Alma Mater Studiorum · University of Bologna

School of Science
Department of Physics and Astronomy
Master Degree in Physics

THESIS TITLE

Supervisor:
Prof.Enrico Giampieri

Co-supervisor: (optional)
Dr. Lidia Strigari

Submitted by:
Lorenzo Spagnoli

Academic Year 2020/2021

Todo list

Abstract

Since the start of 2020 *Sars-COVID19* has given rise to a world-wide pandemic. In an attempt to slow down the fast and uncontrollable spreading of this disease various prevention and diagnostic methods have been developed. In this thesis, out of all these various methods, the attention is going to be put on Machine Learning methods used to predict prognosis that are based, for the most part, on data originating from medical images.

The techniques belonging to the field of radiomics will be used to extract information from images segmented using a software available in the hospital that provided the clinical data as well as the images. The usefulness of different families of variables will be evaluated through their performance in the methods used, namely Lasso regularized regression and Random Forest. Dimensionality reduction techniques will be used to attain a better understanding of the dataset at hand.

Following a first introductiory chapter in the second chapter a basic theoretical overview of the necessary core concepts that will be needed throughout this whole work will be provided and then the focus will be shifted on the various methods and instruments used in the development of this thesis. The third is going to be a report of the results and finally some conclusions will be derived from the previously presented results. It will be concluded that the segmentation and feature extraction step is of pivotal importance in driving the performance of the predictions. In fact, in this thesis, it seems that the information from the images adds no significant predictive power to that derived from the clinical data. This can be interpreted in two ways: first it can be taken as a symptom that the more complex *Sars-COVID19* cases are still too difficult to be segmented automatically, or semi-automatically by untrained personnel, which will lead to counter-intuitive results further down the analysis pipeline. Secondly it can be taken to show that the performance of clinical variables can be reached by radiomic features alone in a semi-automatic pipeline, which could aid in reducing the workload imposed on medical professionals in case of pandemic.

Contents

Chapter 1

Introduction

Nowadays everybody knows of *Sars-COVID19* which, since the start of 2020, has made necessary a few world-wide quarantines forcing everybody in self-isolation. It is also well known that, among the main complications and features of this virus, symptoms gravity as well as the rate of deterioration of the conditions are some of the most relevant and problematic. In some cases asymptomatic or near to asymptomatic people may, in the span of a week, get to conditions that require hospital admission. This peculiarity is also what heavily complicates the triage process, since trying to predict with some degree of accuracy the prognosis of the patient at admission is a thoroughly complex task.

In this thesis the aim will be to use data, specifically including data that cannot be easily interpreted by humans, to try various methods to predict a couple of clinical outcomes, namely the death of the patient or the admission in the Intensive Care Unit (ICU), while assessing their performance.

These analyses will be carried out on a dataset of 434 patients with different variables associated to every person. A part of the variables, which will be called clinical and radiological, are defined by humans and are generally discrete in nature but mostly boolean. The most part of the available variables, however, will be image-derived following the approaches used in the field of radiomics. While the utility of clinical variables, such as age, obesity and history of smoking, is very straightforward it's interesting and helpful to understand the basis behind the utility of radiomic and radiological features.

Generally speaking it's clear that images have the ability to convey a slew of useful images, this is expecially true in the medical field where digital images are used to inspect also the internal state of the patient giving far more detailed information than that obtainable by visual inspection at the hand of medical professionals.

Among the ways in which *Sars-COVID19* can manifest himself the one that is most relevant to the scopes of this thesis is pneumonia and the complications that stem from it. Some of these complications, which are not specific of *Sars-COVID19* but can happen in any pneumonia case, display very peculiar patterns when visualizing the lungs through CT exams.

These patterns are due to the pulmonary response to inflammation which may lead to thickening of the bronchial and alveolar structures up to pleural effusions and collapsed lungs. Without going too much in clinical detail what is of interest is how these condition manifest themselves in the CT exams:

1. **Ground Glass Opacity(GGO)**:

Small diffused changes in density of the lung structure cause a hazy look in the affected region. This complicates the individuation of pulmonary vessels.

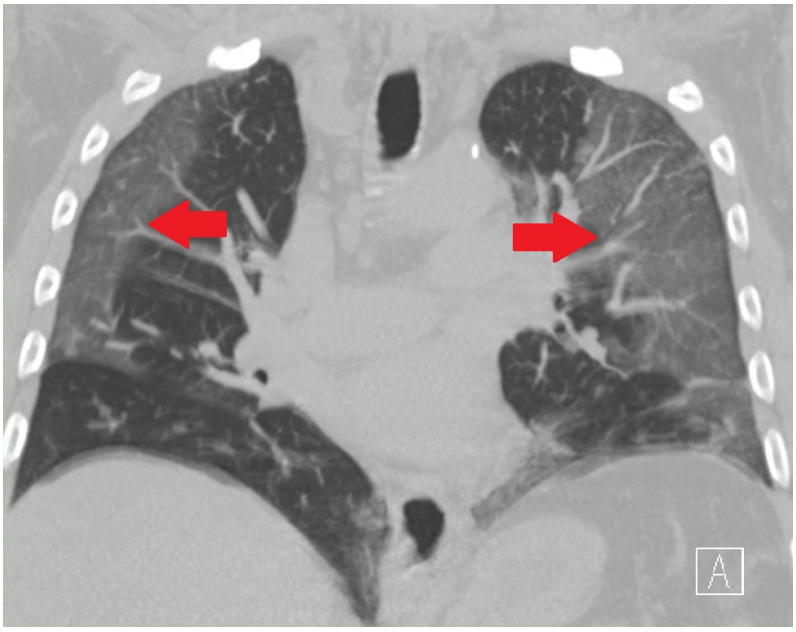
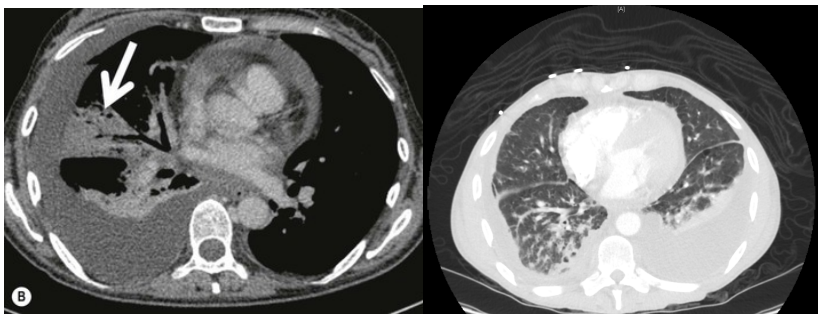


Figure 1.1: Example of GGO

2. **Lung Consolidations**:
   Heavier damage reflects in whiter spots in the lung as the surface more closely resembles outside tissue instead of normal air. The consolidation refer to presence of fluid, cells or tissue in the alveolar spaces

3. **Crazy paving**:
   When GGOs are superimposed with inter-lobular and intra-lobular septal thickening.



(a) **(b)**

Figure 1.2: Differences between a collapsed lung (a) and pleural effusion(b)

4. **Collapsed Lungs and Pleural Effusion**:
   Both of these manifest themself as regions of the lungs that take the same coloring as that of tissue outside the lung. The main difference between the two is that

collapsed lungs are somewhat rigid structures, they can occur in singular lobes of the lung and stay where they occur. Pleural effusions, however, are actually fluid being located in the lung instead of air. As such these lesions usually are located 'at the bottom' of the lung in which they happen and migrate to the lowest part of the lung according to the position of the patient.

Having these manifestation it's clear that they are mainly textural and intensity-like changes in the normal appearance of the lungs. However, whereas these properties can be easily described in a qualitative and subjective way, it's rather complex to describe them in a quantitative and objective way.

The field of radiomics, when coupled with digital images and preprocessing steps, which must include image segmentation, is exactly what undertakes this daunting task. Radiomics comes from the the combination of radiology and the suffix -*omics*, which is characteristic of high-throughput methods that aim to generate a large quantity of numbers, called biomarkers or features. As such it uses very precise and strict mathematical definitions to quantify in various ways either shape, textural or intensity based properties of the radiological image under analysis.

Given the large numerosity of the features produced by radiomics it's necessary to analyze these kinds of data with methods that rely on Machine Learning and their ability to address high-dimensional problems, be it in a supervised or unsupervised way.

Starting from these premises this thesis will be divided in a few chapters and sections. The first step will be taken by providing the general theoretical background regarding the aforementioned topics and techniques, this will be followed by a desscription of the data in use as well as a presentation of the analysis methods and resources used. Finally the results of the methods described will be presented and from them a set of concluding remarks will be set forth.

1.1 Medical Images

In this section the objective is to simply provide a set of basic definitions pertaining to images as well as a general introduction to the methods used to create said images. Firstly images are a means of representing in a visual way a physical object or set thereof, when talking about them it's common to refer specifically to digital images.

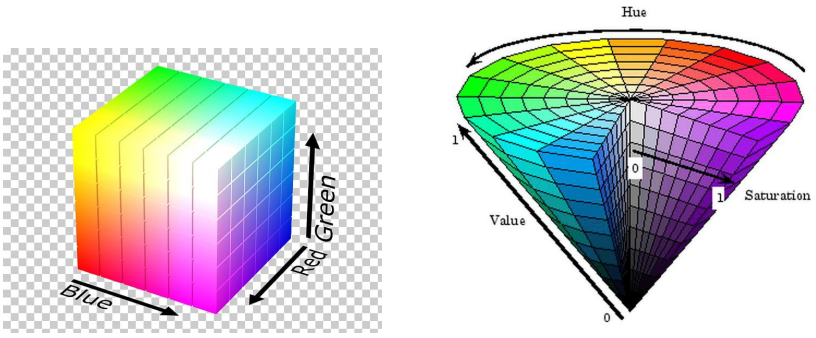
Definition 1.1.1 (Digital Image)**.** A numerical representation of an object; more specifically an ordered array of values representing the amount of radiation emitted (or reflected) by the object itself. The values of the array are associated to the intensity of the radiation coming from the physical object; to represent the image these values need to be associated to a scale and then placed on a discrete 2D grid. To store these intensities the physical image is divided into regular rectangular spacings, each of which is called pixel [1] , to form a 2D grid; inside every spacing is then stored a number (or set thereof) which measures the intensity of light, or color, coming from the physical space corresponding to that grid-spacing.

The term digital refers to the discretization process that inherently happens in storage of the values, called pixel values, as well as in arranging them within the grid. It's

[1]The term pixel seems to originate from a shortening of the expression Picture's (pics=pix) Element(el). The same hold for voxel which stands for Volume Element

possible to generalize from 2D images to 3D volumes, simply by stacking images of the same object obtained at different depths. In this context, the term pixel is substituted by voxel, however since they are used interchangeably in literature they will, from now on, be considered equivalent.

Generally pixel values stored as integers p$\in$ [0,2^n-1] with p,n$\in \mathbb{N}$ or as p$\in$[0,1] with p$\in \mathbb{R}$, the type of value stored within each pixel changes the nature of the image itself.
A single value is to be intended as the overall intensity of light coming from the part of the object contained corresponding to the gridspace and is used for a gray-scale representation, a set of three[2] or four[3] values can be intended as a color image.



(a) RGB color space on a cube **(b)** HSV cone

Figure 1.3: Examples of color spaces

There are a lot of possible scales for representation[4], which are sometimes called color-spaces, however the most noteworthy in the scope of this work is the Hounsfield unit (HU) scale.

Definition 1.1.2 (Hounsfield unit (HU))**.** A scale used specifically to describe radiodensity, frequently used in the context of CT (Computed Tomography) exams. The values are obtained as a transformation of the linear attenuation coefficient 1.5 of the material being imaged and, since the scale is supposed to be used on humans, it's defined such that water has value zero and air has the most negative value -1000. For a more in depth discussion refer to [18]

$$HU = 1000 * \frac{\mu - \mu_{H_2O}}{\mu_{H_2O} - \mu_{Air}} \tag{1.1}$$

The utility of this scale is in it's definition. Since the pixel value depends on the attenuation coefficient it's possible to individuate a set of ranges that identify, within good reason, the various tissues in the human body: for example lungs are [-700, -600] while bone can be in the [500, 1900] range.

[2]The three values correspond each to the intensity of a single color, the most commonly used set of colors is the RGB-scale (Red, Green, Blue). Further information can be found by looking into Tristimulus theory[33]

[3]Same as RGB but with four colors, the most common scale is CMYK (Cyan, Magenta, Yellow, blacK). This spectrum is mainly used in print.

[4]Besides RGB and CMYK 1.3a the most common color spaces are CIE (Commision Internationale d'Eclairage) and HSV fig:1.3b (Hue,Saturation and Value). Refer to [14] for further details

A more in depth discussion of the topics relative to Hounsfield units is going to be carried out at a later point throughout this chapter, in the meantime it's necessary to clarify what are the most important characteristics of an image:

- Spatial Resolution: A measure of how many pixel are in the image or, equivalently, how small each pixel is; a larger resolution implies that smaller details can be seen better fig:1.4.

  Can be measured as the number of pixel measured over a distance of an inch ppi(Pixel Per Inch) or as number of line pairs that can be distinguished in a mm of image lp/mm (line pair per millimeter).

- Colorr quantization: The range of the pixel values, a classic example is an 8-bit resolution which yields 256 levels of gray. A better resolution allows a better distinction of colors within the image fig:1.4.

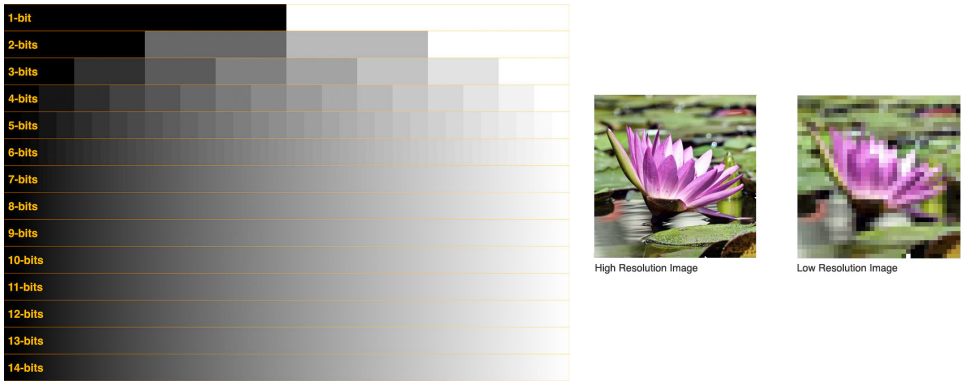


Figure 1.4: Example of visual differences in Gray-level (left) and spatial (right) resolution

- Size: Refers to the number of pixel per side of the image, for example in CT-derived images the coronal slices are usually 512x512. These numbers depend on the acquisition process and instrument but in all cases these refer to the number of rows and columns in the sampling grid as well as in the matrix representing the image.

- Data-Format: How the pixel values are stored in the file of the image.

  The most commonly used formats are .PNG and .JPG, however there are a lot of other formats. In the context of this work, which is going to be centered on medical images, the most interesting formats are going to be the nii.gz (Nifti) and the .dcm (DICOM). The first contains only the pixel value information hence it's a lighter format, it originates in the field of Neuroimaging[5], it is used mainly in Magnetic resonance images of the brain but also for CT scans and, since it contains only numeric information, it's the less memory consuming option out of the two.

  The second contains not only the image data but also some data on the patient, such as name and age, and details on how the exam was carried out, such as machine used and specifics of the acquisition routine. This format is heavier than

[5]In fact Nifti stands for Neuroimaging Informatics Technology Initiative (NIfTI)

the previous one and, for privacy purposes, is much more delicate to handle which is why anonymization of the data needs to be taken in consideration.

For a thorough description of the DICOM standard refer to [2].

The format in which the image is saved depends on the compression algorithm used to store the information within the file. These algorithms can be lossy, in which case some of the information is lost to reduce the memory needed for storage, or lossless which means that all the information is kept at the expense of memory space.

The first set of methods is preferred for storage of natural images, these are cases in which details have no importance, whereas the second set of methods is used where minute details can make a considerable difference such as in the medical field[6].

Given this set of characteristics it should now be clear that images can be thought of as array of numbers, for this reason they are often treated as matrices and, as such, there is a well defined set of valid operations and transformations that can be performed on them. All these operations and transformations, in a digital context[7], are performed via computer algorithms which allow almost perfect repeatability and massive range of possible operations.

Given the list-like nature of images one of the most natural things to do with the pixel values is to build an histogram to evaluate some of the characteristic values of their distribution, such as average, min/max, skewness, entropy.... The histogram of the image, albeit not being an unambiguous way to describe images, is very informative. When looking at an histogram it's immediately evident whether the image is well exposed and if the whole range of values available is being used optimally.

[6] A detailed description of compression algorithms is beyond the scopes of this thesis, for this reason please refer to [36] for more information

[7] As opposed to analog context, which would mean the chemical processes used at the start of photography to develop and modify the film on which the image was stored

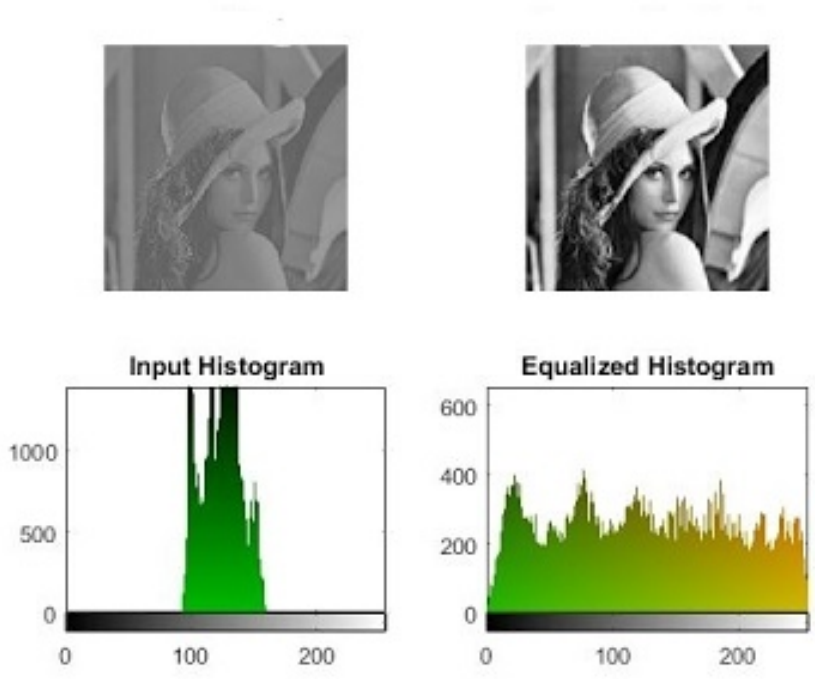


Figure 1.5: Example of differences in contrast due to histogram equalization

This leads us to the concept of *Contrast* which is a quantification of how well different intensities can be distinguished. If all the pixel values are bundled in a small range leaving most of the histogram empty then it's difficult to pick up the differences because they are small. However, if the histogram has no preferentially populated ranges then the differences in values are being showed in the best possible way fig:1.5. Note also that if looking at the histogram there are two(or more) well separated distributions it's possible that these also identify different objects in the image, which will for example allow for some basic background-foreground distinction.

Assuming they are being meaningfully used [8] all mathematical operations doable on matrices can be performed on images for this reason it would be useless to list them all. However, it's useful to provide a list of categories in which transformations can be subdivided:

1. Geometric Transformations: These are transformations that involve the following steps:
   (a) Affine transformations: Transformations that can can be performed via matrix multiplication such as rotations, scaling, reflections and translations. This step basically involves computing where each original pixel will fall in the transformed image
   (b) Interpolation: Since the coordinates of the transformed pixel might not fall exactly on the grid it might become necessary to compute a kind of average contribution of the pixel around the destination coordinate to find a most

[8] For example adding/subtracting one image to/from another can be reasonably understood, multiplying/dividing are less obvious but still used e.g. in scaling/mask imposition and change detection respectively

believable value. Examples of such methods are linear, nearest neighbour and bicubic.

2. Gray-level (GL) Transformations: Involve operating on the value stored within the pixel, these can be further subdivided as:

   (a) Point-wise: The output value at specific coordinates depends only on the output value at those same specific coordinates. Some examples are window-level operations, thresholding, negatives and non-linear operations such as gamma correction which is used in display correction. Taken p as input pixel value and q as output and given a number $\gamma \in \mathbb{R}$, gamma corrections are defined as:

   $$q = p^{\gamma} \tag{1.2}$$

   (b) Local: The output value at specific coordinates depends on a combination of the original values in a neighbourhood around that same coordinates. Some examples are all filtering operation such as edge enhancement, dilation and erosion. These filtering methods are based on performing convolutions in which the output value at each pixel is given by the sum of pixel-wise multiplication between the starting matrix and a smaller (usually 3x3 or 5x5) matrix called kernel. The output image is obtained by moving the kernel along the starting matrix following a predefined stride for example a (2,2) stride will move the kernel 2 pixels to the right and 2 down. When moving near the borders the behavour is defined by the padding of the image, most common choice for padding is zero padding, in which the image is considered to have only zeros outside of it, or no padding at all. Stride **S**, kernel shape **K** [9] and padding **P** determine the shape of the output matrix given the input dimension **W** via the following formula:

   $$OutputShape = \left[\frac{W - K + P}{S}\right] + 1 \tag{1.3}$$

   (c) Global: The output value at specific coordinates depends on all the values of the original images. Most notable operation in this category is the Discrete Fourier Transform and it's inverse which allow switching between spatial and frequency domains. It's worth noting that high frequency encode patterns that change on small scales whereas low frequencies encode regions of the image that are constant or slowly varying.

Having seen what constitutes and image and what can be done with one it becomes interesting to explore how images are obtained. The following discussion is going to introduce briefly some of the most widely used methods to obtain medical images, getting more in depth only on the modality used to obtain all the images that will be analyzed in this thesis which is Computed Tomography.

This technique was used because it's the only one that can provide information on the internal structure of an organ which is very low in density and that has parts that

[9] This formula works for square kernels, images and strides so a kernel MxM will have K=M.

are deep in the patients body. Since no metabolic process of interest is in play PET is not advisable, Ultra Sounds are used for superficial soft tissue which is not the case of the lungs and MRI, despite it's clear advantege in avoiding ionizing radiation, still has close to no acquisition protocol dedicated to lungs.

1. Magnetic Resonance Imaging (MRI): This technique is based on the phenomenon of Nuclear Magnetic Resonance(NMR) which is what happens when diamagnetic atoms are placed inside a very strong uniform magnetic field are subject to a Radio Frequency (RF) stimulus. These atoms absorb and re-emit the RF and supposing this behaviour can somehow be encoded with a positional dependence then it's possible to locate the resonant atoms given the response frequency measured. Suffices to say that this encoding is possible however the setup is very complex and the possible images obtainable with this method are very different and can emphasize very different tissue/material properties. Nothing more will be said on the topic since no data obtained with this methodology will be used. More details can be found in [8]

2. Ultra-Sound (US): The images are obtained by sending waves of frequency higher to those audible by humans and recording how they reflect back. This technique is used mainly in imaging soft peripheral tissues and the contrast between tissues is given by their different responses to sound and how they generate echo.

   The main advantages such as low cost, portability and harmlessness come at the expense of explorable depth, viewable tissues, need for a skilled professional and dependence on patient bodily composition as well as cooperation.

3. Positron Emission Tomography (PET): In this case the images are obtained thanks to the phenomenon of annihilation of particle-antiparticle, specifically of electron-positron pairs.

   The positrons come from the β^+ decay of a radio-nucleide bound to a macromolecule, which is preferentially absorbed by the site of interest [10]. Once the annihilation happens a pair of (almost) co-linear photons having (almost) the same energy of 511 keV is emitted, the detection of this pair is what allows the reconstruction of the image representing the pharmaceutical distribution within the body. The exam is primarily used in oncology given the greater energy consumption, hence nutrients absorption, of cancerous tissue and secondly this technique can be combined with CT scans to obtain a more detailed representation of the internal environment of the patient

The last technique that is going to be mentioned is Computed Tomography however, given it's relevance inside this thesis work, it seems appropriate to describe it in a dedicated section.

[10] Most commonly Fluoro-DeoxyGlucose FDG which is a glucose molecule labelled with a ^{18}F atom responsible of the β^+ decay. In general these radio-pharmaceuticals are obtained with particle accelerators near, or inside, the hospital that uses them. They are characterized by the activity measured as decay/s $\doteq$ Bq (read Becquerel) and half-life $\doteq$ $T_{\frac{1}{2}}$ which is how long it takes for half of the active atoms to decay

1.1.1 X-ray imaging and Computed Tomography (CT)

It's well known that the term x-rays is used to characterize a family of electromagnetic radiation defined by their high energy and penetrative properties. Radiation of this kind is created in various processes such as characteristic emission of atoms, also referred to as x-ray fluorescence, and Bremsstrahlung, braking radiation[11].

The discovery that "A new kind of ray"[37] with such properties existed was carried out by W. C. Roentgen in 1895, which allowed him to win the first Nobel prize in physics in the same year. Clearly the first imaging techniques that involved this radiation were much simpler than their modern counterpart, first of all they were planar and analog in nature, as well as not as refined in image quality. The first CT image was obtained in 1968 in Atkinson Morley's Hospital in Wimbledon.

Tomography indicates a set of techniques[12] that originate as an advancement of planar x-ray imaging; these techniques share most of the physical principles with planar imaging while overcoming some of it's major limitations, main of which being the lack of depth information. X-ray imaging, both planar and tomographic, involves seeing how a beam of photons changes after traversing a target, the process amounts to a kind of average of all the effects occurred over the whole depth travelled.

The way in which slices are obtained is called focal plane tomography and, as the name suggests, the basic idea is to focus in the image only the desired depth leaving the unwanted regions out of focus. This selective focusing can be obtained either by taking geometrical precautions while using analog detectors, such as screen-film cassettes, or by feeding the digital images to reconstruction algorithms to perform digitally the required operations[13].

In both planar and tomographic setting the rough description of the data acquisition process can be summarized as follows:

First x-rays are somehow generated by the machine, the quality of these x-rays is optimized with the use of filters then focused and positioned such that they mostly hit the region that needs imaging. The beam then exits the machine and starts interacting with the imaged object[14], this process causes an attenuation in the beam which depends on the materials composing the object itself. Having then travelled across the whole object it interacts with a sensor, be it film, semiconductor or other, which stores the data that will then constitute the final image. In a digital setting this final step has to be performed following a (tomographic) reconstruction algorithm which given a set of 2D projections returns a single 3D image.

In this light the interesting processes are how the radiation is created and shaped before hitting the patient and how said radiation then interacts with the matter of both the patient's body and the sensor beyond it. To explore these topics it's necessary to see:

- How these x-ray imaging machines are structured

[11] From the German terms *Bremsen* "to brake" and *Strahlung* "radiation"

[12] from the greek *Tomo* which means "to cut" and suffix -graphy to denote that it's a technique to produce images

[13] In the first case the process is referred to as *Geometric Tomography* while in the second case as *Digital Tomosynthesis*

[14] In this work it's always going to be a patient, however this process is general and is also used in industry to investigate object construction

- How x-ray and matter interact as the first traverses the second

For more information on reconstruction algorithms refer to [21] and [44].

1.1.2 Generation and management of radiation: digital CT scanners

As of the writing of this thesis, seven generations of CT scanners with different technologies exist. The conceptual structure of the machines is mostly the same, and the differences between generations also make evident those between machines. Exploiting this fact the structural description is going to be only one followed by a brief list of notable differences between generations.

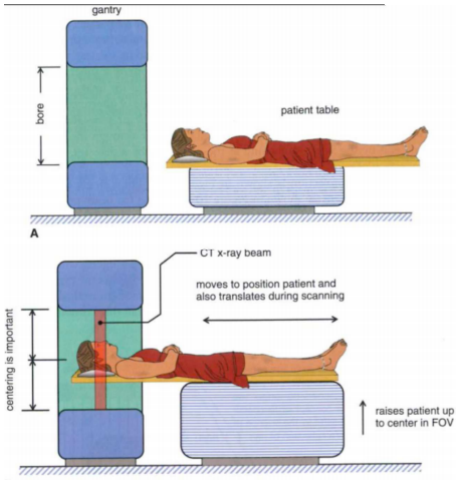


Figure 1.6: General set-up of a CT machine

The beam is generally created by the interaction of high energy particles with some kind of material, so that the particle's kinetic energy can be converted into radiation. In practice this means that an x-ray tube is encapsulated in the machine. Inside this vacuum tube charged particles[15] are emitted from the cathode, accelerated by a voltage differential and shot onto a solid anode[16]. This creation process implies that the spectrum of the produced x-rays is composed of the almost discrete peaks of characteristic emission, due to the atoms composing the target, superimposed with the continuum Bremsstrahlung radiation.

