SCUOLA NORMALE SUPERIORE

Classe di Scienze

Corso di perfezionamento in
Data Science

XXXIV ciclo

Development and discussion of deep learning algorithms for breast density classification and for COVID-19 lesions quantification on CT scans

Towards a real-time multidisciplinary approach

FIS07 Fisica Applicata

Candidata
dr.ssa Francesca Lizzi

Supervisors:

Maria Evelina Fantacci
Piernicola Oliva
Mauro Capocci
Davide Bacciu
Sara Colantonio
Fosca Giannotti
Nadia Pisanti

A.A. 2021/2022

Contents

Acronyms

ACR - American College of Radiology
AI - Artificial Intelligence
ALARA - As Low As Reasonably Achievable
ANN - Artificial Neural Network
ANOVA - Analysis of Variance
AOUP - Azienda Ospedaliero-Universitaria Pisana
BB - Bounding Box
BI-RADS - Breast Imaging Reporting And Data System
BMI - Body Mass Index
CAD - Computer Aided Detection
CADx - Computer Aided Diagnosis
CC (projection) - Cranio Caudal
CNN - Convolutional Neural Network
CODP - Chronic Obstructive Pulmonary Disease
CR - Chest Radiograph
CT - Computed Tomography
CT-SS - CT Severity Score
DBT - Digital Breast Tomosynthesis
DICOM - Digital Imaging and Communications in Medicine
DL - Deep Learning
DNN - Deep Neural Network
ED - Emergency Department
EMI - Electric and Musical Industries
FBP - Filtered Back Projection
FCNN - Fully Convolutional Neural Network
FFDM - Full-Field Digital Mammography
FOV - Field Of View
GAN - Generative Adversarial Network
GE - General Electric
GGO - Ground Glass Opacification
GLM - Generalized Linear Model
grad-CAM - gradient Class Activation Map
HIPAA - Health Insurance Portability and Accountability Act
HPC - High Performance Computing
HU - Hounsfield Unit
IR - Iterative Reconstruction

IRB - Institutional Review Board
LCTSC - Lung CT Segmentation Challenge
MAE - Mean Absolute Error
ML - Machine Learning
MLO (projection) - Medio Lateral Oblique
MRI - Magnetic Resonance Imaging
MSE - Mean Squared Error
NCI - National Cancer Institute (USA)
NIfTI - Neuroimaging Informatics Technology Initiative
NLST - National Lung Screening Trial
NN - Neural Netowrk
NSCLS - Non Small Cell Lung Cancer
PACS - Picture Archiving and Communication Systems
PCA - Principal Component Analysis
PHI - Protected Health Information
PNG - Portable Network Graphics
sDSC - surface Dice Similarity Coefficient
SVM - Support Vector Machine
TCIA - The Cancer Imaging Archive
TFT - Thin Film Transistors
UAMS - University of Arkansas for Medical Sciences
vDSC - volumetric Dice Similarity Coefficient
VR (DICOM) - Values Representation
WHO - World Health Organization
XAI - Explainable Artificial Intelligence

Introduction

Machine and deep learning methods applied to medical images seem to be a promising way to improve the performance in solving many issues: the diagnosis of a specific disease, the contouring of organs or lesions, the prediction of the prognosis, they offer the possibility of analyzing many patients' data in a reproducible way and they can be applied to carry out follow up and radiomic studies. In particular, the advent of deep learning algorithms in the field of medical image analyses is leading to a change in supporting physicians in their role. Many different applications have been explored [89] successfully. However, developing an algorithm with the aim of applying it in clinical practice is a complex task which should take into account the context in which the software is developed and should be used. In the first report of the World Health Organization (WHO) about the ethics and governance of Artificial Intelligence (AI) for health published in 2021 [154], it has been stated that AI may improve health care and medicine all over the world only if ethics and human rights are a main part of its development. WHO recognizes that ethical guidance based on the shared perspectives of the different entities that develop, use or oversee such technologies is critical to build trust in these technologies, to guard against negative or erosive effects and to avoid the proliferation of contradictory guidelines. Involving ethics in technology development means to take into account several issues. First, understanding how scientific method is changing should be at least taken into account and discussed, when developing a medical software. This is directly connected to the epistemological change due to the intensive use of deep and machine learning. According to Kitchin [68] and Hey [52], in fact, epistemology is moving towards a new paradigm called the "fourth paradigm" or "exploration science". In this evolution, some fundamental rules of traditional science are deeply changing and they should be taken into account since they are useful to establish the limits and the possibilities of these new rising methods. The assumptions that are made during the development of an algorithm are critical to define the model itself and the boundaries in which it should be applied. Second, involving ethics means that AI should be built taking care of sampling population in order to prevent social and technological biases. Third, most of deep learning algorithms work in a way that is not easy to explain or interpret. Since a deep learning algorithm is usually made of many hidden layers, it is not straightforward to understand how it comes to a decision. For this reason, in recent years, the explanation of the functioning of an algorithm is a very interesting field of study which goes under the name of Explainable Artificial Intelligence (XAI) [47]. Furthermore, AI applied to medical images should also take into account the process of image production which includes manufacturers, acquisition parameters and also the interactions with physicians. When developing a deep learning based algorithm, in fact, we should always compare the results with a ground truth that defines the objective we want to reach. The ground truth on medical images is usually made by medical doctors opinions or by a consensus among them. However, it always suffers from a certain grade of variability which should be kept under control [15, 121]. The use of a peculiar imaging modality and of a specific imaging system may affect the capability of having a reliable ground truth and aggregate data from different sources is a challenge that still need to be addressed. In fact, publicly available data sets of medical images usually contain small data sets that need to be aggregated in order to obtain a set with a sufficient number of samples to train a deep learning algorithm [92]. Even if it is possible to collect private image data sets from hospitals, the process is very time consuming for both the collection and the labeling. Moreover, the publication of such data sets may not be possible reducing the chance of reproducing results obtained by other studies. Publishing the data is not easy because database maintenance is expensive and the privacy of patients has to be managed rightfully. In this context, the application of AI to medical images needs a special care since its wrong use may harm not only people but also health care systems [140]. Developing an algorithm that takes into account all the issues is a very complex task and the aim of this work is to discuss it with the support of two deep learning based algorithms developed on medical images. In Chapter 1, an overview of the X-ray imaging principles is reported. In particular, the description of X-rays production and their interaction with matter is included. Since the following two use cases are made on mammography and Computed Tomography (CT), these two specific imaging modalities are deeply described. Moreover, Chapter 1 includes an overview of how deep learning works and an introduction to the explainability problem. Understanding how medical images are produced is important to develop a deep learning algorithm. In Chapter 2, a wide discussion and description of the technical and ethical issues of medical algorithm development is discussed. This chapter includes a discussion on the changing scientific paradigm and on how medical images data are collected, labelled and published. Moreover, a discussion on the meaning of the work ”validation” is reported along with a brief dissertation on how deep learning algorithm may be included in a hospital workflow. Finally, a process to assess trustworthiness on medical algorithm, called ”Z-inspection®” [161] is presented. In the following chapters, the two algorithms I developed during my PhD are presented. The first one is included in Chapter 3 and it has been applied on mammography for the classification of breast density [93] [127]. Breast density is defined as the ratio between fibroglandular tissue and fat tissue as seen on a mammographic exam. It is an interesting patient feature because it is responsible for the masking effect, which means it may cover a malignant mass, and radiation dose depends on it. Moreover, breast density is an inherent risk factor for cancer. Since its classification is usually assessed by radiologists following qualitative guidelines according to the BI-RADS Atlas, a Convolutional Neural Network (CNN) has been developed to solve this task and the grad-CAM algorithm has been used to explain the CNN. The performances obtained in terms of accuracy, recall and precision compare well with the literature [91]. Data have been collected from the Azienda Ospedaliero-Universitaria Pisana (AOUP) and labelled by one radiologist with experience in reading mammograms and, hence, data come from a private dataset. This collection made possible to have a sufficient number of exams to train a CNN. During the collection, the sampling of the population was only based on the date of exams and this led, as a result, to a very imbalanced data set. Moreover, it contains only images and class labels, avoiding the possibility of studying important characteristics of the population, such as the ethnicity. It has been studied [68] that once a data set has been acquired, it is quite impossible to add entries to better stratify populations. The ground truth has been made by one radiologist and it has been studied that the variability in BI-RADS classes assessment is not negligible. Furthermore, the lack of a public data set that contains digital mammograms does not allow a fair comparison with other algorithms. Finally, the explanation was made with the Grad-CAM algorithm that highlights the areas of the image used to perform the classification. However, there not exists a method to quantify systematically whether the algorithm has been well explained. For this reason, I propose a simple correlation study between the pixel intensities and the activation map, since it is expected that the classifier should look at denser regions to perform the classification and denser region should be more intense in mammograms.

The second algorithm, reported in Chapter 4, has been developed on Computed Tomography (CT) scans to segment the lung parenchyma, the COVID-19 lung lesions and it returns as output also the CT Severity Score

(CT-SS), which is a classification system based on the percentage of infected lung [90]. Lung Computed Tomography (CT) is an imaging technique useful to assess the severity of COVID-19 infection in symptomatic patients and to monitor its evolution over time while the diagnosis is not possible through CTs because other forms of pneumonia may appear very similar to the COVID-19 one. Since the system outputs the infected areas on the image, the software can be used to compute radiomic features for the prediction of clinical variables. The pipeline is made of a cascade of three CNN:

1. The first module is a CNN which infers through regression 6 points of a bounding box that includes the lungs.
2. The second module is a U-net which takes as input the CT scans cropped at the bounding box and it is devoted to lung segmentation.
3. The third module is a U-net with the same architecture of the previous one which segments, instead, the lesions (ground glass opacities, consolidations and so on).

The software computes the physical volumes of the lesions and the lungs and their ratio in order to obtain the CT-SS and a simple post-processing based on watershed transformation is used to separate right and left lung. The pipeline has been trained on publicly available data sets and some of them have been collected for other purposes. Moreover, since data have been released in the NifTI format, whose header contains only spatial and orientation information, all the acquisition parameters were lost. Thus, they could not be used during data pre-processing to standardize images. In addition, no information has been published regarding the type of population sampled. The limited number of samples forced us to aggregate several data sets in order to have a sufficient number of images in the training set. Each data set has been labelled using different guidelines and label noise has an effect on the performance of the algorithm. As regard the lung segmentation, lung CTs of patients not affected by COVID-19 have been used because there not exists a dataset of COVID-19 patients labelled with the lung contours with the exception of the COVID-19-CT-Seg dataset [98] that has been kept apart for testing the software. This operation biased our algorithm and its performance in lung segmentation may not be satisfactory on COVID-19 patients. To overcome this lack, the masks of the lungs that are returned in output are joined to those of the lesions. However a data set on COVID-19 that includes lung labels could improve the segmentation performance of the software. The lesion segmentation algorithm has been trained with all the available published labelled data which contain mostly mild cases of COVID-19 pneumonia. As a result, the algorithm systematically underestimates the injured areas. This work has been then sent to several Italian hospitals and we are carrying out a study on the agreements between physicians and the software in order to validate the algorithm.

Throughout the critical analysis of these two algorithms, it is possible to underline how theoretical issues affect the reliability, the performance and the fairness in practice and, since they concern many different domains of knowledge, the thesis that it is necessary to involve many expertises including physicists, physicians, lawyers, sociologists, computer scientists, computer engineers and so on, is discussed. Taking into account the issues that concern the application of algorithms in clinical practice is a challenging task which should not be considered only a posteriori. Moreover, it deals also with how institutions that manage health and research interact among each other. Designing an experiment that considers as much as possible the issues presented above is necessary for producing a Data Science which is more impacting, fair, reliable and with high performance.

Chapter 1

Applications of deep learning methods to medical images

The application of deep learning techniques to medical images seems to be a very promising way to improve diagnosis performance. In order to correctly use these methods, it is important to understand what a medical image is and which deep learning methods may be more appropriate than others. For this reason, in Chapter 1, the physical principle of X-Ray imaging, including X-Rays production, their interaction with matter and the principles of detecting them, is firstly presented (Sections 1.1, 1.2 and 1.3). Since the use cases of this work concern mammography and Computed Tomography, in Sections 1.4 and 1.5 these two imaging modalities are presented. In Section 1.6 and 1.7, Convolutional Neural Networks (CNN), the NN used in the use cases, are presented along with a brief literature review of their application on medical images. The specific literature for each use case is presented in later chapters. Finally, the "black box" problem is presented in the last section with a review of explanation methods for CNNs.

1.1 Where does medical imaging come from?

Even if the historical origins of medical physics are traced from the first use of weighing as a means of monitoring health by Sanctorius in the early seventeenth century, the first appearance of the term "Medical Physics" dates back to 1778 in Paris. It was intended as the study of physical principles applied to medicine. Even if the first traces of experimenting physics in medicine goes back to ancient Egypt, in the 18^{th} century, this new discipline began its systematical foundation [31]. Medical imaging is a branch of medical physics which aims to represent the human body using different physical principles. Its beginning can be dated back to the X-ray discovery made by Wilhelm Conrad Röntgen in November 1895 and his finding was the first moment, in human history, in which looking into a living human body was possible. The use of X-ray and radiograph was soon replicated in many physics laboratories in Europe and America [128] and it was widely used during the First World War. Marie Curie, who won the Nobel prize few years earlier for her research on radiation, drove a truck equipped with a portable X-ray machine through the French battlefront, allowing not only a more precise diagnosis of broken bones but also of the effects of gas gangrene [128]. In the very next years, X-ray radiographs were used also to image the lungs in order to study and detect tubercolosis. However, the medium and long term effects of radiation was quantified only few years later, in the late 1940s. Nowadays, many imaging modalities are used in hospitals to diagnose, monitor and screen patients. Each of them exploits a physical principle to represent the human body. X-ray imaging uses the same principle discovered by Röntgen and applied by Marie Curie; Magnetic Resonance Imaging (MRI), which was invented in 1971 by Paul Lauterbur, exploits the possibility to orient the nuclear spins of human body and to reconstruct a 3D image using the back-propagation algorithm; ultrasound imaging, instead, is made with sound waves that are sent to the patient through a transducer and can return morphology information measuring their echos; nuclear imaging exploits the metabolic and chemical behaviour of human body to image the functionality of organ through the use of radiopharmaceutical, which contains radioactive isotopes. Many applications have been developed in the last 50 years and one of the most interesting frontier research field is the multimodality imaging, which searches how to build imaging devices that allows to acquire a body representation using more than one imaging technique mentioned above.

1.2 Physical principles of X-Ray imaging

In this thesis, two main case studies are presented: the first one is related to mammography while the second one is related to Computed Tomography (CT). For this reason, the explanation of this two imaging modalities will be analyzed more in depth. Both mammography and computed tomography are based on the use of X-rays but their applications and their scopes are different.

1.2.1 X-rays production and detection

X-ray is an electromagnetic radiation whose energy is between 124 eV and 124 keV [26]. In this energy range, which is the diagnostic one, X-rays interact with human tissues in two main ways: the photoelectric effect and the Compton scattering. There are other kind of interactions that occurs which are not relevant in this exposition. Every human tissue and organ interacts differently with X-rays and their differential absorption of X-rays, due to their atomic composition, is the basis of the image production. The principles used to produce X-rays are mostly not changed over time but X-rays tube has been refined to achieve the required performance for imaging. The production of X-rays is made by the bombardment of a thick target with energetic electrons. When the electrons reach the target materials, collision and scattering processes happen and, as a result, we have X-ray production. Its spectrum is essentially due to the bremsstrahlung radiation and to the characteristic radiation. When an electron goes across matter, it is slowed down. In particular, when it is enough close to an atomic nucleus, it interacts through the Coulomb force and changes its trajectory. An electron that changes trajectory emits electromagnetic radiation called *bremsstrahlung*. The energy of the emitted radiation is subtracted from the kinetic energy of the incident electron and the energy of the emitted photon depends on the Coulomb forces, hence, it depends on the distance between the electron and the nucleus. Classically, an electron colliding with a thin target yields a constant energy fluence from zero up to the initial kinetic energy of the electron. As a result, its spectrum is a simple rectangular function. A thick target can be seen, classically, as a stack of several layers of thin targets. As the electron is slowed down through target layers, it loses its energy until it reaches the rest state. The spectrum of such interactions results in a stack of rectangular functions which can be represented as a triangular function. The classical theory does not take into account the attenuation processes and, moreover, quantum mechanics tells that the energy spectrum for an electron crossing a thin layer is not rectangular. For these reasons, the bremsstrahlung spectrum represented as a triangular function is an ideal classical representation of the process. The other main process that contributes to X-ray spectrum is the characteristic radiation. A fast electron that collides with a shell electron could knock out the shell electron if the kinetic energy of the fast one is greater than the binding energy of the shell electron. When the shell electron is knocked out, it leaves a vacancy which is filled with an electron coming from an outer shell. This transition is accompanied along with X-ray emission. The energy of the emitted radiation depends on the binding energies of the involved shells. Binding energy is greater in the most inner shell (K) and it decreases in outer shell (L,M and so on). Moreover, binding energies are characteristic and unique for each element and for this reason the emitted radiation is called *characteristic radiation.* The sum of the radiation produced in this two processes composes the X-ray spectrum (Figure 1.1).

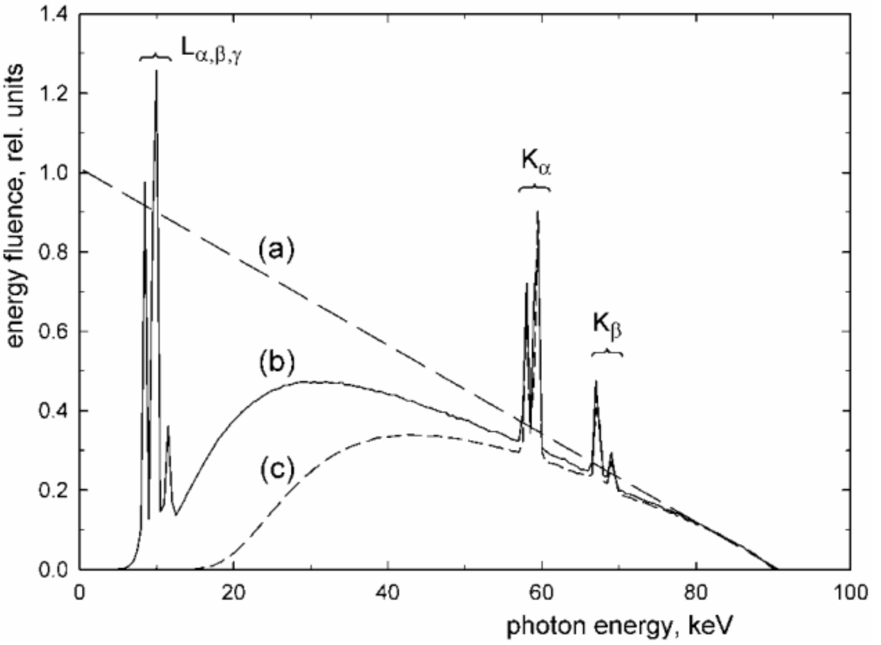

Figure 1.1: (a) Ideal bremsstrahlung spectrum for a tungsten anode (tube voltage 90 kV), (b) an actual spectrum at the beam exit port, including characteristic X rays (anode angle 20°, inherent filtration 1 mm Be) and (c) the spectrum filtered with an equivalent of 2.5 mm Al. This image has been taken from [26]

X-rays are not generated in the surface of the material but within it; for this reason the X-ray beam is also attenuated by the material itself (selfabsorption), especially at low energies. The energy fluence depends on the atomic number of the target, on the current and the square of the potential difference of the X-ray tube. Hence, having higher bremsstrahlung energy requires the use of target materials with higher atomic number (Z). X-rays are usually produced with a X-ray tube. A simplified version of a X-ray tube is made by several elements. First, we have the current supply used to produce energetic electrons through the thermionic emission on a filament, which constitutes the cathode of the tube. Then one of the fundamental element for the X-ray spectrum production is the target on which the energetic electrons collide and it is the anode of the tube. The choice of the anode defines important characteristics of the spectrum and it depends on the radiographic application. Moreover the process of X-ray production does not have high efficiency and, hence, the anode material should also have good thermal properties. The optimal choice in common radiology for anode material is the tungsten (W, Z=74). In mammography, the choice for anode materials may be different because mammography is an imaging technique which inspects a particular soft tissue. For this reason, the energy of the X-ray spectrum required for it is usually lower than any other diagnostic exam. In mammography, the X-ray radiation produced by characteristic emission is higher than the bremsstrahlung one, ensuring good image quality and low dose delivery. Typical anode materials for mammography are molybdenum (Mo, Z=42) and rhodium (Rh, Z=45) but it is possible to find mammographic systems with tungsten anode too. Finally, another fundamental part in X-ray production is the use of filtration. In fact, lower energy X-rays contribute to dose delivering without any improvement to image formation at all. For this reason, filters are usually used at the exit of the tube in order to select the X-rays for the specific imaging task. Another pivotal part of the imaging is the X-ray detection. X-ray detectors can be divided into three categories: film based, indirect digital and direct digital detectors. In the last few years, film detectors have been replaced with digital radiography for many reasons: digital radiography gives advantages of immediate image preview and availability, deletes the cost of film processing, guarantees a wider dynamic range and allows to apply special image processing techniques that enhance overall display quality of the image. Digital radiography is usually made with flat panel detectors, which can be direct or indirect. In the former case, the X-rays interact with a scintillator, which converts them into visible light, and then light is converted into electrons by a system made of amorphous silicon photodiodes and read by Thin Film Transistors (TFTs). One of the best material for the scintillator is the Thallium activated Caesium iodide because it has a good quantum efficiency and its crystal structure ensures good spatial resolution. In the latter case, the direct detectors are made of amorphous Selenium which directly converts X-rays into charges that are subsequently read by TFTs. The amorphous Selenium is the material used in commercial imaging systems since it has good qualities such has a good quantum efficiency.

1.3 Matter X-ray interactions

In order to understand how it is possible to create an image with X-rays, it is important to discuss how this radiation interacts with the matter. In radiology, the range of X-rays energy goes from about 10 keV to about 150 keV [26]. Since the wavelength of the highest energy usually used in radiology is comparable to the atom radius, the interactions occur between the electromagnetic radiation and the electrons at atomic scale. Each photon interaction is expressed in terms of cross sections and attenuation coefficients, as concerns the passage through a medium. The interactions between the X-ray radiation and matter are the photoelectric effect, the Rayleigh scattering and the Compton scattering. In the photoelectric effect, when the photons hit an atom, they may cause the emission of an electron if they have an energy greater than the electrons binding energy. The emitted electron is called photoelectron and its kinetic energy is equal to:

$$T = h\nu - E_s \tag{1.1}$$

where h is the Planck constant, ν is the photon frequency and E$_s$ is the electron binding energy. This happens only if the energy of the incident photon exceeds the binding energy of the electron in that shell. The probability of the interaction is difficult to be calculated and it implies the use of quantum mechanics. However, in the diagnostic energy range, the photoelectric cross section per atom can be written as:

$$\tau = \frac{Z^4}{(h\nu)^3} \tag{1.2}$$

where Z is the atomic number. The left vacancy is then usually filled by an electron of an higher shell. There are two scattering processes at this energy that are Rayleigh, coherent and Compton, incoherent. In Rayleigh scattering, a photon collides on a bounded electron and there is no energy transfer between particles involved in the interaction. Rayleigh scattering has a low probability in the diagnostic range. However, as any scattering process, it degrades the image quality since the scattered photons may be revealed on the detector in a wrong position. Compton scattering is the predominant type of interaction in the lower energy diagnostic range. It is an incoherent scattering process and, hence, there is transfer of energy between interacting particles. The photon that hits on a shell electron changes its direction and energy and provokes the emission of the electron, leaving the atom in an excited state. The ejected electron loses its kinetic energy through the ionization of surrounding tissues, contributing to the radiation dose. For the Compton scattering, the cross-section is proportional to $\frac{Z}{E}$ where E is the photon energy. Scattering is a degradation source for medical images because scattered photons deviate their path and impress the detector in a false position. The X-Ray attenuation depends on the beam energy: the photoelectric effect is predominant at low energy in the diagnostic range, while the Compton effect is predominant in the intermediate energy range.

The interaction processes described above are useful to understand the physical processes behind the image production. Anyway, also the macroscopic effects due to photons crossing matters are important. Linear attenuation coefficient is a coefficient which gives information about the primary photons and material interactions. If we consider a thin slab of material of thickness dx irradiated normally by a beam of photon, the radiation may be absorbed, scattered or may pass without interacting. The probability that a photon interacts with the slab is given by:

$$N_a \sigma dx \tag{1.3}$$

where N_a is the number of interaction centers per unit volume and σ is the total cross-section per atom. The quantity $N_a\sigma$ is called linear attenuation coefficient and is usually denoted by μ. If we consider a thick slab of a certain material and the fluence $\Phi(x)$ of non interacting photons, the expected change in the fluence, $d\Phi$, after the passage through the medium is:

$$d\Phi = -\Phi\mu dx \tag{1.4}$$

The integration of the above equation brings to:

$$\Phi = \Phi_0 e^{-\mu x} \tag{1.5}$$

This equation describes the attenuation of the photon beam and it is known as Lambert-Beer law. If we consider a homogeneous material of thickness t and attenuation coefficient μ_1, crossed by a monochromatic radiation , which contains an insert of thickness x and attenuation coefficient μ_2, the difference in intensity can be written as:

$$\Delta I = I_0 e^{-\mu_1 t}(1 - e^{-(\mu_2 - \mu_1)x}) \tag{1.6}$$

1.4 Mammography

Breast cancer is the most diagnosed women cancer worldwide and the second cause of women death for oncological disease [135]. Mammography is a radiographic procedure optimized for breast examination, performed with X-rays of appropriate energy and which measures the X-rays attenuation through breast tissues [26]. Breast cancer signs, that should be represented and visible on a mammogram, are:

- morphology of the tumor mass, which includes irregular margins or spiculation (Figure 1.2, left);
- mineral deposits of calcium hydroxyapatite or phosphate, which can be seen as little grains called microcalcifications (Figure 1.2, right);
- architectural distortion of the normal breast pattern, which can be seen as straight lines radiating from a central area and retraction or bulging of a contour (Figure 1.3, left);
- asymmetry in corresponding regions of the left and right breast (Figure 1.3, right).

To better visualize such signs, a mammogram has to show a high contrast between breast structures and background and contrast is generated by differences between attenuation coefficients among different tissues. In Figure 1.4, X-ray attenuation coefficients over energy are shown for the three main tissues in the breast: adipose tissue, fibroglandular tissue and infiltrating ductal carcinoma [59].

As energy increases, differences in attenuation between breast tissues decrease. Furthermore, as shown in Figure 1.4, the attenuation coefficients

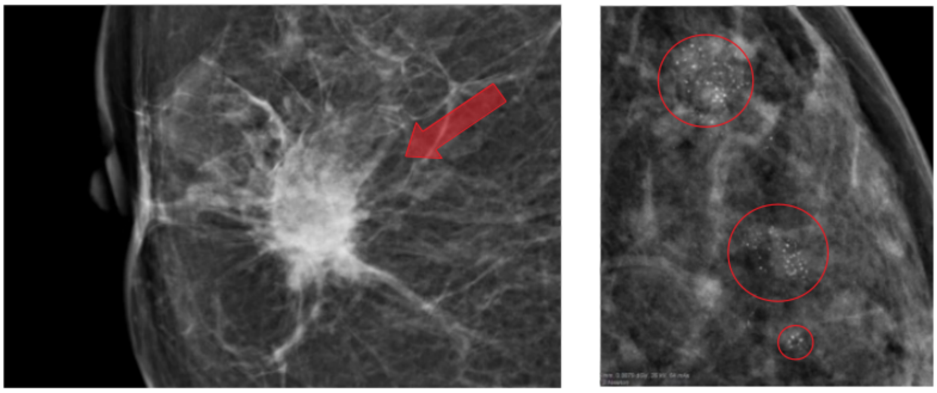

Figure 1.2: On the left, a hyperdense mass with an irregular shape and a spiculated margin. It has been proved to be an invasive ductal carcinoma. On the right, microcalcification clusters which has been proved to be multifocal DCIS (Ductal Carcinoma In Situ) with areas of invasive carcinoma

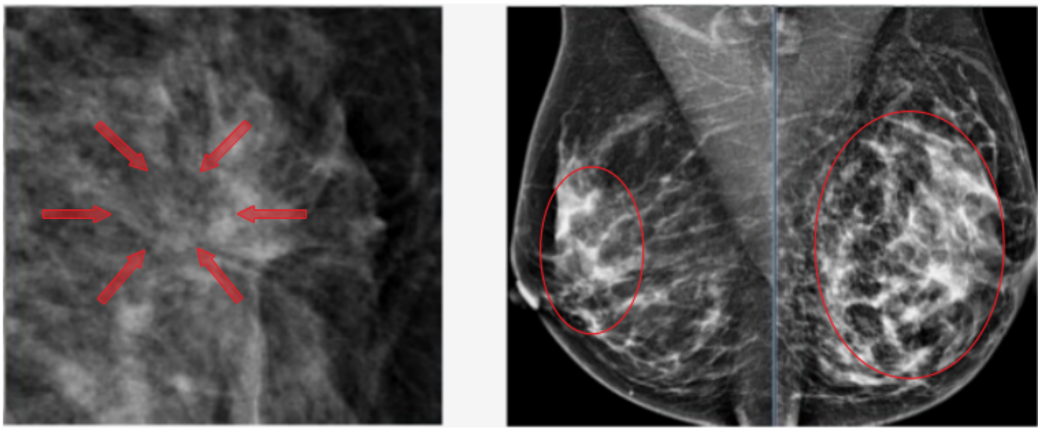

Figure 1.3: On the left, an example of an architectural distortion. On the right, an asymmetrical distortion between left and right breast that has been proved to be adenocarcinoma.

of fibroglandular tissue and cancer tissue are very similar. This similarity makes cancer detection not easy. In order to have a sufficient diagnostic power, mammography needs high spatial resolution, especially to visualize margins of masses. In fact, the irregularities on the edges of masses are in the order of magnitude of 50 μm. Furthermore, breast tissue is radiosensitive.

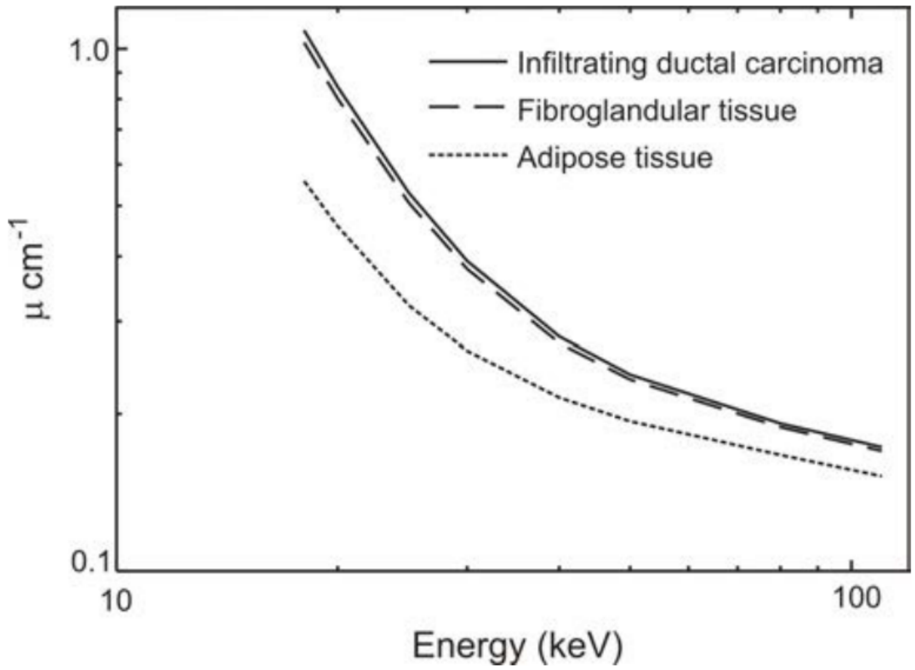

Figure 1.4: Attenuation coefficient versus X-ray energy. It can be noticed that the difference between adipose and fibroglandular tissues is remarkable while the difference in attenuation coefficients between fibroglandular tissue and cancer tissue is not so significant.

For this reason, in an optimal mammographic examination, in particular in screening programs, dose delivering should be kept as low as possible, maintaining a high diagnostic quality of the image. The amount of fibroglandular tissue, or dense tissue, with respect fat tissue as seen on a mammographic exam is called breast density. Since the attenuation coefficient of dense tissue is similar to cancer one, the sensitivity of mammography depends on the breast density. In fact, a malignant mass that is located behind dense tissue may be not detected during the examination: this effect is called "masking effect". Moreover, breast density is an inherent risk factor in developing the disease [102]. The assessment of breast density is made by the radiologist who reads the exam and this measurement suffers from high inter-observer variability. For this reason, the development of an automated method to measure breast density is highly desirable.

X-rays for Mammography

As said above, the X-ray tube is specifically designed for the mammography task and its combination with filters produces the required energy spectrum [97]. Mammographic X-ray tubes are made with rotating anodes in order to have good thermal property of the system. The rotation, in fact, reduces the accumulation of heat of the anode. The most commonly used materials for anode are molybdenum (Mo, Z=42), rhodium (Rh, Z=45) and tungsten (W, Z=74). The choice of these materials is due to their spectra. The spectra are mainly made of bremsstrahlung radiation and characteristic X-rays specific to the target materials. Characteristic X-rays are particularly important in mammography. In fact, characteristic radiation energy is 17.5 and 19.6 keV for molybdenum and 20.2 and 22.7 keV for rhodium. These energies are the required ones to produce the right image contrast for discriminating cancer and normal tissues in mammography. In Figure 1.5, molybdenum spectrum at 25 kVp and 1 mGy of final air kerma is reported.

The low energy bremsstrahlung X-rays deliver a high dose amount with little contribution to the diagnostic power of the image. Furthermore, high energy bremsstrahlung X-rays make subject contrast decrease. For these reasons, filters are used on X-rays to reduce the low and high bremsstrahlung photons. In Figure 1.6, molybdenum spectrum with a 30 μm molybdenum filter is reported.

These filters are often made with the same material of anode, i.e. molybdenum and rhodium because they stop undesired X-rays and transmit characteristic X-rays. As showed in Figure 1.6, molybdenum filter attenuates both X-rays in the low energy range and those above its own K-absorption edge, while the characteristic X-rays pass through the filter with high efficiency. Since atomic number of rhodium is higher than molybdenum one, its spectrum is harder. Thus rhodium anodes offer advantages for thicker and denser breast.

1.4.1 Mammographic systems and standard projections

Full-Field Digital Mammography (FFDM) is the widely accepted methods to perform screening programs. In digital mammography, X-rays are captured on a designed digital detector that converts them in an electronic signal. The

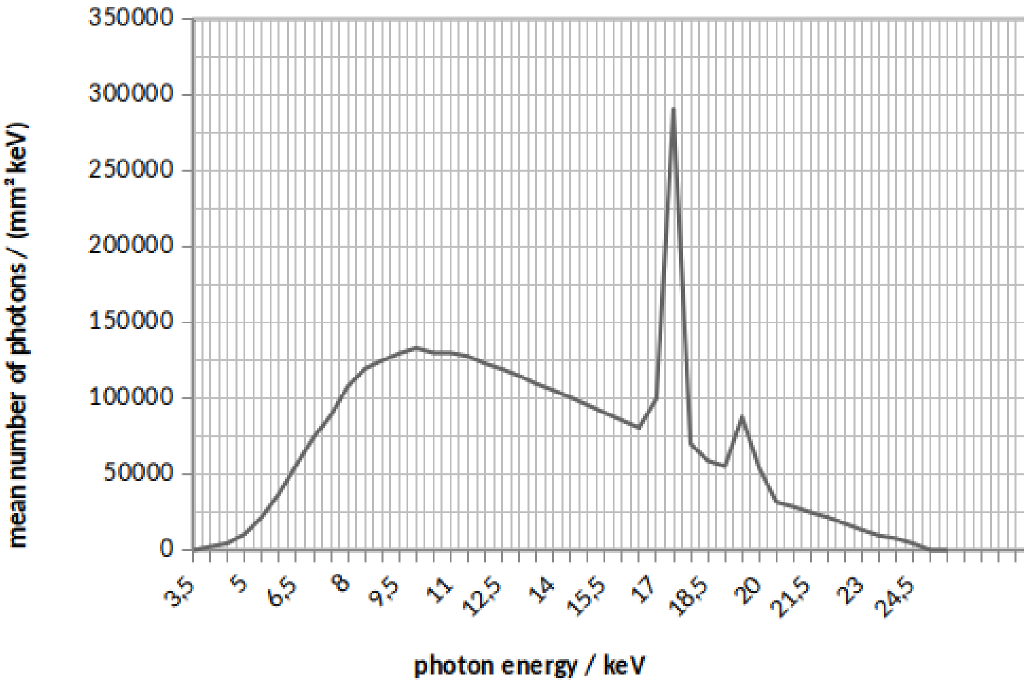

Figure 1.5: Molybdenum spectrum at 25 kVp and 1 mGy of final air kerma obtained with a simulation on https://health.siemens.com/booneweb/index.html

digital image can be visualized on a high resolution monitor and the physician can use tools to manipulate it. In this work, all the imaging systems used are digital. The system is mainly made of the X-ray tube, a compression plate, a support for the breast, an anti-scattering grid and the detector. The goal of mammography is to achieve the image quality required for a given detection task, while keeping the absorbed dose As Low As Reasonably Achievable (ALARA principle). To achieve this goal, the mammographic unit is specifically designed for the examination of breast tissues. The patient can be examined standing or sitting with her breast resting on a support plate. The X-ray tube and support plate are built on a support which can rotate in order to achieve different projection angles. The two standard projections, shown in Figure 1.7, are the craniocaudal and the mediolateral oblique. A mammographic exam is made, when possible, of the four standard projections.

