods that will hopefully lead to an improved understanding of the course of the disease. This work proves that machine learning (ML) can characterize segmented lesions, even employing a relatively medium/low capacity model. Thus, the next chapter, Deep Learning for TB Manifestation Classification [98], proposes a DL-based method capable of identifying lesions from a whole volume without relying on previously segmented lesions. Meanwhile, in Chapter 5, Translational Lung Imaging Analysis Through Disentangled Representations [99], we investigate the ability of a DL model to not only identify pathological lungs for different species and disease models (increase generalization capacities between domains). But as well, to synthesize realistic images of lungs damaged by Tuberculosis by introducing tools developed under the framework of graphical causality. Such innovation initiated through the preliminary work presented in this chapter escalates the models' skill for characterization and quantification of TB to a new paradigm of infinite possibilities within clinical practice.

# Chapter 4

# Deep Learning for TB Manifestation Classification

The results obtained in the previous chapter show the possibility of extracting quantitative information from the Computed Tomography (CT) images through statistical descriptors (texture features) that allow us to formalize radiological descriptions of the manifestations of Tuberculosis (TB).
These features enable automatic manifestations identification using machine learning (ML) to characterize complex relationships among them. Such an approach requires previous manifestation delimitation. The difficulty of the delimitation process depends on the animal and disease model due to their biological variability. Indeed, experts face complications when performing this process manually in clinical practice or generating the ground truths needed to train supervised ML algorithms. Uncertainty in delimitation directly translates to the automatization performance.

To tackle such issues, this chapter presents our previously published work as the *A Multi-Task Self-Normalizing 3D-CNN to Infer Tuberculosis Radiological Manifestations* [98], in which we propose to identify the presence of lesions without prior segmentation.
We hypothesize that since a few handcrafted texture features can capture meaningful information to classify delimited lesions through a *Random Forest*, much more powerful Deep Learning (DL) models should perform a suitable classification without the need to delimit lesions and extract handcrafted features.
As pointed out in the AI Learning Principles with Emphasis in Lung Analysis, DL models are end-to-end. Therefore, the features do not have to be defined. Oppositely these are learnt, and their effectiveness is well demonstrated in object localization tasks both in computer vision [72, 166] and medical imaging fields [78, 188, 351].

Specifically, we investigate this hypothesis by building a model to mimic the radiologist generated reports by inferring the presence of TB manifestations on thoracic Computer Tomography scans.

Our model exploits the well-known advantages of three-dimensional Convolutional Neural Networks (3D-CNNs). In particular, we adapt the *V-Net* encoder to distinguish among five different radiological manifestations of TB at each lung lobe.

Specifically, since usually TB manifestations do not appear in the infected lungs isolated (i.e., nodules, conglomeration or cavities appear together in the lung parenchyma, see Section 1.2.1), we propose a multi-task model (Section 1.4.2) designed to identify single instances. A joint force strategy is established to overtake the issues (e.g., exploiding/vanishing gradients, lack of sensibility) that generally appear when training complicated deep 3D models with limited size datasets and large medical imaging volumes.

Our proposal employs: 1) At the network architecture level, the *scaled exponential linear unit* (SELU) activation, which allows the self-normalization of the network, and 2) at the learning phase, multi-task learning with a loss function weighted by the task *homoscedastic* uncertainty. This is performed independently of the binary or regressive nature of the task.

## 4.1 Introduction

The automation of the radiologists' reports has been pursued during the last three decades. Most of the classical works in the literature approach the challenge as an image segmentation problem. Hand-crafted features are extracted to segment the lung parenchyma [203] and to identify TB lesions [97, 338]. However, these techniques are usually limited to a particular application and are quite sensitive to the high implicit variance of medical images. Fortunately, in recent years, the use of DL techniques has drastically improved the automation of the radiologist reports generation, reaching performances close to the human error [78, 188, 288, 327]. DL has opened a new paradigm in the medical imaging field [123] with remarkable results when expert knowledge is properly incorporated into the models [27, 90]. For the application at hand, knowledge is usually injected into the models in form of manually segmented masks of the lung-damaged Volumes-of-Interest (VOI). This technique yields to good results when 2D Convolutional Neural Networks are employed [108, 141, 171, 264] and promising ones [221], when using complex 3D CNN models [53, 111, 210]. However, to the best of our knowledge, there is a lack of studies directly employing the expertise acquired by radiologists through years of clinical practice as synthesized in tabular reports. For this reason, in this chapter, we aim to employ the information stored in the radiologists' reports.

With this aim, our methodology integrates the following techniques:

1. A 3D-CNN based on *V-Net* architecture [210] is employed to extract distinctive features from whole CT volumes. The extracted features are used to detect the presence or the quantity of specific TB manifestations employing as ground truth the radiologist reports. This approach makes use of network models and learning principles that are common in the literature

2. The 3D-CNN architecture is modified to leverage a better regularization and act as *Self-Normalizing Neural Networks* (SNNs) [164].

3. Our deep network is configured as a multi-task model to perform multi-label classification, acknowledging that TB manifestations do not appear isolated. Moreover, uncertainty is used to weigh the influence of each task loss [158] (see uncertainty definition at Section 1.4.1).

## 4.2 Material and Methods[1]

### 4.2.1 Material

The experiments (see 4.3) were accomplished on a dataset constituted by 56 chest CT-scans acquired from 14 male Cynomolgus macaques at 3, 7, 11 and 16 weeks after TB aerosol exposure. The voxel is 0.26 mm × 0.26 mm × 0.63 mm with an in-plane resolution of 512 pixels × 512 pixels and 201 to 270 slices, which are preprocessed to feed the model (see 4.2.5). We employ the quantitative report elaborated by a radiologist with 20 years of experience as labelled data. These reports are tabular and contain the number of detected nodules (see Fig. 1.6a), which fluctuate between 0 and 15 depending on disease stage, and boolean annotations about the presence or absence of the most common TB manifestations [217] -namely, cavitations (Fig. 1.6e), conglomerations (Fig. 1.6a), consolidations (Fig. 1.6c) and trees in bud (Fig. 1.6b). The reports contain the disaggregated information per lung lobe (i.e., right superior, right middle, right inferior, left superior and left inferior) given a total of 25 manifestations to predict per subject.

[1] The `Python/Tensorflow` implemtation of the methods can be found under the following GitHub Repository

### 4.2.2 Model Architecture

Our model implementation aims to exploit the ability of state-of-the-art architectures to extract fine-grained features from chest CT-Images. We plan to use the extracted features as the input to Fully Connected Layers (FCLs) for multi-label classification among the TB manifestations aforementioned. In this context, the *V-Net* [210] has been proven quite successful at segmentation tasks [188, 221] and even more importantly for our purpose, at the generation of synthetic lung nodules from these features [145].

We adapt the *V-Net* encoder to implement our Self-Normalizing Neural Network (SNN, explained in detail in the next section) or to include Batch Normalization layers [140] for the comparison experiments.

The *V-Net* encoder iteratively convolves inputs of preprocessed lung CT-Volumes (see 4.2.5) with a size of $128 \times 128 \times 64$ voxels with four codification stages, each of them halves the resolution and adds channels up to 256, conducting to the generation of 1376256 $(8 \times 8 \times 4 \times 256)$ features, which are flattened to feed our first FCL, referred as $\text{FCL}_1$.

This first FCL is task-shared. Therefore, it is in charge of characterizing the complex relationship among the low-level features obtained with the *V-Net* in order to produce more abstract features common to the twenty-five tasks of our problem. $\text{FCL}_1$ is built up with four layers of 4096, 2048, 1024 and 256 units, respectively. All the parameters corresponding to the modified *V-Net* encoder and $\text{FCL}_1$ are common to the prediction of each manifestation; this fact enables a more efficient training due to the possibility of modelling conditional relationships among the manifestations related.

However, to achieve a particular prediction for each task, we complete the architecture by employing the last 256 features of $\text{FCL}_1$ as input of two independent FCLs, $\text{FCL}_R$ and $\text{FCL}_B$. $\text{FCL}_R$ predicts the regression tasks (nodules counting), while $\text{FCL}_B$ predicts the binary tasks. Each specific FCL is formed by two layers of 256 and 128 units and the final output units composed by 5 *Rectified Linear Units* (ReLU) for $\text{FCL}_R$ and 20 *sigmoid* activated units for $\text{FCL}_B$.

It is essential to mention that dropout or batch normalization layers are included where is needed, as shown in 4.1. When employing SNN, batch normalization is not needed (see next section for details).

### 4.2.3 Self-Normalizing Neural Networks

As explained above, we only implement half of the *V-Net* architecture, the encoder, which results appropriate for our application but introduce a new issue. As Milletari et al. [210]

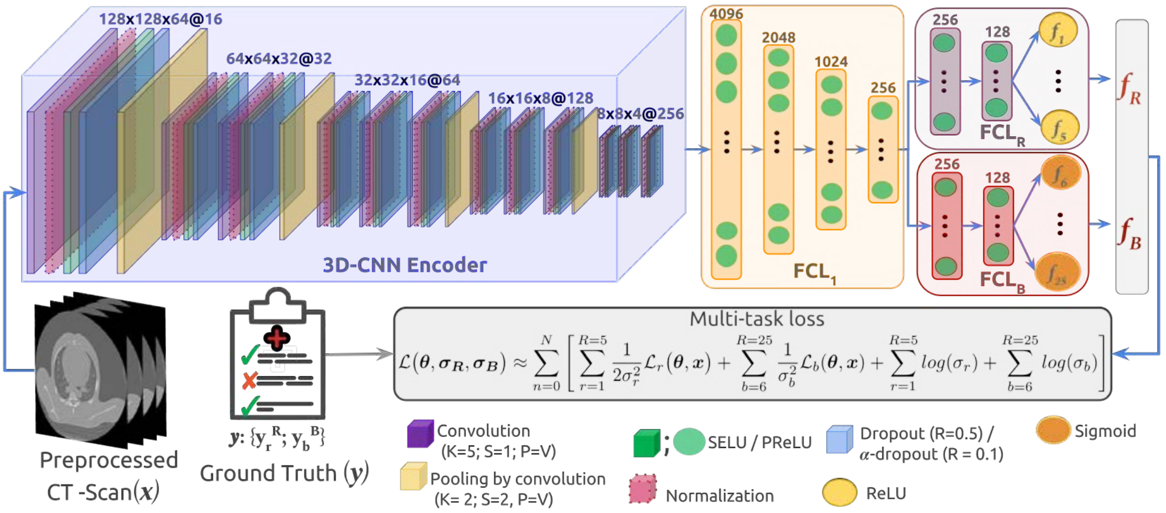

128x128x64@16
64x64x32@32
32x32x16@64
16x16x8@128
8x8x4@256
3D-CNN Encoder
4096
2048
1024
256
FCL$_1$
256
128
$f_1$
$f_5$
FCL$_R$
256
128
$f_6$
$f_{25}$
FCL$_B$
$f_R$
$f_B$
Multi-task loss
$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\sigma_R}, \boldsymbol{\sigma_B}) \approx \sum_{n=0}^{N} \left[ \sum_{r=1}^{R=5} \frac{1}{2\sigma_r^2} \mathcal{L}_r(\boldsymbol{\theta}, \boldsymbol{x}) + \sum_{b=6}^{R=25} \frac{1}{\sigma_b^2} \mathcal{L}_b(\boldsymbol{\theta}, \boldsymbol{x}) + \sum_{r=1}^{R=5} log(\sigma_r) + \sum_{b=6}^{R=25} log(\sigma_b) \right]$
Preprocessed CT -Scan($\boldsymbol{x}$)
$\boldsymbol{y}$: $\{y_r^R; y_b^B\}$
Ground Truth ($\boldsymbol{y}$)
Convolution (K=5; S=1; P=V)
Pooling by convolution (K= 2; S=2, P=V)
; SELU / PReLU
Normalization
Dropout (R=0.5) / $\alpha$-dropout (R = 0.1)
ReLU
Sigmoid

Fig. 4.1 The model is composed by a three-dimensional Covolutional Neural Network (3D-CNN) based on the V-Net [210] and three Fully Connected Layers (FCLs): $FCL_1$ with tasks-shared parameters and, $FCL_R$ and $FCL_B$ for prediction of regression and binary tasks, respectively. The encoder is built upon the following layers: 1) Convolution characterized by their kernel size (K), stride (S) and the Valid (V) padding (P), in purple or in yellow for pooling; 2) Dropout (alpha-dropout for our proposed model) with a rate (R), in blue; 3) Activations: *Scaled Exponential Linear Unit* (SELU) [164] for our model, Parametric Rectified Linear Unit (PReLU) [115] for comparison experiments), in green; 4) Normalization (not employed in our model but for comparison), in pink. The final outputs units take *Rectified Linear Unit*(ReLU) and *sigmoid* activation to predict the labels used in the loss function [158].

stated, *V-Net* forwards the features extracted from the first levels of the encoder to feed the same levels of the decoder in order to improve the convergence of the model.
The vanishing/exploiding gradients effect is remarkably avoided trough the regularization effect produced by the partial deep Feed-forward Neural Networks (FNNs) conforming such connections usually referred as *skip connections*. Our model lacks these connections, making necessary the use of a different effective regularization technique.

In recent years, *Batch Normalization* (BN) [140] has been the most widely used regularization technique [188]. The successful BN application depends on employing relatively large batch sizes during training since each BN layer needs to extract (ideally) unbiased statistics (mean and variance) from each input batch.
However, the usually small amount of available data and the large size of medical volumes force optimization of the network parameters on mini-batches. The use of mini-batches causes perturbations and high variance in the training error particularly, when the residual connections are not present in the network [164].

To alleviate this problem, we propose to include the *Self-Normalizing Neural Networks* (SNNs) strategy [164] to our 3D-CNN + FCLs model.
SNNs leads to automatically normalized networks by preserving the activation close to zero mean and unit variance. Besides, Klambauer et al. [164] proved that even in the worst-case scenarios, gradients could neither vanish nor explode.

Therefore, we implement our model to adjust it to the design conditions defined in [164]. Namely, employing the following specific *activation function*, *weights initialization* and *dropout*:

*Activation Function*: *Scaled Exponential Linear Units* (SELUs) given by

$$\mathrm{SELU}(x) = \lambda \begin{cases} x & if \quad x > 0 \\ \alpha e^{x} - \alpha & if \quad x \leqslant 1 \end{cases} \tag{4.1}$$

being $\lambda$ = 1.0507 and $\alpha$ = 1.6733, which are fixed to assure and activation of unit variance and zero mean. A different configuration implies different normalization parameters.

*Weights Initialization*: It is needed to assure that at initialization, the first and second moments of the weights are zero and one, respectively. Because of this, the weights are drawn from the normal distribution,

$$\mathscr{N}\left(0, \frac{1}{n}\right), \tag{4.2}$$

being $n$ the number of units at each layer.

*Alpha Dropout*: Standard dropout does not work properly with SELU activation (it is not possible to keep the mean and variance at the original values). The proposed alpha dropout technique changes the activation, $x$, as

$$x = a\big(\alpha \cdot d + \alpha'(1-d)\big) + b \tag{4.3}$$

$$d \sim \mathscr{B}(1,q), \qquad q := \text{dropout rate} \tag{4.4}$$

$$a = \big(q + \alpha'^2 q(1-q)\big)^{-\frac{1}{2}} \tag{4.5}$$

$$b = -\big(q + \alpha'^2 q(1-q)^{-\frac{1}{2}}\big)\big((1-q)\alpha'\big), \tag{4.6}$$

being $\alpha' = -\lambda \cdot \alpha = -1.7581$ and $\mathscr{B}$ a binomial distribution.

### 4.2.4 Learning Principle: Uncertainty Weighted Multi-task Loss

The use of multi-task networks provides several significant advantages towards a more capable training since similar tasks modelled together leverage model convergence in contrast with single-task models [268, 269]. However, the most employed learning strategies do not consider these facts explicitly promoting the appearance of the new methodology to exploit such property and tackle multi-task issues. [32, 71, 157].

In this work, we propose the inference principle of uncertainty weighting to cope a non-trivial aspect: the choice of the influence of each task in the final loss.
Traditionally, each task is associated with a weight selected either manually or after an exhaustive grid search. Then, the final loss is computed by the linear combination of each specific task loss and its weight, namely

$$\mathscr{L} = \sum_i w_i \mathscr{L}_i, \tag{4.7}$$

being $w$ and $\mathscr{L}$ the weight and loss of each particular task, $i$.
This approach is highly influenced by the units and the scale of each task and is extremely computationally intensive and time-consuming.

Recently, this problem has been addressed by Kendall et al. [158] proposing the guidelines to compute the weight guiding each specific task loss by the *task-dependent* or *homoscedastic* uncertainty of the predictions (see further details at the description 1.4.1). Besides, this work leverages the quantification of such uncertainty employing DL estimators, which is a hot topic in the literature due to the intractable nature of such models which do not respond to closed solutions. [212].

Thus, following the success of other medical imaging applications of approaches embracing uncertainty [2] (e.g., radiotherapy planning [32], brain imaging [71, 218]). In this work, we adapt to our multi-label classification problem.
For this, we apply the principles of estimation theory (see Equation 1.4) to derive the following loss function for our model [158]:

$$p\big(\mathbf{y}\big|\mathbf{x}\big) = p\big(\mathbf{y}\big|\mathbf{f}(\mathbf{x});\theta\big) = \prod_{t=1}^{T=25} p\big(y_t\big|f_t(\mathbf{x});\theta\big) \tag{4.8}$$

where $\mathbf{y}$ are the 25 labels assigned to each particular example (see 4.3 below), $\mathbf{f}(\mathbf{x})$ are the outputs of our network when a chest CT image, $\mathbf{x}$, is employed as input. The parameters (to be learned) $\theta$ are used as network weights. Therefore, the uncertainty in our predictions, $f_t$, for each task $t$ is measured by a probability distribution, Namely:
In the case of the nodule counting tasks, the probability takes the form:

$$p\big(y\big|\mathbf{x};\theta\big) = \mathcal{N}\left(f_t(\mathbf{x})\big|\theta,\sigma_t^2\right) = \frac{1}{2\sigma_t^2}\exp\left(-\frac{\big(y-f_t(\mathbf{x},\theta)\big)^2}{\sigma_t}\right) \tag{4.9}$$

While for the binary tasks, the squashed version of a *sigmoid*, $S$, is employed

$$p\big(y\big|\mathbf{x};\theta\big) = S\left(\frac{f_t(\mathbf{x})\big|\theta}{\sigma_t^2}\right) = S\left(\frac{f_t(x,\theta)}{\sigma_t^2}\right)^y \left(1 - S\left(\frac{f_t(x,\theta)}{\sigma_t^2}\right)\right)^{1-y} \tag{4.10}$$

The task-specific noise parameters, $\sigma_t$, are inferred to model the amount of noise in the ouputs. Taking this into consideration, Eq. 4.8 can be expressed as:

$$p\big(\mathbf{y}\big|\mathbf{f}(\mathbf{x});\theta\big) = \prod_{r=1}^{R=5}\mathcal{N}\left(f_r(\mathbf{x})\big|\theta,\sigma_r^2\right)\prod_{b=6}^{B=25} S\left(\frac{f_b(\mathbf{x})\big|\theta}{\sigma_b^2}\right), \tag{4.11}$$

where $r$ and $b$ correspond with the regression and binary tasks, respectively.

Given the model of Eq. 4.11, we establish as loss to optimize, the following log-likelihood function of the parameters (see [158] for details), $\mathscr{L}(\theta,\sigma_{\mathbf{R}},\sigma_{\mathbf{B}})$:

$$\mathscr{L}\left(\theta,\sigma_{\mathbf{R}},\sigma_{\mathbf{B}}\right) \approx \sum_{n=0}^{N}\left[\sum_{r=1}^{R=5}\frac{1}{2\sigma_r^2}\mathscr{L}_r\left(\theta,\mathbf{x}\right)+\sum_{b=6}^{R=25}\frac{1}{\sigma_b^2}\mathscr{L}_b(\theta,\mathbf{x})+\sum_{r=1}^{R=5}log(\sigma_r)+\sum_{b=6}^{R=25}log(\sigma_b)\right], \tag{4.12}$$

being $N$ the number of examples and $\mathscr{L}_r(\theta,\mathbf{x}) = \left|\left|y_r - f_r(x,\theta)\right|\right|^2$, the loss associated with the regression tasks. While the binary loss is defined as the *Cross-Entropy* (CE), $\mathscr{L}_r(\theta,\mathbf{x}) = CE\left(y = [0,1], f(\mathbf{x},\theta)\right)$. It is important to remark that the defined loss function (Eq. 4.12) provides an additional regularization term by including the logarithms of the noise factors for all the tasks, $\sigma$'s, which is added to the regularization implicit to the network (see section above). This way the weights of the network, $\theta$, are preserved to zero mean and unit variance, which can be added to the model as $p(\theta) = \mathscr{N}(\mathbf{0},\mathbf{1})$.

### 4.2.5 Pre-procesing

The proposed architecture takes as inputs CT scans of 128 × 128 × 64 voxels (see 4.2.2). Because of this, the original CT volumes are cropped. In this process, we extract the lungs by preserving the rib cage VOI as described in Section 2.2.2 [95]. Then, the resulting volume is sampled to the required size.

Besides, we augment the original data to amend the need of thousands of CT-scans to learn the model parameters in an environment with a minimal dataset. This process is performed online. During training, each input volume is augmented applying three transformations available in the *DLTK* framework [237]: elastic transformation, the addition of Gaussian noise and flipping in the three spatial directions.

## 4.3 Experiments and Results

In order to measure the performance of the proposed model, a 5-fold Cross-Validation (CV) is employed. Each fold is composed of 4 CT volumes of each single subject, which lead to training sets of 13 subjects and 52 CT scans.
Besides, to compare the model behaviour against the state-of-the-art approaches, we perform the CV over the proposed model (*SELU*) and a modified version which employs Parametric Rectified Linear Unit (PReLU) [115], BN [140] and standard dropout [297], referred as *BN+PReLU*.

In both cases, models are trained through 10.000 iterations via ADAM optimizer, a learning rate of $10^{-5}$ and a mini-batch size of $n = 1$. In the case of *BN+PReLU*, we employed the standard parameters for ADAM [159] and a dropout rate of 0.5, while for the proposed model $\beta_2$ and $\varepsilon$ were modified to 0.9 and 0.01 and alpha-dropout rate was settled to 0.1 [164].

In 4.2, we show the loss of the validation data per fold when employing the BN+PReLU (in red) and the proposed model, referred to as SELU (in blue). The SELU loss is always more extensive at the first iterations, although at the final iterations, it is always smaller than PReLU.

The inference error at convergence is estimated by the Root Mean Square Error (RMSE) for the five nodules count tasks and the $F_1$-score for the twenty binary tasks. Table 4.1 presents the results per lung lobe and 4.2, by fold. The five top rows represent lobes or folds; the last one, the average per column, shows the results for each manifestation when employing the *BN+PReLU* or the proposed model.
Results are very similar for both models: no significant statistical differences were found, $p \nleq 0.05$, for paired t-test of each kind of manifestation. Nevertheless, the average results are better for the proposed model, presenting a higher variance in most cases.
The tables present a *RMSE* around the unit for nodule counting and a good balance between the precision and the recall for the binary tasks. This way, the nodule counting task reaches *RMSE*'s between $1.81 - 0.5$ $(2.21 - 0.92)$ at lobe level and $1.09 - 0.45$ $(1.22 - 0.41)$ at fold level for the proposed model (*BN+PReLU*). The binary tasks reaches $F_1$-scores within $0.98 - 0.85$ $(0.98 - 0.74)$ at lobe level and $0.98 - 0.79$ $(0.97 - 0.83)$ for *SELU* (*BN+PReLU*).

| **Manifestation/** | **Nodules** [RMSE] | | **Cavitations** [$F_1$] | | **Conglomeration** [$F_1$] | | **Consolidations** [$F_1$] | | **Tree in bud** [$F_1$] | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Lobe** | BN+PReLU | SELU | BN+PReLU | SELU | BN+PReLU | SELU | BN+PReLU | SELU | BN+PReLU | SELU |
| **Left Inf.** | $1.08_{0.66}$ | $0.94_{0.56}$ | $0.97_{0.13}$ | $0.95_{0.18}$ | $0.95_{0.15}$ | $0.95_{0.16}$ | $0.89_{0.18}$ | $0.88_{0.19}$ | $0.94_{0.14}$ | $0.93_{0.15}$ |
| **Left Sup.** | $0.92_{0.64}$ | $1.25_{0.84}$ | $0.95_{0.15}$ | $0.95_{0.16}$ | $0.86_{0.22}$ | $0.91_{0.17}$ | $0.90_{0.19}$ | $0.89_{0.23}$ | $0.94_{0.16}$ | $0.99_{0.07}$ |
| **Right Inferior** | $2.21_{0.38}$ | $1.28_{0.89}$ | $0.98_{0.08}$ | $0.95_{0.13}$ | $0.94_{0.12}$ | $0.96_{0.13}$ | $0.91_{0.14}$ | $0.96_{0.1}$ | $0.93_{0.15}$ | $0.89_{0.2}$ |
| **Right Middle** | $1.02_{0.77}$ | $0.5_{0.58}$ | $0.96_{0.11}$ | $0.98_{0.13}$ | $0.90_{0.12}$ | $0.91_{0.11}$ | $0.96_{0.13}$ | $0.93_{0.16}$ | $0.94_{0.13}$ | $0.85_{0.21}$ |
| **Right Supeior** | $2.20_{1.01}$ | $1.81_{0.88}$ | $0.88_{0.20}$ | $0.94_{0.18}$ | $0.87_{0.19}$ | $0.93_{0.14}$ | $0.89_{0.16}$ | $0.93_{0.15}$ | $0.74_{0.25}$ | $0.89_{0.18}$ |
| **Total** | $1.34_{0.69}$ | $\mathbf{1.16_{0.75}}$ | $0.95_{0.13}$ | $0.95_{0.16}$ | $0.91_{0.16}$ | $\mathbf{0.93_{0.14}}$ | $0.91_{0.16}$ | $\mathbf{0.92_{0.17}}$ | $0.90_{0.17}$ | $0.91_{0.16}$ |

Table 4.1 Predictions of the reported Tuberculosis manifestations per lung lobe (rows), and compared models, *Batch Normalization and PReLU*(BN+PReLU) and our model, referred as SELU (columns).