[15] Most commonly electrons
[16] Typical materials can be Tungsten, Molybdenum

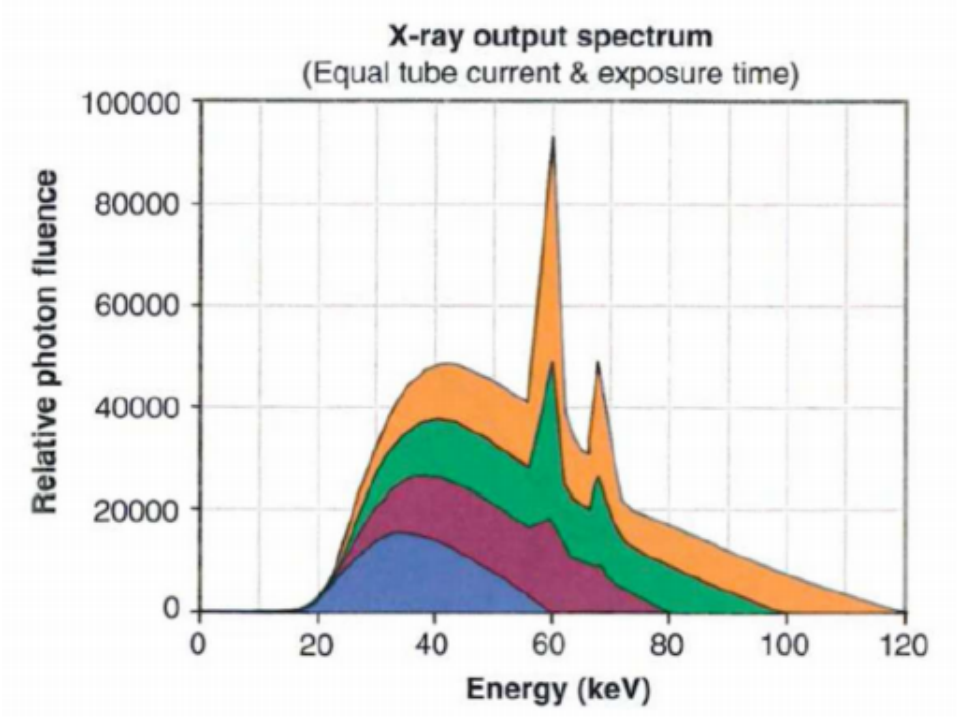


Figure 1.7: X-ray spectrum, composed of characteristic peaks and Bremsstrahlung continuum, computed at various tube voltages

Some of the main characteristics of the x-ray beam are related to this stage in the generation, the Energy of the beam is due to the accelerating voltage in the tube whereas the photon flux is determined by the electron current in the tube. Worth noting, en passant, that these two quantities can be found in the DICOM image of the exam as *kiloVolt Peak (kVP)* and *Tube current mA* and can be used to compute the dose delivered to the patient.

Other relevant characteristics in the tube are the anode material, which changes the peaks in the x-ray spectrum and time duration of the emission, which is called exposure time and influences dose as well as exposure[17].

The electron energy is largely wasted ($\sim$ 99%) as heat in the anode, which then clearly needs to be refrigerated. The remaining energy, as said before, is converted into an x-ray beam which is directed onto the patient. To reduce damage delivered to the tissues it's important that most of the unnecessary photons are removed from the beam.

Exploiting the phenomenon of beam hardening a filter, usually of the same material as the anode, is interposed between the beam and the patient to block lower energy photons from passing through thereby reducing the dose conveyed to the patient. At this point there may also be some form of collimation system which allows further shaping of the dose delivered. Having been collimated the beam traverses the patient and gets to the sensor of the machine, which nowadays are usually solid-state detectors.

[17]Exposure is a term used to identify how much light has gotten in the imaging sensor. Too high an exposure usually means the image is burnt, i.e. too bright and white, while lower exposures are usually associated to darker images. Exposure is proportional to the product of tube current and exposure time, measured in mA*s. Generally the machine handles the planning of exposure time according to treatment plan

At this point is where the differences between generations arise which, loosely speaking, can be found in the emission-detection configuration and technology.

- 1^{st} generation-Pencil Beam: A single beam is shot onto a single sensor, both sensor and beam are translated across the body of the patient and then rotated of some angle. The process is repeated for various angles. Main advantages are scattering rejection and no need for relative calibration, main disadvantage is time of the exam

- 2^{nd} generation-fan Beam: Following the same process as the previous generation the main advantage is the reduction of the time of acquisition by introducing N beam and N sensors which don't wholly cover the patient's body so still need to translate.

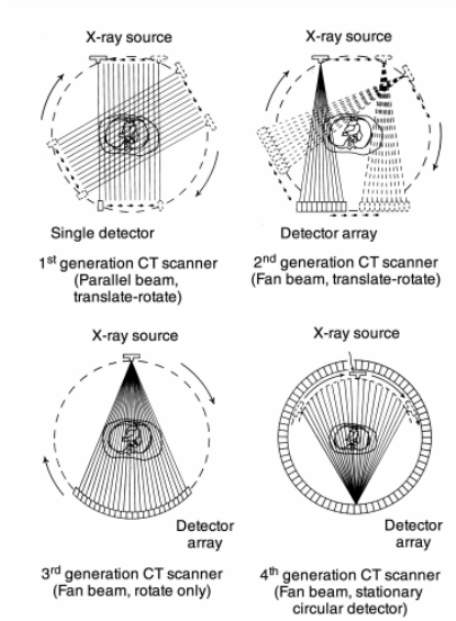


Figure 1.8: First four generations of CT scanners

- 3^{rd} generation-Rotate Rotate Geometry: Enlarging the span of the fan of beams and using a curved array of sensor a single emission of the N beams engulfs the whole body so the only motion necessary is rotation of the couple beam-sensor array around the patient.

- 4^{th} generation-Rotate Stationary Geometry: The sensors are now built to completely be around the patient so that only the beam generator has to rotate around the body

- 5^{th} generation-Stationary Stationary Geometry: The x-ray tube is now a large circle that is completely around the patient. This is only used in cardiac tomography for more information refer to [20]

- 6^{th} generation-Spiral CT: Supposing the patient is laying parallel to the axis of rotation, all previous generations acquired, along the height of the patient, a single slice at a time. In this generation as the tube rotates around the patients the bed

on which they're laying moves along the rotation axis so that the acquisition is continuous and not start-and-stop. This further reduces the acquisition time while significantly complicating the mathematical aspect of the reconstruction.

It's necessary to add another important parameter which is the pitch of the detector[18]. This quantifies how much the bed moves along the axis at each turn the tube makes around the patient. Pitches smaller than one indicate oversampling at the cost of longer acquisition times, pitches greater than one indicate shorter acquisition times at the expense of a sparser depth resolution.

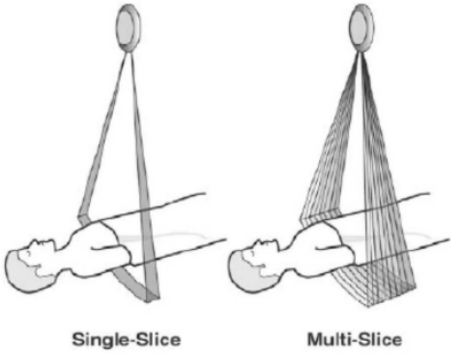


Figure 1.9: $7^{th} generation setup$

- 7^{th} generation-MultiSlice: Up to this seventh generation height-wise slice acquisition was of a singular plane, be it continuous or in a start and stop motion. In this final generation mutliple slices are acquired. Considering cilindrical coordinates with z along the axis of the machine the multiple slice acquisition is obtained by pairing a fanning out along θ and one along z of both sensor arrays and beam. This technique returns to a start and stop technology in which only $\sim 50\%$ of the total scan time is used for acquisition

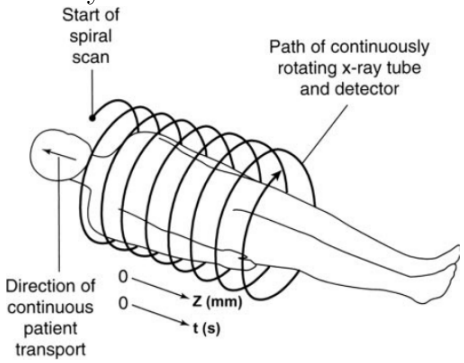


Figure 1.10: $7^{th} generation setup$

The machines used to obtain the images used in this thesis, all belonging to *IRCSS Azienda ospedaliero-universitaria di Bologna - Policlinico Sant'Orsola-Malpighi* , were distributed as shown in Figure 1.11:

[18] Once again the important parameters, such as this, can be accessed in the DICOM file resulting from the exam.

		counts	freqs
KVP	100.000	23	5,28%
	120.000	398	91,28%
	140.000	15	3,44%
Convolutional kernel	A	15	3,44%
	B	2	0,46%
	BONE	13	2,98%
	BONEPLUS	130	29,82%
	LUNG	27	6,19%
	SOFT	1	0,23%
	STANDARD	8	1,83%
	YB	29	6,65%
	YC	210	48,17%
	YD	1	0,23%
Machine	Ingenuity CT	245	56,19%
	LightSpeed VCT	179	41,06%
	iCT SP	12	2,75%
Slice Thickness	1	257	58,94%
	1,25	179	41,06%

Figure 1.11: Acquisition parameter and machine distribution. KVP indicates the KiloVoltPeak used during the acquisition, Convolutional kernels are used by the factories to indicate which reconstruction algorithm is used. Machine indicated the name of the instrument that provided the images and slice thickness indicates the vertical width of the pixel.

1. Ingenuity CT (Philips Medical Systems Cleveland): $\sim$ 56% of the exams were obtained with this machine

2. Lightspeed VCT (General Electric Healthcare, Chicago-Illinois): $\sim$ 41% of the exams in study come from this machine

3. ICT SP (Philips Medical Systems Cleveland): $\sim$ 3% of the exams were performed with this machine

1.1.3 Radiation-matter interaction: Attenuation in body and measurement

Having seen the apparatus for data collection the remaining task is to see how the information regarding the body composition can be actually conveyed by photons.

Let's first consider how a monochromatic beam of x-rays would interact with an object while passing through it. All materials can be characterized by a quantity called attenuation coefficient μ which quantifies how waves are attenuated traversing them, this energy dependent quantity is used in the Beer-Lambert law which allows computation of the surviving number of photons, given their starting number N_0 and μ:

$$N(x) = N_0 e^{-\mu(E)*x} \tag{1.4}$$

At a microscopic level the absorption coefficient will depend on the probability that a photon of a given energy E interacts with a single atom of material. This can be expressed using atomic cross section σ as:

$$\mu(E) = \frac{\rho * N_A}{A} * (\sigma_{Photoelectric}(E) + \sigma_{Compton}(E) + \sigma_{PairProduction}(E)) \tag{1.5}$$

Where ρ is material density, N_A is Avogadro's number, A is the atomic weight in grams and the distinction among the various possible interaction processes for a generic photon of high energy E is made explicit. Overall the behaviour of the cross section is the following

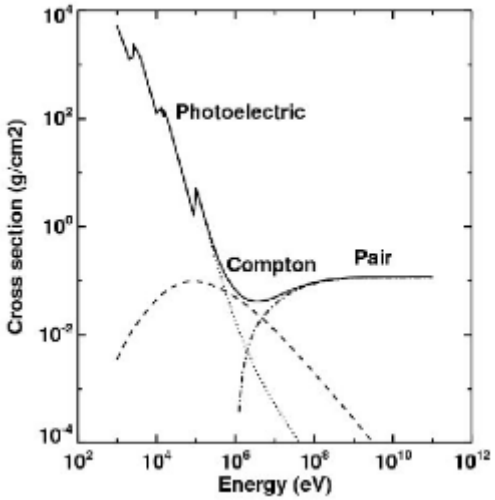
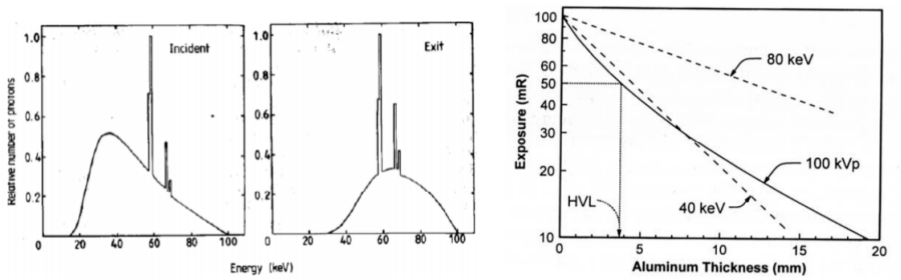


Figure 1.12: Photon cross-section in Pb

Overall, given eq:1.4, it's clear to see that the attenuation behaviour of a monochromatic beam would be linear in semi-logarithmic scale hence the name for μ "*Linear Attenuation Coefficient*".

The first complication comes from the fact that, given their generation method, the x-rays are not monochromatic but rather polychromatic. This introduces a further complication which is the phenomenon of beam hardening: lower energy x-rays interact much more likely than those at higher energies which implies that as it crosses some material the mean energy of the whole beam increases. This behaviour is exploited still within the machine, filters are interposed between anode and patient to reduce the useless part of the spectrum as shown in fig: 1.13a:



(a) Beam Filtering

(b) Polychromatic beam attenuation

Figure 1.13: Polychromatic beam behaviour

Another effect of the beam being polychromatic is that the graphical behaviour of the attenuation instead of being linear gets bent as shown in fig: 1.13b. Since the image brightness is related to the number of photons that get on the sensor it's still possible to define the contrast between two pixel p_1, p_2 as:

$$C(p_1, p_2) = \frac{N_{\gamma,p_2} - N_{\gamma,p_1}}{N_{\gamma,p_2}} \tag{1.6}$$

This formula, that connects the beam to the image, together with eq: 1.4, which connects the beam property to the patient's composition, make clear the processes by which the beam carries patient information. Another complication arises in the context of this last equation due to the phenomenon of scattering which reduces the contrast by changing the direction of the beam and introducing an element of noise. Anti-scattering grids are positioned right before the sensors to reduce this effect by allowing to reach the sensor to only the photons with the correct direction.

Damage can be classified as primary, due to ionization events within the nucleus of the cell, or secondary, due to chemical changes in the cell environment. The energy deposited per unit mass is called dose and is measured in Gy(Gray) and, as said before, depends on exposure time, current and kVP of the tube. Most of contemporary machines for CT self-regulate exposure time during the acquisition automatically using Automatic Exposure Control(AEC). Having the dose it's possible to estimate the fraction of surviving cells and, to do so, various models are used. In the clinical practice it's common to find, still within the DICOM image metadata, the information regarding Dose delivered such as CTDI (Computed Tomography Dose Index) from which it's possible to obtain the DLP (Dose Length Product) taking into consideration the total length of irradiated body. For an introduction to one of these models, the Linear Quadratic (LQ) refer to [30].

1.2 Artificial Intelligence (AI) and Machine Learning(ML)

Having clarified the type of data that will be used in this work, and having seen the general procedure used to gather it, it becomes interesting to discuss what kind of techniques will be used to analyze it.

Starting from the definition given by John McCarthy in [27] "[AI] is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.". Machine Learning (ML) is a sub-branch of AI and contains all techniques that make the computer improve performances via experience in the form of exposure to data. Practically speaking this finds it's application in classification problems, image/speech/pattern recognition, clustering, autoencoding and others.

The general workflow of Machine Learning is the following: given a dataset, the objective is to define a model or function which depends on some parameters which is able to manipulate the data in order to obtain as output something that can be evaluated via a predefined performance metric. The parameters of the model are then automatically adjusted in steps to minimize or maximize this performance metric until a stable point at which the model with the current parameters is considered finalized; one of the main

problems in this procedure is being sure that the stable point found is global and not local. The whole procedure is carried out keeping in mind that the resulting model needs to be able to generalize it's performance on data that it has never seen before, for this reason usually ML is divided in a training phase and a testing phase.

The training phase involves looking at the data and improving the performance of the model on a specific dataset[19], the testing phase involves using brand new data to evaluate the performance of the model obtained in the preceding phase. Machine Learning techniques can be further grouped into the following categories:

1. Supervised Learning: In this type of ML the model is provided with the input data as well as the correct expected output, which is hence called *label*. The objective of the model is to obtain an output as similar to the labels as possible, while also retaining the best possible generalization ability in predicting never seen before data. Some problems that benefit from the use of these techniques are regression and classification problems.

2. Unsupervised Learning: As the name suggests this category of models trains on the data alone, without having the labels available by minimizing some metric defined from the data. For example clustering techniques try to find a set of groups in the data such that the difference within each group is minimal while the difference among groups is maximal, ideally producing dense groups, called clusters, that are each well separated from all the others. Other techniques in this family are Principal Component Analysis (PCA) and autoencoding but the general objective is to infer some kind of structure within the data and the relation between data points.

3. Reinforcement Learning: This kind of ML is well suited for data which has a clear sequential structure in which the requiered task is to develop good long term planning. Broadly speaking the general set-up is that given a set of (state$_t$, action, reward, state$_{t+1}$) these techniques try to maximize the cumulative reward[20]. The main applications of these techniques are in Autonomous driving and learning how to play games

The following methods are those directly involved in this work.

1.2.1 Regression, Classification and Penalization

Regression and Classification are methods used to make predictions via supervised learning by understanding the input-output relation in continuous and discrete cases respectively.

The most basic example of regression is linear regression which consists in finding the slope and intercept of a line passing through a set of points. A branch of classification thats similar to linear regression is logistic regression, this translates in finding out wether or not each point belongs to a certain category given it's properties and can be practically

[19] Usually in studies there is a single dataset which is split into a train-set and a test-set, in some cases if the model is good it can be validated prospectively, which means that it's performance is evaluated on data that did not yet exist at the time of birth of the model

[20] i.e. the sum of the rewards obtained at all previous time steps

thought of, for a binary classification,as a fitting procedure such that the output can be either 0 for one category and 1 for the other, this is usually done using the logistic funtion, also called sigmoid, which compresses $\mathbb{R}$ in [0,L] as seen in 1.14. L represents the maximum value desidered, x_0 is the midpoint and k is the steepness.

Note that for multiple categories it would conceptually suffice to add sigmoids with different centers to obtain a step function in which each step corresponds to a category, in this case the procedure is usually called multinomial regression.

$$\sigma_{(x)} = \frac{L}{1 + e^{-k(x-x_0)}} \tag{1.7}$$

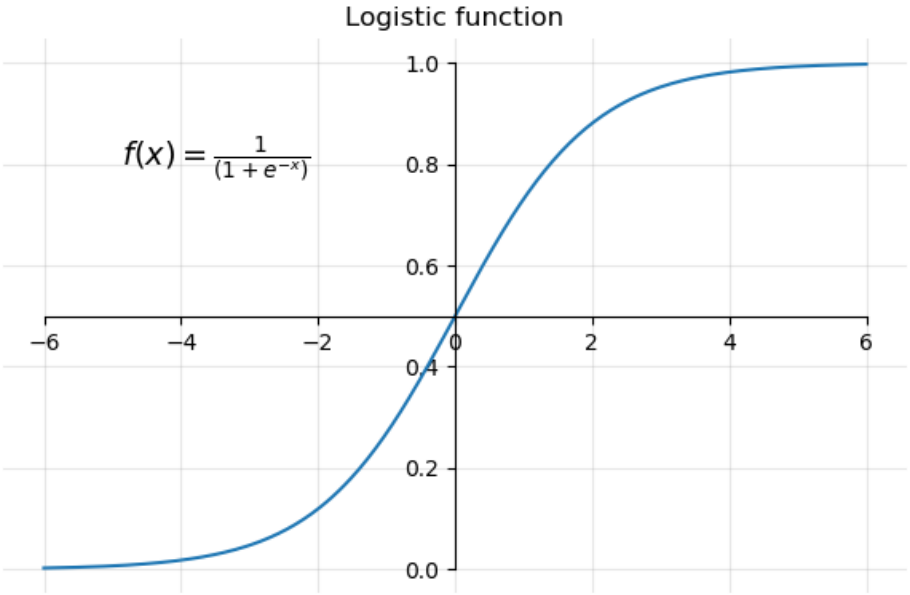


Figure 1.14: sigmoid function with L=1, x_0=0, k=1

To give a general intuition of the procedure, without going in unnecessary details, let's only consider linear regression; the theory on which it is founded is based on the assumption that the residuals, i.e. the distance model-label, are normally distributed. Under this assumption the parameters can be found by changing them and trying to minimize the sum of squared residuals.

When each datapoint is characterized by a lot of different features the procedure is called multiple linear regression, the task becomes finding out how much each feature contributes in predicting the output within a weighted linear combination of features. Practically speaking this can be done in matricial form, supposing that each of the m datapoints has n features associated.

Let $\mathbf{X}$ be a matrix with n+1[21] columns and m rows and let $\mathbf{Y}$ be a vector with m entries. Let also θ be a vector of n+1 entries, one for each feature plus the intercept θ_0, then we are supposing that:

$$y_i = \Sigma_{j=0}^{n+1} x_{i,j} * \theta_i + \epsilon_i => \mathbf{Y} = X * \theta + \epsilon \tag{1.8}$$

[21] n+1 because we have n features but we also want to estimate the intercept of the line, so in practice the first column will be of all ones to have the correct model shape in the following matrix multiplication

Where ϵ is the array of residuals of the model which, as said before, is supposed to contain values that are normally distributed. Minimizing the squared residuals corresponds to minimizing the following cost function:

$$J_\theta = (\mathbf{Y} - X * \theta)^T * (\mathbf{Y} - X * \theta) \tag{1.9}$$

By setting $\frac{\delta J}{\delta \theta}$=0 it can be shown that the best parameters θ^* are :

$$\theta^* = (\mathbf{X}^T * \mathbf{X})^{-1} * \mathbf{X}^T\mathbf{Y} \tag{1.10}$$

It's evident that to obtain a result from the previous operation it's necessary that $(\mathbf{X}^T * \mathbf{X})$ be invertible, which in turn requires that there be no correlated features and that the features be less than the datapoints.

To solve this problem the first step is being careful in choosing the data that goes through the regression, which may even involve some preprocessing.

The second step is called Regularization, it involves adding a penalty to the cost function by adding a small quantity along the diagonal of the matrix.

The nature of this small quantity changes the properties of the regularization procedure, the most famous penalties are Lasso, Ridge and ElasticNet. In practical terms the shape of the penalty determines how much and how fast the slopes relative to the features can be shrunk.

1. Ridge: Adds $\delta^2 * \Sigma_{j=1}^{n+1}\theta_j^2$, is called also L^2 regularization since it adds the L^2 norm of the parameter vector. This penalty can only shrink parameters asymptotically to zero but never exactly, which means that all features will always be used, even with very small contributions

2. Lasso: Adds $\frac{1}{b} * \Sigma_{j=1}^{n+1}|\theta_j|$, is called also L^1 regularization since it adds the L^1 norm of the parameter vector. This penalty can shrink parameters to exactly zero, getting rid of the useless variables within the model.

3. ElasticNet: Adds $\lambda * [\frac{1-\alpha}{2} * \Sigma_{j=1}^{n+1}\theta_j^2 + \alpha * \Sigma_{j=1}^{n+1}|\theta_j|]$, evidently this is a midway between the Lasso and Ridge methods, where the balance is dictated by the value of α.

There are no clear overall advatages in the choice between Ridge and Lasso regularization, however there are substantial differences that can aid in the choice.

Ridge regression shrinks the parameters but never exactly to zero hence it does not perform feature selection and when correlated features are used their coefficients will be similar, rather than shrunk to zero. Hence Ridge still simplifies the model but it doesn't reduce the number of features, this is ideal in cases in which one wants to keep all of the available features or when one expects that most predictors drive the response.

Lasso regularization has the ability to shrink parameters exactly to zero and does so in cases in which variables are correlated with one another. However the choice on which feature to keep is random if the variables are highly correlated. The advantages of Lasso, namely the feature selection it produces while also improving the prediction on the data, come at the cost of difficult to interpret results in some cases as well as a limit in the maximum number of feature that can survive the procedure. [22]. This means that

[22]Out of k features relative to m datapoints, with k m only m features can survive a Lasso regularization even if all k are relevant.

practivally one would choose Lasso regression in cases in which the expectation is that only a few variables drive the behaviour of the response.

Elastic net, which was born from a critique of the too data-dependent shrinkage of parameters, is a method that combines Lasso and Ridge. In this sense the model will still be simplified as it happened in Ridge and Lasso. The surviving parameters will be less then those estimated by Ridge and more than those obtained with a Lasso and the value of the coefficients will be smaller than those in Lasso but larger than those obtained in the case of Ridge.

For more in depth information refer to [17].

Given these last considerations, as well as the number of features and their interpretability, the choice was made to use Lasso regression in order to keep as few variables as possible while still maintaining good predictive power.

Since the whole foundation is the normality of the residuals it's important that the data behaves somewhat nicely in this regard. Changing the data to modify the residuals is not straightforward, a proxy for this procedure is to preprocess the data by manipulating the distribution of the features to make them as close to normal as possible. To this end some of the operations that can be done are:

1. Standard Scaling: This amounts to subtracting the mean of the distribution to the feature and dividing by the standard deviation so that the resulting distribution is somewhat centered around zero and has close to unitary standard deviation

2. Boxcox transform: When distributions are heavy-tailed, like a gamma distribution would be, this transform is used to find the optimal power-law that more closely turns the data in a normal distribution. More specifically the data is transformed according to:

$$Data_{Transformed}(\lambda) = \begin{cases} \frac{data^{\lambda}-1}{\lambda} & if \lambda \neq 0; \\ log(data) & if \lambda = 0; \end{cases} \tag{1.11}$$

   λ is varied from -5 to 5. The best value is chosen so that the transformed data approximates a normal distribution as closely as possible.

   Being that the exponent can be positive or negative this transform cannot handle distributions with negative values. For more information refer to [10]

Practical application of these procedures can be found in cap:2. These considerations conclude the theoretical background on regression, classification and penalization thereof.

1.2.2 Decision Trees and Random Forest

Apart from logistic and multinomial regression there are various other supervised classifiying methods such as Support Vector Machines (SVM), Neural Networks and Decision Trees. In this thesis, among the aforementioned algorithms, the chosen one was a particular evolution of DecisionTrees called RandomForest (RF).

To provide some insight in the method it's necessary to first explain how decision trees work, specifying what problems they face and what are their strong points.

Since it's a classification method let's consider the simple case of binary classification with categorical[23] features.

The task of the decision tree is to approximate to the best of it's abilities the labels contained in the training set using all the features associated with each datapoint, this is done by building a graph-like structure in which the nodes represent the features and the links departing from them are the possible values the feature takes. This graph is built in a top-down approach by choosing at every step the feature that best separates the data in the label categories, this process is done along each branch until a node in which the separation of the preceding feature is better then that provided by all remaining features or all features have been considered.

The first node is called root node, while the nodes that have no branches going out of them are called leaves, the graph represents a tree hence the name of the method. Note that at every node only the subset of data corresponding to all previous feature categories is used.

At each node the separation between the two label categories c_1, c_2 due to the feature is commonly measured with Gini impurity coefficient, which is computed as:

$$\begin{aligned} G &= 1 - p_1^2 - p_2^2 \\ &= 1 - \left[\frac{N_{c_1 \in node}}{N_{samples \in node}}\right]^2 - \left[\frac{N_{c_2 \in node}}{N_{samples \in node}}\right]^2 \end{aligned} \tag{1.12}$$

The Gini impurity for a feature is then computed as an average of the gini coefficients of all the deriving nodes weighted by the number of samples in each of the nodes.

This method can be obviously generalized to cases in which features are continuous by thresholding the features choosing the value that best improves the separation of the deriving node, a way to choose possible thresholds is to take all the means computed with all adjacent measurements. It's important to note firstly that there is no restriction on using, along different branches, different thresholds for the same feature and, secondly, that the same featue can end up at different depths along diffenrent branches.

The strength of this method is it's performance on data it has seen however it has very poor generalization abilities [17].

Random Forests algorithms are born to overcome this problem. As the name suggests the idea is to build an ensemble of Decision Trees in which the features used in the nodes are chosen among random subsamples of all the available features, the final result is obtained as a majority vote over all trained trees. Each tree is also trained on a bootstapped dataset created from the original, this procedure might exacerbate some problems of the starting dataset by changing the relative frequency of classes seen by each tree and can be corrected by balancing the bootstrap procedure.

This method vastly improves the performance and robustness of the final prediction while retaining the simplicity and ease of interpretation of the decision trees, naturally there are methods, such as AdaBoost, to deploy and precautions, such as having balanced dataset, to take to further improve the performance of RF classifiers.

[23] Categorical is to be intended as features with discrete value, opposed to continuous variables which can potentially take any value in $\mathbb{R}$

1.2.3 Synthetic Minority Oversampling TEchnique (SMOTE)

In the context of this thesis it will become necessary to take care of balancing the input dataset, to do so one could randomly oversample, by duplicating instances in the minority class, or undersample, by removing instances within the majority class. The choice that was made was to use Synthetic Minority Oversampling TEchnique (SMOTE)[9] to rebalance the dataset.

This technique considers a user-defined number of nearest neighbours of randomly chosen points in the minority class and populates the feature space by generating samples on the lines that connect the chosen sample with a random neighbour[24].