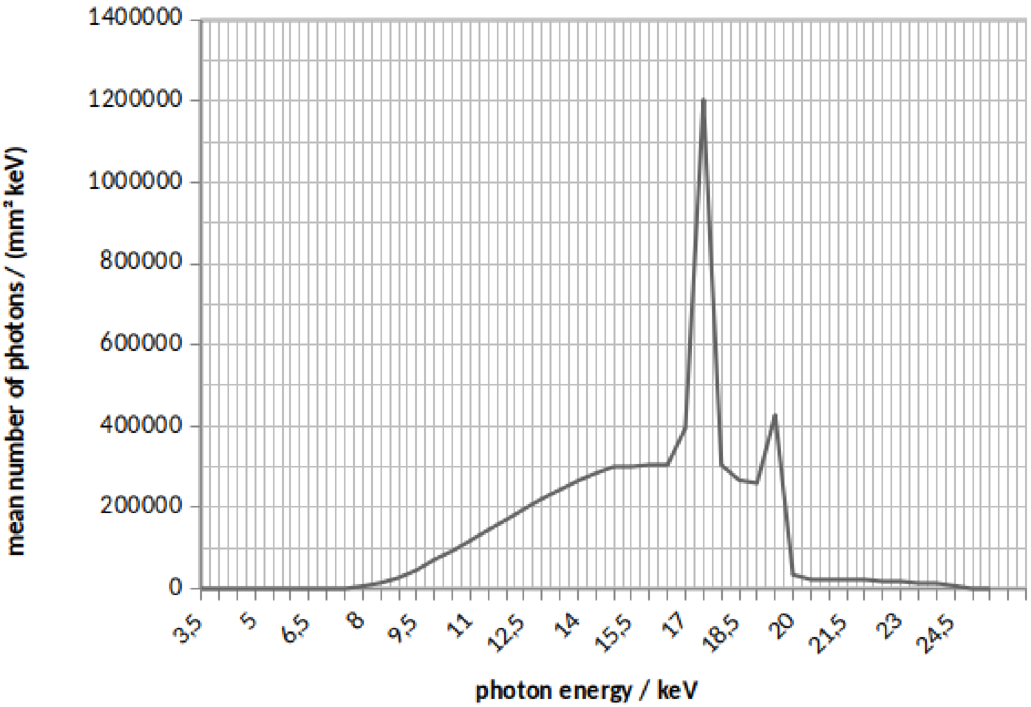

Figure 1.6: Molybdenum spectrum with a 30 μm molybdenum filter obtained with a simulation on https://health.siemens.com/booneweb/index.html

An anti-scatter grid is placed between the breast support and the image receptor. It allows to lower scattering effects on images. A compression is applied to breast using a plastic compression plate. Thanks to compression, overlapping of structures and motion artifacts are minimized.

1.5 Computed Tomography (CT)

1.5.1 What have the Beatles got to do with CT?

Computer Tomography (CT) is a 3D whole body imaging modality for a wide range of clinical applications. It has been invented in the early '70 and it is commonly believed that the revenues from the selling of the Beatles' records allow Electric and Musical Industries (EMI) to develop the CT scanner. EMI, starting from the end of the Second World War, was a company with

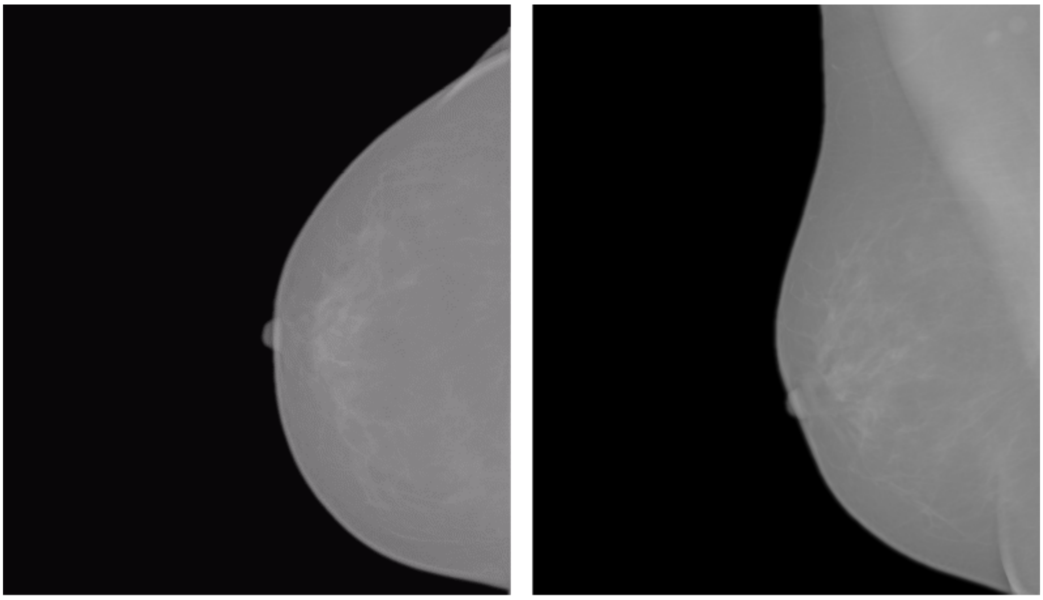

Figure 1.7: On the left: the cranio-caudal projection which is an up-down view of the breast. On the right: the mediolateral oblique projection which is a lateral view of breast. A mammographic exam is made, when possible, of the four standard projections.

experience in electronics and tried to become a leading computing company in Britain [100]. In 1955, they acquired the Capitol Records in the United States and the success of their recordings, including the Beatles one, put the company in a very strong financial position. In 1963, Allan Cormack, a South African physicist published on the *Journal of Applied Physics* [102], a paper with the theoretical solution of the problem of representing an object through its line integrals with radiological applications, which would have been the basis for CT image reconstruction. In 1967 Godfrey Hounsfield, who was an EMI-CRL researcher at that time, unaware of Cormack's work, conceived the idea of a reverse-radar [141] and he thought that one of the most promising field to apply its invention was the radiology one. In this context, the first new CT scanner was proposed in 1968 by a team made by the aforementioned Hounsfield, Stephen Bates (programming), Peter Langstone (electronics), and Mel King (mechanics). The first version was made of a translating and rotating gamma-ray source, Americium 95, around bottles or Perspex jars with a photon counter as detector, placed on the other side.

It needed 9 days to collect sufficient information of 28,000 measurements and 2.5 h to reconstruct the image. Later in the same year, the gamma-ray source was removed and replaced with a X-ray tube, reducing the acquisition time from 9 days to 9 hours. At the end of 1969, the first prototype of the scanner was built. However the analysis of market and the evolution of the EMI, the lack of expertise in medical tools and in medical electronics market and the high cost of such machinery suggests Hounsfield and EMI to ask for the assistance of the British Department of Health and Social Security (DHSS). They received a founding equal to 600000 Pounds to develop 4 scanners and, thanks to this support, the manufacturing of the first scanner by EMI was presented in 1972. So, considering this founding and the research costs, including the team salary, it can be claimed that most of the CT development costs have been paid by the British Government. Complex processes led to the creation of CT scanner, involving not only the EMI company (and the Beatles' success), but also classical academic research and public funding. Moreover, the connection between the Beatles and CT is usually described as the gift the group made to medicine, while there is not any evidence that the group was even involved in the process. Despite the true history, there is a positive aspect of this misconception: it helped to keep in memory the name of the company which developed the CT scanner. Hounsfield and Cormack received the Nobel prize for medicine in 1979.

1.5.2 Principles of CT

CT is used in practical clinic to diagnose, to monitor, to follow-up patients and also to plan radiotherapy treatments. It measures the X-ray transmission profile through a patient. The profile is made by using a X-ray tube and a detector arc made of about 800-900 detection elements. The arc, rotating around the patient, is able to acquire the X-ray transmission at different angles. Moreover, the X-ray detector system slides through the patient in order to acquire different portion of the body. Hence, a CT scanner is able to acquire a large number of views by rotating and sliding the X-ray detection system. The acquired views are then used to reconstruct the 3D image. CT is a digital imaging modality and it assigns to each pixel of the image values that are associated with the attenuation of the corresponding tissue. The physical law that drives the attenuation for a monoenergetic radiation is the

Lambert-Beer law (Equation 1.7):

$$I(x) = I_0 e^{-\mu x} \tag{1.7}$$

where I(x) is the intensity of attenuated X-ray, I_0 is the initial X-ray intensity, μ is called *attenuation coefficient* (m^{-1}) and x is the space crossed by the X-rays. The Lambert-Beer law does not take into account the fact that X-ray radiation is not monoenergetic but it produces a spectrum. However, in CT reconstruction, the average energy of the spectrum is considered and this simplification may lead to inaccuracies in reconstruction and to hardening artefacts. As an X-ray beam is sent to a patient, it goes through several types of tissue. For example, if we want to image the lungs, the X-ray beam will encounter the skin, the soft tissues of the chest, the bones of the rib cage, the air in the lungs and so on. If the path crossed by X-ray goes from 0 to a distance d, the Lambert-Beer law can be written as (Equation 1.8):

$$I(d) = I_0 e^{-\int_0^d \mu(x)dx} \tag{1.8}$$

We can represent a patient as a matrix of different attenuation coefficient as reported in Figure 1.8.

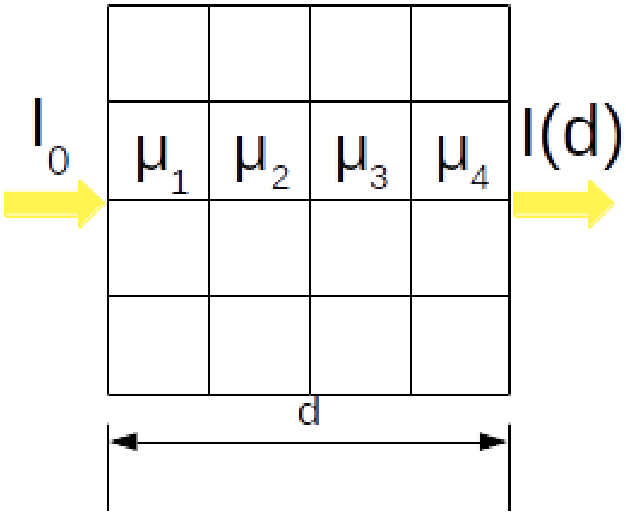

Figure 1.8: Simple representation of a CT system with a matrix of four different attenuation coefficients.

We can hence apply to the system in Figure 1.8 the discretized Lambert-Beer law as shown in Equation 1.9:

$$I(d) = I_0 e^{-\sum_{i=1}^{i=4} \mu_i \Delta x} \tag{1.9}$$

The attenuation coefficients are then translated in a corresponding matrix, which is the image matrix, in the so-called Hounsfield Units (HU). The Hounsfield Unit scale is a linear transformation of the attenuation coefficients relative to attenuation coefficient of water (μ_{water}) at standard pressure and temperature and it is defined in Equation 1.10:

$$HU_{material} = \frac{\mu_{material} - \mu_{water}}{\mu_{water}} \cdot 1000 \tag{1.10}$$

From the HU definition, it is clear that the values that a single CT voxel may have is a relative quantity. Moreover, since the attenuation coefficients depend on the energy of the spectrum, the voxel value depends not only on the specific imaged material but also on the X-ray tube voltage. In fact, as a function of the photon energy, different substances show a non linear behaviour of their linear attenuation coefficients relative to water. This effect is most present as the bigger is the atomic number of the imaged materials. It is so more effective, for example, in contrast medium imaging. The minimum bit depth that should be assigned to a pixel is 12 because it allows to have a Hounsfield unit range from -1024 to +3073. This range includes all the possible relevant clinical values. In Table 1.1 the typical values in Hounsfield Units for different human tissues are reported. Hence, once the CT has been acquired and reconstructed, a windowing is applied to the image depending on the part of the body we want to visualize. In Figure 1.9, the effect of the windowing has been reported for three different windows on an image reconstructed with a lung filter.

1.5.3 3D Image Reconstruction

The CT image reconstruction is an ill-posed inverse problem because image model is not invertible and there is not a unique solution. In order to reconstruct a 3D CT scan, several measurements are required. The collected information is the basis for the process of the reconstruction. The most used algorithm is the Filtered Back Projection (FBP) and to explain the process we need to introduce three interrelated domains: the object space, which is

Table 1.1: In this Table the typical reference values for HU ranges for different tissues or substances are reported. The actual HU units depends on temperature, tube voltage and composition of the imaged material.

Substance/Tissue	HU center (range)
Bone	+1000 (+300,+2500)
Liver	+60 (+50, +70)
Blood	+55 (+50, +60)
Kidneys	+30 (+20, +40)
Muscle	+25 (+10, +40)
Brain, grey matter	+35 (+30, +40)
Brain, white matter	+25 (+20, +30)
Water	0
Fat	-90 (-100, -80)
Lung	-750 (-950, -600)
Air	-1000

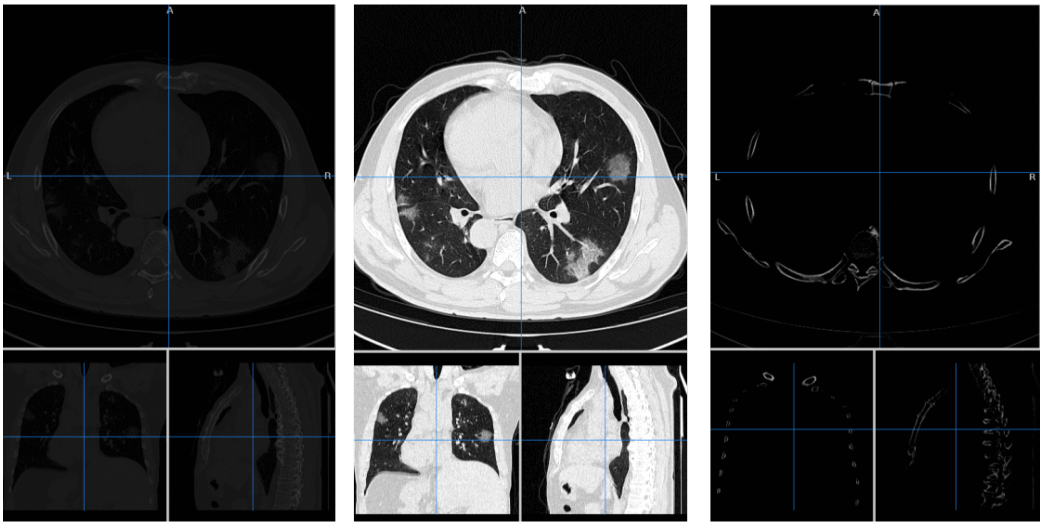

Figure 1.9: The image is the coronacases002.nii taken from the public dataset COVID-19-CT-Seg dataset [99]. Left: the CT without any windowing. The HU range goes from -1023 to 9567. Center: the lung windowing is applied in a range from -1023 to +150. Right: the same image is reported with a bone windowing from +300 to +2500.

made of the linear attenuation coefficients, the Radon Space, which is the projection space and it is also called sinogram space if it is reported in cartesian coordinates, and the Fourier space, which can be computed with a 2D Fourier transform of the space object. What we want to obtain is the object space, while the acquisition is in the projection space, i.e. the Radon space. The necessary steps to pass from the 2D Radon space to the object space are:

1. a Fourier Transform is applied to raw data in the Radon space, resulting in many 1D Fourier Transforms;
2. a high pass filter is then applied to the 1D Fourier Transforms;
3. an inverse Fourier Transform is applied to the filtered data;
4. The back projection is computed in order to obtain the image in the object space.

In Figure 1.10 the process of image reconstruction is shown.

Filtering the 1D Fourier Transforms is necessary in order to avoid artifacts and it is possible to use different filters, depending on the required quality characteristics. The filter, also called convolutional kernel, that yields to the theoretical optimal reconstruction is the Ramachandran–Lakshminarayanan filter, also called the Ram–Lak or ramp filter. However, it yields also to high noise level in the reconstructed images and in clinical practice this filter is usually used for bone reconstruction and it is called sharp filter. Sometimes it is necessary to use filters which roll off at higher frequencies. The so called normal filter or Shepp-Logan filter achieves this characteristic and images results less noisy and with a better low contrast resolution. On the other hand, it decreases spatial resolution. This filter is usually used for soft tissues. It is also possible to reconstruct the same acquisition using different filters in order to have the possibility of analyzing both soft tissues and bones.

Another possibility to reconstruct images is to use Iterative Reconstruction (IR) and this method is now commonly used in CT. The currently available IR algorithms are mostly considered proprietary and they are only partially revealed [41]. In general, the image reconstruction problem can be posed as:

$$p = Hf + n \tag{1.11}$$

where p is the acquisition, i.e. projections, f are the real data (attenuation coefficients), H is the projection process and n is the noise. The solution of

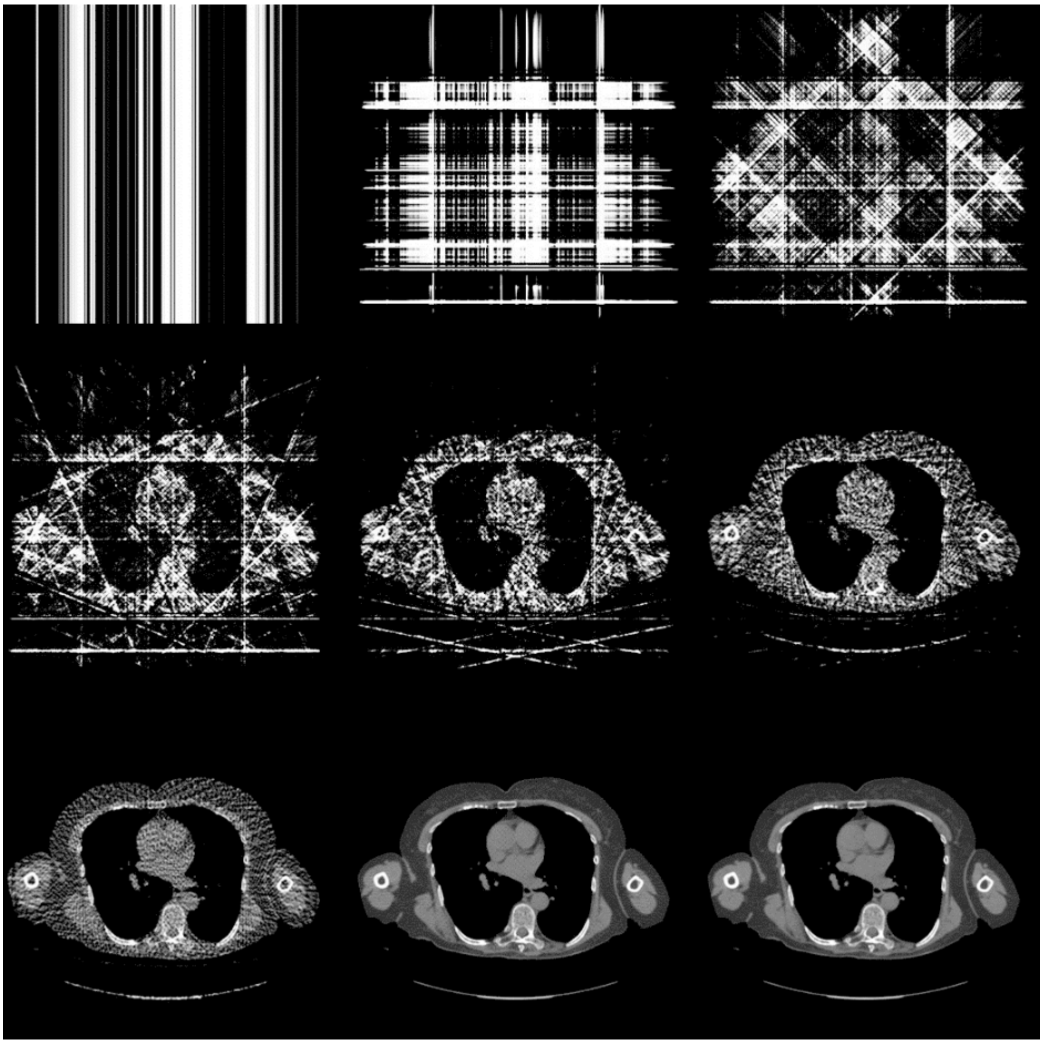

Figure 1.10: This image has been take from [26]. It represents the successive filtered reconstruction made of 1, 2, 4, 8, 16, 32, 64, 256, 512 and 1024 different acquisition angles.

this equation can be find in two main ways: algebraic methods and statistical methods. The principles of IR is mainly made of 6 steps:

1. the projections are acquired;
2. a first image estimate is generated from the projections;

3. a X-ray beam is simulated via forward projection in order to have simulated projection data that are compared to the measured one;

4. in case of discrepancy, the first image estimate is updated according to the underlying algorithm;

5. the process is repeated until a condition is satisfied;

6. the algorithm converge and we obtain the reconstructed image

Iterative techniques reduce noise and may reduce some specific artifacts, particularly when few angles are acquired, but they may be affected by other kinds of artifacts, such as the aliasing patterns and overshoots in areas of sharp intensity transition. Moreover, IR may affect quantitative measures of specific problems and may potentially lead at diverging results when compared with FBP.

Finally, hybrid reconstruction algorithms, which combine analytical and iterative methods, can be used.

To sum up, a CT can be reconstructed using different convolution kernels or filters and after that a windowing is applied to visualize the image. Both the reconstruction process and the preparation for visualization have an effect on the image quality and the observer performance. These effects are due, for example, to reconstructed slice thickness, reconstruction filters, tube voltage, tube current or windowing.

1.6 Deep learning and Medical Imaging

In the last few years, deep learning has been applied to solve many problems, including the ones related to medical imaging. Deep learning consists mainly in Artificial Neural Networks (ANN) with representation learning. There exist many architectures such as convolutional neural networks, recurrent neural networks or multilayers perceptron, suited for different scopes. In this PhD thesis, Convolutional Neural Networks (CNN) have been used to tackle different tasks: classification, regression and segmentation.

1.6.1 Convolutional Neural Networks (CNN)

A CNN is a neural network used to analyze structured data which has in its architecture convolutional layers. In Figure 1.11, Alexnet, the CNN archi-

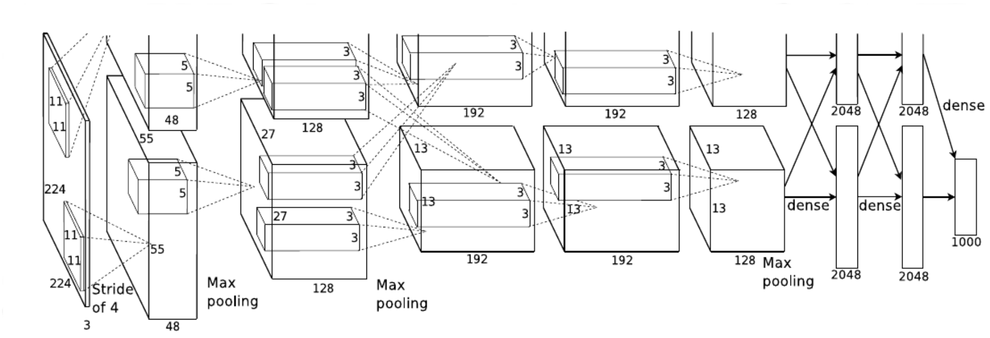

Figure 1.11: This is a graphical representation of AlexNet, the convolutional neural network that won the ImageNet competition in 2012. This image has been taken from [72]

tecture that won the Imagenet competition in 2012, is shown [72]. If A is a matrix of $M \cdot N$ dimension and H is a squared matrix $k \cdot k$ where k is an odd number, convolution between A and H is:

$$C_{AH} = A \otimes H = \sum_{p=0}^{k-1} \sum_{q=0}^{k-1} A(i-p, j-q)H(p,q) \tag{1.12}$$

H is the filter, commonly called kernel, that slides through the entire images in steps whose size can be chosen and it is called *stride*. The result of the convolution between the input image and a kernel is called *activation map*. The introduction of convolutions in an ANN is important because it deals with three ideas that are fundamental for machine learning: sparse connectivity, parameters sharing and equivariant representation.

When we use 2D or 3D images, because of high dimensionality, it is impractical to connect neurons to all neurons in the previous layer. As an example, if we have a 500x500 pixels image, we have 250000 pixels. If we fully connect all these pixels to a hidden layer of a hundred neurons, we will have about 25 millions of connections. In deep networks, more than one hidden layer is usually used. So if we connect neurons to all neurons, we will have an unmanageable number of parameters. For this reason, we connect each neuron only to a local region of the input. The extent of this region is called *receptive field* and it is a hyperparameter that is equivalent to k in Equation 1.12. The idea of a receptive field comes from some biological considerations. In 1968, Hubel and Wiesel [56] studied the response of the striate cortex in monkeys. They found that any small part of the striate cortex can be activated or suppressed in response to specific visual stimuli. Parameters sharing means that every parameter is used for more than one function in the model. Convolution, in fact, is made between the input and the kernels, which have a smaller size than the input image, at every input position. This means that rather than learning a separate set of parameters for every location, we learn only one set. Lastly, since kernels are applied at every input positions, convolution guarantees translation invariance such that a specific pattern in an image is recognized despite its location in the image. These three characteristics allow to reduce the storage requirements for the model and the runtime. Spatial dimensions are treated in an asymmetric way: the connections are local in space (along width and height), but always full along the entire depth of the input volume. This means that if we have a three channel image (RGB), we can choose the kernel height and width but the depth of the filters will always be three, in the first layer. A typical layer of a CNN consists in three steps: first, convolutions are applied to the input image, then a non-linear function is applied to the convoluted image and, in the last step, a pooling is usually used to further modify the structure of the output. Pooling is an operation in which pixel values are aggregated through an invariant for permutation function such as a maximum or an average. The output of these layers is a downsampled image. Pooling helps to make representation invariant to small translations of the input, which can be a useful property. CNN may be used for different goals. In this PhD thesis they have been used for regression, classification and segmentation. The single network architectures are described in Chapters 3 and 4 as well as the other hyperparameters specific for each task.

Basically, the network structure for regression and classification problem is quite the same. A CNN for these scopes, in fact, is made by several convolutional, non-linear activation and pooling layers with a flattening function at the bottom. The flattened data are then sent to one or more fully connected layers that are trained to solve the given task. The last layer is crucial to define the task we want to solve. In regression problems, it must have a number of neurons equal to the number of variable we want to infer. As an example, in chapter 4, a CNN for regression has been used to predict two points that define a bounding box which should contain the lungs. Since CT scans are 3D images, we need to infer 6 coordinates for two points. In this case, the last layer of the CNN has exactly 6 neurons. Moreover, the last layer of a CNN for regression has a linear activation function because we want to predict numerical values without any transformation. Similarly, the last layer of a CNN trained for classification has a number of neurons equal to the number of classes we want to predict. In this case, the activation function of the last layer should not be linear and it could be a sigmoid or a softmax depending on the number of classes we want to predict. As regard segmentation, instead, the CNNs have a different architecture with respect to the previous ones. In a segmentation problem we want to assign to each pixel or voxel a class in order to obtain a mask of the desired areas. Typically, Fully Convolutional Neural Networks (FCNN) are suited for segmentation tasks. The main difference with CNN is that there are not fully connected layers at the bottom of the network. It is possible to use a standard CNN, such as a VGG [138] or a ResNet [51], to perform the segmentation but the chain of convolution and downsampling makes the resulting map in low resolution. For this reason, there are several different methods that have been developed to tackle this task. In this study, U-net [123] has been used to perform the segmentation. U-nets are FCNNs which architecture resembles the shape of a U: in the left path, also called compression path, there are several strided convolution blocks while in the right path, also called decompression path, several deconvolution operations are applied to the images in order to upsample them till the input image size. In order to maintain the fine grained information, skip connections are used to pass the output of each block of the compression path to the corresponding block of the decompression path.

1.7 Applications to medical images

Artificial Intelligence methods, especially deep learning based ones, are playing an increasing role in biomedical research and they have a potential in many applications from risk modelling to diagnosis, prognosis and prediction to response to therapy. In standard Machine Learning (ML) hand-crafted features are usually extracted from segmented data; then they are pre-processed, normalized and selected before the training of a predictive model. This approach applied to medical images is called Radiomics and it represents the bridge between medical imaging and personalized medicine [78]. Radiomics approach has shown a great potential in many fields and one of the most substantial is the oncological one. It has been applied to improve the understanding of tumor biology, such as tumor heterogeneity [84] [46], and to better implement personalized medicine in many ways. Radiomics has been applied to improve diagnosis and prognosis: for example, in [50], it has been shown that such approach applied to low-dose CT for lung cancer screening may help in assessing cancer risk and in [46] it has been studied that the tumor shape complexity is a patient survival predictor. Moreover, it has been found that radiomics helps in predicting treatment response and disease monitoring and survelliance [87].

Deep Learning (DL) models applied to images instead learn by themselves to extract the features to be used for the predictions. However, a huge amount of labelled data is usually required to train, validate and test DL models and labelled medical images datasets are scarce and underpopulated. For this reason, data augmentation and/or transfer learning are typically used to overcome the data limitations issue. It is also possible to generate synthetic data with Generative Adversarial Networks (GAN) but the high resolution that characterizes medical images and their scarcity make it difficult to create whole images. As an example, in this study [70], a progressive GAN has been trained to obtain "high" resolution mammograms but the authors had access to a private dataset of more than 1 million mammograms. However, it is possible to generate patches of images but this should be used when it is appropriate. In the last few years, the concept of federated learning is being discussed: since data exchange has to be compliant to ethical and legal constraints, the idea is to decentralize the learning process, making it directly to the data sources, e.g. hospitals. By implementing decentralized data models, it is possible to perform multicentric studies sharing the models instead of data [19]. It should be noticed that such studies modality requires a huge effort to standardize data and process them coherently in every hospital. Moreover, a huge economic investment is required to build the necessary infrastructure to perform federated learning and such effort would also deliver a substantial improvement of HPC resources in healthcare environments. The imaging fields in which deep learning can be applied are many from neurological field to cancer diagnosis characterization, prognosis and therapy outcome predictions, pathology, microscopy and radiotherapy. In [126], a CNN has been trained to distinguish between lesion and nonlesion on Digital Breast Tomosynthesis (DBT). In [69], a Computer Aided Diagnosis (CADx) software has been trained to recognize benign cyst and soft tissue lesions on mammography. FCNN has been used, for example, for automated multi-organ segmentation which may help to plan radiotherapy treatments [124]. The specific literature for the two cases developed in this work is reported in Chapter 3 and 4.

1.8 Deep learning and black boxes

Deep Neural Networks (DNN) and machine learning algorithms have the capacity of reaching high performance in terms of accuracy and have been applied in a wide range of research fields. However, despite their high performance, they may take decisions that are not explainable and, for this reasons, they are called "black boxes" [47]. Understanding how a neural network or a machine learning algorithm, made by many learnable parameters, comes to a decision is not a trivial task. In order to understand how important may be to control the automated decision making process it is interesting to present some proven examples of how and why black boxes can be dangerous. Lowry et al.[96] studied a computer program used to screen job applicants for the St. George's Hospital Medical School in London, which has been trained without any reference to ethnicity. They found that the software unfairly discriminates against women and ethnic minorities by inferring this information from names and places of birth, resulting in a lowered possibility of being selected. In [45], a review of the state of art of the research on skin cancer detection has been reported and they affirm that there is a significant difference in the performances between caucasian and darker skin people, which leads to misdiagnose cancer in Hispanics and Blacks. In radiology and medical images analyses, this issue should be taken into account too. In [79], the effect of gender imbalance in training data set of chest X-ray on Artificial Intelligence based methods has been studied and they report that training a DNN with gender imbalanced data sets lead to significant different performances on the underrepresented gender. Moreover, in their case, training the algorithm using a perfectly balanced data set allows to obtain the best performance for both genders. The main question that should be tackled is: is AI able to recognize gender or ethnicity directly from medical images without any other information and why? In [7], the ability of CNNs to recognize race has been studied. In order to perform the study, they used large medical images data sets acquired with different modalities (chest X-ray, CT scans, mammography) and trained a classifier to classify self-reported race. Moreover, they tried to understand why and how an AI can recognize the patient race. They found that race can trivially be inferred from all the modalities they studied. They also trained a classifier to predict a diagnosis on chest X-ray to demonstrate that an AI not specifically trained for ethnicity classification is able to predict race. First, the algorithm has been trained to predict the disease and then they used the penultimate layer of the CNN and added softmax layer to measure whether it is possible to identify race from the features learned to make the diagnosis. Moreover Banerjee at al. [7] tried to understand why an AI is biased by race, given also that physicians can not recognize ethnicity from medical images. They studied whether it depends on body habitus, tissue density (mammograms), disease labels, bone density, age and sex and whether it depends on some image characteristics such as image quality, image resolutions, if the information is localized on a specific anatomic region, finding that it is not easy to mitigate or isolate the effect due to race. Despite bias effects, it is important to also have in mind that the diagnosis assessment should be a transparent process that both the physician and the patient should understand. This issue is strictly connected to the accountability problem.

1.8.1 Explanation of Convolutional Neural Networks

Methods to explain models can be roughly divided in two branches: the reverse engineering and the design of explanation [47]. The first one is based on the explanation of an outcome or a model trained for a specific task while the second one consists in developing an interpretable predictor model together with its explanation. The first modality can be addressed in three ways:

- model explanation: the aim is to understand the overall logic behind a black box;
- outcome explanation: the scope is to explain the correlation between input data and outcomes;
- model inspection: it is a modality in the middle of the two previous one and depends on the specific problem we want to explain.

As regard the second modality, i.e. the transparent box design problem, the aim is to build a locally or globally interpretable predictor to solve the specific task. An example of interpretable predictor is the decision tree in which the explanation is simply the rule chain of the tree. All these approaches to explanation are task dependent. In the last few years, agnostic approaches are being developed for explaining the black boxes. In this context, agnostic means that the this approach aims to explain the black boxes despite the specific task or model.

One of the most intuitive way to explain a classifier trained with tabular data is to understand what are the most important features the classifier uses to assign a class. However when the classifier is a CNN which learn by itself the features it is not trivial to interpret even the single feature. Given also that a CNN is made by several layers in which many kernels learn different patterns, it is not easy to design an explanation for their decision making. In [131], it is reported that there are three main levels in which a CNN can be explained: first layer, intermediate layers and last layer levels. In the first layer of a CNN, low-level features are usually stored such as orientations and edges while in the subsequent layer higher level features are learned. Since the first layer features are computed through a direct inner product with the input image it is possible to roughly understand what these filters are looking for just visualizing them. However, there is not a quantitative method to measure the amount of explained algorithm. Differently from the first layer, intermediate layers are not so easy to be visualized since they are the results of products with the previous layer which is not the input image. However, there exist methods to understand CNNs intermediate layers based on gradient-based, such as grad-CAM, and activation maximization approach. One of these approaches has been presented by Zeiler and Fergus [159], who studied how to visualize intermediate features by exploring which image patch activates most the neurons. They propose to visualize the patches corresponding to the part of the image that causes the maximum activation and use deconvolution made by Guided Back Propagation (only positive gradient are backpropagated). The main idea is to obtain a saliency map, which is a synthetic image, of a specific neuron. Another approach to intermediate layer explanation is through the extra-features [108]. While the previous approach assumes that each neuron can detect only one type of feature, in this one it is assumed instead that neurons can be *multifaceted*. In order to obtain a neuron's multiple facets, they use a k-means algorithm to cluster different images that highly activate the neuron and then the activation maximization is applied to produce a synthetic image. Finally, it is possible to study the explanation of a CNN through the last layer. Considering a CNN-based classifier, the flattened layer at the end of the convolutional path produces a vector that summarizes the input image. One approach to explain last layer is to compute the nearest neighbors of the last hidden layer for different images and to assume that CNN considers similar the images whose features vectors are nearest. Another way to visualize the last layer is to use techniques to reduce dimensionality such as the Principal Component Analysis (PCA) in order to obtain 2D or 3D data that can be plotted. However, using linear projections is usually impossible since it is difficult to represent high dimensional data in such representations. Agnostic methods to explain the CNNs have been also explored even if their functioning is not straightforward [35]. For example, Zolna et al. [163] propose an agnostic approach to find the pixels of an image that if obscured can confuse an unknown classifier. To produce such images they used an encoder-decoder approach plus a classifier to produce masked-in, masked-out and inpainted masked-out image in an architecture which resembles to a Generative Adversarial Network (GAN) [44].

Chapter 2

Inside the Complexity

Nowadays, Data Science is having one of its most expansive moment in the history of every developed country. Using data driven algorithms to infer, predict, evaluate and build models appears to be a real revolution in the epistemology field and it is changing the rules of the classical scientific method. Artificial Intelligence-based methods have been and will be applied in many fields, from social sciences to theoretical physics. One of the most interesting point of this change is to study and understand how scientific method is evolving and, in particular, which are the possible scenarios for science and technology. Since this PhD thesis is focused on medical image analysis, the complexity of using data driven algorithms in this domain is discussed in the following sections. In Section 2.1 the hypothesis problem is presented along with a literature review on the different possible scenarios and other issues such as reproducibility of data driven algorithm. In Section 2.2, the issues concerning the labeling and hence the ground truth definition in medical image analysis are presented. In Section 2.3, the differences and potentialities of private and public data are discussed along with a discussion on the most used medical image formats and on the biggest public database of medical images, The Cancer Imaging Archive. In Section 2.4 the concepts of statistical and clinical validation are discussed. Furthermore, the role that AI systems may have in hospitals and their consequences have been discussed. Finally, in Section 2.5, the conclusions of all the above presented issues are discussed. Moreover, Z-inspection project is presented as a possible solution to the problem of complexity of developing ethical and high performance algorithms for medicine.

