

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:
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 has been 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 have been 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 has then been evaluated through their performance in the methods used, namely Lasso regularized regression and Random Forest. Dimensionality reduction techniques have also been used to attain a better understanding of the dataset at hand.

Following a first introductory 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 achieves the same predictive power that can be derived from the clinical data. This can be interpreted in three ways: first it can be taken as a symptom of the fact even the more complex *Sars-COVID19* cases can be segmented automatically, or semi-automatically by untrained personnel, which will leading to competing results with other methodologies. 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. Finally it can be taken as proof that the method has room to improve the performances by more carefully investing in the segmentation phase.

Contents

1	Introduction	1
1.1	Digital Images and Image Processing	3
1.2	Medical Images	8
1.2.1	X-ray imaging and Computed Tomography (CT)	9
1.2.2	Generation and management of radiation: digital CT scanners	10
1.2.3	Radiation-matter interaction: Attenuation in body and measurement	15
1.3	Artificial Intelligence (AI) and Machine Learning(ML)	18
1.3.1	Regression, Classification and Penalization	19
1.3.2	Decision Trees and Random Forest	22
1.3.3	Synthetic Minority Oversampling TEchnique (SMOTE)	23
1.3.4	Dimensionality reduction and clustering	25
1.4	Combining radiological images with AI: Image segmentation and Radiomics	28
1.4.1	Image Segmentation	28
1.4.2	Radiomics	30
1.5	Survival Analysis	33
1.5.1	Kaplan-Meier(KM) curves and log-rank test	35
1.5.2	Cox Proportional-Hazard (CoxPH) model	35
2	Materials and methodologies	37
2.1	Data and objective	37
2.2	Preprocessing and data analysis	43
3	Results	49
3.1	Predicting and classifying the outcome DEATH	50
3.1.1	Feature selection through Lasso regularization and clinical outcome prediction using regression	50
3.1.2	Classification of patients using Random forests	57
3.2	Predicting and classifying the outcome ICU ADMISSION	60
3.2.1	Feature selection through Lasso regularization and clinical outcome prediction using regression	60
3.2.2	Classification of patients using Random forests	64
3.3	Using survival analysis	66
4	Discussion	70
5	Conclusion	74

6 Appendix	76
6.1 Additional Results and complete tables relative to Random Forest	76
6.2 Using Dimensionality reduction to further investigate the dataset	85
6.2.1 Explaining total variance using PCA	86
6.2.2 Exploring data structure with UMAP	86
6.2.3 Predicting clinical outcome using PLS-DA	88
Bibliography	95

¹ Chapter 1

² Introduction

³ Nowadays everybody knows of *Sars-COVID19* which, since the start of 2020, has made
⁴ necessary a few world-wide quarantines forcing everybody in self-isolation. Among the
⁵ main complications and features of this virus, symptoms gravity as well as the rate of
⁶ deterioration of the conditions are some of the most relevant and problematic.

⁷ In some cases asymptomatic or near to asymptomatic people may, in the span of a
⁸ week, get to conditions that require hospital admission. This peculiarity is also what
⁹ heavily complicates the triage process, since trying to predict with some degree of accu-
¹⁰ racy the prognosis of the patient at admission is a thoroughly complex task.

¹¹ In this thesis the aim has been to use data, specifically including data that cannot
¹² be easily interpreted by humans, to try various methods to predict a couple of clinical
¹³ outcomes, namely the death of the patient or the admission in the Intensive Care Unit
¹⁴ (ICU), while assessing their performance.

¹⁵ These analyses have been carried out on a dataset of 436 patients with different
¹⁶ variables associated to every person. A part of the variables, which has been referred
¹⁷ to as either clinical or radiological, are defined by humans and are generally discrete in
¹⁸ nature but mostly boolean. The most part of the available variables, however, have been
¹⁹ derived from images following the approaches used in the field of radiomics.

²⁰ While the utility of clinical variables, such as age, obesity and history of smoking,
²¹ is very straightforward it's interesting and helpful to understand the basis behind the
²² utility of radiomic and radiological features.

²³ Images have the ability to convey a slew of useful images, this is especially true in
²⁴ the medical field where digital images are used to inspect also the internal state of the
²⁵ patient giving far more detailed information than that obtainable by visual inspection
²⁶ at the hand of medical professionals.

²⁷ Among the ways in which *Sars-COVID19* can manifest itself the one that is most
²⁸ relevant to the scopes of this thesis is pneumonia and the complications that stem from
²⁹ it. Some of these complications, which are not specific of *Sars-COVID19* but can happen
³⁰ in any pneumonia case, display very peculiar patterns when visualizing the lungs through
³¹ CT exams.

³² These patterns are due to the pulmonary response to inflammation which may lead to
³³ thickening of the bronchial and alveolar structures up to pleural effusions and collapsed
³⁴ lungs. Without going too much in clinical detail what is of interest is how these condition
³⁵ manifest themselves in the CT exams:

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

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

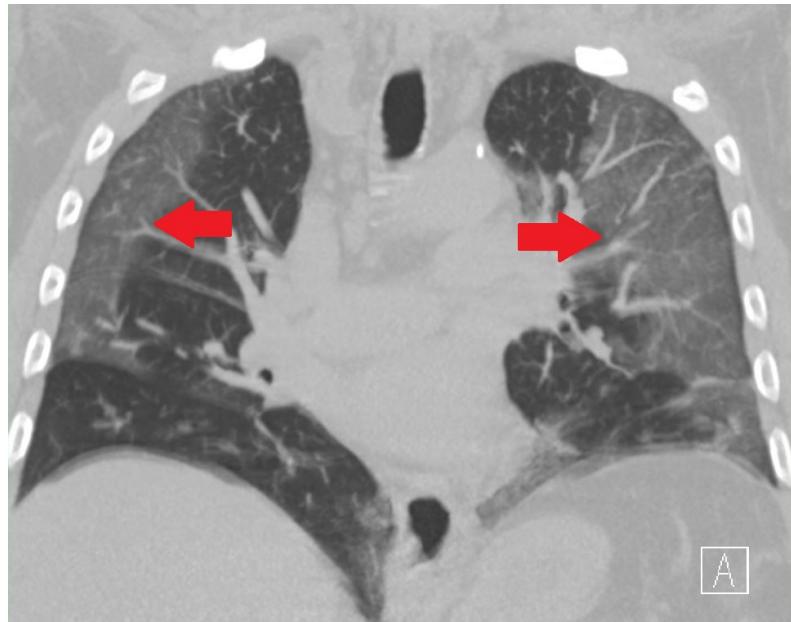


Figure 1.1: Example of GGO

38

39 **2. Lung Consolidations:**

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

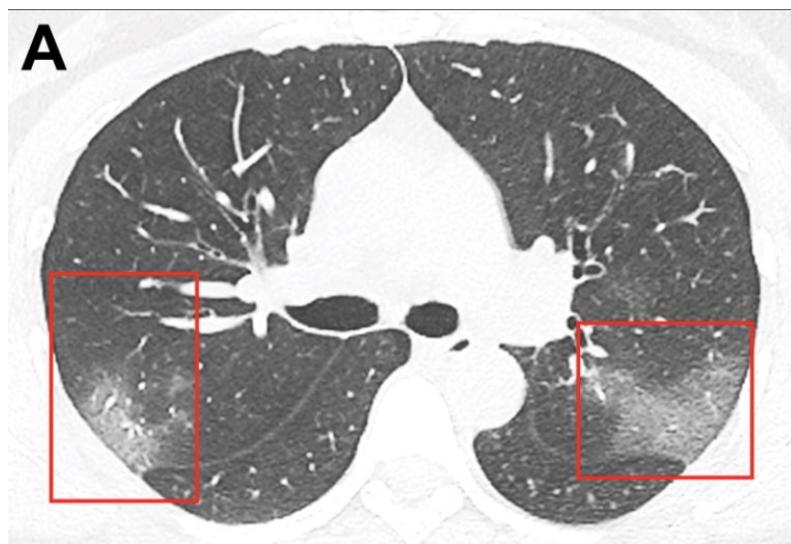


Figure 1.2: Example of consolidations

42

43 **3. Crazy paving:**

44 When GGOs are superimposed with inter-lobular and intra-lobular septal thickening.
45

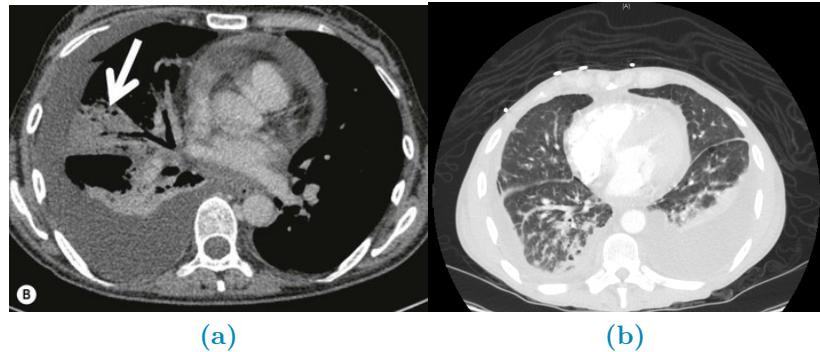


Figure 1.3: Differences between a collapsed lung (a) and pleural effusion(b)

46 **4. Collapsed Lungs and Pleural Effusion:**

47 Both of these manifest themselves as regions of the lungs that take the same coloring
 48 as that of tissue outside the lung. The main difference between the two is that
 49 collapsed lungs are somewhat rigid structures, they can occur in singular lobes of
 50 the lung and stay where they occur. Pleural effusions, however, are actually fluid
 51 being located in the lung instead of air. As such these lesions usually are located
 52 'at the bottom' of the lung in which they happen and migrate to the lowest part
 53 of the lung according to the position of the patient.

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

58 The field of radiomics, when coupled with digital images and preprocessing steps,
 59 which must include image segmentation, is exactly what undertakes this daunting task.

60 Radiomics comes from the combination of radiology and the suffix *-omics*, which
 61 is characteristic of high-throughput methods that aim to generate a large quantity of
 62 numbers, called biomarkers or features. As such it uses very precise and strict mathe-
 63 matical definitions to quantify in various ways either shape, textural or intensity based
 64 properties of the radiological image under analysis.

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

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

74 1.1 Digital Images and Image Processing

75 In this section the objective is to simply provide a set of basic definitions pertaining
 76 to images as well as a general introduction to the methods used to create said images.

77 Firstly images are a means of representing in a visual way a physical object or set thereof,
78 when talking about them it's common to refer specifically to digital images.

79 A **Digital Image** is a numerical representation of an object; more specifically an
80 ordered array of values representing the amount of radiation emitted (or reflected) by
81 the object itself.

82 The values of the array are associated to the intensity of the radiation coming from
83 the physical object; to represent the image these values need to be associated to a scale
84 and then placed on a discrete 2D grid. To store these intensities the physical image
85 is divided into regular rectangular spacings, each of which is called pixel¹, to form a
86 2D grid; inside every spacing is then stored a number (or set thereof) which measures
87 the intensity of light, or color, coming from the physical space corresponding to that
88 grid-spacing.

89 The term digital refers to the discretization process that inherently happens in storage
90 of the values, called pixel values, as well as in arranging them within the grid. It's
91 possible to generalize from 2D images to 3D volumes, simply by stacking images of the
92 same object obtained at different depths. In this context, the term pixel is substituted
93 by voxel, however since they are used interchangeably in literature they will, from now
94 on, be considered equivalent.

95 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
96 $p \in \mathbb{R}$, the type of value stored within each pixel changes the nature of the image itself.

97 A single value is to be intended as the overall intensity of light coming from the
98 part of the object contained corresponding to the gridspace and is used for a gray-scale
99 representation, a set of three² or four³ values can be intended as a color image.

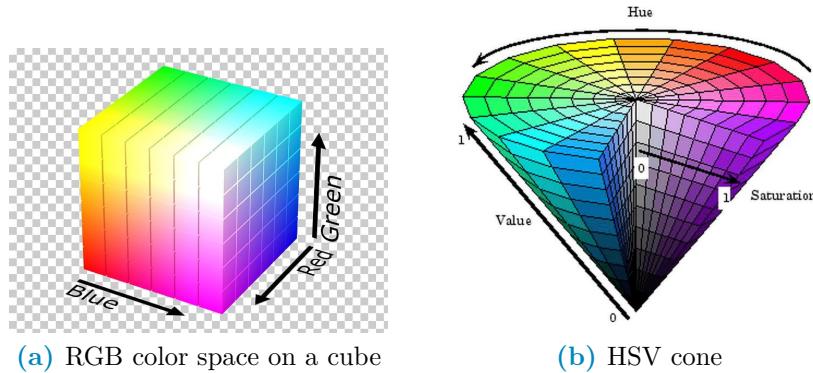


Figure 1.4: Examples of color spaces

100 There are a lot of possible scales for representation⁴, which are sometimes called
101 color-spaces, however the most noteworthy in the scope of this work is the Hounsfield

¹The 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

²The 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[39]

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

⁴Besides RGB and CMYK 1.4a the most common color spaces are CIE (Commision Internationale d'Eclairage) and HSV fig:1.4b (Hue,Saturation and Value). Refer to [15] for further details

102 unit (HU) scale which will be presented in the section relative to x-rays.

103 Digital images can be characterized by the following quantities:

- 104 • Spatial Resolution: A measure of how many pixels are in the image or, equivalently,
105 how small each pixel is; a larger resolution implies that smaller details can be seen
106 better in Figure 1.5.

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

- 110 • Color quantization: The range of the pixel values, a classic example is an 8-bit
111 resolution which yields 256 levels of gray. A better resolution allows a better
112 distinction of colors within the image in Figure 1.5.