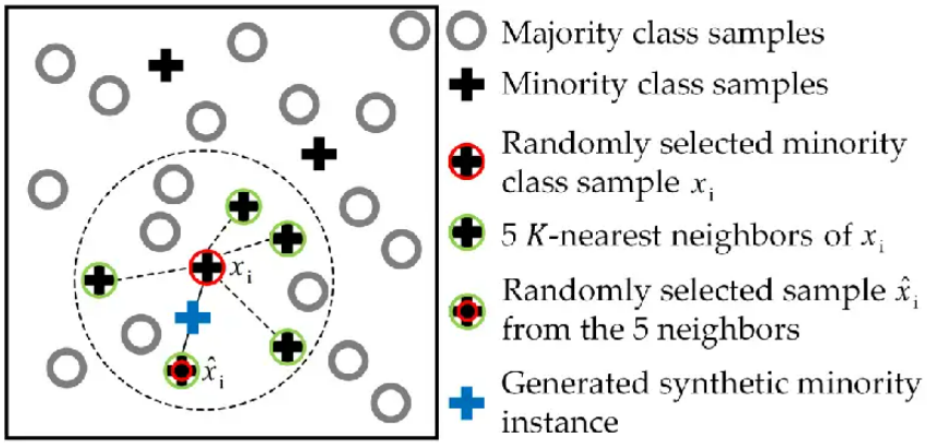
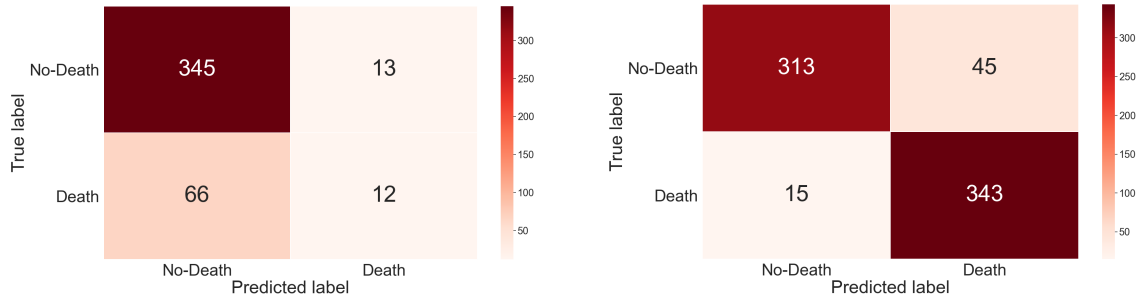


Figure 1.15: Example of SMOTE with 5 nearest neighbours

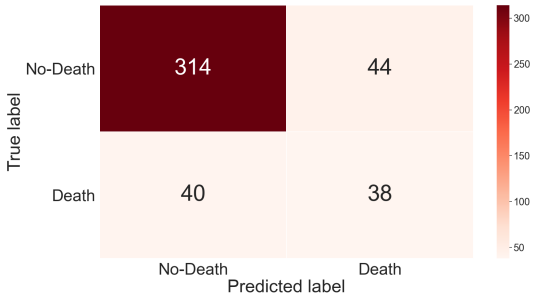
Worth noting that this has been done using the library imblearn in python [26]. To preview some of the data, looking at the performance of a vanilla random forest implementation with all preset parameters, the effect of the position of oversamplling is the following.

[24]This procedure is formally called convex combination, which is a peculiar linear combination of vector in which the coefficients sum to one. Particularly all convex combination of two points lay on the line that connects them, and for three point lay within the triangle that has them as vertices



(a) without smote

(b) SMOTE-before train-test split



(c) SMOTE after train-test split

Figure 1.16: Confusion matrix used to evaluate performance done using datasets with various combinations of SMOTE position

It's clear to see that in unbalanced case, like the one analysed in this thesis in which the label classes are 15%-85%, oversampling the data always determines an improvement in performance. However performing it in the wrong place it clearly makes a far too optimistic evaluation of the performance and also adds false datapoints even in the testing phase.

This highlights the point that all preprocessing should be done with care when train-test splitting the data, specifically it's very important that the oversampling, as well as all other data handling, be performed after the train-test split of the data on the train data alone.

It's clear to see that what's being observed is a leakage phenomenon, in which the testing data contains information regarding the training data and viceversa. In practice, especially because the points are created as convex combination of existing data[25], when the synthetically generated points end up in the testing they depend, at least partially, on the data used in the training procedure.

1.2.4 Dimensionality reduction and clustering

When dataset are composed of many features, namely more than three, it becomes difficult if not impossible to visualize the distribution of data.

[25]Note that this would happen with any other over/under-sampling methods, such as random over-sampling which randomly duplicates data points

There are various techniques that can mitigate or solve this problem, they are grouped under the umbrella term of "Dimensionality reduction techniques" and they are generally based on ML learning methods. As such , following the same reasoning and definitions given in regard to ML, it's possible to introduce a sub-categorization of the whole family of techniques in supervised and unsupervised techniques. Dimensionality reduction, beyond providing useful insight in the general structure of the data, can also be used as a preprocessing step in some analysis pipelines.

A first example could be to find more meaningful features before analyzing with methods such as Random Forests or Regression techniques, a second example could be using the reduced representation that keeps the most information possible to ease in clustering analysis by highlighting the differences in subgroups within the dataset.

The most common dimensionality reduction techniques are:

1. **Unsupervised methods**

   - Principal Component Analysis (PCA): Linearly combines the pre-existing features to obtain new ones and orders them by decreasing ability to explain total variance of data. The first principal component is the combination of features that most explains the variance in the original dataset. Given it's nature it focuses on the global characteristics of the dataset.
   - t-distributed Stochastic Neighbor Embedding (t-SNE): Keeps a mixture of local and global information by using a distance metric in the full high dimensional space and trying to reproduce the distances measured in the lower dimensional space which can be 2- or 3-dimensional.

     Let's define similarity between two points as the value taken by a normal distribution in a point away from the centre, corresponding to the first point of interest, the same distance as the two points taken in consideration.

     Fixing the first point the similarities with all other points can be computed and normalized. Doing this procedure for all points it's possible to build a normalized similarity matrix. Then the whole dataset is projected in the desired space, usually $\mathbb{R}^i$ with i=1,2 or 3, in which a second similarity matrix is built using a t-distribution instead of a normal distribution [26]. At this point, in iterative manner, the points are moved in small steps in directions such that the second matrix becomes more similar to the one computed in the full space.

2. **Supervised methods**

   - Partial Least Squared Discriminant Analysis (PLS-DA): This technique is the classification version of Partial Least Squares (PLS) regression. Much like PCA the idea is to find a set of orthonormal vectors as linear combination of the original features in the dataset. However in PLS and PLS-DA is necessary to add the constraint that the new component, besides being perpendicular to all previous ones, explains the most variability in a given target variable, or set thereof.

[26] The use of this t-distribution gives the name to the technique, the need for this choice is to avoid all the points bunching up in the middle of the projection space since t-distributions have lower peaks and are more spread out than normal distributions. For more details refer to [**?**]

For an in depth description refer to [5] while for a more modern review refer to [25]

3. **Mixed techniques**

   - Uniform Manifold Approximation and Projection (UMAP): Builds a network using a variable distance definition on the manifold on which the data is distributed then uses cross-entropy as a metric to reproduce a network with the same structure in the space with lower dimension.

     This technique maintains very local information on the data and allows complete freedom of choice in the final embedding space as well as the definition of distance metrics in the feature space[27], it's also implemented to work an a generic pandas dataframe in python so it can take in input a vast range of datatypes. The math behind this method is much beyond the scopes of this thesis, as such refer to [28] for more in depth information.

It should be noted that dimensionality reduction is not to be taken as a necessary step, however it can reduce the noise in the data by extrapolating the most informative features while easing in visualization and reducing computational costs of subsequent data analysis.

These upsides become particularly relevant in the field of clustering, where the objective is to group data in sets with similar features by minimizing the differences within each group while maximizing the differences between different groups. Clustering techniques can be roughly divided in:

1. Centroid based techniques: A user defined number of points is randomly located in the data space, datapoints are then assigned to groups according to their distance from the closest center. The main technique in this category is k-means clustering, and some, if not most, other techniques include it as step in the processing pipeline[28]. The main problems are firstly that these techniques require prior knowledge, or at the very least a good intuition, on the number of clusters in the dataset while also assuming that the clusters are distributed in spherical gaussian distribution, which is not always the case.

2. Hierarchical clustering: The idea is to find the hierarchical structure in the data using a bottom-up or a top-down approach, as such this category further subdivides in agglomerative and divisive methods. In the first each point starts by itself and then points are agglomerated using a similarity or distance metric, this build a dendrogram in which the k^{th} level roughly corresponds to a k-centrouid clustering. In the latter the idea is to start with a unique category and then divide it in subgroups.

3. Density based techniques: These methods use data density to define the groups by looking for regions with larger and lower density as clusters and separations.

In order to evaluate the performance of these methods it's necessary to define metrics that evaluate uniformity within clusters and separation among them, the choice in the

[27] Actually to keep the speed in performance the distance function needs to be Numba-jet compilable

[28] An example of such techniques is: affinity propagation

definition of metric should be taken in careful consideration since the different task may imply very different optimal metrics or, put differently, the optimal technique for the task at hand may very well depend on the metric used.

It's worth mentioning that when working in two, or at most three, dimensions humans are generally good at preforming clustering yet it's nearly impossible to do in more dimensions. Computers, on the other hand, require more careful planning even in low dimension because the generally good human intuition on the definition of cluster is not so easily translated in instruction to a machine but are much more performing in higher dimensions. So to obtain good results with clustering techniques it necessary to work with care, expecially with a good understanding of the dataset in use and it's overall structure.

In the context of this thesis, supposing the data suggested a clear cut distinction of two populations in the dataset, as could be male-females or under- vs normal- vs over-weight individuals, then it might become necessary to analyse these groups as different cohorts in order to more accurately predict their clinical outcome.

1.3 Combining radiological images with AI: Image segmentation and Radiomics

Having seen the kind of data that will be of interest throughout this thesis, and having a set of techniques used to describe and make predicition on the data at hand, the final step in this theoretical background chapter will be to combine these two notion in describing first how images can be treated in general terms and then, more specifically, how medical images can be analysed to exploit as much as possible the vast range of information they contain.

Image analysis seems, at first glance, very intuitive since for humans it's very easy to infer qualitative information from images. However upon closer inspection this matter becomes clearly non trivial due to the subjectivity involved in the process as well as in the intuition behind it. More specifically, in the context of this thesis, the same image of damaged lungs contains very different informations to the eyes of trained professional versus those of an ordinary person as well as to the eyes of different professionals.

The first big obstacle in this task is the definition of region of interest: not all people will see the same boundary in a damaged organ, sometimes the process of defining a boundary between organ and tissue may need to account for the final objective it has to achieve. If the objective is to evaluate texture of a damaged lung then the lesion needs to be included whereas in other cases these regions may only be unwanted noise. Generally speaking finding regions of interest in an image is a process called image segmmentation.

The next step would be to quantify the characteristics of the region identified, as such it should be clear that finding ways to derive objective information from images it's of paramount importance, expecially when this information can aid in describing the health of a patient. It should also be clear that medical images are a kind of high dimensional data as such, as it's fashion with fields that occupy themselves with big biological data, the field that studies driving quantitative information out of radiological images is called radiomics.

1.3.1 Image Segmentation

Generally speaking image segmentation is a procedure in which an image is divided in smaller sets of pixels, such that all pixel inside a certain set have some common property and such that there are no overlaps between sets. These sets can then be used for further analysis which could mean foreground-backgroud distinction, edge detection as well as object detection, computer vision and pattern recognition.

Image segmentation can be classified as:

- Manual segmentation: The regions of interest are manually defined usually by a trained individual. The main advantage of this it's the versatility, on the other hand this process can be very time consuming

- Semi-automatic segmentation: A machine defines as best as it can the shape of the region of interest, however the process is then thought to receive intervention of an expert to correct the eventual mistakes or refine the necessary details. This provides the best compromise between time needed and accuracy obtained and becomes of interest in fields in which finer details are important.

- Automatic segmentation: A machine performs the whole segmentation without requiring human intervention

On a practical level these techniques can be used in various fields, as illustrated by the variety of aforementioned tasks, but the one that interests this work the most is the medical field.

Nowadays in medicine, where most of imaging exams are stored in digital form, the ability to automatically discern specific structures within the images can provide a way to aid clinical professionals in their everyday decision making their workload lighter and helping in otherwise difficult cases.

A staggering example connected to this thesis is the process of organ segmentation in CT scans: usually these scans are n[29] stacked images in 512x512 resolution. To have a contour of the lungs in a chest CT a radiologist would need to draw by hand the contour of the lung in each slice of the scan. Even if some shorcuts exist to reduce the number of slices to draw on this very boring, time consuming and repetitive task can occupy hours if not days of work to a human while a machine can take minutes to complete a whole scan. Then considering lesion detection in medical exams having a machine that consistently finds lesions that would otherwise be difficult to discern by a human eye can be of paramount importance in diagnosis as well as treatment.

In the medical field the difficulty comes from the fact that different exams have different types of image formats which means that an algorithm that works well on CT may not work as intended on MRI or other procedures. Automatic image segmentation can be performed in various ways:

1. Artificial Neural Network (ANN): These techniques belong to a sub-field in ML called Deep Learning, they involve building network structures with more layers in which each node is a processing unit that takes a combination of the input data and

[29] n clearly depends on the exam required and slice thickness, some common values for thoracic CTs are around 200-300 but can range up to 900 slices

gives an output according to a certain activation function. These structures are called Neural Networks because they resemble and are modeled after the workings of neuron-dendrite structures while the deep in deep learning refers to the fact that various layers composed of various neurons are stacked one after the other to complete the structure. Learning is obtained by changing how each neuron combines the inputs it recieves. In the case of images these structures are called Convolutional Neural Networks because the first layers, which are intended to extract the latent features or structures in the images, perform convolution operation in which the parameters are the pixel values of convolutional kernels

2. Thresholding: These techinques involve using the histogram of the image to identify two or more groups of pixel values that correspond to specific parts/objects within the image. An obvious case would be a bi-modal distribution in which the two sets can be clearly identified but there are no requirements on the histogram shape. In the case of CT this has a very simple and clear interpretation since HU depend on tissue type it's reasonable to expect that some tissues can be differentiated with good approximation by pixel value alone

3. Deformable models and Region Growing: Both these technique involve setting a starting seed within the image, in the first case the seed is a closed surface which is deformed by forces bound to the region of interest, such as the desired edge.

   In the second case the seed is a single point within the region of interest, step by step more points are added to a set which started as the seed alone according to a similarity rule or a predefined criteria.

4. Atlas-guided: By collecting and summarizing the properties of the object that needs to be segmented it's possible to compare the image at hand with these properties to identify the object within the image itself.

5. Classifiers: These are supervised methods that focus on classifying by focusing on a feature space of the image. A feature space can be obtained by applying a function to the image, an example of feature space could be the histogram. The main distinction from other methods is the supervised approach

6. Clustering: Having a starting set of random clusters the procedure computes the centroids of these clusters, assigns each point to the closest cluster and recomputes centroids. This is done iteratively until either the point distribution or the centroid position doesn't change significantly between iterations.

For a more in depth review of the main methods used in medical image segmentation refer to [35]. Worth noting, at this point, that the semi-automatic segmentation software used in this work uses probably a mixture of region growing and thresholding methods, maybe guided by an atlas. The segmentation process generally produces a boolean mask which can be used to select, using pixel-wise multiplication, the region of interest in the image. The next step in image analysis would be to derive information from the region defined during segmentation, in general this step is called feature extraction[30] and it's objective is to find non-redundant quantities that meaningfully summarize as much properties of the original data as possible.

[30]Even if the term is used in pattern recognition as well as machine learning in general

1.3.2 Radiomics

When the images are medical in nature and when referring to high-throughput quantitative analysis the task of finding these features fall in the realm of radiomics, which uses mathematical tools to describe properties of the images that would otherwise be unquantifiable to the human eye.

The features that can be computed from images are of various types, some of them can be understood somewhat easily through intuition since they are close to what humans generally use to describe images, others are much more complex in definition and quantify more difficulty percieved properties of the image.

All of the features used in this work have been rigorously defined and described in [46] which, being born as a reference manual, covers the founding concepts of radiomics in an attempt to standardize the procedures of the field.

Generally speaking features can be then roughly classified in different families:

1. Morphological features: These features describe only the shape of the region of interest, as such they are independent of the pixel values inside the region and hence, to be computed, require only a boolean mask of the segmented region. These features can be further subcategorized as two or three dimensional features based on whether they focus on single slices or whole volumes. Most of these features compute volumes, lengths, surfaces and shape properties such as sphericity, compactness, flatness and so on.

2. First order features: These features depend strictly on the gray levels within the region of interest since they evaluate the distribution of these values, as such they need that the boolean mask of the segmentation be multiplied pixel-wise with the original image to obtain a new image with only the interesting part in it. Most of these features are commonly used quantities, such as Energy, Entropy, Minimum and Maximum value which have been adapted to the imaging context using the histogram of the original image or by considering intensities within an enclosed region.

3. Higher order features: All the other fatures fall in this macro-category which can be clearly subdivided in other smaller categories in which the features are obtained following the same guiding principle or starting point.

   Generally these describe more texture-like properties of the image and, to do so, use particular matrices derived from the original image which contain specific information regarding order and relationships in pixel value positioning within the image.

   These matrices have very precise definition, as such only the general idea behind them will be reported here redirecting to [46] for a more strict and in detail description. The matrices from which the features are computed also give name to the smaller categories in this family, these categories are:

   - Gray Level Co-occurence Matrix (GLCM) features: This matrix expresses how combination of pixel values are distributed in a 2D or 3D region by considering connected all neighbouring pixel in a certain direction with respects to the one in consideration.

Using all possible directions for a set distance, usually δ=1 or δ=2, various matrices are obtained and from these a probability distribution can be built and evaluated. It should be noted that before computing these matrices the intensities in the image are discretized.

- Gray Level Run Length Matrix (GLRLM) features: Much like before the task of these features is to quantify the distribution of relative values in gray levels throughout the image, as the name suggests what this matrix quantifies is how long a path can be built by connecting pixel of the same value along a single direction. This time information from the matrices computed by considering different directions are aggregated in different ways to improve rotational invariance of the final features.
- Gray Level Size Zone Matrix (GLSZM) features: This matrix counts the number of zones in which voxel have the same discretized gray level. The zones are defined by a notion of connectedness most commonly first neighbouring voxel are considered as connectable if they have the same value, this leads to a 26 neighbouring voxel in 3 dimensions and to 8 connected pixels in 2 dimensions[31]. The matrix contains in position (i,j) the number of zones of size j in which pixel have value i.
- Neighbouring Gray Tone Difference Matrix (NGTDM) features: Born as an alternative to GLCM these features rely on a matrix that contains the sum of differences between all pixel with a given pixel value and the average of the gray levels in a neighbourhood around them.
- Gray Level Dependence Matrix (GLDM) features: The aim of these features is to capture in a rotationally invariant way the texture and coarsness of the image. This matrix requires the already seen concept of connectedness with a given distance as well as dependence among pixel.

  Two voxel in a neighbourhood are dependent if the absolute value of the difference between their discretized value is less then a certain threshold. The number of dependent voxel is then counted with a particular approach to guarantee that the value be at least one

In talking about the previous feature groups the concept of discretization of data which is already digital, and hence a discretized, has emerged. This is often a required step to make the computations of the matrices tractable and consists in further binning together the pixel values, which is commonly done in two main ways: either the number of bins is fixed or the width of the bins is fixed preceeding the discretization process. It should be evident that the results of the feature extraction procedure is heavily dependent on the choices made in all the steps that preceed it, main of which being the segmentation, the eventual re-discretization of the image leading to the algorithm used to compute the feature themselves.

For this reason recently the International Biomarker Standardization Initiative (IBSI [46]) wrote a *"reference manual"* which details in depth the definitions of the features, description of data-processing procedures as well as a set of guidelines for reporting

[31] To visualize, imagine a 2D grid: the 8 pixel are the four at the sides of each square and the four at the corners.

results. This was done in an attempt to reduce as much as possible the variability and lack of reproducibility of radiomic studies.

In the past the main attention in radiological research was focused on improving machine performances and evaluating acquisition sequence technologies, however the great developements in artificial intelligence and performance of computers have brought a lot of attention to the field. Various papers, such as [24] and [4] have been written with the objective of presenting the general workflow of radiomics.

By design the topics in this thesis have been presented to resemble the general order of the pipeline, which can be summarised as:

- Data acquisition
- Definition of the Region Of Interest (ROI)
- Pre-processing
- Feature extraction
- Feature selection
- Classification

Another interesting possibility offered by the biomarkers computed following radiomics is the ability to quatify the differences between successive exams of the same patient. This specific branch of radiomics is called Delta-radiomics, referencing to the time differential that becomes the main focus of the analysis.

1.4 Survival Analysis

Survival analysis is a particular field that tries to determine the probability of a certain event happening before a certain time. As the name suggests one of it's main application is determining the risk of death due to a disease as time progresses from diagnosis, however it can be used to estimate time needed to recover after a surgical procedure, lifetime before breakdown of machines, time needed for criminals to commit new crimes after being released ....

The main concepts and terminologies in this field are:

1. Event: This is the phenomenon under analysis, generally it could be death, remission from recovery, recovery and so on. It is common to refer to the event as failure since usually the event is negative in nature.

2. Censoring: When collecting data to develop a model that describes survival it may happen that the individual drops out of the study without incurring in the event under analysis. For example when looking at effectiveness of a drug the patient develops an adverse reaction to the drug and needs to stop using it, hence falling out of the study.

   In the context of this thesis all individuals that got sent home from the hospital are censored since their survival is known only up until the dropout and not after. Cases like this when the start of the followup is well known but dropout happens

are called right-censored, because only the right side of the timeline is abruply interrupted. Cases can also be left-censored, e.g. a patients with unknown time of contraction of a disease, and interval-censored, e.g. when the contraction of the disease can be restricted to an interval as it may happen when a negative and positive test happen at successive times.

Usually this variable is called **d** and is a binary variable where 1 indicates that the event occurred and 0 indicates all possible censoring causes.

3. Time: This is to be intended as time, be it days, weeks, months or years, since the start of the followup to either the event or the censoring. It usually is indicated with **T**, is referred to as survival time and, being a random variable that indicates time, it cannot be negative.

4. Survivor function **S(t)**: This is to be intended as the probability that the subject in the study survives a time t before incurring in the event. In theory this function is smooth from zero to infinity, it's strictly non-increasing, it starts at S(0)=1 and ends up at S(∞)=0.

   These assumptions are all reasonable since at no point the survival probability can increase, since everybody is alive at the start of observing them and since nobody can live to infinity. However, when it comes to practice, these properties are not necessarily verified. Since no study can continue to infinity the last value need not be zero and since the timesteps at which it's possible to perform a checkup are discrete the curve is actually a step function.

5. Hazard Function **h(t)**: This function is difficult to explain practically, citing [23] **"The hazard function h(t) gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t"**. The mathematical definition is:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{1.13}$$

   Dividing by a time interval the hazard function can also be intended as conditional failure rate, where the conditional refers to the "given that the individual has survived up to time t".

   Being a rate this quantity need not be bound in [0,1] but ranges in [0, ∞] and depends on the time unit used. This will be interpreted as h(t) events per unit time.

   The hazard function is non-negative and has no upper bounds, since it can be related to the survival function [32] the possible shapes give different names to the final model. These could be increasing or decreasing Weibull, exponential or lognormal survival models

[32]The hazard function can be computed as time derivative of the survival function divided by the survival function changed of sign. The survival function is the exponential of minus the integral in [0,t] of the hazard function. On this note, an exponential model is given by a constant hazard function. loosely speaking the Weibulls and LogNomal respectively come from increasing, decreasing and somewhat bell-shaped hazard functions.

A further step in the analysis could be to try to measure the differences in survival between two groups, i.e. using a single variable expected to drive the differences in survival.

The most basic example would be looking at the effectiveness of a drug by dividing in group A and B the patients given the actual medicine and the placebo respectively. However this could be done even for males vs females or, in the case of continuous variables, for patients with above or below threshold values [33]. When the analysis is univariate in nature then the Kaplan-Meier survival curves are used in conjuction with the log-rank test, when the analysis is multivariate then the Cox Proportional-Hazard model is used. The Cox model is very similar to linear and logistic regression.

1.4.1 Kaplan-Meier(KM) curves and log-rank test

Kaplan-Meier curves can be built following a well defined procedure:

The data is separated in sets using the label of the groups that need to be used then the patients in each set are ordered in ascending order of permanence in the study, which is the time variable.

At each failure time T_f the conditional probability of surviving past T_f given availability is computed as ratio of subjects left in the study right after T_f divided by the number of availabe people at that same time [34] Note that this takes in account also the possible censoring by reducing the number of available people at times of event, these censoring event will be drawn on the curve using ticks at the corresponding time. The survival probability is computed at each event time as the product of the probability at the previous failure time with the one computed as explained before at the time of the current event.

Let's consider an imaginary study with 4 patients with one failing at week one, one dropping out at week 2, one failing at week 3 and one surviving after 4.

The KM curve will start at 1, at week one it will drop at $\frac{3}{4}$ since 3 people will be alive out of 4 available, at week two no drop will happen but a tick will be put on the curve and, finally at week 3 it will drop at value $\frac{3}{4} * \frac{1}{2}$ since only one out of the two available will survive.

As a rule of thumb, two Kaplan-Meier curves that do not intersect at any point indicate good separation among the groups, this can then be formally evaluated performing a log-rank test on the data used to build the curves.

The null hypothesis of this specific test is that "...there is no overall difference between the two survival curves" which can be tested with the log-rank test. This basically consist in performing a large-sample χ^2 test that uses the ordered failure times for the entire dataset as expected values vs those observed in the subsets. From this testing procedure a p-value can be obtained to reject the null hypothesis, for more in depth information refer to [23].

[33] Generally, when no obvious threshold is available, the median of the variable is chosen.

[34] Effectively this corresponds to dividing thenumber of available minus dead individuals by the number available at each timestep

1.4.2 Cox Proportional-Hazard (CoxPH) model

As it was mentioned before the shape of the hazard curve, an hence of the survival curve, implies a specific shape for the model used to describe it. Some of the possible models are increasing Weibull, decreasing Weibull, Exponential and log-normal. All of these models are parametric because known the parameters the distribution of the outcome can be known.

When it comes to Cox PH model the distribution cannot be known because part of the model, namely the baseline hazard, remains unestimated.

Generally the data has an optimal parametric model that describes it and, if this information were known, it would make perfect sense to use said function. However this knowledge is not always obtainable, which give Cox PH model an occasion to shine. Cox Proprotional Hazard models can be described as robust in the sense that the results that it gives will approximate those obtained by the correct model, without needing to know which of the model needs to be used. For a more in depth description refer to [23]

Cox PH models rely on the use of the formula in Equation 1.14 to express the risk of a patient with a set of k characteristic variables $\mathbf{X}$ at time t.

$$h(t, \mathbf{X}) = h_0(t) e^{\sum_{i=1}^{k} \beta_i X_i} \tag{1.14}$$

The quantity $h_0(t)$ is called baseline hazard and it hides one of the main hypotheses behind this method which is the *Proportional Hazard* assumption. This assumption is that the baseline hazard $h_0(t)$ depends on time alone and not on all the other variables $\mathbf{X}$, note also that the exponent is time-independent since the $\mathbf{X}$s are supposed to be constant in time [35].

The β in the exponent represents the weights assigned to each variable and, very much like what happened in linear regression, these coefficients can be a proxy of importance in the model: coefficients close to zero signify no particular importance of the variable whereas the more the value is distant from zero the more important the variable can be considered. The reason why it's common practice to report the exponentiated cefficient for features is that 1-exp[β] represents the difference in likelyhood to die of individuals separated using the feature relative to β.

For example, considering a binary feature AGE OVER 50 with exp[β]=1.6, the exponential of the parameter would indicate that people over 50 are 60% more likely to die when adjusting for all other variables. It is also common procedure to provide, for each coefficient, the 95% confidence interval of the parameter and a p-value relative to the hypothesis that the parameter be equal to zero.

Cox models provide a way to predict hazard for patients, this predict method can be used to evaluate the model built by identifying groups in the patient cohort using hazard quantiles The division in the resulting groups can be used as proxy for the quality of the score built using the Cox model.

[35] Dropping the time independence of the $\mathbf{X}$s is possible while still keeping the same shape, yet the model would then be called extended Cox model.

Chapter 2

Materials and methodologies

In this section there's going to be an explanation of the dataset as well as instruments and methodologies used to analyze it's properties, as such the first step is going to be an in depth discussion of the data available and a general overview of the final use. The following step is going to be a description of the preliminary work done to the data itself and to the results of this preliminary analysis in order to select the important features. The final step of this chapter is going to be an explanation of the methods used to derive the final results and to evaluate them.

Worth noting that the data has been collected and used with the approvation of the ethical commettee.

2.1 Data and objective

The objective of this thesis will be to compare how different methods perform in predicting clinical outcomes in covid patients, while also determining if different kind of input data imply different performances of the same methods. All images are non segmented, as such all of them have been semi-automatically segmented via a new software being tested in the medical physics department called *Sophia Radiomics* [1] which seems to be built around region growth algorithm mixed with thresholding.

Statistical analyses have been performed in python using libraries such as scikit-learn [34] and imblearn [1] [26], pandas [29], numpy [16], scipy [42], statsmodels[39].

Lifelines [11] has been used for survival analysis and combined to scikit-learn by coding wrappers that made compatible the functions from lifelines with the API of scikit-learn.

Finally all of the management of the graphs obtained as part of the analysis has been performed with either seaborn[43] or matplotlib[19].

The starting dataset was a list of all the patients that, from 02/2020 to 05/2021, were hospitalized as COVID-19 positive inside the facilities of *IRCSS Azienda ospedaliero-universitaria di Bologna - Policlinico Sant'Orsola-Malpighi* .

As far as exclusion criteria go the main deciding factors, except unavailability of the feature related to the patient, were visibly damaged and lower quality images, for example images with cropped lungs. The first set of selection criteria were:

[1] This library is born to handle cases of imbalanced learning and offers functions and objects compatible with the API offered by scikit-learn.

- All patients that had undergone a CT exam which was retrievable via the PACS (Picture Archiving and Communication System) of *IRCSS Azienda ospedaliero-universitaria di Bologna - Policlinico Sant'Orsola-Malpighi*
- All patients that had a all of the clinical and laboratory features, listed in fig:2.1, which have been found informative of the outcome during the clinical practice.
- Since all patient had at least 2 CT exams only the closest date to the hospital admission date was taken. When more exams were performed on the same date all of them were initially taken. At first only chest or abdomen CTs were taken regardless of the acquisition protocol used.

In tables 2.1,2.2 and 2.3 are reported all of the variables available for all patients with some information on their distribution. These have been divided in: Clinical features used for the models Table 2.1, variables determined by radiologist by examining the CT scan of the patient, called Radiological fetures Table 2.2. Finally in Table 2.3 are contained all the variables that represent the outcome of the illness and not a property of the patient at admission which is why these won't be used in building the predictive models.