2.1 Changing the scientific paradigm: are we back to 1500?

The coming of the Big Data has provoked a revolution not only in the scientific methods but also on the common perception of science and the relationship with technology. As it lays its foundations on using a big amount of data in order to build models, the advent of data driven science has brought to speak about a "New form of Empiricism" [68] as well as "The End of the Theory". According to this interpretation, epistemology is changing too, as regard the knowledge production, the processes of research, the information flows and the nature of categorizing reality. For this reason, the development of algorithms to support diagnosis or to analyze medical images, requires a special attention on scientific premises and hypothesis. Designing a comprehensive history of the scientific method is a goal which is beyond this work. However, it is possible to broadly describe some of the scientific paradigms starting from the definition of paradigm itself and going through some selected papers. In 1962 "The structure of Scientific Revolutions" [75] was published by Kuhn, introducing the notion of paradigm in science. According to Kuhn ("The Nature of Normal Science"), a paradigm is an accepted model or pattern whose fundamental components, for a certain period, remain substantially undisputed. In science, it assumes the shape of an object for further articulations and specifications under new and stricter conditions. When a new paradigm arises within a disciplinary field, according to Kuhn, it is very limited in both its scope and precision. It gains a dominant status when it is more successful than other paradigms. Kuhn means as success of a paradigm its capability of being able to solve problems considered extremely important by the relevant scientific community. However, at the start, it is mostly a "promise of success discoverable in selected and still incomplete examples". Normal science, the sort of scientific activity deployed after the establishment of a paradigm is, according to Kuhn, the actualization of this promise and it allows to expand knowledge in many directions. One of the most important drawback of this approach to describe science is that it may force nature to fit into the paradigm and it may lead to ignore what does not fit into it. Even if Kuhn recognizes that normal science usually investigate a limited area of knowledge, he affirms that those same restrictions were born from the confidence in a paradigm and they are essential to the development of the discipline. This way, science is able to do research in a very detailed and deep way. Moreover, when a paradigm works, the nature of the objects of inquiry changes: the paradigm broaden its scope, and it is applied to issues beyond its initial reach. Finally, according to Kuhn, a part of these achievements prove to be permanent. So, we can summarize the idea of the paradigm according to Kuhn as an accepted way of interrogating the world and produce knowledge which is common to a substantial proportion of researcher in a discipline at any one moment in time [68]. A classical critique to the Kuhnian approach is that, in some academic domains, there is only a little evidence of this modus operandi. Furthermore, taking into account just a paradigmatic approach produces too clean and linear stories on how disciplines evolve, deleting the pluralism of the history of science. However, the definition of paradigm elaborated by Kuhn has been very influential and allows for more clarity in the discussion on the epistemology of data science. In fact, big data and deep learning algorithms introduce a new epistemological approach, testing a theory by analysing relevant data and inferring the theory itself from data. According to Kitchin [68] and Hey et al [52], we can delineate a very simplified scheme to classify how scientific paradigms evolved. The first one, called ”Experimental Science”, can be dated back to the pre-Renaissance and it is based on a pure empiricism based on the observation of natural phenomena. The second paradigm is the so-called ”Theoretical Science” and it dates back to pre-computer era. It consists in moving towards a broader generalization through the theoretical modelling. The third one is referred to as ”Computational Science” and it dates back to the pre-Big Data era. This paradigm is based on the simulation of complex phenomena. One example in Physics may be found in the Monte Carlo Methods which have been invented, in their modern version, by Stanislaw Ulam while he was working on nuclear weapons projects at the Los Alamos National Laboratory. The ”Fourth Paradigm” [52] is the ”exploration science” which is the paradigm that the intense use of big data and the data mining techniques are designing nowadays. As written in the transcription of a talk given by Jim Gray to the NRC-CSTB1 in Mountain View, CA, on January 11, 2007 and reported in [52], Exploration Science, also called e-science, is mainly based on unifying theories, experiments and simulations using data taken by instruments or simulations and analyzing them with some software. For example, in the medical images analysis domain, especially for diagnosis, prognosis and follow up studies, data are usually taken by instruments from hospitals and, with some exceptions, simulations are usually used for dosimetric studies and evaluation. This way, the information and also the knowledge is stored in computers and the scientists analyze databases and files using data management and statistics. Gray affirms that e-science is changing the world of science itself, arguing that the techniques and the technologies are defining a data-intensive science which is a radical extension of the established scientific method. Kitchin states also that there are others that look at the Fourth Paradigm as the new era of empiricism, underlining that the main difference between a pure empiricist approach and other kinds of approach concerns the places in which Big Data are used, i.e. industry versus academy. There are many voices and opinions which try to define the e-science paradigm, its boundaries and its potentials. In 2008, Anderson [4] in the essay titled "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" states that "Correlation is enough", so that, in the Big Data era, correlation overcomes causation. Prensky [119], similarly to Anderson, affirms that data mining techniques can extract the complete set of patterns and effects, producing scientific conclusions without any further experiment. In 2013 Steadman [68] comes to affirm that data analysts should not propose or even "bother" themselves with hypothesis anymore. Even if these positions about e-science are typical of industry, its critical discussion should be taken into account even in the academic research. We can summarize this way of intending the Fourth Paradigm as:

- Big data can capture an entire domain of knowledge with full resolution of all the involved processes;
- There is no need of a priori model, theory or hypothesis;
- The application of data mining is agnostic, data can speak for themselves, i.e. data are inherently meaningful and truthful;
- Meaning transcends the specific domain such that anyone, with minimum statistic knowledge, can interpret the results.

This way of intending Data Science can be properly called a pure empiricist and inductive approach which really mirrors to pre-Renaissance empiricism. This approach may be dangerous for several reasons. Data are always taken and acquired using sampling techniques and data selection, which always introduce a bias. The interpretation of the models can not be done without theory. Finally, affirming that everyone is able to interpret the results without expertise in that specific domain is a reductionist and functionalist approach that ignores the socio-political context of the techno-scientific practice. In the medical images domain, this approach is risky and raises several issues. First, it poses a scientific problem. Neglecting bias and data sampling limits lead to the inability to define the boundary conditions in which algorithms we develop can properly work. For example, we can have an algorithm which is able to diagnose a certain disease without specifying on which population it has been trained and tested; on which imaging system manufacturers the algorithm works or which image acquisition characteristics it needs in order to function properly and it may be commercialized and used in health institutes as an universal tool for every-body. Furthermore, as this PhD thesis wants to demonstrate, the use of algorithms in medical domain should be supported not only by one specific expertise but by many expertises, sharing language, methods and cooperative approach, mindful that a such complex task needs many points of view to substantiate the application of algorithms in clinical practice. Finally, developing a tool or an algorithm in the medical images domain it is not something which is unrelated to the social context since it deals with hospitals, physicians availability, financial support, privacy management and also with how institutions that menage health, academic research and technology in general interact among themselves.

2.1.1 Hypothesis and epistemological claims in healthcare and medical image analysis

The Fourth Paradigm, as pure inductive empiricism brought by the use of Big Data and also by the rising of the deep learning methodologies, has the potential to undermine the scientific legitimacy of the machine learning [34]. As an example, Campolo and Crawford [17] uses the Enchantment theory of Max Weber to describe a broader epistemological diagnosis of modernity. They affirm that not understanding the motivation that leads a deep learning based model to a decision could produce the effect of considering that algorithm as something magical. These considerations do not come only from humanities but also from "hard" sciences. Stuart J. Russell, a well-known professor of computer science from Berkeley University, in 2018, spoke about deep learning and described it as "a kind of magic" since we cannot understand when and why the deep learning hypothesis is consistent. According to Campolo and Crawford, the use of terms like "magic" or "alchemy" can create a hype which is also beyond the epistemological problems and involve the entire society. However, despite the social hype that this terminology may generate, it also undermines the scientific process basis. For this reason, it is interesting to discuss the process of knowledge generation, evidence and causation in particular in the healthcare domain. In [140], a critically and healthcare centered review of epistemological claims is presented. The healthcare field is characterized by an institutionalized set of epistemological principles and generally accepted scientific methodologies [140, 10] which are challenged by the data science practices. The language used to describe the applications of algorithm in healthcare can be an interesting way of analyzing whether there are different ways of using big data in this specific domain. Stevens et al., studying systematically the editorials on the use of Big Data practices in healthcare, describes five ideal typical discourses, naming them using the relations between implicit assumptions about evidence and knowledge and the diverse epistemological positions. The five categories they design are: the modernist, the instrumentalist, the pragmatist, the scientist and the critical-interpretative. In the modernist discourse, big data are often not defined, described as a positive development and their benefits are stressed. In this type of discourse the use of Big Data is completely recommended, optimistic and accompanied along with words like "explosion", and "world-changing possibilities". They also create a sense of urgency of using this technology in contrast to a slow, conservative and old-fashioned medicine. Lastly, there is almost no attention on the possible negative sides of Big Data, such as privacy issues. This approach is based on an epistemological model that tend to naturalize the existence of data treating them like other natural resources. In this frame, data and knowledge seem to be equal as in a pure empiricist model. The modernist discourse supports a radical change in medical knowledge, rejecting all the traditional one. In the second discourse, the instrumentalist one, big data are presented as a set of analytical tools, such as pattern recognition or machine learning. The tone of these editorials are mainly positive and they typically discuss how such techniques should be used, with reference to missing data problem, correlated features and the separation of training and validation sets. Similar to the modernist discourse, in this second type, data seem to exist and are viewed as something with an intrinsic value; on the contrary, in the instrumentalist editorials, the fact that information can only be extracted from data with different techniques is emphasized. The epistemological assumptions are that traditional methods for knowledge generation are outdated and inefficient and that knowledge increases together with the set of used techniques. Regarding the relationship with healthcare domain, this approach seems to treat the Big Data techniques as a reliable source for decision making and to envision them as a tool to solve problems, which is valid to the extent that it helps to make accurate predictions. The third ideal-type is the pragmatist one and, in this frame, Big Data are seen as a positive useful managerial instrument for problem-solving and decision-making in healthcare. The advent of Big Data is a phenomenon that is and will stay here and people are presumed to have a significant role in the way they will be used, as opposed to the more technological determinist way they are implied by the two previous discourses. Pragmatist discourse editorials are focused on the training, recruitment and introduction of the data scientist role and on the cultural factors, the new rules and regulations that need to be made to introduce the data mining techniques in the healthcare practices. As regard the epistemological implications, data, as the two previous ideal types, are seen as something that exists but they need to be translated in information and knowledge. The pragmatist approach sees the new Big Data techniques and the traditional approach as complementary. Similar to the instrumentalist discourse, the pragmatist one considers data as a source for decision making next to traditional knowledge production approaches. However, also in this discourse the epistemological changes due to the use of Big Data are not exposed. The fourth ideal type is called the Scientist discourse in which Big Data are considered as a new trend that concerns data collection, analysis and outcomes in a less rigorous way with respect to the traditional approach. The editorials speak about the possibility of using them to generate hypothesis and to explore data sets. The tone of the scientist approach is critical since data can be used to hint the possible directions for traditional research methods which remain essential to knowledge production. As regard the epistemological assumptions, it seems that Big Data can lead to reliable and valid knowledge if and only if they are selected. In this approach, data are not given as natural or pre-existing in the world. Another important point of this discourse is that more data is not equal to better knowledge contrarily to all the other previous ideal types. Despite this criticism, the epistemological position seems to be similar to the modernist and instrumentalist one since the positivistic notion that truth can be found in data is present. However, in the scientist approach, it is clear that data cannot capture an entire domain of knowledge and hence the process of hypothesis and theory formulation is still valid. A deep difference with respect to the previous discourses is that the scientist approach does not claim for a radical change in healthcare since Big Data are not reliable as knowledge source. The only proper way to produce knowledge is the use of strict scientific methods. The last ideal type is the critical-interpretative discourse where Big data are presented as an oversimplified representation of reality. The critiques made by this approach are both epistemological and societal: Big Data are dismissed because they are a too simplified and reductionist way for representing the reality and unable to properly capture and account for the richness of human experience. The critical-interpretative editorials focus the attention on the importance of letting the results be interpreted by skilled physicians in order to avoid dangerous decision making. The epistemological assumptions of this approach is based on the needing of constructing data: data are no longer presented as something given but as a result of the social and political processes that created them and hence not only they cannot be complete but they necessarily emphasize some aspects while leaving out others. In the critical-interpretative discourse, Big Data will always generate limited knowledge and should be carefully used when applied to healthcare since their use may cause harm to people and healthcare systems. Making a complete essay on hypothesis, its role and use through the history of science, even limited to the healthcare domain, is a task beyond this PhD thesis. However, a discourse on the possible scenarios that may arise in Big Data science, that considers also the hypothesis, is very interesting to deeply understand some insights of applying data mining to medical images. In recent years, many practices, from machine learning to deep learning, have been widely used in the healthcare domain and their application in the radiological field seems to be very effective [88]. Even if we do not discuss about epistemological questions when, as scientists, we develop an algorithm, it does not mean that we are not making assumptions but that we are doing it implicitly. Questioning how and whether knowledge is produced is instead a pivotal moment to address scientific research towards knowledge production and a fair use of the data mining techniques in the healthcare domain. While a simple positivistic, hypothesis-and-theory free and purely empiricist approach seems to be a way of making the use of data simpler in the clinical practice, it is a trap. Regarding the field of medical image analysis and consequently the radiological medical domain, we know that data are not given, natural or pre-existing. Medical images are the results of:

1. A traditional scientific process: their production is based on physical

studies on the interaction between matter and radiation, human body and radiation, on the physical processes of X-ray, magnetic fields and ultrasounds production as regards radiology and radioactivity and all the issues linked to it as regard the nuclear imaging;

2. A technological development history: the image production deals with the detectors improvements, the materials used for detecting photons, the electronics which, simplifying, determine for example the spatial resolution, the contrast and other image quality characteristics. When we deal with 3D images such as Computed Tomography (CT) scans we should consider the image reconstruction algorithms which are a mix of traditional scientific research, especially mathematical one (for example Radon transform or Fourier signal analysis) and pure technological improvements such as the sliding contacts.

3. An industrial process: medical imaging systems are not equally distributed around the world and their production is highly costly producing as a result that there are few vendors that deals with the imaging machinery market. Moreover, as a result of an industrial process, some parts of the medical images production are protected by patents which inhibits the complete knowledge on how an image is produced (See section: 1.5);

4. A function-based process: medical images are made on the basis of their utility and improved following their possible uses in hospitals. The choice of using a specific imaging modality depends on the scope (morphology or functionality and diagnosis, follow up, radiotherapy planning, ...) and on the part of the body that needs to be imaged. They are made to be presented to physicians in a way that medical doctors can interpret and taking into account the specific medical formation process they attended. Moreover, contrarily to natural images, most of medical imaging modality implies the delivering of a radiation dose to the patients, making their use a dynamic equilibrium between costs and risks, in terms of capital and health, and benefits.

For all these reasons, it is unacceptable to consider medical images data as pre-existing or natural. Moreover, applying Big Data techniques, such as machine and deep learning, cannot be considered as a free hypothesis science. Even if the hypothesis is a complex hypothesis and it is very far from the pre-Renaissance way of formulating it, we are always assuming that, given the constructed data and the context, there is a model which may solve the given task we want to study. This means that we are using data that contain already the solution. Moreover, since medical images are constructed data, the use of data mining techniques is not free of theory. This means we should keep in mind that the choices researchers make are always guided by social and epistemological assumptions on data, which should be taken into account within the research framework. Characterizing the data science as comprehensive and intrinsically unbiased can be misleading rather than helpful in shaping scientific as well as public perceptions of the features, opportunities and dangers associated with data-intensive research [82]. Finally, as will be discussed in the following, medical images can rarely be considered as Big Data and hence the application of techniques developed on them should be even more careful as regard, for example, the generalization goal. What is at stake is our ability to produce knowledge not only in a traditional scientific way but making it from a critical position, avoiding the accusation of practising "magic" or "alchemy". It deals with knowing and assuming how much complex is to create algorithms, especially deep learning one, with the scope of applying them in hospitals and, by assuming it, proceed towards a fair, scientific, active and impacting application of data science to medicine.

2.1.2 Beyond Hypothesis: reproducibility

The problem of framing the Fourth Paradigm is bigger than the hypothesis problem. Deep learning is a set of techniques which includes many types of neural networks. The term "deep" [43] means that the learning made by a neural network is hierarchical and made to extract information from data, at different levels, organizing it into layers that are usually numerous and connected to each other. In image analysis, this way of elaborating images is different from the classical one. While Convolutional Neural Networks (CNN) act by automatically extracting important features in a hierarchical order, in the classical image analysis the idea was mainly to focus on some well-known image characteristics, such as borders, and to define a discrete mathematical operator able to capture those characteristics. The image features were defined a-priori and, once they were extracted, an algorithm could be trained to perform the decision. This modality for image analysis suffers from the difficulty of designing the features that should contain the information we need. On the other hand, the algorithms that are used to solve the task are quite simple to be understood. So, the process of elaborating a decision is sufficiently clear and many times fully explainable algorithms, such as Generalized Linear Model (GLM), can be used. The use of deep learning simplifies consistently the task of designing the features since, for example in the case of a classifier, we do not need to compute the features a-priori but only let the algorithm learn them from data. However, the transparency of the decision process is lost and it is not possible to explain why a deep learning algorithm, especially CNN, makes a certain choice. As written in Sec 1.8(Deep learning and black boxes), though there are several and different methods to try to explain how a deep learning algorithm works, explanation and interpretation are different tasks. At the state of the art, we know some methods to explain deep learning algorithms but interpreting them is a more complex task. This is why, sometimes, we refer to this kind of algorithm as "black boxes". This difficulty opens a pivotal question which deals with epistemology: are these algorithms reproducible and reliable? Reproducibility is the possibility of obtaining the same results of other researchers given the same experiment. Since neural networks training deals with stochastic computation, it is important to discuss how it is possible to reproduce an experiment. Moreover, the number of hyperparameters, that are the non learnable parameters of a Neural Network, is usually very high and it is quite impossible to describe all of them in a scientific report. Furthermore, the possibility of reproducing an experiment is strictly connected to data availability. As will be discussed in the following, data may be private or public and even if public datasets ensure the possibility to reproduce an algorithm, their publication implies some drawbacks such as the possible deletion of important information. Hence, the problem of reproducibility is another issue that is added to the epistemological problem making it a very challenging question to be addressed [142].

2.2 What Drives Medical Images algorithms?

When we train a deep learning algorithm for classification, segmentation or regression, we mainly try to solve a particular optimization problem. It means that we define a cost function or loss function which has to be minimized. The cost function measures the error with respect to a "truth" that the algorithm makes when performing its own task and an optimizer works to reduce as much as possible this error. Since the loss function needs to respond to some mathematical boundaries, in deep learning training another function to measure the performance of the algorithm is usually used. This separation between loss function and performance measure makes the deep learning training different from a standard optimization process. Moreover, the main task when we optimize a deep learning algorithm is not simply to find a minimum of the loss function, but also to maximize its generalization capability. This means that we want the algorithm to be able to work on data which differ from the training data set. A loss function is usually computed as an average over the training set or over a subset of it (batch) and it is defined as the expectation taken from the empirical data distribution of the cost functions computed on the training set. We want to minimize the expectation of cost function taken from the data generating distribution and not only from a finite set of data and this quantity is called **risk** [43]. However, it is often quite impossible to know the underlying data distribution and hence the risk is computed on the training set, transforming the problem in an empirical risk minimization problem. Unfortunately, this kind of problems is prone to overfitting if the loss functions used have unusable derivatives (zero or not defined everywhere). For this reason, in deep learning training we usually minimize a quantity which is different from the quantity we truly want to optimize. So, the question is what is the ground truth on which the cost function is computed in medical image analysis? What drives the training of medical images algorithms?

2.2.1 Labeling Medical Data

In Chapters 3 and 4, two use cases with different data and scopes are presented. In both of them, the learning paradigm is the supervised learning. This means that the loss function is computed between the predicted value of the algorithm and the true value and that the task of the optimizer is to make the gap between them as little as possible. It is important to underline that the final task is not just to minimize the error between the true and the predicted value but to let the performance of the algorithm be satisfactory on the test set. However, it is important to discuss how the true value, also called ground truth, is built in the medical image analysis domain. The ground truth usually depends on the task we want to solve and its creation is a pivotal step for algorithm development. There are different ways to label a data set of medical images. If the task we want to solve is a classification task, the ground truth consists in assigning a class to each image or patient in the data set. A patient image taken at different time points may belong to a different class because human body changes over time. The way the classes are defined is mainly based on medical protocols. For example, in Chapter 3, the classification of breast density on mammograms is discussed and the ground truth is based on the Fifth Atlas of Breast Imaging Reporting And Data System (BI-RADS) [133] which defines four classes through textual descriptions and image examples. In this case, the ground truth has been decided by a specialized radiologist who has looked at every image in the data set and assigned the class to that image. He was not supported by quantification tools and acted just looking at the four standard mammographic projections in order to produce his ground truth. Even if this labelling process seems to be very fast, when a huge amount of labelled data is required, the process is very time consuming for doctors. Another way for labelling medical images is to assign to each pixel or each voxel a certain class. This kind of labelling is suited for solving segmentation problems. For example, in Chapter 4, an automatic way to quantify the pulmonary damage due to COVID-19 is presented and the ground truth is made by masks that contours the lung lesions and the lung itself. A medical image usually contains many pixel and voxel and this characteristic makes the labelling extremely time consuming. If we suppose to have a standard lung Computed Tomography (CT) scan with size 512x512x100 the number of elements to be labelled is more than 26 millions! There are some tools to help physicians in this task but they may introduce a bias in the labelling. In order to reduce the cost of labelling, the use of non-expert people has been employed in the field of natural images but [77] the use of such kind of labelling process leads to highly noise data sets. In the medical images domain, in which the objects to be identified are usually small and are difficult to be identified, this process is even harder. Having the availability of large labeled data sets of medical images is currently a real challenge even despite the labelling process. Medical images data sets, in fact, are usually small and their collection is not easy because of privacy issues and institutional policies.

What is label noise?

When we refer to label noise, we do not refer to image or signal noise. The widespread of deep learning techniques brought with itself a variety of different forms of imperfections or corruptions on labels. In classical machine learning classification problems, a data set is typically defined through attributes or features and class labels. Making a classification means to select the attributes that better characterize a class. This process is based on two assumptions [160]:

1. There exists a correlation between attributes and classes. Not every attribute contributes with the same weight to the classification but there are some feature that are more important and others less.

2. We assume that there is a weak interaction among attributes. This assumption is important for many classifier which are trained based on a conditionally independent or even independent relationship among features.

Unfortunately, real world data do not always comply with these two assumptions. In fact, given a data set, it may contain attribute with very low correlation to the class or which strongly interact with other attributes. For this reason, in classical machine learning, we can identify three categories of label noise [62]: class-independent, class-dependent and class and features dependent. There are several techniques to reduce this noise and they are based on:

- model selection or design. The algorithm is chosen on the basis of its robustness to noise itself;

- reducing the label noise on the training data. These methods are very similar to the outliers detection methods;

- methods that train classifier and model the labels at the same time.

As regards deep learning, it usually needs a huger amount of labeled data which leads to a higher amount of label noise. Label noise is, in this case, not really easy to be defined. It can be described as the presence of incorrect labels or ambiguous one and it is unavoidable in many medical image data sets. The label noise is caused by low attention or limited expertise of the annotator, by the subjectivity of the thing we want to label or by errors in computerizing the labeling systems. Many studies have demonstrated that label noise degrades significantly the performance of a deep learning algorithm [62]. As an example, a CT scan and its labelling taken from the COVID-19 Challenge data set is represented in Figure 2.1.

There are several methods used to reduce as much as possible this kind of noise, such as label cleaning or data re-weighting but they have been

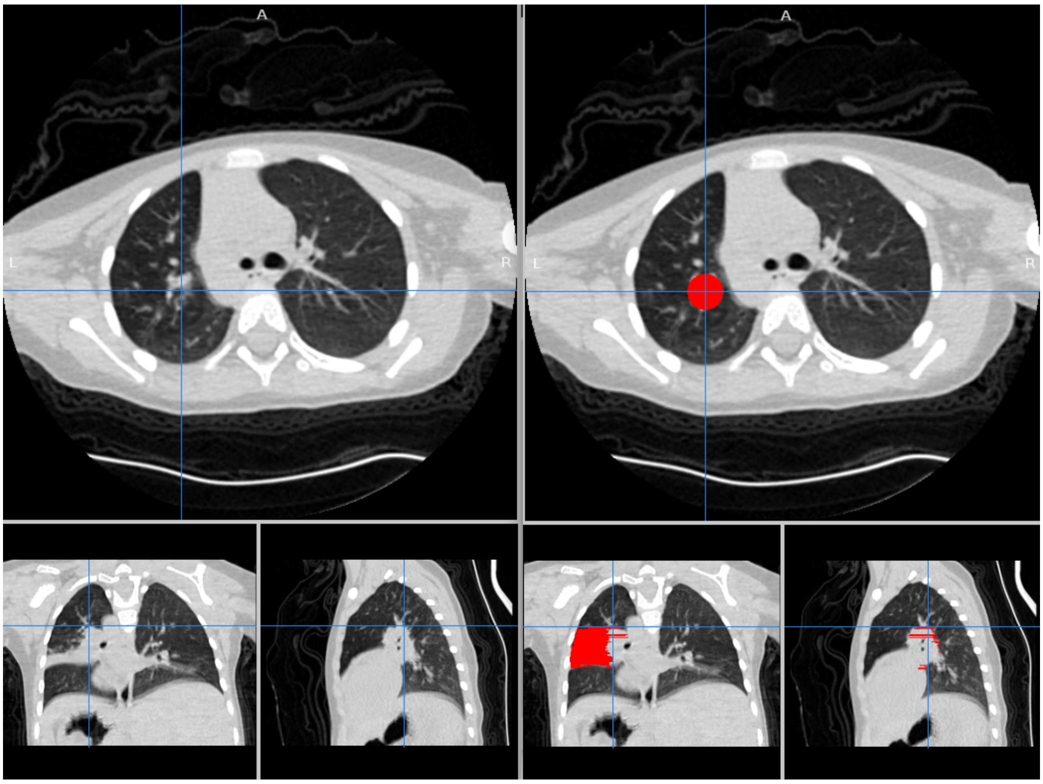

Figure 2.1: On the left: the original CT scan of patient volume-covid19-A-0013_ct represented in the HU window $[-1000, 300]$. On the right: the same CT scan with the ground truth overlay. In the axial plane: the labelling is made of a perfect circle and this can be considered a form of label noise. The presence of the perfect circle is a consequence of the labelling process made with the support of some tool. In the coronal and sagittal plane: it can be observed a strong discontinuity along z axis which represents another kind of label noise.

developed on very large natural images data sets and their use on medical images should be applied carefully.

2.2.2 Inter observer agreement variability

Another source of noise in the medical images labelling process is the inter and intra observer variability. Since labelling medical images requires expertise and it can be a very hard job, an image can be differently annotated by different radiologists. This kind of noise is always present, despite the fact that the labelling was made for both classification or segmentation. As regard the former, when making the labelling, a physician looks at the entire image or sometimes at more than one image related to a patient and assign a unique value, the class, to that image. There not always exists a true measurable and/or collectable class to be used as the ground truth in medical image classification problems. For example, in mammography, the breast density assessment suffers from the inter observer variability because it is usually not possible to have the real ratio between dense and fat tissue. The imaging exam that gives the best measure of them is the Magnetic Resonance Imaging (MRI) with medium contrast which is usually performed, in Italian hospitals, as second or third level of examination. This means that a woman undergoes to MRI only if she had been positive to at least another exam, which can be a mammogram, a ultrasound scan and/or a biopsy. Since it is made only on women who have a high probability of having breast cancer, it is very difficult to find MRI performed on healthy women and hence to build a balanced dataset. Moreover, building labels by crossing more exams related to a woman introduces not negligible privacy issues. Also segmentation problems present inter observer variability. In this case, labels are made by contouring a specific organ or a specific pathology, hence, assigning to each pixel or voxel a specific class. One possible solution to overcome the inter observer variability in both segmentation and classification is to produce the ground truth in a consensus modality, i.e. building the ground truth using a large number of experts that delineate the annotations. This way implies the use of massive financial resources and also logistical resources that are not easy to be obtained in many fields [62]. One interesting example in literature about label variability is the case of the LDCI-IDRI data set [5]. This data set consists in Lung CT with nodules and it was collected to study whether it is possible to use CAD for screening lung cancer. The annotation has been made in two steps by four radiologists: first they annotated the nodules in a blind modality, i.e. without knowing the answers of other radiologists, and then they labelled again the data set reviewing the annotation made by the others. The possible annotations were lung nodule bigger than 3 mm, lung nodule littler than 3 mm and no lung nodule. The LDCI-IDRI data set contains 2996 lung nodules bigger than 3 mm but only 928 (34.8%) of these lesions received nodule 3 mm marks from all four radiologists. The problem of variability also opens another issue which concerns whether it is possible to obtain a consensus and how is it possible to measure it. The most used method to measure consensus is the Kappa correlation coefficient [24] and [15] there are tasks, such as classification or diagnosis, which work with littler variability, and others, like segmentation, which may reach a very low agreement. This opens an epistemological and scientific problem. It has been established that the use of medical protocols may help in reducing the variability because it standardizes the way an image is read in some specific domains [63]. Even if guidelines may decrease the inter observer variability, it seems to be impossible to completely eliminate it. This issue leads to two possible interpretations: 1) the guidelines are inaccurate or insufficient and 2) there is an underlying variability associated to complex clinical tasks that even guidelines can not delete [15]. Another way to maximize the consensus is to train and educate physicians to label images [22] [33] [130] [86]. Discussing the epistemological address of health research is important to face the problem of variability. We can frame it in the positivist approach or in the constructivist one. The most recent research adopts a positivistic approach with the fundamental assumption that there exists a single truth [15]. In this view, there is the underlying assumption that the gold standard can be made by the opinion of an expert. However, guidelines and training have shown the capability to reduce the variability without erasing it at all. Several studies showed that the reliance on an expert opinion is not reliable [15] [20] while the lack of a true gold standard make the absolute comparison difficult [148]. So, we can conclude that variability exists and there is no way to delete it. Approaching this problem in a constructivist frame means to assume that there are several ”truths” that depend on inherent biases, experience or judgments not only among different individuals but also within the same individual at different time points. Bridge et al. [15] states that clinicians should embrace the variability in the constructivist approach, suggesting that, instead of deleting it, they should study what are the acceptable variability amount for the specific clinical task.

For all the reasons discussed above, it is not easy to build models for diagnosis, prognosis or to assess the follow-up of a patient and doing it in the right way requires the collaboration of many experts.

2.3 Public and Private Data

Deep learning algorithms and machine learning ones obviously need data. Deep learning in particular needs a huge amount of data which are not easy to be obtained. Medical images data can come from private or public collections and accessing to them is one of the biggest challenge of the medical image analysis domain.

2.3.1 Are Medical Images data sets big?

The term ”big data” etymologically come from the mid-1990s and it was used to refer to the handling and the analysis of massive datasets. The term ”Big Data” refers not only to data themselves but to a rapidly evolving use of technologies and practices and there is no agreed academic or industrial definition of this term [67]. The most common definition makes reference to the 3V [162] which stands for:

- huge in Volume;
- huge in Velocity, created nearly in real-time;
- diverse in Variety, being structured or unstructured in nature and often temporally and spatially referenced.

Beyond the 3V definition, in literature, other main characteristics describe big data:

- exhaustive in the scope, which means that big data capture entire population or system characteristics;
- fine-grained in resolution, which means that big data contain detailed information;
- relational in nature, which means they contain common fields across different datasets;
- flexible, which means they can be extended and scalable.

Given these characteristics of big data, it is interesting to question whether medical image datasets can be considered big. The Cancer Imaging Archive (TCIA) [23] is taken into account for discussing this issue. TCIA is an Open

Data Portal made available by The National Cancer Institute (NCI), under the supervision of the University of Arkansas for Medical Sciences (UAMS), with the aim of encouraging cross-disciplinary science and increasing transparency and reproducibility in cancer imaging research. It is a project which publishes medical images data and it is one of the most important source for academic research. The collections are mainly divided in two categories (for an accurate description of TCIA access to data see **Sharing medical images: The Cancer Imaging Archive** section): access free data and data which can be accessed only with specific usage policy agreement. It was born with the aim of supporting research about cancer but, in the very last months, it published also data related to COVID-19. It contains 147 collections of several imaging modalities and for several scopes. The most populated dataset is the National Lung Screening Trial (NLST) which contains 26254 Computed Tomography (CT) scans of lung. This dataset was made to study the feasibility of a screening program for lung cancer. The second most populated dataset is the Breast Cancer Screening-DBT which contains 5060 Digital Breast Tomosynthesis (DBT) images for studying whether it is possible to use DBT as a screening tool. Including these two datasets, 8 published datasets contain more than 1 thousand subjects; 51 datasets contain a number of subjects between 100 and 1000; the other datasets contain less than 100 subjects. So, medical images data sets can rarely be considered as big data since their volume is considerably small. However, even if their creation happens in a nearly real-time, their collection does not because of privacy issues and also of storing criticalities. They can be considered various because they are structured data with temporally and, sometimes, spatially, referenced. As regards the other characteristics, datasets which contain so little amount of subjects can not be considered as exhaustive at all and they are not flexible since it is not easy to add new fields or information to already collected data. However, they may contain fine-grained in resolution information and be in relation one to each other, given some boundaries. These boundaries concern the imaging modality, the imaged subjects, the scopes of the collection and across the same modality, subject and scope there may be technical issues, due to, for example, reconstruction algorithms and also timing issues since some collected data were made with technologies that are not used anymore. Hence given the above considerations, in particular the one that refers to the volume, it is very rare that a medical image data can be considered as pure Big Data. For this reason, the application of big data practices to this kind of data should be made carefully and considering all these issues.

2.3.2 Differences between public and private data

Data may come from public or private collections. Both of these two modalities have weaknesses and strengths which are going to be discussed in the following. First of all, public data may be effectively public, i.e. accessible to every one, or they may be accessed through a specific agreement. Private data are instead data which can not be used or accessed in any case. Data may not be accessible for many reasons. One of the most problematic is bound to privacy. In order to better understand the risks of publishing data it is interesting to discuss the most used image formats. This is important because medical image formats usually contain a header with patient and physician information. There are mainly two image formats typically used for medical images and they are the Neuroimaging Informatics Technology Initiative (NIfTI) [109] and the Digital Imaging and Communications in Medicine (DICOM) [29]. NIfTI format was created in the field of neuroimaging and it is a standard which contains a header with only information about orientation, voxel size and image visualization. 3D images, for example CT scans or MRI scans, can be stored in this format which defines uniquely the correct orientation and the physical volume. The Digital Imaging and Communications in Medicine (DICOM) standard [29] is the global convention used by manufacturers to define and store diagnostic imaging data. DICOM images are encoded as a set of elements; public elements are defined by the DICOM standard, and private elements are defined on an individual basis by each manufacturer. A DICOM data element or attribute is made of:

- a tag that identifies the attribute, usually in the format (XXXX,XXXX) with hexadecimal numbers, and may be divided further into DICOM Group Number and DICOM Element Number;

- a DICOM Value Representation (VR) that describes the data type and format of the attribute value.

The fields of the DICOM header contain many information from the patient ID, which is a number that uniquely identifies the patient, the Patient's Birth Name (0010,1005), the Patient's Age (0010,1010), the Patient's Size (0010,1020), the Patient's Address (0010,1040) or even the Patient's Mother's

Birth Name (0010,1060). Moreover it contains information about the referring physician name, the date of the exam and so on. All these data are a problem when we deal with privacy because they may allow a complete re-identification of subjects. On the other hand the DICOM format contains also information on the acquisition parameters such as the reconstruction kernel, the imaging system used, exposure time, X-ray tube current, the field of view (FOV) size or the reconstructed FOV. These characteristics are less prone to be problematic as regards privacy and they are very useful for algorithms development. However, in most of published medical images data all this information is lost. This is mainly due to the fact that it is not so easy to treat privacy and DICOM standard since the number of tags that may be contained is very large. Moreover, making studies on humans imply not only privacy related issues but ethical issues too. For these reasons accessing to Italian hospital data requires a strict protocol to be carried out. Modality manufacturers use private elements to encode acquisition parameters that are not yet defined by the DICOM standard or that they consider proprietary. Modality manufacturers also define and include private elements that contain Protected Health Information (PHI). These PHI private elements can be as obvious as the name of a patient and as subtle as an identifier string that could be tracked back to a patient by someone with access to the departmental image archive. A DICOM conformance statement is a document published by a manufacturer that contains technical information concerning data exchange with a specific type of device (e.g. an imaging unit, workstation, printer, image archive). The conformance statement provides the mechanism for a manufacturer to publish the set of private elements that are stored in the DICOM files created by an imaging system. Manufacturers do not document and publish all of their private elements. For these reasons, the de-identification process should meet two conflicting requirements: (i) any PHI must not be included in exported data and (ii) the system must retain all data that describe the acquisition, such as physical parameters for individual images, as well as other parameters such as series description. Acquisition parameters change according to the image modality: as an example, tube voltage and slice thickness are important for a CT, while the magnetic field strength is essential for MRI. De-identifying a DICOM collection is not easy and there are several technical challenges to satisfy the requirements [105]. In summary, they are:

1. DICOM standard elements with well-defined semantics are abused dur-

ing the collection. In fact, some elements are written by the radiology technician at the console. Instead of using the field for intended purpose, such as “Image Comment”, the technologist may enter PHI.

2. Vendors use private elements to encode acquisition parameters not yet documented by the DICOM standard. Furthermore, they may use private elements to record demographic information.

3. DICOM sequences provide a mechanism to nest data elements at different levels in DICOM objects and PHI may be encoded at these lower levels.

4. Manufacturers do not document all private elements and private elements may contain important acquisition parameters.

5. Image providers remove information from the images that identifies the vendor model and software.

6. The users and managers of the de-identification system may not be able to discuss the collections of images with the original imaging center.

De-identifying a DICOM image is a challenging task that carries the risk of leaving in the header PHI or meta-data that makes the re-identification possible. On the other hand, the NIfTI format has been invented to have not patient information in the header but it does not allow to store important technical parameters. It could be interesting to study a new image format standard suitable for AI and deep learning algorithm which contains all the technical information while keeping the privacy risk as lower as possible.

Sharing medical images: The Cancer Imaging Archive

The Cancer Imaging Archive (TCIA) [23] is an Open Data Portal to share medical image data sets. At the NCI, researchers from TCIA collect and curate clinical and pre-clinical radiology and pathology images, clinical trial data, annotations and image derived features and other type of clinical research data. The database was born to share data about cancer but in the last few months it has been used to share COVID-19 data too. It is organized in different collections and it is possible to query the database selecting the collection name, cancer type, location, species, subjects, image type, supporting data, access, status and last update time fields. TCIA contains mostly biomedical images, such as mammograms or histopathological images, or 3D imaging data such as Magnetic Resonance Imaging (MRI) or Computed Tomography (CT). Sometimes datasets are accompanied along with clinical information about patients. There is no literature about the direct re-identification of individual from images but it can be made with meta-data written on DICOM header. Despite the re-identification through meta-data or clinical data, the possibility of identifying a subject from a public image dataset is strictly connected to the imaging modality used. For example, it is quite impossible to recognize a subject from a leg radiography while re-identification can be achieved from a head MRI image which can be used to reconstruct the contours of the face. A proposed solution to this problem is Federated Learning in which data are not moved from their original acquisition site and are used to train algorithm locally [61].

TCIA data may already be published or released in public domain. In addition, confidential information may be posted which has not yet been published or is subject of patent applications yet to be filed. In fact, some data sets, such as the Curated Breast Imaging Subset of DDSM (CBIS-DDSM, public domain), were already published and they have been published on TCIA with updates or different standardization. Other data sets have been collected and published directly on TCIA. Data has been collected by researchers and published with the ethic committee agreement.

Data may also be subject to copyright and commercial use may be protected under United States and foreign copyright laws. Other parties may retain rights to publish or reproduce these documents. In addition, some data may be the subject of patent applications or issued patents, and you may need to seek a license for its commercial use. Most data are freely available to browse, download, and use for commercial, scientific and educational purposes as outlined in the Creative Commons Attribution 3.0 Unported License or the Creative Commons Attribution 4.0 International License. In rare circumstances commercial use may be prohibited using Attribution-Non Commercial 3.0 Unported (CC BY-NC 3.0) or Creative Commons Attribution-Non Commercial 4.0 International (CC BY-NC 4.0). Furthermore, registration is often not required to access the data. It may happen that small subsets of a collection require a user registration as specified in the access database field. Since TCIA offers a space to publish and share data, it may happen that it is used by an institution to share data among its members (such as QIN Quantitative Imaging Network). Despite final uses, any user accessing TCIA data sets has to not attempt to identify individual human participants from whom the data were obtained according to TCIA policy. Moreover, acknowledging in all oral or written presentations, disclosures or publication the used dataset is required. Citation guidelines can be found in the "Citation and Data Usage Policy" attached to each collection.