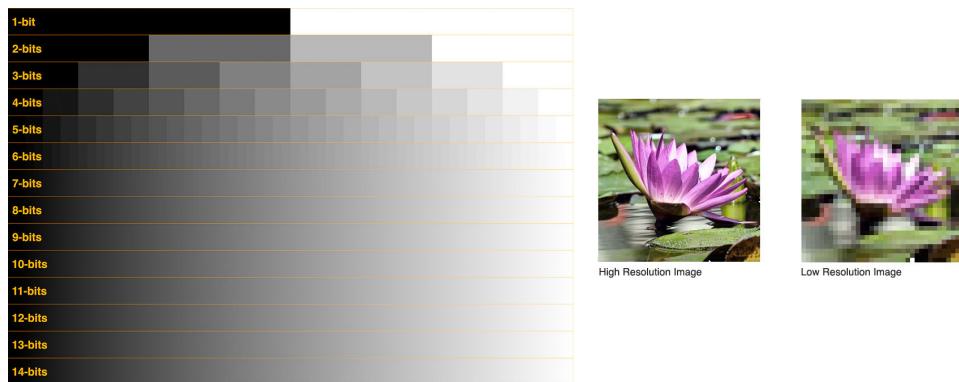


Figure 1.5: Example of visual differences in Gray-level (left) and spatial (right) resolution

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

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

119 The most commonly used formats are .PNG and .JPG, however there are a lot of
120 other formats. In the context of this work, which is going to be centered on medical
121 images, the most interesting formats are going to be the nii.gz (Nifti) and the .dcm
122 (DICOM). The first contains only the pixel value information hence it's a lighter
123 format, it originates in the field of Neuroimaging⁵, it is used mainly in Magnetic
124 resonance images of the brain but also for CT scans and, since it contains only
125 numeric information, it's the less memory consuming option out of the two.

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

⁵In fact Nifti stands for Neuroimaging Informatics Technology Initiative (NIFTI)

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

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

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

136 The first set of methods is preferred for storage of natural images, these are cases
137 in which details have no importance, whereas the second set of methods is used where
138 minute details can make a considerable difference such as in the medical field⁶.

139 Images can then be thought of as array of numbers, for this reason they are often
140 treated as matrices and, as such, there is a well defined set of valid operations and trans-
141 formations that can be performed on them. All these operations and transformations,
142 in a digital context⁷, are performed via computer algorithms which allow almost perfect
143 repeatability and massive range of possible operations.

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

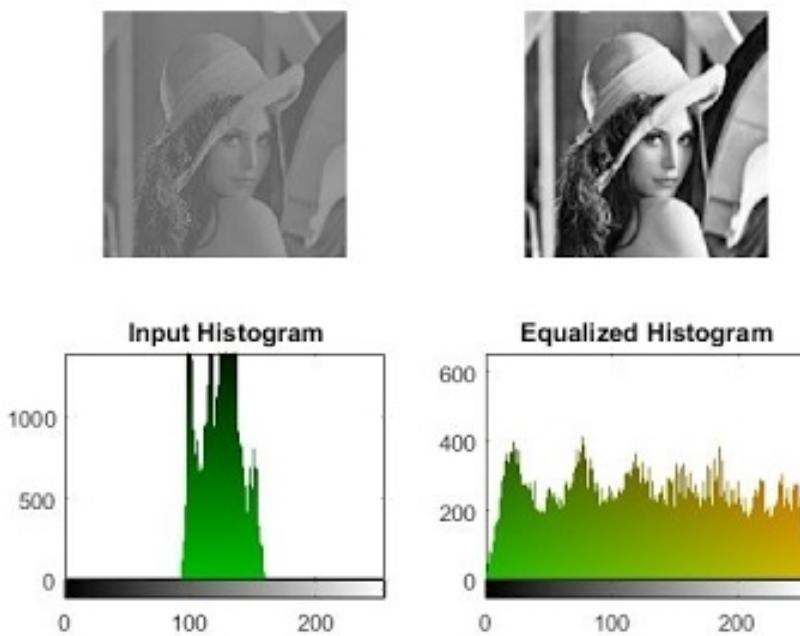


Figure 1.6: Example of differences in contrast due to histogram equalization

⁶A detailed description of compression algorithms is beyond the scopes of this thesis, for this reason please refer to [43] for more information

⁷As 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

150 This leads us to the concept of *Contrast* which is a quantification of how well different
 151 intensities can be distinguished. If all the pixel values are bundled in a small range leaving
 152 most of the histogram empty then it's difficult to pick up the differences because they
 153 are small. However, if the histogram has no preferentially populated ranges then the
 154 differences in values are being showed in the best possible way as can be seen in Figure
 155 1.6. Note also that if looking at the histogram there are two(or more) well separated
 156 distributions it's possible that these also identify different objects in the image, which
 157 will for example allow for some basic background-foreground distinction.

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

162 1. Geometric Transformations: These are transformations that involve the following
 163 steps:

164 (a) Affine transformations: Transformations that can be performed via matrix
 165 multiplication such as rotations, scaling, reflections and translations. This
 166 step basically involves computing where each original pixel will fall in the
 167 transformed image

168 (b) Interpolation: Since the coordinates of the transformed pixel might not fall
 169 exactly on the grid it might become necessary to compute a kind of average
 170 contribution of the pixel around the destination coordinate to find a most
 171 believable value. Examples of such methods are linear, nearest neighbour and
 172 bicubic.

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

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

$$q = p^\gamma \quad (1.1)$$

181 (b) Local: The output value at specific coordinates depends on a combination of
 182 the original values in a neighbourhood around that same coordinates. Some
 183 examples are all filtering operation such as edge enhancement, dilation and
 184 erosion. These filtering methods are based on performing convolutions in
 185 which the output value at each pixel is given by the sum of pixel-wise mul-
 186 tiplication between the starting matrix and a smaller (usually 3x3 or 5x5)
 187 matrix called kernel. The output image is obtained by moving the kernel
 188 along the starting matrix following a predefined stride for example a (2,2)

⁸For 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

stride will move the kernel 2 pixels to the right and 2 down. When moving near the borders the behaviour 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 \mathbf{S} , kernel shape \mathbf{K} ⁹ and padding \mathbf{P} determine the shape of the output matrix given the input dimension \mathbf{W} via the following formula:

$$OutputShape = \left\lceil \frac{W - K + P}{S} \right\rceil + 1 \quad (1.2)$$

- (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 its 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.

1.2 Medical Images

Having seen what constitutes an 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 are deep in the patient's 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 its clear advantage 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 and 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 than 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.

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

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

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

234 The positrons come from the β^+ decay of a radio-nucleide bound to a macro-
235 molecule, which is preferentially absorbed by the site of interest ¹⁰. Once the an-
236 nihilation happens a pair of (almost) co-linear photons having (almost) the same
237 energy of 511 keV is emitted, the detection of this pair is what allows the re-
238 construction of the image representing the pharmaceutical distribution within the
239 body. The exam is primarily used in oncology given the greater energy consump-
240 tion, hence nutrients absorption, of cancerous tissue and secondly this technique
241 can be combined with CT scans to obtain a more detailed representation of the
242 internal environment of the patient

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

246 1.2.1 X-ray imaging and Computed Tomography (CT)

247 It's well known that the term x-rays is used to characterize a family of electromagnetic
248 radiation defined by their high energy and penetrative properties. Radiation of this kind
249 is created in various processes such as characteristic emission of atoms, also referred to
250 as x-ray fluorescence, and Bremsstrahlung, braking radiation¹¹.

251 The discovery that "A new kind of ray"^[45] with such properties existed was carried
252 out by W. C. Roentgen in 1895, which allowed him to win the first Nobel prize in physics
253 in the same year. Clearly the first imaging techniques that involved this radiation were
254 much simpler than their modern counterpart, first of all they were planar and analog in
255 nature, as well as not as refined in image quality. The first CT image was obtained in
256 1968 in Atkinson Morley's Hospital in Wimbledon.

257 Tomography indicates a set of techniques¹² that originate as an advancement of planar
258 x-ray imaging; these techniques share most of the physical principles with planar imaging
259 while overcoming some of its major limitations, main of which being the lack of depth
260 information. X-ray imaging, both planar and tomographic, involves seeing how a beam
261 of photons changes after traversing a target, the process amounts to a kind of average of
262 all the effects occurred over the whole depth travelled.

¹⁰Most commonly Fluoro-DeoxyGlucose FDG which is a glucose molecule labelled with a ¹⁸F 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 \div Bq (read Becquerel) and half-life \div $T_{\frac{1}{2}}$ which is how long it takes for half of the active atoms to decay

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

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

263 The way in which slices are obtained is called focal plane tomography and, as the
264 name suggests, the basic idea is to focus in the image only the desired depth leaving the
265 unwanted regions out of focus. This selective focusing can be obtained either by taking
266 geometrical precautions while using analog detectors, such as screen-film cassettes, or by
267 feeding the digital images to reconstruction algorithms to perform digitally the required
268 operations¹³.

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

271 First x-rays are somehow generated by the machine, the quality of these x-rays is
272 optimized with the use of filters then focused and positioned such that they mostly hit
273 the region that needs imaging. The beam then exits the machine and starts interacting
274 with the imaged object¹⁴, this process causes an attenuation in the beam which depends
275 on the materials composing the object itself. Having then travelled across the whole
276 object it interacts with a sensor, be it film, semiconductor or other, which stores the
277 data that will then constitute the final image. In a digital setting this final step has to
278 be performed following a (tomographic) reconstruction algorithm which given a set of
279 2D projections returns a single 3D image.

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

- 284 • How these x-ray imaging machines are structured
- 285 • How x-ray and matter interact as the first traverses the second

286 For more information on reconstruction algorithms refer to [25] and [56].

287 1.2.2 Generation and management of radiation: dig- 288 ital CT scanners

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

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

¹⁴In 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

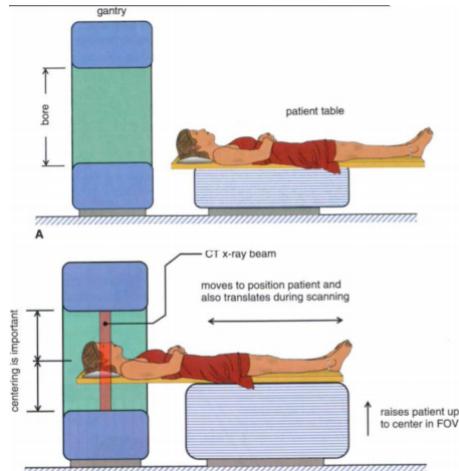


Figure 1.7: General set-up of a CT machine

294 The beam is generally created by the interaction of high energy particles with some
 295 kind of material, so that the particle's kinetic energy can be converted into radiation.
 296 In practice this means that an x-ray tube is encapsulated in the machine. Inside this
 297 vacuum tube charged particles¹⁵ are emitted from the cathode, accelerated by a voltage
 298 differential and shot onto a solid anode¹⁶. This creation process implies that the spec-
 299 trum of the produced x-rays is composed of the almost discrete peaks of characteristic
 300 emission, due to the atoms composing the target, superimposed with the continuum
 301 Bremsstrahlung radiation.

¹⁵Most commonly electrons

¹⁶Typical materials can be Tungsten, Molybdenum

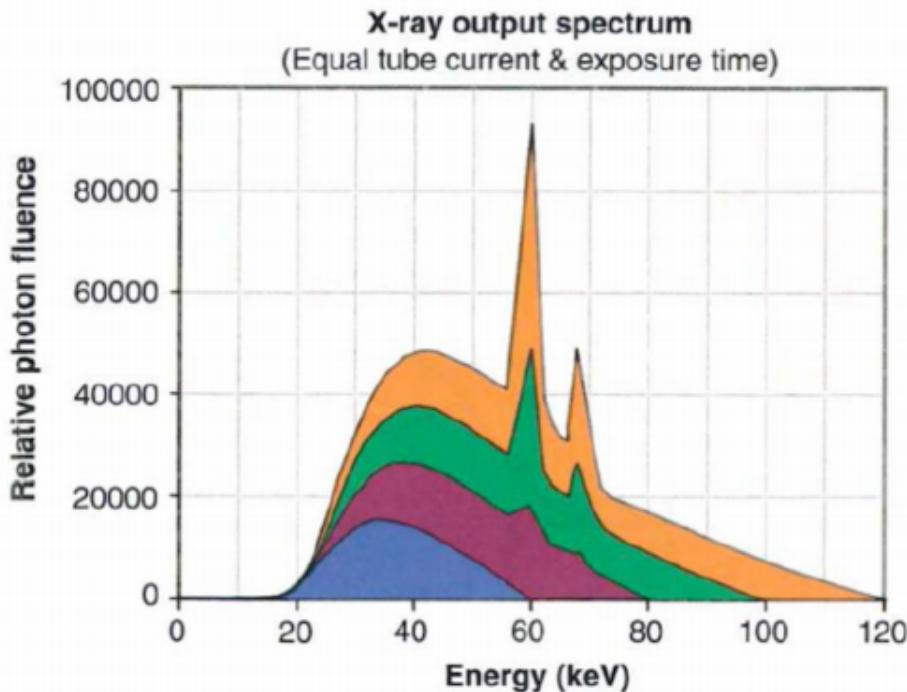


Figure 1.8: X-ray spectrum, composed of characteristic peaks and Bremsstrahlung continuum, computed at various tube voltages

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 *kilo Volt 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 exposure¹⁷.

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.

¹⁷Exposure 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

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

- 323 • 1st generation-Pencil Beam: A single beam is shot onto a single sensor, both sensor
 324 and beam are translated across the body of the patient and then rotated of some
 325 angle. The process is repeated for various angles. Main advantages are scattering
 326 rejection and no need for relative calibration, main disadvantage is time of the
 327 exam
- 328 • 2nd generation-fan Beam: Following the same process as the previous generation
 329 the main advantage is the reduction of the time of acquisition by introducing N
 330 beam and N sensors which don't wholly cover the patient's body so still need to
 331 translate.