Table 2.1: Clinical variables used in the analysis. These are all very self-explanatory

Variable name	count	mean	std	min	median	max
Age (years)	436	67.45	15.08	21	68.50	99
Respiratory Rate	436	21.24	6.80	10	20	98

Variable name	total	unique	top	top count
Hypertension	436	2	1	241
History of smoking	436	2	0	347
Obesity	436	2	0	363
Sex	436	2	Male	286
Fever	436	2	1	251

Table 2.2: Radiological features, boolean expression of the findings of radiologists upon close examination of the CT scan of the patient. 1 indicated that the damage was found, 0 otherwise

Variable name	count	unique	top	freq
Lung consolidation	436	2	1	225
Ground-glass	436	2	1	382
Crazy Paving	436	2	0	336
Bilateral Involvement	436	2	1	403

Most clinical features are pretty self-explanatory, a brief explanation will be provided for those that could appear obscure to an outsider, and that were not explained in 1.

Table 2.3: Clinical variables that indicate the treatment used for the patient, these cannot be used in building the model because they mostly represent outcomes and not "a priori" knowledge available on the patient

Variable name	count	unique	top	top count
DNR	436	2	0	413
ICU Admission	436	2	0	359
Sub-intesive care unit admission	436	2	0	336
Death	436	2	0	358
O2-therapy	436	2	1	370
cPAP	436	2	0	367
Bilateral Involvement	436	2	1	403
Respiratory Failure	436	2	0	231
NIV	436	2	0	371

1. DNR : Acronym for "Do Not Resuscitate", used to indicate the wish of the patient or their relatives that cardiac massage not be performed in case of cardiac arrest.

2. NIV: Acronym for "Non Invasive Ventilation", it's a form of respiratory aid provided to patients.

3. cPAP: Acronym for "continuous Positive Airway Pressure", another form of respiratory aid.

4. ICU: Acronym for "Intensive Care Unit". When patients are in really severe conditions they are treated in these facilities.

5. Clinical Scores: When available values from laboratory analyses and/or patient conditions are summarised in scores that represent the gravity of the state of the patient, as such these can be somewhat correlated and could be treated as comprehensive values to substitute an otherwise large set of obscure clinical features. At admission, or closely thereafter, a set of clinical questions regarding the patient receives a yes or no answer, each answer has an additive contribution towards the final value of the score.

   These scores differ in how much they add for each condition and the set of symptoms the check for.

   (a) MulBSTA: This score acconts for **Mul**tilobe lung involvement, absolute **L**ymphocyte count, **B**acterial coinfection, history of **S**moking, history of hyper**T**ension and **A**ge over 60 yrs. [15]

   (b) MEWS: Modified Early Warning Score for clinical deterioration. Computed considering systolic blood pressure, heart rate, respiratory rate, temperature and AVPU(Alert Voice Pain Unresponsive) score. [40]

   (c) CURB65: **C**onfusion, blood **U**rea Nitrogen or Urea level, **R**espiratory Rate, **B**lood pressure, age over **65** years. This score is specific for pneumonia severity [45]

(d) SOFA: **S**equential **O**rgan **F**ailure **A**ssessment score. Considers various quantities from all systems to assess the overall state of the patient, PaO_2/FiO_2[2] for respiratory system, Glasgow Coma scale[3] for nervous, mean pressure for cardiovascular, Bilirubin levels for liver, platelets for coagulation and creatine for kidneys [3]

(e) qSOFA: **q**uick SOFA. Only considers pressure, high respiratory rate and the low values in the Glasgow scale.

Table 2.4: All the available clinical scores, generally computed adding 1 or 0 by looking if values of laboratory exams are above or below a certain threshold

Variable name	count	unique	top	top count
qSOFA	436	4	0	225
SOFA score	436	9	2	143
CURB65	436	5	1	143
MEWS score	436	8	1	143
MulBSTA score total	436	17	9	110

Finally it's also useful to see a distribution of the days of permanence in the Hospital, abbreviated dOS for **d**ays **O**f **H**ospitalzation, which can be seen in Figure 2.1.

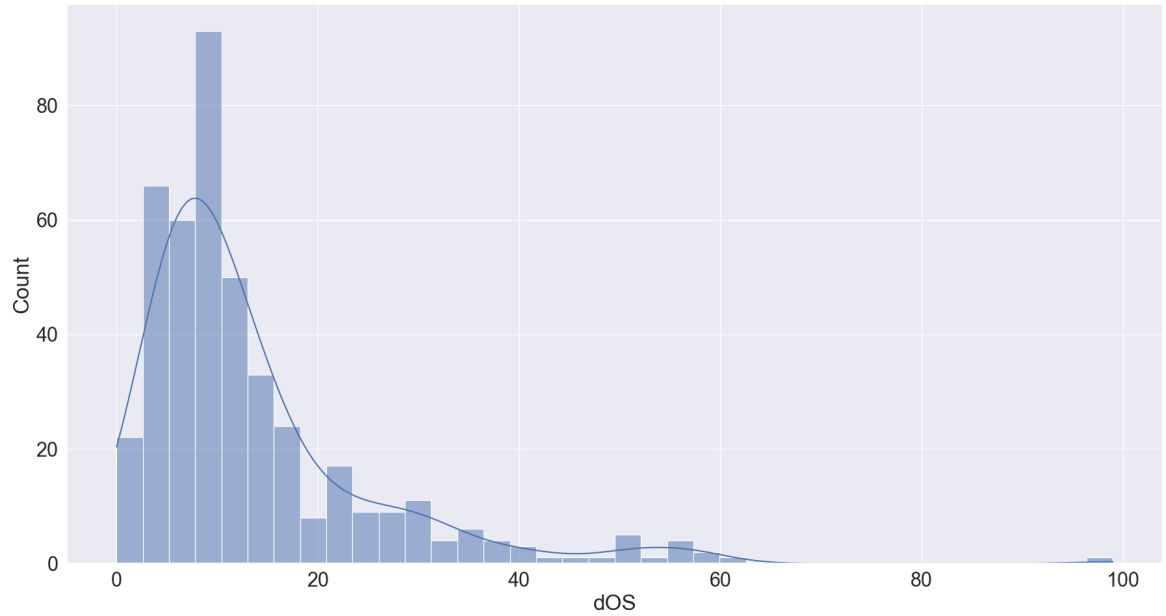


Figure 2.1: Distribution of the days of hospitalization for the patients included in the study.

[2]Very unrefined yet widely used indicator for lung disfunction

[3]GCS for short, proposed in 1974 by Graham Teasdale and Bryan Jennet. Evaluates what kind of stymulus is necessary to obtain motor and verbal reactions in the patient as well as what's necessary for the patient to open their eyes

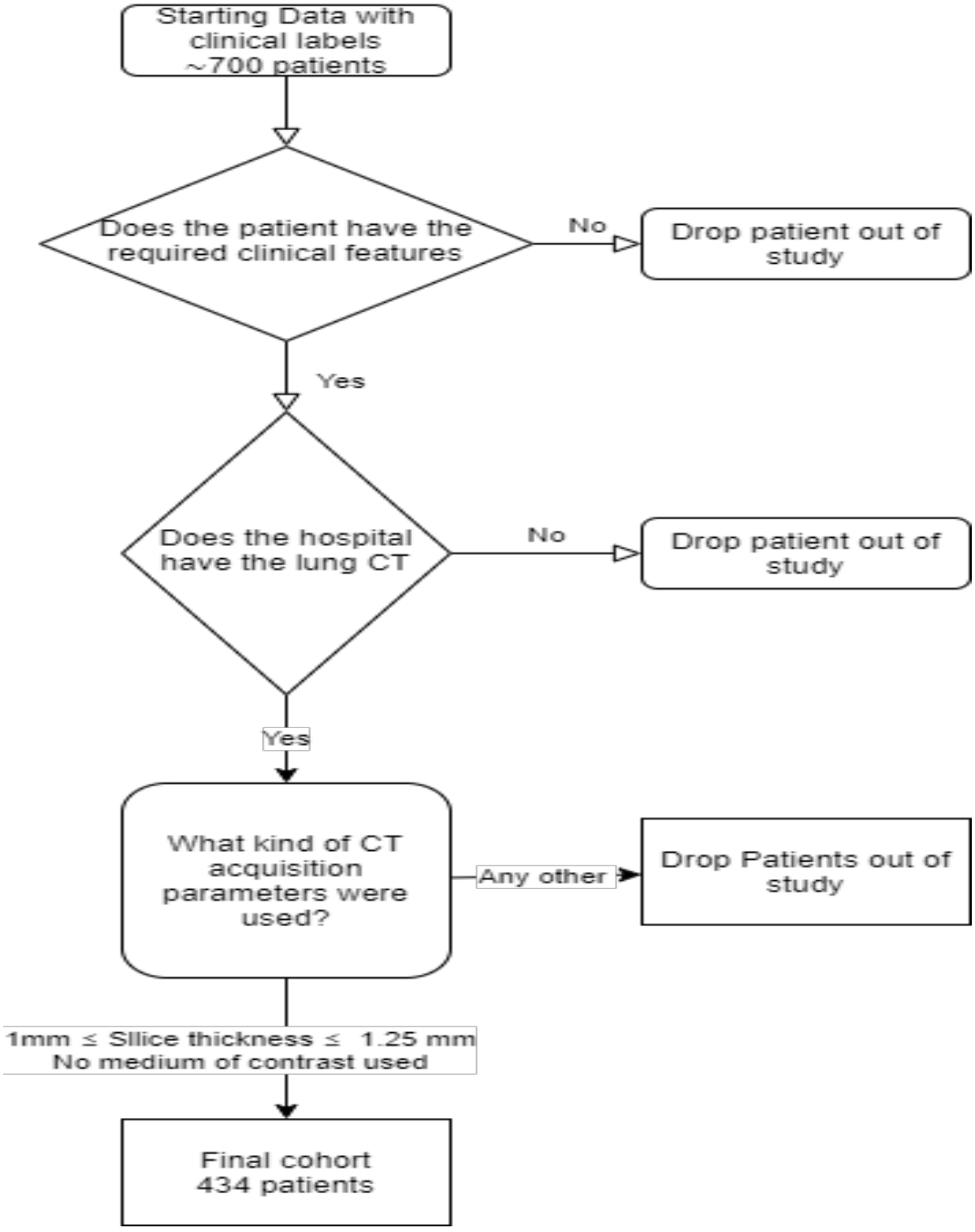


Figure 2.2: Flowchart of the patient selection procedure

This procedure, summarized and represented in Figure 2.2, produced a starting cohort of ∼700 patients which, having all various images available, created a huge set of ∼2200 CT scans. Since this analysis is focused on radiomics there is an evident need for as much consistency as possible in the images analysed. For this reason all CTs taken with medium of contrast were excluded, since they would have brightnesses not indicative of the disease, and for every patient only images with thin slice reconstruction were considered.

More specifically only images with slice thickness of 1 o 1.25 mm [4] along the z-axis were taken into consideration, which meant excluding all the 1.5,2,2.5 and 5 mm slice thicknesses.

All images were segmented using SOPHiA DDM for radiomics [1] which is a tool provided by *IRCSS Azienda ospedaliero-universitaria di Bologna - Policlinico Sant'Orsola-Malpighi* . This tool was chosen as one of the most IBSI-compliant available softwares, obtained as result in [6].

Overall this left the final study cohort to be composed of 436 patients, all descriptions and analyses are related to this cohort.

The same software used for segmentation allowed the extraction of the radiomic features form the segmented volumes, even if it did not allow the extraction of the segmentation masks nor any changes in the segmentation parameters. For this reason all the image analysis in this thesis is reliant on said software which has been treated as a black-box.

So, having segmented all the images and extracted all the features supported in the software, the next step is the definition of the actual analysis pipeline. The whole dataset was comprised of ∼200 features, their distribution has been presented in Tables2.1,2.2 and 2.3.

From now on all of the features used for model construction will be considered divided in three subgroups as follows:

1. Clinical: All these features are derived from the admission procedure in the hospital.

    - The continuous are AGE TAKEN AT THE DATE OF THE CT EXAM and RESPIRATORY RATE defined as number of breaths in a minute.
    - The discrete one were the aforementioned scores and the boolean ones were SEX OF THE PATIENT, OBESITY STATUS, IF THE PATIENT HAD A FEVER[5] as of hospital admission and whether or not the patient suffered of HYPERTENSION

[4]This meant that only exams called 'Parenchima' or 'HRCT' were included. Throughout the internship 'parenchima' has always appeared in contrast with 'mediastino'. These two keywords are used in the phase of reconstruction of the raw data to identify reconstructions with specific properties. Parenchima is used for finer reconstruction of lung specifically, the requiring professional uses these images to look for small nodules with very high contrast and, to do so, the reconstruction allows some noise to achieve the best resolution possible. Mediastino is used in the lung, as well as other regions, to look for bigger lesions but with low contrast. As such the 'mediastino' reconstruction compromises a worse spatial resolution for a better display of contrast, visually speaking the first images are more coarse and noisy while the second are smoother. It should be noted that even with the same identifier, be it HRCT parenchima or others, the machines on which the exams were made were different and had different proprietary convolutional kernels used for reconstruction.

[5]Defined as body temperature>38°

- The remaining features, namely those in 2.1 as well as the DEATH status of the patient, were either used as labels or not used at all because they refer to treatments used and not characteristics of the patient. As such these features, while plausibly correlated to the clinical outcome, are not really descriptive of the patient as of admission and are not information that can be used to aid professionals at admission to assess the situation

2. Radiomic: These features were all the ones supported by the segmentation software and are pretty much most of those described in [46] with the addition of fat and muscle surface, computed as cm^2 by counting pixel identified via threshold as fat or muscle tissue in thoracic slices taken at height of vertebra T-12

3. Radiological: These features are those that can be derived from CT exams by humans. Namely acquisition parameters, such as KVP and CURRENT, were used to search for eventual correlations between image quality and predictive power of the feature derived from the image while boolean features, such as BILATERALITY of lung damage, presence of GROUND GLASS OPACITIES (GGO), LUNG CONSOLIDATIONS as well as CRAZY PAVING were used to see if they were sufficient in determining outcome.

2.2 Preprocessing and data analysis

Before any preprocessing a choice was made to exclude all of the clinical scores, this was done to avoid having them mask other, more straightforward variables.

The first step in the analysis of this data is going to be a lasso regularized regression using either DEATH or ICU ADMISSION as target. As mentioned before when operationg with regressions it's a necessity that the residuals be normally distributed, a common way to get as close as possible to this hypothesis is to boxcox transform the data. Apart from the data which contained negative values, which cannot be fed into the boxcox transform, all variables have been transformed using this method.

The quatile-quantile plots have been used for a visual check of normality, normally distributed data will populate the bisector of the graph while deviations are symptoms of non-normality. Heavy and light tails are visible as deviations respectively above and belox the bisector.

It should be noted that when the data is normally distributed then the transform doesn't change much the distribution while, in cases with more heavy tailed distributions, the improvement is clear to see as it can be seen in Figures 2.4 and 2.5.

The next preprocessing step has been to apply a *StandardScaler* to all of the features, which corresponds to subtracting the mean and dividing by the standard deviation, in order to center all the features around zero.

The final step in preprocessing has been to reduce the features by using a correlation threshold which means that all variables that correlate with another more, in absolute value, than a certain threshold, which has been set to 0.6 in this work, are dropped a priori. A rather important thing to notice is that correlation has been computed using Spearman correlation and not Pearson since the first is invariant under monotone transformations, such as boxcox, and the second is not.

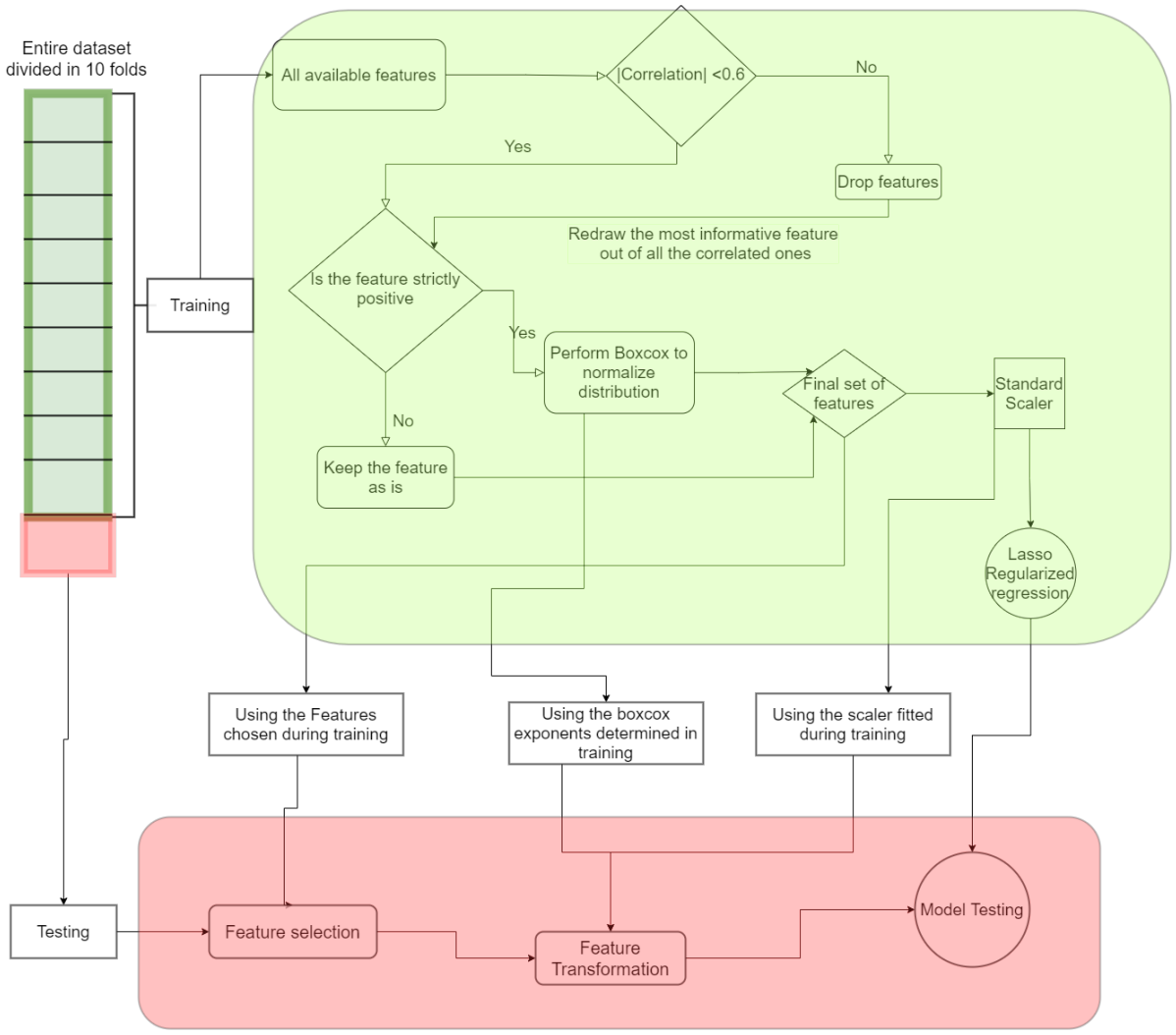


Figure 2.3: Flowchart for the preprocessing steps before preceeding a Lasso regularized regression.

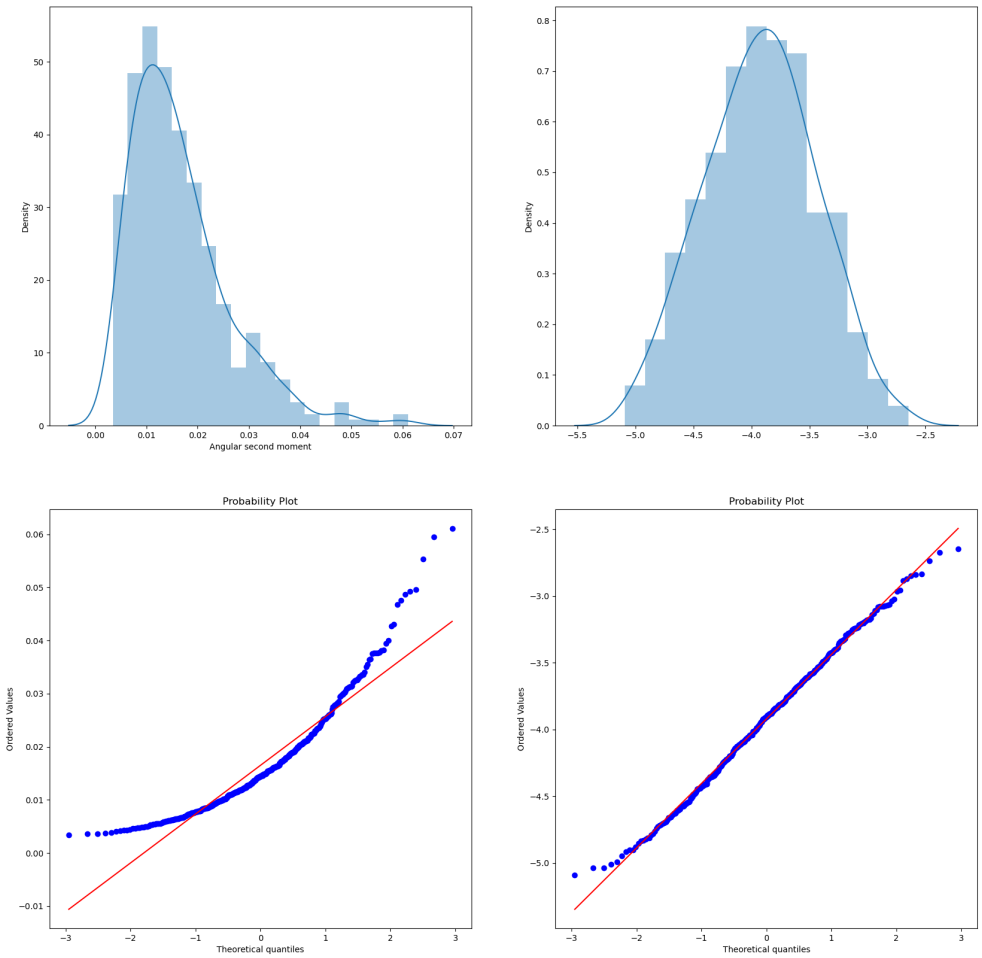


Figure 2.4: Example of boxcox applied to the radiomic feature *Angular second moment* which has a heavy tailed distribution. The graphs contain the original distribution (top-left) the transformed distribution (top-right) and the two respective quantile-quantile plots(bottom)

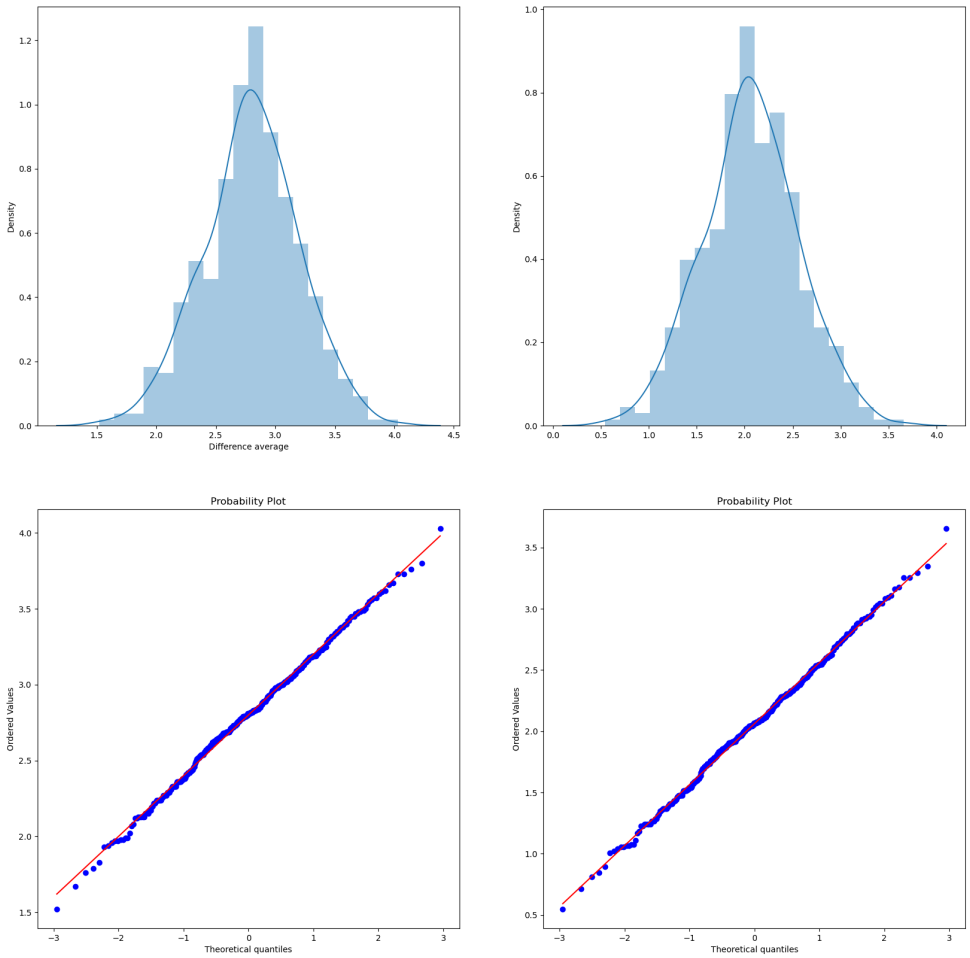


Figure 2.5: Example of boxcox applied to the radiomic feature *Difference Average* which has a close to normal distribution. The graphs contain the original distribution (top-left) the transformed distribution (top-right) and the two respective quantile-quantile plots(bottom)

Given the large number of features it's very plausible that at least one of the eliminated features is correlated with all other dropped features but with none of the remaining ones, since a Lasso regularization will be used introducing a few redundant features is not too damaging and the possible benefits outweight the risks. For this reason a redrawing method has been implemented to add one of the dropped features, this has been done by choosing the one that most correlates with the label being used.

A pivotal point in all of this analysis is that, to obtain reasonable values in the cross-validation procedures and to avoid leakage[6] problems, all of the preprocessing steps have been done after train-test splitting the data on the train set and then applied as defined during training on the test dataset.

When it comes to cross-validation procedure the choice was made to use a stratified k-fold approach with k=10. The data is split in 10 parts with the same percentages of labels[7] then a model is built by training on 9 of the folds and it's performance is then tested on the remaining fold. To use the whole dataset for testing a prediction of it has been built by combinimg the predictions on the 10^{th}" fold for ten different models trained on the respective 9 remaining folds.

The lasso model from training on the 9 folds is actually chosen as the model with the hyperparameters that give the best performance with another 10-fold crossvalidation. This has been obtained by using *cross_val_predict* on a model obtained by including a *LassoCV* step inside a *pipeline* from scikit-learn library in python.

The performance of the cross-validated predictions that, when built this way, is much more representative of the real-world performance of the model, has been evaluated using ROC curves and AUC. Different models have been compared with a Delong test[12] for the significance of difference in the ROC curves.

The second analysis method used was RandomForest. As said before in this case no preprocessing was needed nor has been done, however particular care was taken in handling the imbalances in the dataset by using SMOTE [9] once again being careful to avoid leakage.

The performance of this model was evaluated using confusion matrices which, at a glance, provide very much information on the situation of the data. To facilitate the comparison of the results of the Random Forests with those obtained using Lasso regularized regression ROC curves were also made.

As a standalone method a Cox Proportional-Hazard model was used on the standard scaled variables remaining after the feature reduction performed through correlation thresholding. The score obtained with this procedure was then divided using different percentiles and tested using the log-rank test on Kaplan-Meier curves relative to the groups built. In the case of this thesis the time variable was represented by the days of hospitalization computed using the dates of admission and discharge from the hospital provided by the hospital itself. The idea of resorting to Kaplan-Meier curves was also used with other single variables to see if they had any effect.

Finally a few dimensionality reduction techniques, namely the unsupervised PCA[13] and Umap [28] and the supervised PLS-DA[5], have been used to understand better

[6]This term is used in the field of Machine Learning. It refers to models being created on information that comes from outside the training data.

[7]The stratified in the name refers to this property. This method is useful when dealing with unbalanced datasets, such the one under analysis, in which the label has an uneven 15-85% frequency of occurences of the two labels

the state of the data and further explain some of the obtained results. These peculiar analyses will be reported in the Appendix

A seconda di come viene decisa la struttura finale questa frase va cambiata e aggiunta la reference al punto preciso

This concludes the discussion on the methodologies used and leads perfectly in the discussion of the results.

Chapter 3

Results

In this chapter the result obtained with the methods explained in the previous chapters will be briefly presented, to ease in weaving through the quantity of results a structure of the result presentation is reported in fig 3.1.