## 4.4 Discussion

The models training in this work presents a promising inference of the radiologist reports. Although there are no significant statistical differences between the compared models for our

| **Manifestation/** | **Nodules** [RMSE] | | **Cavitations** [$F_1$] | | **Conglomeration** [$F_1$] | | **Consolidation** [$F_1$] | | **Tree in bud** [$F_1$] | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Fold** | BN+PReLU | SELU | BN+PReLU | SELU | BN+PReLU | SELU | BN+PReLU | SELU | BN+PReLU | SELU |
| **1** | $0.73_{0.84}$ | $0.85_{0.35}$ | $0.88_{0.11}$ | $0.88_{0.12}$ | $0.90_{0.13}$ | $0.92_{0.12}$ | $0.88_{0.18}$ | $0.83_{0.22}$ | $0.83_{0.22}$ | $0.79_{0.23}$ |
| **2** | $1.15_{0.89}$ | $1.09_{0.83}$ | $0.86_{0.23}$ | $0.88_{0.22}$ | $0.94_{0.17}$ | $0.93_{0.18}$ | $0.93_{0.15}$ | $0.93_{0.18}$ | $0.97_{0.08}$ | $0.97_{0.08}$ |
| **3** | $0.41_{0.34}$ | $0.23_{0.39}$ | $0.85_{0.12}$ | $0.87_{0.11}$ | $0.89_{0.19}$ | $0.97_{0.11}$ | $0.96_{0.11}$ | $0.98_{0.03}$ | $0.95_{0.14}$ | $0.96_{0.12}$ |
| **4** | $1.22_{0.6}$ | $0.78_{0.74}$ | $0.94_{0.15}$ | $0.9_{0.19}$ | $0.9_{0.19}$ | $0.92_{0.15}$ | $0.88_{0.18}$ | $0.87_{0.18}$ | $0.87_{0.21}$ | $0.94_{0.15}$ |
| **5** | $0.41_{0.8}$ | $0.45_{0.8}$ | $0.93_{0.17}$ | $0.94_{0.17}$ | $0.94_{0.14}$ | $0.96_{0.12}$ | $0.90_{0.18}$ | $0.94_{0.14}$ | $0.91_{0.18}$ | $0.92_{0.17}$ |
| **Total** | $0.78_{0.69}$ | $\mathbf{0.68_{0.62}}$ | $0.89_{0.16}$ | $\mathbf{0.90_{0.17}}$ | $0.91_{0.16}$ | $\mathbf{0.94_{0.16}}$ | $0.91_{0.16}$ | $0.91_{0.15}$ | $0.91_{0.17}$ | $\mathbf{0.92_{0.15}}$ |

Table 4.2 Predictions of the reported Tuberculosis manifestations per fold lobe (rows), and compared models, *Batch Normalization and PReLU*(BN+PReLU) and our model, referred as SELU (columns).

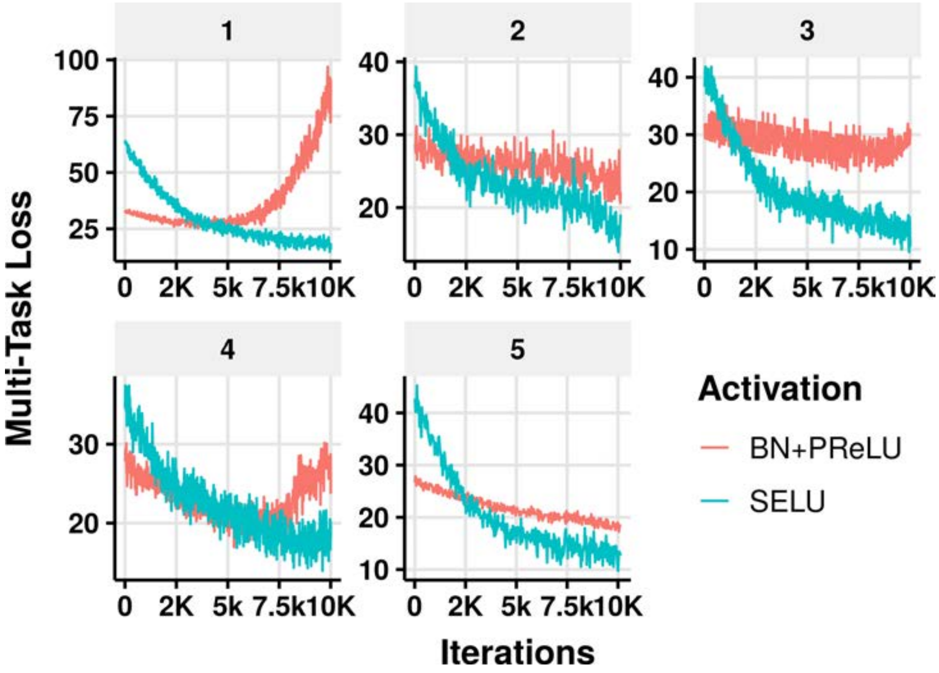


Fig. 4.2 Validation loss for 5-fold Cross-Validation. In red, the standard model which employs *Batch Normalization* [140] and *BN+PReLU*. In blue, our proposed model. *SELU*.

reduced dataset experiment, our model presents some advantages.

By avoiding normalization layers, the number of parameters of the model is reduced, so it is the computational complexity. The novel loss function presents a better model convergence that assures a more robust training and avoids possible overfitting problems common to the state-of-the-art approaches.

Besides, the results are consistent with known facts about the disease. Specifically, TB spreads inside the lungs starting at the right superior lobe, usually more affected by diverse lesions and many nodules, creating difficulties in the report generation, which is reflected as poorer predictions at the region. Similarly, the inference of the severely diseased subjects' reports is poorer when compared to those moderately affected by the disease, as can be observed for subjects #2 and #3 ( 4.2).

We acknowledge that the number of subjects used to validate the study is limited, and further validation is needed.
Namely, prospective studies are a must for new applications to ease the black box effect produced by the DL models (learnt from limited samples) that limits their explainability and generalization (see AI Learning Principles with Emphasis in Lung Analysis).

The lack of generalization is not only due to the scarcity of observations but also to their simplified characterization as $p(y|x)$ instead of $p(x,y)$ since, as explained (see 1.4.1), the latter is hardly tractable using the statistical learning framework.
Although it is not ideal due to the data scarcity, circumvention of the generalization problem is possible by implementing a model for each new domain after applying *transfer learning* [255] or from scratch (e.g., datasets belonging to other animal models in the context of ERA4TB or application to other lung diseases). Remarking such an approach is valid when the model just intends to achieve the highest possible predictive power on data belonging to the same distribution [34, 275].
On the contrary, establishing informative (explainable) models can be vital for the correct understanding of the disease. However, the statistical learning framework is insufficient to establish causal relationships between the correlations found by the mere statistical characterization $p(y|x)$ or $p(x,y)$.

Not in vain, the following chapter presents an approach based on graphical causality embedding statistical learning.
The meeting of both modelling frameworks establishes a causal model governing $p(x,y)$, as shown by the results of the method in terms of generalization and ability to generate realistic images for the specific case of TB-infected lung imaging in different animal models.

## 4.5 Conclusion

The chapter introduces a novel methodology for multi-label classification, which enables the inclusion of *Self-Normalizing Networks* within 3D CNNs. This approach allows an improved extraction of relevant features on large medical volumes through multi-task learning guided by the uncertainty in the model predictions. Therefore, demonstrating that DL models with designs adapted to the context of this thesis allow the extraction of essential information for the characterization of TB. This fact represents a significant advance in the field, even bearing in mind the limitations mentioned for *explicability* and *generalization* terms addressed in the next chapter.

# Chapter 5

# Translational Lung Imaging Analysis Through Disentangled Representations

The development of new treatments often requires experiments with translational animal models using (pre)-clinical imaging to characterize inter-species pathological processes (see Project Framework: ERA4TB).
To accelerate the development, this thesis presents a collection of novel methods to automatize related tasks and promote explainable Tuberculosis (TB) models for imaging biomarkers extraction.
Thus, the previous chapter (Deep Learning for TB Manifestation Classification) shows how Deep Learning (DL) models are commonly used to automate relevant information extraction from the images. However, as usual for automation tasks, the proposed DL model is specific for a domain (animal model) (see Section 1.4.1). This is due to low generalizability and explainability, a product of their entangled design (see AI Learning Principles with Emphasis in Lung Analysis).
Consequently, it is quite complicated to take advantage of the proven high capacity of DL to discover statistical relationships [10, 194] from inter-species images. Note that discovering such relationships would be essential to establish a shared disease marker beyond the particular manifestations of TB for each animal model.

In this chapter, we present a model to leverage such capacities. Concretely, we extract it from our publicly available work: *Translational Lung Imaging Analysis Through Disentangled Representations* [99].

This model extracts disentangled information from images of different animal models. Such an approach allows characterizing common mechanisms of the data generation, as is proven synthesizing realistic chest Computed Tomography (CT) volumes.

Our method stands at the intersection between deep generative models, disentanglement and causal representation learning. In this thesis context, it is optimized from images of pathological lung infected by Tuberculosis and is able: a) from an input slice, infer its position in a volume, the animal model to which it belongs, the damage present and even more, generate its lung delimitation mask (similar overlap measures to the *nnU-Net* [141]), b) generate realistic lung images by setting the above variables and c) generate counterfactual images, as healthy versions of a damaged input slice.

## 5.1 Introduction

Understanding disease progression is essential to develop new treatments ( From a binary perspective to continuous spectrum of diagnosis). The longitudinal characterization of animal models in (pre-)clinical experiments is crucial [339]. For this, we need to extract comparable biomarkers in similar phase of the pathology (Fig. 1.5). We also need to proof the existence of similar pathophysiological mechanisms modulating common causal factors, that give rise to the variability of trial outcomes (see Eradicating Tuberculosis: The need for continuous assessment).

In this context, medical imaging techniques enable the extraction of indicators (imaging biomarkers) from *in vivo* studies [330]. For example, the number of *Mycobacterium tuberculosis* (*Mtb.*) colonies present in a subject can be approximated[1] from the damaged lung volume in an image of a human, primate, or mouse [339] (see 1.3).

The images contain meaningful information to interpret the mentioned physiological process. However, their analysis is tedious and automation is advantageous to process the vast amount of data produced during the trials. Thus, developing Artificial Intelligence (AI) systems that can not only automate the extraction of particular markers for each animal model (e.g., the damaged lung volume) but are also capable of inferring the common agents of such particular indicators (e.g., bacterial burden) is essential (see Computer Aided Diagnosis: The way to automated quantification).

Although AI, has eased the process [123, 351], some design premises has lessened its inference capabilities. In particular, DL models excel at extracting the statistical dependence between input-output pairs, i.e.,$(x_i, y_i) \in \mathscr{X}, \mathscr{Y}$, from assumed *independent and identically distributed (i.i.d.)* observational data [249] (see 1.4.1). Such success has leaned the model

[1] The number of colonies correlates with some radiological manifestations of TB

designs towards an insufficient representation learning strategy [22]. Namely, the discovery of statistical dependence between specific data pair samples is priorized rather than the understanding of the physical model generating the whole data population (e.g., physiological mechanisms).

Since the i.i.d. assumption is fragile, data scarcity (especially for labelled data) and data mismatch are characteristic of the medical imaging field, well-known distribution shifts [42] between data employed at training, validation/test phases and “real world" data are usual. Under this scenario, the models tend to learn correlated representations that only hold for specific environments or domains, namely *spurious correlations* [10]. Since (as a mantra)*“correlation does not imply causation"*, such flaws cause ruinous effects [62, 261] for generalisation, transferability and explainability purposes [275].

More formally, naive DL models maximize a joint distribution, $p(X,Y)$ or $p(X)$ (self-supervision), characterized by an entangled representation of the input. Namely, if $X$ and $Y$ correlate during training without necessarily derive from a causal representation ($X \rightarrow Y$), $p(X,Y)$ can adopt numerous (specific domain dependant) factorization forms [100]. Thus, there is a need to implement independent models to process the information in related domains (particularly, lung CT images of TB animal models). Such models are put in common through *posthoc* analysis, losing possible data synergies.
In general, state-of-the-art learning strategies, mitigate this issue by shrinking the $p(X,Y)$ solutions space. To this aim, models are enriched injecting inductive biases (e.g., CNNs assume spatial correlation [74], equivariant transformations [36]), to facilitate the discovery of more meaningful and disentangled representations [191].
The above mentioned techniques simulate human cognition. Under the realms of causality [240], human cognition arranges the proper biases to extract a limited number of relevant factors related to a task holding among different environments.

Designing AI systems can follow a similar causal perspective. We can introduce specific biases to shrink the solution space. Thus, in this chapter, we consider: a) the strongly hierarchical nature of the human visual system and b) the data generation process. Such an approach intends to mimic the radiologists’ tasks, who take into account specific patient factors (i.e., clinical history, sex, age) beyond the image *per se*.
This approach yields more effective disentangled representations of the input [275].

In particular, we intend to identify the unique mechanisms in the generation of translational imaging of lung Computed Tomography (CT) images and their corresponding segmentation masks (Fig. 5.1). We employ three different animal models (mouse, primate and human) infected by *Mtb.*[235].

From a simplified radiological point of view, mammals' lungs share texture and shape features. We model these shared characteristics as an effect of the same causative factors, for example, the bacterial load (see 1.3).

To prove the benefits of our strategy, we show how after optimizing the model employing a small limited number of volumes, our design can:

- Produce a very accurate reconstruction of the input images and generate suitable segmentation masks (Fig.5.7, Table 5.3 and Table 5.4).
- Generate new realistic images of the three models controlling the lung damage on each, which implies the proper characterization of the disentangled variables (Fig. 5.3).
- Generate counterfactual images [58, 276] of damaged lungs. Namely, the model is able to capture the meaningful representations of an input image to convert it into a healthy version by intervening on the damage variable value.

## 5.2 Methods

We define a generative model in which the high dimensional texture and shape features that can be extracted from lung CT images and their corresponding segmentation masks are a result of the causal Direct Acyclic Graph (DAG) presented in Fig. 5.1.

The proposed DAG simplify the physical image generation for obvious reasons. All the possible elementary causative factors (specific scanner, comorbidities, subject age, sex, etc.) are reduced to three: the animal model, $A$, the observed lung axial slice, $S$, and the lung damage, $D$. The causative factors are modelled as three groups of independent variables, $\mathbf{z}^0$, under the noise term, $\varepsilon_{\{A,S,D\}}$, which comprises noise and unconsidered variables. The primary variables govern the generative process which follows a part-whole hierarchy [124] from low-level representations of the texture and shape features, $\mathbf{z}^1$, to high dimensional ones, $\mathbf{z}^k$, the observed image, $\mathbf{x}$ and the segmentation mask, $\mathbf{y}$. This part-whole hierarchy resembles brain columns functioning [66, 192]. Variable superscripts, $\mathbf{z}^k$, symbolize hierarchy levels at the DAG.

The plate notation at the DAG represents such upsampling generation. The DAG implements two paths diverging at the first hierarchy level (shared representation path), $\mathbf{z}^1$. The division forces, during optimization, to generate a disentangled representation of shape, $\mathbf{z}_L$ and texture,

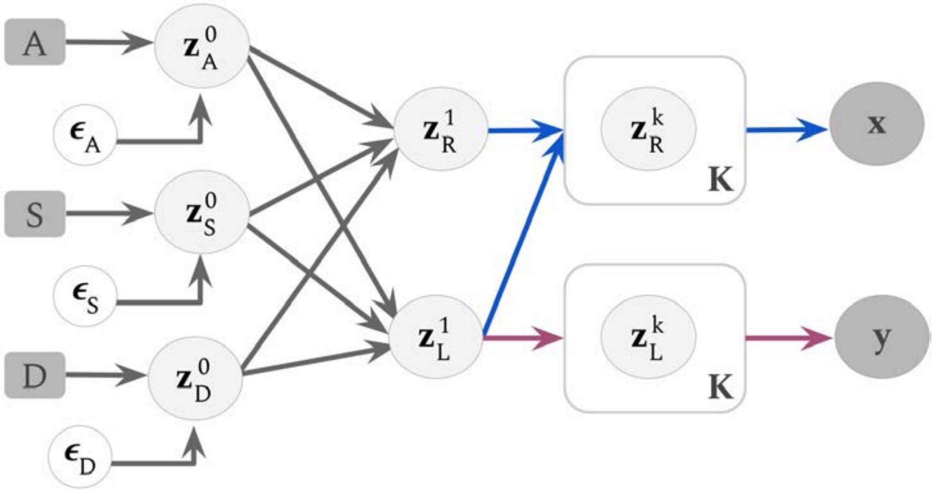


Fig. 5.1 Direct Acyclic Graph (DAG) representing the generation of pathological lung CT images **x**, and their segmentation masks **y**. Both generated from a latent variables hierarchy at different resolutions scales, *K*, governed by three factors, i.e., animal model, *A*, the relative position of the axial slice, *S*, and the estimated lung damage caused by *Mtb.*, *D*.

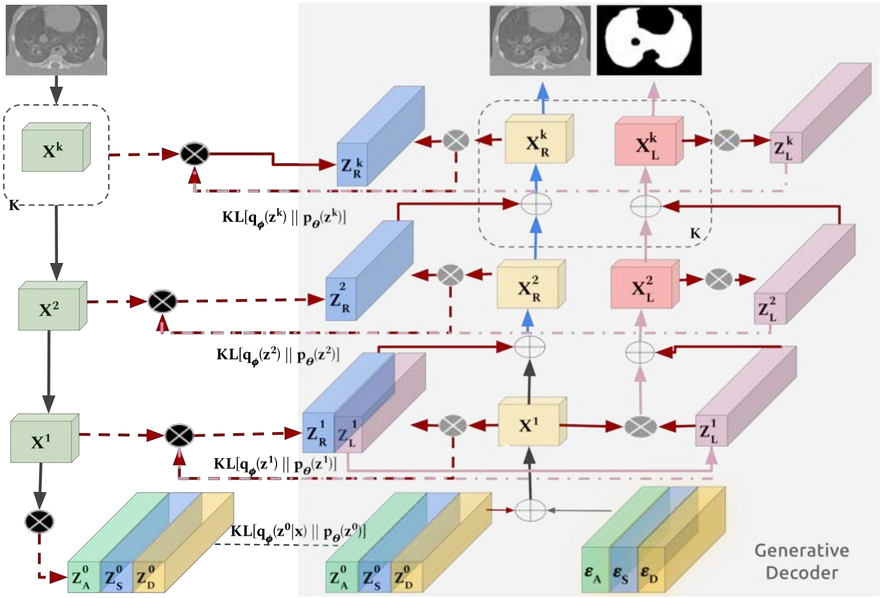

$X^k$
K
$X^2$
$X^1$
$Z^k_R$
$X^k_R$
$X^k_L$
$Z^k_L$
$KL[q_\phi(z^k) || p_\theta(z^k)]$
K
$Z^2_R$
$X^2_R$
$X^2_L$
$Z^2_L$
$KL[q_\phi(z^2) || p_\theta(z^2)]$
$Z^1_R$
$Z^1_L$
$X^1$
$Z^1_L$
$KL[q_\phi(z^1) || p_\theta(z^1)]$
$KL[q_\phi(z^0|x) || p_\theta(z^0)]$
$Z^0_A$ $Z^0_S$ $Z^0_D$
$Z^0_A$ $Z^0_S$ $Z^0_D$
$\varepsilon_A$ $\varepsilon_S$ $\varepsilon_D$
Generative Decoder

Fig. 5.2 DAG representing the generation of pathological lung CT images **x**, and their segmentation masks **y**. Both generated from a latent variables hierarchy at different resolutions scales, *K*, governed by three factors, i.e., animal model, *A*, the relative position of the axial slice, *S*, and the estimated lung damage caused by *Mtb.*, *D*.

$\mathbf{z}_R$. CT images depend on both shape and texture variables (blue path), while the segmentation masks only depend on shape variables (pink path). Then, assuming the independence of the noise terms, the (*independent causal mechanism* (ICM) principle is fulfilled [275]) and the following disentangled factorization arise:

$$p(\mathbf{x},\mathbf{y},\mathbf{z}) = p(\mathbf{x}\,|\mathbf{z}_R^K)p(\mathbf{y}\,|\mathbf{z}_L^K)p(\mathbf{z}_R^k,)p(\mathbf{z}_L^k)p(\mathbf{z}_R^2|\mathbf{z}_R^1,\mathbf{z}_L^1)p(\mathbf{z}_R^1,\mathbf{z}_L^1|\mathbf{z}^0)p(\mathbf{z}^0) \tag{5.1}$$

being

$$p(\mathbf{z}_R^k) = \prod_{k=3}^{K} p(\mathbf{z}_R^k|\mathbf{z}_R^{k-1}); \qquad p(\mathbf{z}_L^k) = \prod_{k=2}^{K} p(\mathbf{z}_L^k|\mathbf{z}_L^{k-1}); \quad p(\mathbf{z^0}) = p(\mathbf{z}_A^0)p(\mathbf{z}_S^0)p(\mathbf{z}_D^0); \tag{5.2}$$

### 5.2.1 Model optimization

For the above equations, each conditional distribution is parametrized by depthwise convolutional decoders with parameters $\theta$, leveraging a high capacity model (Fig. 5.2) allowing to characterize the unobservable causes of variation ($\varepsilon$) consistent with the available data (lung CT images) [238, 249]. Once the model is optimized, is it possible to modify the disentangled variables to obtain new generated images (5.3.3) and counterfactual images [58, 276] (see 5.3.4).

The computation of the parameters requires optimization through training of the posterior probability, $p_\theta(\mathbf{z}\,|\,\mathbf{x},\mathbf{y})$, which is intractable. To tackle this issue, we adapt the particular factorization in eq:factorization to the methodology developed for deep Variational Autoencoders (deep VAEs) [49, 162]. In this way, we obtain the best approximate amortized posterior distribution, $q_\phi(z|x)$, being $\phi$ the parameters of the encoder. Notice that the distribution is amortized just from $\mathbf{x}$ (not from $\mathbf{y}$), so we force the model to extract the meaningful mechanism to generate the segmentation masks just from the self-supervisory signal of the image [175]. Indeed, we add a segmentation branch in the architecture (Fig. 5.2), dependent on the main branch.