It is possible to submit data to TCIA repository, certifying that you are the original source of the submitted data and you are authorized to release the data by your local Institutional Review Board (IRB) or independent Ethics Committee. It is needed to certify that the Technology development office of your institution has been consulted before posting or disclosing confidential information which can be patentable. TCIA does not charge a fee for sharing data except in rare circumstances where dataset are extremely large. Applications are reviewed every month by the TCIA advisory group to assess their utility to user community. They give a strong preference to fully public data sets and to ones which contain supporting non-image data, such as patient outcomes, training labels and tumor segmentation. If approved, data submitter must sign the TCIA UAMS Data Transfer Agreement or the TCIA Non-Commercial Data Submission Agreement, in the case submitters are not legally permitted to allow commercial use of their data. However, NCI and UAMS do not warrant or assume any legal liability or responsibility for accuracy, completeness or usefulness of information in this archive. In order to ensure that Protected Health Information (PHI) is not used or disclosed inappropriately, PHI from images are going to be removed by the submitter and again by tested automatic de-identification processes by the University of Arkansas for Medical Sciences according to the Health Insurance Portability and Accountability Act (HIPAA). All data is fully de-identified in accordance with international standards, US laws and UAMS IRB protocol requirements. Data is anonymized to the fullest extent possible and then encrypted prior to the trasmission to UAMS. Incoming data is stored in a quarantine system and treated as if it contains PHI. Data is analyzed and completely de-identificated and then moved to a separate public repository in order to make it available to the research community. This process has been reviewed by the UAMS Chief Security Officer. In order to help data sharing, TCIA provides data de-identification, curation and hosting services. TCIA uses a standards-based approach to de-identification of Digital Imaging and Communications in Medicine (DICOM) international standard for medical images, following the industry best practices. DICOM is most commonly used for storing and transmitting medical images enabling the integration of medical imaging devices such as scanners, servers, workstations, printers, network hardware and Picture Archiving and Communication Systems (PACS) from multiple manufacturers. Mainly this process removes or replaces with a hash the image metadata fields that can held to an individual identification. Patient name, ID, geographic information, dates, exam identifiers, patient demographics, free text entry fields, vendor private tags are removed to minimize the possibility of being able to uniquely identify an individual. Universal IDentifiers (UID), which are used in DICOM, are replaced with a hash since it may be possible to identify subjects if the user has access to the PACS system. Date and Date-Time fields in DICOM header have been offset based on a random number but the longitudinal relationship between dates is maintained. As example, it is possible to preserve the information about the amount of time between an exam and its follow up without knowing the exam date. Patient demographic characteristics, such as patient's sex, age and weight, may be useful for research purpose and it is possible to keep this information. TCIA represents a landmark in this research field and this thesis has been possible thanks to their work.

2.4 Statistical Validation versus Clinical Validation

Statistical validation is the task of confirming that the outputs of a statistical model are acceptable with respect to the real-data generation process. In machine and deep learning there are several ways to statistically validate an algorithm. The simplest way is to divide the data set into training, validation and test set. The algorithm is trained using only training data and then it is evaluated, during training, on the validation set. The introduction of the validation set allows also the best model selection as the training can be stopped at the epoch of the best performance on this set. Since the results are influenced by the samples in the validation set, the algorithm should be evaluated also on a set of data, test set, that is completely separated with respect to the training and validation one. This method is simple and it does not add computational time. However, it may happen that the validation and the test set contain some specific subset of population. In particular, this can be considered true when we do not have any further information about population. In order to overcome this issue, it is possible to perform a k-fold cross-validation. In this case the entire data set is divided in k subsets of data and the model is cyclically tested on one of these groups and trained on all the remaining data. This method has the advantage to test also the stability of the algorithm performance and it reduces the possibility of evaluating it on a biased sample. On the contrary, it increases the computing time and power so that it is usually a hard method to be used in deep learning validation. Another way to validate algorithms is the leave-one-out method which is similar to the cross validation but instead of a group of samples, just one sample is taken apart during training. While this method reduces randomness in sample choices to zero, it increases the computation time depending on the number of samples in the training set. So, it is recommended only on small datasets. Once the algorithm has been trained and validated using one of the above methods, it is interesting to question whether it can be clinically validated and how. AI technology is expected to be of substantial help in medicine with innovative solutions. There is a wide range of AI devices for healthcare; most of them are diagnostic tools, such as CAD, CADx or clinical decision support systems. However, machine learning methods, especially deep learning ones, are prone to overfitting because of their high dimensionality and complexity so that their performances deteriorates when applied to external data [112]. Moreover, public datasets usually differ from real data so that it is not easy to translate algorithms into practice. Even if patients have the same disease, other characteristics such as age, sex, comorbidities often differ across different hospitals. Moreover, different hospitals usually have different devices to acquire images and different scanners impress different characteristics on the images. Healthcare equipment advancements are continuously evolving producing as a result that an algorithm trained on older images may not work properly on more recent ones. For all these reasons, a key step to pass from statistical to clinical validation is to perform external validation using external data independent from training and internal validation data sets. It is also important to maintain an internal test set in order to evaluate the algorithm performance also on internal data. This may help to understand whether a tested performance on an external validation data set is reliable or spurious since it is logical to expect that the algorithm performs better on data of the same type of training data. Unlike in the fields of medicine and health, in the field of artificial intelligence and machine learning, the term validation often refers to the fine-tuning stage of model development, and another term, test, is used instead to mean the process of verifying model performance. External validation can be carried out in two different modalities: diagnostic case-control study and diagnostic cohort studies. In the first case, samples with and without the disease are collected separately and, this way, prevalence is artificially designed, unlike natural prevalence in the real-world settings. This selection bias results in a different disease spectrum and it affects the algorithm performance. In the cohort studies, patients are selected based on some predefined criteria and this allows to select natural spectrum and prevalence data. However, the choice of eligibility criteria is pivotal to represent prevalence in a good way. Another issue that deals with the clinical validation is that dividing the population in subjects who clearly have disease and subjects who clearly do not have disease (i.e., two opposite extremes in the disease and non disease spectra) would inflate diagnostic or predictive performance [111]. Robust clinical verification of the performance of a diagnostic or predictive artificial intelligence model requires external validation (validation as verification of a model's performance) in a clinical cohort that adequately represents the target patient population, and the use of prospectively collected data is desirable. In conclusions, there are many steps to pass from statistical to clinical validation and the process implies the collaboration among many hospitals and institutions to achieve a sufficient proof of the algorithm generalization capability.

2.4.1 Can really physicians improve their performance with algorithms?

Over the last 10 years, publications on AI in radiology have increased from 100–150 per year to 700–800 per year [115] and the interest in the medicine field is continuously increasing. AI and deep learning studies mainly focus their scopes in increasing the accuracy of diagnosis when compared to the physicians performances. However high accuracy does not necessarily mean that an AI algorithm improves clinical outcomes. It is, in fact, important to assess whether its use in clinical practice can be integrated in the hospital workflow and how much the impact is, not only on the outcomes, but also on the physicians training. In order to perform this kind of analysis, clinical trial modality is needed and clinical trial studies are usually time consuming and expensive. As an example, in [49], the performances of a CNN classifier for skin cancer have been compared to the dermatologists' one. They found that CNN outperforms most readers. In a later letter in which the authors characterized better the readers experience and provenance [48], it has been shown that, even if it is true that the CNN outperforms the dermatologists, this behaviour significantly depends on the reader experience and on skin cancer detection specialization. It is pivotal to question which could be the role of AI in the medical and clinical workflow, especially in the radiology field which seems to be the most explored field of medicine. It is also interesting to discuss the role of a radiologist in the hospital workflow and whether they can be replaced by an artificial intelligence or be supported by it. In [115], a group of radiologists reflects on what it means to let an AI make a diagnosis and what are the differences between the human evaluation and the AI one. AI and especially deep learning functioning in radiology is based on a principle that is very similar to the clinical one: ”the more images you see, the more examinations you report, the better you get” and this may be the reason why AI is successfully applied to radiology. Since the comparison between the radiologist's and AI performance depends on the radiologist experience and also on the quality of the developed AI, it is not straightforward to state whether and when one is better than the other. When image analysis takes too much time with respect the necessity of the patient, i.e. a very urgent clinical evaluation is necessary, AI may be very helpful in a hospital workflow. As an example, in this study [64], the application of a deep learning-based assistive technology in the Emergency Department (ED) context has been studied on Chest Radiographs (CRs). CR interpretation is a difficult task that requires both experience and expertise because various anatomical structures tend to overlap when captured on a single two-dimensional image, different diseases may have a similar presentation and specific diseases may be present with different characteristics. For these reasons, the CR interpretation suffers from a significant possibility of misinterpretation (22% according to [30]). ED physicians perform worse than trained radiologists in reading images. However, radiologists may not be available, especially during nights and weekend and CR interpretation in the ED settings is given to ED physicians. For all these reasons, [64] Kim et al. studied whether an ED physician supported by a deep-learning based algorithm for CR interpretation performs better than the single ED physician. They found that ED departments may benefit from the use of AI even if this experiment needs at least an external validation study. This is an example that shows clearly how much it is important to know the healthcare domain and practice in order to structure a deep learning experiment. Despite the improvements deep learning may produce to healthcare, another pivotal question concerns the problem of accountability. When an AI is used to make a decision in clinical practice, it is not trivial to understand who is responsible for the diagnosis. In this work [107], a radiologist supported by an AI is depicted as responsible for the diagnosis if they are trained on the use of AI since they are responsible for the actions of machines. Moreover, it is necessary to deepen the research field of explainability in order to let the radiologists understand the AI behaviour. Furthermore, the use of AI may bias the radiologist decisions. Lastly, the public discussion on the introduction of AI systems as possible substitutes of the physicians themselves can be dangerous and produce a paradox effect: since radiologists are going to be replaced by AI, there will be a lack of motivation for young doctors to pursue a career in radiology. For all these reasons, building and even bring in the public debate deep learning models to be applied to radiology is a delicate task.

2.5 The urgency of a real-time interdisciplinary approach

In this chapter, the many aspects that deal with the creation of a deep learning algorithm applied to medicine and, specifically, to imaging have been discussed. The difficulty of taking into account all these issues is clear and they relate to many fields of knowledge. In the first section, the changing scientific paradigm has been discussed as well as the problem of the hypothesis and the reproducibility. How researchers pose their research questions and which epistemological assumptions they embrace are fundamental to understand the kind of research they are doing. This process cannot be done without looking at the social processes that leads to the data collection and the data generation. In the second section, the process to define a ground truth on medical images is discussed within potentialities and limitations. Typically, the ground truth on medical images is made by the physician opinion or by a consensus among many medical doctors. When made with the second modality, the ground truth always suffers from the inter-observer variability that is difficult to be erased. The quality of an algorithm strictly depends on the quality of the ground truth but having a large number of physicians is economically expensive and requires a high grade of coordination and collaboration among research and health institutions. The quality of the algorithms depends also on the quality of data that can be private or public. Public data guarantees the possibility of testing different algorithms on the same data set but, in order to make them publishable, important information on, for example, acquisition protocols or scanners, may be lost. Private data has the advantage to be designed for the specific experiment and taken following inclusion criteria decided by the collector. When released, this kind of data can be designed to contain the information on acquisition that could be useful and meaningful for the analysis. In any case, medical images data are scarce and they may lack of label quality. This issue is one of the most limiting in deep learning algorithm development. Finally, when an algorithm is developed to be used in clinical practice, it has to be validated not only statistically but also clinically. Validation is a word that can be misunderstood since it has different meanings in medicine and in computer science field. The validation set is a specific training-dependent set for algorithm developers and the performances computed on it are not independent since the algorithm has been chosen on the basis of performance on the validation set. The test set is instead an independent set of data which is not used during training and that is taken apart to evaluate the final performance. However, the test set is not sufficient to claim clinical advancements since it belongs to the same data set used for the training and the validation. The algorithm, in fact, needs to be tested also on at least an independent external data set to evaluate its generalization capability. The external data should be taken from another medical center and should contain the information on acquisition and scanners in order to make possible the analysis of the image characteristics that may confuse the algorithm. This process can be done in two modalities, case-control and clinical trial studies, and both of them may suffer from the issues to correctly represent the population. Once the algorithm has been externally validated, it should be integrated in the hospital workflow and its performance should be evaluated also in this context. It has been established that the capacity of an algorithm to outperform a physician is strictly connected to the experience of the physician to solve that specific task. For this reason, there exist situations in which the application of an algorithm may be really helpful to both increase performance and save time. In this context, it is interesting to question who is responsible for the diagnosis when an algorithm is used to support physicians or directly to diagnose a certain disease. In order to solve this issue, we need juridical instruments that helps the application of algorithms in clinical practice. Building responsibility means also to train physicians to the use of AI in order to make them mindful of its use and to produce an informed consent that patients can really understand. All these issues suggest that the development of an algorithm for clinical applications need a deep and widespread knowledge in all of the cited fields: medicine, radiology, healthcare processes, laws, computer science, computer engineering, physics, social sciences, philosophy and so on. What is at stake is to develop a high performance and trustworthy Artificial Intelligence.

2.5.1 Z-Inspection® project

The promise of applying AI to medicine is inherently bound to the capacity to develop and deploy trust in these new instruments. Assessing trustworthy AI is a difficult task [161], in fact “the real-life ethical impact that a technology will have on people, their communities and the planet, can only be fully understood once the product or service is in real-world use” [116]. In this context, the study of applied ethics plays a central role. For this reason, the process called ”Z-inspection®” [161] has been designed to assess if an AI system is trustworthy based on the definition of trustworthy AI given by the high-level European Commission expert group on AI [53]. The Z-inspection® project is made of independent researchers who come from all over the world and its aim is to apply the process to assess ethical, social, technical and legal risks when implementing an AI. The process is based on the integration of two well-known approaches: the first one is a holistic approach, which aim is to capture the whole without considering the single parts while the second is the analytical approach, which considers instead all the parts of the problem domain. The Z-inspection® process consists of three main phases:

1. the Set Up phase: in this phase, the pre-conditions for participating in the process are verified (initial questions, absence of conflict of interest and so on) and a multidisciplinary team is chosen to have the required skills and expertise. In order to conduct an independent AI ethical assessment, the absence of conflict of interest both direct and indirect is required. Finally, the boundaries and the context are defined to delineate an ecosystem. The concept of ecosystem is particularly important in this framework and it is defined as a set of sectors and parts of society, level of social organization, and stakeholders within a political and economic context [161]. This definition takes in consideration the following hypothesis: AI is not a single element; AI is not in isolation;

AI is dependent on the domain where it is deployed; AI is part of one or more (digital) ecosystems; AI is part of processes, products, services, etc.; and AI is related to people and data.

2. the Assess phase: this phase begins with the analysis of socio-technical scenarios. Fixing the usage scenarios is useful to describe the aim of the system, the actors, their expectations and goals, the technology and the context. This analysis is carried out by relevant stakeholders, including designers (when available), domain, technical, legal and ethics experts. The analysis of the scenarios consists in the classification of the AI domain and usage, in the review of domain-specific frameworks, regulation and laws, in the development of an evidence base by analyzing and verifying the authors' claims and in making a list of potential ethical issues and tensions. The output of the analysis described above is the list of ethical issues that are called *flags*. The flags are then described and classified following the dilemmas definitions of Whittlestone et al. [153]. When some ethical issues do not fit into one or more predefined example, it can be described using free text. From this mapping, a plan of investigation is created: each issue is assigned to one of the four ethical principles, rooted in fundamental rights, that are 1) respect for human autonomy, 2) prevention of harm, 3) fairness, and 4) explicability and to one of the seven requirements established by the EU High-Level Experts Guidelines for Trustworthy AI [53]. After the mapping, at the execution stage, the group chooses a strategy to perform the inspection and defines paths to do in the system evaluation. Then, after the evaluation, the group provides feedbacks that are used to reassess ethical issues and flags. The Assess phase is repeated until a consensus is reached.

3. the Resolve phase: at this point a score is given to the system if possible. The tensions detected at the previous phase are solved when possible. The ethical issues and flags are prioritized using Whittlestone definition [153] and then the team of inspectors may give recommendations. Lastly, the needing of an ethical maintenance over time is assessed.

The Z-inspection$^{®}$ process can be applied before or even after the algorithm development and it helps to understand whether an AI system has ethical and scientific consistency. So it can be applied on a specific case and domain and the process has to consider many variables that relates to many knowledge domains. In order to perform the assessment, the context and the actors who will use the system have to be clear. In particular, actors are not only the developers or the radiologists but also the patients who should be informed and give consent to the use of AI for their disease treatments. Moreover, the context is something that have to be defined and its definition is not straightforward. Z-inspection uses the term "ecosystem" to define a context which means that stakeholders, institutions, data, analysis procedures and so on vary along time and space and, even if it can be a closed system, it is in a dynamical equilibrium among all its part. Defining a context and the stakeholders for a determined problem is a central part of the scientific hypothesis creation. Moreover, the reliability of an algorithm depends on the ecosystem on which it is applied, making its applicability to other ecosystem a very challenging horizon for the research. Another point that is interesting to stress is the relationship between algorithm's ecosystem and the improvement of clinical outcomes. This issue deals with the definition of what we intend to improve with the algorithm itself. As described in Section 2.2, training an algorithm and evaluating its performance is based on a ground truth which is usually made of physicians' opinion. This is certainly a starting crucial point in algorithm development since this modality of truth assessment suffers from an indelible variability. Despite this difficulty there exist mathematical instruments to assess the improvement in precision, accuracy and recall in, for example, the diagnosis of a certain disease. However, when the AI system is applied into the hospital workflow, its performance should be continuously evaluated. The improvements of the figures of merit due to algorithms may not correspond to an improvement in clinical practice.

During my participation to the project, we analyzed an algorithm, called BS-Net, for COVID-19 severity assessment trained on Chest X-Ray (CXR) images [136]. The BS-Net system [136] is an end-to-end AI system able to estimate the severity of damage in a COVID-19 patient's lung by assigning the corresponding Brixia score to a chest CXR image. We tried to answer the following questions:

1. Is the AI system trustworthy?
2. Is the use of this AI system trustworthy?
3. What does "trustworthy AI" in time of a global pandemic mean?

Considering the pandemic context and the final use of the algorithm is important to define the ground truth and the performance goal. In fact, the system has been developed to help tired and exhausted radiologists and doctor and hence its performance should be evaluated in this context. First, the experts have been chosen to start the evaluation of the algorithm and after that divided into working groups. Then, the socio-technical scenarios have been defined as well as the actors involved in the algorithm. Subsequently, the socio-technical scenarios have been evaluated from different points of view: the medical doctor's one, the radiologists', the technical and the legal/lawyers one. As I participated in the technical group, what we found on this algorithm is, in summary, that:

1. The data could be too small to capture the problem's variance and the algorithm needs further external validation to test its generalization capability. Data are male-biased and patient's age is skewed towards older patients. Ethnicity is dominated by Italian demographic and since further ethnic information was not collected from patients, ethnic representation could not be verified. In addition, a very limited set of device manufacturers has been used.

2. As regard the data labelling, they used the Brixia scores defined within the Brescia Hospital which does not rely on a "hard" ground truth. Moreover, the scores describe the pulmonary damage as seen on the CXRs which is not COVID-19 specific. Finally, the labelling may be biased from the fact that radiologists came from the same hospital.

3. The explanation of the algorithm could be better performed since they used a LIME-based procedure which produces high variable explanations.

After the identification of flags, the mapping has been performed and the final recommendations has been given. They concern the use of a bigger, diverse, high-quality images curated from multiple institutions and different geographic areas, the inclusion of patients in the algorithm evaluation, a detailed risk management plan and governance and so on. The Z-inspection process for this use case presents some limitations. The group which participated in the evaluation of BS-Net was made of about 60 people from many disciplinary fields and from all over the world. Despite the efforts to make an objective analysis, the evaluation is western-culture based. Keeping in mind that ethics and legal issues are always bound to a specific culture is a hypothesis which should always be considered. Moreover, the process had to consider many points of view and even if the number of people involved in the project is quite high, it can not be ensured that every point of view has been considered. During the evaluation, it has not been investigated whether and how the AI system actually influenced the radiologists routine and decision-making and both the mappings and the consolidation of the mappings involve subjective decision-making components. Assessing practical ethics is a real challenge that needs many experts and expertises, the capacity of analyzing the many involved aspects and the ability to communicate across all the cited disciplinary fields.

Chapter 3

Breast Density Classification with Convolutional Neural Networks: from Performance to Explanation Insights

3.1 Research problem

Breast cancer is the most frequently diagnosed cancer among women worldwide and it is the second leading cause of death [134]. It has been evaluated that one woman in eight is going to develop breast cancer in her life and early diagnosis is one of the most powerful instruments we have in fighting the disease [95]. Full Field Digital Mammography (FFDM) is a non-invasive highly sensitive method for early-stage breast cancer detection and diagnosis, and represents the reference imaging technique to explore the breast in a complete way [26] [144]. One of the major issues in cancer detection is due to the presence of breast dense tissue. Breast density is defined as the amount of fibroglandular parenchyma or dense tissue with respect to the fat one as seen on a mammographic exam [133]. Since x-ray absorption coefficient for dense and cancerous tissues are similar, a mammogram with a very high percentage of fibroglandular tissue is less readable. In order to have a sufficient sensitivity in denser breasts, a higher radiation dose has to be delivered to the patient [103]. Moreover, breast density is an intrinsic risk factor in developing the disease [102] [12] [145]. For these reasons, a density standard has been established by the American College of Radiology (ACR) in 2013 [133] and it is reported on the Breast Imaging Reporting and Data System (BI-RADS) Atlas. The standard defines four qualitative classes: almost entirely fatty (A), scattered areas of fibroglandular density (B), heterogeneously dense (C) and extremely dense (D). Since mammographic density assessment made by radiologists suffers from a non-negligible intra and inter-observer variability [22], some automatic methods have been developed in order to make the classification reproducible. Many approaches use a two-step classification [76] [11] [6] [110]: first, either they extract features from the images or apply a segmentation method and, afterwards, they train a classifier with a Support Vector Machine (SVM) or other machine learning methods. In [146], a fully automated algorithm has been developed: the breast is segmented, density features are extracted and used to train and evaluate SVM classifiers with an accuracy of 84.47% on the miniMIAS dataset. In [117], Petroudi et al. conceived a method based on the statistical distribution of rotationally invariant filter responses in a low dimensional space, following the Third Edition of the BI-RADS standard (1998). In the last few years, deep learning-based methods have been developed with success in a wide range of medical image analysis problems [88]. The main advantage of deep learning-based classifier stands in their capability of analyzing data from different sources automatically extracting image related features. Since features represent image properties which cannot be analytically described, they are not easily intelligible. Moreover, it is not straightforward to explain how such algorithms perform the classification. The detailed study of deep learning applications to medical images and their explainability is a challenge that can help medical physicists on tasks such as the data quality control and validation [71]. Explaining a deep learning based classifier is crucial in order to understand whether the classification is correct. In fact, since Convolutional Neural Networks are trained directly on images, it may happen that they focus their attention on uncorrelated or wrong part of the images [122], introducing a bias in the classification. Moreover, there exists no standard to quantify when an algorithm is well explained or not. The lack of huge public labelled mammographic datasets is a major issue when dealing with deep machine learning models applied to mammography, because it implies the impossibility of comparing models using the same data [71]. As an example, one of the most used public analogic datasets of mammograms, called miniMIAS [143], is labeled by three qualitative classes that are Fatty (F), Fatty-Glandular (G) and Dense-Glandular (D) which are obsolete nowadays.

In these works [93] [91] [94] [127], a residual Convolutional Neural Network (CNN) classifier was trained and a widely used explanation method to assess which are the main factors affecting the classifier performance not only in terms of accuracy but also in terms of a posteriori explanation, was applied. Both the figures of merit and the saliency maps produced by the grad-CAM algorithm have been considered and the goodness of our model has been measured by computing the Spearman's rank correlation between the input images and their saliency maps. This method has never been used on natural images which are 3-channel images; on the contrary, a mammogram can be interpreted as a result of a pre-processed signal. The classification performances have been studied considering different proportions of density class labels in the train and test datasets and different pre-processing pipelines.

3.1.1 Mammographic Density Standards

This work was born in the framework of RADIOMA project ("RADiazioni IOnizzanti in MAmmografia") [139] whose aim was to develop a personalized dose index in mammography. The assessment of breast density is important to develop in the future a personalized dose index since the amount of radiation dose depends on the quantity of dense tissue in the breasts. Medical research towards the prevention of breast cancer has shown that breast parenchymal density is a strong indicator of cancer risk. Specifically, the risk of developing breast cancer is increased only by 5% related to mutations in the genetic biomarkers BRCA 1 and 2; this risk, on the other hand, is increased by 30% for breast densities higher than 50% [13] [147]. A higher breast density is also responsible for a low sensitivity on mammograms because dense tissue has about the same absorption coefficient of cancer. Defining and sharing a classification standard is a fundamental starting point to study the correlation between a high breast density and the risk of having cancer. In 1976, Wolfe [155] empirically defined four classes of density, showing some classified mammograms and describing few features on them. Beyond controversial efficiency of this first classification method, Wolfe had the merit of laying the basis to study the effective correlation between breast density and the increase of risk in developing a cancer. Nowadays the worldwide recognized standard has been established by the American College of Radiology (ACR) and it is called BI-RADS Atlas (Breast Imaging-Reporting And Data System) [133]. These classes have been established to standardize mammographic reports in order to reduce interpretative confusion on mammograms.

In the previous BI-RADS Edition, published in 2003, the four density classes were identified with percentage indication as follow:

1. B1: It refers to lower dense breast with fibroglandular tissue less than 25%
2. B2: Class with a percentage of fibroglandular tissue between 25% and 50%
3. B3: Class with a percentage of fibroglandular tissue between 50% and 75%
4. B4: It refers to highest dense breast with fibroglandular tissue more than 75%

The first difficulty of this density standard is the definition of which pixel represents fat tissue and which one represents instead fibroglandular tissue. In fact, the high tissue variability among women and the different conditions in which mammograms can be performed make threshold methods not efficient: the pixel value in a woman that is assessed as "fat" can mean "fibroglandular" on another woman. Tissue variability is a problem not only among different women but also on the same individual over time and depends on several factors such as Body Mass Index (BMI), age, use of hormonal therapies, weight and diet. The second main problem is the lack of reproducibility [39]. Studies on inter-reader agreement with k-statistic showed a low value of accordance [22]. For these reasons, in the fifth edition of BI-RADS Atlas [133], percentage indication has been replaced with guideline based on text description of mammograms. This standard is widely used in North America and in Europe and it plays an important role in assessing the relations between breast density and cancer detectability. At the same time, automated methods for assessing density classes have been developed to overcome limitations of area-based evaluations that are subjective and time-consuming and hence not suited for large epidemiological studies. Furthermore, automated classification software makes the density assessment really reproducible. Some of these software are already available [2] and tested. The most known is CumulusV (University of Toronto), which is an interactive software to segment and to estimate breast density according to BI-RADS standard. Cumulus is not completely automatic but it is operator-dependent and this means that it suffers from the variability discussed above.

Other software for breast density assessment are available such as Volpara (Volpara Solutions) and Quantra (Hologic, Danbury, Conn) but they do not classify density in BI-RADS standard. However, these automated software have the merit of making density measures really reproducible unlike the one assessed by radiologists. The inherent problem is that we want to measure the mammographic density, which is a 3D quantity, from mammograms that are 2D projections. This kind of difficulty may be overcome defining a volumetric density standard using new breast imaging techniques such as MRI but mammography is still the most used breast imaging system all over the world.

BIRADS guidelines for breast density assessment

In the fifth edition of BI-RADS Atlas, density assessment is defined as an overall assessment of the volume of attenuating tissues in the breast. Density evaluation helps to indicate the relative possibility that a lesion could be obscured by normal tissue and that the sensitivity of examination thereby may be compromised by dense breast tissue. Since mammography does not detect all breast cancers, clinical breast examination is a complementary element of screening. The four density categories are named "A", "B", "C" and "D" and they are defined by the visually estimated content of fibroglandular-density tissue within the breasts. If breasts, on the same individual, are not apparently belonging to the same density class, the denser one should be considered in the assessment. The less dense class is "A" and breasts belonging to this class are almost entirely fatty. In this case, mammography shows the highest sensibility possible and the probability of masking effect is really low. In Figure (left) 3.1, a mammogram of an almost entirely fatty breast is reported. In the second density class "B", there are breasts with scattered areas of fibroglandular density which can not be considered as mammographic findings. In Figure 3.1 (right), a mammogram classified B is reported.

The category "C" includes heterogeneously dense breasts. It is common that some areas of breast are relatively dense while other areas are almost fat. In these cases, it is useful to describe locations of denser areas in the medical density report. In fact, in these areas, small uncalcified lesions may be obscured. Some text examples are reported in BI-RADS Atlas such as "The dense tissue is located anteriorly in both breasts, and the posterior portions are mostly fatty" or "Primarily dense tissue is located in the upper outer quadrants of both breasts; scattered areas of fibroglandular tissue are

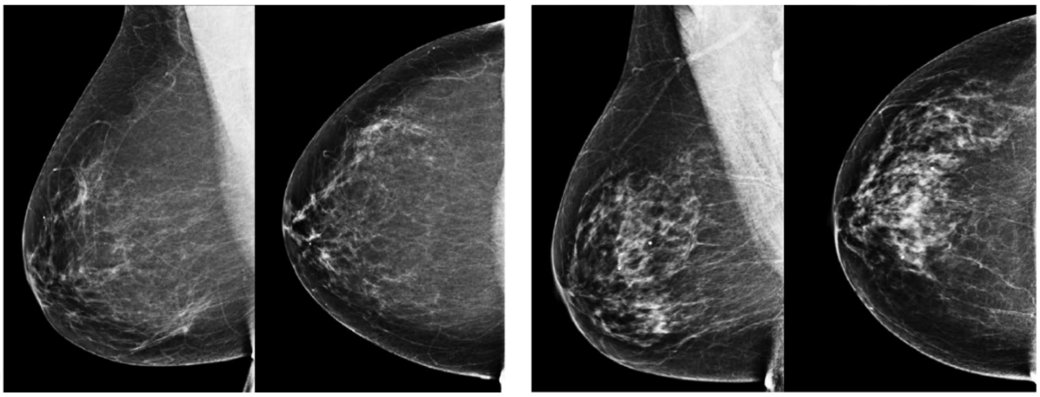

Figure 3.1: On the left: the CC and MLO projections of a class A breast; the breasts are almost entirely fatty. On the right: the CC and MLO projections of a class B breast; there are scattered areas of fibroglandular density. The images have been taken from the BI-RADS Atlas.

present in the remainder of the breasts". In Figure 3.2 (left), an example of a C-classified breast is shown. The denser class is "D" and includes breasts with such an extreme density that lowers the sensitivity of mammography. An example of dense breasts is reported in Figure 3.2 (right).

The historical empirical distribution of density classes of 3,865,070 screening mammography examinations over 13 years is reported in Figure 3.3.

The fourth edition of BI-RADS, unlike previous editions, indicated quartile ranges of percentage dense tissue (increments of 25% density) for each of the four density categories, with the expectation that the assignment of breast density would be distributed more evenly across categories than the historical distribution of 10% fatty, 40% scattered, 40% heterogeneously, and 10% extremely dense. However, it has since been demonstrated in clinical practice that there has been essentially no change in this historical distribution across density categories, despite the 2003 guidance provided in the BI-RADS Atlas.

3.2 Data

Due to the lack of public research databases populated with digital mammograms to use in AI applications devoted to density class identification [143] [80], I analyzed Full-Field Digital Mammograms (FFDM) collected within the

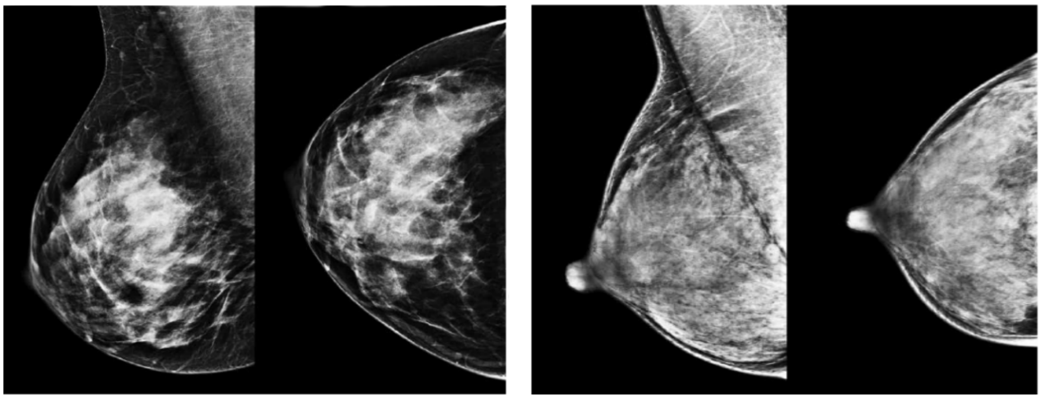

Figure 3.2: On the left: the CC and MLO projections of a class C breast; the breasts are heterogeneously dense, which may obscure small masses. On the right: the CC and MLO projections of a class D breast; the breasts with such an extreme density that lowers the sensitivity of mammography. The images have been taken from the BI-RADS Atlas.

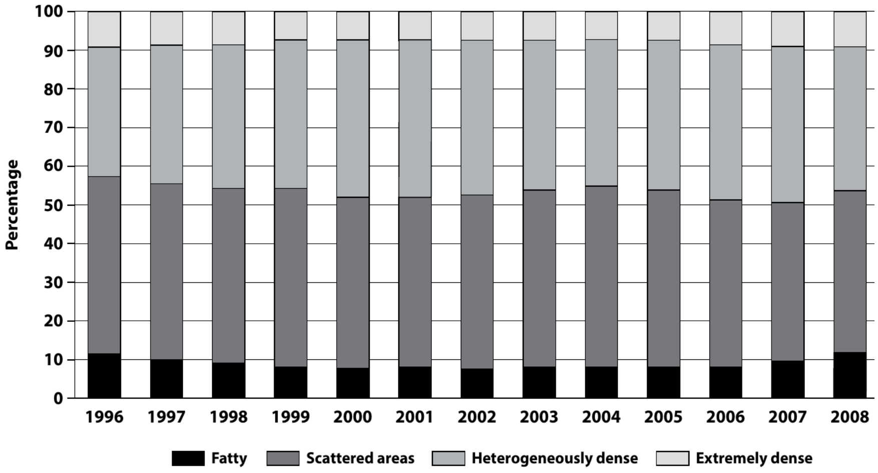

Figure 3.3: The real data distribution reported in the BIRADS Atlas made on more than 3,800,000 screening examinations made by U.S. Radiologists' Use of BI-RADS Breast Density Descriptors, 1996-2008

RADIOMA project and described in [139]. This private repository includes data from 1662 subjects (6648 images) acquired at the University Hospital of Pisa (Azienda Ospedaliero-Universitaria Pisana AOUP, Pisa, Italy). Informed consent was obtained from all the participants included in the present study. The team, which included physicists, radiologists and a radiology Technician, that worked at the data collection, has implemented and applied the following inclusion criteria to select images from the wider clinical database available:

- All exam reports were required to be negative. Whenever possible, a later mammographic exam in medical records has been examined to verify the current state of health of women.
- Badly exposed X-ray mammograms were not collected.
- Only exams including all the four projections usually acquired in mammography (cranio-caudal –CC– and medio-lateral oblique –MLO– of left and right breast) were chosen.

The exams were acquired with the GE Senograph DS imaging systems available at the University Hospital. For each exam, data annotation, which is the assessment of density class, has been performed by a radiologist with specific expertise in mammography, who relied also on the medical report already available within the routine clinical evaluation. The distribution of the 1662 exams over the 4 density classes is reported in Table 3.1, where the average age is reported for each class. As expected, both the average and median age of the cohorts of subjects increase as the breast density increase.

Table 3.1: Dataset population and age distribution (described in terms of the mean, standard deviation and median values) of the exams over the four BI-RADS density classes (A,B,C,D).

	A	**B**	**C**	**D**
N. of exams	200	473	804	185
Average age (years)	61	57	51	46
Standard Deviation (years)	11	11	9	7
Median (years)	62	55	49	45

3.3 Methods

Fully understanding a CNN behavior is a non-trivial problem and there are currently no protocols or guidelines establishing a strict and robust validation method. Furthermore, most datasets used in published works are not accessible and, hence, comparisons among different methods and algorithms are hampered. Lastly, two of the most used data sets, miniMIAS [143] and the CBIS-DDSM [80] contain only digitized analog mammograms. Using different datasets in a reproducibility test does not guarantee the achievement of the same results. For further testing the consistency of AI-based results, it is advisable to investigate which characteristics of the images, of the acquisition protocols and of the manipulation pipelines sensibly affect the performances of deep learning algorithms. Studying the roboustness of algorithms is important in order to understand the boundaries in which the classifier can be applied. Since the data set of this study has been collected from a clinical database, it is crucial to study whether and in which conditions it may be applied on a screening population. To this purpose, I trained from scratch a residual CNN to classify breast density in four categories, according to the Fifth Edition of BI-RADS standard, and systematically evaluated the impact on the CNN performance of:

- the different proportion of mammograms belonging to the four density categories in the training and test sets;
- either including or not an image pre-processing step.

The effect of the latter on the model interpretability is also studied and discussed. Finally, a simple metric to quantify the appropriateness of the chosen explainability framework is proposed. The choice of a CNN to perform the classification was due to the change of the BI-RADS classes definition from the 4^{th} Atlas to the 5^{th} one. In fact, in the last edition, the definition through the quantification criteria based on the percentage of dense and fat tissue has been abandoned and the classes were defined through image examples and textual description. In order to capture this new definition, the deep learning based methods seem to be the most appropriate.

Data preparation and pre-processing

Mammograms have been extracted from the DICOM files using DICOM-ToolKit, since they were stored in a jpeg lossless compression format. Then, the images have been converted from 12-bit to 8-bit. By visually inspecting the dataset, I found out that the images acquired have some burnt pixels which always assume the maximum grayscale value, while most of the signal is in another part of the histogram. Furthermore, having the dataset been extracted from a clinical sample instead of a screening one, clips, that are used after a biopsy, are represented in many images. These clips appear whiter than the expected maximum breast signal intensity. As the mammograms could not be normalized to the maximum intensity values, I set a maximum threshold to 3500 for the pixel value, then the image pixel values have been linearly scaled between the minimum intensity and the maximum of 3500; finally the values have been converted to 8 bits and the exams have been stored in the Portable Network Graphics (PNG) format. All the PNG images have been inspected one by one in order to eliminate some problematic images which were not correctly acquired.