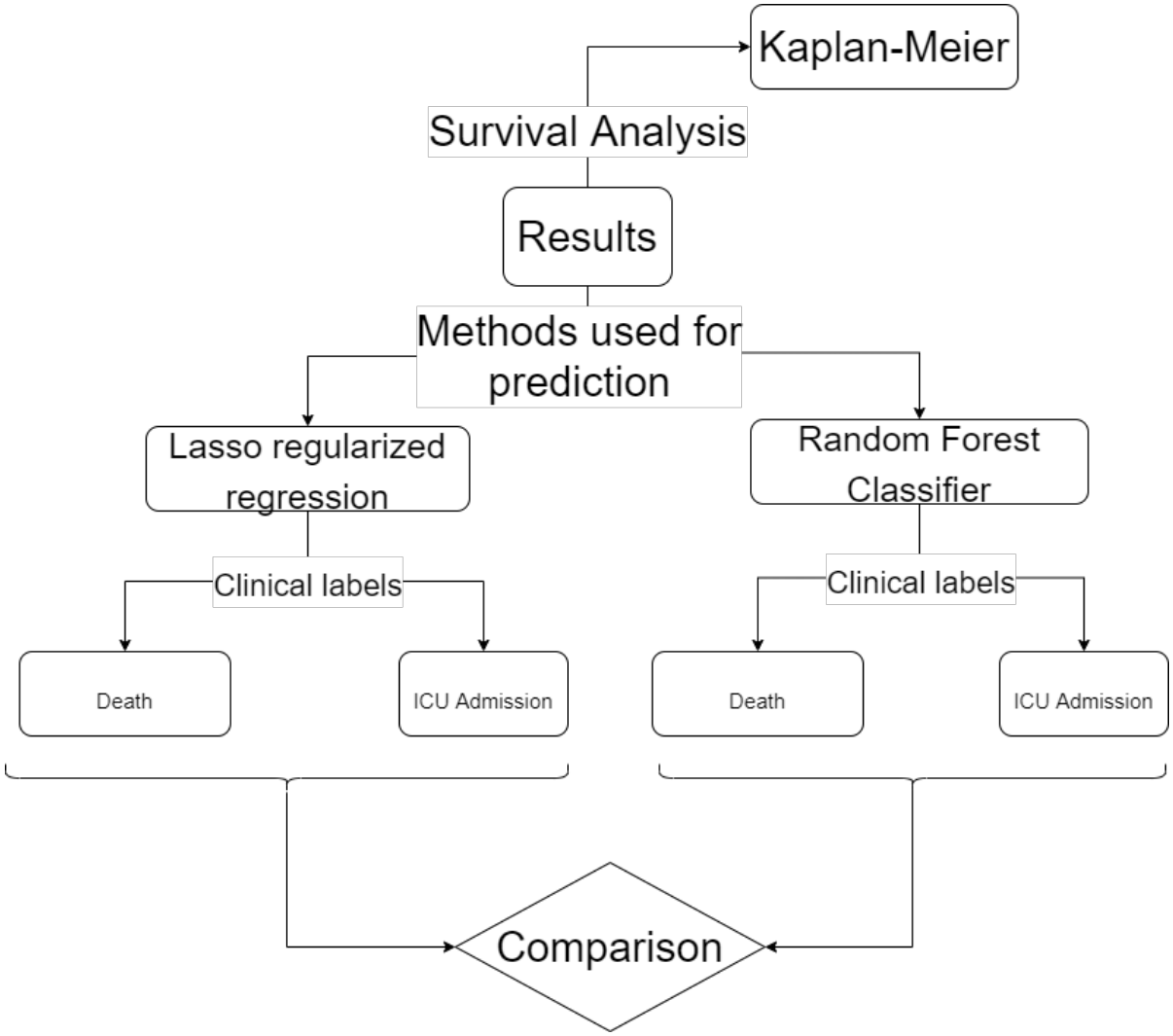


Figure 3.1: Logical structure used in the presentation of the results.

3.1 Feature selection through Lasso regularization and clinical outcomes prediction using regression

When it comes to lasso regression usually graphs are reported that show the convergence of the parameters to the final value. Since the real information of this process is the value to which the coefficients converge only these values will be presented in tables and the ROC curves, with respective AUCs, will be provided. An example of the aforementione graph is the one visible in Figure 3.2.

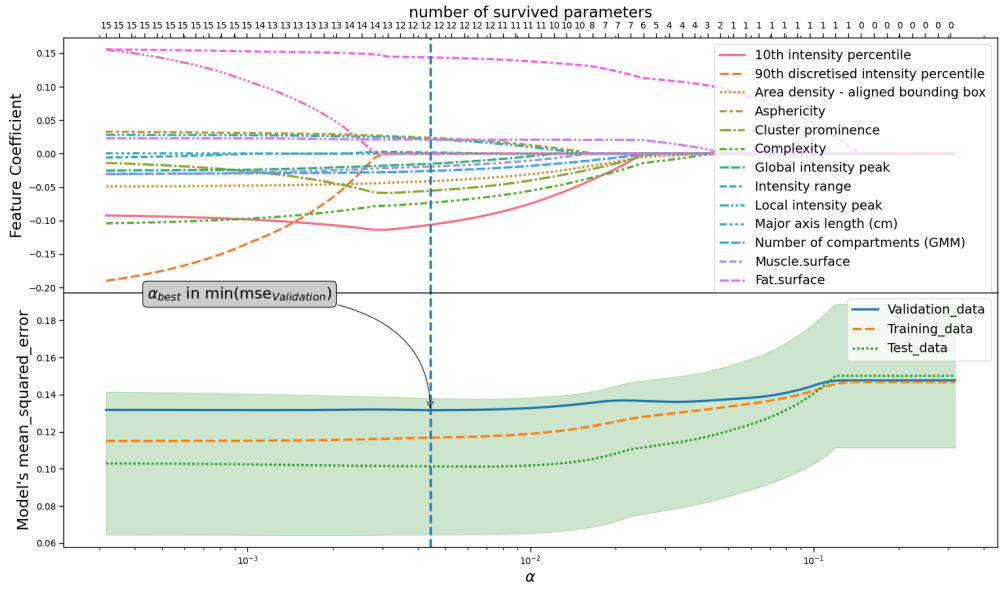


Figure 3.2: Example of graph representing the convergence of the coefficients in a lasso procedure. The final model is chosen by looking at the lowest value in mean squared error computed on the testing dataset. This graph is relative to only radiomic features with the model using DEATH as label.

Finally it seems useful to report the following contingency table that gives an idea on how superimposed the labels on DEATH and ICU ADMISSION are.

	ICU ADMISSION	
	0	1
DEATH		
0	311	47
1	48	30

3.1.1 Using Death as predicted outcome

Starting to predict the death outcome of the patient different groups of features have been used, first of all only the radiomic features have been used. The ROC curve obtained with the radiomic features is the one in Figure 3.3. The curve in bold is an average curve

obtained by aggregating the ten curves relative to each of the folds used in testing and the gray band represents a ± 1 standard deviation.

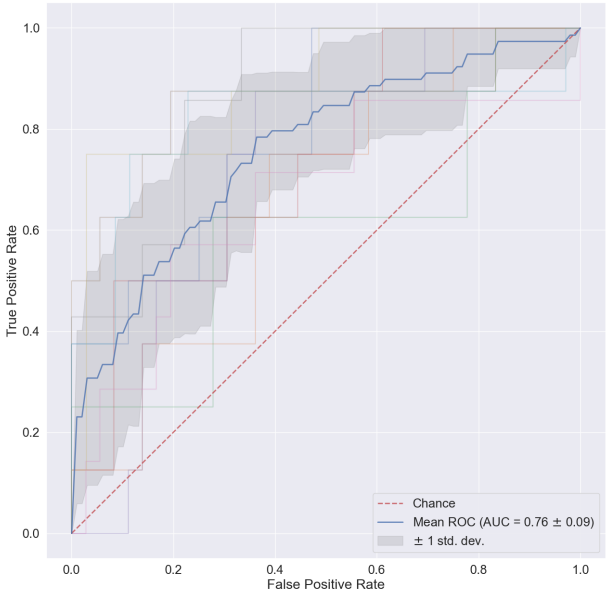


Figure 3.3: ROC curves obtained with crossvalidation procedure using the radiomic features alone. In bold is the mean ROC with gray bands of width equal to the standard deviation.

The model that was built using the coefficients reported in 3.1 reaches a AUC = 0.76±0.09. The features inside the tabular, as well as those in the tabulars that follow, will have the coefficients in descending order by absolute value so that the top features are the most relevant within it's relative model.

Before commenting more in depth the performance of this model it seems appropriate to see at least the other models built with singular features groups, so, when it comes to the clinical features, the results reported in Figure3.4 and Table3.7 have been obtained. All the results will be put together for ease in Table 3.5.

And finally, considering the radiological features Figure3.5 and Table 3.8 are obtained.

The first thing to notice is that the radiological features, when considered alone, have close to null predictive power. This is reasonable for at least part of the features because there is no reason for acquisition parameter to actually influence the outcome of the patient. When it comes to the radiologically determined quantities, such as GGO, Crazy paving, lung consolidation and bilaterlaity even if one would expect these to be relevant their distribution across the dataset is not condusive the good predictions. In fact 88% of patients had GGO, 50% of all patient had Lung consolidation, 77% of all patients did not have Crazy paving and 92% had bilateral involvment.

When it comes to clinical features, categorys that performs better when considered singularly, nothing ground breaking has been obtained. Age, Respiratory rate and sex are the most relevant features and are all very much in concordance with what is expected. Finally radiomic features perform slightly worse than the clinical features. To see if the

Table 3.1: Coefficients used in the linear combination estimated by a Lasso regularization relative to the radiomic features in modelling DEATH . All values are in descending order of absolute value

Feature Name	Importance
Intercept	0.178899
10th intensity percentile	-0.125094
Intensity-based interquartile range	0.103349
Complexity	-0.102924
Cluster prominence	-0.064690
Area density - aligned bounding box	-0.039374
Entropy	0.033002
Number of compartments (GMM)	-0.032441
Asphericity	0.028517
Local intensity peak	0.028478
Global intensity peak	-0.024832
Intensity range	0.012509
Fat.surface	0.007267
Major axis length (cm)	0.000000
Number of voxels of positive value	0.000000

Table 3.2: Coefficients used in the linear combination estimated by a Lasso regularization relative to the clinical features in modelling DEATH . All values are in descending order of absolute value

Feature Name	Importance
Intercept	0.178899
Age (years)	0.116771
Respiratory Rate	0.082292
Sex	-0.037591
Febbre	-0.022923
Hypertension	-0.000000
History of smoking	-0.000000
Obesity	0.000000

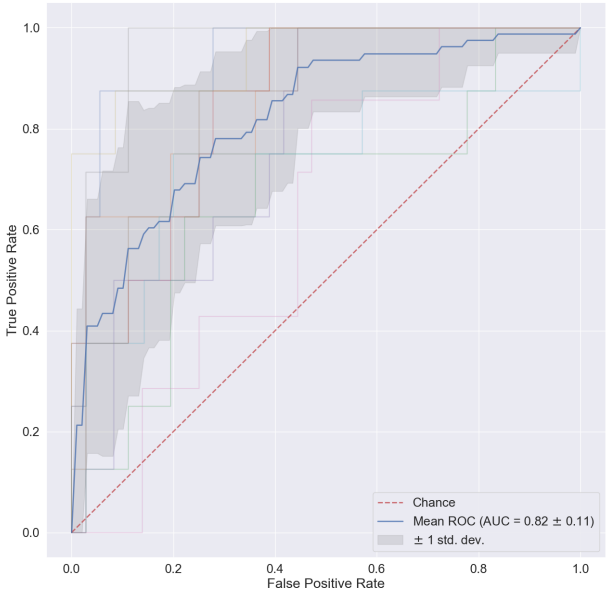


Figure 3.4: ROC curves obtained with crossvalidation procedure using the clinical features alone. In bold is the mean ROC with gray bands of width equal to the standard deviation.

Table 3.3: Coefficients used in the linear combination estimated by a Lasso regularization on a linear regression model of death, relative to the radiological features. Values are in descending order of absolute value.

Feature Name	Importance
Intercept	0.178899
Ground-glass	-0.043875
Lung consolidation	0.038143
XRayTubeCurrent	-0.017264
KVP	0.004995
Crazy Paving	-0.000000
Bilateral Involvement	0.000000
SliceThickness	0.000000

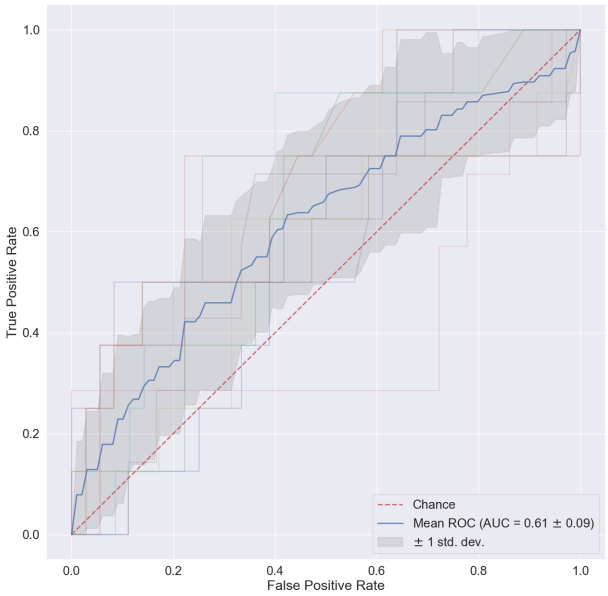


Figure 3.5: ROC curves obtained with crossvalidation procedure using the radiolocical features alone. As before in bold is the mean ROC with bands of width equal to the standard deviation.

feature obtained are, at least, reasonable a quick explanation is needed. This will be done only for the top performing features while deferring to [46] for the complete description:

- 10^{th} intensity percentile and Intensity based interquartile range are both intensity based statistics.
- Complexity: A complex image is one that presents many rapid changes in intensity and is heavily non-uniform because it has a lot of primitive components.
- Cluster Prominence: GLCM feature which measures the symmetry and skewness of the matrix from which it derives. When this is high the image is not symmetric-
- Area density aligned bounding box: This is a ratio of volume to surface.
- Entropy: Measures the average quantity of information needed to describe the image. In other words it quantifies randomness in the image, the more random the more info is needed to describe it.

To summarize, since all of the features are computed on the whole lung segmentation, it seems that the some information on the distribution of gray levels as well as some textural information inside the whole lung are important. It also seems that some information on the shape of the organ itself is also relevant.

One would expect that when combining all of the available features, i.e. by building a model using the previous clinical, radiomic and radiological features, the performance should somewhat rise especially given the fact that clinical and radiomic features have

Table 3.4: Coefficients used in the linear combination estimated by a Lasso regularization predicting death event relative to all available features features. Values are in descending order of absolute value

Feature Name	Importance
Intercept	0.178899
Age (years)	0.092963
Intensity-based interquartile range	0.057260
Respiratory Rate	0.049603
Ground-glass	-0.031423
Sex_bin	-0.028895
Complexity	-0.028606
Lung consolidation	0.017272
Febbre	-0.016933
XRayTubeCurrent	-0.016908
Area density - aligned bounding box	-0.009676
Cluster prominence	-0.006663
Fat.surface	0.004984
Number of compartments (GMM)	-0.001448
Local intensity peak	0.000195
Obesity	0.000000
Number of voxels of positive value	0.000000
Hypertension	0.000000
Intensity range	0.000000
Global intensity peak	-0.000000
Asphericity	0.000000
Crazy Paving	-0.000000
Bilateral Involvement	-0.000000
SliceThickness	0.000000
KVP	0.000000
10th intensity percentile	-0.000000
Entropy	0.000000
History of smoking	-0.000000

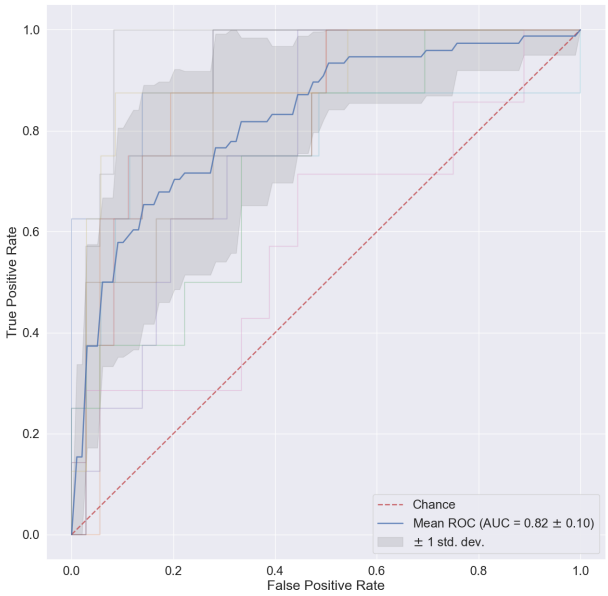


Figure 3.6: ROC curves obtained with crossvalidation procedure using all the available features. Model obtaine with Lasso regularization of a linear regression modelling death

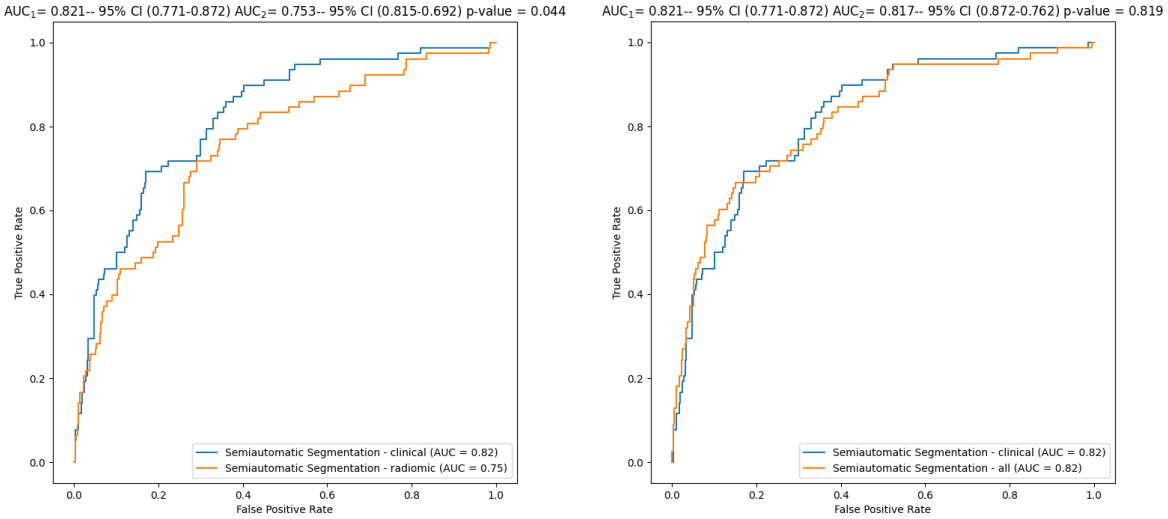
Table 3.5: Recap table with the performance of the various models predicting on different groups of features predicting DEATH

Features used	mean AUC ±std
Radiomic	0.76 ± 0.09
Clinical	0.82 ± 0.11
Radiological	0.61 ± 0.09
All	0.82 ± 0.10

almost the same performance. The combined results can be seen in Figure 3.6 and Table 3.9.

In spite of what the expectations were, the performance not only doesn't improve but it doesn't even change. To be sure of this claim a Delong test was used to compare pairwise the receiver operator curves and their respective AUCs.

The null hypothesis of this test is that the two models are the same, hence a p-value smaller than 0.05 means that the curves and their AUCs are statistically different. The results from this analysis can be seen in 3.7



(a) Comparison clincal-radiomic

(b) Comparison clinical-all

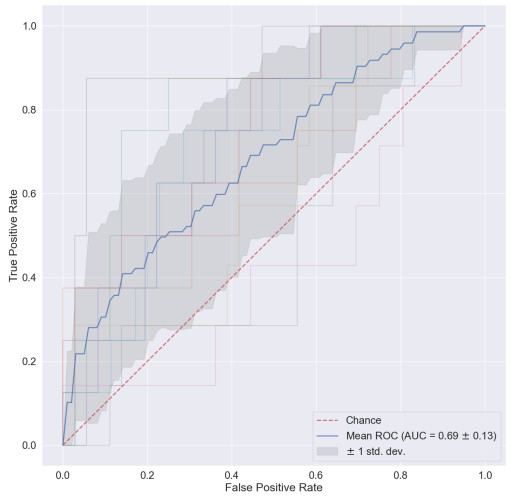
Figure 3.7: Comparison between ROC curves for clinical vs radiomic curves (a) and clinical vs all (b). The p-values are obtained with a Delong Test

As one would have expected the models from radiomic and clinical features are different while the ones built using clinical and all features are not statistically different. Inspecting the coefficient of the parameters there are a few perplexing things to notice and a few reassuring ones. First of the reassuring facts is that the most relevant clinical features are still relevant in this combined model. Then, as one would expect, the radiological features retain some importance when combined with the others. However, when it comes to perplexing behaviours, the most concerning fact is that the radiomic features have mostly lost all relevance in the model which is surely unexpected.

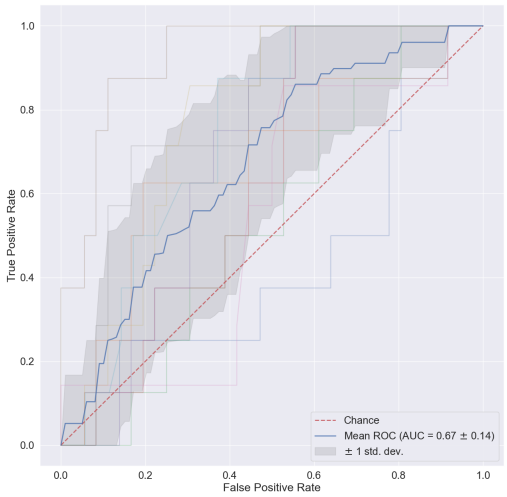
Some possible explanations will be given in the concluding remarks at the end of this subsection. as well as in the final chapter of the thesis.

3.1.2 Using ICU Admission as predicted outcome

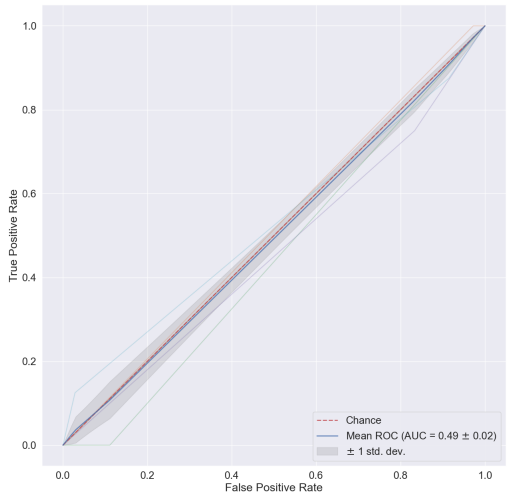
In trying to predict if the patient will be admitted in the Intensive Care Unit of the hospital the same procedure as before has been used. The ROC curves obtained with the various features are reported in Figure 3.8. Just like before the curve in bold is an average curve obtained by aggregating the ten curves relative to each of the folds used in testing and the gray band represents a $\pm$ 1 standard deviation. Following the blueprint of the previous subsection, all of the results will be presented and then briefly discussed



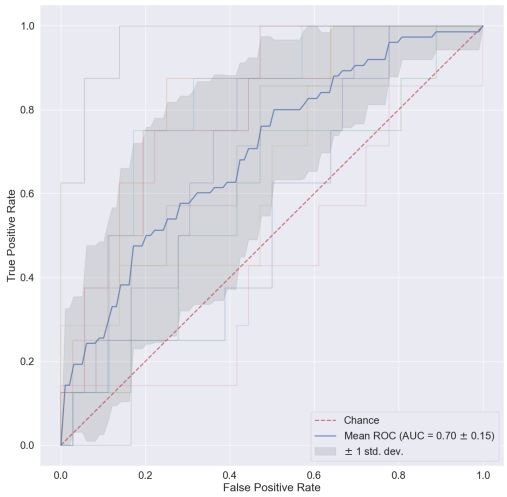
(a) Radiomic Features



(b) Clinical Features



(c) Radiological Features



(d) All Features

Figure 3.8: Performances of all the models represented using ROC curves. Each of these has in bold the mean ROC curve over the 10-fold originating from a stratified k-fold cross-validation procedure

Compared to before the performance is definitely worse. The radiological features have the same performance of a random variable, which is not that concerning given their expected impact on gravity of the clincial picture of the patient. Even if superfluous a Delong test was used to confirm that the hypothesis of the curves being equal could not be rejected. When it comes to the relevant features in each model the following can be deduced:

- For the clinical features all of them have a role in the prediction. The only suprising

Table 3.6: Coefficients used in the linear combination estimated by a Lasso regularization of a model predicting ICU Admission relative to the radiomic features. Values in descending order of modulus

Feature Name	Importance
Intercept	0.176605
Number of voxels of positive value	0.160751
Intensity range	-0.144834
Entropy	0.128999
Cluster prominence	-0.122290
Complexity	-0.093416
10th intensity percentile	-0.081133
Area density - aligned bounding box	-0.037373
Major axis length (cm)	-0.035723
Dependence count entropy	-0.029603
Fat.surface	0.027308
Asphericity	-0.023645
Local intensity peak	-0.019619
Global intensity peak	-0.016209
Number of compartments (GMM)	-0.000157

Table 3.7: Coefficients used in the linear combination estimated by a Lasso regularization of a model predicting ICU Admission relative to the clinical features. Values in descending order according to modulus

Feature Name	Importance
Intercept	0.176606
Respiratory Rate	0.045510
Febbre	0.038332
History of smoking	0.036888
Hypertension	0.034547
Sex_bin	-0.031504
Obesity	0.030716
Age (years)	-0.014646

Table 3.8: Coefficients used in the linear combination estimated by a Lasso regularization of a model predicting ICU Admission relative to the radiological features

Feature Name	Importance
Intercept	1.766055e-01
XRayTubeCurrent	0
Lung consolidation	0
Ground-glass	0
Crazy Paving	0
Bilateral Involvement	0
SliceThickness	0
KVP	0

Table 3.9: Coefficients used in the linear combination estimated by a Lasso regularization of a model predicting ICU Admission relative to all available features features. Values in ascending absolute value order

Feature Name	Importance
Intercept	0.176605
Number of voxels of positive value	0.109947
Dependence count entropy	0.070527
Cluster prominence	-0.069526
Intensity range	-0.057924
Febbre	0.044149
Hypertension	0.039127
SliceThickness	-0.037174
Complexity	-0.033393
History of smoking	0.032812
Age (years)	-0.032579
XRayTubeCurrent	-0.032182
Respiratory Rate	0.028504
Obesity	0.027580
Local intensity peak	-0.026135
Area density - aligned bounding box	-0.023027
Asphericity	-0.022999
Global intensity peak	-0.018280
Fat.surface	0.014179
Sex_bin	-0.004316
Crazy Paving	-0.003428
Ground-glass	0.003077
Lung consolidation	-0.001329
Bilateral Involvement	-0.001047
Number of compartments (GMM)	-0.000000
KVP	0.000000
10th intensity percentile	-0.000000

Table 3.10: Recap table with the performance of the various models built for different families of features when predicting ICU Admission

Features used	mean AUC ±std
Radiomic	0.69 ± 0.13
Clinical	0.67 ± 0.14
Radiological	0.49 ± 0.02
All	0.70 ± 0.15

fact, even if it keeps a certain degree of plausibility, is that age is the less relevant out of the available features when it comes to ICU Admission .

- None of the radiological features have virtually any impact
- The radiomic features still value intensity measurements and disorder in the image as primary origins of information. However it seems that shape of the lung now has more relevance in the whole model.

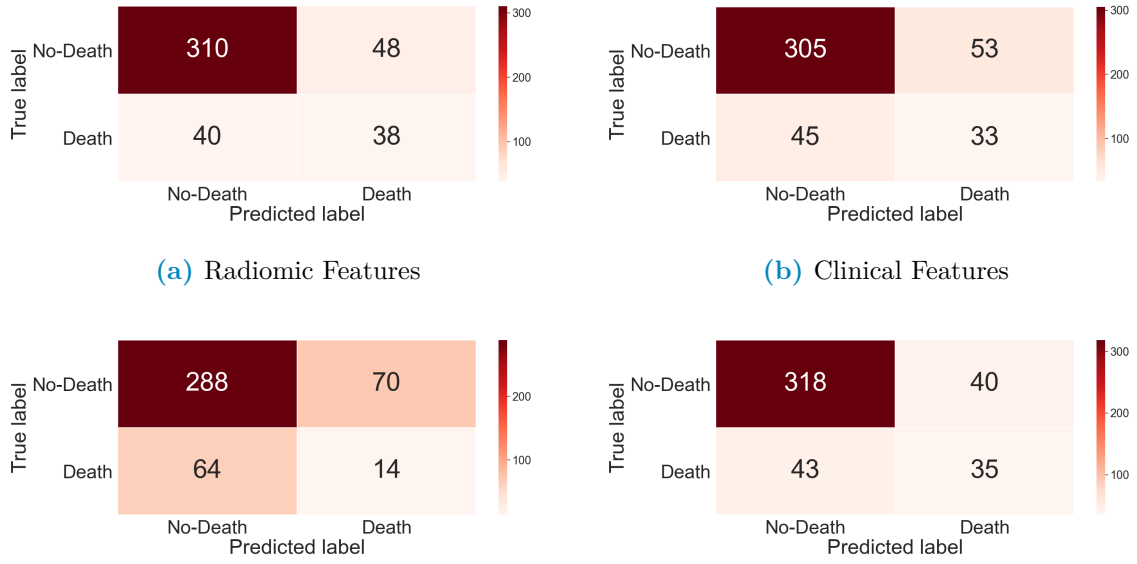
3.2 Classification of patients using Random forests

When it comes to random forests the approach followed was similar, in the sense that the various combinations of features were tried, yet different, because the preprocessing step consisted in simply applying a standard scaler to the data. The performance of the models has been looked at using both confusion matrices and ROC curves, the last of which has the only objective of comparing RF classifiers to the previously described Lasso regression.

Once again, following the description scheme used in the section dedicated to Lasso, the results will be divided using the two labels ICU Admission and Death .

3.2.1 Classifying the oucome Death

For the sake of brevity all of the results will be reported and then discussed. Since RF classifiers use all of the availalble features it is very space consuming to report a table with all of the importances for the radiomic features as well as those used in the models with all the features. These two will be found in the appendix 6.1 while those relative to clinical and radiological features will be reported here in Table **??**.