Namely, we adopt the Noveau VAE (NVAE) [315]. This architecture is carefully designed for hierarchical models. Moreover, it has proven efficacy in approximating posteriors by introducing as inductive bias in the image generating process a deeply hierarchical architecture. To this aim, the set of $\mathbf{z}$ variables at each representation level $k$, is divided in smaller sets, $m_k$, to get a total of $M$ groups of latent variables. They establish a hierarchical structure within each resolution too to help narrowing the solutions space, being **z** the set:

$$\mathbf{z} = \left\{\{(\mathbf{z}_A, \mathbf{z}_S, \mathbf{z}_D)_0, \mathbf{z}_1, \mathbf{z}_2 \ldots, \mathbf{z}_{m_{k=0}}\}^0, \{(\mathbf{z}_L, \mathbf{z}_R)_{m+1}, \ldots, \mathbf{z}_{m_{k=1}}\}^1, \ldots, \{\mathbf{z}_{m+1}, \ldots, \mathbf{z}_{m_k}\}^k, \{\mathbf{z}_{m+1}, \ldots, \mathbf{z}_M\}^K\right\} \tag{5.3}$$

Its prior and approximate posterior probability are given by:

$$p_\theta(\mathbf{z}) = \prod_m p_\theta(\mathbf{z}_m|\mathbf{z}_{m-1}) \qquad q_\phi(\mathbf{z}|\mathbf{x}) = \prod_m q_\phi(\mathbf{z}_m|\mathbf{z}_{m-1}, \mathbf{x}). \tag{5.4}$$

Following this formulation, from marginalization of the log of 5.1 and rearranging terms, we obtain the variational lower bound to optimize (subscripts colors denote each optimization branch):

$$\mathscr{L}(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[log p_\theta(\mathbf{x}|\mathbf{z})\right] - KL(q_\phi(\mathbf{z}_0|x)||p_\theta(\mathbf{z}_0)) + \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[log p_\theta(\mathbf{y}|\mathbf{z})\right] - \mathbb{E}_\mathbf{z}\left[KL_\mathbf{z}\right] - \mathbb{E}_\mathbf{z}\left[KL_\mathbf{z}\right] \tag{5.5}$$

being $KL$ the Kullback–Leibler divergence and

$$\mathbb{E}_\mathbf{z}\left[KL_\mathbf{z}\right] = \sum_m^M \mathbb{E}_{q_\phi(\mathbf{z_{m-1}}|\mathbf{x})}\left[KL(q_\phi(\mathbf{z}_m|\mathbf{z}_{m-1}, \mathbf{x})||p_\theta(\mathbf{z}_m|\mathbf{z}_{m-1}))\right] \tag{5.6}$$

Being $q_\phi(\mathbf{z}_{m-1}|\mathbf{x})$ the approximate posterior through the hierarchy of $m_{k-1}$ group.
Since, NVAE convergence depends on the proper approximation of KL terms (see [315]). To this aim, all priors and posterior probabilities are approximated as Normal distributions. Thus, we can write:

$$p(\mathbf{z}_A^0) \sim \mathscr{N}(\mu(a), \sigma(a)); \qquad p(\mathbf{z}_S^0) \sim \mathscr{N}(\mu(s), \sigma(s)); \qquad p(\mathbf{z}_D^0) \sim \mathscr{N}(\mu(d), \sigma(d)); \tag{5.7}$$

## 5.3 Experiments and Results

### 5.3.1 Datasets description

The model is optimized employing small datasets: ten lung CT volumes per animal model ($\sim$ 2000 slices). The data used for the optimization phase (training) are axial slices from three models of pathological lungs infected by *Mtb*. The dataset names identify: the animal model, $A$, the data source and the phase as follows $A^{source}_{phase}$). Namely, the human volumes, $H^{CLE}_{tr}$ , corresponds to the validation data of the 2019 ImageClefMed TB task [68]. The mice images, $M^{GSK}_{tr}$, are provided from *GlaxoSmithKline plc.* (GSK) within the context of the ERA4TB project [76], similarly to the primate ones, $P^{PHE}_{tr}$, from the *Public Health of England* (PHE) [95, 98]. For testing (twenty volumes per model), $P^{PHE}_{ts}$ and $P^{GSK}_{ts}$, are selected from different cohorts of $P_{tr}^{PHE}$ and $P_{tr}^{GSK}$, while the human dataset, $H_{ts}^{CLE}$ is a partition of the mentioned data. The remaining sets are included to evaluate the model generalisation and transferability capabilities. $M_{ts}^{EXM}$ belongs to a publicly available dataset from the Institute for Experimental Molecular Imaging (ExMI) [265] which contains healthy subjects at low resolution. Finally, the human dataset, $H_{ts}^{RAD}$, presents subjects with lung damage caused by COVID-19 [57].

Note that all datasets include segmentation masks produced/corrected by trained experts.
A detailed description of the different datasets is presented in Table 5.1.

| Dataset ID | Phase | Animal Model | Source | Voxel Spacing [mm] | Resolution |
|---|---|---|---|---|---|
| $M_{tr}^{GSK}$ | Training | Mouse | GSK | $0.087 \times 0.087$ | $500 \times 500$ |
| $M_{ts}^{GSK}$ | Test | | | | |
| $M_{ts}^{EXM}$ | | | ExMI | $0.282 \times 0.282$ | $144 \times 100$ |
| $P_{tr}^{PHE}$ | Training | Primate | PHE | $0.235 \times 0.235$ | $512 \times 512$ |
| $P_{ts}^{PHE}$ | Test | | | | |
| $H_{tr}^{CLE}$ | Training | Human | ImageClef | 0.60-0.75 × 0.60-0.75 | $512 \times 512$ |
| $H_{ts}^{CLE}$ | Test | | | | |
| $H_{ts}^{RAD}$ | | | Radiopedia | 0.68-0.75 × 0.68-0.75 | 512-630 × 430-630 |

Table 5.1 Datasets description

### 5.3.2 Implementation details

The model is optimized employing six scales, $K = 6$, with 18 latent variables per scale, partitioned in $m_k$ groups per scale as follows, $m_k = [2,2,2,3,6,9]$ The three $\mu_A$, $\mu_S$ and $\mu_D$ per prior are known during training ($\mu_A = [-1,0,1]$, $\mu_D = (0,1)$, $\mu_S = (0,1)$), fix at image generation and inferred for image reconstruction and segmentation mask generation employing $KL\big(q_\phi(z^0)||\mathcal{N}(0,1)\big)$. Note that $\mu_D$ during optimization is given by the relative volume of the healthy lung (extracted by simple thresholding) with respect to the whole ground truth mask volume.

### 5.3.3 Pathological Lungs Generation

After optimization, the model is able to generate realistic images, such as those shown in fig:generated, by choosing the mean values of $\mathbf{z}_A^0$, $\mathbf{z}_S^0$, $\mathbf{z}_D^0$ factors. To illustrate this capacity in Fig. 5.3, we set a relative slice position of 0.5, the animal model is fixed for each row and, the effect of the lung damage variable is modulated from lower to higher in each column.

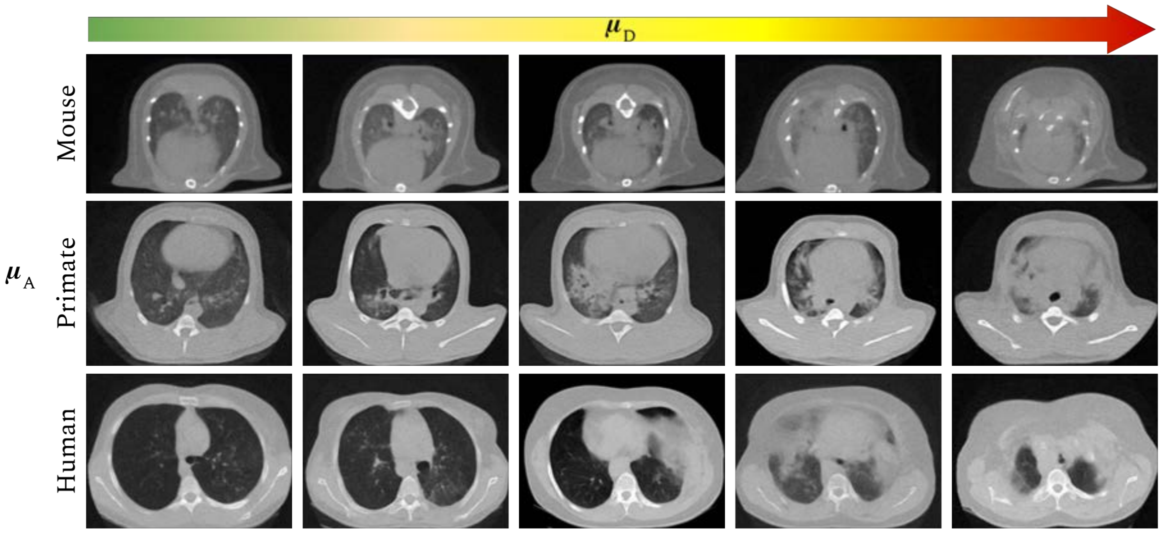


Fig. 5.3 Synthetic lung CT images generated by our model. Images are generated with a fix slice relative position, ($\mu_S$). For each row, the animal model $\mu_A$ is fixed to $-1, 0, 1$, respectively, while for each column, the damage $\mu_D$ is increased [0-1].

#### 5.3.3.1 Pathological Lungs Generation: Varying the slice position

This appendix shows generated slices instances fixing the damage and varying the relative slice position. This experiment extends the previous section, in which axial slices belong to a fixed relative slice position.

Since our chest CT volumes orientation is cephalic to caudal, the model generates axial images of the upper airways (trachea) and the corresponding per animal model surrounding tissues at the lowest slice position, as shown in the first column of the Fig. 5.4. This way, the second column shows the corresponding generated anatomy for the superior lungs, while the third and fourth columns accordingly show the middle and inferior regions. Finally, the fifth column depicts the generated version at the beginning of the abdominal anatomy.

### 5.3.4 Counterfactual Images

The first column of each row in Fig. 5.5 shows a real image of a damaged lung corresponding to a given animal model. When no actions are performed, the model infers the disentangled image representation of the causative variables ($\mathbf{z}_A^0$, $\mathbf{z}_S^0$, $\mathbf{z}_D^0$) through the encoder. Subsequently, the image is reconstructed and a segmentation mask is generated employing the optimized decoder (Fig. 5.2). The second column shows a healthy counterfactual of the input images. Each counterfactual image is generated after intervening on the inferred

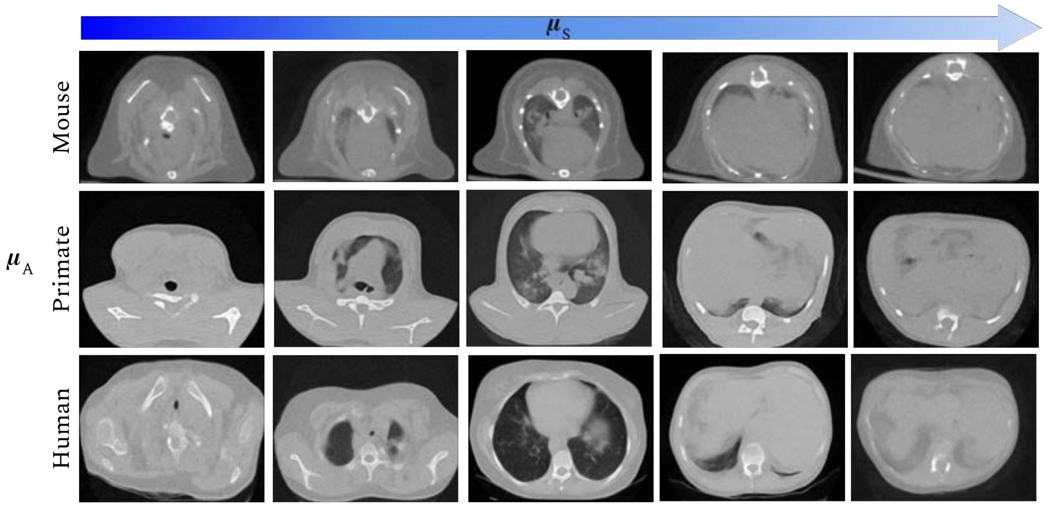


Fig. 5.4 Synthetic lung CT images generated by our model. Images are generated with a fix relative damage, $\mu_D = 0.5$. For each row, the animal model $\mu_A$ is fixed to $-1, 0, 1$, respectively, while for each column, the relative slice position $\mu_S$ is increased between 0 and 1.

damage variable, $\mathbf{z_D^0}$. To this aim, its mean value is set to 0. The decoder is fed with the zero-mean counterfactual image and the rest (unaltered) inferred causal variables.

#### 5.3.4.1 Counterfactual Images: Extended Assessment

This appendix extends the qualitative results presented in Section 5.3.4. The former section shows the model capacity generating counterfactual images and their respective segmentation masks.

Here, we evaluate how realistic are the generated images. For that, we compare the Hounsfield Units (HU) of real CT slices with two cases: a) the reconstructed slice from the variable inferred by the encoder without modification of any of these values, and b) the counterfactual image, namely, after intervening on the inferred damage value. We compute the voxel-wise Root Mean Square Error (RMSE) for the reconstructed images per test dataset. Table 5.2 shows these results with an average $RMSE = 18.73 \pm 2.16$.

Voxel-wise evaluation is not suitable for counterfactual images. Previous manual delimitation of comparable regions is needed, which is a priority for our future work.

To illustrate similarities and differences in the HU scale, in Fig. 5.6, we plot the HU profile belonging to the damaged regions shown in Fig. 5.5. Respectively, the first three rows contain 1) the original axial slice from the different test datasets (the image is generated from the $\mu_a$, $\mu_s$ and $\mu_d$ inferred by our model), with the profile horizontal line in green, 2) the reconstructed slice (the image is generated maintaining $\mu_a$, $\mu_s$ inferred by our model

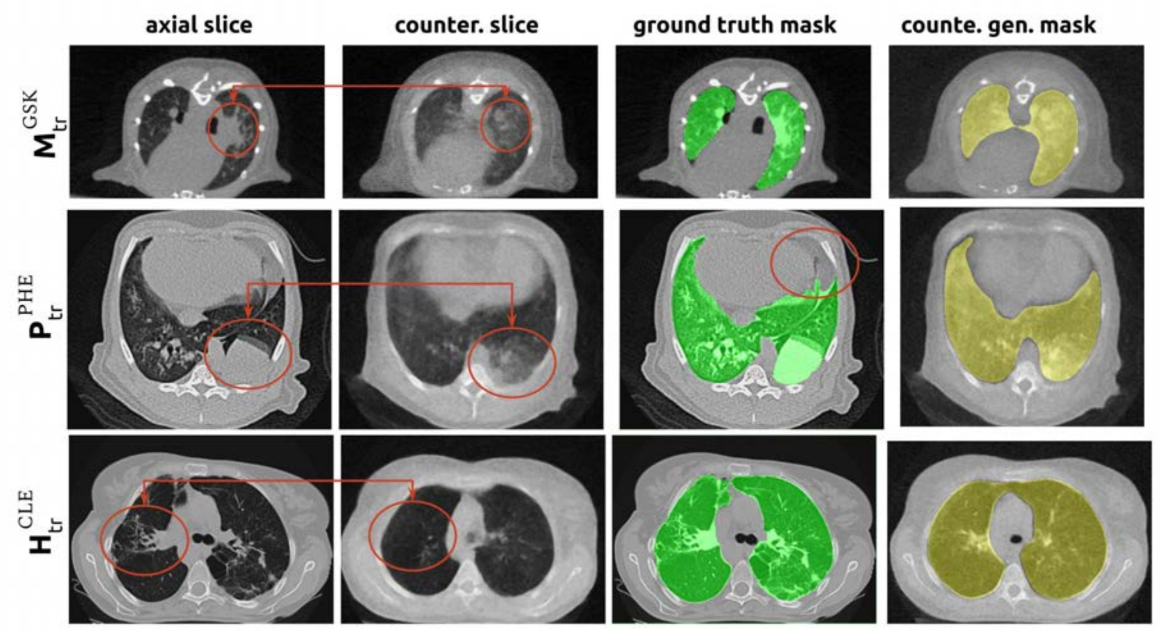


Fig. 5.5 he encoder infers the real image (axial slice) disentangled representation, $\mathbf{z}_A^0$, $\mathbf{z}_S^0$, $\mathbf{z}_D^0$. By setting the damage variable $\mathbf{z}_D^0$ to 0 the decoder generates the healthy counterfactual (counterfactual slice) and its respective mask (counterfactual mask).

and correcting $\mu_d$), with profile line in yellow and 3) the counterfactual after modifying the inferred expected damage, with the profile line in blue.

The last row shows the HU plot for each profile-specific colour. HU values are similar for the three slices except for those regions where the slice counterfactual version replaces the damage with healthy tissue-like. We highlight such changes framing them in vertical dashed red lines.

Besides, it is important to note that the original and reconstructed images present more noisy patterns than the counterfactual version, as was expected from its blurrier appearance and the thickening of the soft tissue for the mice dataset.

Table 5.2 Root Mean Square Error (RMSE) between the real images and the image reconstructed from the $\mu_a$, $\mu_s$ and $\mu_d$ inferred by our model for the test datasets

| RMSE[HU] | | | | |
|---|---|---|---|---|
| $M_{ts}^{GSK}$ | $M_{ts}^{EXT}$ | $P_{ts}^{PHE}$ | $H_{ts}^{CLE}$ | $H_{ts}^{COV}$ |
| 21.26 | 18.75 | 20.12 | 17.89 | 15.63 |

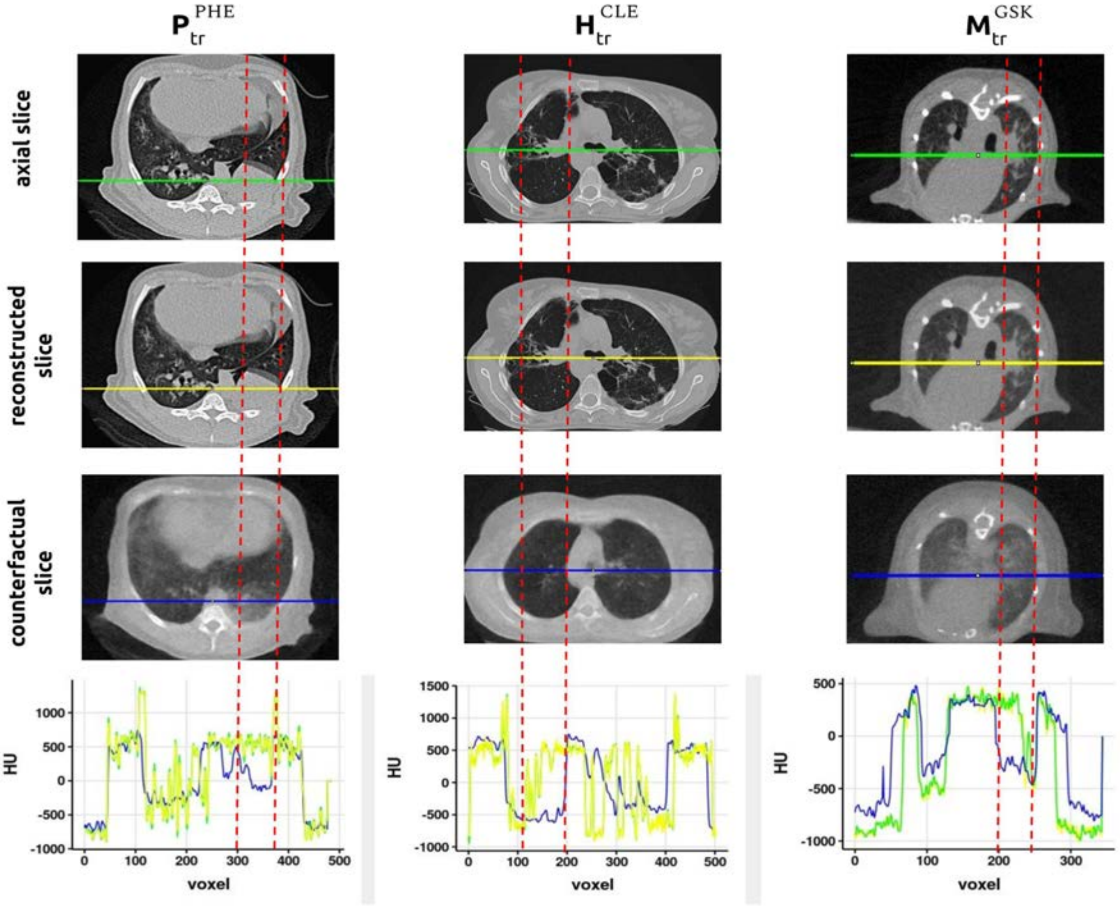


Fig. 5.6 Hounsfield Units (HU) plots for profiles at regions damaged in original test axial slices. Each column contains instances of each dataset, previously employed in Section 5.3.4. The first rows depict the original, reconstructed and counterfactual slices with the profile line green, yellow and blue, respectively. The last row draws the HU profiles per voxel. Vertical dashed lines highlight big differences between real/reconstructed and counterfactual slices.

### 5.3.5 Segmentation employing counterfactual images

Pathological lung segmentation is an important task for drug development studies. Unfortunately, it a complex task due to the difficulty of discrimination between lesions and other neighborhood tissues with the added difficulty of the diversity of the biological data [126]. In this experiment, we retrain the optimized model with counterfactual images, such as the obtained in the previous experiment (see 5.3.4), to generate the segmentation masks from the test datasets 5.3.1. To measure the strengths and weaknesses of this generative approach, we compare the obtained results, $our_c$, with the segmentation masks calculated by our method optimized before employing counterfactual images, $our_{nc}$, and the state of the art full supervised method, *nnU-Nnet* [141].

| | $\text{DSC}_{\pm SD}$ | | | | |
|---|---|---|---|---|---|
| | $M_{ts}^{GSK}$ | $M_{ts}^{EXT}$ | $P_{ts}^{PHE}$ | $H_{ts}^{CLE}$ | $H_{ts}^{COV}$ |
| nnU-Net | $0.845_{\pm 0.10}$ | $0.851_{\pm 0.11}$ | $0.957_{\pm 0.06}$ | $0.978_{\pm 0.04}$ | $0.973_{\pm 0.03}$ |
| $our_{nc}$ | $0.849_{\pm 0.10}$ | $0.843_{\pm 0.12}$ | $0.949_{\pm 0.06}$ | $0.963_{\pm 0.06}$ | $0.963_{\pm 0.06}$ |
| $our_c$ | $0.877_{\pm 0.08}$ | $0.859_{\pm 0.11}$ | $0.955_{\pm 0.06}$ | $0.977_{\pm 0.06}$ | $0.968_{\pm 0.04}$ |

Table 5.3 Mean and standard deviation (SD) of the Dice Similarity Coefficient (DSC) between the ground truth masks and mask obtained from the methods indicated at rows (*nnU-Nnet*, proposed method before employing counterfactual images ($our_{nc}$), and after ($our_c$)) for each test dataset (columns).

The Table 5.3 and Table 5.4 show the mean and standard deviation for Dice Similarity Coefficient (DSC) and Hausdorff Distance (HD) between each segmentation method and the ground truth masks for each test dataset. The results present an improvement for all measures and datasets when employing counterfactual images, yielding similar results to the *nnU-Nnet*.

| | $\text{HD}_{\pm SD}$ [mm] | | | | |
|---|---|---|---|---|---|
| | $M_{ts}^{GSK}$ | $M_{ts}^{EXM}$ | $P_{ts}^{PHE}$ | $H_{ts}^{CLE}$ | $H_{ts}^{COV}$ |
| nnU-Net | $1.737_{\pm 1.01}$ | $1.90_{\pm 1.52}$ | $3.30_{\pm 3.96}$ | $9.37_{\pm 15.14}$ | $8.31_{\pm 10.71}$ |
| $our_{nc}$ | $1.948_{\pm 1.11}$ | $2.06_{\pm 1.82}$ | $3.81_{\pm 4.10}$ | $10.12_{\pm 18.32}$ | $10.56_{\pm 10.77}$ |
| $our_c$ | $1.519_{\pm 0.89}$ | $1.88_{\pm 1.53}$ | $2.95_{\pm 3.54}$ | $8.78_{\pm 16.11}$ | $9.48_{\pm 9.89}$ |

Table 5.4 Mean and standard deviation (SD) of the Hausdorff Distance (HD) between the ground truth masks and mask obtained from the methods indicated at rows (*nnU-Nnet*, proposed method before employing counterfactual images ($our_{nc}$), and after ($our_c$)) for each test dataset (columns).

The differences are due to subtle changes in most of the cases or even small imperfections in the ground truth masks as it is shown in Fig. 5.7.