Standard image pre-processing step

The GE Senograph mammograms are 1914×2294 pixel images, where the breast representation often occupies about half of the image width. To limit the data processing time (i.e. to minimize the number of input nodes of the CNN and thus the weights to be learned during the training process) I decided to crop the images according to the minimum bounding box enclosing the breast view. To this purpose, I attempted to recognize the skin line of the breast using a marching-square algorithm for 2D images [152][101], available within the scikit-learn Python package [113]. To properly identify the breast margin, the starting threshold has been set at the intensity level of 50 while leaving the other parameters to default values; then images has been cropped to the minimum bounding box including the margin, as shown in Figure 3.4.

Additional pre-processing step: pectoral muscle removal

As an additional pre-processing step, an algorithm to remove the pectoral muscle that appears on medio-lateral oblique projections has been designed. First, all the medio-lateral oblique projections have been oriented in the same way, i.e. left ones have been flipped horizontally. Then images have been cropped at the half of height and width in order to obtain a square which contains the pectoral muscle. A Gaussian filter has been applied to all the selected regions in order to reduce noise ($\sigma = 1.1$ as computed by cv2 for a

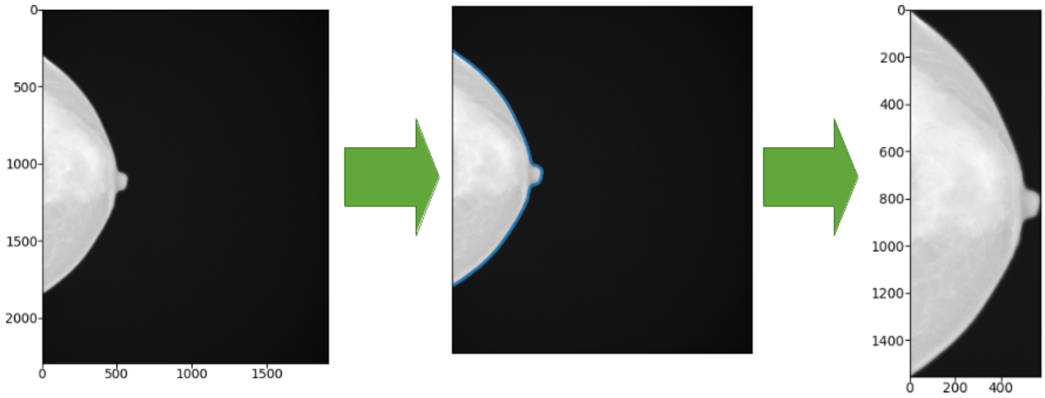

Figure 3.4: Left: original image. Center: the blue line shows the contour identified by the marching-square algorithm. Right: cropped image according to the bounding box enclosing the breast view.

kernel size equal to 5x5). For each image, the regions have been binarized with an adaptive threshold method based on inverted binary thresholding and Otsu's binarization and the mask containing the pectoral muscle (white) and the rest of the breast (black) have been produced. The coordinates of the points at the edge of the pectoral muscle have been fitted with a linear function and the values of all the pixels above the edge have been replaced with the mean gray level of the breast. In Figure 3.5, an example of these operations is reported .

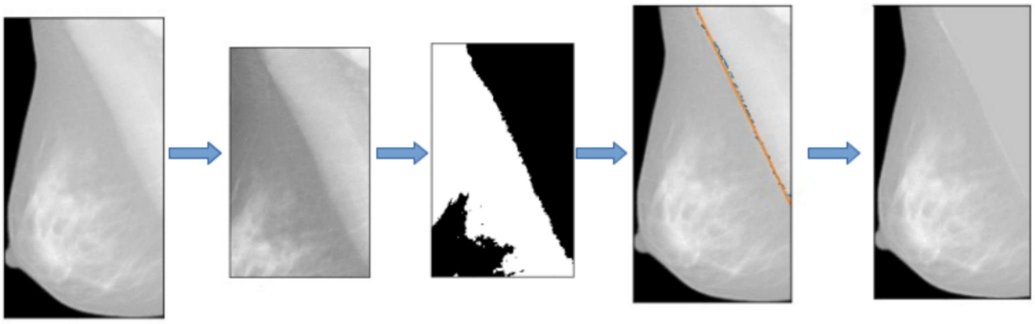

Figure 3.5: Main steps describing the pectoral muscle segmentation pipeline.

This procedure works for the 80% of the images of our dataset. In Figure 3.6, there is an example of a mammogram on which the pectoral muscle segmentation did not achieve a good result. On problematic images, segmentations have been manually fixed, being the robustness of the pectoral muscle segmentation algorithm not one of the main objectives of this work.

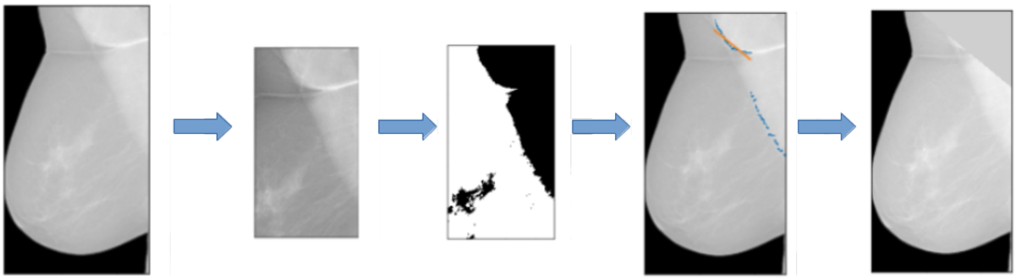

Figure 3.6: Example of a mammogram on which the pectoral muscle segmentation did not work properly. The algorithm considers the very first points of the muscle and, as a result, the segmentation does not include the muscle below.

Data augmentation for CNN training

The last step of the pre-processing of images for the CNN consists in data augmentation [114]. In fact, although our dataset contains about 6600 images, this amount may not be sufficient to avoid overfitting and to achieve, at the same time, good performances in terms of accuracy [132]. I used the Keras built-in class ImageDataGenerator which applies random transformations to the input data at runtime. The chosen transformations are:

- random zoom in a range of 0.2;
- width shift in a range of 0.2 of the whole input image;
- height shift in a range of 0.2 of the whole input image;
- random rotations with a range of 10 degrees.

Classifier training

In order to train, fit and evaluate the CNN, Keras -a Python API- with Tensorflow in backend [21] has been used. I implemented a model based on a very deep residual convolutional neural network [51]. The architecture of our model [93] was made of 41 convolutional layers, organized in residual blocks, and it had about 2 millions learnable parameters. The input block consists of a convolutional layer, a batch normalization layer [57], a leakyReLU as activation function and a 2D-max pooling. The output of this block has been fed into a series of four blocks, each made of 3 residual modules. In Figure 3.7, the architecture of one of the four block is shown.

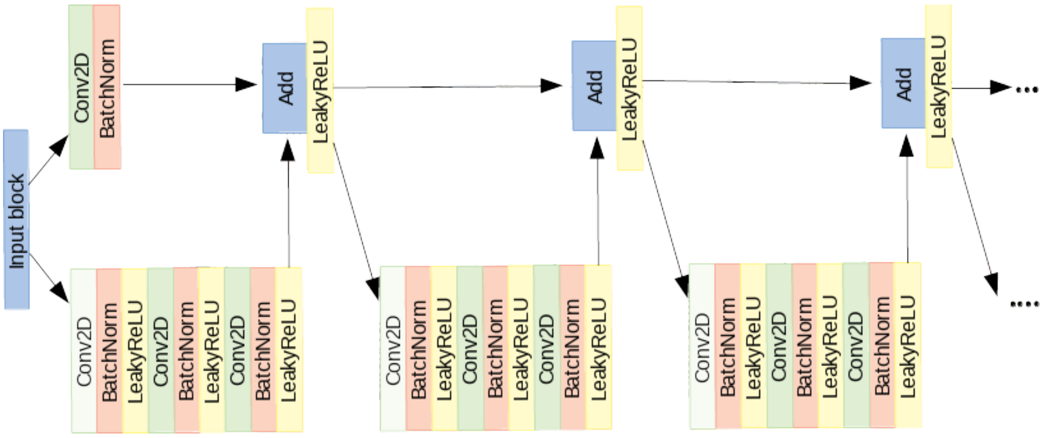

Figure 3.7: One of the four blocks made of 3 residual modules.

The input of each of the four blocks is shared by two branches: in the first, it passes through several convolutional, batch normalization, activation and max pooling layers while in the other branch it passes through a convolutional layer and a batch normalization. The outputs of these two branches are then added together to constitute the residual block [51]. The sum goes through a non-linear activation function and the result passes through two identical modules. The architecture of the left branch of these last modules is the same as the first one. In the right branch, instead, no operation is performed. At the exit of the module, the two branches are summed together. At the end of the network, the output of the last block is fed to a global average pooling and to a fully-connected layer with softmax as activation function. Data have been split randomly into training set (80%), validation set (10%) and test set (10%). To evaluate the performance on the test set, the accuracy, the recall and the precision has been computed as figures of merit. They are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

where TP is the number of true-positive, TN the number of true-negative, FP the number of false-positive and FN the number of false-negative detections. Moreover, the Cohen coefficient has been computed in order to compare our results with others published in literature. The Cohen coefficient, also called Cohen's kappa, is a more robust measure than simple percent agreement calculation, as it takes into account the possibility of the agreement occurring by chance. The Cohen's kappa is defined as follows:

$$K = \frac{p_o - p_e}{1 - p_e} \tag{3.4}$$

where p_o is the observed agreement and p_e is the hypothetical probability of agreement by chance. The CNN has been trained for 100 epochs and the reported results refer to the epoch with the best validation accuracy. The best selected model has been evaluated also in terms of Kappa coefficient to measure the accordance with the physician evaluation. The main hyperparameters are:

- 41 convolutional layers organized in 12 similar blocks;
- training performed in batches of 4 images;
- Loss function: Categorical Cross-Entropy;
- Optimizer: Stochastic Gradient Descent (SGD);
- Regularization: Batch Normalization;
- Learning rate = 0.1, Decay = 0.1, Patience = 15, Monitor = validation loss.

In order to consider all the four projections related to a subject, four CNNs have been separately trained on each projection, on a K80 Nvidia

GPU. Finally, the classification scores (i.e. the CNN output) have been averaged separately for right and left breast and, in case of asymmetry, the higher class has been assigned since breast density is an overall evaluation of the projections and, in clinical practice, the radiologist assigns the higher class to subjects with density asymmetry.

Model explanation

I aimed to characterize the models in a transparent modality and wanted to create an explanation framework for the outcome of a deep CNN to identify which pixels and salient regions in the image influence the most the final prediction. This was done through off-line visualization techniques, which means analysing an already trained model without altering its architecture. I used the *visualizecam* utility function, provided by Keras, to generate a gradient based class activation map that maximizes the outputs of filters within a specified layer and returns an image indicating the regions of the input whose changes would most contribute towards maximizing the output. This function implements a way of visualizing attention over input, which is known as grad-CAM. The basic idea of class activation mapping technique is to identify the importance of image regions by projecting back the weights of the output layer onto the convolutional feature maps. A weighted sum of the feature maps of the last convolutional layer is computed to obtain class activation maps. Grad-CAM uses the gradient information flowing into the last convolutional layer of the CNN to assign importance values to each neuron for a particular decision of interest. In order to obtain the class-discriminative localization map Grad-CAM $L^c_{Grad-CAM} \in R^{u \times v}$ of width u and height v for any class c, I first compute the gradient of the score for class c, y^c (before the softmax), with respect to feature map activations A^k of a convolutional layer, i.e. $\frac{\partial y^c}{\partial A^k}$. These gradients flowing back are global-average-pooled over the width and height dimensions (indexed by i and j respectively) to obtain the neuron importance weights α^c_k:

$$\alpha^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}. \tag{3.5}$$

This weight highlights the ‘importance’ of feature map k for a target class c. Then, a weighted combination of forward activation maps, followed by a

ReLU results in:

$$L^c_{Grad-CAM} = ReLU(\sum_k \alpha^c_k A^k). \tag{3.6}$$

To sum up, the places where the gradient is large allow us to define the region that has a large impact on the final score decision.

Evaluation of the explanation framework

There is no standard procedure to quantify the quality of the saliency maps. The grad-CAM algorithm is usually used to visually assess the correctness of the classification. This means that it is used as an observer-dependent measure. The heatmaps produced by the grad-CAM highlights the most important part of the image the classifier looks at when it performs the classification. Since the breast density classification is an intensity-based classification, I propose to quantify if the highlighted regions in the heatmap correspond to the denser regions in the original image through the Spearman correlation. In fact, it is possible to directly study the correlation between the mammograms and the saliency maps in order to quantify at least whether there is a monotonic dependence between the images and their explanation. For this reason, I computed the Spearman's rank correlation between the pre-processed images, which actually contain the information strictly related to the breast density provided in input to the CNN, and their relative saliency maps. Since mammograms are gray-scaled the Spearman's rank correlation has been computed between the pixel intensities and the gray-scaled map intensity values to test whether they are in an increasing monotonic relationship, as expected. The value for the perfect increasingly monotonic relationship between two variables is 1.

3.4 Results

3.4.1 Evaluation of the effect of sample composition on CNN training

The CNN model has been trained with different class distributions in order to understand whether it is possible to use the maximum available number of images and how much the probability distribution of classes affects the results. Three different distributions have been considered: the native one of the dataset (A: 12%, B:29%, C:48%, D:11%), which is the distribution of the classes in the original dataset as collected from the AOUP; the BI-RADS one (A: 10%, B:40%, C:40%, D:10%), the density class distribution provided in the BI-RADS Atlas; and a uniform one (A: 25%, B:25%, C:25%, D:25%), i.e. a distribution including the same proportion of the four density classes. The CNN was trained and tested on samples with these three different distributions of class labels. In Tables 3.2 the performance metrics results are shown.

Table 3.2: Final results of CNN trained on different training set and tested on different test sets.

		AOUP Test set	BI-RADS Test set	Uniform Test set
BI-RADS Training set	test accuracy (%)	79.1	83.1	73.6
	recall (%)	75.2	80.1	73.6
	precision (%)	82.6	87.9	79.0
AOUP Training set	test accuracy (%)	78.5	79.7	73.6
	recall (%)	74.2	77.9	73.6
	precision (%)	81.2	83.0	79.4
Uniform Training set	test accuracy (%)	72.8	72.9	77.8
	recall (%)	78.9	79.9	77.8
	precision (%)	69.5	68.8	78.0

From Table 3.2, it can be observed that the best accuracy, precision and recall in the classification are achieved by training the CNN on the BI-RADS distribution of samples (A: 10%, B:40%, C:40%, D:10%) and testing it on the same BI-RADS distribution. Moreover, this distribution is the closest to the real data distribution. In fact, it is the one reported in the BI-RADS Atlas made on more than 3,800,000 screening examinations and so it is the most representative of what we can observe in clinical practice. Moreover, our dataset includes a small number of mammograms of class D. This means that we should have a dataset with a very small total size to have a uniform distribution, and this size is too small to train our deep network. The best performance is achieved when the classifier is trained on a set of images with the BI-RADS distributions of classes and tested on a set with the same distribution. Although maintaining the proportion among classes reported in the BI-RADS Atlas, representative of the screening practice, forced us to use a reduced dataset size, this did not penalize the results. However, training the network on a dataset with a uniform distribution of the classes, therefore with an even smaller size, gives worse results. I conclude that the dataset size does affect the obtained results and the probability distribution of classes is an influencing factor as well.

3.4.2 Implementation and visual assessment of the grad-CAM technique

The heatmaps obtained through the grad-CAM technique have been used to establish if the classifier effectively makes its predictions based on the presence of dense areas in the mammogram. This fits into the more general purpose of assessing trust in predictions from our algorithm. The heatmap evaluation has been done qualitatively, which means by visually estimating if the highlighted regions in the heatmap correspond to the denser regions in the original image. The analysis consisted in visualizing and comparing the maps generated using the input images of the four classes. The maps have been produced for all the images in the test set and for all the four projections constituting the mammographic exam. In Figure 3.8 an example of a comparison of the heatmaps of the four density classes obtained from a model trained on right cranio-caudal projections is reported.

The activated regions in the maps match reasonably well with the dense regions in the original mammogram for B, C and D classes. The grad-CAMs prove that the "attention" of the classifier is focused on the dense region as expected. An important remark resulting from analyzing all the maps is that for class A mammograms the active area is almost always at the edge of the breast. This is reasonable because the A class is the one corresponding to the lowest density and it seems like the classifier, not recognizing any dense region, focuses its attention on a different feature, such as the edge.

3.4.3 Evaluation of the impact of pectoral muscle removal

The CNN model has been trained with and without the pectoral muscle. The images with the pectoral muscle removed have been obtained after applying the algorithm described in the Methods section. This algorithm was efficient

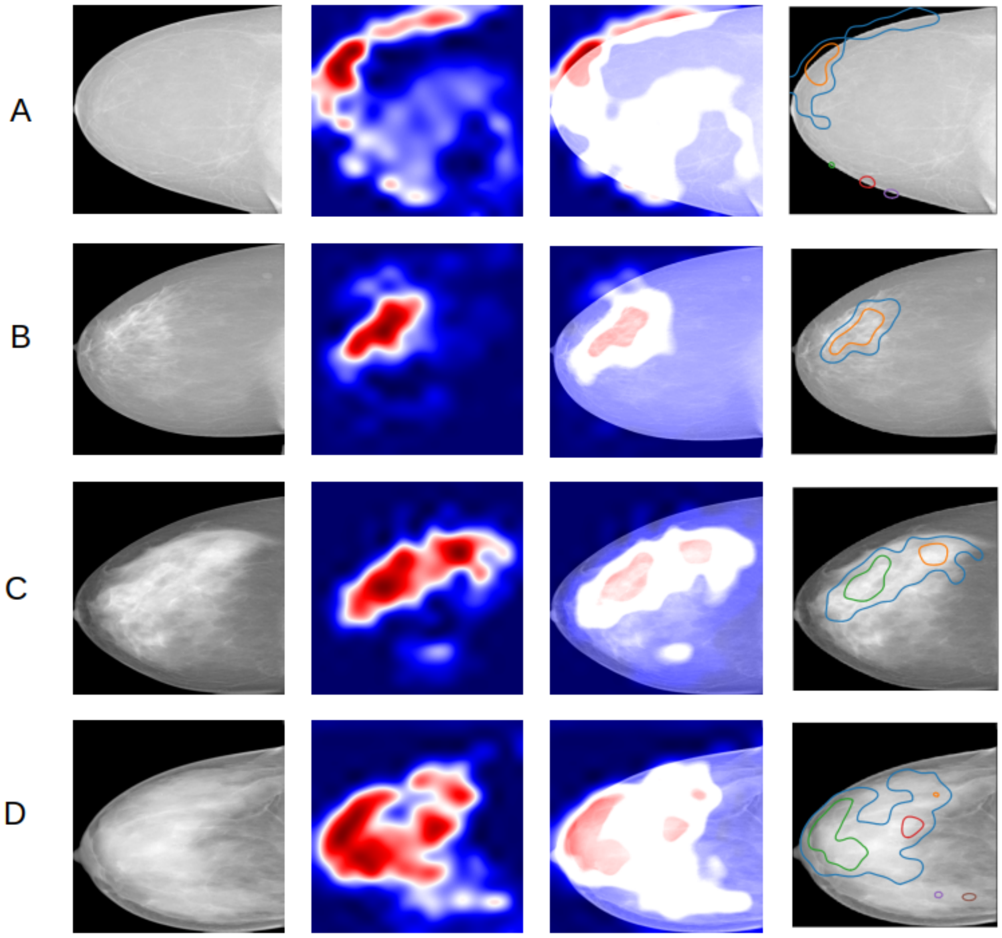

Figure 3.8: Comparison of heatmaps of the four density classes A, B, C, D with one example per class, obtained from a CNN trained on right CC projections. From left to right, the input image, the grad-CAM, the overlay of the map on the input image, the overlay of edges of red activated areas in the map on the input image.

on 80% of the available exams. The 20% of exams on which the segmentation algorithm failed, i.e. the muscle edge was not correctly identified in at least one projection, were manually segmented. By grad-CAM visualization, it can be noticed that, for some MLO projection images, the related maps activate at the pectoral muscle visible in these projections. I then trained the CNN on MLO projections with the muscle removed, to check if in this case the classifier performance and the heatmaps improve. In terms of performance metrics, training the model with and without the pectoral muscle gave the results reported in Table 3.3.

Table 3.3: Performances of the CNN trained with images with (with PM) and without the pectoral muscle (without PM)

	Precision	**Recall**	**Accuracy**
with PM	81.1%	78.1%	79.9%
without PM	83.3%	80.3%	82.0%

Grad-CAM maps have been generated in the two cases and they have been compared. In most cases, muscle removal helps in guiding the network to focus on the right breast area and after segmentation the pixels forming part of the muscle are no longer highlighted and activated (Figure 3.9). Therefore, segmentation and removal of pectoral muscle in the image pre-processing phase help in the performance improvement.

3.4.4 Quantitative evaluation of the explanation framework

From a visual inspection of a number of examples, including those shown in Figure 3.9, it seems that to predict the breast density category the CNN is actually “looking” at the appropriate image information, namely the higher-intensity regions of the mammograms. A possibility could be to quantitatively compare the area of the saliency map over a predefined threshold and a hand-crafted pixel-wise ground truth for dense areas generated by a radiologist. That would be an extremely time-consuming task; thus I discarded this option and proposed a straightforward method to evaluate whether the maps and the higher-density breast areas are spatially correlated. Moreover, for this classification task it is not fair to use pixel wise ground truth since the class assessment by physicians is made by observing the entire image and not pixel wise, i.e. breast density assessment is a classification task and not a segmentation one. To quantify the extent of this hypothesized direct relationship, I computed the Spearman’s rank correlation coefficient **r** between

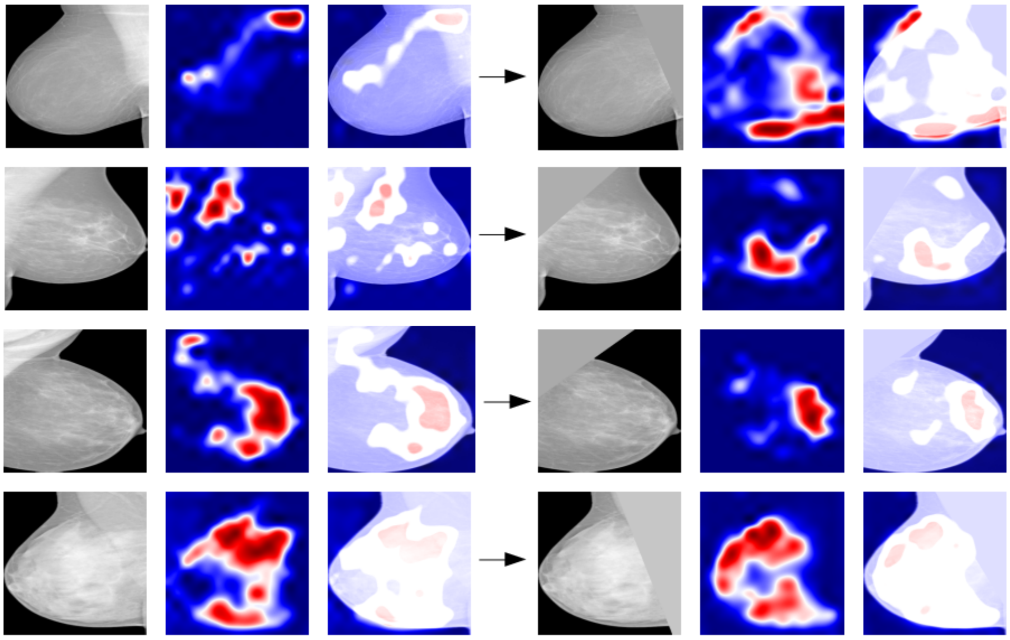

Figure 3.9: Example of comparison between grad-CAM maps obtained with the original image (on the left) and with the segmented image (on the right) for various density classes.

the grey-scale image and the grad-CAM map. The box plots obtained for the correctly classified mammograms in the test set of the four density categories are shown in Figure 3.10 separately for the CC and MLO projections.

The Kruskal–Wallis test [74], which is a non-parametric ANOVA test, has been performed to measure whether there is a significant difference in the Spearman's rank correlations among the four classes. I obtained a p-value less than 0.05 for the tests made on for CC and MLO projections separately. Hence, I can affirm that there is a significant difference among the classes. However, the Kruskal–Wallis test does not state if all the groups are significantly different. For this reason, the Dunn test [32] has been performed with correction for multiple comparisons, which is the post hoc analysis for Kruskal–Wallis test. The Dunn test showed a significant difference among A, B and C classes, while this is not true for the D class (Table 3.4).

From the boxplots, it can be noticed that high median values of $\mathbf{r}$ are generally obtained on mammograms belonging to higher density classes. For

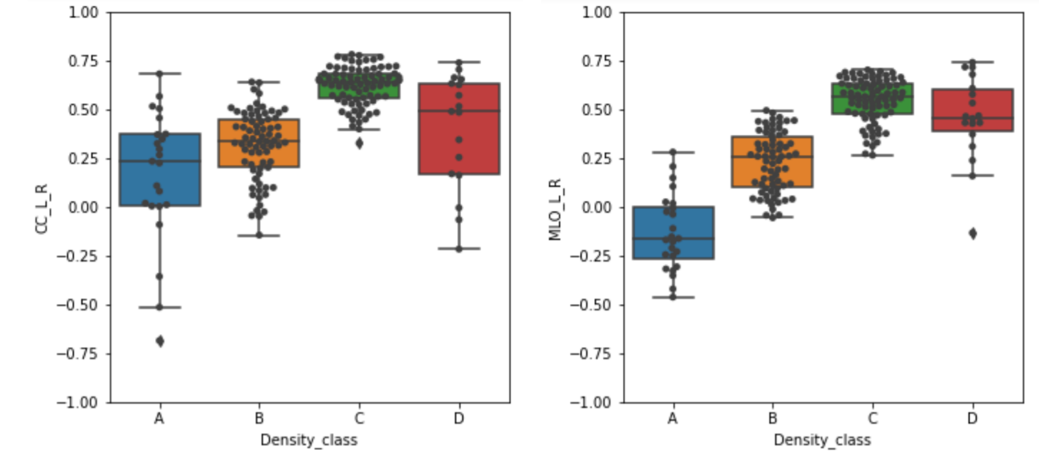

Figure 3.10: Box plots of the Spearman's rank correlation coefficients **r** obtained for the correctly classified mammograms across the four density categories. The correlation values obtained on the left and right cranio-caudal (CC_L_R) and left and right medio-lateral-oblique (MLO_L_R) projections are separately shown. (The box plots are centered on the median and the boxes represent the interquartile range.)

Table 3.4: Top: results of the Dunn test, corrected for multiple comparisons, computed on CC projections. Bottom: results of the Dunn test, corrected for multiple comparison, computed on MLO projections.

	A	**B**	**C**	**D**
A	1	$p = 0.43$	$p < 0.05$	$p = 0.12$
B	$p = 0.43$	1	$p < 0.05$	$p = 0.16$
C	$p < 0.05$	$p < 0.05$	1	$p < 0.05$
D	$p = 0.12$	0.16	$p < 0.05$	1

	A	**B**	**C**	**D**
A	1	$p < 0.05$	$p < 0.05$	$p < 0.05$
B	$p < 0.05$	1	$p < 0.05$	$p < 0.05$
C	$p < 0.05$	$p < 0.05$	1	$p = 0.20$
D	$p < 0.05$	$p < 0.05$	$p = 0.20$	1

the C class the correlation values show the highest median value and the most compact distribution, thus indicating that the CNN classifier is actually considering the higher density areas as the ones to take into account to assign the mammogram to a breast density category. For mammograms of the D class (i.e., those with higher density) this is not always true. The grad-CAM is not in general activated consistently with the higher-density areas of the breast, as depicted in the mammograms. However, the large spread of the **r** values for this category hampers drawing generalized conclusions. For mammograms of the B category, a positive median r value still indicates a systematic overlap between the higher-intensity areas of the grad-CAM maps and the mammograms. By contrast, the situation is controversial for the mammograms belonging to the A category. In that case, the median **r** value for CC projection is positive (about 0.25), thus suggesting a systematic overlap between the higher-intensity areas of maps and mammograms, whereas the median **r** value for MLO projection is negative (about -0.20), thus indicating an opposite relationship. Namely, as visible in the line corresponding to the A example of Figure 3.9, the grad-CAM map activates in the breast areas complementary to the high-intensity ones. The hypothesized direct monotonic relationship between the pixel intensity values between the original pre-processed breast mammograms and the saliency maps is thus verified in most cases, namely for the higher-density categories (B, C and D) with median **r** values above 0.25. For the lower-density A category, the behavior of the CNN seems instead to be different in the interpretation of CC and MLO mammograms, exploiting, in the latter case, the complementary density information.

3.5 Discussion

As regards the comparison with the previous classifier [93] where no pre-processing was implemented, I obtained better results in terms of the figures of merit and activation maps. In fact, the CNN reaches an accuracy of 82%. Moreover, the CNN compares very well with the literature [81] (where an accuracy of 77% is obtained) with a classifier trained on about 60,000 exams. Compared to [6, 11], the CNN based classifier achieves better performances in terms of accuracy (respectively 47% and 71%). As regards the study by Oliver et al. [110], our classifier works better also in terms of Cohen (Kappa) coefficient on the four classes problem, since the presented algorithm reaches a K equal to 0.76 with respect to 0.67. Other studies [40, 104] reach better accuracy on the classification of two classes: dense versus non-dense and BI-RADS 2 versus BI-RADS 3, respectively. It is, hence, not possible to compare our results with theirs. In [125], the accuracy on the 4 BI-RADS classes is equal to 99% which is higher than the accuracy reached in this work. However, their method is not explained and not explainable. As a general consideration, the comparison with other different methods is not performed on the same dataset, making it not completely fair. Since our breast density classifier was trained on digital mammograms, the method cannot be applied to MIAS and DDSM which contain analog digitized mammograms. Moreover, analog mammography is not used in hospitals anymore. This represents one of the issues described in the previous chapter. In fact, health instruments change as the technology of mammographic systems, in this case, changes. This issue makes the comparison with other studies harder and suggests the needing for curated and updated data sets. Moreover, it is interesting to notice that our classifier has been trained on data coming from the clinical routine: public data sets are usually made of cleaned data which rarely represents the real prevalence of radiological findings. On the other hand, since our data set has been collected from a clinical database it suffers from the problem of not representing a true screening population.

I found that pre-processing has a crucial impact not only on the accuracy, but also on the explainability of the classifier. In fact, the grad-CAM activation maps showed a good localization capability once the pectoral muscle has been removed from the image. For this reason, I believe that CNN classifiers should be trained on medical images which comes from hospitals and screening or clinical routines, paying particular attention not only to the classification performances but also to obtaining reasonable activation maps. In fact, in order to make the classifier understandable for physicians and patients, it should show a good behaviour not only in terms of accuracy, precision and recall but also in selecting the right part of the image.

As regards the training on different sample compositions, the discussion on the more appropriate strategy to be used in training ML algorithms in case of unbalanced data set is highly debated [60]. Both the balanced and the natural distribution approach can be actually used [151]. Even if training a classifier on a balanced data set can ensure a better performance evaluation, real data sets are not balanced at all. I found out that the CNN performs better on the BI-RADS distribution in terms of accuracy, precision and recall. This distribution is the closest to the natural one reported on the BI-RADS Atlas. This result was not unexpected as the native distribution is strongly unbalanced over the density classes, while the uniform one forces us to use far fewer images than the other two. I then visualized the saliency maps computed on the test set to check whether the classifier is looking at the dense part of the breast to perform the classification. I found out that for the medio-lateral oblique projections the saliency maps highlighted more the regions of the pectoral muscle than the dense parenchyma. For this reason, the pectoral muscle has been segmented and the classifier retrained . Then, I compared the saliency maps obtained with and without the muscle and found out that segmentation helps in identifying the correct dense region as shown in Figure 3.9. Furthermore, the performance in terms of figures of merit increases for the classifier trained with the segmentation. Finally, I computed the Spearman's rank correlation to assess whether the pre-processed images and the relative saliency maps are in a direct monotonic relationship. The Kruskal-Wallis test and the Dunn test, which is its post hoc test, have been computed. The tests have been performed to confirm the trend highlighted in the boxplots of Figure 3.10. I found out a correlation for the B, C and D classes while I obtained a controversial result for the A class. The visual inspection of saliency maps and Spearman's rank correlation computed for different classes show a mutual accordance with our hypothesis. I underline that it is important to evaluate both visually and quantitatively the maps to reach an optimal performance. The main drawbacks of our work are the use of a single mammographic system, a ground truth made by only one radiologist and the use of a clinical dataset instead of a screening one. Moreover, the algorithm should be tested also on an external data set.

3.6 Conclusion

In this study, I presented a detailed study of a CNN trained on mammograms in an explainable way. I trained a CNN classifier on a wide set of clinical mammograms to classify them according to breast density and then I implemented an explanation algorithm to explore the CNN behavior on different input data. The CNN performance has been evaluated using different distributions of class labels in the training and test sets, and different preprocessing steps, taking into account the accuracy, precision and recall figures of merit, and the saliency maps obtained with the grad-CAM algorithm. This approach can be extended to other medical images in the attempt to provide clinicians with reliable and explainable AI-based decision support tools.

Chapter 4

Fully Automated DL-based Algorithm for Segmentation of Lungs and COVID-19 Lesions

4.1 Research problem

More than 340 million of cases of SARS-CoV-2 over the world have been registered since the beginning of the pandemic; the virus is affecting more than 200 countries and caused the death of more than 5 million people by January 2022 [1].

Computed tomography (CT) has a high sensitivity in the identification of lung lesions, including those related (but not-specific) to COVID-19 pneumonia. It has a key role in monitoring the clinical course of patients and in the evaluation of disease severity. The extent of lung involvement in the disease has been shown to be predictive of the patients' need of intensive care unit support [25, 36]. Thus, the quantification of the extent of abnormal lung tissue with respect to the subject's whole pulmonary volume is a fundamental information for the management of the emergency due to the pandemic. To this purpose, a standardized assessment scheme for the reporting of radiological findings in chest CT of subjects suspected of COVID-19 has been defined [120]. It is based on a five-level scale of increasing suspicion of pulmonary involvement. Another scoring system, directly based on the extent of lung involvement is the CT Severity Score (CT-SS), which has been demonstrated to be directly correlated with disease severity [158]. The estimation of the percentage P of affected lung parenchyma is used to assign a CT-SS score to a chest CT scan: CT-SS=1 for P<5%, CT-SS=2 for 5% $\leq$ P<25%, CT-SS=3 for 25% $\leq$ P<50%, CT-SS=4 for 50% $\leq$ P<75%, CT-SS=5 for P $\geq$ 75%.

However, the mere visual assessment of lung CT can hardly provide a reliable and reproducible estimate of the percentage of lung involvement. To facilitate this task, an Artificial Intelligence (AI)-based support tool is highly desirable. The quantification problem that needs to be solved is actually a segmentation problem. To estimate the percentage of the affected lung in COVID-19 pneumonia it is necessary to accurately segment both the subject's lungs and the COVID-19 related lesions.

The task of lung segmentation has been addressed over the years with several different techniques, including grey-value thresholding, region growing, isosurface triangulation, morphological operations, and combinations of them [9, 28, 42, 27, 16]. However, most traditional approaches fail when abnormalities introduce changes in the normal lung density [150], especially in the specific case where abnormalities are adjacent to the pleura surface. The latter is actually the case of most CT of subjects with COVID-19 lesions. Traditional medical image segmentation methods have gradually given way to data-driven approaches mainly based on Machine Learning (ML) and Deep Learning (DL) in the specific field of thoracic imaging [149] and in medical image analysis in general [129]. U-nets [123] are currently outperforming other AI-based methods in the image segmentation task in many research fields. They are also becoming widespread in medical imaging to identify organs, lesions and other regions of interest across several imaging modalities [58, 85]. The main drawback of DL approaches to image segmentation is their need of large annotated datasets for training the models. Collecting data and reliable annotations is particularly difficult and time-consuming especially for image segmentation tasks, where pixel/voxel-level ground truth is required. DL-based lung segmentation approaches demonstrated to be efficient in the accurate identification of lung parenchyma even in case of compromised lung appearance due to COVID-19 infection [156], or to Chronic Obstructive Pulmonary Disease (COPD) [55], or to any routine clinical condition affecting the lungs [54]. The challenging task of lung lobe segmentation is tackled in the paper by Xie *at al.* [156], where the transfer learning of a model trained on thousands of subjects with COPD was applied on a sample of hundreds of subjects affected by COVID-19 pneumonia. Lobe segmentation reference was acquired for all subjects, as it is a fundamental information for model train, test and evaluation. Such large and annotated data samples are not publicly available at present.

The task of segmenting the abnormalities of the lung parenchyma related to COVID-19 infection is a typical segmentation problem that can be addressed with methods based on DL. CT findings of patients with COVID-19 infection may include bilateral distribution of ground-glass opacifications (GGO), consolidations, crazy-paving patterns, reversed halo sign and vascular enlargement [18]. Due to the extremely heterogeneous appearance of COVID-19 lesions in density, textural pattern, global shape and location in the lung, an analytical approach is definitely hard to code, whereas it is preferable to learn directly from examples. The potential of DL-based segmentation approaches is particularly suited in this case, provided that a sufficient number of annotated examples are available for model training.

Few fully automated software tools for the segmentation of COVID-19 lung abnormalities and quantification of lung involvement have been recently proposed [83, 36, 98]. The approach proposed by Lessmann *et al.* [83] for lesion segmentation is based on a U-net model trained on semi-automatically annotated COVID-19 cases. Then, the authors combined the output of this system with the lung lobe segmentation algorithm reported in Xie *et al.* [156]. The approach proposed in Fang *et al.* [36] implements the automated lung segmentation method provided in the work of Hofmanninger *et al.* [54], together with a lesion segmentation strategy based on multiscale feature extraction [37].

The specific problem related to the development of fully automated DL-based segmentation strategies with limited annotated data samples has been explicitly tackled by Ma *et al.* [98]. The authors studied how to train and evaluate a DL-based system for lung and COVID-19 lesion segmentation on poorly populated samples of CT scans. They also made the data publicly available, allowing for a fair comparison with their system.

In this work, a DL-based fully automated system to segment both lungs and lesions associated with COVID-19 pneumonia, the *LungQuant* system, is presented which provides the part of lung volume compromised by the infection. It is an extension of the study proposed by Ma *et al.* [98] with a focus on the investigation and the discussion on the impact of using different datasets and different labeling styles. Data can be highly variable in terms of acquisition protocols and machines when they are gathered from different sources. This poses a serious problem of dependence of the segmentation performances on the training sample characteristics. Despite advanced data harmonisation strategies could mitigate this problem [38], this approach is not applicable in absence of data acquisition information, as it is in this study for the available CT data. Nevertheless, DL methods, when trained with sufficiently large samples of heterogeneous data, can acquire the desired generalization ability by themselves. In this analysis, I implemented an inter-sample cross-validation method to train, test and evaluate the generalization ability of the *LungQuant* DL-based segmentation pipeline across the different available datasets. Finally, the effect of using larger datasets to train, validate and test this kind of algorithm has been quantified too.