(a) Radiomic Features

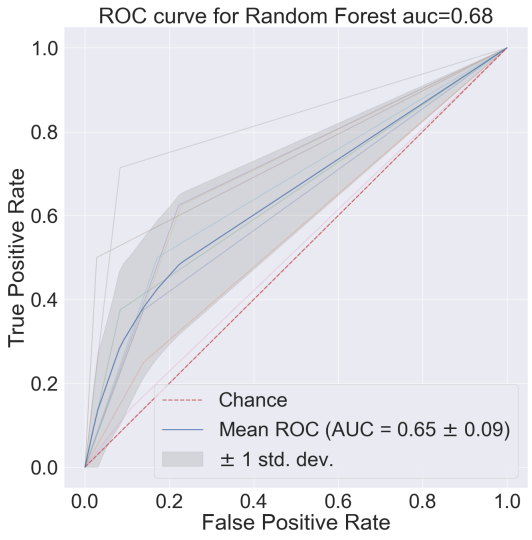
(b) Clinical Features

(c) Radiological Features

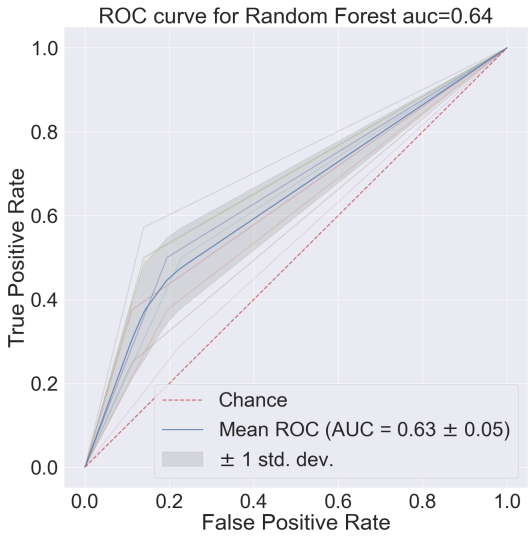
(d) All Features

Figure 3.9: Confusion matrices for Random Forest cross-validated predictions after training on Synthetically oversampled data to predict DEATH . All of the available feature families are reported

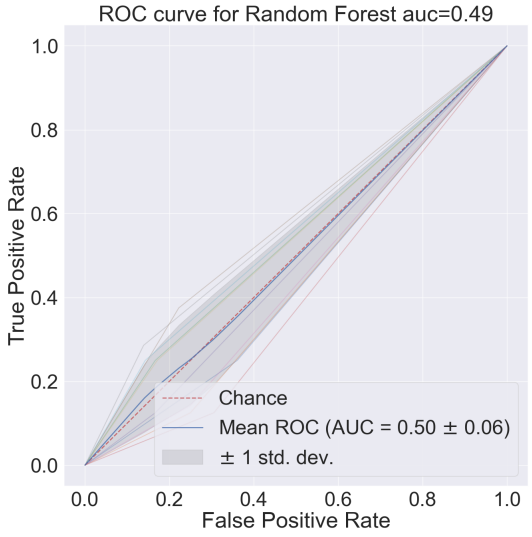
Even without looking at the ROC curves it's plain to see that the data at hand is proving to be difficult for this model. Much like before radiological features alone are useless. In evaluating these confusion matrices it should be kept in mind that the data is heavily unbalanced, since only ∼15% of the patient died or were admitted in the ICU. Even when using SMOTE in the training phase to correct this probelm it seems that the classifiers learns that it's optimal to guess that someone is alive. When it comes to the ROC curves Figure 3.10 and Table 3.11 summarises the results.



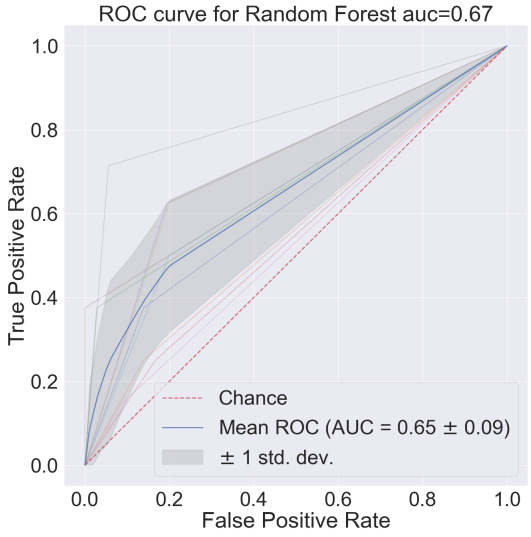
(a) Radiomic Features



(b) Clinical Features



(c) Radiological Features



(d) All Features

Figure 3.10: Cross-validated ROC curves built with Random forest classifier predictions of DEATH . Performances of allvariable families are reported

Table 3.11: Recap table with the performance of the various families of features

Features used	mean AUC ±std
Radiomic	0.65 ± 0.13
Clinical	0.64 ± 0.06
Radiological	0.50 ± 0.07
All	0.66 ± 0.8

	RF_importances
Age (years)	0.463821
Respiratory Rate	0.287735
Febbre	0.084276
Sex_bin	0.079571
Hypertension	0.033435
History of smoking	0.029404
Obesity	0.021758

(a) Radiological Features

	RF_importances
XRayTubeCurrent	0.800101
Lung consolidation	0.043942
KVP	0.039768
Crazy Paving	0.032511
SliceThickness	0.031362
Ground-glass	0.030423
Bilateral Involvement	0.021893
HRCT performed	0.000000

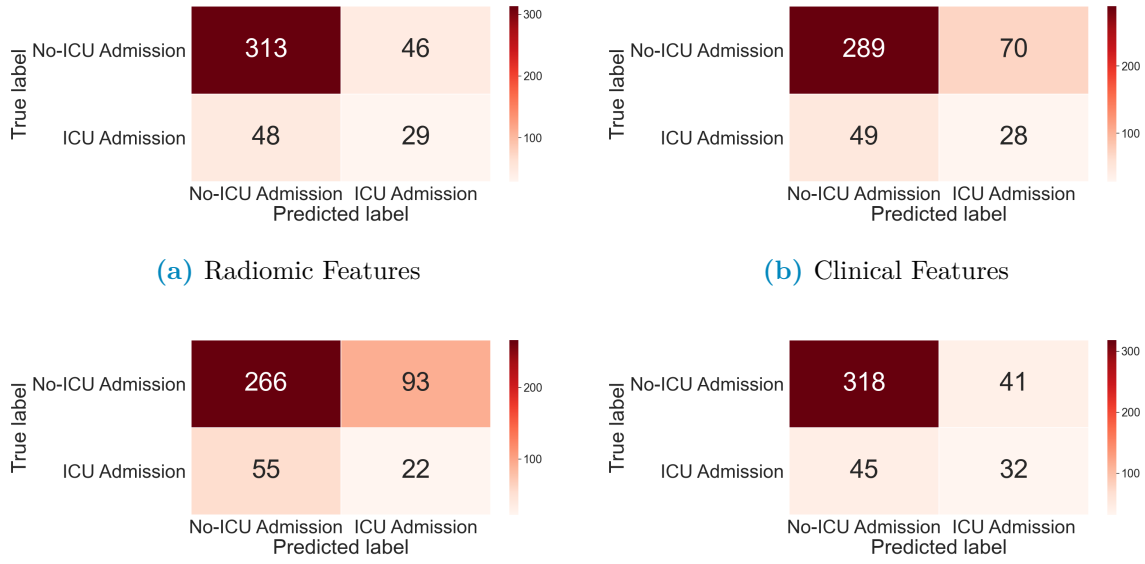
(b) Clinical Features

Figure 3.11: Importances estimated by random forest

Once again the curves are evidently not statistically different. This counterintuitive behaviour seems to be constant across the two implemented methods and also across different labels tried. An attempt to explain this phenomenon will be postponed to the end of the next subsection

3.2.2 Classifying the outcome ICU Admission

Even when using the admission in the ICU the performance remains pretty much the same, so the comments would still be the same as before



(a) Radiomic Features

(b) Clinical Features

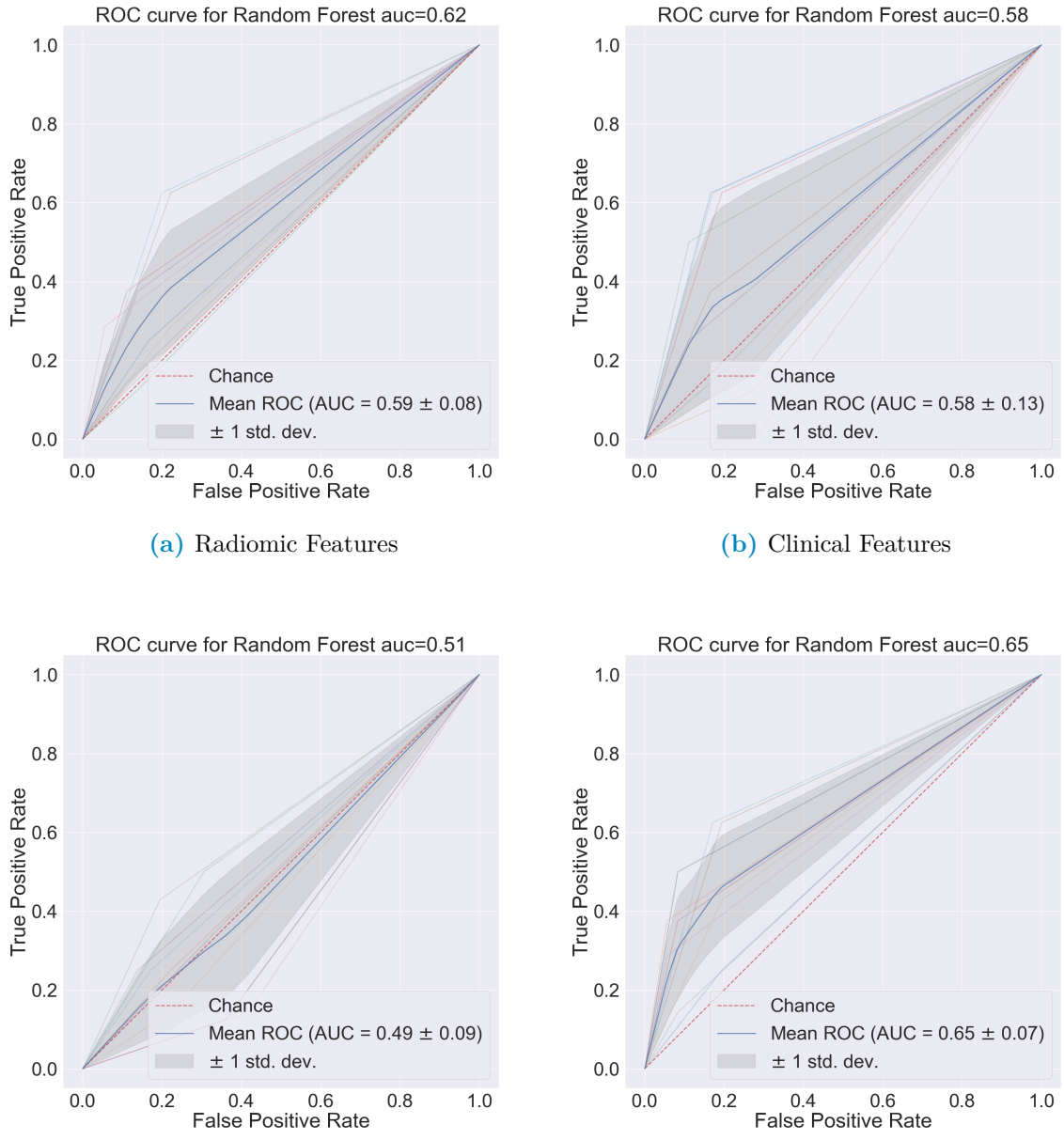
(c) Radiological Features

(d) All Features

Figure 3.12: Confusion matrices for Random Forest cross-validated predictions after training on Synthetically oversampled data predicting ICU Admission . All of the available feature families are the reported

Table 3.12: Recap table with the performance of the various families of features

Features used	mean AUC ±std
Radiomic	0.62 ± 0.08
Clinical	0.56 ± 0.08
Radiological	0.51 ± 0.11
All	0.64 ± 0.09



(a) Radiomic Features

(b) Clinical Features

(c) Radiological Features

(d) All Features

Figure 3.13: Cross-validated ROC curves built with Random forest classifier predictions of DEATH . Performances of allvariable families are reported

Table 3.13: Results obtained with CoxPH fitter from lifelines library

covariate	coef	exp(coef)	se(coef)	p	-log2(p)
Lung consolidation	0.166411	1.181058	0.142506	0.242908	2.041517
Ground-glass	0.100946	1.106217	0.134109	0.451619	1.146822
Crazy Paving	0.064744	1.066886	0.140817	0.645680	0.631108
Bilateral Involvement	-0.026048	0.974288	0.121990	0.830918	0.267222
SliceThickness	-0.008439	0.991597	0.164518	0.959092	0.060259
KVP	0.376184	1.456715	0.139983	0.007202	7.117333
XRayTubeCurrent	-0.272076	0.761796	0.178915	0.128335	2.962010
Age (years)	-0.016550	0.983587	0.160470	0.917858	0.123657
Hypertension	0.292450	1.339705	0.160211	0.067940	3.879603
History of smoking	-0.083840	0.919578	0.139121	0.546747	0.871054
Obesity	0.066777	1.069057	0.170581	0.695451	0.523979
Respiratory Rate	-0.010716	0.989341	0.154003	0.944525	0.082338
Sex_bin	-0.366086	0.693443	0.172651	0.033974	4.879413
Febbre	0.115458	1.122387	0.139878	0.409134	1.289354
10th intensity percentile	0.324188	1.382908	0.230611	0.159790	2.645753
Area density - aligned bounding box	-0.169228	0.844316	0.185602	0.361884	1.466401
Asphericity	-0.470777	0.624517	0.174449	0.006962	7.166236
Cluster prominence	-0.011242	0.988821	0.234222	0.961720	0.056312
Complexity	0.214330	1.239032	0.248380	0.388186	1.365179
Global intensity peak	0.138264	1.148279	0.165033	0.402146	1.314210
Intensity range	-0.196751	0.821395	0.281701	0.484901	1.044237
Local intensity peak	0.132819	1.142044	0.150161	0.376419	1.409589
Number of compartments (GMM)	0.197884	1.218821	0.140554	0.159164	2.651410
Number of voxels of positive value	0.436757	1.547680	0.284518	0.124764	3.002721
Fat.surface	-0.425033	0.653748	0.208060	0.041069	4.605815
Normalised zone distance non-uniformity	0.589917	1.803838	0.242437	0.014963	6.062489

3.3 Using survival analysis

Following the preprocessing steps delineated before, a Cox Proportional-Hazard produces the results presented in Table 3.13.

As explained in section 1.4 the relevalt columns are the coeff column, that expressed percentual difference of survival, and the p column, that indicate the significance of the first value. It turns out that, out of the reduced variables fed to the Cox model, the most relevant are: Sex, Asphericity, FatSurface, Normalized zone distance non-uniformity and KVP.

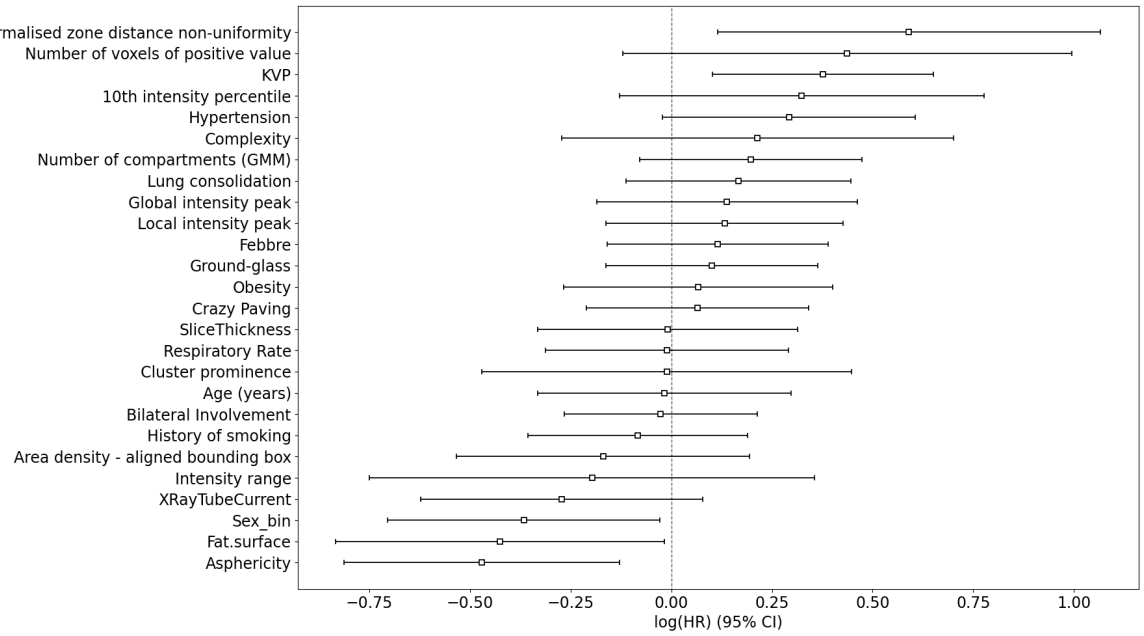
Zone distance non uniformity measures distribution of zone counts over the different zone distances, it is low when the count relative to the zones are equally distributed along zone distances

Asphericity quantifies how much the segmented region deviates from a sphere.

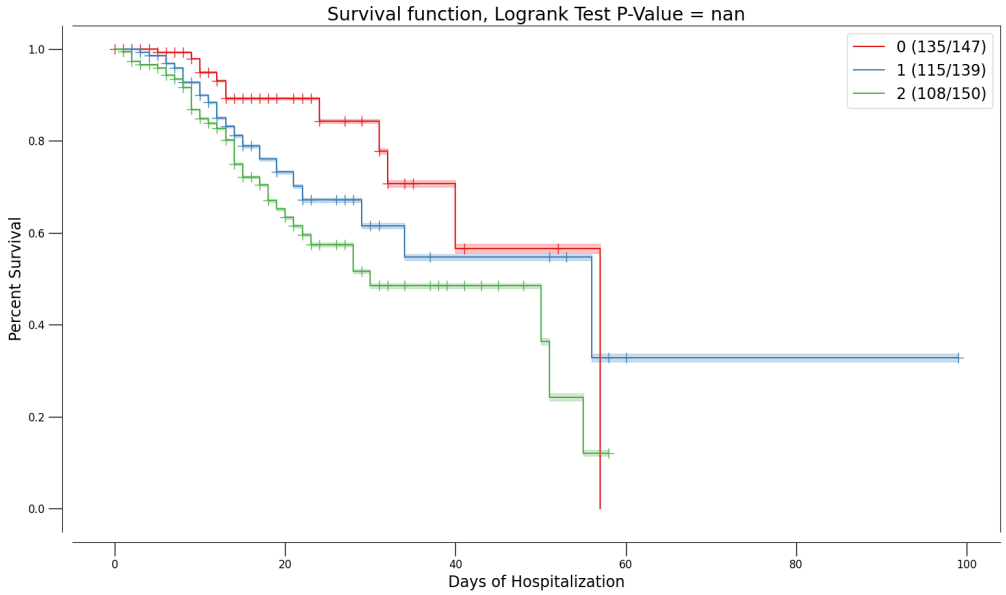
In this caseSex and FatSurface and Normalized zone distance non-uniformity can be reasonable variables to expect, however KVP, Asphericity seem quite strange.

A score was built automatically using the predict method of the CoxPH fitter and assigned to each patient of the dataset using the previously described cross-validated prediction procedure. To see if the prediction was representative of differences in the individuated populations first the Kaplan-Meier curves according to thirds in the score

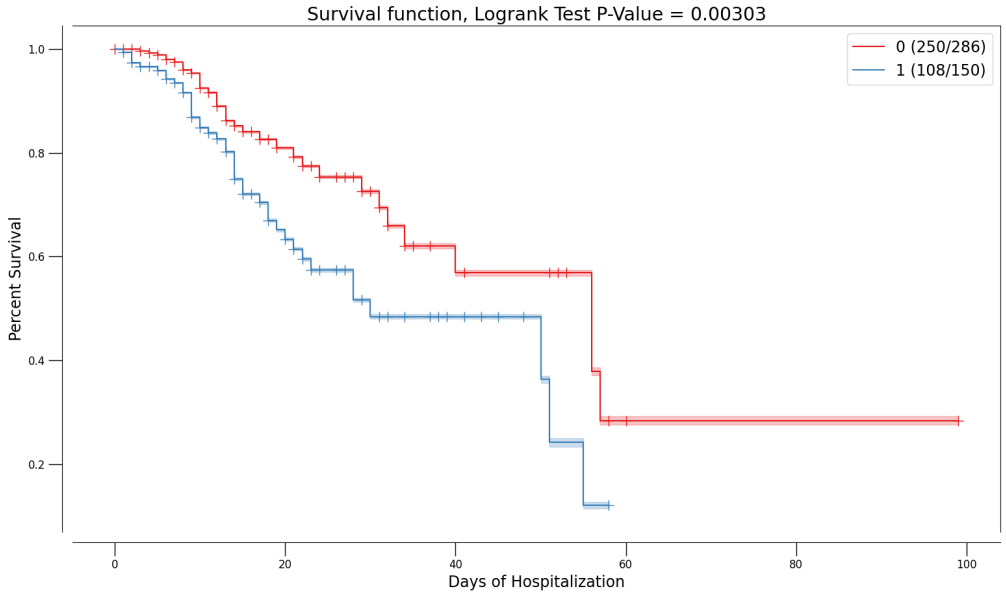
Figure 3.14: Graph that represents the coefficient values estimated by the CoxPH model with their respective 95% confidence intervals



distribution were used and then the score was binarized using the 66^{th} percentile in the score distribution as threshold. The results of this procedures are reported in Figure 3.15



(a) Population divided according to score tertiles



(b) Population divided in two groups 0-66th percentile and 66th to 100th

Figure 3.15: Kaplan-Meier curves for pupulations divided using either tertiles in the predicted hazard by the cox model (a) or binarized using 66th percentile as threshold (b) . The prediction on the whole database is obtained with the aforementioned cross-validation procedure

It can be seen that the groups built in this ways can be used to drive some differences in survival, when binarizing the score obtained with Cox the curves also turn out to be significantly different.

Finally, in order to see if there were differences in treatment or in survival between the two waves of admission, the population was divided in two subgroups according to

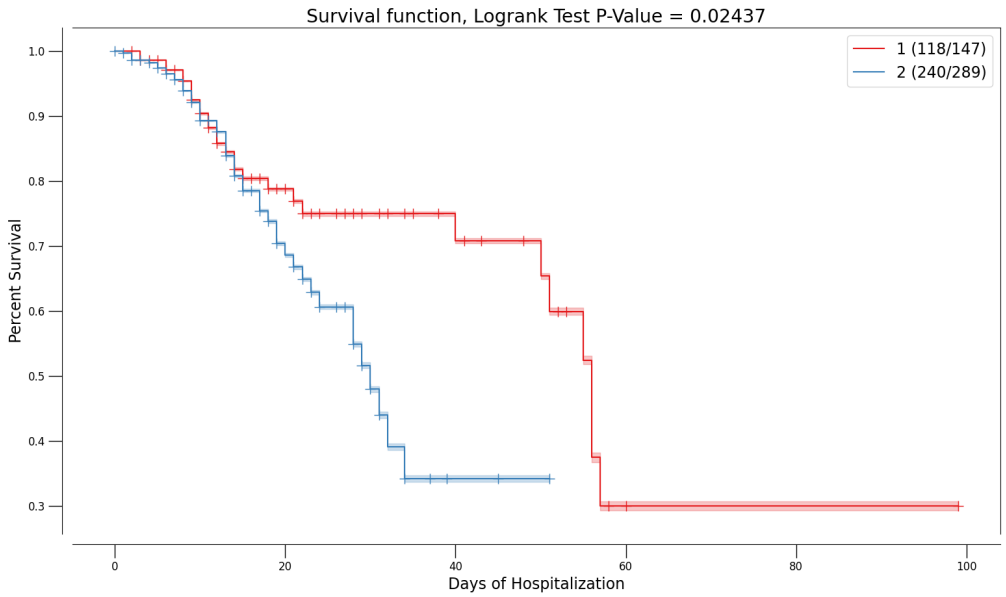


Figure 3.16: Kaplan-Meyer curves for patient admitted before (red curve) and after (blue curve) 20/07/2020

the date of admission. The two groups were representatively called 1^{st} and 2^{nd} wave and the division was drawn on the 20^{th} of July 2020 and the results can be seen in Figure 3.16.

It can be seen that there is a statistical difference in survival between the patients admitted in the first wave vs those admitted in the second. Furthermore this difference is quite perplexing as it seems to indicate that people in the second wave died more than people in the first wave, which seems counter-intuitive given that one would expect the experience from the previous wave to improve performance. The most reasonable explanation for this fact is the change in admission policy as time advanced. Probably in the first wave, when still little was known on *Sars-COVID19* patients, more people were addmitted in less problematic condition whereas in the second wave, having understood better what were the most dangerous cases as well as in an attempt to admitt only those strictly in need, most of the admitted patients were in more critical condition.

It's also possible that this result that has been obtained could be a symptom of subtle differences in the two *Sars-COVID19* manifestations, as if to indicate different variants. Further analysis in this direction could be a followup work of this thesis.

Chapter 4

Summarizing and discussing the results; possibilities for further develpement

Having presented all of the results obtained in this thesis it has been seen that:

- The chosen preprocessing method followed by Lasso regularized regressions perform overall well when it comes to predicting either Death or ICU Admission
- Random Forest classifiers are consistently worse than Lasso regularized regressions, probably mainly due to the unbalancedness of the dataset at hand.
- Cox proportional Hazard allows us to distinguish at least two groups with statistically different Survival curves.
- Combining clinical, radiomic and radiological information does not determine an improvement when compared to clinical features alone. This result is quite perplexing and it could have multiple causes

It is very difficult to quantify what the expectations for each model were a priori, this transposes to difficulties in trying to diagnose problems in the models when considered singularly. However when combining all of the available features it's very strange that no improvement in performance can be achieved. This would mean that 8 clinical variables provide the exact same amount of information as the total $\sim$200 features most of which derived from images which, at least ideally, contain much more accurate information.

One could surmise that the analysis methods have been implemented in an incorrect way, yet when two methods implemented differently and separately obtain the same result it reinforces the idea that the problem lies somewhere before the analysis. All of these analses started from the following hypotheses:

1. Clinical labels are informative of the final prognosis of the patient
2. Radiological images contain a lot of useful information
3. Radiomics can extract these information
4. This information is informative of the prognosis of the patient

It is very clear that the first three hypotheses are verified with a caveat, the performance of the radiomic pipeline hinges on the quality of the images and of the segmentation procedure performed on them. Since the images used in this thesis were, are and will be used by the hospital in routine processes it's close to impossible that all of them have problems, expecially because of the preliminary screening done before segmentation.

A first possibility is that when trying to segment lungs affected by *Sars-COVID19* the peculiar patterns developed make the task much harder then expected. In fact considering that the method used for this thesis relies on region growth and thresholding methods it's possible that the worst cases end up with unrepresentative segmentations.

Another possibility is that the images are not really representative of the situation of the patient. Since one of the prevailing properties of *Sars-COVID19* is the speed with which the clinical picture of the patient can change it's possible that images taken at admission are not as informative of the final prognosis. This problem is, however, unavoidable in the setting of this thesis which has as aim the construction of a model that, exactly at admission, can discriminate between serious and easier cases.

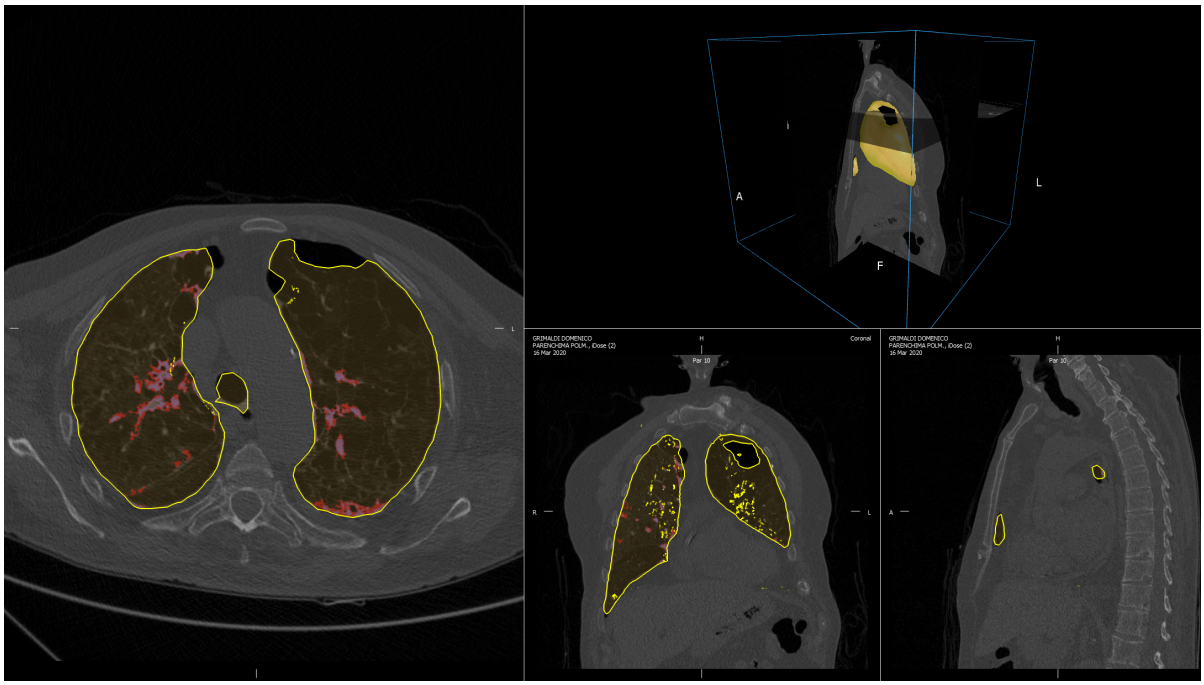
Another possibility is that there are two or more subgroups in the patient cohort and the performance on these is widely different, determiniing an average performance below the expectations. To diagnose if this is the case a few dimensionality reduction techniques have been used to visualize the data and to prepare for clustering in case of need, all of the results of these procedure will be presented in the followig section.