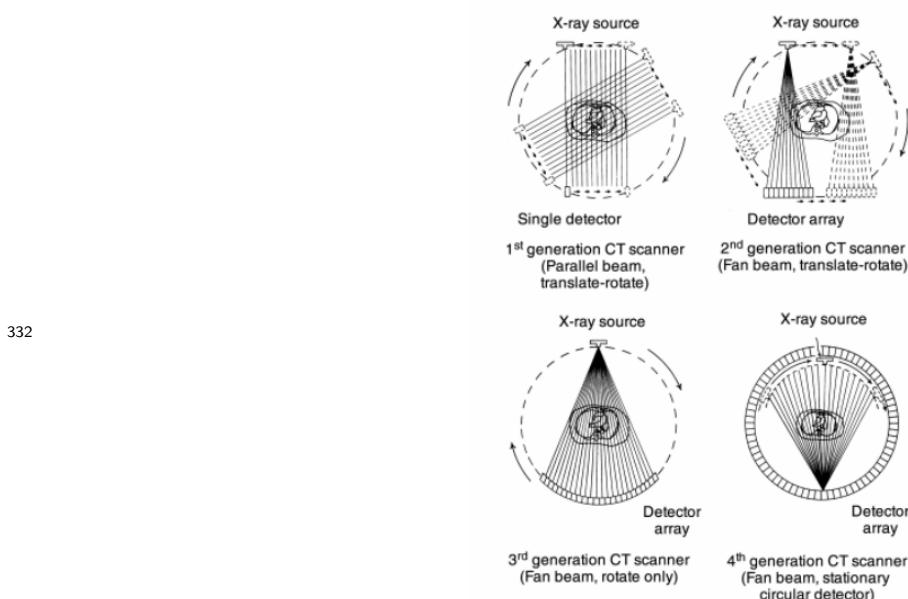


Figure 1.9: First four generations of CT scanners

- 332 • 3rd generation-Rotate Rotate Geometry: Enlarging the span of the fan of beams
 333 and using a curved array of sensor a single emission of the N beams engulfs the
 334 whole body so the only motion necessary is rotation of the couple beam-sensor
 335 array around the patient.
- 336 • 4th generation-Rotate Stationary Geometry: The sensors are now built to com-
 337 pletely be around the patient so that only the beam generator has to rotate around
 338 the body
- 339 • 5th generation-Stationary Stationary Geometry: The x-ray tube is now a large circle
 340 that is completely around the patient. This is only used in cardiac tomography for
 341 more information refer to [24]
- 342 • 6th generation-Spiral CT: Supposing the patient is laying parallel to the axis of
 343 rotation, all previous generations acquired, along the height of the patient, a single
 344 slice at a time. In this generation as the tube rotates around the patients the bed
 345

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

350 It's necessary to add another important parameter which is the pitch of the de-
351 tector¹⁸. This quantifies how much the bed moves along the axis at each turn the
352 tube makes around the patient. Pitches smaller than one indicate oversampling
353 at the cost of longer acquisition times, pitches greater than one indicate shorter
354 acquisition times at the expense of a sparser depth resolution.

355

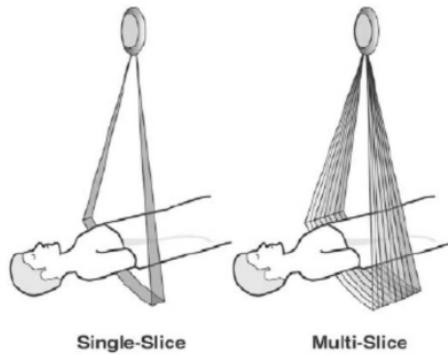


Figure 1.10: 7th generation setup

- 356 • 7th generation-MultiSlice: Up to this seventh generation height-wise slice acquisi-
357 tion was of a singular plane, be it continuous or in a start and stop motion. In this fi-
358 nal generation mutliple slices are acquired. Considering cilindrical coordinates with
359 z along the axis of the machine the multiple slice acquisition is obtained by pairing a
360 fanning out along θ and one along z of both sensor arrays and beam. This technique
361 returns to a start and stop technology in which only $\sim 50\%$ of the total scan time is

362

used for acquisition

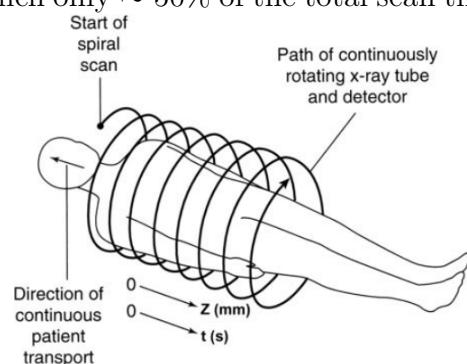


Figure 1.11: 7th generation setup

363 The machines used to obtain the images used in this thesis, all belonging to *IRCSS*
364 *Azienda ospedaliero-universitaria di Bologna - Policlinico Sant'Orsola-Malpighi* , were
365 distributed as shown in Figure 1.12:

¹⁸Once 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 1.12: 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.

- 366 1. Ingenuity CT (Philips Medical Systems Cleveland): $\sim 56\%$ of the exams were
- 367 obtained with this machine
- 368 2. Lightspeed VCT (General Electric Healthcare, Chicago-Illinois): $\sim 41\%$ of the
- 369 exams in study come from this machine
- 370 3. ICT SP (Philips Medical Systems Cleveland): $\sim 3\%$ of the exams were performed
- 371 with this machine

372 1.2.3 Radiation-matter interaction: Attenuation in 373 body and measurement

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

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

$$N(x) = N_0 e^{-\mu(E)*x} \quad (1.3)$$

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

$$\mu(E) = \frac{\rho * N_A}{A} * (\sigma_{Photoelectric}(E) + \sigma_{Compton}(E) + \sigma_{PairProduction}(E)) \quad (1.4)$$

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

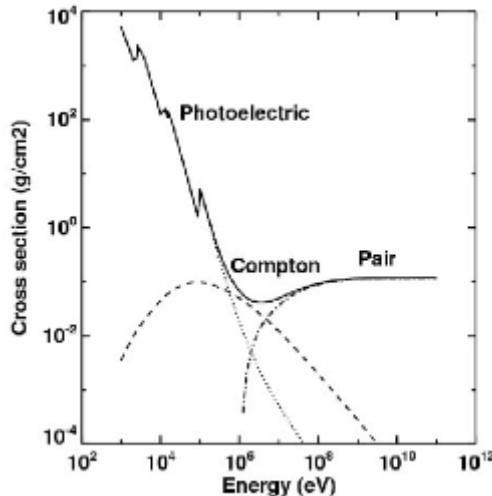
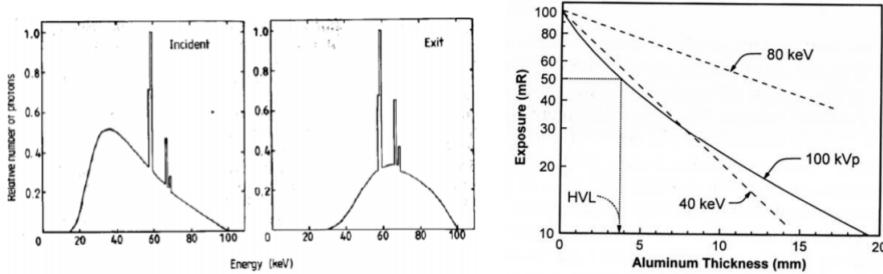


Figure 1.13: Photon cross-section in Pb

Overall, given eq:1.3, the attenuation behaviour of a monochromatic beam is expected to be linear in shape when plotted 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 Figure 1.14a:



(a) Difference in unfiltered(left) and filtered(right) relative number of photons
(b) Polychromatc beam attenuation

Figure 1.14: Polychromatc beam behaviour. The attenuation behaviour of a polychromatc beam displays, as shown in (b), a slight curvature which is in contrast to a purely linear behaviour in case of a monochromatc beam

Another effect of the beam being polychromatc is that the graphical behaviour of the curve describing the attenuation, instead of being linear, gets bent as shown in Figure 1.14b. 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:

$$C(p_1, p_2) = \frac{N_{\gamma, p_2} - N_{\gamma, p_1}}{N_{\gamma, p_2}} \quad (1.5)$$

This formula, that connects the beam to the image, together with eq: 1.3, 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.

The scale used for pixel values is the Hounsfield scale which is a scale used specifically to describe radiodensity. The values are obtained as a transformation of the linear attenuation coefficient 1.4 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 [21].

$$HU = 1000 * \frac{\mu - \mu_{H_2O}}{\mu_{H_2O} - \mu_{Air}} \quad (1.6)$$

The utility of this scale is in its 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.

In any case the tissue that is traversed by photons undergoes a process that produces some damage, which 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

425 of surviving cells and, to do so, various models are used. In the clinical practice it's
426 common to find, still within the DICOM image metadata, the information regarding Dose
427 delivered such as CTDI (Computed Tomography Dose Index) from which it's possible
428 to obtain the DLP (Dose Length Product) taking into consideration the total length of
429 irradiated body. For an introduction to one of these models, the Linear Quadratic (LQ)
430 refer to [36].

431 1.3 Artificial Intelligence (AI) and Machine Learn- 432 ing(ML)

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

436 Starting from the definition given by John McCarthy in [32] "[AI] is the science and
437 engineering of making intelligent machines, especially intelligent computer programs. It
438 is related to the similar task of using computers to understand human intelligence, but
439 AI does not have to confine itself to methods that are biologically observable.". Ma-
440 chine Learning (ML) is a sub-branch of AI and contains all techniques that make the
441 computer improve performances via experience in the form of exposure to data. Practi-
442 cally speaking this finds it's application in classification problems, image/speech/pattern
443 recognition, clustering, autoencoding and others.

444 The general workflow of Machine Learning is the following: given a dataset, the
445 objective is to define a model or function which depends on some parameters which is
446 able to manipulate the data in order to obtain as output something that can be evaluated
447 via a predefined performance metric. The parameters of the model are then automatically
448 adjusted in steps to minimize or maximize this performance metric until a stable point
449 at which the model with the current parameters is considered finalized; one of the main
450 problems in this procedure is being sure that the stable point found is global and not
451 local. The whole procedure is carried out keeping in mind that the resulting model needs
452 to be able to generalize it's performance on data that it has never seen before, for this
453 reason usually ML is divided in a training phase and a testing phase.

454 The training phase involves looking at the data and improving the performance of the
455 model on a specific dataset¹⁹, the testing phase involves using brand new data to eval-
456 uate the performance of the model obtained in the preceding phase. Machine Learning
457 techniques can be further grouped into the following categories:

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

¹⁹Usually 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

- 464 2. Unsupervised Learning: As the name suggests this category of models trains on the
 465 data alone, without having the labels available by minimizing some metric defined
 466 from the data. For example clustering techniques try to find a set of groups in
 467 the data such that the difference within each group is minimal while the difference
 468 among groups is maximal, ideally producing dense groups, called clusters, that
 469 are each well separated from all the others. Other techniques in this family are
 470 Principal Component Analysis (PCA) and autoencoding but the general objective
 471 is to infer some kind of structure within the data and the relation between data
 472 points.
- 473 3. Reinforcement Learning: This kind of ML is well suited for data which has a
 474 clear sequential structure in which the required task is to develop good long term
 475 planning. Broadly speaking the general set-up is that given a set of (state_t , action,
 476 reward, state_{t+1}) these techniques try to maximize the cumulative reward²⁰. The
 477 main applications of these techniques are in Autonomous driving and learning how
 478 to play games

479 The following methods are those directly involved in this work.

480 1.3.1 Regression, Classification and Penalization

481 Regression and Classification are methods used to make predictions via supervised learn-
 482 ing by understanding the input-output relation in continuous and discrete cases respec-
 483 tively.

484 The most basic example of regression is linear regression which consists in finding the
 485 slope and intercept of a line passing through a set of points. A branch of classification
 486 that's similar to linear regression is logistic regression which consist of looking at data
 487 and predicting it's belonging to one category within a set of possible categories, in the
 488 case of two classes this corresponds to having a 0-1 boolean output.

489 More specifically logistic regression translates in finding out whether or not each point
 490 belongs to a certain category given it's properties and can be practically thought of, for
 491 a binary classification, as a fitting procedure such that the output can be either 0 for one
 492 category and 1 for the other, this is usually done using the logistic function, also called
 493 sigmoid, which compresses \mathbb{R} in $[0, L]$ as seen in 1.15. L represents the maximum value
 494 desired, x_0 is the midpoint and k is the steepness.

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

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

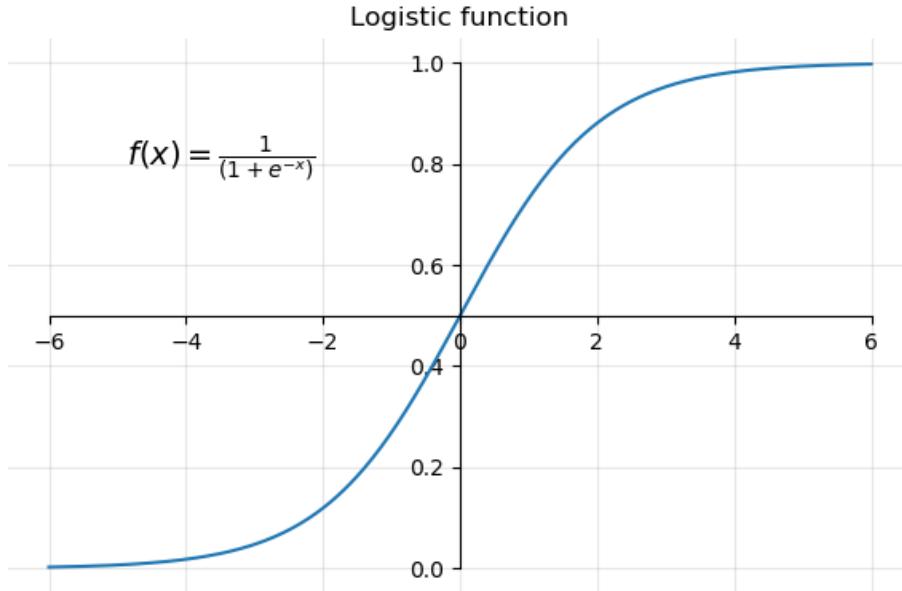


Figure 1.15: sigmoid function with L=1, $x_0=0$, k=1

495 To give a general intuition of the procedure, without going too much in depth, let's
 496 only consider linear regression redirecting to [18] for more details; the theory on which
 497 it is founded is based on the assumption that the residuals, i.e. the distance model-
 498 label, are normally distributed. Under this assumption the parameters can be found by
 499 changing them and trying to minimize the sum of squared residuals.