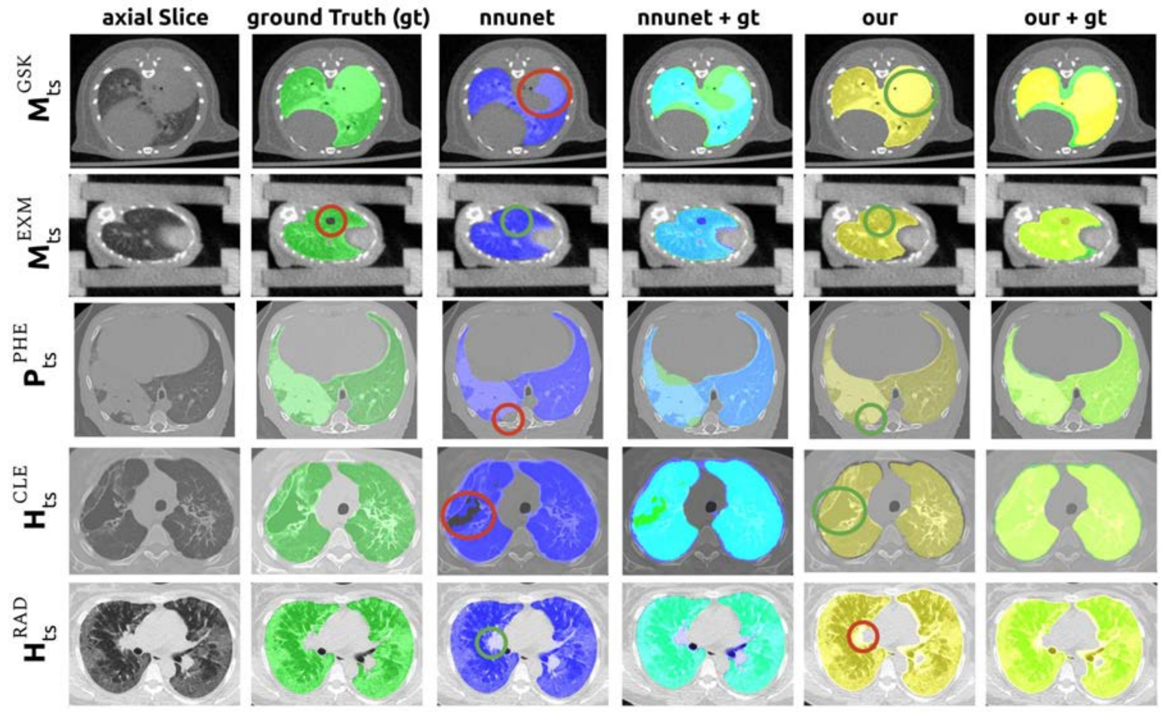


Fig. 5.7 Comparison of methods. Each row contains axial slices and segmentation masks of each test dataset. Columns show the original CT image, ground truth mask (gt), *nnU-Net* mask, overlay of *nnU-Net* and ground truth, the mask with our method employing counterfactual images during training and the overlay with the ground truth.

#### 5.3.5.1 Comparison with Chapter 2 Rule-Based Method

For completeness, these paragraphs recover the results obtained with the traditional method of Chapter 2. In that chapter, we show the qualitative performance of DL and the rules-based approaches for segmentation tasks to illustrate the issues of traditional methods under new environments.

Actually, the instance slices in the Fig. 2.11 coincide with the ones in the Fig. 5.7 except for $M_{ts}^{EXM}$ (not suitable for the traditional approach) and the row for the mild TB primate model, $P^{PHE_1}$, in figure 2, already included in it.

Here, we complete the analysis for the 156 slices specifically selected for Chapter 2 experiments. To this aim, the Table 5.5 adds the DSC and HD to the previous chapter results

(Table 2.1) employing the DL methods, both the *nnU-Net* and the two variants of our method. As can be seen, the DL method overpasses the performance of traditional ones.

| | $\text{DSC}_{\pm SD}$ | $\text{HD}_{\pm SD}$ [mm] |
|---|---|---|
| | $P^{CH_2}$ | $P^{CH_2}$ |
| No DL (Section 2.2.2) | $0.933_{\pm 0.03}$ | $8.642_{\pm 7.36}$ |
| nnU-Net | $0.981_{\pm 0.05}$ | $5.954_{\pm 5.11}$ |
| $our_{nc}$ | $0.977_{\pm 0.06}$ | $6.897_{\pm 5.87}$ |
| $our_{c}$ | $0.982_{\pm 0.06}$ | $6.142_{\pm 6.01}$ |

Table 5.5 DSC and HD for the 156 slices defined at Appendix A and employed at the experiments in Chapter 2, $P^{CH_2}$ (see 2.2.2) belonging to a mild TB primate model, which results appears at Fig. 2.11 and Table 2.1.

## 5.4 Conclusions

The methodology proposed in this work yields promising results for obtaining the shared factors between animal models that characterize the pathophysiological processes. Beyond the existing limitations, such as the use of isolated axial slices instead of more informative whole three dimensional images or the characterization of damage based simply on the damaged volume and not on the specific manifestations of the disease for each animal model, our model is capable of inferring meaningful disentangled representations.

The model can generate synthetic axial slices by setting the values of the modelled factors. Even more relevant, it produces counterfactual versions of existing axial slices by testing the effective disentanglement. This capability leverages potential benefits in the field of translational image analysis. Extending the diversity of existing data, essential for automatic segmentation, or providing the damage variable as a possible (to be validated) inter-species biomarker.

# Chapter 6

# Conclusions and Prospective work

## 6.1 Conclusions

This thesis develops different methods to automate fundamental CT image analysis tasks for TB infected lungs. We also facilitate the interpretation of CT images belonging to various animal models of TB infection, which is essential for the definition of helpful imaging biomarkers for anti-TB new drugs development. The works are presented from lower to a higher degree of complexity in response to the considered problems nature. The developed models embed in different Artificial Intelligence (AI) strategies and fulfil the thesis objectives specified in the introductory chapter.

The method based on a traditional formulation presented in Chapter 2 allows segmenting damaged lungs automatically. This initial approach enables both subsequent quantification analyses and the extraction of fundamental knowledge in the implementation of higher capacity Deep Learning (DL) algorithms. The method presents limitations, such as the inability to provide lung segmentation for significant organ damage or animal models other than a macaque. However, it can assist experts during lung delimitation of other mammals by readjusting the adequate parameters (see Section 2.2). The method validity in key experiments relies on the quantification results given in the second part of the same chapter. In Section 2.3, we introduce an automatic quantification method of TB burden. The procedure, which may be considered simple in terms of specificity to TB manifestations, is vital in some translational analyses. Namely, when the animal models are not fully defined (i.e., there is no clear radiological taxonomy of the manifestations), the volume of the damaged lung can be used as a proxy.

Even for models with sufficient image quality for radiological characterization of the lesions, their identification by experts is tedious, time-consuming and prone to errors. This fact motivates the development of the methods presented in Chapters 3 and 4 to help the task automatization. Thus, Chapter 3 uses statistical descriptors that feed a machine learning (ML) model. These statistical descriptors result in lower interpretability than the previous approach but greater predictive power. The ML algorithms allow discerning between specific TB manifestations on the previously delimited whole lung of the macaque model. To avoid possible errors during delimitation, the model proposed in Chapter 4 extends the ability to detect lesions. To this aim, we train a high-capacity Deep Learning model capable of extracting descriptors automatically from complete image volumes to infer the presence or the number of manifestation types. This way eliminates the need to delineate the lungs in a previous step.

The methods in Chapters 2, 3 and 4 allow classifying tissues. They have the ability to capture relevant information from images beyond human capabilities, as proved for the mild-TB macaque model. Such approaches are easily extendible to different datasets/animal models, as shown in the recent literature, allowing the experts to obtain specific dataset outcomes. Subsequently, multidisciplinary experts can combine individual outputs in the drug design context to provide model translational explanations. This idea reflects the traditional workflow for disease understanding and drug design relying on computer-assisted image analysis. It also reveals the lack of automation tools implemented to operate with multidomain data, mimicking human intelligence.

In this context, given the ability of current computational models to find relevant relationships or subtle information, introducing artificial intelligence systems that can exploit inter-species data is fundamental to overcoming the limitations mentioned above. Therefore, in the previous chapter, a method that assumes a causal structure in data generation from different animal models is employed to promote the formulated translational analyses. The proposed DL model leverages shared features between images of distinct animal models. Apart from providing proper segmentation masks, the model infers similar levels of lung damage for comparable slices of the different animal models and alternatively generate such images when intervening in the model to fix a given level of damage.

## 6.2 Prospective work and Perspectives

Throughout this work, we have presented different methodologies based on Artificial Intelligence that enrich the extraction of relevant information from pathological images. Introducing the proposed methodologies in the drug assays pipelines is essential to facilitate diagnosis, disease longitudinal monitoring, and understanding disease etiology. From a mere technical point of view, what is missing in the current work for this inclusion is an extended validation.

Fortunately, the context of ERA4TB (see 1.2.2) enables different ways to accomplish it. Prospective trials based on a higher subject number will allow us to extract further conclusions about the methods generalization capabilities. More importantly, such studies will enable to use of the benefits of computer assistance for understanding Tuberculosis pathophysiology.

In addition, the project context allows validation through triangulation, which is fundamental for qualitative research [40]. In ERA4TB, not only CT images are used for longitudinal analysis of infected animals, but also vital data produced from other imaging modalities (see 1.3): Positron Emission Tomography (PET), pathological microscopy, Matrix-Assisted Laser Desorption Ionization (MALDI), etc., together with molecular and DNA analyses that greatly extend the description at the microscopic level of the bacteria interactions with the new drugs.

It is important to note that the method developed in Chapter 5 is not only extensible to new animal models (e.g., rat, rabbit, pig) but also to all the information sources mentioned above. Therefore, extending the current state of the art (SOTA) computational methods towards mimicking much more human intelligence. Note that SOTA DL models usually rely on domain-specific assumptions. For instance, inference depends only on CT images. However, when a human expert makes decisions employing information extracted from images is acting in an *imagined space*[1]. Namely, the expert consciously or implicitly considers all previous knowledge, meaning, disease model, subject demographics, comorbidity, treatment, etc. Therefore, our computational models need to resemble such mechanisms to assist during reasoning. The most meaningful works in recent literature already point in such a direction that will be a fundamental point in our future job [143, 148]. This prospective framework will enable the simulation of the longitudinal progress of the disease and determine the causal factors of treatments efficacy. Actually, for our future work, we have coined the term CaFE (Causal Factors of Efficacy) extended it to ICaFE (Imaging CaFE) when the factors are derived just from medical imaging data and TICaFE (Translational ICaFE) or TCaFE when the factors are translational.

Undoubtedly the range of possible technical work that can arise in this context is extensive, especially if tools integration in the workflow is effective. However, this process goes beyond technicalities. Like any other new automation technology, there is an existing reluctance in many strata of society that strongly constrain its integration.

On the one hand, this rejection is due to the operation in itself of the new technologies [311]. For example, it is unfortunately common to find cases in which Artificial Intelligence is employed without any control and let make biased decisions disfavoring certain social

[1] *Imagined space* concept was coined by Lorenz [195], and it is employed by Scholkopf et al. [275] for AI

groups [21, 50]. It is also of note the multiple instances in which the introduction of Artificial Intelligence can lead to the destruction of jobs without alternatives for workers [1]. Dealing with these problems necessarily involves making political decisions that contextualize the use of new tools within the normative ethical values that (*a priori*) formalize societies [30]. Although legislation development usually lets behind, measures to introduce a control are becoming more common [122].

However, the rejection within the research context attains to different matters. The medical imaging field copes (for too long already), its exciting oversized version of a *paradigm shift* [168]. It keeps an unnecessary struggle between the reluctance from the old "capos", the ones too embedded in the old fashioned clinical practices, and the necessity to boost the branch towards the Artificial Intelligence community direction [309, 351] supported by those with *"hands-on"* the actual predicament. It is horrific to witness how some self-interested negationists in the field take advantage of their political positions to badly exploit common resources that more than ever need to learn from the scientific community. Therefore, the long term work must go beyond technical aspects and penetrate ethics and education.

# References

[1] The future of work. *Nature*, 550(7676):315, 10 2017. ISSN 14764687. doi:10.1038/550315A.

[2] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 12 2021. ISSN 1566-2535. doi:10.1016/j.inffus.2021.05.008.

[3] Bariqi Abdillah, Alhadi Bustamam, and Devvi Sarwinda. Image processing based detection of lung cancer on CT scan images. *Journal of Physics: Conference Series*, 893(1):012063, 10 2017. ISSN 1742-6596. doi:10.1088/1742-6596/893/1/012063.

[4] Ibrahim Abubakar, Francis Drobniewski, Jo Southern, Alice J. Sitch, Charlotte Jackson, Marc Lipman, Jonathan J. Deeks, Chris Griffiths, Graham Bothamley, William Lynn, Helen Burgess, Bobby Mann, Ambreen Imran, Saranya Sridhar, Chuen Yan Tsou, Vladyslav Nikolayevskyy, Melanie Rees-Roberts, Hilary Whitworth, Onn Min Kon, Pranab Haldar, Heinke Kunst, Sarah Anderson, Andrew Hayward, John M. Watson, Heather Milburn, Ajit Lalvani, D. Adeboyeku, N. Bari, J. Barker, H. Booth, F. Chua, D. Creer, M. Darmalingam, R. N. Davidson, M. Dedicoat, A. Dunleavy, J. Figueroa, M. Haseldean, N. Johnson, S. Losewicz, J. Lord, J. Moore-Gillon, G. Packe, M. Pareek, S. Tiberi, A. Pozniak, and F. Sanderson. Prognostic value of interferon-$\gamma$ release assays and tuberculin skin test in predicting the development of active tuberculosis (UK PREDICT TB): a prospective cohort study. *The Lancet Infectious Diseases*, 18(10):1077–1087, 10 2018. ISSN 1473-3099. doi:10.1016/S1473-3099(18)30355-4.

[5] Rolf Adams and Leanne Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1994. ISSN 01628828. doi:10.1109/34.295913.

[6] Wafaa Alakwaa, Mohammad Nassef, and Amr Badr. Lung cancer detection and classification with 3D convolutional neural network (3D-CNN). *International Journal of Biology and Biomedical Engineering*, 11(8):66–73, 2017. ISSN 19984510. doi:10.14569/ijacsa.2017.080853.

[7] Janee Alam, Sabrina Alam, and Alamgir Hossan. Multi-Stage Lung Cancer Detection and Prediction Using Multi-class SVM Classifie. *International Conference on Computer, Communication, Chemical, Material and Electronic Engineering, IC4ME2 2018*, 9 2018. doi:10.1109/Ic4me2.2018.8465593.

[8] Amine Amyar, Romain Modzelewski, Hua Li, and Su Ruan. Multi-task deep learning based CT imaging analysis for COVID-19 pneumonia: Classification and segmentation. *Computers in Biology and Medicine*, 126:104037, 11 2020. ISSN 0010-4825. doi:10.1016/j.compbiomed.2020.104037.

[9] Diego Ardila, Atilla P. Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J. Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, David P. Naidich, and Shravya Shetty. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine*, 25(6):954–961, 6 2019. ISSN 1078-8956. doi:10.1038/s41591-019-0447-x.

[10] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. In *ArXiv*, 7 2020. URL http://arxiv.org/abs/2002.04692.

[11] Samuel G. Armato and William F. Sensakovic. Automated lung segmentation for thoracic CT: Impact on computer-aided diagnosis1. *Academic Radiology*, 11(9): 1011–1021, 9 2004. ISSN 1076-6332. doi:10.1016/j.acra.2004.06.005.

[12] Miguel A. Arroyo-Ornelas, Ma. Concepción Arenas-Arrocena, Horacio V. Estrada, Victor M. Castaño, and Luz M. López-Marín. Immune Diagnosis of Tuberculosis Through Novel Technologies. In *Understanding Tuberculosis - Global Experiences and Innovative Approaches to the Diagnosis*. IntechOpen, 2 2012. ISBN 978-953-307-938-7. doi:10.5772/31421.

[13] Xabier Artaechevarria, Daniel Pérez-Martín, Mario Ceresa, Gabriel De Biurrun, D Blanco, L M Montuenga, Bram Van Ginneken, Carlos Ortiz-De-Solorzano, and A M Muñoz-Barrutia. Airway segmentation and analysis for the study of mouse models of lung disease using micro-CT. *Phys. Med. Biol*, 54(November):7009–7024, 2009. ISSN 0031-9155. doi:10.1088/0031-9155/54/22/017.

[14] Xabier Artaechevarria, David Blanco, Daniel Pérez-Martín, Gabriel De Biurrun, Luis M. Montuenga, Juan P. De Torres, Javier J. Zulueta, Gorka Bastarrika, Arrate Muñoz-Barrutia, and Carlos Ortiz-De-Solorzano. Longitudinal study of a mouse model of chronic pulmonary inflammation using breath hold gated micro-CT. *European Radiology*, 20(11):2600–2608, 2010. ISSN 09387994. doi:10.1007/s00330-010-1853-0.

[15] Xabier Artaechevarria, Daniel Pérez-Martín, Joseph M Reinhardt, Arrate Muñoz-Barrutia, and Carlos Ortiz-De-Solórzano. Automated Quantitative Analysis of a Mouse Model of Chronic Pulmonary Inflammation using Micro X-ray Computed Tomography. In *Medical Image Computing and Computer Assisted Intervention Society (Pulmonary Imaging Workshop)*, 2010. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.148.5532.

[16] Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. 9 2019. URL https://arxiv.org/abs/1909.03012v2.

[17] Peter Auguste, Alexander Tsertsvadze, Joshua Pink, Rachel Court, Noel McCarthy, Paul Sutcliffe, and Aileen Clarke. Comparing interferon-gamma release assays with tuberculin skin test for identifying latent tuberculosis infection that progresses to active tuberculosis: systematic review and meta-analysis. *BMC Infectious Diseases 2017 17:1*, 17(1):1–13, 3 2017. ISSN 1471-2334. doi:10.1186/S12879-017-2301-4.

[18] S. Avinash, K. Manjunath, and S. Senthil Kumar. An improved image processing analysis for the detection of lung cancer using Gabor filters and watershed segmentation technique. In *Proceedings of the International Conference on Inventive Computation Technologies, ICICT 2016*, volume 2016. Institute of Electrical and Electronics Engineers Inc., 2016. doi:10.1109/inventive.2016.7830084.

[19] Clifton E Barry 3rd, Helena Boshoff, Véronique Dartois, Thomas Dick, Sabine Ehrt, Joanne Flynn, Dirk Schnappinger, Robert J Wilkinson, and Douglas Young. The spectrum of latent tuberculosis: rethinking the goals of prophylaxis. *Nature reviews. Microbiology*, 7(12):845–855, 2009. doi:10.1038/nrmicro2236.

[20] J. Baxter. A Model of Inductive Bias Learning. *Journal of Artificial Intelligence Research*, 12:149, 6 2000. doi:10.1613/jair.731.

[21] Emily M Bender, Timnit Gebru, Angelina Mcmillan-Major, Shmargaret Shmitchell, and Shmar-Garet Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? CCS CONCEPTS • Computing methodologies → Natural language processing. ACM Reference Format. In *Association for Computing Machinery*, pages 610–623, New York, NY, USA, 3 2021. ISBN 9781450383097. doi:10.1145/3442188.3445922.

[22] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi:10.1109/tpami.2013.50.

[23] M Bevk and I Kononenko. A Statistical Approach to Texture Description of Medical Images: A Preliminary Study. In *Proceedings of 15th IEEE Symposium on Computer-Based Medical Systems (CBMS 2002)*, pages 239–244, 2002. doi:10.1109/cbms.2002.1011383.

[24] Ashu Seith Bhalla, Ankur Goyal, Randeep Guleria, and Arun Kumar Gupta. Chest tuberculosis: Radiological review and imaging recommendations. *The Indian Journal of Radiology & Imaging*, 25(3):213, 8 2015. doi:10.4103/0971-3026.161431.

[25] Abhir Bhandary, G. Ananth Prabhu, V. Rajinikanth, K. Palani Thanaraj, Suresh Chandra Satapathy, David E. Robbins, Charles Shasky, Yu Dong Zhang, João Manuel R.S. Tavares, and N. Sri Madhava Raja. Deep-learning framework to detect lung abnormality – A study with chest X-Ray and lung CT scan images. *Pattern Recognition Letters*, 129:271–278, 1 2020. ISSN 0167-8655. doi:10.1016/j.patrec.2019.11.013.

[26] Siddharth Bhatia, Yash Sinha, and Lavika Goel. Lung Cancer Detection: A Deep Learning Approach. *Advances in Intelligent Systems and Computing*, 817:699–705, 2019. doi:10.1007/978-981-13-1595-4_55.

[27] Christopher M. Bishop. Model-based machine learning. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, 371(2):1–17, 2013. doi:10.1098/rsta.2012.0222.

[28] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006. ISBN 9780387310732. doi:10.1017/CBO9781107415324.004.

[29] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518): 859–877, 2017. ISSN 1537274X. doi:10.1080/01621459.2017.1285773.

[30] Jason Borenstein and Ayanna Howard. Emerging challenges in AI and the need for AI ethics education. *AI and Ethics 2020 1:1*, 1(1):61–65, 10 2020. ISSN 2730-5961. doi:10.1007/S43681-020-00002-7.

[31] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM. ISBN 0-89791-497-X. doi:10.1145/130385.130401.

[32] Felix J S Bragman, Ryutaro Tanno, Zach Eaton-Rosen, Wenqi Li, David J Hawkes, Sebastien Ourselin, Daniel C Alexander, Jamie R Mcclelland, and M Jorge Cardoso. Uncertainty in multitask learning: joint representations for probabilistic MR-only radiotherapy planning. *ArXiv*, 2018. URL https://arxiv.org/pdf/1806.06595.pdf.

[33] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 08856125. doi:10.1023/A:1010933404324.

[34] Leo Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3): 199–215, 2001. ISSN 08834237. doi:10.1214/ss/1009213726.

[35] Ernesto Bribiesca. An easy measure of compactness for 2D and 3D shapes. *Pattern Recognition*, 41(2):543–554, 2008. doi:10.1016/j.patcog.2007.06.029.

[36] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. 4 2021. URL http://arxiv.org/abs/2104.13478.

[37] Thomas Bülow, Cristian Lorenz, and Steffen Renisch. A General Framework for Tree Segmentation and Reconstruction from Medical Volume Data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 533–540, 2004. doi:10.1007/978-3-540-30135-6_65.

[38] Joshua Burrill, Christopher J. Williams, Gillian Bain, Gabriel Conder, Andrew L. Hine, and Rakesh R. Misra. Tuberculosis: A Radiologic Review. *RadioGraphics*, 27 (5):1255–1273, 2007. ISSN 0271-5333. doi:10.1148/rg.275065176.

[39] Yukun Cao, Zhanwei Xu, Jianjiang Feng, Cheng Jin, Xiaoyu Han, Hanping Wu, and Heshui Shi. Longitudinal assessment of covid-19 using a deep learning–based quantitative ct pipeline: Illustration of two cases, 3 2020. ISSN 26386135.

[40] Nancy Carter, Denise Bryant-Lukosius, Alba Dicenso, Jennifer Blythe, and Alan J. Neville. The use of triangulation in qualitative research. *Oncology nursing forum*, 41 (5):545–547, 9 2014. ISSN 1538-0688. doi:10.1188/14.ONF.545-547.

[41] Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. Geodesic Active Contours. *International Journal of Computer Vision*, 22(1):61–79, 1997. doi:10.1023/A:1007979827043.

[42] Daniel C. Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 12 2020. ISSN 20411723. doi:10.1038/s41467-020-17478-w.

[43] M Ceresa, X Artaechevarria, A Munoz-Barrutia, and C Ortiz-De-Solorzano. Automatic Leakage Detection and Recovery For Airway Tree Extraction in Chest CT Images. In *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 568–571. IEEE, 2010. doi:10.1109/ISBI.2010.5490282.

[44] Chia-Hung Chen, Chih-Kun Chang, Chih-Yen Tu, Wei-Chih Liao, Bing-Ru Wu, Kuei-Ting Chou, Yu-Rou Chiou, Shih-Neng Yang, Geoffrey Zhang, and Tzung-Chi Huang. Radiomic features analysis in computed tomography images of lung nodule classification. *PLOS ONE*, 13(2):e0192002, 2 2018. ISSN 1932-6203. doi:10.1371/journal.pone.0192002.

[45] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. 2021. URL http://arxiv.org/abs/2102.04306.

[46] Jun Chen, Lianlian Wu, Jun Zhang, Liang Zhang, Dexin Gong, Yilin Zhao, Qiuxiang Chen, Shulan Huang, Ming Yang, Xiao Yang, Shan Hu, Yonggui Wang, Xiao Hu, Biqing Zheng, Kuo Zhang, Huiling Wu, Zehua Dong, Youming Xu, Yijie Zhu, Xi Chen, Mengjiao Zhang, Lilei Yu, Fan Cheng, and Honggang Yu. Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography. *Scientific Reports 2020 10:1*, 10(1):1–11, 11 2020. ISSN 2045-2322. doi:10.1038/s41598-020-76282-0.