A questo punto magari cito l'appendice con la parte di clustering quando sarà pronta

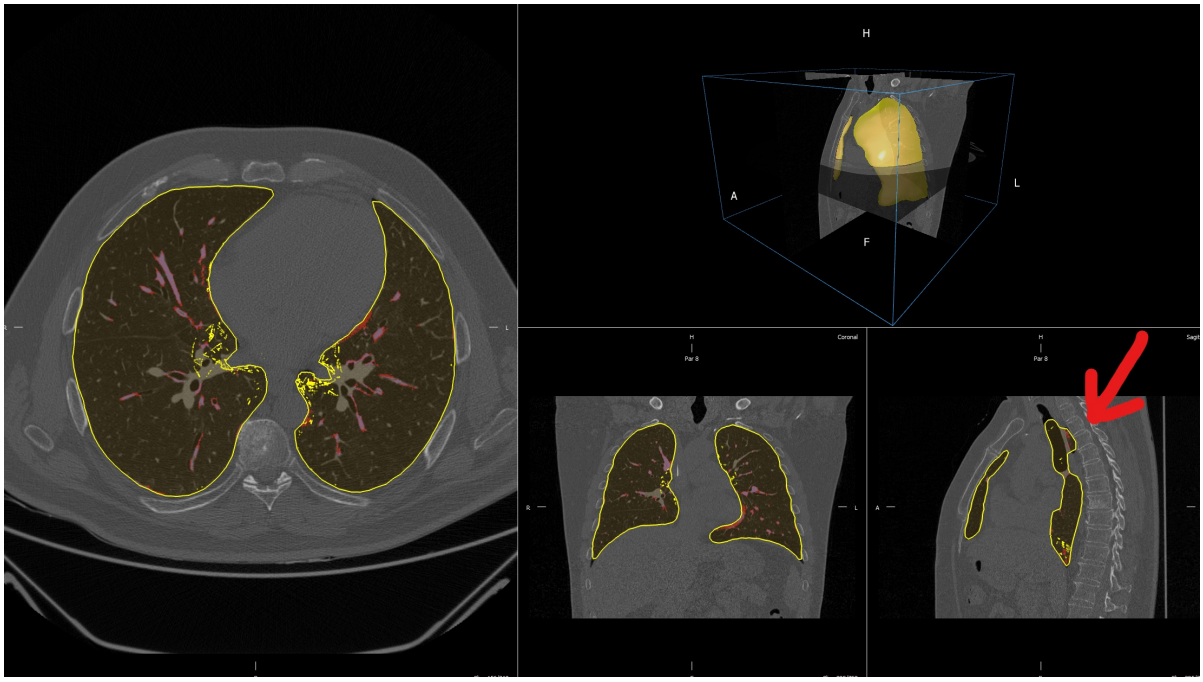
To give a qualitative idea of the situation regarding the segmentations $\sim$70 were looked at and evaluated as good, unsure or bad.[1]

Good segmentations are those in which an untrained professional could not see any problems, unsure are those in which there are small inaccuracies, such as lungs that connect in some points and the inclusion of the trachea. Finally segmentation were classified as bad in cases with obvious errors, such as part of the intestine being labelled as lung, damages in the lung being labelled as outside tissue or holes in what is supposed to be lung.

[1]This was done on the first patients in alphabetical order which was considered to be equivalent to random since there is no reasonable motive for surnames to be correlated with segmentation quality.



(a) Example of segmentation classified as bad



(b) Example of segmentation classified as dubious

Figure 4.1: Example of segmentations being classified as bad (a) and dubious (b). In the first case (a) there is a clear hole in the lung which cannot be exact, in the second (b) there are small portion of outside tissue being labelled as lung as well as the whole trachea.

The result of this qualitative analysis can be seen in Table 4.1 in which the incidence of both DEATH and ICU ADMISSION labels is computed in all possible segmentation categories.

This table suggests that there is large difference of very severe individuals which

Table 4.1: Contingency table with number of DEATH and ICU ADMISSION labels in all segmentation groups.

ICU Admission	Segmentation Status Death	Subject Bad	Good	Unsure
0	0	8	11	24
	1	5	0	6
1	0	2	0	2
	1	1	1	1

end up being not perfectly segmented. It's possible, even if unlikely, that the sample of analysed segmentations is biased in some way however, given the numbers in Table 4.1, it's reasonable to fear that in the most important cases, which means those that correspond to 1 labels in DEATH or ICU ADMISSION , the values of the radiomic features used in this thesis were not the best possible.

Another thing that can be noted is that in all models that included radiomics feature there are, among the important variables, features that quantify volumetric or shape information regarding the lung.

A priori it's very difficult to imagine how lung size by itself could determine the prognosis of the patient. However it's possible that most of the severe cases, corresponding to the most difficult segmentation to perform, end up being associated to lung shapes that are significanlty different from almost healthy patients. In this light it's possible that the model is using these shape information, which is naturally unrepresentative of reality, to infer the gravity of the situation and is finding something simply because the information is the most unrealistic in the most severe cases.

Chapter 5

Conclusion

In this thesis various methods were used in an attempt to predict the prognosis of *Sars-COVID19* patients.

Regularized regression was used to predict clinical outcome using various families of variables to compare the information hidden in each of them and to evaluate if CT exams add any value to a small set of clinical variables.

Random forest classifiers were used with the same aim. As a by-product these two methods can be compared to see which, given the same data, can extract the most information.

Survival analysis was used, mostly by itself, to see if the data could be divided in smaller groups with different survival functions.

In the first two lines of developement it was found that, in the specific case of data available for this thesis, radiomic features added no information to the clinical variables. This was taken as a (grim) reminder that the quality of the segmentations is very relevant in determining the results obtained by radiomics analysis and that current commercial segmentation techniques based on thresholding and region growing are not yet ready to face the problem posed by *Sars-COVID19* pneumonia lungs.

However the results obtained can be looked at from another perspective. The method developed, which also shows potential for improvement, has shown performances on par with those obtained using clinical variables. This means that, especially in times of system overload due to increased accesses during a pandemic, this system could be integrated in PACS systems of the hospital so bring to attention some of the patients in worst condition. This, especially thanks to the objectivity of radiomics, the user-independence and, most of all, to the semi-automatic nature of the pipeline could be done on a large scale to aid in hospitals that don't have the facilities , or lack the personnel, to analyze in detail all cases. Finally some directions in which future works could start from this thesis are:

- The implementation of a *Sars-COVID19* specific segmentation method using the vast amount of available data, similar to what has been done in [7].
- It would be interesting to test the pipeline, as well as the deriving models, on manually segmented CT scans to compare the results and performances.
- Given the statistical difference found in the survival of patients in the first and second waves, defined here as before and after 20/07/2020, it might be very interesting

to investigate the causes of this finding and to prospectively continue this analysis with the data from the third wave and eventual next ones that might occur.

- Eventualmente se esce roba dal clustering

Chapter 6

Appendix

6.1 Additional Results and complete tables relative to Random Forest

Here are the tables with all of the features from random forest. There is no real utility in providing them for all possible feature combination, hence only the importances for all features will be given. This will be done for both labels used, i.e. Death and ICU Admission.

Feature Name	Importance estimated by Random Forest
Age (years)	0.056516
CURB65	0.023775
Intensity histogram quartile coefficient of dis...	0.021053
Discretised interquartile range	0.017892
Ground-glass	0.015137
Dependence count entropy	0.014432
Intensity-based interquartile range	0.014269
Small zone emphasis	0.014180
Zone size entropy	0.013493
Normalised zone size non-uniformity	0.013321
Skewness	0.013009
Dependence count energy	0.012736
Information correlation 1	0.011409
Information correlation 2	0.011391
Intensity-based median absolute deviation	0.011173
Quartile coefficient of dispersion	0.010367
Respiratory Rate	0.010176
Entropy	0.009969
Intensity histogram median absolute deviation	0.009406
Run entropy	0.009264
Volume density - enclosing ellipsoid	0.009191
Intensity histogram robust mean absolute deviation	0.009046
Uniformity	0.008612
Discretised intensity skewness	0.008386

Intensity-based robust mean absolute deviation	0.008134
Maximum histogram gradient intensity	0.007806
Grey level variance (GLDZM)	0.007664
Intensity-based mean absolute deviation	0.007580
Normalised grey level non-uniformity (NGLDM)	0.007455
Fat.surface	0.007220
Febbre	0.007157
Sum entropy	0.006892
Local intensity peak	0.006887
Minor axis length (cm)	0.006875
Area density - enclosing ellipsoid	0.006777
Grey level variance (GLSZM)	0.006726
Angular second moment	0.006453
Cluster shade	0.006377
XRayTubeCurrent	0.006247
Max value	0.006077
Zone distance non-uniformity	0.005951
Normalised zone distance non-uniformity	0.005922
Small distance emphasis	0.005790
Cluster prominence	0.005772
RECIST (cm)	0.005728
Large distance high grey level emphasis	0.005609
Normalised grey level non-uniformity (GLRLM)	0.005479
Low dependence emphasis	0.005422
Small distance low grey level emphasis	0.005405
Normalised grey level non-uniformity (GLSZM)	0.005318
Grey level non-uniformity (NGLDM)	0.005307
Volume density - convex hull	0.005301
Volume at intensity fraction 90%	0.005267
Large distance emphasis	0.005247
Normalised homogeneity	0.005180
Dependence count non-uniformity	0.005174
Small zone low grey level emphasis	0.005110
Number of grey levels	0.005006
Area density - convex hull	0.004984
Low grey level zone emphasis.1	0.004983
10th intensity percentile	0.004967
Intensity histogram mean absolute deviation	0.004965
Intensity median value	0.004960
Discretised intensity kurtosis	0.004954
Energy	0.004950
High dependence low grey level emphasis	0.004895
Integrated intensity	0.004888
Small distance high grey level emphasis	0.004863
Normalised inverse difference	0.004814
Zone distance entropy	0.004801
Normalised grey level non-uniformity (GLDZM)	0.004749

Difference average	0.004743
Thresholded area intensity peak (50%)	0.004735
Centre of mass shift (cm)	0.004683
Minimum histogram gradient	0.004634
Number of voxels	0.004592
Low grey level zone emphasis	0.004470
Area density - oriented bounding box	0.004465
Volume density - aligned bounding box	0.004453
High dependence emphasis	0.004453
Intensity-based coefficient of variation	0.004442
Thresholded area intensity peak (75%)	0.004426
Discretised intensity uniformity	0.004342
Low grey level count emphasis	0.004310
Grey level non-uniformity (GLDZM)	0.004261
Contrast (GLCM)	0.004247
Difference entropy	0.004174
Kurtosis	0.004151
Grey level non-uniformity (GLRLM)	0.004114
Number of compartments (GMM)	0.004110
Intensity-based energy	0.004093
Small zone high grey level emphasis	0.004086
Least axis length (cm)	0.004086
Intensity histogram mode	0.004073
Volume density - oriented bounding box	0.004064
Inverse variance	0.004055
Difference variance	0.004024
Surface to volume ratio	0.003966
Run length variance	0.003913
Variance	0.003910
Correlation	0.003908
Muscle.surface	0.003907
High grey level zone emphasis	0.003879
Number of voxels of positive value	0.003857
Inverse elongation	0.003853
Cluster tendency	0.003820
Intensity range	0.003804
Normalised run length non-uniformity	0.003799
Large zone high grey level emphasis	0.003776
Long run low grey level emphasis	0.003766
Area density - aligned bounding box	0.003687
Zone percentage (GLDZM)	0.003660
Asphericity	0.003657
Grey level variance (NGLDM)	0.003643
Intensity at volume fraction 90%	0.003617
Volume at intensity fraction 10%	0.003610
Major axis length (cm)	0.003604
Low dependence low grey level emphasis	0.003570

Run length non-uniformity	0.003548
Strength	0.003459
Long run high grey level emphasis	0.003433
Mean discretised intensity	0.003413
Low dependence high grey level emphasis	0.003409
Dissimilarity	0.003395
High grey level count emphasis	0.003394
SliceThickness	0.003389
Grey level non-uniformity (GLSZM)	0.003381
Volume fraction difference between intensity fr...	0.003381
Grey level variance (GLRLM)	0.003369
Short run low grey level emphasis	0.003347
Maximum histogram gradient	0.003338
High dependence high grey level emphasis	0.003321
Compactness 2	0.003317
Long run emphasis	0.003316
Autocorrelation	0.003261
Joint maximum	0.003254
Global intensity peak	0.003237
Sum average	0.003233
Low grey level run emphasis	0.003187
Dependence count variance	0.003182
Intensity at volume fraction 10%	0.003128
Large distance low grey level emphasis	0.003125
Zone percentage (GLSZM)	0.003088
Intensity fraction difference between volume fr...	0.003034
Zone distance variance	0.002949
Maximum 3D diameter (cm)	0.002940
Normalized dependence count non-uniformity	0.002937
Inverse difference	0.002912
Intensity histogram coefficient of variation	0.002872
Coarseness	0.002836
Run percentage	0.002805
Flatness	0.002788
Standard deviation	0.002760
Joint variance	0.002714
Busyness	0.002711
Intensity mean value	0.002706
Homogeneity	0.002698
Large zone low grey level emphasis	0.002665
Joint average	0.002614
KVP	0.002597
90th discretised intensity percentile	0.002525
90th intensity percentile	0.002523
Contrast (NGTDM)	0.002511
Joint Entropy	0.002488
Spherical disproportion	0.002455

Sphericity	0.002423
Discretised intensity standard deviation	0.002377
Compactness 1	0.002372
Crazy Paving	0.002324
Area under the IVH curve	0.002259
High grey level run emphasis	0.002258
Complexity	0.002223
Short run emphasis	0.002206
Discretised intensity variance	0.002196
Large zone emphasis	0.002158
Short run high grey level emphasis	0.002158
High grey level zone emphasis.1	0.002006
Median discretised intensity	0.001935
Min value	0.001904
Quadratic mean	0.001870
Obesity	0.001766
Discretised intensity entropy	0.001680
Sex_bin	0.001662
Minimum histogram gradient intensity	0.001547
Sum variance	0.001496
Lung consolidation	0.000751
Bilateral Involvement	0.000694
History of smoking	0.000513
Hypertension	0.000493
Discretised max value	0.000000
Discretised min value	0.000000
Discretized intensity range	0.000000
Dependence count percentage	0.000000
Number of grey levels after quantization	0.000000
HRCT performed	0.000000

Table 6.1: Importances determined by RandomForest predicting death using all available features. The values are in descending order.

	RF_importances
Age (years)	0.043174
Sex_bin	0.020316
Fat.surface	0.017942
Dependence count entropy	0.017538
Dependence count energy	0.015580
Respiratory Rate	0.012740
Intensity histogram quartile coefficient of dis...	0.012675
Flatness	0.012201
Discretised interquartile range	0.012029
Small zone high grey level emphasis	0.011807
Run entropy	0.010264

Intensity histogram median absolute deviation	0.010097
Least axis length (cm)	0.009921
Muscle.surface	0.009778
Dependence count variance	0.009557
Angular second moment	0.009544
Quartile coefficient of dispersion	0.009437
Large distance high grey level emphasis	0.009423
Joint Entropy	0.009381
Low dependence high grey level emphasis	0.009338
SliceThickness	0.009026
Inverse elongation	0.008748
Intensity-based interquartile range	0.008110
Information correlation 2	0.008001
Run length variance	0.007324
Dependence count non-uniformity	0.007150
Energy	0.006935
Global intensity peak	0.006818
Normalized dependence count non-uniformity	0.006794
RECIST (cm)	0.006689
Maximum 3D diameter (cm)	0.006610
Lung consolidation	0.006506
Centre of mass shift (cm)	0.006447
Max value	0.006395
Information correlation 1	0.006361
Compactness 1	0.006337
Run percentage	0.006314
Long run emphasis	0.006191
Short run emphasis	0.006061
Zone distance non-uniformity	0.006042
Sphericity	0.005936
Normalised run length non-uniformity	0.005929
Autocorrelation	0.005853
High grey level zone emphasis.1	0.005797
Volume density - convex hull	0.005778
Integrated intensity	0.005719
Volume density - oriented bounding box	0.005678
Intensity histogram robust mean absolute deviation	0.005671
Volume at intensity fraction 90%	0.005601
Normalised homogeneity	0.005591
Inverse difference	0.005562
Local intensity peak	0.005533
Area density - oriented bounding box	0.005528
Inverse variance	0.005505
Intensity-based energy	0.005463
Crazy Paving	0.005460
Sum entropy	0.005449
Homogeneity	0.005447

Small distance low grey level emphasis	0.005413
Asphericity	0.005396
Thresholded area intensity peak (50%)	0.005375
Minimum histogram gradient	0.005369
High dependence high grey level emphasis	0.005369
Contrast (GLCM)	0.005365
Zone distance variance	0.005334
Surface to volume ratio	0.005209
Volume density - aligned bounding box	0.005162
Spherical disproportion	0.005159
Sum average	0.005132
High dependence low grey level emphasis	0.005112
Area density - convex hull	0.005111
Grey level variance (GLDZM)	0.005107
Compactness 2	0.004996
Number of voxels of positive value	0.004984
Discretised intensity uniformity	0.004979
KVP	0.004974
Cluster shade	0.004964
Thresholded area intensity peak (75%)	0.004958
Grey level non-uniformity (GLDZM)	0.004918
Normalised zone distance non-uniformity	0.004912
Number of grey levels	0.004905
Grey level non-uniformity (NGLDM)	0.004786
Dissimilarity	0.004767
Large zone high grey level emphasis	0.004756
Complexity	0.004710
Cluster prominence	0.004679
Low dependence low grey level emphasis	0.004673
Area density - aligned bounding box	0.004666
Long run high grey level emphasis	0.004659
Grey level variance (GLSZM)	0.004627
Normalised zone size non-uniformity	0.004610
Strength	0.004605
Normalised grey level non-uniformity (NGLDM)	0.004563
Difference variance	0.004546
Correlation	0.004544
CURB65	0.004524
Normalised grey level non-uniformity (GLRLM)	0.004522
Small zone emphasis	0.004512
Volume at intensity fraction 10%	0.004447
Large distance emphasis	0.004423
Minor axis length (cm)	0.004406
Zone distance entropy	0.004374
XRayTubeCurrent	0.004373
Area density - enclosing ellipsoid	0.004333
Small distance high grey level emphasis	0.004326

Contrast (NGTDM)	0.004271
Low dependence emphasis	0.004255
Short run high grey level emphasis	0.004209
Small zone low grey level emphasis	0.004182
Difference average	0.004180
Intensity range	0.004178
High grey level zone emphasis	0.004157
Intensity-based robust mean absolute deviation	0.004130
Intensity histogram coefficient of variation	0.004124
Difference entropy	0.004122
Major axis length (cm)	0.004113
Volume fraction difference between intensity fr...	0.004113
Low grey level count emphasis	0.004092
Intensity median value	0.004051
Uniformity	0.004049
Grey level non-uniformity (GLRLM)	0.004033
High grey level run emphasis	0.004027
Number of voxels	0.004015
Joint variance	0.003956
Run length non-uniformity	0.003941
Discretised intensity entropy	0.003905
Zone percentage (GLDZM)	0.003893
Large zone low grey level emphasis	0.003886
Sum variance	0.003878
Low grey level zone emphasis	0.003853
Zone percentage (GLSZM)	0.003821
High grey level count emphasis	0.003811
Busyness	0.003795
High dependence emphasis	0.003757
Grey level non-uniformity (GLSZM)	0.003727
Large distance low grey level emphasis	0.003638
Volume density - enclosing ellipsoid	0.003633
Zone size entropy	0.003558
Joint maximum	0.003558
Discretised intensity skewness	0.003552
Skewness	0.003496
Grey level variance (NGLDM)	0.003480
Area under the IVH curve	0.003470
Normalised grey level non-uniformity (GLDZM)	0.003468
Intensity histogram mean absolute deviation	0.003431
Normalised inverse difference	0.003320
Short run low grey level emphasis	0.003299
Mean discretised intensity	0.003293
Low grey level zone emphasis.1	0.003285
Discretised intensity kurtosis	0.003275
Small distance emphasis	0.003148
Joint average	0.003141

Intensity-based median absolute deviation	0.003112
90th discretised intensity percentile	0.003094
Large zone emphasis	0.003052
90th intensity percentile	0.003030
10th intensity percentile	0.003021
Low grey level run emphasis	0.002930
Kurtosis	0.002860
Cluster tendency	0.002743
Intensity at volume fraction 10%	0.002725
Grey level variance (GLRLM)	0.002673
Long run low grey level emphasis	0.002640
Intensity-based coefficient of variation	0.002634
Min value	0.002620
Intensity mean value	0.002595
Entropy	0.002562
Normalised grey level non-uniformity (GLSZM)	0.002561
Variance	0.002516
Minimum histogram gradient intensity	0.002509
Discretised intensity standard deviation	0.002474
Maximum histogram gradient intensity	0.002467
Standard deviation	0.002411
Maximum histogram gradient	0.002410
Intensity at volume fraction 90%	0.002390
Quadratic mean	0.002362
Intensity fraction difference between volume fr...	0.002215
Intensity-based mean absolute deviation	0.002203
Discretised intensity variance	0.001964
Intensity histogram mode	0.001620
Coarseness	0.001438
Median discretised intensity	0.001337
Number of compartments (GMM)	0.001055
Obesity	0.000973
History of smoking	0.000896
Bilateral Involvement	0.000893
Ground-glass	0.000859
Hypertension	0.000497
Febbre	0.000358
HRCT performed	0.000000
Discretized intensity range	0.000000
Discretised min value	0.000000
Discretised max value	0.000000
Dependence count percentage	0.000000
Number of grey levels after quantization	0.000000

Table 6.2: Importances determined by RandomForest predicting ICU Admission using all available features. The values are in descending order.

6.2 Using Dimensionality reduction to further investigate the dataset

For these analyses the data was always fed into a standard scaler before applying the technique of choice, furthermore a custom gravity score by classifying as 4 the dead individuals and then by assigning a progressive score form 1 to 3 by looking at the time of permanence was built as follows:

1. Gravity 1: Survived individuals with permanence from 0^{th} percentile to 25^{th} percentile
2. Gravity 2:Survived individuals with permanence from 25^{th} percentile to 75^{th} percentile
3. Gravity 3:Survived individuals with permanence from 75^{th} percentile to 100^{th} percentile
4. Gravity 4: Dead individuals without regard for permanence in the hospital

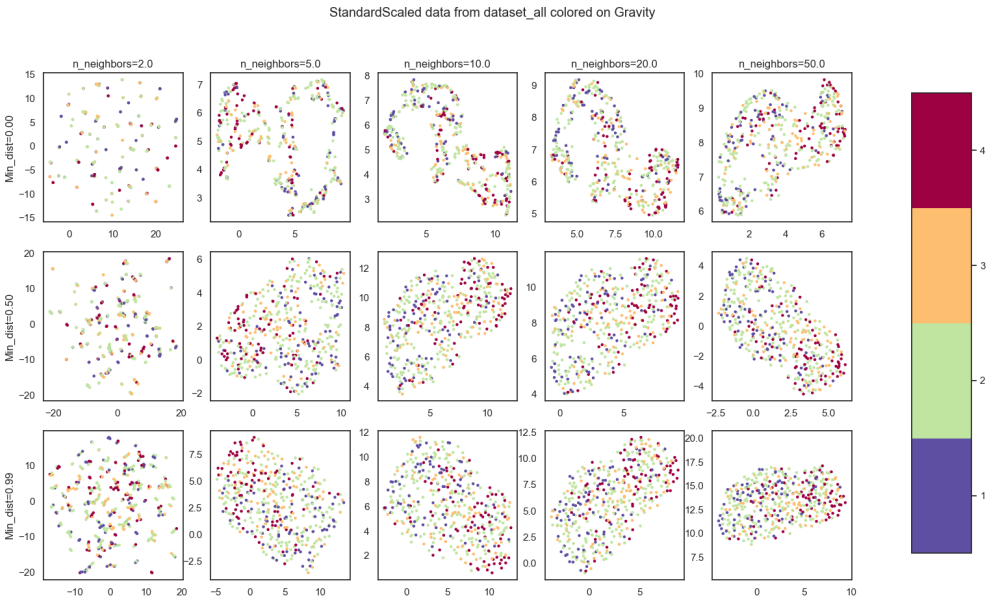


Figure 6.1: Possible combination for umap hyperparameters "number of neighbours" and "minimum distance". Color coding is done with aforementioned gravity score and all the features, i.e. clinical radiomic and radiological, were used.

Also, before proceding, the hyperparameter space for umap was explored since it's the method that allows the most control over rather intuitive parameters. Changing the value of the minimum distance of points in the final space from 0 to 0.99 changes how the structure is projected, while changing the number of neighbours changes how much the local or global structure of the data influences the final projection. Some of the combinations of these parameters can be seen in Figure 6.1

6.2.1 Explaining total variance using PCA

Starting from PCA, the data was reduced to either two or three dimensions considering clinical and radiomic features, both separated and together. In this first example there seems to be a kind of left leaning polarization of the dead individuals, however there are no clear separations in the data.

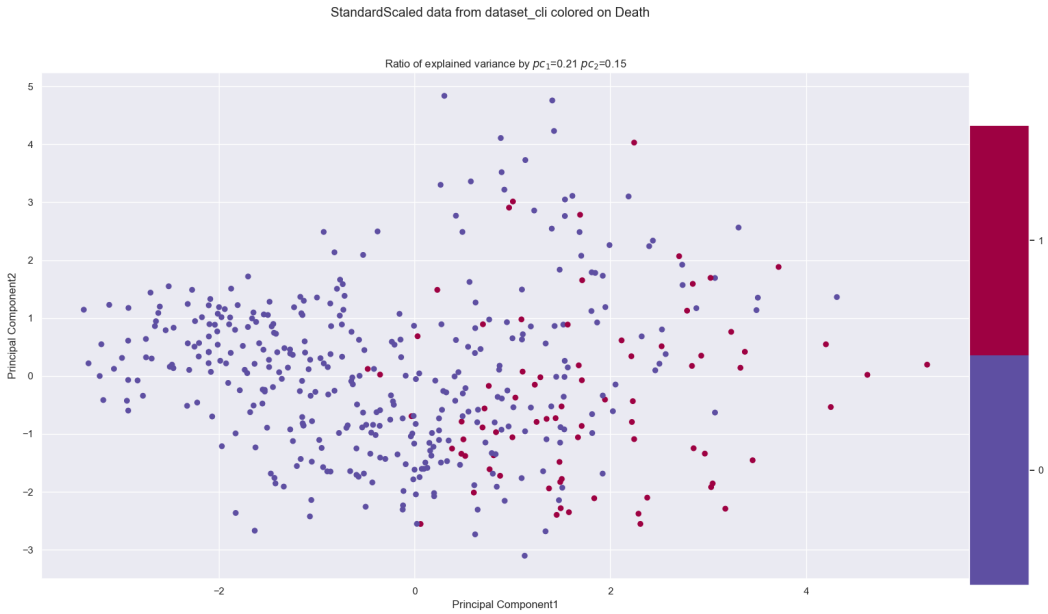


Figure 6.2: 2-Principal Component on clinical features.

Working on the clinical dataset it can also be noted that the first two components of the PCA explain only 36% of the total variance. This leads to the conclusion that changes in the data cannot be explained by a single, nor a few, features or linear combination thereof. The next approach was using the first three principal components using various labels available, most relevant of which being ICU admission, Death and Gravity score.

In all cases it seems like introducing the radiomic features causes the loss of the polarization structure that could be seen in the PCA on the clinical dataset alone in fig6.2.

Since there are no visible clusters proceeding with cluster analysis would mean incurring in the risk of finding non meaningful results so it seemed appropriate to try other dimensionality reduction techniques.