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

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

$$y_i = \sum_{j=0}^{n+1} x_{i,j} * \theta_i + \epsilon_i \Rightarrow \mathbf{Y} = \mathbf{X} * \theta + \epsilon \quad (1.8)$$

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

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

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

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

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

512 It's evident that to obtain a result from the previous operation it's necessary that
513 $(\mathbf{X}^T * \mathbf{X})$ be invertible, which in turn requires that there be no correlated features and
514 that the features be less than the datapoints.

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

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

519 The nature of this small quantity changes the properties of the regularization proce-
520 dure, the most famous penalties are Lasso [51], Ridge [20] and ElasticNet [61].

521 In practical terms the shape of the penalty determines how much and how fast the
522 slopes relative to the features can be shrunk.

- 523 1. Ridge: Adds $\delta^2 * \sum_{j=1}^{n+1} \theta_j^2$, is called also L^2 regularization since it adds the L^2 norm
524 of the parameter vector. This penalty can only shrink parameters asymptotically
525 to zero but never exactly, which means that all features will always be used, even
526 with very small contributions
- 527 2. Lasso: Adds $\frac{1}{b} * \sum_{j=1}^{n+1} |\theta_j|$, is called also L^1 regularization since it adds the L^1 norm of
528 the parameter vector. This penalty can shrink parameters to exactly zero, getting
529 rid of the useless variables within the model.
- 530 3. ElasticNet: Adds $\lambda * [\frac{1-\alpha}{2} * \sum_{j=1}^{n+1} \theta_j^2 + \alpha * \sum_{j=1}^{n+1} |\theta_j|]$, evidently this is a midway
531 between the Lasso and Ridge methods, where the balance is dictated by the value
532 of α .

533 There are no clear overall advantages in the choice between Ridge and Lasso regular-
534 ization, however there are substantial differences that can aid in the choice.

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

540 Lasso regularization has the ability to shrink parameters exactly to zero and does
541 so in cases in which variables are correlated with one another. However the choice on
542 which feature to keep is random if the variables are highly correlated. The advantages
543 of Lasso, namely the feature selection it produces while also improving the prediction on
544 the data, come at the cost of difficult to interpret results in some cases as well as a limit
545 in the maximum number of feature that can survive the procedure.²². This means that
546 practically one would choose Lasso regression in cases in which the expectation is that
547 only a few variables drive the behaviour of the response.

548 Elastic net [61], which was born from a critique and improvement upon Lasso regu-
549 larization, is a method that combines Lasso and Ridge. In this sense the model will still
550 be simplified as it happened in Ridge and Lasso. The surviving parameters will be less
551 than those estimated by Ridge and more than those obtained with a Lasso and the value
552 of the coefficients will be smaller than those in Lasso but larger than those obtained in
553 the case of Ridge.

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

554 For more in depth information and contextualization of all three refer to [18].

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

558 In regression the whole foundation is the normality of the residuals it's important
559 that the data behaves somewhat nicely in this regard. It has also been found empirically
560 that regression procedures have problems dealing with asymmetric distributions in the
561 data as well as with outliers in the dataset.

562 Changing the data to modify the residuals is not straightforward, however it's possible
563 to change and improve the shape of the distribution at hand to make them as close to
564 normal as possible. To this end some of the operations that can be done are:

- 565 1. Standard Scaling: This amounts to subtracting the mean of the distribution to the
566 feature and dividing by the standard deviation so that the resulting distribution is
567 somewhat centered around zero and has close to unitary standard deviation
- 568 2. Boxcox [11] transform: When distributions are asymmetric, like a gamma distribu-
569 tion would be, this transform is used to find the optimal power-law that turns the
570 data into a distribution that most closely resembles a normal distribution. More
571 specifically the data is transformed according to:

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

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

574 Being that the exponent can be positive or negative this transform cannot handle
575 distributions with negative values.

576 Practical application of these procedures can be found in Chapter 2. These consider-
577 ations conclude the theoretical background on regression, classification and penalization
578 thereof.

579 1.3.2 Decision Trees and Random Forest

580 Apart from logistic and multinomial regression there are various other supervised classi-
581 fying methods such as Support Vector Machines (SVM) [10], Neural Networks[33] and
582 Decision Trees[?]. In this thesis, among the aforementioned algorithms, the chosen one
583 was a particular evolution of DecisionTrees called RandomForest (RF)[19].

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

586 Since it's a classification method let's consider the simple case of binary classification
587 with categorical²³ features.

588 The task of the decision tree is to approximate to the best of it's abilities the labels
589 contained in the training set using all the features associated with each datapoint, this is

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

590 done by building a graph-like structure in which the nodes represent the features and the
 591 links departing from them are the possible values the feature takes. This graph is built in
 592 a top-down approach by choosing at every step the feature that best separates the data
 593 in the label categories, this process is done along each branch until a node in which the
 594 separation of the preceding feature is better than that provided by all remaining features
 595 or all features have been considered.

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

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

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

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

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

610 The strength of this method is it's performance on data it has seen however it has
 611 very poor generalization abilities [18].

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

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

623 1.3.3 Synthetic Minority Oversampling TEchnique 624 (SMOTE)

625 In the context of this thesis it will become necessary to take care of balancing the input
 626 dataset, to do so one could randomly oversample, by duplicating instances in the minority
 627 class, or undersample, by removing instances within the majority class. The choice

628 that was made was to use Synthetic Minority Oversampling TEchnique (SMOTE)[9] to
629 rebalance the dataset.

630 This technique considers a user-defined number of nearest neighbours of randomly
631 chosen points in the minority class and populates the feature space by generating samples
632 on the lines that connect the chosen sample with a random neighbour²⁴.

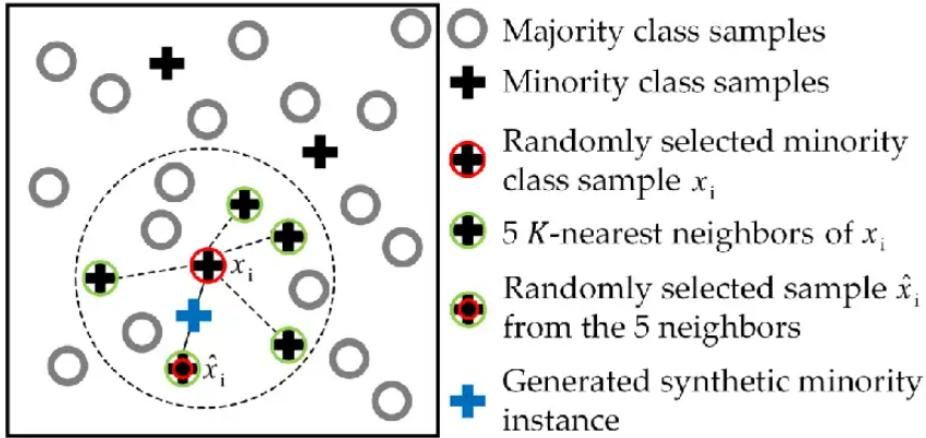


Figure 1.16: Example of SMOTE with 5 nearest neighbours

633 Worth noting that this has been done using the library imblearn in python [30].
634 To preview some of the data, looking at the performance of a vanilla random forest
635 implementation with all preset parameters, the effect of the position of oversamplling is
636 the following.

²⁴This 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

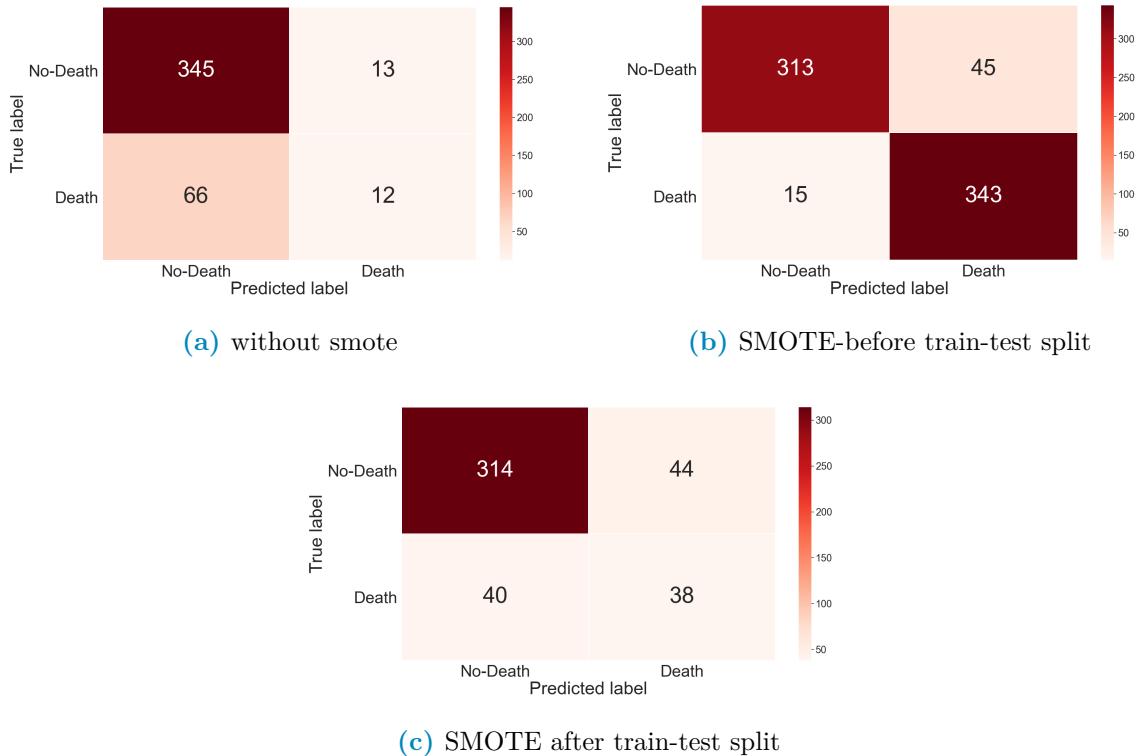


Figure 1.17: Confusion matrix used to evaluate performance done using datasets with various combinations of SMOTE position

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

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

645 What's being observed is a leakage phenomenon, which consists in the testing data
 646 containing information regarding the training data and viceversa. In practice, especially
 647 because the points are created as convex combination of existing data²⁵, when the syn-
 648 thetically generated points end up in the testing they depend, at least partially, on the
 649 data used in the training procedure.

650 1.3.4 Dimensionality reduction and clustering

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

653 There are various techniques that can mitigate or solve this problem, they are grouped
 654 under the umbrella term of "Dimensionality reduction techniques" and they are generally

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

655 based on ML learning methods. As such , following the same reasoning and definitions
656 given in regard to ML, it's possible to introduce a sub-categorization of the whole fam-
657 ily of techniques in supervised and unsupervised techniques. Dimensionality reduction,
658 beyond providing useful insight in the general structure of the data, can also be used as
659 a preprocessing step in some analysis pipelines.

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

664 The most common dimensionality reduction techniques are:

665 1. Unsupervised methods

- 666 • Principal Component Analysis (PCA):[14] Linearly combines the pre-existing
667 features to obtain new ones and orders them by decreasing ability to explain
668 total variance of data. The first principal component is the combination of
669 features that most explains the variance in the original dataset. Given it's
670 nature it focuses on the global characteristics of the dataset.
- 671 • t-distributed Stochastic Neighbor Embedding (t-SNE) [53]: Keeps a mixture
672 of local and global information by using a distance metric in the full high
673 dimensional space and trying to reproduce the distances measured in the
674 lower dimensional space which can be 2- or 3-dimensional.

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

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

686 2. Supervised methods

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

694 For an in depth description refer to [5] while for a more modern review refer
695 to [29]

²⁶The 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 [53]

696 3. Mixed techniques

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

701 This technique maintains very local information on the data and allows complete
702 freedom of choice in the final embedding space as well as the definition
703 of distance metrics in the feature space²⁷, it's also implemented to work an
704 a generic pandas dataframe in python so it can take in input a vast range of
705 datatypes. The math behind this method is much beyond the scopes of this
706 thesis, as such refer to [34] for more in depth information.

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

711 These upsides become particularly relevant in the field of clustering, where the ob-
712 jective is to group data in sets with similar features by minimizing the differences within
713 each group while maximizing the differences between different groups. Clustering tech-
714 niques can be roughly divided in:

- 715 1. Centroid based techniques: A user defined number of points is randomly located
716 in the data space, datapoints are then assigned to groups according to their dis-
717 tance from the closest center. The main technique in this category is k-means
718 clustering, and some, if not most, other techniques include it as step in the pro-
719 cessing pipeline²⁸. The main problems are firstly that these techniques require
720 prior knowledge, or at the very least a good intuition, on the number of clusters
721 in the dataset while also assuming that the clusters are distributed in spherical
722 gaussian distribution, which is not always the case.
- 723 2. Hierarchical clustering: The idea is to find the hierarchical structure in the data
724 using a bottom-up or a top-down approach, as such this category further subdivides
725 in agglomerative and divisive methods. In the first each point starts by itself and
726 then points are agglomerated using a similarity or distance metric, this build a
727 dendrogram in which the k^{th} level roughly corresponds to a k -centroind clustering.
728 In the latter the idea is to start with a unique category and then divide it in
729 subgroups.
- 730 3. Density based techniques: These methods use data density to define the groups by
731 looking for regions with larger and lower density as clusters and separations.