[47] L. Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical Image Analysis*, 58:101539, 12 2019. ISSN 1361-8415. doi:10.1016/j.media.2019.101539.

[48] Ray Y Chen, Lori E Dodd, Myungsun Lee, Praveen Paripati, Dima A Hammoud, James M Mountz, Doosoo Jeon, Nadeem Zia, Homeira Zahiri, M Teresa Coleman, Matthew W Carroll, Jong Doo Lee, Yeon Joo Jeong, Peter Herscovitch, Saher Lahouar, Michael Tartakovsky, Alexander Rosenthal, Sandeep Somaiyya, Soyoung Lee, Lisa C Goldfeder, Ying Cai, Laura E Via, Seung-Kyu Park, Sang-Nae Cho, and Clifton E Barry 3rd. PET/CT imaging correlates with treatment outcome in patients with multidrug-resistant tuberculosis. *Science translational medicine*, 6(265):166–265, 2014. doi:10.1126/scitranslmed.3009501.

[49] Rewon Child. Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images. 11 2020. URL https://arxiv.org/abs/2011.10650v2.

[50] Sahil Chinoy. The Racist History Behind Facial Recognition, 9 2019. ISSN 17439019. URL https://www.nytimes.com/2019/07/10/opinion/facial-recognition-race.html.

[51] Wookjin Choi, Jung Hun Oh, Sadegh Riyahi, Chia-Ju Liu, Feng Jiang, Wengen Chen, Charles White, Andreas Rimner, James G. Mechalakos, Joseph O. Deasy, and Wei Lu. Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer. *Medical Physics*, 45(4):1537–1549, 4 2018. ISSN 2473-4209. doi:10.1002/mp.12820.

[52] Stergios Christodoulidis, Marios Anthimopoulos, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. Multisource Transfer Learning with Convolutional Neural Networks for Lung Pattern Analysis. *IEEE Journal of Biomedical and Health Informatics*, 21(1):76–84, 1 2017. doi:10.1109/JBHI.2016.2636929.

[53] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. 6 2016. URL http://arxiv.org/abs/1606.06650.

[54] William J. Clancey. The epistemology of a rule-based expert system -a framework for explanation. *Artificial Intelligence*, 20(3):215–251, 5 1983. ISSN 00043702. doi:10.1016/0004-3702(83)90008-5.

[55] William S Cleveland and Susan J Devlin. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988. ISSN 01621459. doi:10.2307/2289282.

[56] Joseph Paul Cohen, Mohammad Hashir, Rupert Brooks, and Hadrien Bertrand. On the limits of cross-domain generalization in automated X-ray prediction, 9 2020. ISSN 2640-3498. URL http://proceedings.mlr.press/v121/cohen20a.html.

[57] Joseph Paul Cohen, Paul Morrison, and Lan Dao. COVID-19 Image Data Collection. 3 2020. ISSN 2331-8422. URL http://arxiv.org/abs/2003.11597.

[58] Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P. Lungren, and Akshay Chaudhari. Gifsplanation via Latent Shift: A Simple Autoencoder Approach to Counterfactual Generation for Chest X-rays. *Proceedings of Machine Learning Research*, 2021. URL http://arxiv.org/abs/2102.09475.

[59] Thibaud P. Coroller, Patrick Grossmann, Ying Hou, Emmanuel Rios Velazquez, Ralph T.H. Leijenaar, Gretchen Hermann, Philippe Lambin, Benjamin Haibe-Kains, Raymond H. Mak, and Hugo J.W.L. Aerts. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiotherapy and Oncology*, 114(3): 345–350, 3 2015. ISSN 0167-8140. doi:10.1016/j.radonc.2015.02.015.

[60] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989. ISSN 09324194. doi:10.1007/BF02551274.

[61] Sumona Datta, Lena Shah, Robert H. Gilman, and Carlton A. Evans. Comparison of sputum collection methods for tuberculosis diagnosis: a systematic review and pairwise and network meta-analysis. *The Lancet Global Health*, 5(8):e760–e771, 8 2017. ISSN 2214-109X. doi:10.1016/s2214-109x(17)30201-2.

[62] Alex J. DeGrave, Joseph D Janizek, and Su In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7): 2020.09.13.20193565, 10 2021. doi:10.1038/s42256-021-00338-7.

[63] Mike J Dennis, Simon Parks, Gordon Bell, Irene Taylor, Jody Lakeman, and Sally A Sharpe. A Flexible Approach to Imaging in ABSL-3 Laboratories. *Applied Biosafety*, 20(2), 2015. ISSN 24701246. doi:10.1177/153567601502000204.

[64] Adrien Depeursinge, Antonio Foncubierta-Rodriguez, Dimitri Van De Ville, and Henning Müller. Three-dimensional Solid Texture Analysis in Biomedical Imaging: Review and Opportunities. *Medical Image Analysis*, 18(1):176–196, 2014. ISSN 13618415. doi:10.1016/j.media.2013.10.005.

[65] Thomas Deschamps and Laurent D. Cohen. Fast extraction of minimal paths in 3D images and applications to virtual endoscopy. *Medical Image Analysis*, 5(4):281–299, 2001. doi:10.1016/S1361-8415(01)00046-9.

[66] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186, 10 2018. URL https://arxiv.org/abs/1810.04805v2.

[67] Andreas H. Diacon, Alexander Pym, Martin Grobusch, Ramonde Patientia, Roxana Rustomjee, Liesl Page-Shipp, Christoffel Pistorius, Rene Krause, Mampedi Bogoshi, Gavin Churchyard, Amour Venter, Jenny Allen, Juan Carlos Palomino, Tine De Marez, Rolf P.G. van Heeswijk, Nacer Lounis, Paul Meyvisch, Johan Verbeeck, Wim Parys, Karel de Beule, Koen Andries, and David F. Mc Neeley. The Diarylquinoline TMC207 for Multidrug-Resistant Tuberculosis. *http://dx.doi.org/10.1056/NEJMoa0808427*, 360(23):2397–2405, 12 2009. doi:10.1056/nejmoa0808427.

[68] Yashin Dicente Cid, Vitali Liauchuk, Dzmitri Klimuk, Aleh Tarasau, Vassili Kovalev, and Henning Müller. Overview of ImageCLEFtuberculosis 2019 - Automatic CT-based Report Generation and Tuberculosis Severity Assessment. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. CEUR-WS.org, 2019. URL http://ceur-ws.org/Vol-2380/paper_138.pdf.

[69] Tan N. Doan, Damon P. Eisen, Morgan T. Rose, Andrew Slack, Grace Stearnes, and Emma S. McBryde. Interferon-gamma release assay for the diagnosis of latent tuberculosis infection: A latent-class analysis. *PLOS ONE*, 12(11):e0188631, 11 2017. ISSN 1932-6203. doi:10.1371/journal.pone.0188631.

[70] Tom Doel, David J Gavaghan, and Vicente Grau. Review of automatic pulmonary lobe segmentation methods from CT. *Computerized Medical Imaging and Graphics*, 2015. doi:10.1016/j.compmedimag.2014.10.008.

[71] Reuben Dorent, Thomas Booth, Wenqi Li, Carole H. Sudre, Sina Kafiabadi, Jorge Cardoso, Sebastien Ourselin, and Tom Vercauteren. Learning joint segmentation of tissues and brain lesions from task-specific hetero-modal domain-shifted datasets. *Medical Image Analysis*, 67(December):1–5, 2021. ISSN 13618423. doi:10.1016/j.media.2020.101862.

[72] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 10 2020. URL http://arxiv.org/abs/2010.11929.

[73] Richard Duda and Peter Hart. Use of the Hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1 1972. doi:10.1145/361237.361242.

[74] Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. 3 2018. URL https://arxiv.org/abs/1603.07285v2http://arxiv.org/abs/1603.07285.

[75] David A van Dyk and Xiao-Li Meng. The Art of Data Augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001. doi:10.1198/10618600152418584.

[76] ERA4TB consotium. ERA4TB, 2021. URL https://era4tb.org/the-project/.

[77] Joel D. Ernst. The immunological life cycle of tuberculosis. *Nature Reviews Immunology 2012 12:8*, 12(8):581–591, 7 2012. ISSN 1474-1741. doi:10.1038/nri3259.

[78] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2 2017. ISSN 0028-0836. doi:10.1038/nature21056.

[79] Amal A Farag, Hossam E Abd, El Munim, James H Graham, and Aly A Farag. A Novel Approach for Lung Nodules Segmentation in Chest CT Using Level Sets. *IEEE Transactions on Image Processing*, 22(12):5202–5213, 2013. doi:10.1109/tip.2013.2282899.

[80] José Raniery Ferreira-Junior, Marcel Koenigkam-Santos, Ariane Priscilla Magalhães Tenório, Matheus Calil Faleiros, Federico Enrique Garcia Cipriano, Alexandre Todorovic Fabro, Janne Näppi, Hiroyuki Yoshida, and Paulo Mazzoncini de Azevedo-Marques. CT-based radiomics for prediction of histologic subtype and metastatic disease in primary malignant lung neoplasms. *International Journal of Computer Assisted Radiology and Surgery 2019 15:1*, 15(1):163–172, 11 2019. ISSN 1861-6429. doi:10.1007/S11548-019-02093-Y.

[81] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 9 1936. ISSN 20501420. doi:10.1111/j.1469-1809.1936.tb02137.x.

[82] Nicolas Forcadel, Carole Le Guyader, and Christian Gout. Generalized fast marching method: applications to image segmentation. *Numerical Algorithms*, 48(1-3):189–211, 2008. doi:10.1007/s11075-008-9183-x.

[83] Yarin Gal. Uncertainty in Deep Learning. *PhD Thesis*, (October), 2016. URL https://mlg.eng.cam.ac.uk/yarin/thesis/thesis.pdf.

[84] Craig J Galbán, Meilan K Han, Jennifer L Boes, Komal A Chughtai, Charles R Meyer, Timothy D Johnson, Stefanie Galbán, Alnawaz Rehemtulla, Ella A Kazerooni, Fernando J Martínez, and Brian D Ross. Computed tomography-based biomarker provides unique signature for diagnosis of COPD phenotypes and disease progression. *Nature medicine*, 18(11):1711–1715, 2012. doi:10.1038/nm.2971.

[85] Mary M. Galloway. Texture analysis using gray level run lengths. *Computer Graphics and Image Processing*, 4(2):172–179, 1975. ISSN 0146664X. doi:10.1016/s0146-664x(75)80008-6.

[86] Neel R. Gandhi, Paul Nunn, Keertan Dheda, H. Simon Schaaf, Matteo Zignol, Dick van Soolingen, Paul Jensen, and Jaime Bayona. Multidrug-resistant and extensively drug-resistant tuberculosis: a threat to global control of tuberculosis. *The Lancet*, 375 (9728):1830–1843, 5 2010. ISSN 0140-6736. doi:10.1016/S0140-6736(10)60410-2.

[87] Mingchen Gao, Ulas Bagci, Le Lu, Aaron Wu, Mario Buty, Hoo Chang Shin, Holger Roth, Georgios Z. Papadakis, Adrien Depeursinge, Ronald M. Summers, Ziyue Xu, and Daniel J. Mollura. Holistic classification of CT attenuation patterns for interstitial lung diseases via deep convolutional neural networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging and Visualization*, 6(1):1–6, 1 2018. ISSN 21681171. doi:10.1080/21681163.2015.1124249.

[88] Francesco Di Gennaro, Luigi Pisani, Nicola Veronese, Damiano Pizzol, Valeria Lippolis, Annalisa Saracino, Laura Monno, Michaëla A.M. Huson, Roberto Copetti, Giovanni Putoto, and Marcus J. Schultz. Potential Diagnostic Properties of Chest Ultrasound in Thoracic Tuberculosis—A Systematic Review. *International Journal of Environmental Research and Public Health 2018, Vol. 15, Page 2235*, 15(10):2235, 10 2018. doi:10.3390/ijerph15102235.

[89] Sarah E. Gerard, Jacob Herrmann, David W. Kaczka, Guido Musch, Ana Fernandez-Bustamante, and Joseph M. Reinhardt. Multi-resolution convolutional neural networks for fully automated segmentation of acutely injured lungs in multiple species. *Medical Image Analysis*, 60, 2 2020. ISSN 13618423. doi:10.1016/j.media.2019.101592.

[90] Zoubin Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015. ISSN 0028-0836. doi:10.1038/nature14541.

[91] Robert J. Gillies, Paul E. Kinahan, and Hedvig Hricak. Radiomics: Images Are More than Pictures, They Are Data. *Radiology*, 278(2):563–577, 2016. ISSN 0033-8419. doi:10.1148/radiol.2015151169.

[92] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets, 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.

[93] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Regularization for Deep Learning. In *Deep Learning*, chapter 7. 2016. URL www.deeplearningbook.org.

[94] P. M. Gordaliza, A. Muñoz-Barrutia, L. E. Via, S. Sharpe, M. Desco, and J. J. Vaquero. Computed Tomography-Based Biomarker for Longitudinal Assessment of Disease Burden in Pulmonary Tuberculosis. *Molecular Imaging and Biology*, 21(1):19–24, 2

2019. ISSN 1536-1632. doi:10.1007/s11307-018-1215-x. URL http://link.springer.com/10.1007/s11307-018-1215-x.

[95] Pedro M. Gordaliza, Arrate Muñoz-Barrutia, Mónica Abella, Manuel Desco, Sally Sharpe, and Juan José Vaquero. Unsupervised CT Lung Image Segmentation of a Mycobacterium Tuberculosis Infection Model. *Scientific Reports*, 8(1), 12 2018. ISSN 2045-2322. doi:10.1038/s41598-018-28100-x.

[96] Pedro M. Gordaliza, J J Vaquero, S Sharpe, and M Desco Menéndez. Radiomics for the Discrimination of Tuberculosis Lesions. In *European Molecular Imaging Meeting*, 2018. ISBN 9783540874812. URL http://eventclass.org/contxt_emim2018/online-program/session?s=PS-12.

[97] Pedro M. Gordaliza, Juan José Vaquero, Sally Sharpe, Manuel Desco, and Arrate Munoz-Barrutia. Towards an informational model for tuberculosis lesion discrimination on X-ray CT images. In *International Symposium on Biomedical Imaging*, 2018. ISBN 9781538636367. doi:10.1109/ISBI.2018.8363570.

[98] Pedro M. Gordaliza, Juan José Vaquero, Sally Sharpe, Fergus Gleeson, and Arrate Muñoz-Barrutia. A Multi-Task Self-Normalizing 3D-CNN to Infer Tuberculosis Radiological Manifestations. In *Medical Imaging with Deep Learning (MIDL)*, pages 1–5, 7 2019. URL http://arxiv.org/abs/1907.12331.

[99] Pedro M. Gordaliza, Juan José Vaquero, and Arrate Munoz-Barrutia. Translational Lung Imaging Analysis Through Disentangled Representations. In *Proceedings of Machine Learning Research-Under Review*, pages 1–13, 12 2021. URL https://openreview.net/forum?id=16efiNAl_VA.

[100] Anirudh Goyal and Yoshua Bengio. Inductive Biases for Deep Learning of Higher-Level Cognition. 11 2021. URL http://arxiv.org/abs/2011.15091.

[101] Ophir Gozes, Maayan Frid-Adar, Nimrod Sagie, Huangqi Zhang, Wenbin Ji, and Hayit Greenspan. Coronavirus Detection and Analysis on Chest CT with Deep Learning. 4 2020. URL http://arxiv.org/abs/2004.02640.

[102] Leo Grady. Random Walks for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1768–1783, 2006. doi:10.1109/tpami.2006.233.

[103] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang Zhong Yang. XAI-Explainable artificial intelligence. *Science Robotics*, 4(37), 12 2019. doi:10.1126/scirobotics.aay7120.

[104] Sebastian M. Gygli, Chloé Loiseau, Levan Jugheli, Natia Adamia, Andrej Trauner, Miriam Reinhard, Amanda Ross, Sonia Borrell, Rusudan Aspindzelashvili, Nino Maghradze, Klaus Reither, Christian Beisel, Nestani Tukvadze, Zaza Avaliani, and Sebastien Gagneux. Prisons as ecological drivers of fitness-compensated multidrug-resistant Mycobacterium tuberculosis. *Nature Medicine 2021 27:7*, 27(7):1171–1177, 5 2021. ISSN 1546-170X. doi:10.1038/s41591-021-01358-x.

[105] Robert M. Haralick. Statistical and Structural Approaches to Texture. *Proceedings of the IEEE*, 67(5):786–804, 1979. doi:10.1109/proc.1979.11328.

[106] Robert M. Haralick, K. Shanmugan, and I. H. Dinstein. Textural features for image classification. *Systems, Man and Cybernetics*, 6:610–621, 1973. doi:10.1109/tsmc.1973.4309314.

[107] G J Harper and J D Morton. The respiratory retention of bacterial aerosols: experiments with radioactive spores. *The Journal of hygiene*, 51(3):372–85, 1953. ISSN 0022-1724. doi:10.1017/S0022172400015801.

[108] Adam P. Harrison, Ziyue Xu, Kevin George, Le Lu, Ronald M. Summers, and Daniel J. Mollura. Progressive and multi-path holistically nested neural networks for pathological lung segmentation from CT images. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10435 LNCS, pages 621–629. Springer, Cham, 9 2017. ISBN 9783319661780. doi:10.1007/978-3-319-66179-7_71.

[109] Akane Haruna, Shigeo Muro, Yasutaka Nakano, Tadashi Ohara, Yuma Hoshino, Emiko Ogawa, Toyohiro Hirai, Akio Niimi, Koichi Nishimura, Kazuo Chin, and Michiaki Mishima. CT scan findings of emphysema predict mortality in COPD. *Chest*, 138(3):635–640, 9 2010. ISSN 19313543. doi:10.1378/chest.09-2836.

[110] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009. URL https://link.springer.com/book/10.1007/978-0-387-84858-7.

[111] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. UNETR: Transformers for 3D Medical Image Segmentation. 2021. URL http://arxiv.org/abs/2103.10504.

[112] Samuel Hawkins, Hua Wang, Ying Liu, Alberto Garcia, Olya Stringfield, Henry Krewer, Qian Li, Dmitry Cherezov, Robert A. Gatenby, Yoganand Balagurunathan, Dmitry Goldgof, Matthew B. Schabath, Lawrence Hall, and Robert J. Gillies. Predicting Malignant Nodules from Screening CT Scans. *Journal of Thoracic Oncology*, 11 (12):2120–2128, 12 2016. ISSN 1556-0864. doi:10.1016/j.jtho.2016.07.002.

[113] Haibo He and Edwardo A. Garcia. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. ISSN 10414347. doi:10.1109/tkde.2008.239.

[114] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:770–778, 12 2015. ISSN 10636919. doi:10.1109/CVPR.2016.90.

[115] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034. 2015 International Conference on Computer Vision, ICCV 2015, 2 2015. ISBN 9781467383912. doi:10.1109/ICCV.2015.123.

[116] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October: 2980–2988, 12 2017. ISSN 15505499. doi:10.1109/ICCV.2017.322.

[117] Lan He, Yanqi Huang, Zelan Ma, Cuishan Liang, Changhong Liang, and Zaiyi Liu. Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule. *Scientific Reports 2016 6:1*, 6(1):1–10, 10 2016. ISSN 2045-2322. doi:10.1038/srep34921.

[118] James J. Heckman. Sample selection bias as a specification error. *Applied Econometrics*, 31(3):129–137, 2013. doi:10.2307/1912352.

[119] Claudia I Henschke, David F Yankelevitz, Daniel M Libby, Mark W Pasmantier, and James P Smith. Survival of Patients with Stage I Lung Cancer Detected on CT Screening. *New England Journal of Medicine*, 355(17):1763–1771, 10 2006. ISSN 0028-4793. doi:10.1056/nejmoa060476.

[120] Miguel A Hernán and James M Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020. URL https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

[121] Charlotte C. Heuvelings, Sabine Bélard, Savvas Andronikou, Henrique Lederman, Halvani Moodley, Martin P. Grobusch, and Heather J. Zar. Chest ultrasound compared to chest X-ray for pediatric pulmonary tuberculosis. *Pediatric Pulmonology*, 54(12): 1914–1920, 12 2019. ISSN 1099-0496. doi:10.1002/ppul.24500.

[122] High-Level Expert Group on Artificial Intelligence and European Commission. Ethics Guidelines For Trustworthy AI. Technical report, Brussels, 2019. URL https://ec.europa.eu/digital.

[123] Geoffrey Hinton. Deep Learning—A Technology With the Potential to Transform Health Care. *JAMA*, 320(11):1101, 9 2018. ISSN 0098-7484. doi:10.1001/jama.2018.11100.

[124] Geoffrey Hinton. How to represent part-whole hierarchies in a neural network. 2 2021. URL http://arxiv.org/abs/2102.12627.

[125] Gladys L. Hobby, Audrey P. Holman, Michael D. Iseman, and Julia M. Jones. Enumeration of Tubercle Bacilli in Sputum of Patients with Pulmonary Tuberculosis. *Antimicrobial Agents and Chemotherapy*, 4(2):94, 1973. doi:10.1128/AAC.4.2.94.

[126] Johannes Hofmanninger, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European Radiology Experimental*, 4(1), 12 2020. ISSN 25099280. doi:10.1186/s41747-020-00173-2.

[127] S. A. Hojjatoleslami and J. Kittler. Region growing: A new approach. *IEEE Transactions on Image Processing*, 7(7):1079 – 1084, 1998. ISSN 10577149. doi:10.1109/83.701170.

[128] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312, 7 2019. ISSN 1942-4795. doi:10.1002/WIDM.1312.

[129] Rein M. G. J. Houben and Peter J. Dodd. The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. *PLOS Medicine*, 13(10): e1002152, 10 2016. ISSN 1549-1676. doi:10.1371/JOURNAL.PMED.1002152.

[130] P. V. C. Hough. Machine analysis of bubble chamber pictures. *2nd International Conference on High-Energy Accelerators and Instrumentation*, 73:554–558, 1959. URL https://inspirehep.net/literature/919922.

[131] Shiying Hu, Eric A Hoffman, and Joseph M Reinhardt. Automatic Lung Segmentation for Accurate Quantitation of Volumetric X-Ray CT Images. *IEEE Transactions on Medical Imaging*, 20(6):490–498, 2001. doi:10.1109/42.929615.

[132] Lu Huang, Rui Han, Tao Ai, Pengxin Yu, Han Kang, Qian Tao, and Liming Xia. Serial Quantitative Chest CT Assessment of COVID-19: A Deep Approach. *Radiology: Cardiothoracic Imaging*, 2(2), 3 2020. doi:10.1148/RYCT.2020200075.

[133] Yanqi Huang, Zaiyi Liu, Lan He, Xin Chen, Dan Pan, Zelan Ma, Cuishan Liang, Jie Tian, and Changhong Liang. Radiomics Signature: A Potential Biomarker for the Prediction of Disease-Free Survival in Early-Stage (I or II) Non—Small Cell Lung Cancer. *https://doi.org/10.1148/radiol.2016152234*, 281(3):947–957, 6 2016. doi:10.1148/radiol.2016152234.

[134] D. H. Hubel and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, 160(1):106–154, 1 1962. ISSN 14697793. doi:10.1113/jphysiol.1962.sp006837.

[135] Robin E Huebner, Maybelle F Schein, and Jr. John B. Bass. The Tuberculin Skin Test. *Clinical Infectious Diseases*, 17(6):968–975, 1993. ISSN 10584838. URL http://www.jstor.org/stable/4457498.

[136] Zeshan Hussain, Francisco Gimenez, Darvin Yi, and Daniel Rubin. Differential Data Augmentation Techniques for Medical Imaging Classification Tasks. *AMIA, Annual Symposium proceedings. AMIA Symposium*, 2017:979–984, 2017. ISSN 1942597X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977656/.

[137] Ferenc Huszár. Is Maximum Likelihood Useful for Representation Learning?, 5 2017. URL https://www.inference.vc/maximum-likelihood-for-representation-learning-2/.

[138] Alexey Ignatiev. Towards trustable explainable AI. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2021-Janua, pages 5154–5158, 2020. ISBN 9780999241165. doi:10.24963/ijcai.2020/726.

[139] John P A Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):0696–0701, 2005. ISSN 15491277. doi:10.1371/journal.pmed.0020124.

[140] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2 2015. URL http://arxiv.org/abs/1502.03167.

[141] Fabian Isensee, Paul F. Jaeger, Simon A.A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2 2021. ISSN 15487105. doi:10.1038/s41592-020-01008-z.