6.2.2 Exploring data structure with UMAP

The next technique tried was unsupervised Umap. Following the conclusions derived from fig:6.1 the number of neighbours was set to 10 and the minimum distance was set to 0. Once again the comparison were made between clinical and radiomic dataset as well as different possible labellings. Starting from the clinical dataset, without reporting all labels used, it's clear to see that the dataset seems to indicate very local well separated structures which don't seem correlated to gravity outcome

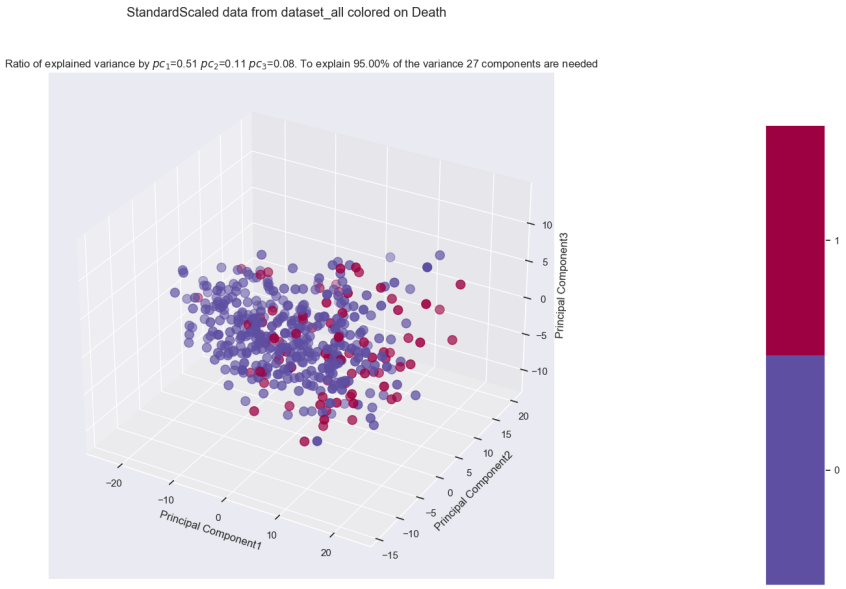


Figure 6.3: 3D PCA of whole dataset, colored with death label

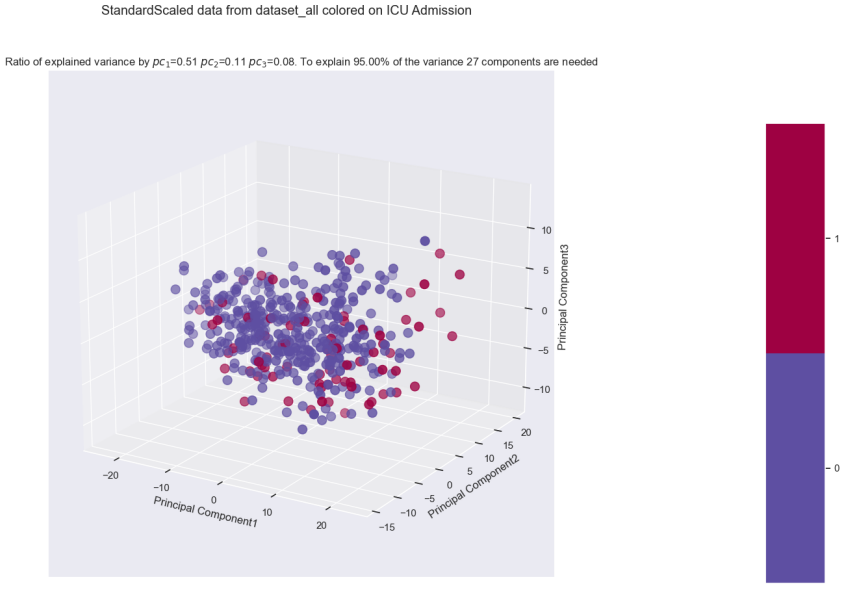


Figure 6.4: 3D PCA of whole dataset, colored with ICU Admission label

There are 9 well defined groups which don't seem to be correlated to any of the available labels. The dimension of these group is also very prohibitive if thinking of further analyses since groups of 35-50 people in a dataset with 15% mortality rate would mostly be very unbalanced if they were to be used for classification. However if the introduction of radiomic features were to unite some of these groups then this embedding could be meaningfully used for analysis. Looking at the 3D embedding for the whole dataset, the results are:

Once again the introduction of the radiomic feature seems to be a confounding factor

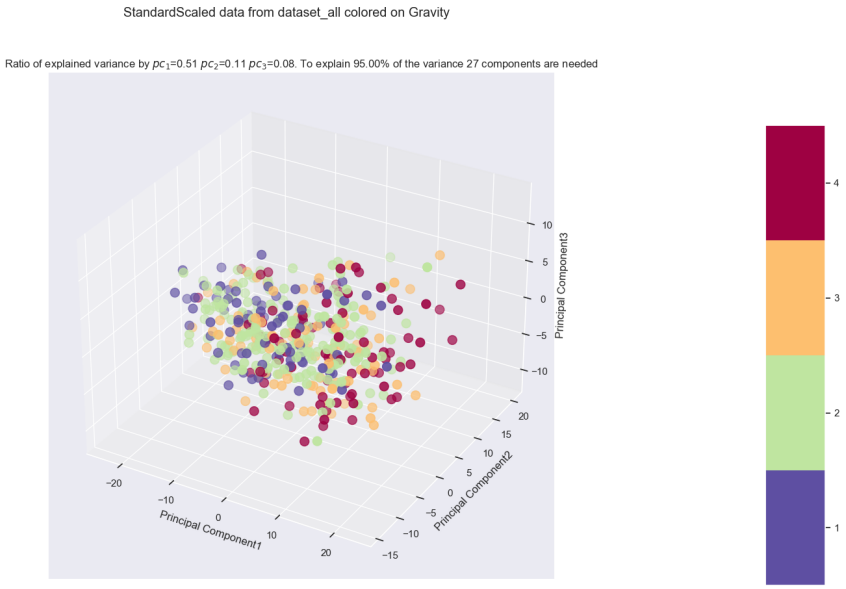


Figure 6.5: Comparison between various colour labels of the top 3 principal components for the entire dataset. Note that to explain 95% of the variance 27 components would be needed

in the seemingly clear-cut order present in the clinical dataset alone. There seems to be a well connected structure, which makes sense because umap sets out with the objective of preserving said structure. However since the variables of interest as label are Death, ICU admission or some kind of combination of them with hospital permanence there seems to be no visual correlation between structure and label. As such the next dimensionality reduction was tried to see if it yielded better results.

6.2.3 Predicting clinical outcome using PLS-DA

Moving on from unsupervised methods to a supervised one, PLS-DA was used giving as label both death and ICU using both whole dataset, and singularly radiomic or clinical features. Starting from the clinical features alone, predicting on death Figure 6.10 can be obtained.

In Figure 6.10 there are a few things to note. The first is the presence, in the top plot, of an outlier which, since PLS-DA is based on minimization of least squares, can ruin a lot the performance of the procedure. For this reason in the second plot the outlier was removed and the algorithm was run again on the cleaned data. The second thing to notice is the coloring used which, in the first plot, was used to highlight that along the one of the two latent variables the data is roughly distributed depending on age while, in the second plot, was used to highlight that the algorithm is able to perfectly separate the subjects with hypertension from those without it. However, by looking at the same embedding labelled with death and ICU admission fig 6.11 can be obtained It's clear to see that, when predicting on death, the PLS-DA algorithm doesn't find any behaviour relevant for ICU admission. It's also clear that there is at least a pattern of points labeled as dead being towards the right of the image, this can be easily explained

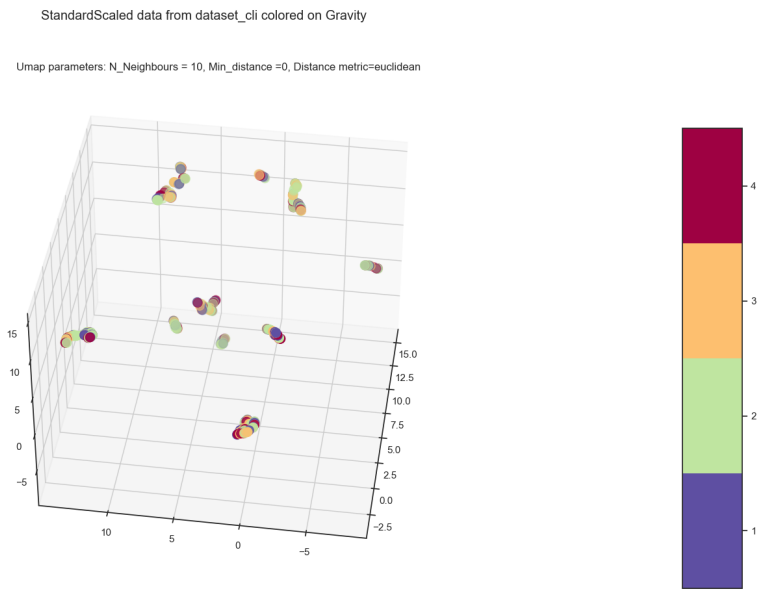


Figure 6.6: 3D umap of clinical dataset, colored based on gravity

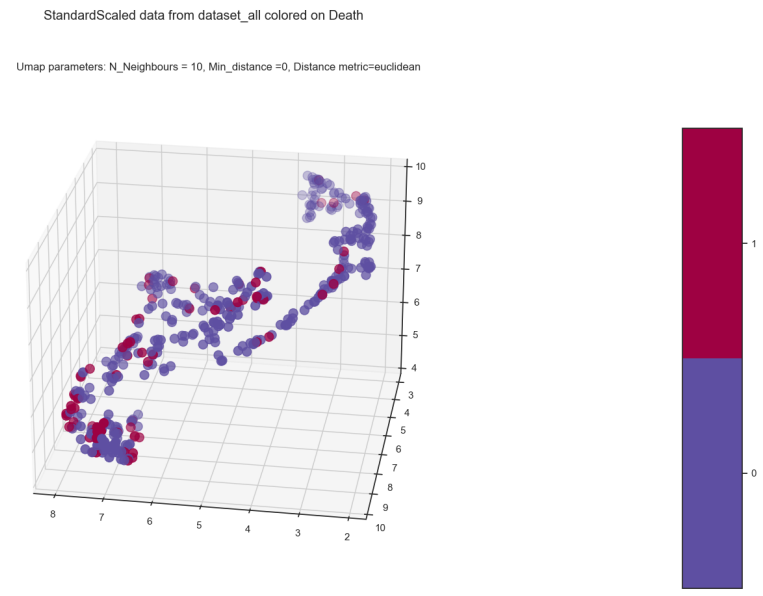


Figure 6.7: 3D umap of whole dataset, colored based on death

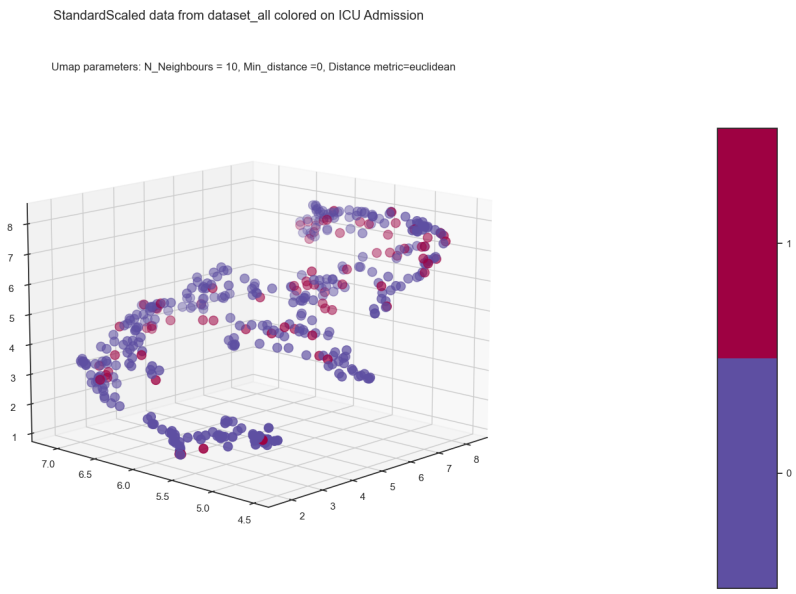


Figure 6.8: 3D umap of whole dataset, colored based on ICU admission

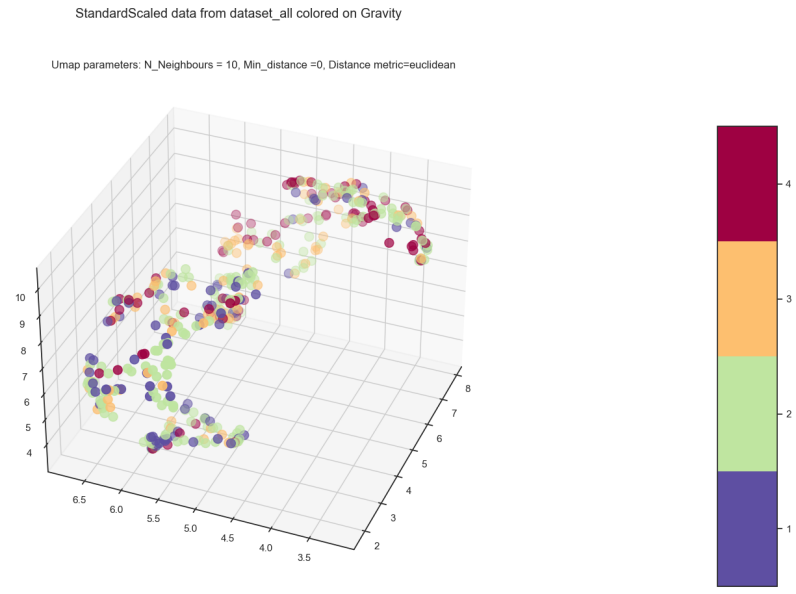


Figure 6.9: 3D umap of whole dataset, colored based on gravity

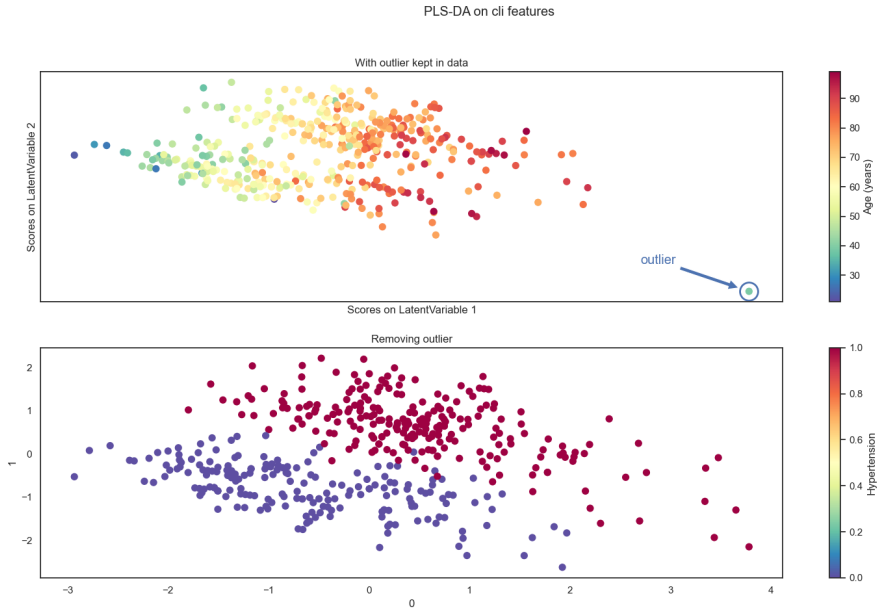


Figure 6.10: PLS-DA predicting on death coloured with age and hypertension

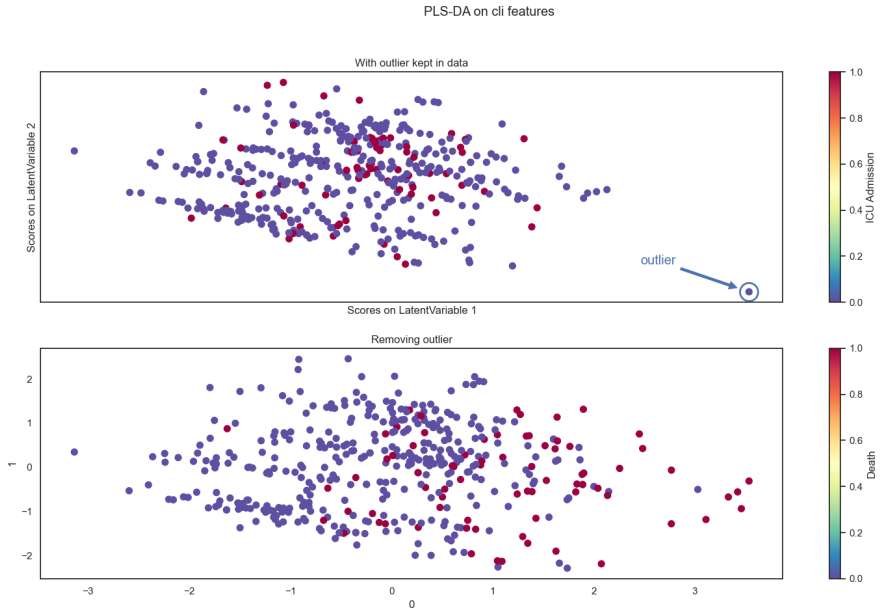


Figure 6.11: PLS-DA predicting on death coloured with death(bottom) and ICU Admission(top)

Table 6.3: PLS-DA feature weights in prediction on death using clinical features

Feature Name	Importance
Respiratory Rate	0.120206
Age (years)	0.116305
Obesity	0.004293
Hypertension	-0.004626
History of smoking	-0.012314
Febbre	-0.045431
Sex_bin	-0.054947

by looking at how the ages are distributed in the first plot of fig: 6.10. From this it's possible to deduce that older individuals tend to die more and that hypertension does not seem to be relevant when considering death as a clinical outcome. If necessary the PLS-DA algorithm allows also to see the weights given to the features in predicting the label. At least for the clinical dataset, which has a reasonable number of features, it's interesting to report it ordering the coefficients by descending absolute value: Doing the exact same procedure on the whole dataset, which means by including the radiomic features, Figure 6.12 can be obtained.

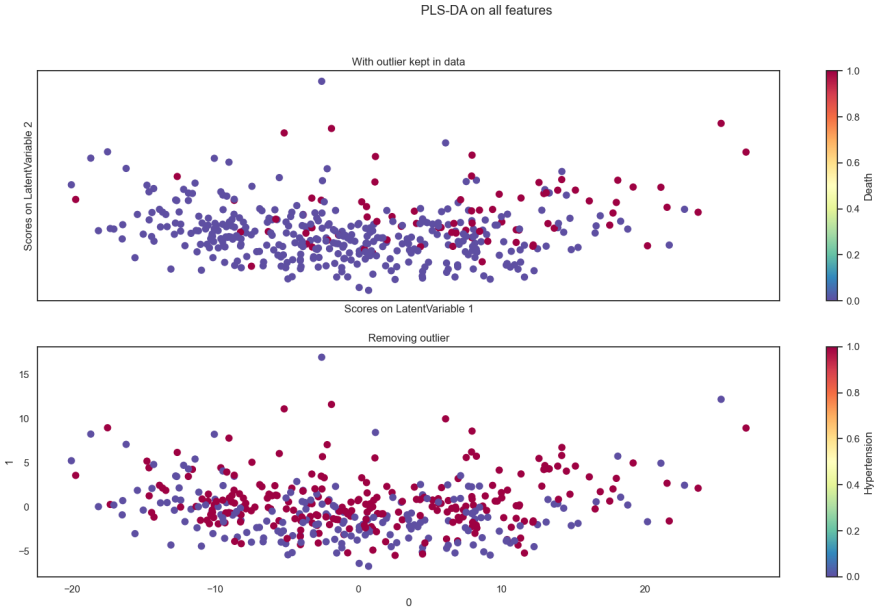


Figure 6.12: PLS-DA predicting on death coloured with death(top) and hypertension(bottom) on whole dataset

Once again adding the radiomic features has evidently introduced noise in the system, which no longer displays any kind of behaviour, pattern nor separation. Doing the same analysis but using ICU Admission as a label Figure 6.13 can be obtained. Now the colors have been chosen to highlight that respiratory rate and age have the main role in determining the latent variables. However, in this case, there doesn't appear to be a clear cut distinction as it happened before with hypertension. Looking at how the points scatter by coloring them according to the two interesting clinical

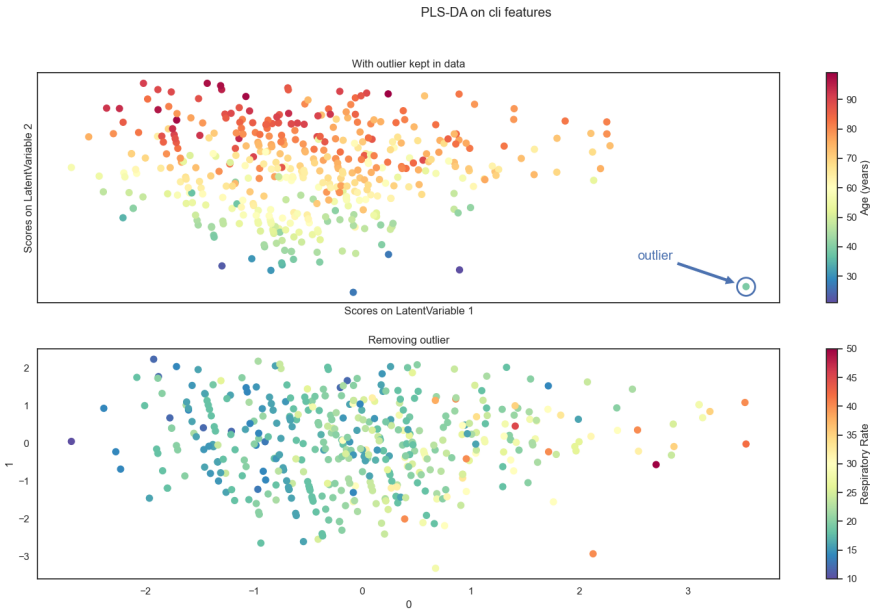


Figure 6.13: PLS-DA predicting on death coloured with Age and respiratory rate on clinical features

labels the following figure can be obtained:

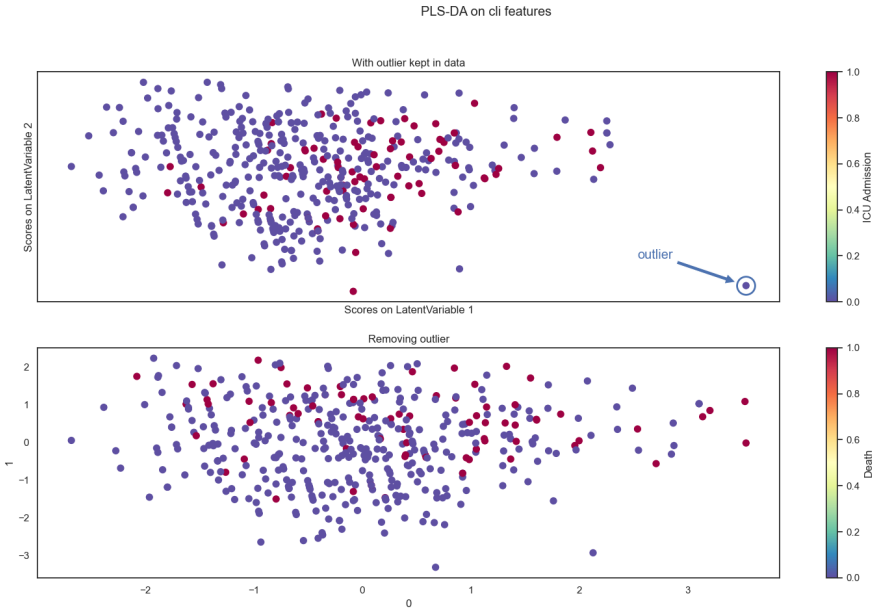


Figure 6.14: PLS-DA predicting on death(bottom) coloured with death and ICU admission(top) on clinical features

Finally, introducing the radiomic features in the analysis the usual effect of reducing separation can be seen can be seen in the figure below:

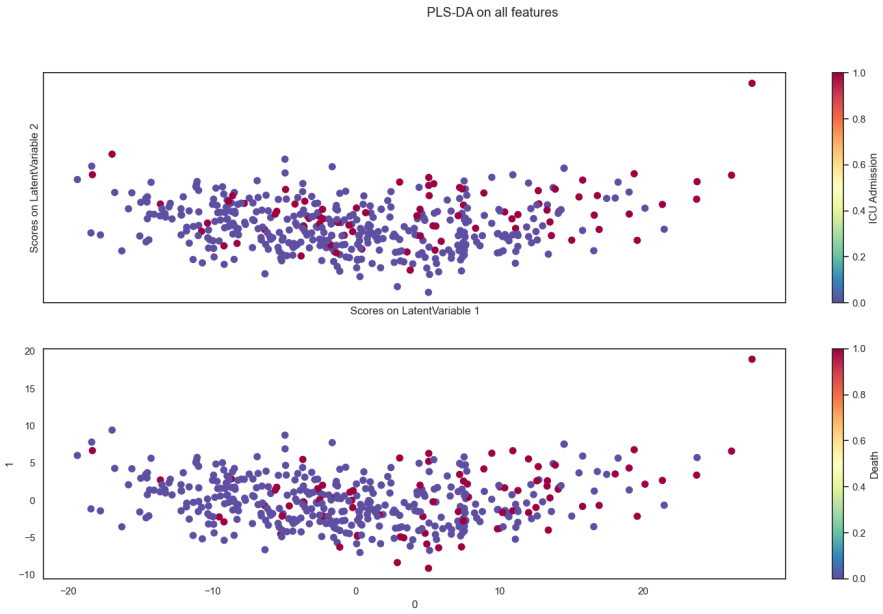


Figure 6.15: PLS-DA predicting on death coloured with death(bottom) and ICU admission(top) on all available features

Bibliography

[1] Sophia ddm for radiomics. https://www.sophiagenetics.com/technology/sophia-ddm-for-radiomics/. Accessed: 26/08/2021.

[2] Nema ps3 / iso 12052, digital imaging and communications in medicine (dicom) standard, national electrical manufacturers association, rosslyn, va, usa. available free at, 2021.

[3] J. L. V. 1, R. Moreno, J. Takala, S. Willatts, A. D. Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, and L. G. Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. on behalf of the working group on sepsis-related problems of the european society of intensive care medicine. *Intensive Care Medicine*, 1996.

[4] U. Attenberger and G. Langs. How does radiomics actually work? – review. *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, 193, 12 2020.

[5] M. Barker and W. Rayens. Partial least squares for discrimination, journal of chemometrics. *Journal of Chemometrics*, 17:166 – 173, 03 2003.

[6] A. Bettinelli, F. Marturano, M. Avanzo, E. Loi, E. Menghi, E. Mezzenga, G. Pirrone, A. Sarnelli, L. Strigari, S. Strolin, and M. Paiusco. A new benchmarking approach to assess the consistency of ibsistandardized features across radiomic tools: a multicenter study. *Radiology*, in press.

[7] R. Biondi, N. Curti, F. Coppola, E. Giampieri, G. Vara, M. Bartoletti, A. Cattabriga, M. A. Cocozza, F. Ciccarese, C. De Benedittis, L. Cercenelli, B. Bortolani, E. Marcelli, L. Pierotti, L. Strigari, P. Viale, R. Golfieri, and G. Castellani. Classification performance for covid patient prognosis from automatic ai segmentation—a single-center study. *Applied Sciences*, 11(12), 2021.

[8] R. W. Brown, Y.-C. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan. *Magnetic Resonance Imaging: Physical Principles and Sequence Design, 2nd Edition*. Wiley-Blackwell, 2014.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, Jun 2002.

[10] T. Daimon. *Box–Cox Transformation*, pages 176–178. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[11] C. Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317, 2019.

[12] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

[13] K. P. F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[14] D. G. George H. Joblove. Color spaces for computer graphics. *ACM SIGGRAPH Computer Graphics*, (12(3), 20–25), 1978.

[15] L. Guo. Clinical features predicting mortality risk in patients with viral pneumonia: The mulbsta score. *Frontiers in Microbiology*, 2019.

[16] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020.

[17] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[18] G. N. Hounsfield. Computed medical imaging. nobel lecture. *Journal of Computer Assisted Tomography*, (4(5):665-74), 1980.

[19] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[20] L. M. J. Cine computerized tomography. *The International Journal of Cardiac Imaging*, 1987.

[21] A. Kaka and M. Slaney. *Principles of Computerized Tomographic Imaging*. Society of Industrial and Applied Mathematics, 2001.

[22] J. Kirby. Mosmeddata: dataset, 09/2021.

[23] D. Kleinbaum and M. Klein. *Survival Analysis: A Self-Learning Text*. Statistics for Biology and Health. Springer New York, 2006.

[24] P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. de Jong, J. van Timmeren, S. Sanduleanu, R. T. H. M. Larue, A. J. G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F. M. Mottaghy, J. E. Wildberger, and S. Walsh. Radiomics: the bridge between medical imaging and personalized medicine. *Nature reviews. Clinical oncology*, 14(12):749—762, December 2017.

[25] L. Lee and C.-Y. Liong. Partial least squares-discriminant analysis (pls-da) for classification of high-dimensional (hd) data: a review of contemporary practice strategies and knowledge gaps. *The Analyst*, 143, 06 2018.

[26] G. Lemaître, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[27] J. McCarthy. What is artificial intelligence? 01 2004.

[28] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[29] W. McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[30] S. J. McMahon. The linear quadratic model: usage, interpretation and challenges. *Physics in medicine and biology*, 2018.

[31] S. Morozov. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *IAU Symp.*, (S227), 2020.

[32] MosMed. Mosmeddata: dataset, 28/04/2020.

[33] Y. Ohno. Cie fundamentals for color measurements. *International Conference on Digital Printing Technologies*, 01 2000.

[34] 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.

[35] D. L. Pham, C. Xu, and J. L. Prince. Current methods in medical image segmentation. *Annual Review of Biomedical Engineering*, 2(1):315–337, 2000. PMID: 11701515.

[36] P. C. Rajandeep Kaur. A review of image compression techniques. *International Journal of Computer Applications*, 2016.

[37] W. C. Roentgen. On a new kind of rays. *Science*, 1986.

[38] A. Saltelli. *Global Sensitivity Analysis. The Primer*. John Wiley and Sons, Ltd, 2007.

[39] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[40] C. P. Subbe. Validation of a modified early warning score in medical admissions. *QJM: An international journal of medicine*, 2001.

[41] L. Tommy. Gray-level invariant haralick texture features. *PLOS ONE*, 2019.

[42] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[43] M. Waskom, O. Botvinnik, D. O'Kane, P. Hobson, S. Lukauskas, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, J. de Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, A. Miles, Y. Ram, T. Yarkoni, M. L. Williams, C. Evans, C. Fitzgerald, Brian, C. Fonnesbeck, A. Lee, and A. Qalieh. mwaskom/seaborn: v0.8.1 (september 2017), Sept. 2017.

[44] S. Webb. The physics of medical imaging. 1988.

[45] L. WS, van der Eerden MM, and e. a. Laing R. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax 58(5):377-382*, 2003.

[46] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. W. L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, and et al. The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338, May 2020.