732 In order to evaluate the performance of these methods it's necessary to define metrics
733 that evaluate uniformity within clusters and separation among them, the choice in the
734 definition of metric should be taken in careful consideration since the different task may
735 imply very different optimal metrics or, put differently, the optimal technique for the
736 task at hand may very well depend on the metric used.

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

²⁸An example of such techniques is: affinity propagation

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

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

749 1.4 Combining radiological images with AI: Image 750 segmentation and Radiomics

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

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

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

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

776 1.4.1 Image Segmentation

777 Generally speaking image segmentation is a procedure in which an image is divided in
778 smaller sets of pixels, such that all pixels 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 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 the number of slices to draw on can be reduced using shortcuts, 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 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

²⁹n 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

821 the inputs it receives. In the case of images these structures are called Convolu-
822 tional Neural Networks because the first layers, which are intended to extract the
823 latent features or structures in the images, perform convolution operation in which
824 the parameters are the pixel values of convolutional kernels

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

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

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

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

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

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

849 For a more in depth review of the main methods used in medical image segmentation
850 refer to [41]. Worth noting, at this point, that the semi-automatic segmentation software
851 used in this work uses probably a mixture of region growing and thresholding methods,
852 maybe guided by an atlas. The segmentation process generally produces a boolean mask
853 which can be used to select, using pixel-wise multiplication, the region of interest in
854 the image. The next step in image analysis would be to derive information from the
855 region defined during segmentation, in general this step is called feature extraction³⁰
856 and its objective is to find non-redundant quantities that meaningfully summarize as
857 much properties of the original data as possible.

858 1.4.2 Radiomics

859 When the images are medical in nature and when referring to high-throughput quan-
860 titative analysis the task of finding these features fall in the realm of radiomics, which

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

861 uses mathematical tools to describe properties of the images that would otherwise be
862 unquantifiable to the human eye.

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

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

871 All of the features used in this work have been rigorously defined and described in
872 [62] which, being born as a reference manual, covers the founding concepts of radiomics
873 in an attempt to standardize the procedures of the field.

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

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

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

890 3. Higher order features: All the other features fall in this macro-category which can
891 be subdivided in a clear-cut way in other smaller categories. These categories are
892 created by grouping within them all the features that are obtainable following the
893 same guiding principle or starting point.

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

898 These matrices have very precise definition, as such only the general idea behind
899 them will be reported here redirecting to [62] for a more strict and in detail
900 description. The matrices from which the features are computed also give name to
901 the smaller categories in this family, these categories are:

- 902 • Gray Level Co-occurrence Matrix (GLCM) features: This matrix expresses how
903 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 $\delta=1$ or $\delta=2$, 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 dimensions³¹. 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 coarseness 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 or the width of the bins is fixed preceding the discretization process. The results of the feature extraction procedure are heavily dependent on the choices made in all the steps that precede 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 [62]) 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

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

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

949 In the past the main attention in radiological research was focused on improving
950 machine performances and evaluating acquisition sequence technologies, however the
951 great developments in artificial intelligence and performance of computers have brought
952 a lot of attention to the field. Various papers, such as [28] and [4] have been written
953 with the objective of presenting the general workflow of radiomics.

954 By design the topics in this thesis have been presented to resemble the shape of the
955 generic radiomic pipeline as outlined in [28] and [4], which can be summarised as:

- 956 • Data acquisition
- 957 • Definition of the Region Of Interest (ROI)
- 958 • Pre-processing
- 959 • Feature extraction
- 960 • Feature selection
- 961 • Classification

962 1.5 Survival Analysis

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

969 The main concepts and terminologies in this field are:

- 970 1. Event: This is the phenomenon under analysis, generally it could be death, remis-
971 sion from recovery, recovery and so on. It is common to refer to the event as failure
972 since usually it is negative in nature.
- 973 2. Censoring: When collecting data to develop a model that describes survival it may
974 happen that the individual drops out of the study without incurring in the event
975 under analysis. For example, when looking at effectiveness of a drug, the patient
976 may develop adverse reactions to the drug and may need to stop using it, hence
977 falling out of the study.

978 In the context of this thesis all individuals that got sent home from the hospital
979 are censored since their survival is known only up until the dropout and not after.
980 Cases like this when the start of the followup is well known but dropout happens
981 are called right-censored, because only the right side of the timeline is abruptly
982 interrupted. Cases can also be left-censored, e.g. a patients with unknown time
983 of contraction of a disease, and interval-censored, e.g. when the contraction of
984 the disease can be restricted to an interval as it may happen when a negative and
985 positive test happen at successive times.

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

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

992 In this thesis the variable used to cover this role was the **DOS** variable, which is
993 computed as the number of days from the admission in the hospital to the discharge
994 from the hospital facilities.

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

999 These assumptions are all reasonable since at no point the survival probability can
1000 increase, since everybody is alive at the start of observing them and since nobody
1001 can live to infinity. However, when it comes to practice, these properties are not
1002 necessarily verified. Since no study can continue to infinity the last value need not
1003 be zero and since the timesteps at which it's possible to perform a checkup are
1004 discrete the curve is actually a step function.

- 1005 5. Hazard Function **h(t)**: This function is difficult to explain practically, citing [27]
1006 "**The hazard function $h(t)$ gives the instantaneous potential per unit**
1007 **time for the event to occur, given that the individual has survived up**
1008 **to time t** ". The mathematical definition is:

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

1009 Dividing by a time interval the hazard function can also be intended as conditional
1010 failure rate, where the conditional refers to the "given that the individual has
1011 survived up to time t ".

1012 Being a rate this quantity need not be bound in $[0,1]$ but ranges in $[0, \infty]$ and
1013 depends on the time unit used. This will be interpreted as $h(t)$ events per unit
1014 time.

1015 The hazard function is non-negative and has no upper bounds, since it can be re-
1016 lated to the survival function³² the possible shapes give different names to the final
1017 model. These could be increasing or decreasing Weibull, exponential or lognormal
1018 survival models

1019 A further step in the analysis could be to try to measure the differences in survival
1020 between two groups, i.e. using a single variable expected to drive the differences in
1021 survival.

³²The 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.

1022 The most basic example would be looking at the effectiveness of a drug by dividing
1023 in group A and B the patients given the actual medicine and the placebo respectively.
1024 However this could be done even for males vs females or, in the case of continuous
1025 variables, for patients with above or below threshold values ³³. When the analysis is
1026 univariate in nature then the Kaplan-Meier survival curves are used in conjunction with
1027 the log-rank test, when the analysis is multivariate then the Cox Proportional-Hazard
1028 model is used. The Cox model is very similar to linear and logistic regression.

1029 1.5.1 Kaplan-Meier(KM) curves and log-rank test

1030 Kaplan-Meier curves can be built following a well defined procedure:

1031 The data is separated in sets using the label of the groups that need to be used then
1032 the patients in each set are ordered in ascending order of permanence in the study, which
1033 is the time variable.

1034 At each failure time T_f the conditional probability of surviving past T_f given avail-
1035 ability is computed as ratio of subjects left in the study right after T_f divided by the
1036 number of available people at that same time ³⁴ Note that this takes in account also the
1037 possible censoring by reducing the number of available people at times of event, these
1038 censoring event will be drawn on the curve using ticks at the corresponding time. The
1039 survival probability is computed at each event time as the product of the probability at
1040 the previous failure time with the one computed as explained before at the time of the
1041 current event.

1042 Let's consider an imaginary study with 4 patients with one failing at week one, one
1043 dropping out at week 2, one failing at week 3 and one surviving after 4.

1044 The KM curve will start at 1, at week one it will drop at $\frac{3}{4}$ since 3 people will be alive
1045 out of 4 available, at week two no drop will happen but a tick will be put on the curve
1046 and, finally at week 3 it will drop at value $\frac{3}{4} * \frac{1}{2}$ since only one out of the two available
1047 will survive.

1048 As a rule of thumb, two Kaplan-Meier curves that do not intersect at any point indi-
1049 cate good separation among the groups, this can then be formally evaluated performing
1050 a log-rank test on the data used to build the curves.

1051 The null hypothesis of this specific test is that there is no overall difference between
1052 the two survival curves [?] which can be tested with the log-rank test. This basically
1053 consist in performing a large-sample χ^2 test that uses the ordered failure times for the
1054 entire dataset as expected values vs those observed in the subsets. From this testing
1055 procedure a p-value can be obtained to reject the null hypothesis, for more in depth
1056 information refer to [27].

1057 1.5.2 Cox Proportional-Hazard (CoxPH) model

1058 As it was mentioned before the shape of the hazard curve, an hence of the survival curve,
1059 implies a specific shape for the model used to describe it. Some of the possible models
1060 are increasing Weibull, decreasing Weibull, Exponential and log-normal. All of these

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

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

1061 models are parametric because known the parameters the distribution of the outcome
1062 can be known.

1063 When it comes to Cox PH model the distribution cannot be known because part of
1064 the model, namely the baseline hazard, remains unestimated.

1065 Generally the data has an optimal parametric model that describes it and, if this
1066 information were known, it would make perfect sense to use said function. However this
1067 knowledge is not always obtainable, which give Cox PH model an occasion to shine. Cox
1068 Proportional Hazard models can be described as robust in the sense that the results that
1069 it gives will approximate those obtained by the correct model, without needing to know
1070 which of the model needs to be used [27].

1071 Cox PH models rely on the use of the formula in Equation 1.14 to express the risk
1072 of a patient with a set of k characteristic variables \mathbf{X} at time t.

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

1073 The quantity $h_0(t)$ is called baseline hazard and it hides one of the main hypotheses
1074 behind this method which is the *Proportional Hazard* assumption. This assumption is
1075 that the baseline hazard $h_0(t)$ depends on time alone and not on all the other variables
1076 \mathbf{X} , note also that the exponent is time-independent since the \mathbf{X} s are supposed to be
1077 constant in time³⁵.

1078 The β in the exponent represents the weights assigned to each variable and, very much
1079 like what happened in linear regression, these coefficients can be intended as a proxy of
1080 importance in the model: coefficients close to zero signify no particular importance of the
1081 variable whereas the more the value is distant from zero the more relevant the variable
1082 can be considered. The reason why it's common practice to report the exponentiated
1083 coefficient for features is that $1 - \exp[\beta]$ represents the difference in likelihood to die of
1084 individuals separated using the feature relative to β .

1085 For example, considering a binary feature AGE OVER 50 with $\exp[\beta]=1.6$, the expo-
1086 nential of the parameter would indicate that people over 50 are 60% more likely to die
1087 when adjusting for all other variables. It is also common procedure to provide, for each
1088 coefficient, the 95% confidence interval of the parameter and a p-value relative to the
1089 hypothesis that the parameter be equal to zero.

1090 Cox models provide a way to predict hazard for patients, this predict method can be
1091 used to evaluate the model built by identifying groups in the patient cohort using hazard
1092 quantiles. The division in the resulting groups can be used as proxy for the quality of the
1093 score built using the Cox model.

1094 Finally, much like the previously presented regression techniques, it's possible to
1095 penalize the regression done with the Cox model.

1096 In this thesis the Cox model was preceeded by the same feature reduction that was
1097 used before the Lasso regularized regression, however no regularization was used.

³⁵Dropping 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*.

₁₀₉₈ **Chapter 2**

₁₀₉₉ **Materials and methodologies**

₁₁₀₀ In this section there's going to be an explanation of the dataset as well as instruments
₁₁₀₁ and methodologies used to analyze it's properties, as such the first step is going to be
₁₁₀₂ an in depth discussion of the data available and a general overview of the final use. The
₁₁₀₃ following step is going to be a description of the preliminary work done to the data itself
₁₁₀₄ and to the results of this preliminary analysis in order to select the important features.
₁₁₀₅ The final step of this chapter is going to be an explanation of the methods used to derive
₁₁₀₆ the final results and to evaluate them.

₁₁₀₇ Worth noting that the data has been collected and used with the approvation of the
₁₁₀₈ ethical commettee.

₁₁₀₉ **2.1 Data and objective**

₁₁₁₀ The objective of this thesis has been to compare how different methods perform in
₁₁₁₁ predicting clinical outcomes in covid patients, while also determining if different kind
₁₁₁₂ of input data imply different performances of the same methods. All images available
₁₁₁₃ were, at the start, not segmented. As such all of them have been semi-automatically
₁₁₁₄ segmented via a new software being tested in the medical physics department called
₁₁₁₅ *Sophia Radiomics* [1] which seems to be built around region growth algorithm mixed
₁₁₁₆ with thresholding.

₁₁₁₇ Statistical analyses have been performed in python using libraries such as scikit-learn
₁₁₁₈ [40] and imblearn¹ [30], pandas [35], numpy [17], scipy [54], statsmodels[47].

₁₁₁₉ Lifelines [12] has been used for survival analysis and combined to scikit-learn by
₁₁₂₀ coding wrappers that made compatible the functions from lifelines with the API of
₁₁₂₁ scikit-learn.

₁₁₂₂ Finally all of the management of the graphs obtained as part of the analysis has been
₁₁₂₃ performed with either seaborn[55] or matplotlib[23].

₁₁₂₄ The starting dataset was a list of all the patients that, from 02/2020 to 05/2021, were
₁₁₂₅ hospitalized as COVID-19 positive inside the facilities of *IRCSS Azienda ospedaliero-*
₁₁₂₆ *universitaria di Bologna - Policlinico Sant'Orsola-Malpighi*.

₁₁₂₇ As far as exclusion criteria go the main deciding factors, except unavailability of
₁₁₂₈ the feature related to the patient, were visibly damaged and lower quality images, for
₁₁₂₉ example images with cropped lungs. The first set of selection criteria were:

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

- 1130 • All patients that had undergone a CT exam which was retrievable via the PACS
 1131 (Picture Archiving and Communication System) of *IRCSS Azienda ospedaliero-*
 1132 *universitaria di Bologna - Policlinico Sant'Orsola-Malpighi*
- 1133 • All patients that had all of the clinical and laboratory features, listed in fig:2.1,
 1134 which have been found informative of the outcome during the clinical practice.
- 1135 • Since all patient had at least 2 CT exams only the closest date to the hospital
 1136 admission date was taken. When more exams were performed on the same date
 1137 all of them were initially taken. At first only chest or abdomen CTs were taken
 1138 regardless of the acquisition protocol used.