[142] Colin Jacobs and Bram van Ginneken. Google's lung cancer AI: a promising tool that needs further validation. *Nature Reviews Clinical Oncology*, pages 7–8, 2019. ISSN 17594782. doi:10.1038/s41571-019-0248-7.

[143] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General Perception with Iterative Attention. 2021. URL http://arxiv.org/abs/2103.03206.

[144] Marcin Janaszewski, Michel Couprie, and Laurent Babout. Hole filling in 3D volumetric objects. *Pattern Recognition*, 43(10):3548–3559, 2010. doi:10.1016/j.patcog.2010.04.015.

[145] Dakai Jin, Ziyue Xu, Youbao Tang, Adam P Harrison, and Daniel J Mollura. CT-Realistic Lung Nodule Simulation from 3D Conditional Generative Adversarial Networks for Robust Lung Segmentation. *ArXiV*, 2018. URL https://arxiv.org/pdf/1806.04051.pdf.

[146] Hans J Johnson, Matthew M Mccormick, and Luis Ibañez. *The ITK Software Guide*. Kitware Inc., fourth edi edition, 2015. ISBN 9781-930934-28-3. doi:10.3389/fninf.2014.00013.

[147] Daniel E. Jonas, Daniel S. Reuland, Shivani M. Reddy, Max Nagle, Stephen D. Clark, Rachel Palmieri Weber, Chineme Enyioha, Teri L. Malo, Alison T. Brenner, Charli Armstrong, Manny Coker-Schwimmer, Jennifer Cook Middleton, Christiane Voisin, and Russell P. Harris. Screening for Lung Cancer with Low-Dose Computed Tomography: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *JAMA - Journal of the American Medical Association*, 325(10): 971–987, 3 2021. ISSN 15383598. doi:10.1001/jama.2021.0377.

[148] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A.A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873*, 596(7873):583–589, 7 2021. ISSN 1476-4687. doi:10.1038/s41586-021-03819-2.

[149] Kaggle Inc. Data Science Bowl 2017 | Kaggle, 2017. URL https://www.kaggle.com/c/data-science-bowl-2017.

[150] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi,

Daniel Rueckert, and Ben Glocker. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10265 LNCS, pages 597–609. Springer, Cham, 2017. ISBN 9783319590493. doi:10.1007/978-3-319-59050-9_47.

[151] Yan Kang, Klaus Engelke, and Willi A. Kalender. A new accurate and precise 3-D segmentation method for skeletal structures in volumetric CT data. *IEEE Transactions on Medical Imaging*, 22(5):586 – 598, 2003. ISSN 02780062. doi:10.1109/TMI.2003.812265.

[152] Davood Karimi, Simon K. Warfield, and Ali Gholipour. Critical Assessment of Transfer Learning for Medical Image Segmentation with Fully Convolutional Neural Networks. 5 2020. URL http://arxiv.org/abs/2006.00356.

[153] Stefan HE Kaufmann. Fact and fiction in tuberculosis vaccine research: 10 years later. *The Lancet Infectious Diseases*, 11(8):633–640, 8 2011. ISSN 1473-3099. doi:10.1016/S1473-3099(11)70146-3.

[154] D Kaushal, S Mehra, P J Didier, and A A Lackner. The non-human primate model of tuberculosis. *Journal of Medical Primatology*, 41(3):191–201, 2012. ISSN 1600-0684. doi:10.1111/j.1600-0684.2012.00536.x.

[155] Baris Kayalibay, Grady Jensen, and Patrick van der Smagt. CNN-based Segmentation of Medical Imaging Data. 1 2017. URL https://arxiv.org/abs/1701.03056v2.

[156] Salome Kazeminia, Christoph Baur, Arjan Kuijper, Bram van Ginneken, Nassir Navab, Shadi Albarqouni, and Anirban Mukhopadhyay. GANs for medical image analysis. *Artificial Intelligence in Medicine*, 109:101938, 9 2020. ISSN 0933-3657. doi:10.1016/J.ARTMED.2020.101938.

[157] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In Guyon I, V Luxburg, S Garnett, Y Bengio, H. Wallach, R. Fergus, and S. Vishwanathan, editors, *Advances in Neural Information Processing Systems (NIPS 2017)*, pages 1–11. Curran Associates, Inc.w, 2017. URL http://papers.nips.cc/paper/7141-what-uncertainties-do-we-need-in-bayesian-deep-learning-for-computer-vision.pdf.

[158] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10, 2018. URL https://arxiv.org/pdf/1705.07115.pdf.

[159] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. 12 2014. URL http://arxiv.org/abs/1412.6980.

[160] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. (Ml):1–14, 2013. ISSN 1312.6114v10. doi:10.1051/0004-6361/201527329.

[161] Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling. Semi-Supervised Learning with Deep Generative Models. *Advances in Neural Information Processing Systems*, 4(January):3581–3589, 6 2014. URL https://arxiv.org/abs/1406.5298v2.

[162] Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving Variational Inference with Inverse Autoregressive Flow. (Nips), 2016. ISSN 10495258. URL http://arxiv.org/abs/1606.04934.

[163] Justin S. Kirby, Samuel G. Armato, Karen Drukker, Feng Li, Lubomir Hadjiiski, Georgia D. Tourassi, Laurence P. Clarke, Roger M. Engelmann, Maryellen L. Giger, George Redmond, and Keyvan Farahani. LUNGx Challenge for computerized lung nodule classification. *Journal of Medical Imaging*, 3(4):044506, 12 2016. ISSN 2329-4302. doi:10.1117/1.JMI.3.4.044506.

[164] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-Normalizing Neural Networks. In *Advances in Neural Information Processing Systems 30*, pages 971–980, 2017. URL https://papers.nips.cc/paper/6698-self-normalizing-neural-networks.pdf.

[165] Nancy A Knechel. Tuberculosis: Pathophysiology, Clinical Features, and Diagnosis. *Critical Care Nurse*, 29(2):34–43, 2009. doi:10.4037/ccn2009968.

[166] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. 2012. URL http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[167] Martin Krzywinski and Naomi Altman. Points of significance: Importance of being uncertain. *Nature Methods*, 10(9):809–810, 2013. ISSN 15487091. doi:10.1038/nmeth.2613.

[168] Thomas S Kuhn. *The Structure of Scientific Revolutions*, volume 2. University of Chicago Press, 1970. ISBN 0226458032. URL https://www.lri.fr/~mbl/Stanford/CS477/papers/Kuhn-SSR-2ndEd.pdf.

[169] Jan Martin Kuhnigk, Volker Dicken, Lars Bornemann, Annemarie Bakai, Dag Wormanns, Stefan Krass, and Heinz Otto Peitgen. Morphological segmentation and partial volume analysis for volumetry of solid pulmonary lesions in thoracic CT scans. *IEEE Transactions on Medical Imaging*, 25(4):417–434, 4 2006. doi:10.1109/TMI.2006.871547.

[170] Jan Kukačka, Vladimir Golkov, and Daniel Cremers. Regularization for Deep Learning: A Taxonomy. 10 2017. URL https://arxiv.org/abs/1710.10686v1.

[171] Rodney LaLonde and Ulas Bagci. Capsules for Object Segmentation. In *Medical Imaging with Deep Learning*, number MIDL, pages 1–9, 2018. URL http://arxiv.org/abs/1804.04241.

[172] Philippe Lambin, Ralph T H Leijenaar, Timo M Deist, Jurgen Peerlings, Evelyn E C de Jong, Janita van Timmeren, Sebastian Sanduleanu, Ruben T H M Larue, Aniek J G Even, Arthur Jochems, Yvonka van Wijk, Henry Woodruff, Johan van Soest,

Tim Lustberg, Erik Roelofs, Wouter van Elmpt, Andre Dekker, Felix M Mottaghy, Joachim E Wildberger, and Sean Walsh. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*, advance on, 10 2017. ISSN 1759-4782. doi:10.1038/nrclinonc.2017.141.

[173] Y. LeCun, Y. Bengio, and G. Hinton. Deep Learning. *Nature*, 521(7553):436–444, 2015. doi:doi:10.1038/nature14539.

[174] Yann LeCun and Yoshua Bengio. The handbook of brain theory and neural networks. *Convolutional Networks for Images, Speech and Time-Series*, 1998. URL https://dl.acm.org/citation.cfm?id=303704.

[175] Yann LeCun and Misram Ishan. Self-supervised learning: The dark matter of intelligence, 3 2021. URL https://ai.facebook.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/.

[176] Hoileong Lee, Tahreema Matin, Fergus Gleeson, and Vicente Grau. Efficient 3D Fully Convolutional Networks for Pulmonary Lobe Segmentation in CT Images. 9 2019. URL http://arxiv.org/abs/1909.07474.

[177] Sang Hwan Lee, Sang Min Lee, Jin Mo Goo, Kwang-Gi Kim, Young Jae Kim, and Chang Min Park. Usefulness of Texture Analysis in Differentiating Transient from Persistent Part-solid Nodules(PSNs): A Retrospective Study. *PLOS ONE*, 9(1):e85167, 1 2014. ISSN 1932-6203. doi:10.1371/JOURNAL.PONE.0085167.

[178] Gaetan Lehmann. Label object representation and manipulation with ITK. Technical report, 2007. URL http://hdl.handle.net/1926/584.

[179] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. *International Journal of Computer Vision*, 127(5): 456–476, 11 2014. URL https://arxiv.org/abs/1411.5908v2.

[180] David M. Lewinsohn, Ian S. Tydeman, Marisa Frieder, Jeff E. Grotzke, Rebecca A. Lines, Sheela Ahmed, Kamm D. Prongay, Steven L. Primack, Lois M A Colgin, Anne D. Lewis, and Deborah A. Lewinsohn. High resolution radiographic and fine immunologic definition of TB disease progression in the rhesus macaque. *Microbes and Infection*, 8(11):2587–2598, 2006. ISSN 12864579. doi:10.1016/j.micinf.2006.07.007.

[181] David Kellogg Lewis. Counterfactuals. *Journal of Symbolic Logic*, 44(2):278–281, 6 1973. ISSN 0022-4812. doi:10.2307/2273738.

[182] Baojun Li, Gary E. Christensen, Eric A. Hoffman, Geoffrey McLennan, and Joseph M. Reinhardt. Establishing a normative atlas of the human lung: Intersubject warping and registration of volumetric CT images. *Academic Radiology*, 10(3):255–265, 2003. doi:10.1016/S1076-6332(03)80099-5.

[183] Lin Li, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, Yi Lu, Zhenghan Fang, Qi Song, Kunlin Cao, Daliang Liu, Guisheng Wang, Qizhong Xu, Xisheng Fang, Shiqin Zhang, Juan Xia, and Jun Xia. Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneumonia on Chest CT. *Radiology*, 296(2):E65–E71, 8 2020. doi:10.1148/RADIOL.2020200905.

[184] Xuechen Li, Linlin Shen, Xinpeng Xie, Shiyun Huang, Zhien Xie, Xian Hong, and Juan Yu. Multi-resolution convolutional networks for chest X-ray radiograph based lung nodule detection. *Artificial Intelligence in Medicine*, 103:101744, 3 2020. ISSN 0933-3657. doi:10.1016/J.ARTMED.2019.101744.

[185] Fangzhou Liao, Ming Liang, Zhe Li, Xiaolin Hu, and Sen Song. Evaluate the Malignancy of Pulmonary Nodules Using the 3-D Deep Leaky Noisy-or Network. *IEEE Transactions on Neural Networks and Learning Systems*, PP:1–12, 2019. ISSN 2162-237X. doi:10.1109/tnnls.2019.2892409.

[186] Wey Wen Lim, Nancy H L Leung, Sheena G Sullivan, Eric J Tchetgen Tchetgen, and Benjamin J Cowling. Distinguishing Causation from Correlation in the Use of Correlates of Protection to Evaluate and Develop Influenza Vaccines. *American Journal of Epidemiology*, 189(3):185–192, 3 2020. ISSN 0002-9262. doi:10.1093/AJE/KWZ227.

[187] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–10, 2018. ISSN 0162-8828. doi:10.1109/TPAMI.2018.2858826.

[188] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 12 2017. ISSN 1361-8415. doi:10.1016/J.MEDIA.2017.07.005.

[189] Caixia Liu, Ruibin Zhao, and Mingyong Pang. A fully automatic segmentation algorithm for CT lung images based on random forest. *Medical Physics*, 47(2): 518–529, 2 2020. ISSN 2473-4209. doi:10.1002/MP.13939.

[190] Shuang Liu, Yiting Xie, and Anthony P Reeves. Segmentation of the sternum from low-dose chest CT images. In *Proc. SPIE*, page 941403. International Society for Optics and Photonics, 2015. ISBN 9781628415049. doi:10.1117/12.2082436.

[191] Xiao Liu, Pedro Sanchez, Spyridon Thermos, Alison Q. O'Neil, and Sotirios A. Tsaftaris. A Tutorial on Learning Disentangled Representations in the Imaging Domain. 8 2021. URL https://arxiv.org/abs/2108.12043v1.

[192] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-Centric Learning with Slot Attention. *Advances in Neural Information Processing Systems*, 2020-December, 6 2020. ISSN 10495258. URL https://arxiv.org/abs/2006.15055v2.

[193] Alex John London. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, 49(1):15–21, 1 2019. ISSN 1552-146X. doi:10.1002/HAST.973.

[194] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Schölkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 58–66, 2017. ISBN 9781538604571. doi:10.1109/CVPR.2017.14.

[195] Konrad Lorenz. *Behind the Mirror: A Search for a Natural History of Human Knowledge.*, volume 12. Piper Verlag, Munich, 1973. ISBN 978-3-492-02030-5. doi:10.2307/2800563.

[196] Octavio Loyola-Gonzalez. Black-box vs. White-Box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113, 2019. doi:10.1109/ACCESS.2019.2949286.

[197] Brian Luna, André Kubler, Christer Larsson, Brent Foster, Ulas Bagci, Daniel J. Mollura, Sanjay K. Jain, and William R. Bishai. In vivo prediction of tuberculosis-associated cavity formation in rabbits. *Journal of Infectious Diseases*, 211(3):481–485, 2 2015. ISSN 15376613. doi:10.1093/infdis/jiu449.

[198] Yunan Luo, Jian Peng, and Jianzhu Ma. When causal inference meets deep learning. *Nature Machine Intelligence 2020 2:8*, 2(8):426–427, 8 2020. ISSN 2522-5839. doi:10.1038/s42256-020-0218-x.

[199] Li Ma and Suohai Fan. CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. *BMC Bioinformatics*, 18(1):169, 2017. ISSN 1471-2105. doi:10.1186/s12859-017-1578-z.

[200] Thomas M. Maddox, John S. Rumsfeld, and Philip R. O. Payne. Questions for Artificial Intelligence in Health Care. *JAMA*, 321(1):31–32, 1 2019. ISSN 0098-7484. doi:10.1001/JAMA.2018.18932.

[201] Suren Makaju, P. W.C. Prasad, Abeer Alsadoon, A. K. Singh, and A. Elchouemi. Lung Cancer Detection using CT Scan Images. *Procedia Computer Science*, 125:107–114, 1 2018. ISSN 1877-0509. doi:10.1016/J.PROCS.2017.12.016.

[202] Awais Mansoor, Ulas Bagci, Brent Foster, Ziyue Xu, Deborah Douglas, Jeffrey M Solomon, Jayaram K Udupa, and Daniel J Mollura. CIDI-lung-seg: A single-click annotation tool for automatic delineation of lungs from CT scans. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 1087–1090, 2014. ISBN 9783902823625. doi:10.1109/EMBC.2014.6943783.

[203] Awais Mansoor, Ulas Bagci, Ziyue Xu, Brent Foster, Kenneth N Olivier, Jason M Elinoff, Anthony F Suffredini, Jayaram K Udupa, and Daniel J Mollura. A Generic Approach to Pathological Lung Segmentation. *IEEE Transactions On Medical Imaging*, 33(12):2293–2310, 2014. doi:10.1109/TMI.2014.2337057.

[204] Awais Mansoor, Ulas Bagci, Brent Foster, Ziyue Xu, Georgios Z Papadakis, Les R Folio, Jayaram K Udupa, and Daniel J Mollura. Segmentation and Image Analysis of Abnormal Lungs at CT: Current Approaches, Challenges, and Future Trends. *Radio-Graphics*, 35(4):1056–1076, 2015. ISSN 0271-5333. doi:10.1148/rg.2015140232.

[205] Paula Martin-Gonzalez, Estibaliz Gomez de Mariscal, M. Elena Martino, Pedro M. Gordaliza, Isabel Peligros, Jose Luis Carreras, Felipe A. Calvo, Javier Pascau, Manuel Desco, and Arrate Muñoz-Barrutia. Association of visual and quantitative heterogeneity of 18F-FDG PET images with treatment response in locally advanced rectal cancer: A feasibility study. *PLOS ONE*, 15(11):e0242597, 11 2020. ISSN 1932-6203. doi:10.1371/journal.pone.0242597.

[206] Robert Matthews. Storks Deliver Babies (p= 0.008). *Teaching Statistics*, 22(2):36–38, 6 2000. ISSN 1467-9639. doi:10.1111/1467-9639.00013.

[207] Fabrizio Menardo, Sebastian Duchêne, Daniela Brites, and Sebastien Gagneux. The molecular clock of mycobacterium tuberculosis. *PLoS Pathogens*, 15(9), 2019. ISSN 15537374. doi:10.1371/journal.ppat.1008067.

[208] Bjoern H Menze, Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. 2009. doi:10.1186/1471-2105-10-213.

[209] Temesguen Messay, Russell C. Hardie, and Timothy R. Tuinstra. Segmentation of pulmonary nodules in computed tomography using a regression neural network approach and its application to the Lung Image Database Consortium and Image Database Resource Initiative dataset. *Medical Image Analysis*, 22(1):48–62, 2015. doi:10.1016/j.media.2015.02.002.

[210] Fausto Milletari, Nassir Navab, and Seyed Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *2016 Fourth International Conference on 3D Vision (3DV)*, 565 - 571:565–571, 2016. doi:10.1109/3DV.2016.79.

[211] Jane L. Mitchell, Edward T. Mee, Neil M. Almond, Keith Cutler, and Nicola J. Rose. Characterisation of MHC haplotypes in a breeding colony of Indonesian cynomolgus macaques reveals a high level of diversity. *Immunogenetics*, 64(2):123–129, 2012. ISSN 00937711. doi:10.1007/s00251-011-0567-z.

[212] Shakir Mohamed. Planting the Seeds of Probabilistic Thinking (Bayesian Learning), 2018. URL https://www.shakirm.com/slides/MLSS2018-Madrid-ProbThinking.pdf.

[213] Grégoire Montavon, Wojciech Samek, and Klaus Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2 2018. ISSN 1051-2004. doi:10.1016/J.DSP.2017.10.011.

[214] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 1 2012. ISSN 0031-3203. doi:10.1016/J.PATCOG.2011.06.019.

[215] Filipa Mota, Ravindra Jadhav, Camilo A. Ruiz-Bedoya, Alvaro A. Ordonez, Mariah H. Klunk, Joel S. Freundlich, and Sanjay K. Jain. Radiosynthesis and Biodistribution of 18F-Linezolid in Mycobacterium tuberculosis-Infected Mice Using Positron Emission Tomography. *ACS infectious diseases*, 6(5):916–921, 5 2020. ISSN 23738227. doi:10.1021/acsinfecdis.9b00473.

[216] Kevin P Murphy. *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. URL https://probml.ai.

[217] Arun C Nachiappan, Kasra Rahbar, Xiao Shi, Elizabeth S Guy, Eduardo Mortani Barbosa Jr, Girish S Shroff, Daniel Ocazionez, Alan E Schlesinger, Sharyn I Katz, and Mark M Hammer. Pulmonary Tuberculosis: Role of Radiology in Diagnosis and Management. *RadioGraphics*, 37(1):52–72, 2017. doi:10.1148/rg.2017160032.

[218] Tanya Nair, Doina Precup, Douglas L. Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for Multiple sclerosis lesion detection and segmentation. *Medical Image Analysis*, 59:101557, 1 2020. ISSN 13618423. doi:10.1016/j.media.2019.101557.

[219] Esperanza Naredo, Javier Pascau, Nemanja Damjanov, Gemma Lepri, Pedro M Gordaliza, Iustina Janta, Juan Gabriel Ovalles-Bonilla, Francisco Javier López-Longo, and Marco Matucci-Cerinic. Performance of ultra-high-frequency ultrasound in the evaluation of skin involvement in systemic sclerosis: a preliminary report. *Rheumatology*, 59(7):1671–1678, 10 2019. ISSN 1462-0324. doi:10.1093/rheumatology/kez439.

[220] Radford M. Neal and Geoffrey E. Hinton. A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants. In *Learning in Graphical Models*, volume 89, pages 355–368. Springer, 1998. doi:10.1007/978-94-011-5014-9_12.

[221] Mohammadreza Negahdar, David Beymer, and Tanveer F. Syeda-Mahmood. Automated volumetric lung segmentation of thoracic CT images using fully convolutional neural network. In *SPIE Medical Imaging*, volume 10575, pages 1–6, 2018. ISBN 9781510616394. doi:10.1117/12.2293723.

[222] Andreas G Nerlich, Christian J Haas, Albert Zink, Ulrike Szeimies, and Hjalmar G Hagedorn. Molecular evidence for tuberculosis in an ancient Egyptian mummy. *The Lancet*, 350(9088):1404, 11 1997. ISSN 0140-6736. doi:10.1016/S0140-6736(05)65185-9.

[223] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning. *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings*, 12 2014. URL https://arxiv.org/abs/1412.6614v4.

[224] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? 8 2020. URL http://arxiv.org/abs/2008.11687.

[225] Jean Claude Semuto Ngabonziza, Chloé Loiseau, Michael Marceau, Agathe Jouet, Fabrizio Menardo, Oren Tzfadia, Rudy Antoine, Esdras Belamo Niyigena, Wim Mulders, Kristina Fissette, Maren Diels, Cyril Gaudin, Stéphanie Duthoy, Willy Ssengooba, Emmanuel André, Michel K. Kaswa, Yves Mucyo Habimana, Daniela Brites, Dissou Affolabi, Jean Baptiste Mazarati, Bouke Catherine de Jong, Leen Rigouts, Sebastien Gagneux, Conor Joseph Meehan, and Philip Supply. A sister lineage of the Mycobacterium tuberculosis complex discovered in the African Great Lakes region. *Nature Communications 2020 11:1*, 11(1):1–11, 6 2020. ISSN 2041-1723. doi:10.1038/s41467-020-16626-6.

[226] Richard Nock and Frank Nielsen. Statistical Region Merging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(11):1452–1458, 2004. doi:10.1109/TPAMI.2004.110.

[227] Norliza Mohd Noor, Joel Chia, Ming Than, Omar Mohd Rijal, N M Noor, J C M Than, and O M Rijal. Performance Evaluation of Lung Segmentation. In *Medical Imaging Technology: Reviews and Computational Applications*, chapter 5, pages 111–127. Springer, 2015. doi:10.1007/978-981-287-540-2_5.

[228] Anna Odone, Rein M.G.J. Houben, Richard G. White, and Knut Lönnroth. The effect of diabetes and undernutrition trends on reaching 2035 global tuberculosis targets. *The Lancet Diabetes & Endocrinology*, 2(9):754–764, 9 2014. ISSN 2213-8587. doi:10.1016/S2213-8587(14)70164-0.

[229] Timo Ojala, Matti Pietikäinen, and David Harwood. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. *Proceedings - International Conference on Pattern Recognition*, 3:582–585, 1994. ISSN 10514651. doi:10.1109/ICPR.1994.576366.

[230] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, Lasse Espeholt, Alex Graves, and Koray Kavukcuoglu. Conditional Image Generation with PixelCNN Decoders. *Advances in Neural Information Processing Systems*, pages 4797–4805, 6 2016. URL https://arxiv.org/abs/1606.05328v2.

[231] Alvaro A. Ordonez, Laurence S. Carroll, Sudhanshu Abhishek, Filipa Mota, Camilo A. Ruiz-Bedoya, Mariah H. Klunk, Alok K. Singh, Joel S. Freundlich, Ronnie C. Mease, and Sanjay K. Jain. Radiosynthesis and PET Bioimaging of 76Br-Bedaquiline in a Murine Model of Tuberculosis. *ACS Infectious Diseases*, 5(12):1996–2002, 12 2019. ISSN 23738227. doi:10.1021/acsinfecdis.9b00207.