This chapter is structured as follows: I list all the publicly accessible data samples used to develop and validate the *LungQuant* system; then, the image analysis pipeline is described along with the training and cross-validation strategies adopted; finally, I show and discuss the quantification performance either against a voxel-wise ground truth or in terms of the CT severity scores, according to the information available for each data sample. Finally, I present the further improvements added in a second version of LungQuant, *LungQuant*2, and briefly describe the on-going research on inter-reader variability of the CTSS, the pseudo-clinical validation of *LungQuant*2 and the future work on radiomics that are going to be performed .

4.2 Data

In this study, only public available data sets have been used to train and evaluate the segmentation pipeline. Five different data sets containing a variable number of cases and annotations have been used. Most of them include image annotations, but each annotation has been associated to patients using different criteria, which are described in the following sections. In Table 4.1, a summary of available labels for each data set is reported. The lung segmentation problem has been tackled using a wide representation of the population and three different data sets: the Plethora, the Lung CT Segmentation Challenge and a subset of the MosMed data set (detailed description below). On the other hand, the number of samples that are publicly available for COVID-19 infection segmentation may not be sufficient to obtain good performances on this task. The currently available data, provided along with infection annotations, have been labelled following different guidelines and released in NIfTI format. They do not contain complete acquisition and population information and they have been stored according

Table 4.1: A summary of the datasets used in this study. The CT Severity Score (CT-SS) information is not available for all datasets, but it can be computed for data which has both lung masks and ground-glass opacification (GGO) masks, and for the MosMed dataset, which provides a similar scale of severity, as reported in Table 4.2.

Dataset name	Lung mask	GGO mask	CT-SS	N. of cases
Plethora [65]	Yes	No	No	402
Lung CT Segmentation Challenge [157]	Yes	No	No	60
COVID-19 Challenge [3]	No	Yes	No	199
MosMed [106]	No	No	No	1110
MosMed (annotated subsample)	No	Yes	Inferable	50
MosMed (in-house annotated subsample)	Yes	No	No	91
COVID-19-CT-Seg [98]	Yes	Yes	Inferable	10

to different criteria. Some of the choices made during the DICOM to NIfTI conversion may strongly affect the quality of data. For example, the MosMed data set as described by Morozov *et al.* [106] preserves only one slice out of ten during this conversion. This operation results in a significantly loss of resolution along z axis with respect to the COVID-19 Challenge data set. Questioning how much such conversion influences the quantitative analysis is important to improve not only the performance but also the possibility of comparing DL algorithm in a fair modality.

4.2.1 The Plethora dataset

The PleThora dataset [65] is a chest CT scan collection with thoracic volume and pleural effusion segmentations, delineated on 402 CT studies of the Non-Small Cell Lung Cancer (NSCLC) radiomics dataset, available through the The Cancer Imaging Archive (TCIA) repository [23]. This dataset has been made publicly available to facilitate improvements of the automatic segmentation of lung cavities, which is typically the initial step in the development of automated or semi-automated algorithms for chest CT analysis. In fact, automatic lung identification struggles to perform consistently in subjects with lung diseases. The PleThora lung annotations have been produced with a U-net based algorithm trained on chest CT of subjects without cancer, manually corrected by a medical student and revised by a radiation oncologist or a radiologist.

4.2.2 The 2017 Lung CT Segmentation Challenge dataset

The Lung CT Segmentation Challenge (LCTSC) dataset consists of CT scans of 60 patients, acquired from 3 different institutions and made publicly available in the context of the 2017 Lung CT Segmentation Challenge [157]. Since the aim of this challenge was to foster the development of auto-segmentation methods for organs at risk in radiotherapy, the lung annotations followed the RTOG 1106 contouring atlas.

4.2.3 The 2020 COVID-19 Lung CT Lesion Segmentation Challenge dataset

The 2020 COVID-19 Lung CT Lesion Segmentation Challenge dataset (COVID-19 Challenge) is a public dataset consisting of unenhanced chest CT scans of 199 patients with positive RT-PCR for SARS-CoV-2 [3]. Each CT is accompanied with the ground truth annotations for COVID-19 lesions. Data has been provided in NIfTI format by The Multi-national NIH Consortium for CT AI in COVID-19 via the NCI TCIA public website [23]. Annotations have been made using a COVID-19 lesion segmentation model provided by NVIDIA, which takes a full CT chest volume and produces pixel-wise segmentations of COVID-19 lesions. These segmentations have been adjusted manually by a board of certified radiologists, in order to give 3D consistency to lesion masks. The dataset annotation was made possible through the joint work of Children's National Hospital, NVIDIA and National Institutes of Health for the COVID-19-20 Lung CT Lesion Segmentation Grand Challenge.

The dataset and the annotations have been made available in the context of a MICCAI-endorsed international challenge which had the aim to compare AI-based approaches to automated segmentation of COVID-19 lung lesions.

4.2.4 The MosMed dataset

MosMed [106] is a COVID-19 chest CT dataset collected by the Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department. It includes CT studies taken from 1110 patients. Each study is represented by one series of images reconstructed into soft tissue mediastinal window. MosMed provides 5 labeled categories, based on the percentage of lung parenchyma affected by COVID-19 lesions. The 5 categories of lung involvement and their correspondence to the CT-SS scale are described in Table 4.2. The first category (CT-0) contains cases with normal lung tissue and no CT-signs of viral pneumonia, whereas the other categories contain GGO (CT-1 and CT-2) and both GGO and regions of consolidation in the higher classes (CT-3 and CT-4).

Table 4.2: MosMed severity categories defined on the basis of the percentage P of lung volume affected by COVID-19 lesions. The correspondence to the CT-SS scale is reported.

MosMed CT category	N. of cases	Percentage P of involved lung parenchyma	Corresponding CT-SS
0	254	$P = 0$	0
1	684	$0 < P \leq 25$	1, 2
2	125	$25 < P \leq 50$	3
3	45	$50 < P \leq 75$	4
4	2	$75 < P \leq 100$	5

A small subset of class CT-1 cases (50 patients) had been annotated by expert radiologists with the support of MedSeg software (2020 Artificial Intelligence AS). The annotations consist of binary masks in which white voxels represent both ground-glass opacifications and consolidations. Both CT scans and annotations were provided in NIfTI format. During the DICOM-to-NIfTI conversion process, only one slice out of ten was preserved and, as a result, MosMed CT scans have a reduced total number of slices with respect to the other datasets.

Generation of a set of reference lung segmentation for model training

As reported in Table 4.1, the available datasets with lung mask annotations, which were necessary to train the U-net for lung segmentation, are mainly of subjects affected by lung cancer (Plethora and LCTSC datasets). To complement this sample with subjects without lesions, and, at the same time, to expose to U-net to the acquisition characteristics of the MosMed CT scans, the lung mask annotations for a subset of subjects of the CT-0 MosMed category has been generated, i.e. subjects without COVID-19 lesions.

An in-house lung segmentation algorithm was developed for this purpose and implemented in *matlab* (The MathWorks, Inc.). It is based on the following steps: 1) CT windowing in the [-1000,1000] HU range; 2) rough segmentation of the lungs on a central coronal slice (Otsu binary thresholding and removal of components connected with the image border) to define the minimum and maximum axial coordinates of the lung region; 3) 2D rough segmentation of the lungs on each axial slice (same procedure as the previous step) to generate a 3D seed mask for the following step; 4) segmentation of the lung parenchyma by an active contour model (*activecontour* matlab function); 5) filling holes (e.g. vessels and airway walls) with 3D morphological operators (*imclose* matlab function). Out of the 254 CT scans belonging to the CT-0 MosMed sample, the 91 CT scan considered here are those on which the in-house segmentation algorithm provided an accurate segmentation, as judged by an experienced medical imaging data analyst. This algorithm, which accurately segments the lung parenchyma in absence of lesions, has very limited performance on CT scans of subjects with COVID-19 lesions.

4.2.5 The COVID-19-CT-Seg dataset

The COVID-19-CT-Seg dataset is a collection of CT scans taken from the Coronacases Initiative and Radiopaedia [98]. It contains 20 CT scans tested positive for COVID-19 infection. This public dataset contains both lung and infection annotations. The ground truth has been made in three steps: first, junior radiologists (1-5 years of experience) delineated lungs and infections annotations, then two radiologists (5-10 years of experience) refined the labels and finally the annotations have been verified and optimized by a senior radiologist (more than 10 years of experience in chest radiology). The annotations have been produced with the ITK-SNAP software. Ten CT images of this dataset were provided in 8-bit depth, therefore, I decided to not use them.

4.3 Methods

4.3.1 *LungQuant*: the Deep-Learning based quantification analysis pipeline

The analysis pipeline, which is hereafter referred to as the *LungQuant* system, provides in output the lung and COVID-19 infection segmentation masks, the percentage P of lung volume affected by COVID-19 lesions and the corresponding CT-SS (CT-SS=1 for P<5%, CT-SS=2 for 5% $\leq$ P<25%, CT-SS=3 for 25% $\leq$ P<50%, CT-SS=4 for 50% $\leq$ P<75%, CT-SS=5 for P $\geq$ 75%).

A summary of our image analysis pipeline is reported in Fig. 4.1. The central analysis module is a U-net for image segmentation [123], which is implemented in a cascade of two different U-nets: the first network, U-net$_1$, is trained to segment the lung and the second one, U-net$_2$, is trained to segment the COVID lesions in the CT scans. In the following sections, the whole process is described step by step.

U-net

For both lung and COVID-19 lesion segmentation, a fully automated method inspired by the U-net developed by Ronneberger *et al.* [123] has been implemented. U-nets are fully-convolutional neural networks for image segmentation. I implemented a U-net using Keras [21], a Python deep-learning API that uses Tensorflow as backend. In Figure 4.2 a simplified scheme of our U-net is reported.

Each block of layers in the compression path (left) is made by 3 convolutional layers, ReLu activation functions and instance normalization layers. The input of each block is added to the block output in order to implement a residual connection. In the decompression path (right), one convolutional layer has been replaced by a de-convolutional layer to upsample the images to the input size. In the last layer of the U-nets, a softmax is applied to the final feature map and then the loss is computed.

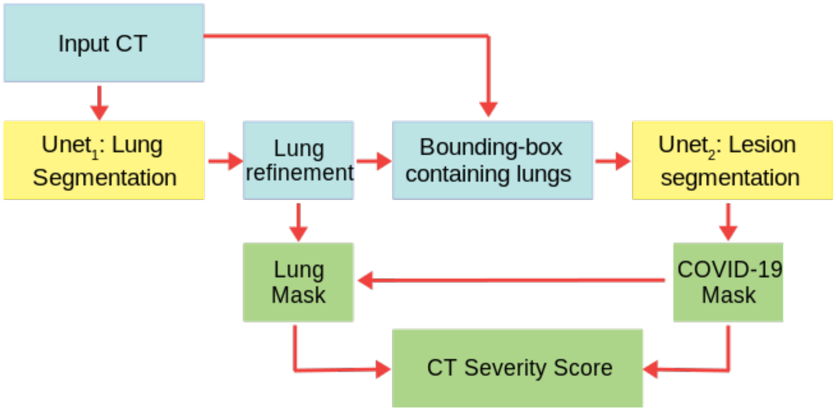
Input CT
Unet$_1$: Lung Segmentation
Lung refinement
Bounding-box containing lungs
Unet$_2$: Lesion segmentation
Lung Mask
COVID-19 Mask
CT Severity Score

Figure 4.1: A summary of the whole analysis pipeline: the input CT scans are used to train U-net$_1$, which is devoted to lung segmentation; its output is refined by a morphology-based method. A bounding box containing the segmented lungs is made and applied to all CT scans for training U-net$_2$, which is devoted to COVID-19 lesion segmentation. Finally, the output of U-net$_2$ is the definitive COVID-19 lesion mask, whereas the definitive lung mask is obtained as the union between the outputs of U-net$_1$ and U-net$_2$. The ratio between the COVID-19 lesion mask and the lung mask provides the CT-SS for each patient.

The U-net cascade for lesion quantification and severity score assignment

I started by training U-net$_1$, which is devoted to lung segmentation, using the three datasets containing original CT scans and lung masks (see Table. 4.1). The input CT scans, whose number of slices is highly variable, are oriented to canonical direction and resampled to matrices of 200x150x100 voxels to match the size of the U-net input layer. The output of U-net$_1$ was then refined using a connected-component labeling strategy, which helps to remove small regions of the segmented mask not connected with the main objects identified as the lungs. A bounding box enclosing the morphologically refined segmented lungs has been built for each CT, adding a conservative padding of 2.5 cm. The bounding boxes were used to crop the training images for U-net$_2$, which has the same architecture as U-net$_1$. The cropped images were resized to a matrix of 200x150x100 voxels to match the size of the U-net input layer.

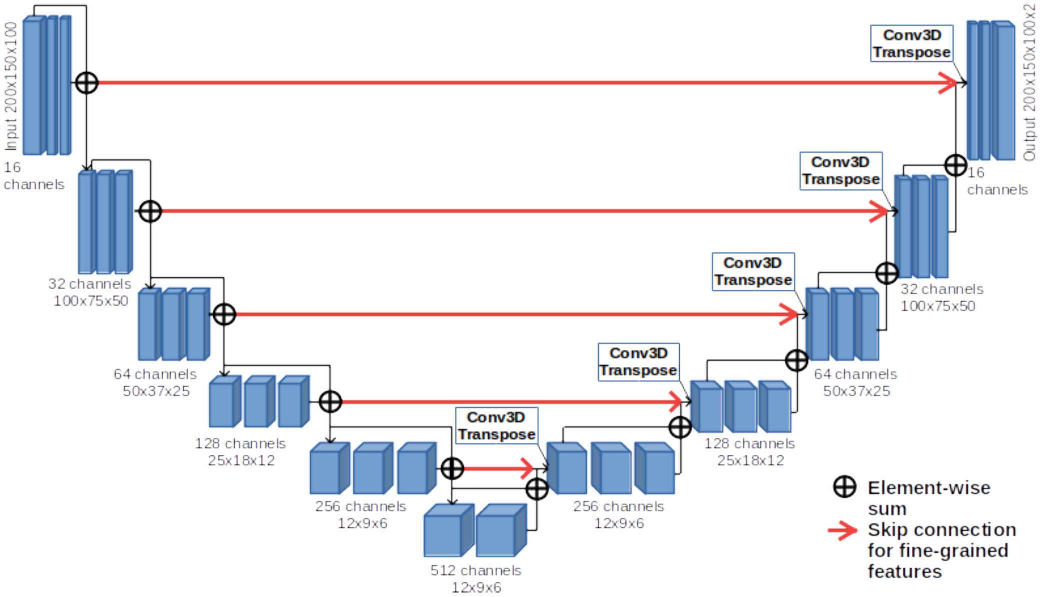

Figure 4.2: U-net scheme: the neural network is made of 6 levels of depth. In the compression path (left), the input is processed through convolutions, activation layers (ReLu) and instance normalization layers, while in the decompression one (right), in addition to those already mentioned, 3D Transpose Convolution (de-convolution) layers are also introduced.

A windowing has been applied on the grey-level values of the CT scans to optimize the image contrast for the two segmentation problems. In particular, I selected the [-1000, 1000] HU window range for the U-net$_1$ and the [-1000, 300] HU range for U-net$_2$. The first window highlights the contrast between the lung parenchyma and the surrounding tissues, whereas the second one enhances the heterogeneous structure of the lung abnormalities related to the COVID-19 infection.

To overcome the fact that the amount of data with COVID-19 lesion annotations is rather limited (see Table. 4.1), and optimize the training phase of the U-net$_2$, a data augmentation strategy has been implemented, relying on the most commonly used data augmentation techniques for DL.

The quantification system developed returns the infection mask as the output of U-net$_2$ and the lung mask as the union between the output of U-net$_1$ and U-net$_2$. This choice has been made *a priori* by design, as U-net$_1$ has been trained to segment the lungs relying on the available annotated data, which are almost totally of patients not affected by COVID-19 pneumonia. Thus, U-net$_1$ is expected to be unable to accurately segment the areas affected by GGO or consolidations; as also these areas are part of the lungs, they should be instead included in the mask. Training U-net$_2$ to recognize the COVID-19 lesions on a conservative bounding box containing only the lungs has two main advantages: it allows to restrict the action volume of the U-net to the region where the lung parenchyma (either normal or affected by COVID-19 lesions) is supposed to be, thus avoiding false-positive findings outside the chest; it facilitates the U-net training phase, as the dimensions of the lungs of different patients are normalized, thus the U-net learning process can be focused on the textural patterns characterizing the COVID-19 lesions.

Finally, once lung and lesion masks have been identified, the *LungQuant* system computes the percentage of lung volume affected by COVID-19 lesions as the ratio between the total number of voxels of the infection mask and the total number of voxels of the lung mask. The system also converts these percentage values into the corresponding CT severity scores.

4.3.2 Training details and evaluation strategy for the U-nets

Training a deep neural network means defining many variables and elements, such as the metrics to be used for model training and validation, the data-splitting strategy between train, validation and test sets, the eventual need of relying on data augmentation strategies. The latter is pivotal in our implementation due to the limited amount of annotated data samples of COVID-19 lesions. All these ingredients are detailed below.

Loss functions and evaluation metrics

U-net$_1$ has been trained with the volumetric Dice Similarity Coefficient (vDSC) as loss function, while U-net$_2$ has been trained using the sum of the vDSC and a weighted cross-entropy as error function in order to balance the number of voxels representing lesions and the background. The vDSC is defined as follows:

$$\mathrm{vDSC}_{loss} = 1 - \frac{2 \cdot |M_{true} \cap M_{pred}|}{|M_{true}| + |M_{pred}|} \tag{4.1}$$

where M_{true} is the true mask, M_{pred} is the predicted mask and the operator $|\cdot|$ is the cardinality. The vDSC loss has been computed only on the foreground (white voxels). This strategy has been used in order to avoid giving excessive weight to the background (black voxels), since the number of black and white voxels is quite unbalanced in favor of the former. For U-net$_2$, a loss function (L) has been used and it consists in the sum of the vDSC and a weighted cross-entropy (CE), defined as follows:

$$L = Dice_{loss} + CE_{weighted} \tag{4.2}$$

$$CE_{weighted} = w(x) \sum_{x \in \Omega} log(M_{true}(x) \cdot M_{pred}(x)) \tag{4.3}$$

where $w(x)$ is the weight map which takes into account the frequency of white voxels, x is the current sample and Ω is the training set. Since the background class is larger than the foreground class on the order 10^3, the weight map $w(x)$ has been computed for each ground-truth segmentation to increase the relevance of the underrepresented class, following the approach implemented by Phan *et al.* [118]. The weight map was defined as $w(x) = w_0/f_j$ where f_j is the average number of voxels of the j^{th} class over the entire training data set ($j = 0, 1$) and w_0 is the the average between the frequencies f_j.

The segmentation performances for both U-nets have been evaluated with the vDSC, computed between the true mask volume (V_{true}) and the predicted mask volume ($V_{predict}$), and with the surface Dice Similarity Coefficient (sDSC), computed between the true surface (S_{true}), and the predicted one defined, ($S_{predict}$) [66], as follows;

$$\text{vDSC}_{metric} = \frac{2 \cdot |V_{true} \cap V_{predict}|}{|V_{true}| + |V_{pred}|} \tag{4.4}$$

$$\text{sDSC}_{metric} = \frac{2 \cdot |S_{true} \cap S_{predict}|}{|S_{true}| + |S_{pred}|} \tag{4.5}$$

The surface metric has been introduced because vDSC inflates as the volume to be segmented is large.

Cross-validation strategy

To train, validate and test the performances of each of the two U-nets, I divided the available datasets into the training, validation and test sets, and the network performance has been evaluated separately and globally on the datasets. The U-net for lung segmentation, U-net$_1$, has been trained and evaluated on CT scans coming from three different datasets: Plethora, MosMed and LCTSC. The U-net for COVID-19 lesion segmentation, U-net$_2$, has been trained and evaluated on samples made of CT scans coming from the COVID-19-Challenge dataset and from the MosMed dataset.

The amount of CT scans used for train, validation and test sets for each U-net is reported in Table 4.3. U-net$_2$ has been trained twice, i.e. on both 60% and 90% of the CT scans of COVID-19-Challenge and Mosmed datasets to investigate the effect of maximizing training set size on the system's ability to properly segment COVID-19 lesions. In the former case, U-net$_2^{60\%}$ training has been evaluated on a validation set made of 20% of cases and tested on the remaining 20%. As regard the latter, U-net$_2^{90\%}$, the remaining 10% of CT scans has been used as validation set.

The trained segmentation networks (U-net$_1$ and both U-net$_2^{60\%}$ and U-net$_2^{90\%}$) have been validated on an external independent validation set consisting of the 10 CT scans of the COVID-19-CT-Seg dataset. The latter is the only public available dataset that contains both lung and infection mask annotations.

Table 4.3: Number of CT scans assigned to the train, validation (val) and test sets used during the training and performance assessment of the U-net$_1$ and the U-net$_2$ networks.

U-net$_1$	train	val	test
Plethora	319	40	40
MosMed (91 CT-0)	55	18	18
LCTSC	36	12	12
COVID-19-CT-Seg	/	/	10
U-net$_2^{60\%}$	train (60%)	val (20%)	test
COVID-19 Challenge	119	40	40
MosMed (50 CT-1)	30	10	10
COVID-19-CT-Seg	/	/	10
U-net$_2^{90\%}$	train (90%)	val (10%)	test
COVID-19 Challenge	179	20	/
MosMed (50 CT-1)	45	5	/
COVID-19-CT-Seg	/	/	10

The global quantification pipeline, the *LungQuant* system, has been set up by integrating all analysis modules, as reported in Fig. 4.1. In this work I built and analyzed two *LungQuant* systems, obtained by integrating alternately U-net$_2^{60\%}$ or U-net$_2^{90\%}$ into the analysis pipeline. The systems have been evaluated in terms of the ability to predict the percentage of affected lung parenchyma and CT-SS on the fully annotated COVID-19-CT-Seg dataset, which is completely independent.

Data augmentation

Data augmentation is a strategy to increase the size of the training set by synthetically generating additional training images through geometric transformations. This technique is particularly important to improve the generalization capability of the model, especially in the case of a limited number of training samples. In this work, data augmentation has been applied during the data pre-processing phase (after defining the bounding boxes enclosing the segmented lungs) in order to generate a fixed number of augmented images for each original sample. I chose an augmentation factor equal to 2 which means that the number of artificially generated images is twice the number of the original training set. For each image in the training set, two of the following geometric transformations were randomly chosen:

- Zooming. The CT image and the ground truth masks were zoomed in the axial plane, using a third-order spline interpolation and the k-nearest neighbors method, respectively. The zooming factor was randomly chosen among the following values: 1.05, 1.1, 1.15, 1.2.

- Rotation. The CT image and the ground truth mask were rotated in the axial plane, using a third-order spline interpolation and the k-nearest neighbors method, respectively. The rotation angle was randomly sampled among the following values: -15°, -10°, -5°, 5°, 10°, 15°.

- Gaussian noise. An array of noise terms randomly drawn from a normal distribution was added to the original CT image. For each image, the mean of the Gaussian distribution was randomly sampled in the [-400, 200] HU range and the standard deviation randomly chosen among 3 values: 25, 50, 75 HU.

- Elastic deformation. An elastic distortion was applied to the original 3D CT and mask arrays following the approach of Simard *et al.* [137].

This transformation has two parameters: the elasticity coefficient which has been fixed to 12 and the scaling factor, fixed to 1000.

- Motion blurring. Slice by slice, the CT image has been convolved with a linear kernel (i.e. ones along the central row and zero elsewhere for a matrix of size $k \times k$) through the function filter2D, defined in the OpenCV Python library [14], keeping the output image size the same as the input image. The filter is applied with a kernel size of 4, 3 and 3 in the anterior-posterior, latero-lateral and cranio-caudal direction, respectively.

An example of the application of these augmentation techniques to one CT scan of the dataset is provided in Fig. 4.3.

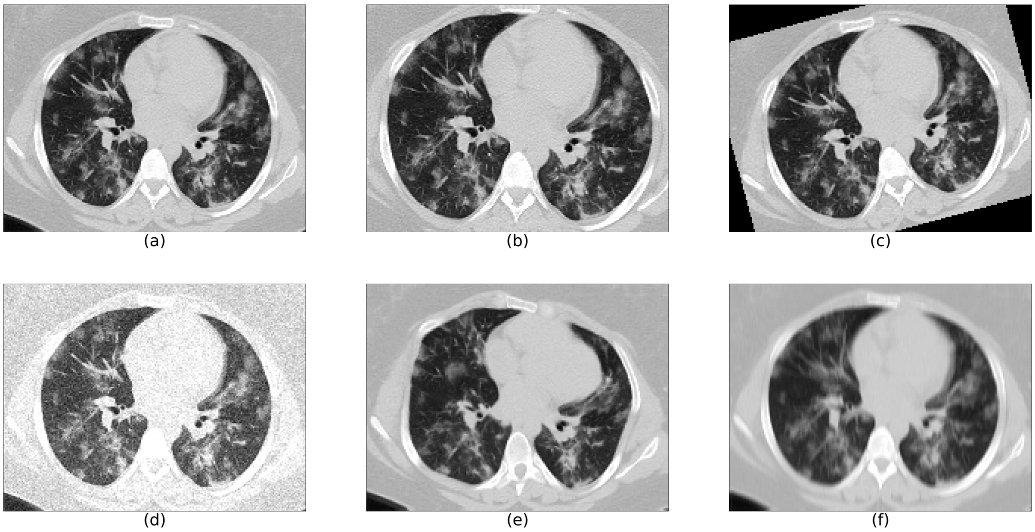

Figure 4.3: Data augmentation to increase the diversity of dataset: a) Image without data augmentation; b) Zooming; c) Rotation; d) Gaussian noise; e) Elastic deformations; f) Motion blurring.

4.4 Results

I report in this section, first, the performance achieved by each of the segmentation networks trained, U-net$_1$ and U-net$_2$, then, the quantification performance of the integrated *LungQuant* system, evaluated on completely independent test sets. Both U-nets have been trained for 300 epochs on a NVIDIA V100 GPU using ADAM as optimizer and the training has been stopped at the epoch with the best evaluation metric on the validation set.

4.4.1 U-net$_1$: Lung segmentation performance

The U-net$_1$ for lung segmentation was trained using three different datasets, as specified in Table 4.3: the Plethora, a subsample of 91 CT-0 cases of the MosMed dataset and the 60 CT scans of the LCTSC datasets. For the MosMed dataset, as reported in Table 4.1, the lung mask annotations were provided by an in-house developed segmentation software.

The data has been split randomly in training set, validation set and test set as described in Table 4.3. The learning process has been stopped at the epoch where the best vDSC metric was obtained on the validation set. Then, I tested U-net$_1$ on each of the three independent test sets, and reported in Table 4.4 the performance achieved in terms of vDSC values computed between the segmented and the reference masks. In order to remove false-positive regions (*i.e.* voxels misclassified as lung parts), at first, the connected components in the lung masks generated by U-net$_1$ has been identified, then, those components whose number of voxels was below an empirically-fixed threshold has been excluded. This threshold was set to the 40% of the foreground mask, and it was reduced to 30% whether the resulting number of voxels was found to be lower than the 65% of the initial mask provided by U-net$_1$. Figure 4.4 shows some examples of how this procedure works on real CT scans.

The lung segmentation performances have been evaluated in three cases: 1) on CT scans and masks resized to the 200x150x100 voxel array size needed match the U-net input/output layer size; 2) on CT scans and masks in the original size before undergoing the morphological refinement step; 3) on CT scans and masks in the original size and after the morphological refinement. Even if segmentation refinement has a small effect on vDSC, as shown in Table 4.4, it is a fundamental step to allow the definition of precise bounding boxes enclosing the lungs, and thus to facilitate the U-net$_2$ learning process.

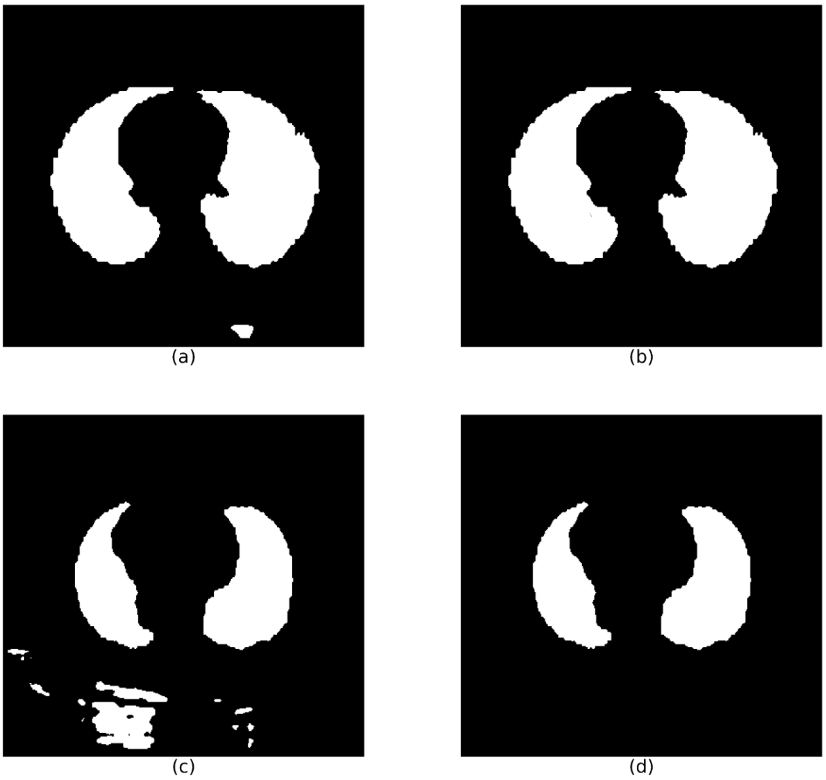

Figure 4.4: Morphological refinement of the U-net$_1$ output: a) and c) lung masks as generated by U-net$_1$; b) and d) refined masks after the connected component selection.

Table 4.4: Performances achieved by U-net$_1$ in lung segmentation on different test sets, evaluated in terms of the Dice metric at three successive stages of the segmentation procedure.

Test set	Masks of U-net size (vDSC)	Masks before refinement (vDSC)	Masks after refinement (vDSC)
Plethora	0.96 ± 0.02	0.95 ± 0.02	0.95 ± 0.04
MosMed	0.97 ± 0.02	0.97 ± 0.02	0.97 ± 0.02
LCTSC	0.96 ± 0.03	0.95 ± 0.03	0.96 ± 0.01
Coronacases	0.96 ± 0.01	0.95 ± 0.01	0.95 ± 0.01

4.4.2 U-net$_2$: COVID-19 lesion segmentation performance

The U-net$_2$ network devoted to COVID-19 lesion segmentation, has been trained and evaluated separately on the COVID-19-Challenge dataset and on the annotated subset of the MosMed dataset, following the train/validation/test partitioning reported in Table 4.3.

The segmentation performances achieved on the test sets are reported in terms of the vDSC metric in Table 4.5. As reported in the table, the performances of U-net$_2$ were evaluated also according to a cross-sample validation scheme.

Table 4.5: Performances achieved by U-net$_2$ in COVID-19 lesion segmentation, evaluated in terms of the Dice metric. The composition of the train and test sets is reported in Table 4.3.

U-net	Trained on	Test set	U-net size (vDSC)	Original CT size (vDSC)
U-net$_2^{60\%}$	COVID-19 Challenge	COVID-19 challenge	0.51 ± 0.24	0.51 ± 0.25
	COVID-19 Challenge	MosMed	0.39 ± 0.19	0.40 ± 0.19
	MosMed	MosMed	0.54 ± 0.22	0.55 ± 0.22
	MosMed	COVID-19 challenge	0.25 ± 0.23	0.25 ± 0.23
	COVID-19 challenge + MosMed	COVID-19 challenge + MosMed	0.49 ± 0.21	0.50 ± 0.21
U-net$_2^{90\%}$	COVID-19 Challenge + MosMed	COVID-19 Challenge + MosMed	0.64 ± 0.23	0.65 ± 0.23

As expected, the U-net$_2$ performances are higher when both the training set and independent test sets belong to the same data cohort. By contrast, when a U-net$_2$ is trained on COVID-19-Challenge data and tested on Mosmed (and the other way around) performances significantly decrease. This effect is due to the fact that the two datasets have been collected and annotated with different criteria and from different sources. A better result has been obtained with the U-net$_2$ trained on the COVID-19 Challenge dataset and tested on the MosMed test set, since the network has been trained on a larger data sample and hence it has a higher generalization capability. The best segmentation performances have been obtained by the U-net$_2$ trained using the 90% of the available data, U-net$_2^{90\%}$, which reaches a vDSC value of 0.65 ± 0.23 on the test set. This result suggests the need to train U-net models on the largest possible data samples in order to achieve higher segmentation performance.

4.4.3 Evaluation of the quantification performance of the *LungQuant* system

Evaluation of lung and COVID-19 lesion segmentations

Once the two U-nets have been trained and the whole analysis pipeline has been integrated into the *LungQuant* system, it has been tested on a completely independent set (COVID-19-CT-Seg dataset) of CT scans. The performances of the whole process were quantified both in terms of vDSC and sDSC with tolerance values of 1, 5 and 10 mm (Table 4.6). A very good overlap between the predicted and reference lung masks is observable in terms of vDSC, whereas the sDSC values are highly dependent on tolerance values, ranging from moderate to very good agreement measures. Regarding the lesion masks a moderate overlap is measured between the predicted and reference lesion masks in terms of vDSC, whereas the sDSC returns measures extremely dependent on tolerance values, that span from limited to moderately good and ultimately satisfactory performances for tolerance values of 1 mm, 5 mm and 10 mm, respectively. Figure 4.5 allows for a visual comparison between the lung and lesion masks provided by the *LungQuant* system integrating U-net$_2^{90\%}$ and the reference ones.

Table 4.6: Performances of the *LungQuant* system on the independent COVID-19-CT-Seg test dataset. The vDSC and sDSC computed between the lung and lesion reference masks and those predicted by the *LungQuant* (LQ) system are reported.

	Lung Segmentation			
Metrics	vDSC	sDSC (1 mm)	sDSC (5 mm)	sDSC (10 mm)
LQ (U-net$_2^{60\%}$)	0.96 ± 0.01	0.66 ± 0.09	0.95 ± 0.02	0.98 ± 0.01
LQ (U-net$_2^{90\%}$)	0.95 ± 0.01	0.65 ± 0.09	0.95 ± 0.02	0.98 ± 0.01
	Infection Segmentation			
Metrics	vDSC	sDSC (1 mm)	sDSC (5 mm)	sDSC (10 mm)
LQ (U-net$_2^{60\%}$)	0.62 ± 0.09	0.29 ± 0.06	0.75 ± 0.11	0.90 ± 0.09
LQ (U-net$_2^{90\%}$)	0.66 ± 0.13	0.36 ± 0.13	0.76 ± 0.18	0.87 ± 0.13

Percentage of affected lung volume and CT-SS estimation

The lung and lesion masks provided by the *LungQuant* system can be further processed to derive the physical volumes of each mask and the ratios

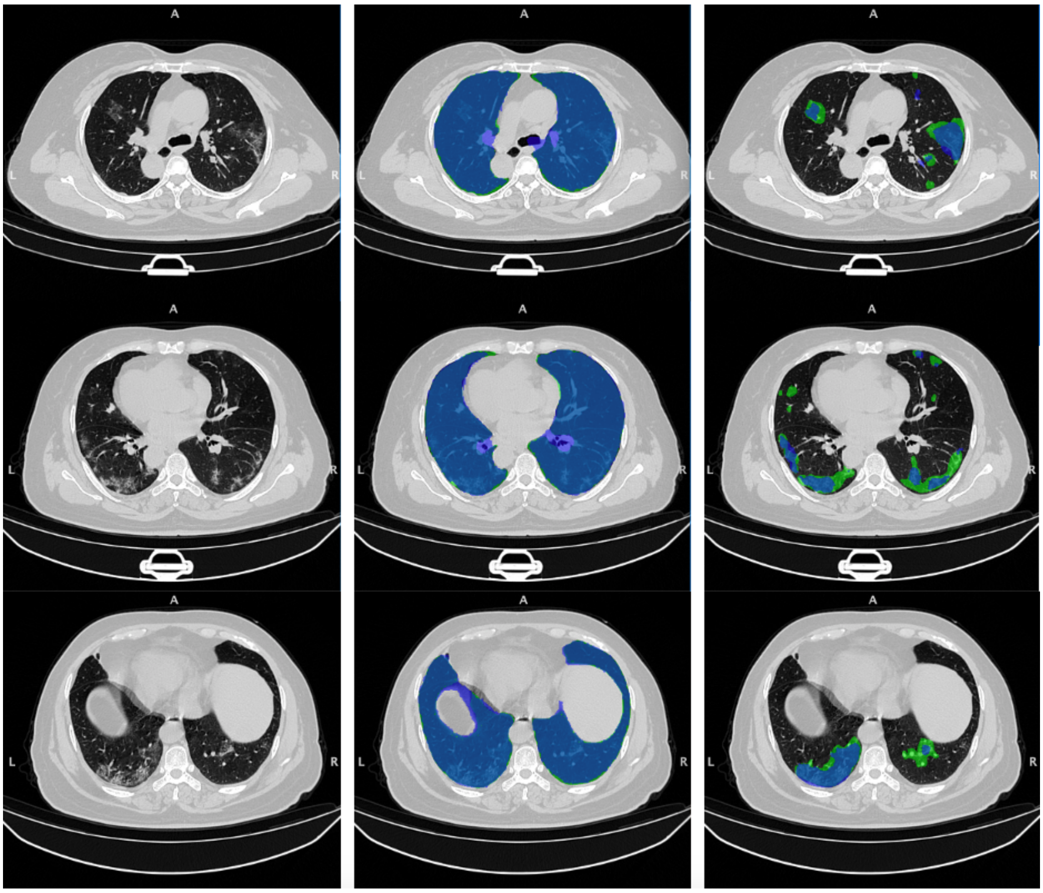

Figure 4.5: On the rows: three axial slices of the first CT scan on the COVID-19-CT-Seg test dataset (*coronacases*001.*nii*) are shown. On the columns: original images (left); overlays between the predicted and the reference lung (center) and COVID-19 lesion (right) masks. The reference masks are in green, while the predicted ones, obtained by the *LungQuant* system integrating U-net$_2^{90\%}$,are in blue.

between the lesion and lung volumes. In Fig. 4.6 the relationship between the percentage of lung involvement as predicted by the *LungQuant* system vs. the corresponding values for the reference masks of the fully independent test set COVID-19-CT-Seg, has been shown for both the *LungQuant* systems with the U-net$_2^{60\%}$ and the U-net$_2^{90\%}$. Despite the limited range of samples to carry out this test, an agreement between the *LungQuant* system output and the reference values is observed for both U-net$_2^{60\%}$ and U-net$_2^{90\%}$. In terms of the Mean Absolute Error (MAE) among the estimated and the reference percentages of affected lung volume (P), I obtained a MAE equal to MAE=4.6% for the LungQuant system with U-net$_2^{60\%}$ and MAE=4.2% for the system with U-net$_2^{90\%}$.