1139 In tables 2.1,2.2 and 2.3 are reported all of the variables available for all patients with
 1140 some information on their distribution. These have been divided in: Clinical features
 1141 used for the models Table 2.1, variables determined by radiologist by examining the
 1142 CT scan of the patient, called Radiological fetures Table 2.2. Finally in Table 2.3 are
 1143 contained all the variables that represent the outcome of the illness and not a property
 1144 of the patient at admission which is why these won't be used in building the predictive
 1145 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	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

1146 Most clinical features are pretty self-explanatory, a brief explanation will be provided
 1147 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

- 1148 1. DNR : Acronym for "Do Not Resuscitate", used to indicate the wish of the patient
1149 or their relatives that cardiac massage not be performed in case of cardiac arrest.
- 1150 2. NIV: Acronym for "Non Invasive Ventilation", it's a form of respiratory aid provided
1151 to patients.
- 1152 3. cPAP: Acronym for "continuous Positive Airway Pressure", another form of respi-
1153 ratory aid.
- 1154 4. ICU: Acronym for "Intensive Care Unit". When patients are in really severe con-
1155 ditions they are treated in these facilities.
- 1156 5. Clinical Scores: When available values from laboratory analyses and/or patient
1157 conditions are summarised in scores that represent the gravity of the state of the patient,
1158 as such these can be somewhat correlated and could be treated as com-
1159 prehensive values to substitute an otherwise large set of obscure clinical features.
1160 At admission, or closely thereafter, a set of clinical questions regarding the patient
1161 receives a yes or no answer, each answer has an additive contribution towards the
1162 final value of the score.

1163 These scores differ in how much they add for each condition and the set of symptoms
1164 the check for.

- 1165 (a) MulBSTA: This score accounts for **M**ultilobe lung involvement, absolute
1166 **L**ymphocyte count, **B**acterial coinfection, history of **S**moking, history of hy-
1167 **p**er**T**ension and **A**ge over 60 yrs. [16]
- 1168 (b) MEWS: Modified Early Warning Score for clinical deterioration. Computed
1169 considering systolic blood pressure, heart rate, respiratory rate, temperature
1170 and AVPU(Alert Voice Pain Unresponsive) score. [49]
- 1171 (c) CURB65: **C**onfusion, blood **U**rea Nitrogen or Urea level, **R**espiratory Rate,
1172 **B**lood pressure, age over **65** years. This score is specific for pneumonia severity
1173 [57]

- 1174 (d) SOFA: Sequential Organ Failure Assessment score. Considers various quan-
 1175 tities from all systems to assess the overall state of the patient, $\text{PaO}_2/\text{FiO}_2^2$
 1176 for respiratory system, Glasgow Coma scale³ for nervous, mean pressure for
 1177 cardiovascular, Bilirubin levels for liver, platelets for coagulation and creatine
 1178 for kidneys [3]
- 1179 (e) qSOFA: quick 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

1180 Finally it's also useful to see a distribution of the days of permanence in the Hospital,
 1181 abbreviated DOS for days Of hospitalization⁴, which can be seen in Figure 2.1.

1182 This procedure, which is the starting part of what is summarized and represented
 1183 in Figure 2.2, produced a starting cohort of ~ 700 patients which, having all various
 1184 images available, created a huge set of ~ 2200 CT scans. Since this analysis is focused
 1185 on radiomics there is an evident need for as much consistency as possible in the images
 1186 analysed. For this reason all CTs taken with medium of contrast were excluded, since
 1187 they would have brightnesses not indicative of the disease, and for every patient only
 1188 images with thin slice reconstruction were considered.

1189 1190 More specifically only images with slice thickness of 1 o 1.25 mm⁵ along the z-axis
 1191 were taken into consideration, which meant excluding all the 1.5, 2, 2.5 and 5 mm slice
 1192 thicknesses.

²Very unrefined yet widely used indicator for lung dysfunction

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

⁴It's common practice to use abbreviations such as dOS, mOS. The common way to intend these acronyms is either days or months Of Survival

⁵This 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.

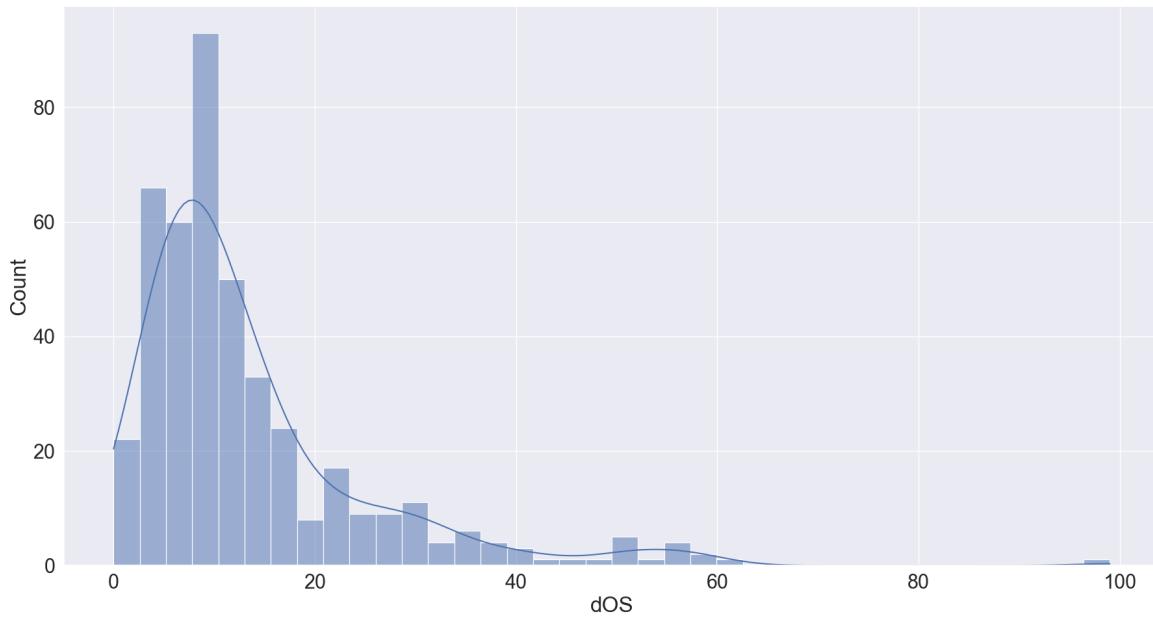


Figure 2.1: Distribution of the days of hospitalization for the patients included in the study.

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 that follow 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 Tables 2.1, 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 FEVER⁶ as of hospital admission and whether or not the patient suffered of HYPERTENSION

⁶Defined as body temperature $> 38^\circ$

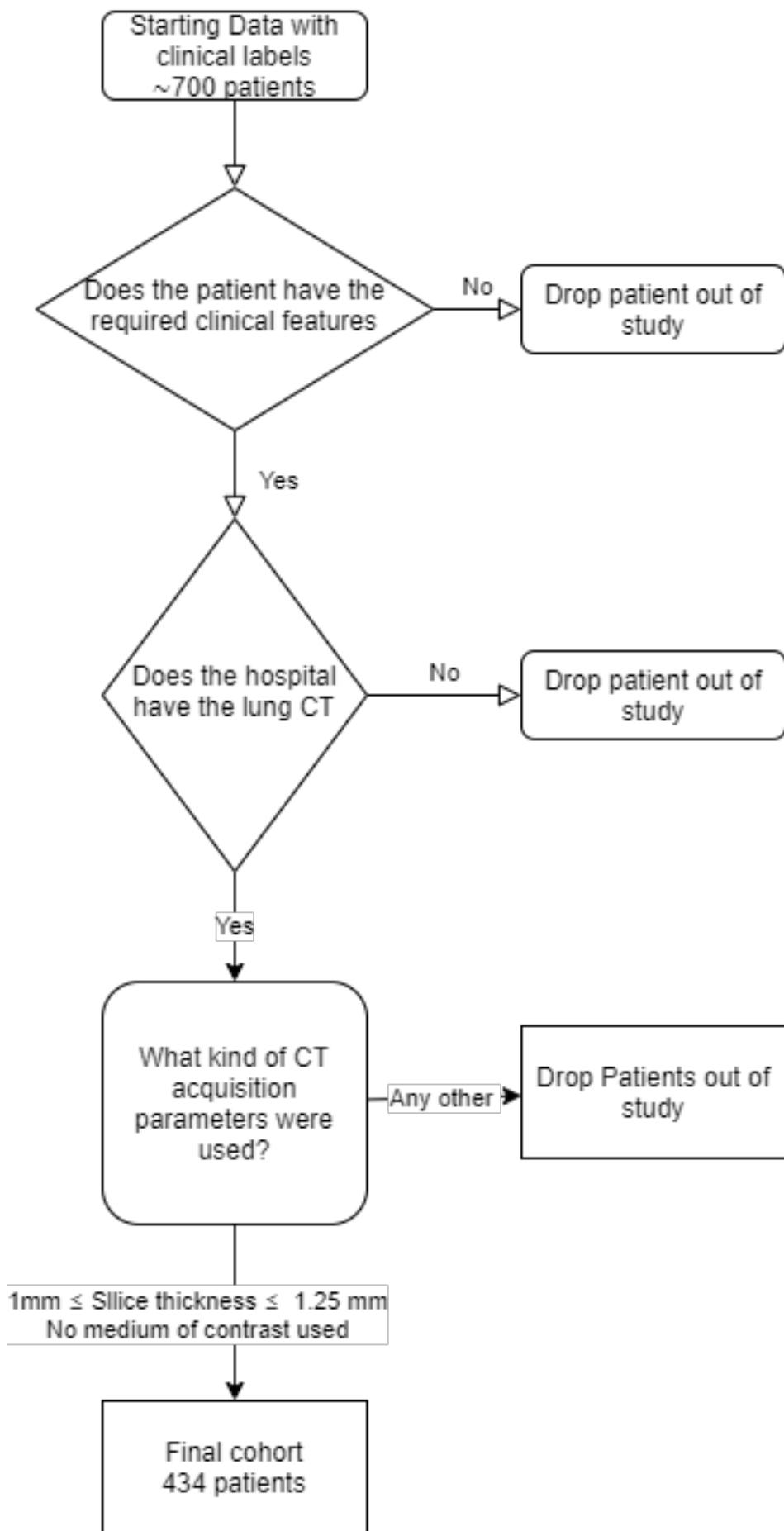


Figure 2.2: Flowchart of the patient selection procedure

1217 • The remaining features, namely those in 2.1 as well as the DEATH status of
1218 the patient, were either used as labels or not used at all because they refer to
1219 treatments used and not characteristics of the patient. As such these features,
1220 while plausibly correlated to the clinical outcome, are not really descriptive
1221 of the patient as of admission and are not information that can be used to aid
1222 professionals at admission to assess the situation

- 1223 2. Radiomic: These features were all the ones supported by the segmentation software
1224 and are pretty much most of those described in [62] with the addition of fat and
1225 muscle surface, computed as cm^2 by counting pixel identified via threshold as fat
1226 or muscle tissue in thoracic slices taken at height of vertebra T-12
- 1227 3. Radiological: These features are those that can be derived from CT exams by hu-
1228 mans. Namely acquisition parameters, such as KVP and CURRENT, were used
1229 to search for eventual correlations between image quality and predictive power of
1230 the feature derived from the image while boolean features, such as BILATERALITY
1231 of lung damage, presence of GROUND GLASS OPACITIES (GGO), LUNG CON-
1232 SOLIDATIONS as well as CRAZY PAVING were used to see if they were sufficient in
1233 determining outcome.

1234 2.2 Preprocessing and data analysis

1235 Before any preprocessing a choice was made to exclude all of the clinical scores, this was
1236 done to avoid having them mask other, more straightforward variables.

1237 The first step in the analysis of this data is going to be a lasso regularized regression
1238 using either DEATH or ICU ADMISSION as target. As mentioned before when operationg
1239 with regressions it's a necessity that at least the feature distributions be as symmetric
1240 as possible, a common way to get as close as possible to this empirical requirement is to
1241 boxcox transform the data. Apart from the data which contained negative values, which
1242 cannot be fed into the boxcox transform, all variables have been transformed using this
1243 method.

1244 The quatile-quantile plots have been used for a visual check of normality, normally
1245 distributed data will populate the bisector of the graph while deviations are symptoms
1246 of non-normality. Heavy and light tails are visible as deviations respectively above and
1247 belox the bisector.

1248 It should be noted that when the data is normally distributed then the transform
1249 doesn't change much the distribution while, in cases with much more pronounced asym-
1250 metries in the distributions, the improvement obtainable can be visualized in Figures 2.4
1251 and 2.5.

1252 The next preprocessing step has been to apply a *StandardScaler* to all of the features,
1253 which corresponds to subtracting the mean and dividing by the standard deviation, in
1254 order to center all the features around zero.