[232] Alvaro A. Ordonez, Hechuan Wang, Gesham Magombedze, Camilo A. Ruiz-Bedoya, Shashikant Srivastava, Allen Chen, Elizabeth W. Tucker, Michael E. Urbanowski, Lisa Pieterse, E. Fabian Cardozo, Martin A. Lodge, Maunank R. Shah, Daniel P. Holt, William B. Mathews, Robert F. Dannals, Jogarao V.S. Gobburu, Charles A. Peloquin, Steven P. Rowe, Tawanda Gumbo, Vijay D. Ivaturi, and Sanjay K. Jain. Dynamic imaging in patients with tuberculosis reveals heterogeneous drug exposures in pulmonary lesions. *Nature Medicine*, 26(4):529–534, 2020. ISSN 1546170X. doi:10.1038/s41591-020-0770-2.

[233] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979. doi:10.1109/TSMC.1979.4310076.

[234] Madhukar Pai, Alice Zwerling, and Dick Menzies. Systematic review: T-cell-based assays for the diagnosis of latent tuberculosis infection: An update. *Annals of Internal Medicine*, 149(3):177–184, 2008. ISSN 00034819. doi:10.7326/0003-4819-149-3-200808050-00241.

[235] Madhukar Pai, Marcel A. Behr, David Dowdy, Keertan Dheda, Maziar Divangahi, Catharina C. Boehme, Ann Ginsberg, Soumya Swaminathan, Melvin Spigelman, Haileyesus Getahun, Dick Menzies, and Mario Raviglione. Tuberculosis. *Nature Reviews Disease Primers*, 2:1–23, 2016. ISSN 2056676X. doi:10.1038/nrdp.2016.76.

[236] Javier Pascau, Juan José Vaquero, Mónica Abella, R Cacho, Eduardo Lage, and Manuel Desco. Multimodality Workstation For Small Animal Image Visualization And Analysys. In *AMI Annual Conference*, volume 8, pages 97–98, 2006. ISBN 1857545672 (pbk.) : ¹7.95. doi:10.1007/s11307-006-0031-x.

[237] Nick Pawlowski, Ira Ktena, Matthew C H Lee, Bernhard Kainz, Daniel Rueckert, Ben Glocker, and Martin Rajchl. DLTK: State of the Art Reference Implementations for Deep Learning on Medical Images. *ArXiv*, 2017. URL https://dltk.github.io/.

[238] Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Deep Structural Causal Models for Tractable Counterfactual Inference. In *Neural Information Processing Systems (NIPS)*, 6 2020. URL http://arxiv.org/abs/2006.06485.

[239] Judea Pearl. *Heuristics: intelligent search strategies for computer problem solving*, volume 1. Addison-Wiley, 1984. doi:10.1002/int.4550010107.

[240] Judea Pearl. *Causality: Models, reasoning, and inference, second edition*. 2011. ISBN 9780511803161. doi:10.1017/CBO9780511803161.

[241] Judea Pearl. The Do-Calculus Revisited. Keynote Lecture, August 17, 2012. UAI-2012 Conference, Catalina, CA. *Proceedings of the Twenty-Eight Conference on Uncertainty in Artificial Intelligence*, (August):4–11, 2012. ISSN 0000-0000. URL https://arxiv.org/pdf/1210.4852.pdf.

[242] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2 2019. ISSN 00010782. doi:10.1145/3241036.

[243] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, NY, USA, 2018. ISBN 9780465097616. doi:10.4324/9780429462764.

[244] Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. *Causal Inference in Statistics*. Wiley, 2016. ISBN 9781119186847.

[245] Juliet C. Peña and Wen Zhe Ho. Monkey models of tuberculosis: Lessons learned. *Infection and Immunity*, 83(3):852–862, 2015. ISSN 10985522. doi:10.1128/IAI.02850-14.

[246] Luis Perez and Jason Wang. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. 12 2017. URL https://arxiv.org/abs/1712.04621v1.

[247] Christian S. Perone, Pedro Ballester, Rodrigo C. Barros, and Julien Cohen-Adad. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11, 7 2019. ISSN 1053-8119. doi:10.1016/J.NEUROIMAGE.2019.03.026.

[248] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 78(5):947–1012, 11 2016. ISSN 14679868. doi:10.1111/rssb.12167.

[249] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, London, England, 2017. ISBN 9780262037310. URL http://web.math.ku.dk/~peters/.

[250] Thomas Piekos. Confidence Connected Segmentation With ITK. Technical report, 2007. URL http://hdl.handle.net/1926/1306.

[251] Kent E. Pinkerton, Laura S. Van Winkle, Charles G. Plopper, Suzette Smiley-Jewell, Elvira C. Covarrubias, and John T. McBride. Architecture of the Tracheobronchial Tree. In *Comparative Biology of the Normal Lung*, pages 33–51. Elsevier, 2015. doi:10.1016/B978-0-12-404577-4.00004-7.

[252] Mithun N Prasad and Arcot Sowmya. Multilevel emphysema diagnosis of HRCT lung images in an incremental framework. In *Medical Imaging 2004*, pages 42–50. International Society for Optics and Photonics, 2004. doi:10.1117/12.533943.

[253] Jiantao Pu, Justus Roos, Chin A. Yi, Sandy Napel, Geoffrey D. Rubin, and David S. Paik. Adaptive border marching algorithm: Automatic lung segmentation on chest CT images. *Computerized Medical Imaging and Graphics*, 32(6):452–462, 9 2008. ISSN 0895-6111. doi:10.1016/J.COMPMEDIMAG.2008.04.005.

[254] Xiaolong Qi, Zicheng Jiang, Qian Yu, Chuxiao Shao, Hongguang Zhang, Hongmei Yue, Baoyi Ma, Yuancheng Wang, Chuan Liu, Xiangpan Meng, Shan Huang, Jitao Wang, Dan Xu, Junqiang Lei, Guanghang Xie, Huihong Huang, Jie Yang, Jiansong Ji, Hongqiu Pan, Shengqiang Zou, and Shenghong Ju. Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: A multicenter study. *medRxiv*, page 2020.02.29.20029603, 3 2020. doi:10.1101/2020.02.29.20029603.

[255] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding Transfer Learning for Medical Imaging. In *Neural Computing and Applications*, 2019. URL https://proceedings.neurips.cc/paper/2019/file/eb1e78328c46506b46a4ac4a1e378b91-Paper.pdf.

[256] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. pages 3–9, 2017. URL http://arxiv.org/abs/1711.05225.

[257] Pedro Pedrosa Rebouças Filho, Paulo César Cortez, Antônio C. da Silva Barros, Victor Hugo Victor, and R. S.João Manuel Tavares. Novel and powerful 3D adaptive crisp active contour method applied in the segmentation of CT lung images. *Medical Image Analysis*, 35:503–516, 1 2017. ISSN 1361-8415. doi:10.1016/J.MEDIA.2016.09.002.

[258] Jacob C. Reinhold, Aaron Carass, and Jerry L. Prince. A Structural Causal Model for MR Images of Multiple Sclerosis. *Lecture Notes in Computer Science*, 12905 LNCS: 782–792, 2021. ISSN 16113349. doi:10.1007/978-3-030-87240-3_75.

[259] Annika Reinke, Matthias Eisenmann, Minu D. Tizabi, Carole H. Sudre, Tim Rädsch, Michela Antonelli, Tal Arbel, Spyridon Bakas, M. Jorge Cardoso, Veronika Cheplygina, Keyvan Farahani, Ben Glocker, Doreen Heckmann-Nötzel, Fabian Isensee, Pierre Jannin, Charles E. Kahn, Jens Kleesiek, Tahsin Kurc, Michal Kozubek, Bennett A. Landman, Geert Litjens, Klaus Maier-Hein, Bjoern Menze, Henning Müller, Jens Petersen, Mauricio Reyes, Nicola Rieke, Bram Stieltjes, Ronald M. Summers, Sotirios A. Tsaftaris, Bram van Ginneken, Annette Kopp-Schneider, Paul Jäger, and Lena Maier-Hein. Common Limitations of Image Processing Metrics: A Picture Story. 4 2021. URL http://arxiv.org/abs/2104.05642.

[260] Norman Ricker. Wavelet Contraction, Wavelet Expansion, and the Control of Seismic Resolution. *Geophysics*, 18(4):769–792, 10 1953. doi:10.1190/1.1437927.

[261] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, Alessandro Ruggiero, Anna Korhonen, Emily Jefferson, Emmanuel Ako, Georg Langs, Ghassem Gozaliasl, Guang Yang, Helmut Prosch, Jacobus Preller, Jan Stanczuk, Jing Tang, Johannes Hofmanninger, Judith Babar, Lorena Escudero Sánchez, Muhunthan Thillai, Paula Martin Gonzalez, Philip Teare, Xiaoxiang Zhu, Mishal Patel, Conor Cafolla, Hojjat Azadbakht, Joseph Jacob, Josh Lowe, Kang Zhang, Kyle Bradley, Marcel Wassin, Markus Holzer, Kangyu Ji, Maria Delgado Ortet, Tao Ai, Nicholas Walton, Pietro Lio, Samuel Stranks, Tolou Shadbahr, Weizhe Lin, Yunfei Zha, Zhangming Niu, James H.F. Rudd, Evis Sala, and Carola Bibiane Schönlieb. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3 (3):199–217, 3 2021. ISSN 25225839. doi:10.1038/s42256-021-00307-0.

[262] Sara Yukie Rodriguez-Takeuchi, Martin Eduardo Renjifo, and Francisco José Medina. Extrapulmonary tuberculosis: Pathophysiology and imaging findings. *Radiographics*, 39(7):2023–2037, 11 2019. ISSN 15271323. doi:10.1148/rg.2019190109.

[263] Kamila Romanowski, Brett Baumann, C. Andrew Basham, Faiz Ahmad Khan, Greg J. Fox, and James C. Johnston. Long-term all-cause mortality in people treated for tuberculosis: a systematic review and meta-analysis. *The Lancet Infectious Diseases*, 19(10):1129–1137, 10 2019. ISSN 1473-3099. doi:10.1016/S1473-3099(19)30309-3.

[264] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science*, volume 9351, pages 234–241. Springer, Cham, 2015. ISBN 9783319245737. doi:10.1007/978-3-319-24574-4_28.

[265] Stefanie Rosenhain, Zuzanna A. Magnuska, Grace G. Yamoah, Wa'El Al Rawashdeh, Fabian Kiessling, and Felix Gremse. A preclinical micro-computed tomography database including 3D whole body organ segmentations. *Scientific Data*, 5(1):1–9, 12 2018. ISSN 20524463. doi:10.1038/sdata.2018.294.

[266] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. QuickNAT: Segmenting MRI Neuroanatomy in 20 seconds, 2018. URL http://arxiv.org/abs/1801.04161.

[267] Arunabha S Roy, Samuel G Armato III, Andrew Wilson, and Karen Drukker. Automated detection of lung nodules in CT scans: false-positive reduction with the radial-gradient index. *Medical physics*, 33(4):1133–1140, 2006. doi:10.1118/1.2178450.

[268] Sebastian Ruder. An overview of gradient descent optimization algorithms. 9 2016. URL http://arxiv.org/abs/1609.04747.

[269] Sebastian Ruder. An Overview of Multi-Task Learning in Deep Neural Networks. *ArXiv*, 6 2017. URL https://arxiv.org/abs/1706.05098.

[270] Yvan Saeys, Thomas Abeel, and Yves de Peer. Robust Feature Selection Using Ensemble Feature Selection Techniques. In *ECML PKDD*, pages 313–325, 2008. ISBN 978-3-540-87481-2. doi:10.1007/978-3-540-87481-2_21.

[271] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from Synthetic Data: Addressing Domain Shift for Semantic Segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3752–3761, 2018. ISBN 9781538664209. doi:10.1109/CVPR.2018.00395.

[272] Charles A Scanga and Joanne L Flynn. Modeling Tuberculosis in Nonhuman Primates. *Cold Spring Harbor Perspectives in Medicine*, 4(3):1–17, 2014. doi:10.1101/cshperspect.a018564.

[273] Charles A. Scanga, Brian J. Lopresti, Jaime Tomko, Lonnie J. Frye, Teresa M. Coleman, Daniel Fillmore, Jonathan P. Carney, Philana L. Lin, Jo Anne L Flynn, Christina L. Gardner, Chengqun Sun, William B. Klimstra, Kate D. Ryman, Douglas S. Reed, Daniel J. Fisher, and Kelly S. Cole. In vivo imaging in an ABSL-3 regional biocontainment laboratory. *Pathogens and disease*, 71(2):207–212, 2014. ISSN 2049632X. doi:10.1111/2049-632X.12186.

[274] T. Schlathoelter, C. Lorenz, I. C. Carlsen, S. Renisch, and T. Deschamps. Simultaneous segmentation and tree reconstruction of the airways for virtual bronchoscopy. In *Proc. SPIE*, pages 103–113, 2002. doi:10.1117/12.467061.

[275] Bernhard Scholkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5):612–634, 5 2021. ISSN 15582256. doi:10.1109/JPROC.2021.3058954.

[276] Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. Using StyleGAN for Visual Interpretability of Deep Learning Models on Medical Images. 1 2021. URL https://arxiv.org/abs/2101.07563v1.

[277] Arnaud A. A. Setio, Colin Jacobs, Jaap Gelderblom, and Bram van Ginneken. Automatic detection of large pulmonary solid nodules in thoracic CT images. *Medical Physics*, 42(10):5642–5653, 2015. doi:10.1118/1.4929562.

[278] Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J. Van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I. Sanchez, and Bram Van Ginneken. Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks. *IEEE Transactions on Medical Imaging*, 35(5):1160–1169, 5 2016. ISSN 1558254X. doi:10.1109/TMI.2016.2536809.

[279] Dominik Seuß. Bridging the Gap Between Explainable AI and Uncertainty Quantification to Enhance Trustability. 5 2021. URL https://arxiv.org/abs/2105.11828v1.

[280] Arash Shaban-Nejad, Martin Michalowski, John Brownstein, and David Buckeridge. Guest Editorial Explainable AI: Towards Fairness, Accountability, Transparency and Trust in Healthcare. *IEEE Journal of Biomedical and Health Informatics*, 25(7): 2374–2375, 7 2021. doi:10.1109/JBHI.2021.3088832.

[281] Fei Shan, Yaozong Gao, Jun Wang, Weiya Shi, Nannan Shi, Miaofei Han, Zhong Xue, Dinggang Shen, and Yuxin Shi. Abnormal lung quantification in chest CT images of COVID-19 patients with deep learning and its application to severity prediction. *Medical Physics*, 48(4):1633–1645, 4 2021. ISSN 2473-4209. doi:10.1002/MP.14609.

[282] Anuj Sharma and Shubhamoy Dey. Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis. *International Journal of Computer Applications*, (June):15–20, 2012. URL https://arxiv.org/pdf/1309.3949.pdf.

[283] Surendra K Sharma, Alladi Mohan, and Mikashmi Kohli. Extrapulmonary tuberculosis. *Expert Review of Respiratory Medicine*, 15(7):931–948, 7 2021. doi:10.1080/17476348.2021.1927718.

[284] S. A. Sharpe, H. McShane, M. J. Dennis, R. J. Basaraba, F. Gleeson, G. Hall, A. McIntyre, K. Gooch, S. Clark, N. E R Beveridge, E. Nuth, A. White, A. Marriott, S. Dowall, A. V S Hill, A. Williams, and P. D. Marsh. Establishment of an aerosol challenge model of tuberculosis in rhesus macaques and an evaluation of endpoints for vaccine testing. *Clinical and Vaccine Immunology*, 17(8):1170–1182, 2010. ISSN 1556679X. doi:10.1128/CVI.00079-10.

[285] Sally Sharpe, Andrew White, Fergus Gleeson, Anthony McIntyre, Donna Smyth, Simon Clark, Charlotte Sarfas, Dominick Laddy, Emma Rayner, Graham Hall, Ann Williams, and Mike Dennis. Ultra low dose aerosol challenge with Mycobacterium tuberculosis leads to divergent outcomes in rhesus and cynomolgus macaques. *Tuberculosis*, 96:1–12, 2016. ISSN 1873281X. doi:10.1016/j.tube.2015.10.004.

[286] Cong Shen, Nan Yu, Shubo Cai, Jie Zhou, Jiexin Sheng, Kang Liu, Heping Zhou, Youmin Guo, and Gang Niu. Quantitative computed tomography analysis for stratifying the severity of Coronavirus Disease 2019. *Journal of Pharmaceutical Analysis*, 10 (2):123–129, 4 2020. ISSN 2095-1779. doi:10.1016/J.JPHA.2020.03.004.

[287] Feng Shi, Liming Xia, Fei Shan, Bin Song, Dijia Wu, Ying Wei, Huan Yuan, Huiting Jiang, Yichu He, Yaozong Gao, He Sui, and Dinggang Shen. Large-scale screening to distinguish between COVID-19 and community-acquired pneumonia using infection size-aware classification. *Physics in Medicine and Biology*, 66(6):065031, 3 2021. ISSN 13616560. doi:10.1088/1361-6560/abe838.

[288] Hoo Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Senior Member, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016. ISSN 0278-0062. doi:10.1109/TMI.2016.2528162.

[289] Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data 2019 6:1*, 6(1):1–48, 7 2019. ISSN 2196-1115. doi:10.1186/S40537-019-0197-0.

[290] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*, 12 2013. URL https://arxiv.org/abs/1312.6034v2.

[291] Gur Amrit Pal Singh and P. K. Gupta. Performance analysis of various machine learning-based approaches for detection and classification of lung cancer in humans. *Neural Computing and Applications 2018 31:10*, 31(10):6863–6877, 5 2018. ISSN 1433-3058. doi:10.1007/S00521-018-3518-X.

[292] Sumedha Singla, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. Explaining the Black-box Smoothly- A Counterfactual Approach. 1 2021. URL http://arxiv.org/abs/2101.04230.

[293] Leen Kiat Soh and Costas Tsatsoulis. Texture analysis of sar sea ice imagery using gray level co-occurrence matrices. *IEEE Transactions on Geoscience and Remote Sensing*, 37(2):780–795, 1999. doi:10.1109/36.752194.

[294] Ahmed Soliman, Fahmi Khalifa, Ahmed Elnakib, Mohamed Abou El-Ghar, Neal Dunlap, Brian Wang, Georgy Gimel'Farb, Robert Keynton, and Ayman El-Baz. Accurate Lungs Segmentation on CT Chest Images by Adaptive Appearance-Guided Shape Modeling. *IEEE Transactions on Medical Imaging*, PP(99):263–276, 2016. ISSN 1558254X. doi:10.1109/TMI.2016.2606370.

[295] Yang Song, Weidong Cai, Yun Zhou, and David Dagan Feng. Feature-based image patch approximation for lung tissue classification. *IEEE Transactions on Medical Imaging*, 32(4):797–808, 2013. ISSN 02780062. doi:10.1109/TMI.2013.2241448.

[296] Ying Song, Shuangjia Zheng, Liang Li, Xiang Zhang, Xiaodong Zhang, Ziwang Huang, Jianwen Chen, Ruixuan Wang, Huiying Zhao, Yunfei Zha, Jun Shen, Yutian Chong, and Yuedong Yang. Deep learning Enables Accurate Diagnosis of Novel Coronavirus (COVID-19) with CT images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2021. doi:10.1109/TCBB.2021.3065361.

[297] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. ISSN 15337928. URL http://jmlr.org/papers/v15/srivastava14a.html.

[298] Joes Staal, Bram van Ginneken, and Max A. Viergever. Automatic rib segmentation and labeling in computed tomography scans using a general framework for detection, recognition and segmentation of objects in volumetric data. *Medical Image Analysis*, 11(1):35–46, 2007. ISSN 13618415. doi:10.1016/j.media.2006.10.001.

[299] Thérèse A. Stukel, Elliott S. Fisher, David E. Wennberg, David A. Alter, Daniel J. Gottlieb, and Marian J. Vermeulen. Analysis of Observational Studies in the Presence of Treatment Selection Bias: Effects of Invasive Cardiac Management on AMI Survival Using Propensity Score and Instrumental Variable Methods. *JAMA*, 297(3):278–285, 1 2007. ISSN 0098-7484. doi:10.1001/JAMA.297.3.278.

[300] K. Suzuki, R. Kohlbrenner, M. L. Epstein, A.M. Obajuluwa, J. Xu, and M. Hori. Computer-aided measurement of liver volumes in CT by means of geodesic active contour segmentation coupled with level-set algorithms. *Medical physics*, 37(5): 2159–2166, 2010. doi:10.1118/1.3395579.

[301] Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. In *Advances in Neural Information Processing Systems*, volume 32, pages 6417–6428, 2019. URL https://proceedings.neurips.cc/paper/2019/file/73c03186765e199c116224b68adc5fa0-Paper.pdf.

[302] Aiham Taleb, Christoph Lippert, Tassilo Klein, and Moin Nabi. Multimodal Self-supervised Learning for Medical Image Analysis. pages 661–673, 6 2021. doi:10.1007/978-3-030-78191-0_51.

[303] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A Survey on Deep Transfer Learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11141 LNCS:270–279, 10 2018. doi:10.1007/978-3-030-01424-7_-27.

[304] Naoya Tanabe, Susumu Sato, Kazuya Tanimura, Tsuyoshi Oguma, Atsuyasu Sato, Shigeo Muro, and Toyohiro Hirai. Associations of CT evaluations of antigravity muscles, emphysema and airway disease with longitudinal outcomes in patients with COPD. *Thorax*, 76(3):295–297, 3 2021. ISSN 14683296. doi:10.1136/thoraxjnl-2020-215085.

[305] Zhenyu Tang, Wei Zhao, Xingzhi Xie, Zheng Zhong, Feng Shi, Jun Liu, and Dinggang Shen. Severity Assessment of Coronavirus Disease 2019 (COVID-19) Using Quantitative Features from Chest CT Images. 3 2020. URL https://arxiv.org/abs/2003.11988v1.

[306] Martin A. Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987. doi:10.1080/01621459.1987.10478458.

[307] Yimo Tao, Le Lu, Maneesh Dewan, Albert Y. Chen, Jason Corso, Jianhua Xuan, Marcos Salganicoff, and Arun Krishnan. Multi-level ground glass nodule detection and segmentation in CT lung images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention. MICCAI*, pages 715–723, 2009. ISBN 3642042708. doi:10.1007/978-3-642-04271-3_87.

[308] Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the National Conference on Artificial Intelligence*, pages 567–573, 2002. URL https://www.aaai.org/Library/AAAI/2002/aaai02-085.php.

[309] Eric J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 1 2019. ISSN 1078-8956. doi:10.1038/s41591-018-0300-7.

[310] Nestor Tsafack, Syam Sankar, Bassem Abd-El-Atty, Jacques Kengne, Jithin K. C., Akram Belazi, Irfan Mehmood, Ali Kashif Bashir, Oh Young Song, and Ahmed A.Abd El-Latif. A New Chaotic Map with Dynamic Analysis and Encryption Application in Internet of Health Things. *IEEE Access*, 8:137731–137744, 2020. doi:10.1109/ACCESS.2020.3010794.

[311] Andreas Tsamados, Nikita Aggarwal, Josh Cowls, Jessica Morley, Huw Roberts, Mariarosaria Taddeo, and Luciano Floridi. The ethics of algorithms: key problems and solutions. *AI & SOCIETY 2021*, 1:1–16, 2 2021. ISSN 1435-5655. doi:10.1007/S00146-021-01154-8.

[312] Robert R. Tucci. Introduction to Judea Pearl's Do-Calculus. 4 2013. URL http://arxiv.org/abs/1305.5506.

[313] Jayaram K. Udupa, Vicki R. LeBlanc, Ying Zhuge, Celina Imielinska, Hilary Schmidt, Leanne M. Currie, Bruce E. Hirsch, and James Woodburn. A framework for evaluating image segmentation algorithms. *Computerized Medical Imaging and Graphics*, 30(2): 75–87, 2006. doi:10.1016/j.compmedimag.2005.12.001.

[314] Michael E. Urbanowski, Alvaro A. Ordonez, Camilo A. Ruiz-Bedoya, Sanjay K. Jain, and William R. Bishai. Cavitary tuberculosis: the gateway of disease transmission. *The Lancet Infectious Diseases*, 20(6):e117–e128, 6 2020. ISSN 14744457. doi:10.1016/S1473-3099(20)30148-1.

[315] Arash Vahdat and Jan Kautz. NVAE: A Deep Hierarchical Variational Autoencoder. *Advances in Neural Information Processing Systems*, 2020-Decem, 7 2020. ISSN 10495258. URL https://arxiv.org/abs/2007.03898v3.

[316] Miss Hetal J Vala and Astha Baxi. A review on Otsu image segmentation algorithm. *International Journal of Advanced Research in Computer Engineering and Technology*, 2(2):387–389, 2013. URL http://ijarcet.org/wp-content/uploads/IJARCET-VOL-2-ISSUE-2-387-389.pdf.