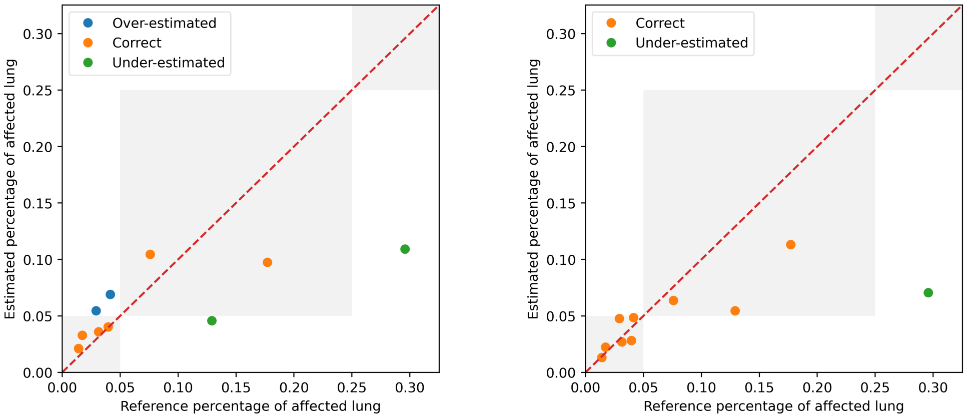

Figure 4.6: Estimated percentages P of affected lung volume versus the ground truth percentages, as obtained by the *LungQuant* system integrating U-net$_2^{60\%}$ (left) and U-net$_2^{90\%}$ (right). The gray areas in the plot backgrounds guide the eye to recognize the CT-SS values assigned to each value of P (from left to right: CT-SS=1, CT-SS=2, CT-SS=3).

The accuracy in assigning the correct CT-SS class is reported in Table 4.7, together with the number of misclassified cases, for the 10 cases of the COVID-19-CT-Seg dataset. The best accuracy achieved by *LungQuant* is of 90% with U-net$_2^{90\%}$. In all cases, the system misclassifies the examples by 1 class at most.

Table 4.7: Classification performances of the whole system in predicting CT Severity Score on MosMed and COVID-19-CT-Seg datasets. The number of misclassified cases is reported.

U-net	Dataset	Accuracy	Misclassified by 1 class	Misclassified by 2 classes
U-net$_2^{60\%}$	COVID-19-CT-Seg	6/10	4/10	0
U-net$_2^{90\%}$	COVID-19-CT-Seg	9/10	1/10	0

4.5 Further improvements: *LungQuant*2.0

As discussed above, the *LungQuant* system has been trained and evaluated on publicly available data. Public data, as discussed in Chapter 2, have the main disadvantage of not containing acquisition information that can be helpful to define the boundaries conditions in which the algorithm can properly work. Once the first version of *LungQuant* was fixed, it has been sent to different hospitals in order to be run on their cases. I noticed that the algorithm failed to segment lung and infections on cases that have a different Field of View (FOV) with respect to the FOV of the public data. CT scans could be, in fact, presented and stored with a reconstructed FOV that can significantly differ from the real FOV. Moreover, most clinicians that gave feedback on the system were interested in studying the left and right lungs separately. For this reason, a second version of the software, *LungQuant*2.0, has been implemented to overcome the following issues:

1. Field of View standardization;
2. Left and right lung separation;
3. Find solutions to linearize the system response with respect to the disease severity.

4.5.1 FOV standardization: BB-Net

Our goal for the second version of the algorithm has been to standardize the FOV of CT scans. In fact, it may happen that the CT scan is reconstructed with a smaller FOV than the acquisition one in order to obtain an enlarged image. For this reason, a third CNN for regression has been introduced on top of *LungQuant*. This CNN extracts six coordinates belonging to two points (x,y,z) from a CT image, which define the bounding box (BB) around the lungs, and it is referred to as BB-net in the following. This bounding box is then used to crop the CT image to the lung volume. The representation of the new pipeline is reported in Figure 4.7.

The network model chosen for selecting the lung bounding box (BB-net) is based on the AlexNet [73]. As shown in Figure 4.8, the model is made up of a series of convolution, max pooling, flattening and dense layers. The final layer of BB-net is a vector with shape 6 which represents the (x,y,z) coordinates of the two points that define the bounding box enclosing the

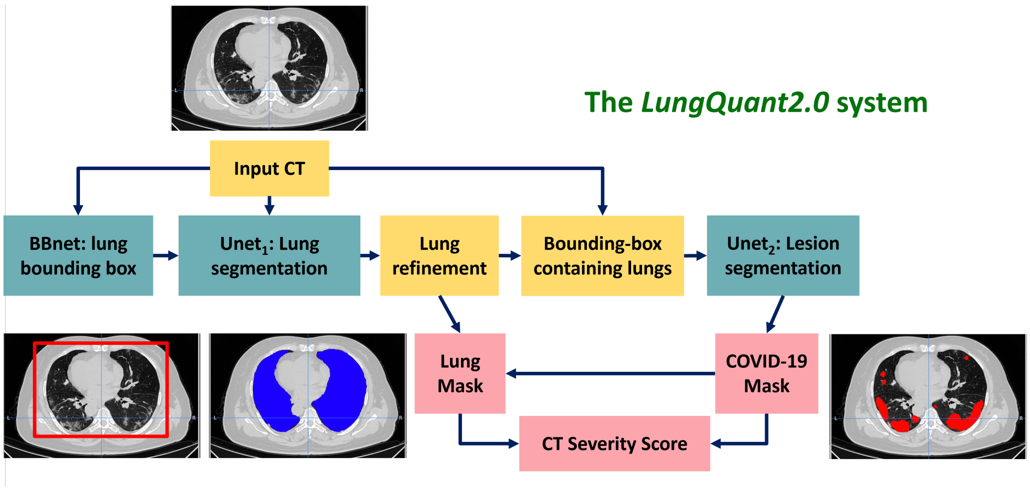
The *LungQuant2.0* system
Input CT
BBnet: lung bounding box
Unet$_1$: Lung segmentation
Lung refinement
Bounding-box containing lungs
Unet$_2$: Lesion segmentation
Lung Mask
COVID-19 Mask
CT Severity Score

Figure 4.7: A sketch of the *LungQuant2.0* analysis pipeline: the input CT scans are processed by the BB-net, which identifies a bounding box enclosing the lungs to be used to crop the images to be provided in input to U-net$_1$, which is devoted to lung segmentation; its output is refined by a morphology-based method; a bounding box enclosing the segmented lungs is identified and used to crop the original CT scan to be then processed by U-net$_2$, which is devoted to COVID-19 lesion segmentation. The *LungQuant2.0* provides as output: the COVID-19 lesion mask (directly provided by U-net$_2$), the lung mask (which is obtained as the logical union between the outputs of U-net$_1$ and U-net$_2$), and the ratio between the COVID-19 lesion and the lung volumes, which provides the percentage of affected lung volume and the CT-SS for each patient.

lungs. The training was performed through a regression and the loss function was the Mean Square Error (MSE). The input image has been windowed in the HU range [-1000, 1000], and then linearly scaled to the [0,1] range. Then, it has been resampled to $100 \times 100 \times 100$ voxels.

BB-net was trained on the data shown in Table 4.1 for which lung masks were available to derive reference bounding boxes for model training. Since the data set is small, not all the available inputs are well represented. In particular, there is an unbalance in the different image FOVs. Most of the publicly available CT scans have large FOVs and a very limited amount of CT scans showed a FOV more focused over the lung volume. For this reason data augmentation was implemented by reducing the FOV, rotating

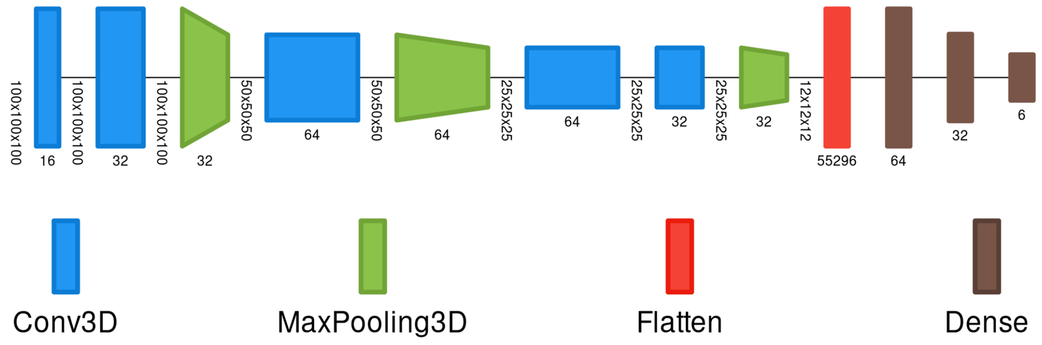

Figure 4.8: Graphical representation of the BB-net, image obtained with Net2Vis software [8]

and displacing the center of the images.

Once the hyperparameters have been optimized through a grid search, the BB-Net has been trained on 80% of the available data (i.e. Plethora, Lung CT Segmentation Challenge, COVID-19 Challenge and MosMed) and its augmentation, while leaving 10% as validation data and 10% as test data. The latter 20% of data was composed only by the original data, i.e. without augmentation. The weights which provided the lowest loss value on the validation set were saved and stored.

BB-net performance

Figure 4.9 shows the training of the BB-net with the optimized hyperparameters. The graph shows the loss (MSE) as a function of the training epoch computed on the training set and the validation set. The minimum value of the validation loss, which is equal to 1.110^{-6}, highlighted in the graph by the vertical dotted gray line, is reached at epoch 758. The weights at this epoch are saved and used to predict the bounding box around the lung shown in Figure 4.10. The red square inside the image shows the predicted bounding box, which nearly perfectly overlaps the true bounding box (yellow square), obtained from the reference lung masks of the annotated CT scans.

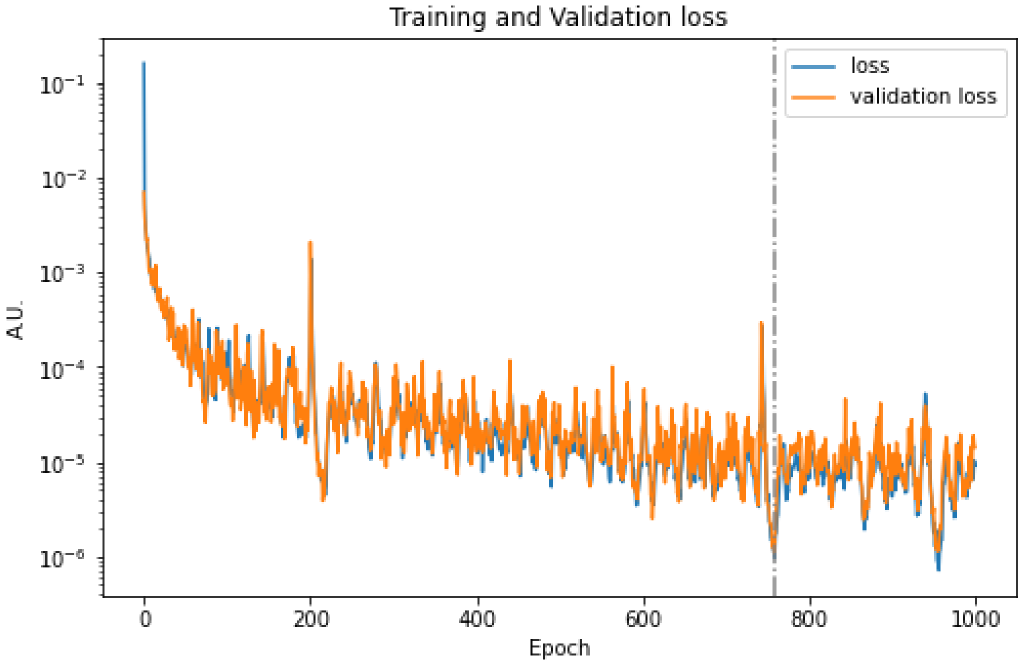

Figure 4.9: BB-net learning curves on the training sample. The blue line is the loss value, computed with the mean square error on the training sample. The yellow line is the loss value calculated on the validation sample. The gray vertical dotted line shows the epoch where the weights of the model were saved.

Results of the *LungQuant2.0* pipeline on the COVID-19-CT-Seg benchmark dataset

The segmentation performance of the two U-nets has been evaluated separately. For both U-net$_1$, which is devoted to lung segmentation, and U-net$_2$, which is used to segment the lesions, the volumetric Dice Similarity Coefficient (vDSC) and the surface Dice Similarity Coefficient (sDSC) at 5 mm of tolerance have been computed on the independent test set COVID-19-CT-Seg. The results are in Table 4.8. Figure 4.11 shows the segmentation outputs computed on a test case (*coronacases008.nii*). Even if the effect on the metrics is negligible, the effect of the introduction of the DNN which infers the bounding boxes containing the lungs is clear looking at the masks applied to images with a different FOV. Figure 4.12 shows the lung segmen-

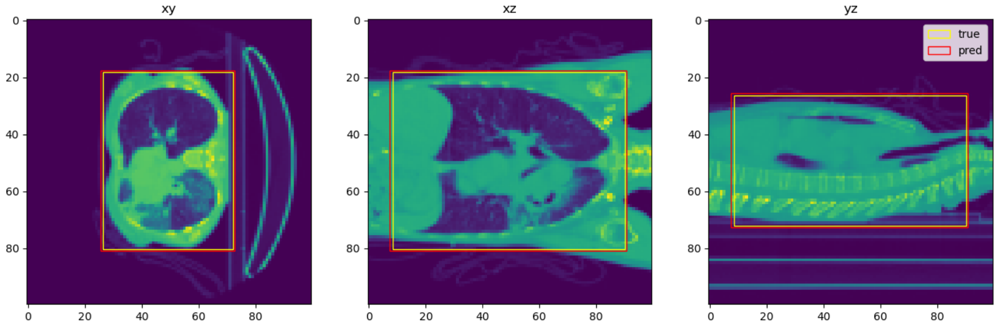

Figure 4.10: BB-net: a predicted bounding box example (red rectangle), compared to the true bounding box (yellow rectangle).

tation of the first and the last version of LungQuant as an example. Datasets which include labelled images with different FOV was not available and hence it was not possible to compute the vDSC and the sDSC to compare the two versions. However from visual assessment of the two outputs I can conclude that the introduction of the bounding box before the lung segmentation has a positive effect.

Table 4.8: Performances of lung segmentation and COVID-19 affected volume made by U-net$_1$ and U-net$_2$ respectively. The metrics are the vDSC and sDSC computed with 5 mm of tolerance.

U-net	vDSC	sDSC
U-net$_1$	0.96 ± 0.01	0.95 ± 0.02
U-net$_2$	0.64 ± 0.14	0.77 ± 0.15

The volumes of the lungs and of the COVID-19 lesions and their ratio have been computed to obtain the CT-SS on the independent test set COVID-19-CT-Seg.

4.5.2 Left and right separation

The algortihm used to separate the left and right lung is based on a watershed transformation. Once the system computes the lung segmentation, the mask is firstly resized at half of its initial size. This was necessary to reduce the

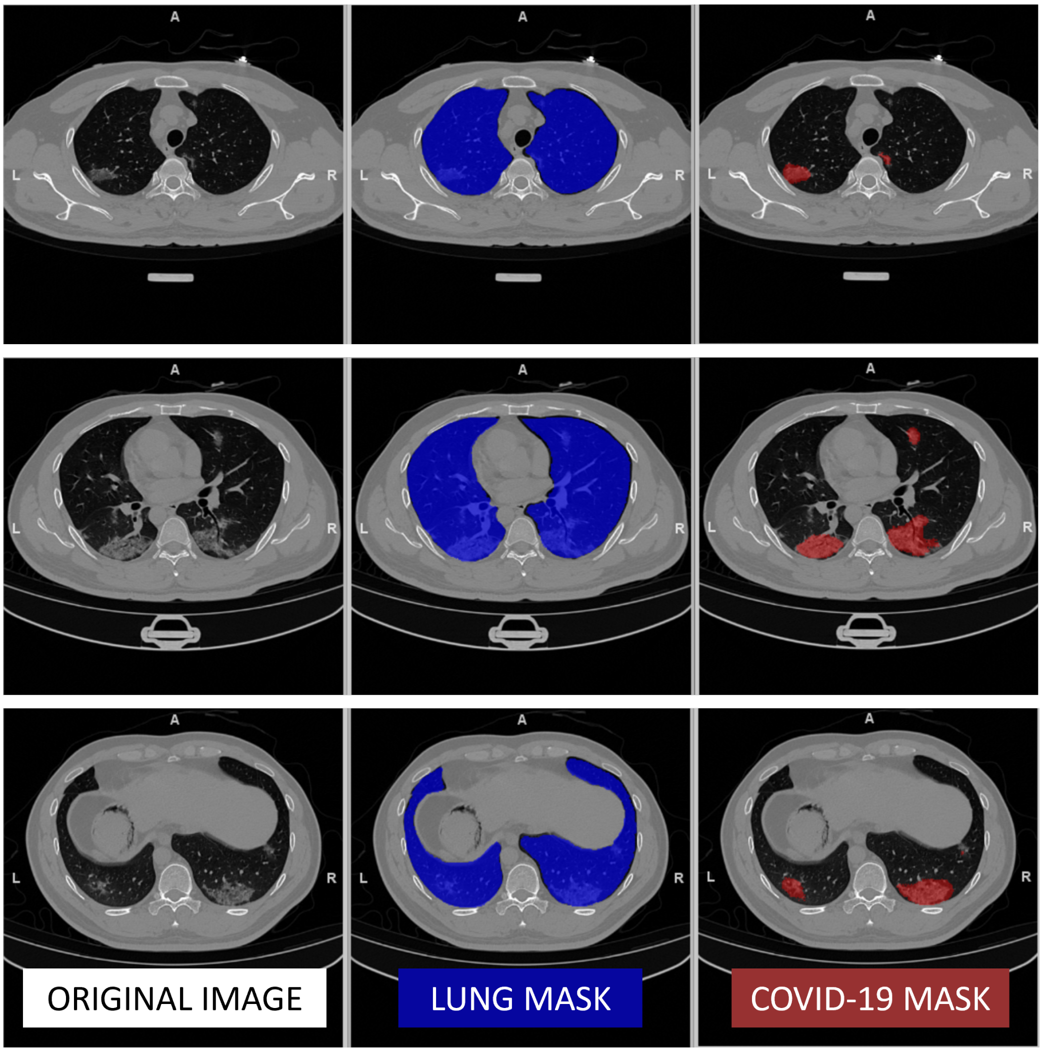

Figure 4.11: *LungQuant*2.0 system: axial slices of case coronacases008.nii from COVID-19-CT-Seg test dataset. On the columns: original images (left), predicted lung (center) and COVID-19 lesion masks (right).

computing time of the following procedure. Then, the euclidean distance transform is applied to the resized lung mask as well as a gaussian filter to reduce noise. Using the peak_local_max function of scikit-image, the local maxima has been computed on the euclidean distance and hence applied the watershed segmentation. Figure 4.13 shows an example of the output of this

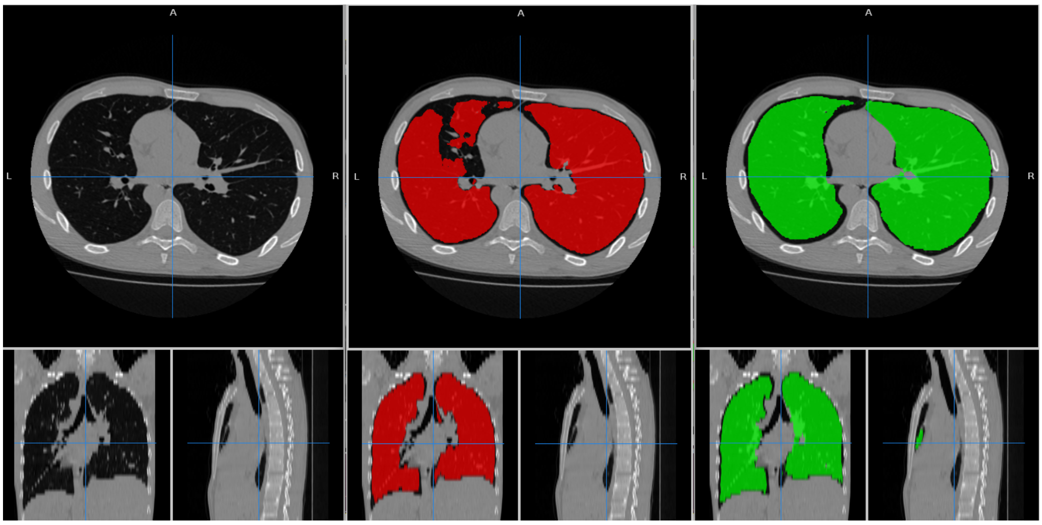

Figure 4.12: Visual assessment of the lung segmentations made with *LungQuant*1.0 and *LungQuant*2.0. On the left: the original image (case study1064.nii from MosMed dataset). On the center: lung segmentation made by LungQuant 1.0. On the right: lung segmentation made by LungQuant 2.0

procedure computed on a case (coronacases005.nii) of the COVID-19-CT-Seg.

4.5.3 Linearization of the response

As a last improvement, since *LungQuant* underestimates the most severe cases, I tried to find a strategy to make the response of the system more linear. This defect was mainly due to the unbalanced data used to train the infection segmentation. In fact, public datasets contain mostly mild cases of COVID-19 pulmonary infection. Moreover, it is not straightforward to imagine a data augmentation which try to augment only severe cases. For this reason, a different loss function has been defined to train the U-Net$_2$ again. The vDSC used to train the previous version in fact is a volumetric metric which inflates when the volumes to be segmented are large. For this reason, the new loss function is less focused on the volumes and more focused on the surface. Moreover, the sDSC, which takes into account the surfaces, is computed in a not efficient way as regards the computing time. For this

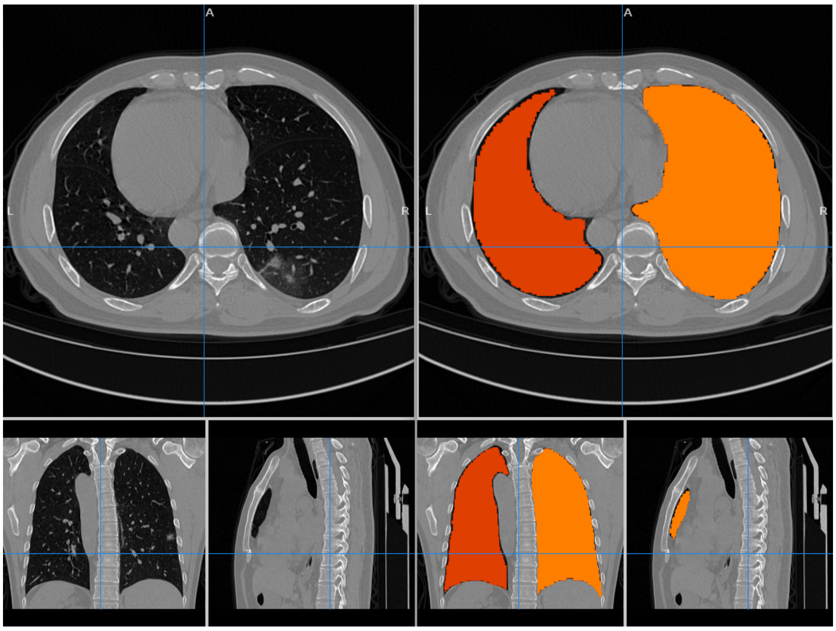

Figure 4.13: On the left: original CT scan of coronacases005.nii.gz with a windowing in [-1000,1000] HU range. On the right: effect of the watershed segmentation to contour left and right lungs separately.

reason, a new term has been added to the loss function and it is defined as follows:

$$L = \sum_{x \in \Omega} F_{pred} \cdot (B_{true} - F_{true}) \tag{4.6}$$

where F_{pred} and F_{true} are the predicted and the reference foreground masks respectively and B_{true} is the reference background mask. The U-Net$_2$ devoted to lesion segmentation has been trained for 150 epochs and the training has been stopped at the best validation vDSC. The performance of the new *LungQuant* has been evaluated on the external independent data set COVID-19-CT-Seg. The results are reported in Figure 4.14 and in Table 4.9 in terms of vDSC and sDSC at 5 mm of tolerance. The MAE for this new system has been computed and it is equal to MAE= 2%. The linearization effect is also clear looking at the right part of Figure 4.14.

The *LungQuant*2.0, the last version of the system developed, is currently under a pseudo-clinical evaluation. We, in fact, started a collaboration with about 12 radiologists coming from 5 different hospitals in Italy and used 120

Table 4.9: Performances achieved by *LungQuant*2 in both lung and infection segmentation on the external independent COVID-19-CT-Seg test dataset.

	Lung	Lung	Infection	Infection
	(vDSC)	(sDSC 5 mm)	(vDSC)	(sDSC 5 mm)
*LungQuant*1	0.95 ± 0.01	0.95 ± 0.02	0.66 ± 0.13	0.76 ± 0.18
*LungQuant*2	0.96 ± 0.01	0.97 ± 0.01	0.69 ± 0.08	0.83 ±0.07

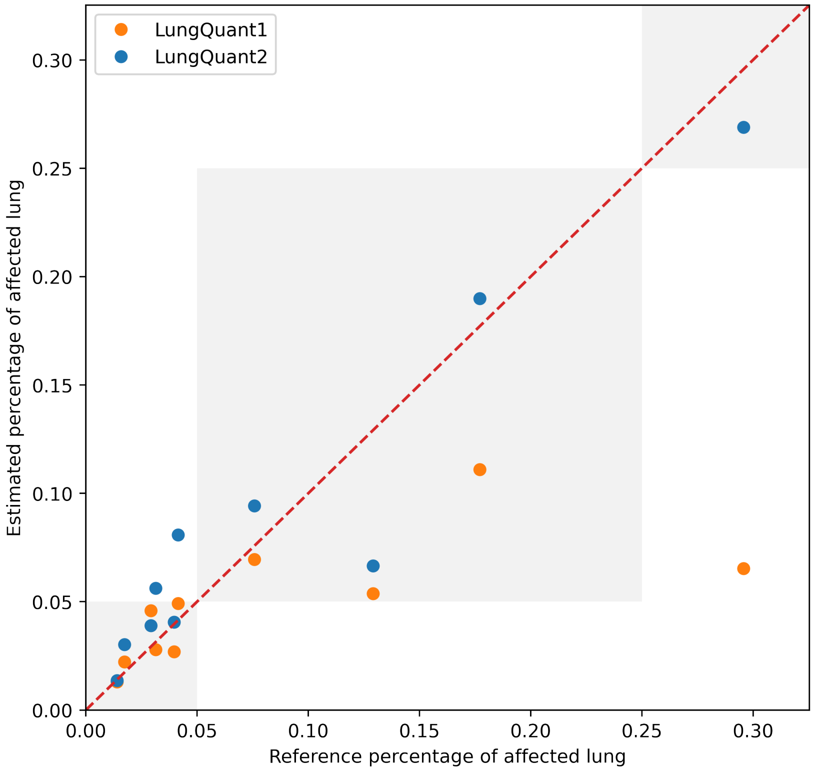

Figure 4.14: In this figure, the estimated percentage of affected lung over the reference one is reported. The Mean Absolute Error obtained with *LungQuant*2.0 is equal to 2%.

public available cases of COVID-19 patient to study both the inter-reader agreement variability of the CTSS and the accordance with our system. This is a pseudo-clinical validation because the used public cases available at the time of this study come from the TCIA "CT Images in COVID-19" collection that has been released without labelling.

4.6 Discussion

In this work [90] a DL-based fully automated analysis pipeline for chest CT scans, the *LungQuant* system, which is able to segment the lungs and the regions of the lung parenchyma affected by COVID-19 infections, has been presented. The system quantifies the percentage of lung tissue affected by COVID-19 lesions and provides the CT-Severity Score for each exam. The *LungQuant* system was developed using only publicly available data to train and test two different U-nets devoted to lung and lesion segmentation, respectively. The whole U-net cascade reaches good performances in terms of both vDSC and sDSC for the lung segmentation, and satisfactory performance for COVID-19 infection segmentation.

As regard the lung segmentation task, *LungQuant* performances compare well with those obtained by Ma *et al.* [98]. Regarding the COVID-19 lesion segmentation, *LungQuant* reaches a vDSC equal to 0.69±0.08 on the independent test set and this result compares well with other fully automated systems, such as the one proposed by Ma *et al.* [98] which used the MosMed annotated subsample as independent test set, obtaining a Dice equal to 0.59±0.21. The *LungQuant* has been evaluated also in terms of sDSC for different values of tolerance. The results obtained at a tolerance of 5 mm, equal to 0.76 ± 0.18, is satisfactory for our purpose, given the heterogeneity of the labelling process. Regarding the correct assignment of the CT-SS, the *LungQuant* system showed an accuracy of 90% on the completely independent test set COVID-19-CT-Seg. Despite this result is encouraging, it was obtained on a rather small independent test set, thus, a broader validation on larger data sample with more heterogeneous composition in terms of disease severity is required.

Training deep learning methods requires a huge amount of labeled data [92]. This problem has been tackled in this work harvesting all accessible, to the best of our knowledge, public datasets. Larger data collections allow training U-nets with a high number of learnable parameters, maintaining their generalization capabilities. In particular, the lung segmentation task has been made feasible thanks to the use of lung CT datasets collected for purposes different from the study of COVID-19 pneumonia. Training a segmentation system on these samples had the effect that when the trained network is used to process the CT scan of a patient with COVID-19 lesions, especially in case ground glass opacities and consolidation are very severe, the lung segmentation is not accurate anymore. In order to overcome this problem, the proposed *LungQuant* system returns a lung mask which is the logical union between the output mask of the U-net$_1$ and the infection mask generated by the U-net$_2$. The lung segmentation module integrated in the *LungQuant* system can actually be improved once lung masks annotation are available for a sample of subjects with COVID-19 lesions. Additionally, balancing training examples according to the severity of radiological findings may also facilitate the learning process of the U-net$_1$ for segmenting the lung.

A similar problem occurs for the segmentation of ground glass opacities and consolidations. The available data, in fact, are very unbalanced with respect to the severity of COVID-19 disease and, hence, the accuracy in segmenting the most severe case is worse. The U-net$_2$ has been trained on very few cases belonging to severe classes, thus it performs better on less severe cases. Training ML systems on balanced datasets is a crucial point to obtain homogeneous performance that are independent from the severity of the disease. The current lack of a large dataset, fully representative of the underlying population, i.e. collected by paying attention to adequately represent all categories of disease severity, limits the possibility to carry out accurate training of AI-based models. Moreover, public available data sets used in this study do not contain demographic or acquisition information. This limit implies that I do not know how some population characteristics may influence the algorithm and hence its application in a hospital workflow should be strictly monitored in order to avoid gender or racial biases. Finally, the lack of acquisition information makes the harmonisation impossible whereas it is well known that it could help to improve both performances and generalization. An additional problem that deals with data, encountered in harvesting COVID-19 data for this analysis, was the difference between the two data sample in the guidelines followed during the collection of images and their annotation. As reported in Table 4.5, the performances of a U-net trained on one specific dataset may decrease significantly when the network is tested on the other dataset. This problem can be overcome by gathering together many images acquired and labelled in different ways, as shown in the last row of Table 4.5. However, in our case, merging the COVID-19-Challenge dataset and the MosMed dataset led to very unbalanced training data and the proposed system underestimates the extension of the infection regions. The possibility to access to more populated and fully annotated data samples is fundamental to push the performance of DL-based image processing models. Despite this issue, I tried to solve it defining a loss function which mitigate the underestimation effect and obtained good results in terms of

MAE with respect the previous version. However, this improvement should be definitely tested on a larger data set since the COVID-19-CT-Seg contains only 10 cases and just one case with a CTSS equal to 3.

As a final consideration, this segmentation and quantification work opens the way to lesion characterization studies. The segmentation of lungs and lesions related to COVID-19 pneumonia is a prerequisite to the extraction of radiomic features that can help to distinguish COVID-19 infection from other non-COVID related pneumonia, and to develop predictive models of patients' outcome. In this direction, the work by Fang *et al.* [36] developed an AI-based method to predict a severity score, which showed the remarkable performance of AUC = 0.813 in predicting the subjects' intensive care unit admission. To evaluate the capability of our *LungQuant* system to enable the development of predictive models of disease progression and patients' outcome, the availability of a fully annotated database with phenotypic and clinical information of patients is required.

Conclusions and Discussion

In this PhD thesis, I faced the problem of developing deep-learning based algorithm applied to medical images. In Chapter 1, an overview of the principle of X-Ray imaging has been presented as well as an introduction to deep convolutional neural networks and their explainability. In Chapter 2, I inserted a discussion which goes even beyond the simple algorithm development. It deals with the meaning of making data science with opaque and non transparent algorithms, especially if they are meant to be applied in clinical practice. As any multidisciplinary science, I support the thesis that we should know the domains that are involved in algorithm development such as medicine, jurisprudence, physics, computer science, philosophy, history and social sciences in general. Even if an experiment that takes into account so many expertise is very expensive and requires huge organization capability, it is fundamental to build a fair and reliable instrument to support physicians. Applied sciences always move on a thin border between making science, intended as the increase of knowledge and research processes, and making a product, to be sold and then used into a hospital. This peculiarity comes before the advent of the fourth paradigm of science and this offers to us a very interesting moment for thinking about how science and especially applied science is changing. Using the Kuhnian definition of paradigm, the fourth paradigm is changing the classic scientific method since it infers a model by the data. The way we intend the data is important to frame the kind of science we are doing: in a pure empiricist approach data are true and natural, while in a constructivist approach data are constructed and the grade of truth they represent depends on how they have been collected. As regards medical images, building a dataset is a challenging task. In fact, data need to be labelled and medical labeling requires time, precision and, usually, tools to support the physician labeling. The ground truth on images usually can not be compared to an objective measure that establishes its goodness and so it relies on radiologist opinions. Unfortunately, this kind of labeling suffers from inter observer variability such that it is not easy to obtain a good ground truth even if the labeling has been made in a consensus modality, i.e. with agreement among many doctors. Furthermore, the data availability is limited: datasets of medical images are usually small and they may be not accessible. In public collections, instead, important acquisition information may be lost due to privacy issues. The access to datasets is important not only to train algorithms but also to test different algorithms on the same dataset. In this context, the application of AI to medical images needs a special care since its wrong use may harm not only people but also health care systems [140]. Developing an algorithm that takes into account all the issues is a very complex task. However all these problems need to be addressed if algorithms we develop are going to be used in hospitals on people. It is possible to evaluate all these issues also a-posteriori and, in Chapter 2, I presented a process called Z-Inspection to assess trustworthiness of algorithm following the EU guidelines on ethics. The process is made of three different phases and its aim is to examine a medical algorithm from many points of view. The issues described in Chapter 2 are then discussed also in the two use cases of Chapters 3 and 4. The first one is a classifier for breast density on mammograms. Breast density is an important feature for three reasons: 1) it is responsible for masking effect, i.e. dense tissue may cover a malignant mass, 2) dense tissue is radiosensitive and 3) higher breast density is associated with a higher risk of developing cancer. The most used standard for breast density classification is the BI-RADS and it is made of four qualitative classes that are defined through example and text descriptions. For this reason, I used a Convolutional Neural Network to perform the classification. Since CNNs are able to extract significant features by themselves, the way they take a decision is opaque and for this reason I also presented a way to qualitatively and quantitatively measure the goodness of explanation through class activation maps. The qualitative explanation is based on the visual assessment of the activation maps to verify that the part of the image that the CNN uses to predict a correct class correspond to the dense tissue. As regards the quantitative explanation, it was not possible to obtain a segmentation of the dense tissues to be directly compared to most activated areas of the maps because it is too time consuming and also because breast density is not defined as a pixel-wise measure. However, breast density is surely correlated to the pixel intensity so that I propose as a quantitative measure for the explanation a simple computation based on the Spearman correlation coefficient. The expectation is that correlation increases as the density class increases. The performance obtained with this classifier compares well with the literature and the accuracy, precision and recall of the classifier are respectively 82%, 83.3% and 80.3%. The trend of the Spearman correlation ranks over the classes generally shows an increase for higher density classes. However, this is not generally true for the most dense class. This behaviour can be explained with the unbalanced dataset used which anyway represents the empirical distribution of breast density classes measured on more than 3,8 millions women of a screening population. In Chapter 4, a deep learning based algorithm for the quantification of lung damage due to COVID-19 infection on CT has been presented. The algorithm consists in three CNNs in a cascade modality: the first one is devoted to the identification of a bounding box that encloses the lungs, the second one is trained to segment the lungs and the last one to segment the infection inside the lungs. To train these CNNs, only publicly available data have been used in order to let the performance be reproducible. However, there are not public datasets that contain CT scans of COVID-19 patients with lungs reference masks and this issue lowers the performance of lung segmentation. Despite the problem of having an adequate dataset to be used for training, the system, which is called *LungQuant* reaches very good performances in terms of volumetric and surface Dice Similarity Coefficient. This algorithm which is able to segment the COVID-19 lesions can be used to extract radiomic features in order to predict patient prognosis. Working on these two use cases allow me to know and understand deeper all the issues described in Chapter 2 even if it could not obviously be possible to address all of them. They, for example, have been trained using different kinds of data, public and private, on different imaging modalities and for different scopes. This thesis is, at the best of my knowledge, the first attempt to organize and discuss the issues linked to medical algorithm development in literature. The difficulty of doing this kind of work is that it deals with many arguments and many technical languages. [1] This requires huge financial funds and organization skills. One

[1] During my PhD, I participated in many meetings with physicians, physicists and computer scientists; I had the possibility to speak with lawyers, historians and sociologists; I had also the chance to have conversations with patients. All these experiences, combined with my scientific study on AI applied to medical images, are the reason why I decided to insert a chapter that contains what I really learned during my PhD experience. And what I really learned is that we need many points of view, many expertises and many experiences to develop an algorithm which could properly be used in a hospital.

of the pivotal problem is the access to data so that I want to conclude my thesis with a call for research and health institutions. While we are moving inside these problems trying to get the sufficient number of images and data by discussing with a specific hospital or a single physician, I think it would be important a collective project or organization such that researchers belonging to public research institutes or universities, could access to a huge amount of anonymized and diverse medical data in order to build a fairer way to develop medical algorithms. This requires an ethical treatment of patients privacy and financial investments for shared servers and their maintenance. In an increasingly digital world and medicine, the access to medical data for researchers to build and validate models, algorithms and software is a necessary but not sufficient way to preserve our health systems which are very precious for having a healthy society.