1255 The final step in preprocessing has been to reduce the features by using a correlation
1256 threshold which means that all variables that correlate with another more, in absolute
1257 value, than a certain threshold, which has been set to 0.6 in this work, are dropped
1258 a priori. A rather important thing to notice is that correlation has been computed

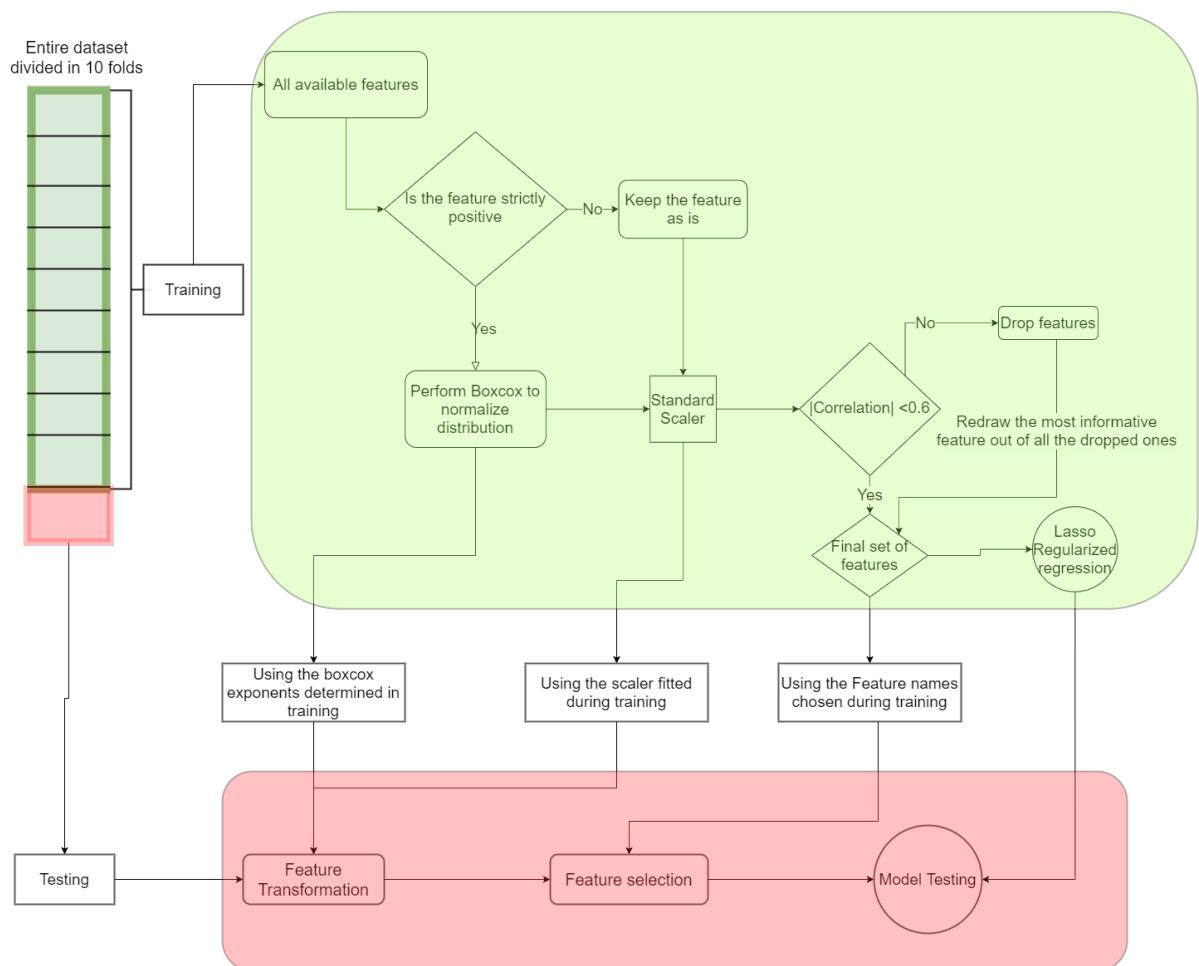


Figure 2.3: Flowchart for the preprocessing steps before proceeding a Lasso regularized regression.

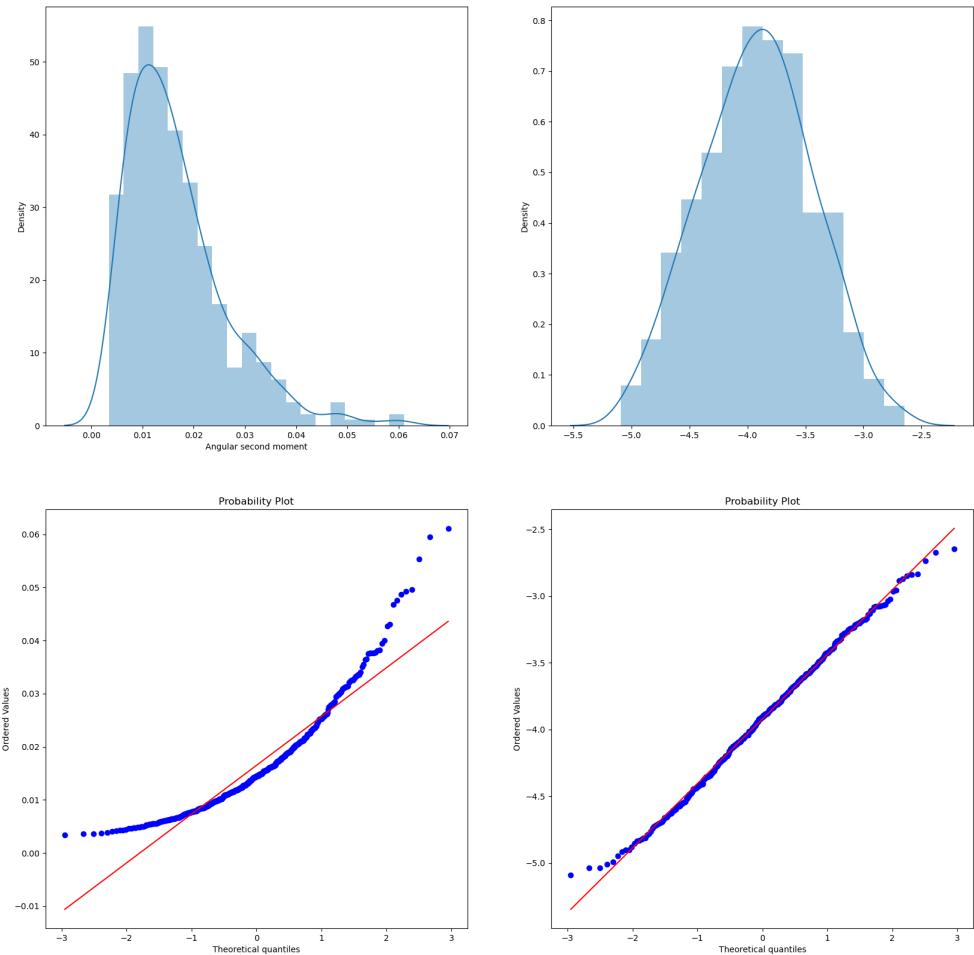


Figure 2.4: Example of boxcox applied to the radiomic feature *Angular second moment* which has a heavy tailed distribution. The graphs contain the original distribution (top-left) the transformed distribution (top-right) and the two respective quantile-quantile plots(bottom)

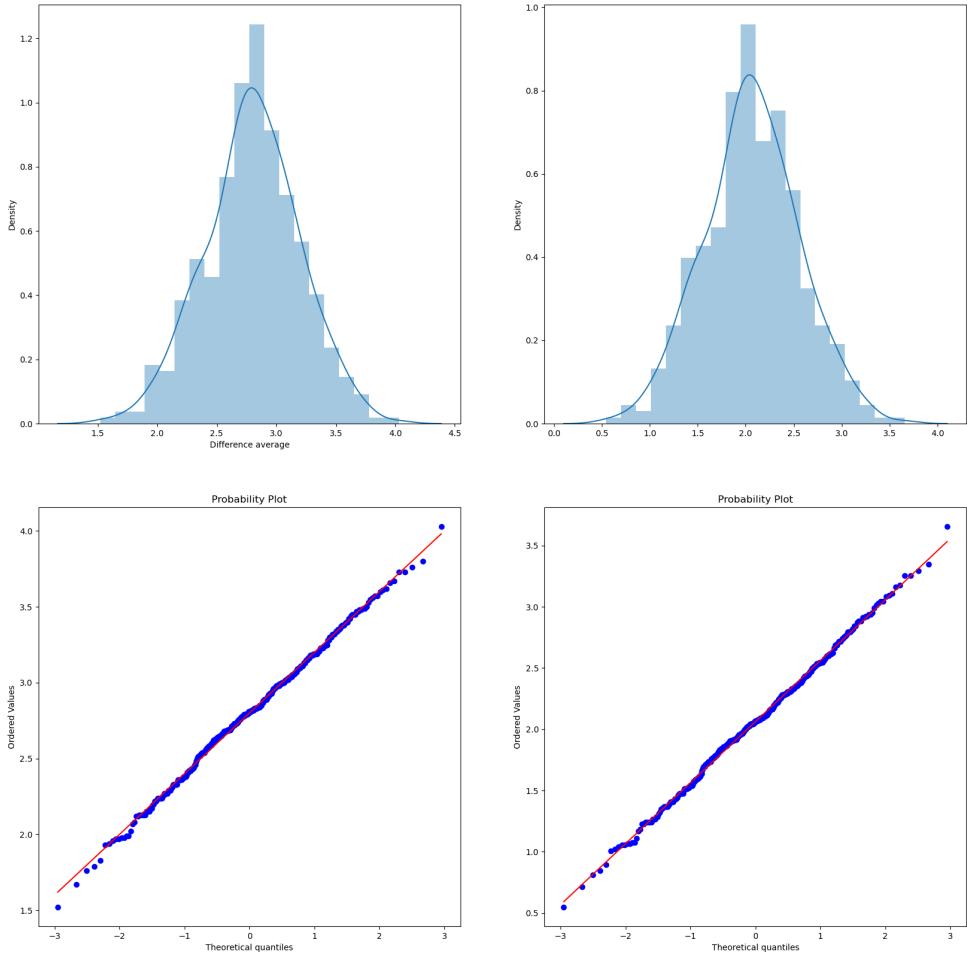


Figure 2.5: Example of boxcox applied to the radiomic feature *Difference Average* which has a close to normal distribution. The graphs contain the original distribution (top-left) the transformed distribution (top-right) and the two respective quantile-quantile plots(bottom)

1259 using Spearman correlation and not Pearson since the first is invariant under monotone
1260 transformations, such as boxcox and standard scaling, whereas the second is not.

1261 Given the large number of features it's very plausible that at least one of the eliminated
1262 features is correlated with all other dropped features but with none of the remaining ones,
1263 since a Lasso regularization will be used introducing a few redundant features is not
1264 too damaging and the possible benefits outweigh the risks. For this reason a redrawing
1265 method has been implemented to add one of the dropped features, this has been done
1266 by choosing the one that most correlates with the label being used.

1267 A pivotal point in all of this analysis is that, to obtain reasonable values in the cross-validation
1268 procedures and to avoid leakage⁷ problems, all of the preprocessing steps have
1269 been done after train-test splitting the data on the train set and then applied as defined
1270 during training on the test dataset.

1271 When it comes to cross-validation procedure the choice was made to use a stratified
1272 k-fold approach with k=10. The data is split in 10 parts with the same percentages of
1273 labels⁸ then a model is built by training on 9 of the folds and it's performance is then
1274 tested on the remaining fold. To use the whole dataset for testing a prediction of it
1275 has been built by combining the predictions on the 10th fold for ten different models
1276 trained on the respective 9 remaining folds.

1277 The lasso model from training on the 9 folds is actually chosen as the model with the
1278 hyperparameters that give the best performance with another 10-fold crossvalidation.
1279 This has been obtained by using *cross_val_predict* on a model obtained by including a
1280 *LassoCV* step inside a *pipeline* from scikit-learn library in python.

1281 The performance of the cross-validated predictions that, when built this way, is much
1282 more representative of the real-world performance of the model, has been evaluated using
1283 ROC curves and AUC. Different models have been compared with a Delong test[13] for
1284 the significance of difference in the ROC curves.

1285 The second analysis method used was RandomForest. As said before in this case
1286 no preprocessing was needed nor has been done, however particular care was taken in
1287 handling the imbalances in the dataset by using SMOTE [9] once again being careful to
1288 avoid leakage.

1289 The performance of this model was evaluated using confusion matrices which, at
1290 a glance, provide very much information on the situation of the data. To facilitate
1291 the comparison of the results of the Random Forests with those obtained using Lasso
1292 regularized regression ROC curves were also made.

1293 As a standalone method a Cox Proportional-Hazard model was used on the standard
1294 scaled variables remaining after the feature reduction performed through correlation
1295 thresholding. The score obtained with this procedure was then divided using different
1296 percentiles and tested using the log-rank test on Kaplan-Meier curves relative to the
1297 groups built. In the case of this thesis the time variable was represented by the days of
1298 hospitalization computed using the dates of admission and discharge from the hospital
1299 provided by the hospital itself. The idea of resorting to Kaplan-Meier curves was also
1300 used with other single variables to see if they had any effect.

⁷This term is used in the field of Machine Learning. It refers to models being created on information that comes from outside the training data.

⁸The 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 occurrences of the two labels

1301 Finally a few dimensionality reduction techniques, namely the unsupervised PCA[14]
1302 and Umap [34] and the supervised PLS-DA[5], have been used to understand better
1303 the state of the data and further explain some of the obtained results. These peculiar
1304 analyses will be reported in the Appendix.

1305 This concludes the discussion on the methodologies used and leads perfectly in the
1306 discussion of the results.

₁₃₀₇ **Chapter 3**

₁₃₀₈ **Results**

₁₃₀₉ In this chapter the result obtained with the methods explained in the previous chapters
₁₃₁₀ will be briefly presented, to ease in weaving through the quantity of results a structure
₁₃₁₁ of the result presentation is reported in fig 3.1 and a final summary will be provided in
₁₃₁₂ the following chapter.