[317] Bram Van Ginneken, Shigehiko Katsuragawa, Bart M. Ter Haar Romeny, Kunio Doi, and Max A. Viergever. Automatic detection of abnormalities in chest radiographs using local texture analysis. *IEEE Transactions on Medical Imaging*, 21(2):139–149, 2002. doi:10.1109/42.993132.

[318] Maarten van Smeden, Timothy L Lash, and Rolf H H Groenwold. Reflection on modern methods: five myths about measurement error in epidemiological research. *International Journal of Epidemiology*, 49(1):338–347, 2 2020. ISSN 0300-5771. doi:10.1093/IJE/DYZ251.

[319] Gijs Van Tulder and Marleen De Bruijne. Combining Generative and Discriminative Representation Learning for Lung CT Analysis With Convolutional Restricted Boltzmann Machines. *IEEE Transactions on Medical Imaging*, 35(5):1262–1272, 5 2016. doi:10.1109/TMI.2016.2526687.

[320] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013. ISBN 978-1-4419-3160-3. URL https://link.springer.com/book/10.1007/978-1-4757-3264-1.

[321] Rama K. Vasudevan, Maxim Ziatdinov, Lukas Vlcek, and Sergei V. Kalinin. Off-the-shelf deep learning is not enough, and requires parsimony, Bayesianity, and causality. *npj Computational Materials*, 7(1):1–6, 12 2021. ISSN 20573960. doi:10.1038/s41524-020-00487-0.

[322] Busireddy Venkata, Ramana Reddy, Ummadi Janardhan Reddy, Busi Reddy, Venkata Ramana Reddy, and Bodi Eswara Reddy. Recognition of Lung Cancer Using Machine Learning Mechanisms with Fuzzy Neural Networks. *Traitement du Signal*, 2019. doi:10.18280/ts.360111.

[323] Laura E. Via, Dan Schimel, Danielle M. Weiner, Veronique Dartois, Emmanuel Dayao, Ying Cai, Young Soon Yoon, Matthew R. Dreher, Robin J. Kastenmayer, Charles M. Laymon, J. Eoin Carny, Jo Anne L. Flynn, Peter Herscovitch, and Clifton E. Barry. Infection dynamics and response to chemotherapy in a rabbit model of tuberculosis using [18F]2-fluoro-deoxy-D-glucose positron emission tomography and computed tomography. *Antimicrobial Agents and Chemotherapy*, 56(8):4391–4402, 2012. ISSN 00664804. doi:10.1128/AAC.00531-12.

[324] Laura E. Via, Danielle M. Weiner, Daniel Schimel, Philana Ling Lin, Emmanuel Dayao, Sarah L. Tankersley, Ying Cai, M. Teresa Coleman, Jaime Tomko, Praveen Paripati, Marlene Orandle., Robin J. Kastenmayer, Michael Tartakovsky, Alexander Rosenthal, Damien Portevin, Seok Yong Eum, Saher Lahouar, Sebastien Gagneux, Douglas B. Young, JoAnne L. Flynn, and Clifton E. Barry. Differential virulence and disease progression following mycobacterium tuberculosis complex infection of the common marmoset (callithrix jacchus). *Infection and Immunity*, 81(8):2909–2919, 2013. ISSN 00199567. doi:10.1128/IAI.00632-13.

[325] Robert S. Wallis, Peter Kim, Stewart Cole, Debra Hanna, Bruno B. Andrade, Markus Maeurer, Marco Schito, and Alimuddin Zumla. Tuberculosis biomarkers discovery: Developments, needs, and challenges. *The Lancet Infectious Diseases*, 13(4):362–372, 2013. ISSN 14733099. doi:10.1016/S1473-3099(13)70034-3.

[326] Shuai Wang, Bo Kang, Jinlu Ma, Xianjun Zeng, Mingming Xiao, Jia Guo, Mengjiao Cai, Jingyi Yang, Yaodong Li, Xiangfei Meng, and Bo Xu. A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19). *European Radiology 2021 31:8*, 31(8):6096–6104, 2 2021. ISSN 1432-1084. doi:10.1007/S00330-021-07715-1.

[327] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017. doi:10.1109/CVPR.2017.369.

[328] Simon K. Warfield, Kelly H. Zou, and William M. Wells. Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 7 2004. ISSN 02780062. doi:10.1109/TMI.2004.828354.

[329] Ted W Way, Lubomir M Hadjiiski, Berkman Sahiner, Heang Ping Chan, Philip N Cascade, Ella A Kazerooni, Naama Bogot, and Chuan Zhou. Computer-aided diagnosis of pulmonary nodules on CT scans: Segmentation and classification using 3D active contours. *Medical Physics*, 33(7):2323–2337, 2006. ISSN 00942405. doi:10.1118/1.2207129.

[330] Jürgen K Willmann, Nicholas van Bruggen, Ludger M Dinkelborg, and Sanjiv S Gambhir. Molecular imaging in drug development. *Nature Reviews Drug Discovery*, 7(7):591–607, 2008. ISSN 14741776. doi:10.1038/nrd2290.

[331] Ryan Wilson and Anand Devaraj. Radiomics of pulmonary nodules and lung cancer. *Translational Lung Cancer Research*, 6(1):86–91, 2017. doi:10.21037/TLCR.2017.01.04.

[332] World Health Organization. Global Tuberculosis Report 2020. Technical report, 2020. URL http://apps.who.int/bookorders.

[333] World Health Organization. Global Tuberculosis Report 2021. Technical report, 2021. URL https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2021.

[334] Yu Huan Wu, Shang Hua Gao, Jie Mei, Jun Xu, Deng Ping Fan, Rong Guo Zhang, and Ming Ming Cheng. JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation. *IEEE Transactions on Image Processing*, 30:3113–3126, 2021. doi:10.1109/TIP.2021.3058783.

[335] Weiyi Xie, Colin Jacobs, Jean Paul Charbonnier, and Bram van Ginneken. Relational Modeling for Robust and Efficient Pulmonary Lobe Segmentation in CT Scans. *IEEE transactions on medical imaging*, 39(8):2664–2675, 8 2020. ISSN 1558254X. doi:10.1109/TMI.2020.2995108.

[336] Xiaowei Xu, Xiangao Jiang, Chunlian Ma, Peng Du, Xukun Li, Shuangzhi Lv, Liang Yu, Qin Ni, Yanfei Chen, Junwei Su, Guanjing Lang, Yongtao Li, Hong Zhao, Jun Liu, Kaijin Xu, Lingxiang Ruan, Jifang Sheng, Yunqing Qiu, Wei Wu, Tingbo Liang, and Lanjuan Li. A Deep Learning System to Screen Novel Coronavirus Disease 2019 Pneumonia. *Engineering*, 6(10):1122–1129, 10 2020. ISSN 2095-8099. doi:10.1016/J.ENG.2020.04.010.

[337] Ziyue Xu, Ulas Bagci, Colleen Jonsson, Sanjay Jain, and Daniel J Mollura. Efficient Ribcage Segmentation from CT Scans Using Shape Features. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2899–2902, 2014. doi:10.1109/EMBC.2014.6944229.

[338] Ziyue Xu, Ulas Bagci, Awais Mansoor, Gabriela Kramer-Marek, Brian Luna, Andre Kubler, Bappaditya Dey, Brent Foster, Georgios Z Papadakis, Jeremy V Camp, and others. Computer-aided pulmonary image analysis in small animal models. *Medical Physics*, 42(7):3896–3910, 2015. doi:10.1118/1.4921618.

[339] Hee-Jeong Yang, Decheng Wang, Xin Wen, Danielle M. Weiner, and Laura E. Via. One Size Fits All? Not in In Vivo Modeling of Tuberculosis Chemotherapeutics. *Frontiers in Cellular and Infection Microbiology*, 0:134, 3 2021. ISSN 2235-2988. doi:10.3389/FCIMB.2021.613149.

[340] Stephen S. F. Yip and Hugo J. W. L. Aerts. Applications and limitations of radiomics. *Physics in medicine and biology*, 61(13):150–66, 2016. ISSN 1361-6560. doi:10.1088/0031-9155/61/13/R150.

[341] Douglas B Young, Hannah P Gideon, and Robert J Wilkinson. Eliminating latent tuberculosis. *Trends in Microbiology*, 17(5):183–188, 2009. doi:10.1016/j.tim.2009.02.005.

[342] Matthew D Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. 2014. URL https://arxiv.org/abs/1311.2901.

[343] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain Adaptation under Target and Conditional Shift, 5 2013. ISSN 1938-7228. URL http://proceedings.mlr.press/v28/zhang13d.html.

[344] Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Ciccarelli, Frederik Barkhof, and Daniel C. Alexander. Disentangling Human Error from the Ground Truth in Segmentation of Medical Images. 7 2020. URL http://arxiv.org/abs/2007.15963.

[345] Li Jun Zhang, He Ming Jia, and Da Peng Jiang. Sliding mode prediction control for 3D path following of an underactuated AUV. In *IFAC Proceedings Volumes*, volume 19, pages 8799–8804. IEEE, 2014. ISBN 9783902823625. doi:10.3182/20140824-6-ZA-1003.00372.

[346] Quanshi Zhang and Song-Chun Zhu. Visual Interpretability for Deep Learning: a Survey. *Frontiers of Information Technology and Electronic Engineering*, 19(1):27–39, 2 2018. URL https://arxiv.org/abs/1802.00614v2.

[347] Yang Zhang, Philip David, and Boqing Gong. Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-Octob, pages 2039–2049, 2017. ISBN 9781538610329. doi:10.1109/ICCV.2017.223.

[348] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. Bridging Theory and Algorithm for Domain Adaptation, 5 2019. ISSN 2640-3498. URL http://proceedings.mlr.press/v97/zhang19i.html.

[349] Chen Zhao, Yan Xu, Zhuo He, Jinshan Tang, Yijun Zhang, Jungang Han, Yuxin Shi, and Weihua Zhou. Lung segmentation and automatic detection of COVID-19 using radiomic features from chest CT images. *Pattern Recognition*, 119:108071, 11 2021. ISSN 0031-3203. doi:10.1016/J.PATCOG.2021.108071.

[350] Chuansheng Zheng, Xianbo Deng, Qiang Fu, Qiang Zhou, Jiapei Feng, Hui Ma, Wenyu Liu, and Xinggang Wang. Deep Learning-based Detection for COVID-19 from Chest CT using Weak Label. *medRxiv*, page 2020.03.12.20027185, 3 2020. doi:10.1101/2020.03.12.20027185.

[351] S. Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S. Duncan, Bram van Ginneken, Anant Madabhushi, Jerry L. Prince, Daniel Rueckert, and Ronald M. Summers. A Review of Deep Learning in Medical Imaging: Imaging Traits, Technology Trends, Case Studies With Progress Highlights, and Future Promises. *Proceedings of the IEEE*, pages 1–19, 2021. ISSN 0018-9219. doi:10.1109/JPROC.2021.3054390.

[352] Zhi Hua Zhou. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53, 1 2018. ISSN 2095-5138. doi:10.1093/NSR/NWX106.

[353] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2017-Octob, pages 2242–2251, 2017. ISBN 9781538610329. doi:10.1109/ICCV.2017.244.

[354] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, 109(1):43–76, 1 2021. doi:10.1109/JPROC.2020.3004555.

[355] Matteo Zignol, Mehran S. Hosseini, Abigail Wright, Catharina Lambregts–van Weezenbeek, Paul Nunn, Catherine J. Watt, Brian G. Williams, and Christopher Dye. Global Incidence of Multidrug-Resistant Tuberculosis. *The Journal of Infectious Diseases*, 194(4):479–485, 8 2006. ISSN 0022-1899. doi:10.1086/505877.

# Appendix A

# Surrogate Truth Extraction

## A.1 Selecting CT Slices With The More Uncertain Boundaries

Segmentation of lungs infected by Mycobacterium tuberculosis (Mtb) in chest computed tomography (CT) images is a complex task. Moreover, it is difficult to establish a suitable ground truth, as generation thereof is very time-consuming, subject to intra- and inter-expert variability, and prone to errors. As discussed, the commonly used measures of similarity do not represent well the unavoidable human variability inherent in a segmentation process. Therefore, in our workflow evaluation, we used a surrogate ground truth built as a consensus between three experts who performed detailed segmentations on 156 slices from our chest CT dataset. These slices were selected from the whole dataset using the procedure described in the next paragraph and were designed to ensure that the surrogate ground truth contains a representative sample of the most uncertain slices.

For the selection, we use the lung segmentation results obtained with the semi-automatic tool. This tool makes it possible to perform a simple interactive segmentation of each chest CT scan. Although the procedure is time-consuming and the results obtained are not ideal, they can be used as a reference to identify which of the lung segmentations computed with our tool have changed more with the refinement procedure. To work with reliable segmentations, we exclude slices for which the *DSC* is below 0.7. In the subset, we measure the Hausdorff distance (pre-refinement and post-refinement), using the semi-automatic lung segmentation as a reference. Finally, we select the slices for which the absolute differences between the Hausdorff distances are larger than $\mu_{\Delta(HD_{pre},HD_{post})} + 3\sigma_{\Delta(HD_{pre},HD_{post})}$ (with $\mu$ and $\sigma$ being the mean and standard deviation of the HD differences, $\Delta$). In Figure A.1, the $HD$ differences are plotted against the *DSC* for all slices with a *DSC* larger than 0.7. The threshold is drawn

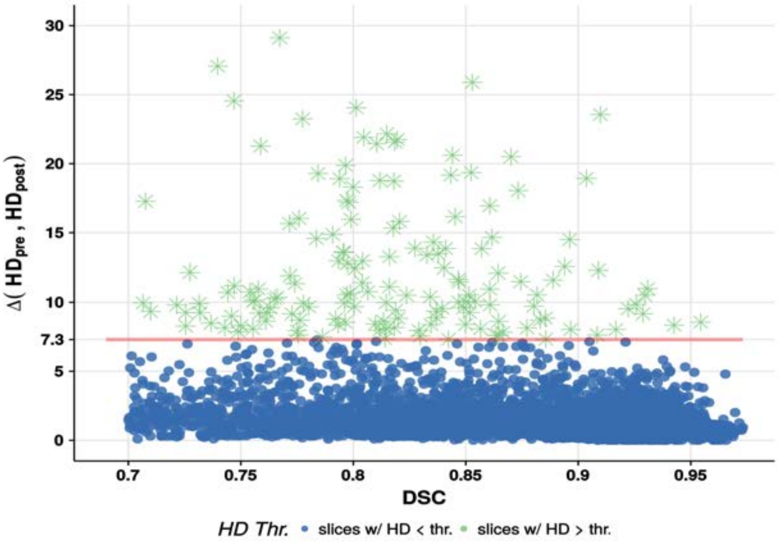


Fig. A.1 Hausdorff distance (*HD*) differences between those corresponding to the lung segmentation performed with our tool before and after the refinement process. *HD* was measured using the semi-auto segmentations as a reference. Only those slices with a Dice similarity coefficient (*DSC*) over 0.7 were included. *HD* corresponding to the slices for which the *HD* difference (Δ) is larger than a given threshold (thr., red line) are drawn in green, while points smaller than the threshold are shown in blue. In this case, the threshold is computed as $\mu_{\Delta(HD_{pre},HD_{post})} + 3\sigma_{\Delta(HD_{pre},HD_{post})}$ and the number of slices with an HD larger than the threshold is 156.

in red, the slices with an *HD* difference under the threshold are shown in blue and those above in green. As observed, the *DSCs* of the latter are uniformly distributed among all the possible *DSC* values, which is an indicator of disagreement at the surface delimitation and not at the complete filled volume.

## A.2 Inter-Expert Variability

In order to characterize the agreement between the lung segmentations performed by the experts, the intraclass correlation coefficient (*ICC*) of each similarity measure was computed. The semi-automatic segmentation was used as reference. In Table A.1, the agreement coefficients are presented. We observed excellent consistency between the experts at the surface similarity measures (Hausdorff distance (*HD*), Hausdorff distance averaged (*HAD*), as was intended, and a good correlation for the volume overlap indicators (Dice similarity coefficient (*DSC*), false-positive error (*FPE*), false-negative error (*FNE*)). Figure A.2 shows the boxplots corresponding to these results.

| Coeff. | ICC | CI (95%) | p-val. |
|---|---|---|---|
| **HD** | 0.88 | 0.84 to 0.90 | <0.001 |
| **HDA** | 0.85 | 0.79 to 0.88 | <0.001 |
| **DSC** | 0.74 | 0.66 to 0.81 | <0.001 |
| **FPE** | 0.71 | 0.27 to 0.86 | <0.001 |
| **FNE** | 0.60 | 0.26 to 0.77 | <0.001 |

Table A.1 Intra-class correlation coefficient (ICC) and 95% confidence intervals (CI) for the similarity coefficients between the three experts' delimitation and the refined masks. Note: Haussdorff distance (HD), Haussdorff distance averaged (HDA), Dice similarity coefficient (DSC), false-positive error (FPE), false-negative error (FNE).

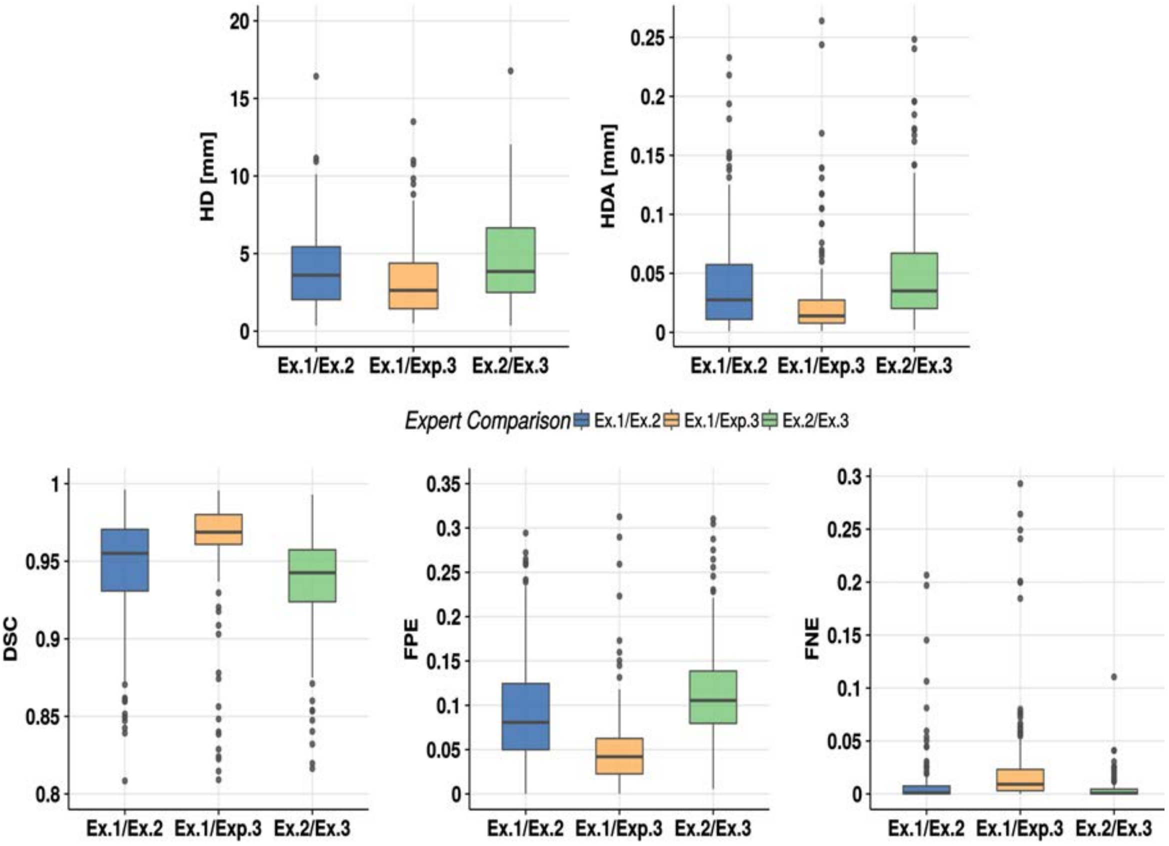


Fig. A.2 Boxplot charts for the similarity coefficients obtained from the 156 most uncertain slices in the experts' delimitations.

# Appendix B

# Texture Features Definition

1. Maximum:

$$f_1 = \max I(i,j)$$

2. Mean:

$$f_2 = \frac{1}{N+M}\sum_{i}^{N}\sum_{j}^{M} I(i,j)$$

3. Minimum:

$$f_3 = \min I(i,j)$$

4. Standard Deviation:

$$f_4 = \frac{1}{N+M}(I(i,j) - f_2)^{\frac{1}{2}}$$

5. Autocorrelation:

$$f_5 = \sum_{i}\sum_{j}(ij)p(i,j)^2$$

6. *Cluster Prominance*:

$$f_6 = \sum_{i}\sum_{j}(i+j-\mu_x-\mu_y)^4 p(i,j)$$

7. *Cluster Shade*:

$$f_7 = \sum_i \sum_j (i + j - \mu_x - \mu_y)^3 p(i,j)$$

8. Contrast:

$$f_8 = \left| \sum_i \sum_j \right|^2 p(i,j)$$

9. Correlation 1:

$$f_9 = \sum_i \sum_j \frac{(i - \mu_x)(j - \mu_y)p(i,j)}{\sigma_x \sigma_y}$$

10. Correlation 2:

$$f_{10} = \frac{\sum_{i=1} \sum_{j=1} (ij) p(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$$

11. Difference Entropy:

$$f_{11} = -\sum_{i=0}^{L-1} p_{x-y}(i) \log \left( p_{x-y}(i) \right)$$

12. Difference Variance:

$$f_{12} = \sum_{i=0}^{L-1} i^2 p_{x-y}(i)$$

13. Dissimilarity:

$$f_{13} = \left| \sum_i \sum_j \right| p(i,j)$$

14. Energy:

$$f_{14} = \sum_i \sum_j p(i,j)^2$$

15. Entropy:

$$f_{15} = -\sum_{i=1}^{L}\sum_{j=1}^{L} p(i,j)\log(p(i,j))$$

16. Homogenety 1:

$$f_{16} = \sum_{i}\sum_{j} \frac{p(i,j)}{1+|i+j|}$$

17. Homogenety 2:

$$f_{17} = \sum_{i}\sum_{j} \frac{p(i,j)}{1+|i+j|^2}$$

18. Information Measure Correlation 1:

$$f_{18} = \frac{f_9 - HXY1}{\max(HX,HY)}$$

19. Information Measure Correlation 2:

$$f_{19} = [1-\exp(-2(HXY2-f_9))]^{1/2}$$

20. Normalized Inverse Difference:

$$f_{20} = \sum_{i=1}^{L}\sum_{j=1}^{L} \frac{1}{1+|i-j|^2/L} p(i,j)$$

21. Normalized Moment Inverse Difference:

$$f_{21} = \sum_{i=1}^{L}\sum_{j=1}^{L} \frac{1}{1+(i-j)^2/L} p(i,j)$$

22. Maximum Probability:

$$f_{22} = \max_{i,j} p(i,j)$$

23. Sum Average:

$$f_{23} = \sum_{i=2}^{2L} i p_{x+y}(i)$$

24. Entropy Sum:

$$f_{24} = -\sum_{i=2}^{2L} p_{x+y}(i)\log(p_{x+y}(i))$$

25. Sum Variance:

$$f_{25} = \sum_{i=2}^{2L} (i - f_8)^2 p_{x+y}(i)$$

26. Sum of Squares:

$$f_{26} = \sum_i \sum_j (i - \nu)^2 p(i,j)$$

## Definitions

- $L$: Quantization level
- $p(i,j)$: Co-ocurrence matrix at position (i,j)
- $\nu = \frac{1}{L}\sum_i^L \sum_j^L p(i,j)$
- $p_x(i) = \sum_{j=1}^L p(i,j)$
- $p_y(j) = \sum_{i=1}^L p(i,j)$
- $p_{x+y}(k) = \sum_{i=1,i+j=k}^L \sum_{j=1}^L p(i,j), \qquad k = 2,3,...,2L$
- $p_{x-y}(k) = \sum_{i=1,|i-j|=k}^L \sum_{j=1}^L p(i,j), \qquad k = 0,1,...,L-1$
- $HX = -\sum_i p_x(i)\log(p_x(i))$
- $HY = -\sum_j p_y(j)\log(p_y(j))$
- $HXY = -\sum_i \sum_j p(i,j)\log(p(i,j))$
- $HXY1 = -\sum_i \sum_j p(i,j)\log(p_x(i)p_y(j))$
- $HXY2 = -\sum_i \sum_j p_x(i)p_y(j)\log(p_x(i)p_y(j))$