Bibliography

[1] https://covid19.who.int/.

[2] Olivier Alonzo-Proulx, Gordon E. Mawdsley, James T. Patrie, Martin J. Yaffe, and Jennifer A. Harvey. Reliability of automated breast density measurements. *Radiology*, 275(2):366–376, 2015.

[3] Peng An, Sheng Xu, Stephanie A. Harmon, Evrim B. Turkbey, Thomas H. Sanford, Amel Amalou, Michael Kassin, Nicole Varble, Maxime Blain, Victoria Anderson, Francesca Patella, Gianpaolo Carrafiello, Baris T. Turkbey, and Bradford J. Wood. CT Images in COVID-19, 2020.

[4] Chris Anderson. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, 2008.

[5] Samuel G. Armato, Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, Ella A. Kazerooni, Heber MacMahon, Edwin J.R. Van Beek, David Yankelevitz, Alberto M. Biancardi, Peyton H. Bland, Matthew S. Brown, Roger M. Engelmann, Gary E. Laderach, Daniel Max, Richard C. Pais, David P.Y. Qing, Rachael Y. Roberts, Amanda R. Smith, Adam Starkey, Poonam Batra, Philip Caligiuri, Ali Farooqi, Gregory W. Gladish, C. Matilda Jude, Reginald F. Munden, Iva Petkovska, Leslie E. Quint, Lawrence H. Schwartz, Baskaran Sundaram, Lori E. Dodd, Charles Fenimore, David Gur, Nicholas Petrick, John Freymann, Justin Kirby, Brian Hughes, Alessi Vande Casteele, Sangeeta Gupte, Maha Sallam, Michael D. Heath, Michael H. Kuhn, Ekta Dharaiya, Richard Burns, David S. Fryd, Marcos Salganicoff, Vikram Anand, Uri Shreter, Stephen Vastagh, Barbara Y. Croft, and Laurence P. Clarke.

The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2):915–931, 2011.

[6] Reyer Zwiggelaar Arnau Oliver, Jordi Freixenet. AUTOMATIC CLASSIFICATION OF BREAST DENSITY Arnau Oliver , Jordi Freixenet Computer Vision and Robotics Group University of Girona Campus de Montilivi s / n . 17071 Girona , Spain Reyer Zwiggelaar Department of Computer Science University of Wales Aberystw. *Ieee*, pages 1258–1261, 2005.

[7] I Banerjee, Bhimireddy A R Ms, Burns J L Ms, Celi La, L Chen, and R Correa. Reading Race : AI Recognizes Patient ' s Racial Identity In Medical Images. *arXiv*.

[8] Alex Bäuerle, Christian Van Onzenoodt, and Timo Ropinski. Net2vis–a visual grammar for automatically generating publication-tailored cnn architecture visualizations. *IEEE transactions on visualization and computer graphics*, 27(6):2980–2991, 2021.

[9] R. Bellotti, F. De Carlo, G. Gargano, S. Tangaro, D. Cascio, E. Catanzariti, P. Cerello, S.C. Cheran, P. Delogu, I. De Mitri, C. Fulcheri, D. Grosso, A. Retico, S. Squarcia, E. Tommasi, and B. Golosio. A CAD system for nodule detection in low-dose lung CTs based on region growing and a new active contour model. *Medical Physics*, 34(12), 2007.

[10] Robert A Beltran. The Gold Standard: The Challenge of Evidence-Based Medicine and Standardization in Health Care. *Journal of the National Medical Association*, 97(1):110, jan 2005.

[11] Keir Bovis. Classification of mammographic breast density using a combined classifier paradigm. *Medical Image Understanding and Analysis*, (c):1–4, 2002.

[12] N F Boyd, J W Byng, R A Jong, E K Fishell, L E Little, A B Miller, G A Lockwood, D L Tritchler, and M J Yaffe. Quantitative Classification of Mammographic Densities and Breast Cancer Risk: Results From the Canadian National Breast Screening Study. *JNCI: Journal of the National Cancer Institute*, 87(9):670–675, 1995.

[13] Norman F. Boyd, Lisa J. Martin, Martin J. Yaffe, and Salomon Minkin. Mammographic density and breast cancer risk: Current understanding and future prospects. *Breast Cancer Research*, 13(6):1–12, 2011.

[14] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[15] Pete Bridge, Andrew Fielding, Pamela Rowntree, and Andrew Pullar. Intraobserver Variability: Should We Worry? *Journal of Medical Imaging and Radiation Sciences*, 47(3):217–220, 2016.

[16] Niccolò Camarlinghi, Ilaria Gori, Alessandra Retico, Roberto Bellotti, Paolo Bosco, Piergiorgio Cerello, Gianfranco Gargano, Ernesto Lopez Torres, Rosario Megna, Marco Peccarisi, and Maria Evelina Fantacci. Combination of computer-aided detection algorithms for automatic lung nodule identification. *International Journal of Computer Assisted Radiology and Surgery*, 7(3):455–464, 2012.

[17] Alexander Campolo and Kate Crawford. Enchanted Determinism: Power without Responsibility in Artificial Intelligence. *Engaging Science, Technology, and Society*, 6:1–19, 2020.

[18] Marina Carotti, Fausto Salaffi, Piercarlo Sarzi-Puttini, Andrea Agostini, Alessandra Borgheresi, Davide Minorati, Massimo Galli, Daniela Marotto, and Andrea Giovagnoni. Chest CT features of coronavirus disease 2019 (COVID-19) pneumonia: key points for radiologists. *Radiologia Medica*, 125(7):636–646, 2020.

[19] Isabella Castiglioni, Leonardo Rundo, Marina Codari, Giovanni Di Leo, Christian Salvatore, Matteo Interlenghi, Francesca Gallivanone, Andrea Cozzi, Natascha Claudia D'Amico, and Francesco Sardanelli. AI applications to medical images: From machine learning to deep learning. *Physica Medica*, 83(November 2020):9–24, 2021.

[20] K. S.Clifford Chao, Shreerang Bhide, Hansen Chen, Joshua Asper, Steven Bush, Gregg Franklin, Vivek Kavadi, Vichaivood Liengswangwong, William Gordon, Adam Raben, Jon Strasser, Christopher Koprowski, Steven Frank, Gregory Chronowski, Anesa Ahamad, Robert Malyapa, Lifei Zhang, and Lei Dong. Reduce in Variation and Improve

Efficiency of Target Volume Delineation by a Computer-Assisted System Using a Deformable Image Registration Approach. *International Journal of Radiation Oncology Biology Physics*, 68(5):1512–1521, 2007.

[21] François Chollet and Others. Keras. https://keras.io, 2015.

[22] S Ciatto, N Houssami, A Apruzzese, E Bassetti, B Brancato, F Carozzi, S Catarzi, M P Lamberini, G Marcelli, R Pellizzoni, B Pesce, G Risso, F Russo, and A Scorsolini. Categorizing breast mammographic density: intra- and interobserver reproducibility of BI-RADS density categories. *The Breast*, 14(4):269–275, 2005.

[23] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (TCIA): Maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013.

[24] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, apr 1960.

[25] Davide Colombi, Flavio C. Bodini, Marcello Petrini, Gabriele Maffi, Nicola Morelli, Gianluca Milanese, Mario Silva, Nicola Sverzellati, and Emanuele Michieletti. Well-aerated Lung on Admitting Chest CT to Predict Adverse Outcome in COVID-19 Pneumonia. *Radiology*, 296(2):E86–E96, 2020.

[26] D R Dance, S Christofides, I D McLean, A D A Maidment, and K H Ng. *Diagnostic Radiology Physics*. Non-serial Publications. INTERNATIONAL ATOMIC ENERGY AGENCY, Vienna, 2014.

[27] Giorgio De Nunzio, Eleonora Tommasi, Antonella Agrusti, Rosella Cataldo, Ivan De Mitri, Marco Favetta, Silvio Maglio, Andrea Massafra, Maurizio Quarta, Massimo Torsello, Ilaria Zecca, Roberto Bellotti, Sabina Tangaro, Piero Calvini, Niccolò Camarlinghi, Fabio Falaschi, Piergiorgio Cerello, and Piernicola Oliva. Automatic lung segmentation in CT images with accurate handling of the hilar region. *Journal of Digital Imaging*, 24(1):11–27, 2011.

[28] Jamshid Dehmeshki, Hamdan Amin, Manlio Valdivieso, and Xujiong Ye. Segmentation of pulmonary nodules in thoracic CT scans: A region

growing approach. *IEEE Transactions on Medical Imaging*, 27(4):467–480, 2008.

[29] DICOM standard. https://www.dicomstandard.org/current, accessed on 07/01/2022.

[30] Jennifer J. Donald and Stuart A. Barnard. Common patterns in 558 diagnostic radiology errors. *Journal of Medical Imaging and Radiation Oncology*, 56(2):173–178, 2012.

[31] Francis A. Duck. The origins of medical physics. *Physica Medica*, 30(4):397–402, 2014.

[32] Olive Jean Dunn. Multiple Comparisons Using Rank Sums. *Technometrics*, 6(3):241–252, aug 1964.

[33] Ernest U. Ekpo, Ujong Peter Ujong, Claudia Mello-Thoms, and Mark F. McEntee. Assessment of interradiologist agreement regarding mammographic breast density classification using the fifth edition of the BI-RADS atlas. *American Journal of Roentgenology*, 206(5):1119–1123, 2016.

[34] Simon Aagaard Enni and Maja Bak Herrie. Turning biases into hypotheses through method: A logic of scientific discovery for machine learning. *Big Data and Society*, 8(1), 2021.

[35] Lijie Fan, Shengjia Zhao, and Stefano Ermon. Adversarial Localization Network. *Nips*, (Nips), 2017.

[36] Xi Fang, Uwe Kruger, Fatemeh Homayounieh, Hanqing Chao, Jiajin Zhang, Subba R. Digumarthy, Chiara D. Arru, Mannudeep K. Kalra, and Pingkun Yan. Association of AI quantified COVID-19 chest CT and patient outcome. *International Journal of Computer Assisted Radiology and Surgery*, 2021.

[37] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Transactions on Medical Imaging*, 39(11):3619–3629, 2020.

[38] Jean-Philippe Fortin, Drew Parker, Birkan Tunç, Takanori Watanabe, Mark A Elliott, Kosha Ruparel, David R Roalf, Theodore D Satterthwaite, Ruben C Gur, Raquel E Gur, Robert T Schultz, Ragini Verma, and Russell T Shinohara. Harmonization of multi-site diffusion tensor imaging data. *NeuroImage*, 161:149–170, 2017.

[39] Phoebe E. Freer. Mammographic breast density: Impact on breast cancer risk and implications for screening. *Radiographics*, 35(2):302–315, 2015.

[40] Ziba Gandomkar, Moayyad E Suleiman, Delgermaa Demchig, Patrick C Brennan, and Mark F McEntee. BI-RADS density categorization using deep neural networks. In *Proc.SPIE*, volume 10952, mar 2019.

[41] Lucas L. Geyer, U. Joseph Schoepf, Felix G. Meinel, John W. Nance, Gorka Bastarrika, Jonathon A. Leipsic, Narinder S. Paul, Marco Rengo, Andrea Laghi, and Carlo N. De Cecco. State of the Art: Iterative CT reconstruction techniques1. *Radiology*, 276(2):339–357, 2015.

[42] Bruno Golosio, Giovanni Luca Masala, Alessio Piccioli, Piernicola Oliva, Massimo Carpinelli, Rosella Cataldo, Piergiorgio Cerello, Francesco De Carlo, Fabio Falaschi, Maria Evelina Fantacci, Gianfranco Gargano, Parnian Kasae, and Massimo Torsello. A novel multithreshold method for nodule detection in lung CT. *Medical Physics*, 36(8):3607–3618, 2009.

[43] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[44] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[45] Manu Goyal, Thomas Knackstedt, Shaofeng Yan, and Saeed Hassanpour. Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities. *Computers in Biology and Medicine*, 127(August):104065, 2020.

[46] Olya Grove, Anders E. Berglund, Matthew B. Schabath, Hugo J.W.L. Aerts, Andre Dekker, Hua Wang, Emmanuel Rios Velazquez, Philippe Lambin, Yuhua Gu, Yoganand Balagurunathan, Edward Eikman, Robert A. Gatenby, Steven Eschrich, and Robert J. Gillies. Quantitative computed tomographic descriptors associate tumor shape complexity and intratumor heterogeneity with prognosis in lung adenocarcinoma. *PLoS ONE*, 10(3):1–14, 2015.

[47] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 2018.

[48] H. A. Haenssle and C. Fink. Reply to the letter to the editor What type of man against machine?' by H. Smith. *Annals of Oncology*, 29(9):2024–2025, 2018.

[49] H. A. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. Ben Hadj Hassen, L. Thomas, A. Enk, L. Uhlmann, Christina Alt, Monika Arenbergerova, Renato Bakos, Anne Baltzer, Ines Bertlich, Andreas Blum, Therezia Bokor-Billmann, Jonathan Bowling, Naira Braghiroli, Ralph Braun, Kristina Buder-Bakhaya, Timo Buhl, Horacio Cabo, Leo Cabrijan, Naciye Cevic, Anna Classen, David Deltgen, Christine Fink, Ivelina Georgieva, Lara Elena Hakim-Meibodi, Susanne Hanner, Franziska Hartmann, Julia Hartmann, Georg Haus, Elti Hoxha, Raimonds Karls, Hiroshi Koga, Ju¨rgen Kreusch, Aimilios Lallas, Pawel Majenka, Ash Marghoob, Cesare Massone, Lali Mekokishvili, Dominik Mestel, Volker Meyer, Anna Neuberger, Kari Nielsen, Margaret Oliviero, Riccardo Pampena, John Paoli, Erika Pawlik, Barbar Rao, Adriana Rendon, Teresa Russo, Ahmed Sadek, Kinga Samhaber, Roland Schneiderbauer, Anissa Schweizer, Ferdinand Toberer, Lukas Trennheuser, Lyobomira Vlahova, Alexander Wald, Julia Winkler, Priscila Wo¨lbing, and Iris Zalaudek. Man against Machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, 2018.

[50] 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, 2016.

[51] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]*, 2015.

[52] Tony Hey, Stewart Tansley, and Kristin M. Tolle. *Fourth Paradigm.* 2021.

[53] High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI, 2021.

[54] Johannes Hofmanninger, Florian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic lung segmentation in routine imaging is a data diversity problem, not a methodology problem. *arXiv*, 2, 2020.

[55] Qinhua Hu, Luís Fabrício Luís, Gabriel Bandeira Holanda, Shara S.A. Alves, Francisco Hércules Francisco, Tao Han, and Pedro P. Rebouças Filho. An effective approach for CT lung segmentation using mask region-based convolutional neural networks. *Artificial Intelligence in Medicine*, 103(July 2019):101792, 2020.

[56] DH Hubel and TN Wiesel. RECEPTIVE FIELDS AND FUNCTIONAL ARCHITECTURE OF MONKEY STRIATE CORTEX. *The Journal of physiology*, pages 215–243, 1968.

[57] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]*, 2015.

[58] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Köhler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein. nnU-Net: Self-adapting framework for u-net-based medical image segmentation. *arXiv*, 2018.

[59] P C Johns and M J Yaffe. X-ray characterisation of normal and neoplastic breast tissues. *Physics in Medicine and Biology*, 32(6):675–695, may 1987.

[60] Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), 2019.

[61] Georgios A. Kaissis, Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

[62] Davood Karimi, Haoran Dou, Simon K. Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

[63] Eric L.H. Khoo, Karlissa Schick, Ashley W. Plank, Michael Poulsen, Winnie W.G. Wong, Mark Middleton, and Jarad M. Martin. Prostate contouring variation: Can it be fixed? *International Journal of Radiation Oncology Biology Physics*, 82(5):1923–1929, 2012.

[64] Ji Hoon Kim, Sang Gil Han, Ara Cho, Hye Jung Shin, and Song-Ee Baek. Effect of deep learning-based assistive technology use on chest radiograph interpretation by emergency department physicians: a prospective interventional simulation-based study. *BMC Medical Informatics and Decision Making*, 21(1):1–9, 2021.

[65] Kendall J. Kiser, Sara Ahmed, Sonja Stieb, Abdallah S.R. Mohamed, Hesham Elhalawani, Peter Y.S. Park, Nathan S. Doyle, Brandon J. Wang, Arko Barman, Zhao Li, W. Jim Zheng, Clifton D. Fuller, and Luca Giancardo. PleThora: Pleural effusion and thoracic cavity segmentations in diseased lungs for benchmarking chest CT processing pipelines. *Medical Physics*, 47(11):5941–5952, 2020.

[66] Kendall J. Kiser, Arko Barman, Sonja Stieb, Clifton D. Fuller, and Luca Giancardo. Novel Autosegmentation Spatial Similarity Metrics Capture the Time Required to Correct Segmentations Better Than Traditional Metrics in a Thoracic Cavity Segmentation Workflow. *Journal of Digital Imaging*, 34(3):541–553, 2021.

[67] R. Kitchin. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences.* SAGE Publications, 2014.

[68] Rob Kitchin. Big Data, new epistemologies and paradigm shifts. *Big Data and Society*, 1(1):1–12, 2014.

[69] Thijs Kooi, Bram van Ginneken, Nico Karssemeijer, and Ard den Heeten. Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network. *Medical physics*, 44(3):1017–1027, 2017.

[70] Dimitrios Korkinof, Tobias Rijken, Michael O'Neill, Joseph Yearsley, Hugh Harvey, and Ben Glocker. High-Resolution Mammogram Synthesis using Progressive Generative Adversarial Networks. (2017), 2018.

[71] Mika Kortesniemi, Virginia Tsapaki, Annalisa Trianni, Paolo Russo, Ad Maas, Hans-Erik Källman, Marco Brambilla, and John Damilakis. The European Federation of Organisations for Medical Physics (EFOMP) White Paper: Big data and deep learning in medical imaging and in relation to medical physics profession. *Physica Medica*, 56:90–93, 2018.

[72] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[73] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[74] William H Kruskal and W Allen Wallis. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260):583–621, dec 1952.

[75] Thomas S Kuhn. *The structure of Scientific Revolution*, volume I,II.

[76] Indrajeet Kumar, Bhadauria H.S., Jitendra Virmani, and Shruti Thakur. A classification framework for prediction of breast density

using an ensemble of neural network classifiers. *Biocybernetics and Biomedical Engineering*, 37(1):217–228, 2017.

[77] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[78] 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. *Nature Reviews Clinical Oncology*, 14(12):749–762, 2017.

[79] Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences of the United States of America*, 117(23):12592–12594, 2020.

[80] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L Rubin. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data*, 4(1):170177, 2017.

[81] Constance D Lehman, Adam Yala, Tal Schuster, Brian Dontchos, Manisha Bahl, Kyle Swanson, and Regina Barzilay. Mammographic Breast Density Assessment Using Deep Learning: Clinical Implementation. *Radiology*, 290(1):52–58, 2019.

[82] S. Leonelli. What difference does quantity make? On the epistemology of Big Data in biology. *Big Data and Society*, 1(1):1–11, 2014.

[83] Nikolas Lessmann, Clara I. Sánchez, Ludo Beenen, Luuk H. Boulogne, Monique Brink, Erdi Calli, Jean Paul Charbonnier, Ton Dofferhoff,

Wouter M. van Everdingen, Paul K. Gerke, Bram Geurts, Hester A. Gietema, Miriam Groeneveld, Louis van Harten, Nils Hendrix, Ward Hendrix, Henkjan J. Huisman, Ivana Išgum, Colin Jacobs, Ruben Kluge, Michel Kok, Jasenko Krdzalic, Bianca Lassen-Schmidt, Kicky van Leeuwen, James Meakin, Mike Overkamp, Tjalco van Rees Vellinga, Eva M. van Rikxoort, Riccardo Samperna, Cornelia Schaefer-Prokop, Steven Schalekamp, Ernst Th Scholten, Cheryl Sital, J. Lauran Stöger, Jonas Teuwen, Kiran Vaidhya Venkadesh, Coen de Vente, Marieke Vermaat, Weiyi Xie, Bram de Wilde, Mathias Prokop, and Bram van Ginneken. Automated assessment of COVID-19 reporting and data system and chest CT severity scores in patients suspected of having COVID-19 using artificial intelligence. *Radiology*, 298(1):E18–E28, 2021.

[84] Hui Li, Yitan Zhu, Elizabeth S. Burnside, Erich Huang, Karen Drukker, Katherine A. Hoadley, Cheng Fan, Suzanne D. Conzen, Margarita Zuley, Jose M. Net, Elizabeth Sutton, Gary J. Whitman, Elizabeth Morris, Charles M. Perou, Yuan Ji, and Maryellen L. Giger. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. *npj Breast Cancer*, 2(1), 2016.

[85] Suiyi Li, Yuxuan Chen, Su Yang, and Wuyang Luo. Cascade Dense-Unet for Prostate Segmentation in MR Images. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11643 LNCS, pages 481–490. Springer Verlag, 2019.

[86] Karen Lim, William Small, Lorraine Portelance, Carien Creutzberg, Ina M. Jürgenliemk-Schulz, Arno Mundt, Loren K. Mell, Nina Mayr, Akila Viswanathan, Anuja Jhingran, Beth Erickson, Jennifer De Los Santos, David Gaffney, Catheryn Yashar, Sushil Beriwal, Aaron Wolfson, Alexandra Taylor, Walter Bosch, Issam El Naqa, and Anthony Fyles. Consensus guidelines for delineation of clinical target volume for intensity-modulated pelvic radiotherapy for the definitive treatment of cervix cancer. *International Journal of Radiation Oncology Biology Physics*, 79(2):348–355, 2011.

[87] E. J. Limkin, R. Sun, L. Dercle, E. I. Zacharaki, C. Robert, S. Reuzé, A. Schernberg, N. Paragios, E. Deutsch, and C. Ferté. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Annals of Oncology*, 28(6):1191–1206, 2017.

[88] 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, 2017.

[89] 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(December 2012):60–88, 2017.

[90] Francesca Lizzi, Abramo Agosti, Francesca Brero, Raffaella Fiamma Cabini, Maria Evelina Fantacci, Silvia Figini, Alessandro Lascialfari, Francesco Laruina, Piernicola Oliva, Stefano Piffer, Ian Postuma, Lisa Rinaldi, Cinzia Talamonti, and Alessandra Retico. Quantification of pulmonary involvement in COVID-19 pneumonia by means of a cascade of two U-nets: training and assessment on multiple datasets using different annotation criteria. *International Journal of Computer Assisted Radiology and Surgery*, 2021.

[91] Francesca Lizzi, Stefano Atzori, Giacomo Aringhieri, Paolo Bosco, Carolina Marini, Alessandra Retico, Antonio C. Traino, Davide Caramella, and M. Evelina Fantacci. Residual convolutional neural networks for breast density classification. *BIOINFORMATICS 2019 - 10th International Conference on Bioinformatics Models, Methods and Algorithms, Proceedings; Part of 12th International Joint Conference on Biomedical Engineering Systems and Technologies, BIOSTEC 2019*, (Biostec):258–263, 2019.

[92] Francesca Lizzi, Francesca Brero, Raffaella Fiamma Cabini, Maria Evelina Fantacci, Stefano Piffer, Ian Postuma, Lisa Rinaldi, and Alessandra Retico. Making data big for a deep-learning analysis: Aggregation of public COVID-19 datasets of lung computed tomography

scans. *Proceedings of the 10th International Conference on Data Science, Technology and Applications, DATA 2021*, (Data):316–321, 2021.

[93] Francesca Lizzi, Francesco Laruina, Piernicola Oliva, Alessandra Retico, and Maria Evelina Fantacci. Residual Convolutional Neural Networks to Automatically Extract Significant Breast Density Features. In Mario Vento, Gennaro Percannella, Sara Colantonio, Daniela Giorgi, Bogdan J Matuszewski, Hamideh Kerdegari, and Manzoor Razaak, editors, *Computer Analysis of Images and Patterns*, pages 28–35. Springer International Publishing, 2019.

[94] Francesca Lizzi, Camilla Scapicchio, Francesco Laruina, Alessandra Retico, and Maria E Fantacci. Convolutional Neural Networks for Breast Density Classification: Performance and Explanation Insights, 2022.

[95] Magnus Løberg, Mette Lise Lousdal, Michael Bretthauer, and Mette Kalager. Benefits and harms of mammography screening. *Breast Cancer Research*, 17(1), 2015.

[96] Stella Lowry and Gordon Macpherson. A blot on the profession? *British Medical Journal (Clinical research ed.)*, 296(6625):865, 1988.

[97] Bing Ma. Mammography. In Muhammad Maqbool, editor, *An Introduction to Medical Physics*, pages 199–219. Springer International Publishing, Cham, may 2017.

[98] Jun Ma, Yixin Wang, Xingle An, Cheng Ge, Ziqi Yu, Jianan Chen, Qiongjie Zhu, Guoqiang Dong, Jian He, Zhiqiang He, Tianjia Cao, Yuntao Zhu, Ziwei Nie, and Xiaoping Yang. Toward data-efficient learning: A benchmark for COVID-19 CT lung and infection segmentation. *Medical Physics*, 2020.

[99] Jun Ma, Yixin Wang, Xingle An, Cheng Ge, Ziqi Yu, Jianan Chen, Qiongjie Zhu, Guoqiang Dong, Jian He, Zhiqiang He, Tianjia Cao, Yuntao Zhu, Ziwei Nie, and Xiaoping Yang. Toward data-efficient learning: A benchmark for COVID-19 CT lung and infection segmentation. *Medical Physics*, 48(3):1197–1210, 2021.

[100] Zeev V. Maizlin and Patrick M. Vos. Do we really need to thank the beatles for the financing of the development of the computed tomography scanner? *Journal of Computer Assisted Tomography*, 36(2):161–164, 2012.

[101] C Maple. Geometric design and space planning using the marching squares and marching cube algorithms. In *2003 International Conference on Geometric Modeling and Graphics, 2003. Proceedings*, pages 90–95, 2003.

[102] V A McCormack. Breast Density and Parenchymal Patterns as Markers of Breast Cancer Risk: A Meta-analysis. *Cancer Epidemiology Biomarkers & Prevention*, 15(6):1159–1169, 2006.

[103] Diana L Miglioretti, Jane Lange, Jeroen J van den Broek, Christoph I Lee, Nicolien T van Ravesteyn, Dominique Ritley, Karla Kerlikowske, Joshua J Fenton, Joy Melnikow, Harry J de Koning, and Rebecca A Hubbard. Radiation-Induced Breast Cancer Incidence and Mortality From Digital Mammography Screening: A Modeling Study. *Annals of Internal Medicine*, 164(4):205, 2016.

[104] Aly A. Mohamed, Wendie A. Berg, Hong Peng, Yahong Luo, Rachel C. Jankowitz, and Shandong Wu. A deep learning method for classifying mammographic breast density categories. *Medical Physics*, 45(1):314–321, 2018.

[105] Stephen M. Moore, David R. Maffitt, Kirk E. Smith, Justin S. Kirby, Kenneth W. Clark, John B. Freymann, Bruce A. Vendt, Lawrence R. Tarbox, and Fred W. Prior. De-identification of medical images with retention of scientific research value. *Radiographics*, 35(3):727–735, 2015.

[106] S P Morozov, A E Andreychenko, N A Pavlov, A V Vladzymyrskyy, N V Ledikhova, V A Gombolevskiy, I A Blokhin, P B Gelezhe, A V Gonchar, and V.Yu. Chernina. MosMedData: Chest CT Scans with COVID-19 Related Findings Dataset. *medRxiv*, page 2020.05.20.20100362, jan 2020.

[107] Emanuele Neri, Francesca Coppola, Vittorio Miele, Corrado Bibbolino, and Roberto Grassi. Artificial intelligence: Who is responsible for the diagnosis? *Radiologia Medica*, 125(6):517–521, 2020.

[108] Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks. 2016.

[109] NIfTI format. https://nifti.nimh.nih.gov/, accessed on 07/01/2022.

[110] A Oliver, J Freixenet, R Marti, J Pont, E Perez, E R E Denton, and R Zwiggelaar. A Novel Breast Tissue Density Classification Methodology. *IEEE Transactions on Information Technology in Biomedicine*, 12(1):55–65, 2008.

[111] Seong Ho Park, Jaesoon Choi, and Jeong Sik Byeon. Key principles of clinical validation, device approval, and insurance coverage decisions of artificial intelligence. *Korean Journal of Radiology*, 22(3):442–453, 2021.

[112] Seong Ho Park and Kyunghwa Han. Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction 1 REVIEW: Evaluation of Artificial Intelligence Tools for Diagnostic or Predictive Analysis Park and Han. *radiology.rsna.org n Radiology Radiology*, 286(3—March), 2018.

[113] F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[114] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[115] Filippo Pesapane, Marina Codari, and Francesco Sardanelli. Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *European Radiology Experimental*, 2(1), 2018.

[116] Dorian Peters, Karina Vold, Diana Robinson, and Rafael A. Calvo. Responsible AI—Two Frameworks for Ethical Design Practice. *IEEE Transactions on Technology and Society*, 1(1):34–47, 2020.

[117] S Petroudi, T Kadir, and M Brady. Automatic classification of mammographic parenchymal patterns: a statistical approach. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439)*, pages 798–801, Cancun, Mexico, 2003. IEEE.

[118] Trong Huy Phan and Kazuma Yamamoto. Resolving Class Imbalance in Object Detection with Weighted Cross Entropy Losses. *arXiv*, 2020.

[119] Marc Prensky. Innovate: Journal of Online Education H. Sapiens Digital: From Digital Immigrants and Digital Natives to Digital Wisdom. *Innovate: Journal of Online Education*, 5(3), 2009.

[120] Mathias Prokop, Wouter Van Everdingen, Tjalco Van Rees Vellinga, Henriëtte Quarles Van Ufford, Lauran Stöger, Ludo Beenen, Bram Geurts, Hester Gietema, Jasenko Krdzalic, Cornelia Schaefer-Prokop, Bram Van Ginneken, and Monique Brink. CO-RADS: A Categorical CT Assessment Scheme for Patients Suspected of Having COVID-19-Definition and Evaluation. *Radiology*, 296(2):E97–E104, 2020.

[121] Félix Renard, Soulaimane Guedria, Noel De Palma, and Nicolas Vuillerme. Variability and reproducibility in deep learning for medical image segmentation. *Scientific Reports*, 10(1):1–16, 2020.

[122] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat]*, 2016.

[123] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9351:234–241, 2015.

[124] Holger R. Roth, Chen Shen, Hirohisa Oda, Masahiro Oda, Yuichiro Hayashi, Kazunari Misawa, and Kensaku Mori. Deep learning and its application to medical image segmentation. pages 1–6, 2018.

[125] Nasibeh Saffari, Hatem A. Rashwan, Mohamed Abdel-Nasser, Vivek Kumar Singh, Meritxell Arenas, Eleni Mangina, Blas Herrera, and Domenec Puig. Fully automated breast density segmentation and classification using deep learning. *Diagnostics*, 10(11):1–20, 2020.

[126] Ravi K. Samala, Heang Ping Chan, Lubomir Hadjiiski, Mark A. Helvie, Jun Wei, and Kenny Cha. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. *Medical Physics*, 43(12):6654–6666, 2016.

[127] C Scapicchio, F Lizzi, and M E Fantacci. Explainability of a CNN for breast density assessment. *Nuovo Cimento*, pages 1–4, 2021.

[128] James H. Scatliff and Peter J. Morris. From Roentgen to magnetic resonance imaging: the history of medical imaging. *North Carolina medical journal*, 75(2):111–113, 2014.

[129] Hyunseok Seo, Masoud Badiei Khuzani, Varun Vasudevan, Charles Huang, Hongyi Ren, Ruoxiu Xiao, Xiao Jia, and Lei Xing. Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications. *Medical Physics*, 47(5):148–167, 2020.

[130] J. M. Seo, E. S. Ko, B. K. Han, E. Y. Ko, J. H. Shin, and S. Y. Hahn. Automated volumetric breast density estimation: A comparison with visual assessment. *Clinical Radiology*, 68(7):690–695, 2013.

[131] Atefeh Shahroudnejad. A Survey on Understanding, Visualizations, and Explanation of Deep Neural Networks. 2021.

[132] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[133] E A Sickles, C J D'Orsi, L W Bassett, and Et al. ACR BI-RADS®Atlas, Breast Imaging Reporting and Data System, 2013.

[134] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2019: Cancer Statistics, 2019. *CA: A Cancer Journal for Clinicians*, 69(1):7–34, 2019.

[135] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70(1):7–30, 2020.

[136] Alberto Signoroni, Mattia Savardi, Sergio Benini, Nicola Adami, Riccardo Leonardi, Paolo Gibellini, Filippo Vaccher, Marco Ravanelli, Andrea Borghesi, Roberto Maroldi, and Davide Farina. Bs-net: learning

covid-19 pneumonia severity on a large chest x-ray dataset. *Medical Image Analysis*, page 102046, 2021.

[137] Patrice Y. Simard, Dave Steinkraus, and John C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, volume 2003-Janua, pages 958–963. IEEE Computer Society, 2003.

[138] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pages 1–14, 2015.

[139] Chiara Sottocornola, Antonio Traino, Patrizio Barca, Giacomo Aringhieri, Carolina Marini, Alessandra Retico, Davide Caramella, and Maria Evelina Fantacci. Evaluation of Dosimetric Properties in Full Field Digital Mammography (FFDM) - Development of a New Dose Index. In *Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 1: BIODEVICES,*, pages 212–217. INSTICC, SciTePress, 2018.

[140] Marthe Stevens, Rik Wehrens, and Antoinette de Bont. Conceptualizations of Big Data and their epistemological claims in healthcare: A discourse analysis. *Big Data and Society*, 5(2):1–21, 2018.

[141] A B Strong and R A A Hurst. EMI patents on computed tomography; history of legal actions. *The British Journal of Radiology*, 67(795):315–316, mar 1994.

[142] Aaron Stupple, David Singerman, and Leo Anthony Celi. The reproducibility crisis in the age of digital medicine. *npj Digital Medicine*, 2(1):1–3, 2019.

[143] J Suckling, J Parker, D Dance, S Astley, I Hutt, C Boggis, I Ricketts, Emmanuel Stamatakis, N Cerneaz, Sl Kok, P Taylor, D Betal, and J Savage. Mammographic Image Analysis Society (MIAS) database v1.21. 2015.

[144] The Independent UK Panel on Breast Cancer Screening, M G Marmot, D G Altman, D A Cameron, J A Dewar, S G Thompson, and M Wilcox.

The benefits and harms of breast cancer screening: an independent review: A report jointly commissioned by Cancer Research UK and the Department of Health (England) October 2012. *British Journal of Cancer*, 108(11):2205–2240, 2013.

[145] Jonathan Tyrer, Stephen W Duffy, and Jack Cuzick. A breast cancer prediction model incorporating familial and personal risk factors. *Statistics in Medicine*, 23(7):1111–1130, apr 2004.

[146] Stylianos D Tzikopoulos, Michael E Mavroforakis, Harris V Georgiou, Nikos Dimitropoulos, and Sergios Theodoridis. A fully automated scheme for mammographic segmentation and classification based on breast density and asymmetry. *Computer Methods and Programs in Biomedicine*, 102(1):47–63, apr 2011.

[147] Giske Ursin and Samera A. Qureshi. Mammographic density - A useful biomarker for breast cancer risk in epidemiologic studies. *Norsk Epidemiologi*, 19(1):59–68, 2009.

[148] Joris Van De Velde, Tom Vercauteren, Werner De Gersem, Johan Wouters, Katrien Vandecasteele, Philippe Vuye, Frank Vanpachtenbeke, Katharina D'Herde, Ingrid Kerckaert, Wilfried De Neve, and Tom Van Hoof. Reliability and accuracy assessment of radiation therapy oncology group-endorsed guidelines for brachial plexus contouring. *Strahlentherapie und Onkologie*, 190(7):628–635, 2014.

[149] Bram van Ginneken. Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning. *Radiological Physics and Technology*, 10(1):23–32, 2017.

[150] Eva M. Van Rikxoort, Bartjan De Hoop, Max A. Viergever, Mathias Prokop, and Bram Van Ginneken. Automatic lung segmentation from thoracic computed tomography scans using a hybrid approach with error detection. *Medical Physics*, 36(7):2934–2947, 2009.

[151] Gary M Weiss and Foster Provost. Learning when training data are costly: The effect of class distribution on tree induction. *Journal of Artificial Intelligence Research*, 19:315–354, 2003.

[152] Rephael Wenger. *Isosurfaces: geometry, topology, and algorithms.* CRC Press, 2013.

[153] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, Kanta Dihal, and Stephen Cave. *Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research.* 2019.

[154] WHO. *Ethics and Governance of Artificial Intelligence for Health Ethics and Governance of Artificial Intelligence for Health 2.* 2021.

[155] John N Wolfe. Risk for breast cancer development determined by mammographic parenchymal pattern. *Cancer*, 37(5):2486–2492, may 1976.

[156] 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*, pages 1–1, 2020.

[157] Jinzhong Yang, Greg Sharp, Harini Veeraraghavan, Wouter van Elmpt, Andre Dekker, Tim Lustberg, and Mark. Gooding. Data from Lung CT Segmentation Challenge. The Cancer Imaging Archive., 2017.

[158] Ran Yang, Xiang Li, Huan Liu, Yanling Zhen, Xianxiang Zhang, Qiuxia Xiong, Yong Luo, Cailiang Gao, and Wenbing Zeng. Chest CT Severity Score: An Imaging Tool for Assessing Severe COVID-19. *Radiology: Cardiothoracic Imaging*, 2(2):e200047, 2020.

[159] Matthew D Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks BT - Computer Vision – ECCV 2014. pages 818–833, Cham, 2014. Springer International Publishing.

[160] Xingquan Zhu and Xindong Wu. Class Noise vs. Attribute Noise: A Quantitative Study. *Artificial Intelligence Review*, 22(3):177–210, 2004.

[161] Roberto V. Zicari, John Brodersen, James Brusseau, Boris Dudder, Timo Eichhorn, Todor Ivanov, Georgios Kararigas, Pedro Kringen, Melissa McCullough, Florian Moslein, Naveed Mushtaq, Gemma Roig, Norman Sturtz, Karsten Tolle, Jesmin Jahan Tithi, Irmhild van Halem, and Magnus Westerlund. Z-Inspection ® : A Process to Assess Trustworthy AI . *IEEE Transactions on Technology and Society*, 2(2):83–97, 2021.

[162] Paul Zikopoulos, Chris Eaton, and IBM. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Osborne Media, 1st edition, 2011.

[163] Konrad Zolna, Krzysztof J. Geras, and Kyunghyun Cho. Classifier-agnostic saliency map extraction. *Computer Vision and Image Understanding*, 196:1–27, 2020.