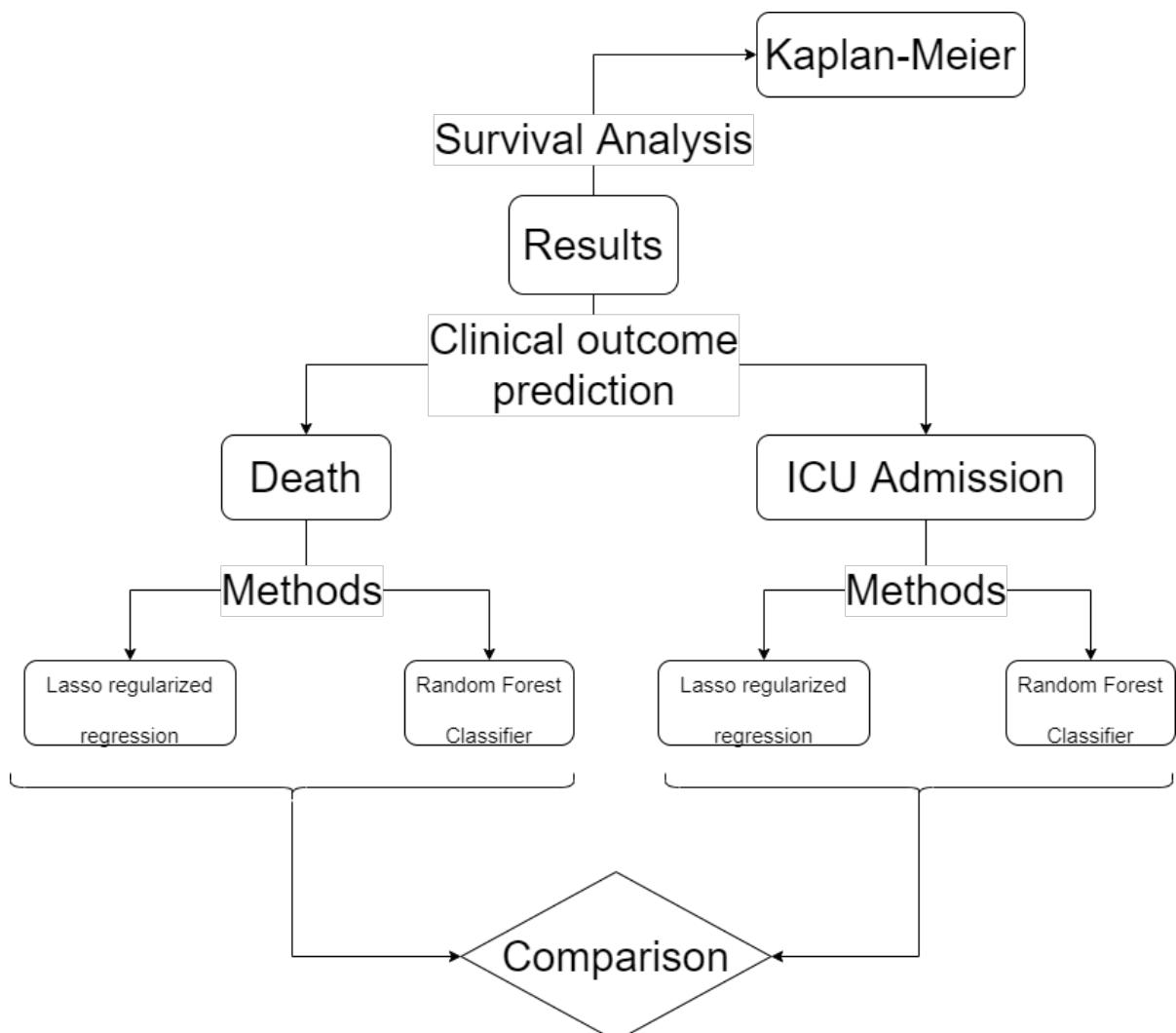


Figure 3.1: Logical structure used in the presentation of the results.

1313 3.1 Predicting and classifying the outcome Death

1314 First step in reporting the results is going to be using DEATH as the clinical outcome of
1315 interest for either Lasso regularized regression or Random Forest classifier.

1316 3.1.1 Feature selection through Lasso regularization 1317 and clinical outcome prediction using regres- 1318 sion

1319 When it comes to lasso regression usually graphs are reported that show the convergence
1320 of the parameters to the final value. Since the real information of this process is the value
1321 to which the coefficients converge only these values will be presented in tables and the
1322 ROC curves, with respective AUCs, will be provided. An example of the aforementioned
1323 graph is the one visible in Figure 3.2.

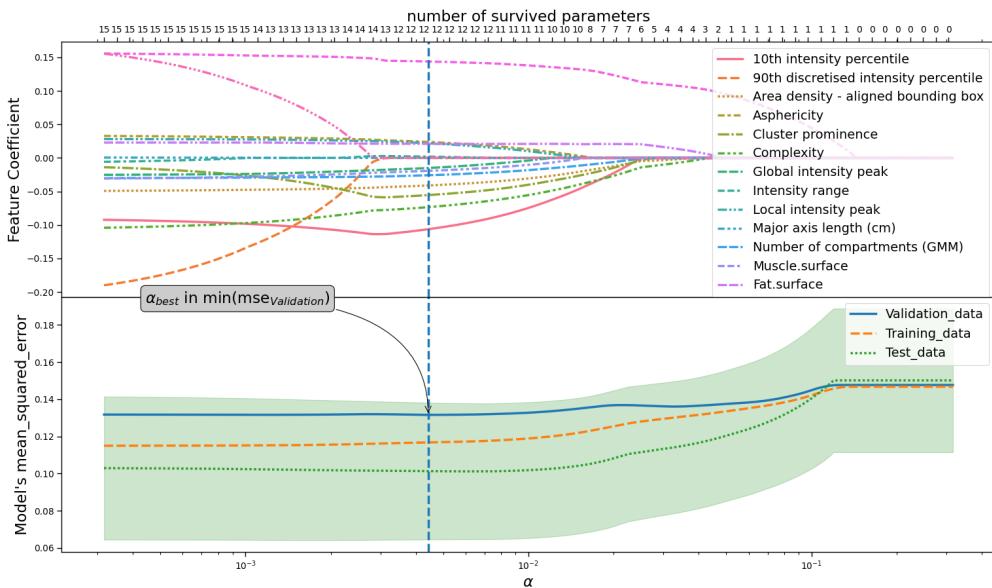


Figure 3.2: Example of graph representing the convergence of the coefficients in a lasso procedure. The top graph represents how the values of the weights of the input features change as the Lasso hyperparameter α changes, the top axis label indicates at each point how many features would have weight different from zero. The bottom graph shows the curves that represent the behaviour of the mean squared error of the model on the validation set, the test set and the training set. One of the ways to find the optimal value of the α parameter is to find the value that minimizes the mean squared error relative to the validation set. The vertical dashed line is a graphical representation of how the values for the feature weight is chosen. This particular graph is relative to only radiomic features and comes from the regularization of a model that uses DEATH as target variable.

1324 Finally it seems useful to report Table 3.1 the contingency table that gives an idea
1325 on how superimposed the clinical outcomes on DEATH and ICU ADMISSION are.

Table 3.1: Contingency table that quantifies overlap between individual accessed in the ICU and Dead individuals.

		ICU ADMISSION	
		0	1
DEATH	0	311	47
	1	48	30

1326 Starting to predict the death outcome, using Lasso regularized regression, of the
 1327 patient different groups of features have been used, first of all only the radiomic features
 1328 have been used. The ROC curve obtained with the radiomic features is the one in Figure
 1329 3.3. The curve in bold is an average curve obtained by aggregating the ten curves relative
 1330 to each of the folds used in testing and the gray band represents a ± 1 standard deviation.

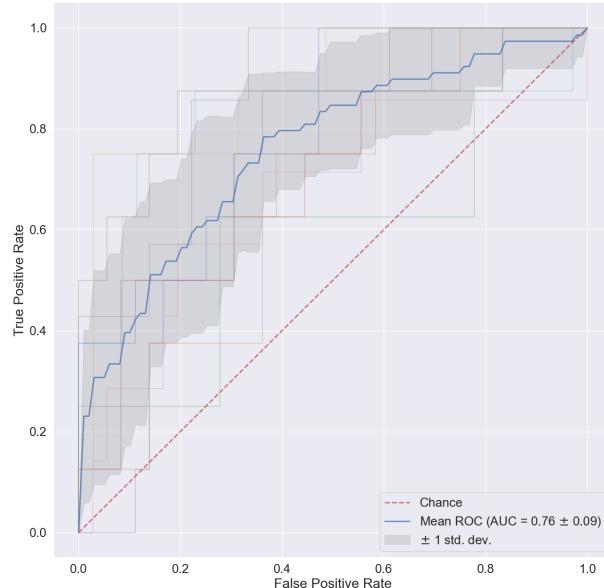


Figure 3.3: ROC curves obtained with crossvalidation procedure using the radiomic features alone. In bold is the mean ROC with gray bands of width equal to the standard deviation.

1331 The model that was built using the coefficients reported in 3.2 reaches a $AUC =$
 1332 0.76 ± 0.09 . The features inside the tabular, as well as those in the tabulars that follow,
 1333 will have the coefficients in descending order by absolute value so that the top features
 1334 are the most relevant within it's relative model.

1335 Before commenting more in depth the performance of this model it seems appropriate
 1336 to see at least the other models built with singular features groups, so, when it comes to
 1337 the clinical features, the results reported in Figure3.4 and Table3.9 have been obtained.
 1338 All the results will be put together for ease in Table 3.6.

Table 3.2: 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 3.3: 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

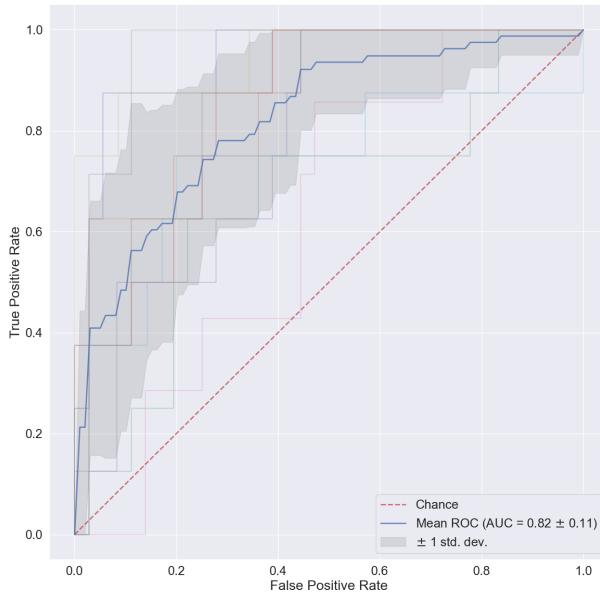


Figure 3.4: ROC curves obtained with crossvalidation procedure using the clinical features alone. In bold is the mean ROC with gray bands of width equal to the standard deviation.

1339 And finally, considering the radiological features Figure 3.5 and Table 3.10 are ob-
 1340 tained.

1341 The first thing to notice is that the radiological features, when considered alone, have
 1342 close to null predictive power. This is reasonable for at least part of the features because
 1343 there is no reason for acquisition parameter to actually influence the outcome of the
 1344 patient. When it comes to the radiologically determined quantities, such as GGO, Crazy
 1345 paving, lung consolidation and bilaterality even if one would expect these to be relevant
 1346 their distribution across the dataset is not conducive to the good predictions. In fact 88%
 1347 of patients had GGO, 50% of all patients had Lung consolidation, 77% of all patients did
 1348 not have Crazy paving and 92% had bilateral involvement.

1349 When it comes to clinical features, category that performs better when considered
 1350 singularly, nothing groundbreaking has been obtained. Age, Respiratory rate and sex are
 1351 the most relevant features and are all very much in concordance with what is expected.
 1352 Finally radiomic features perform slightly worse than the clinical features. To see if the
 1353 feature obtained are, at least, reasonable a quick explanation is needed. This will be done
 1354 only for the top performing features while deferring to [62] for the complete description:

- 1355 • 10th intensity percentile and Intensity based interquartile range are both intensity
 1356 based statistics.
- 1357 • Complexity: A complex image is one that presents many rapid changes in intensity
 1358 and is heavily non-uniform because it has a lot of primitive components.
- 1359 • Cluster Prominence: GLCM feature which measures the symmetry and skewness
 1360 of the matrix from which it derives. When this is high the image is not symmetric-

Table 3.4: 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

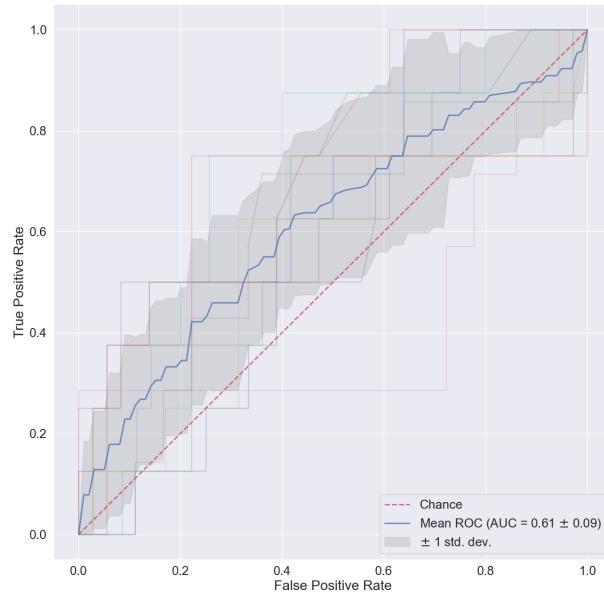


Figure 3.5: ROC curves obtained with crossvalidation procedure using the radiological features alone. As before in bold is the mean ROC with bands of width equal to the standard deviation.

- 1361 • Area density aligned bounding box: This is a ratio of volume to surface.
- 1362 • Entropy: Measures the average quantity of information needed to describe the
- 1363 image. In other words it quantifies randomness in the image, the more random the
- 1364 more info is needed to describe it.

1365 To summarize, since all of the features are computed on the whole lung segmenta-

1366 tion, it seems that the some information on the distribution of gray levels as well as

1367 some textural information inside the whole lung are important. It also seems that some

1368 information on the shape of the organ itself is also relevant.

1369 One would expect that when combining all of the available features, i.e. by building

1370 a model using the previous clinical, radiomic and radiological features, the performance

1371 should somewhat rise especially given the fact that clinical and radiomic features have

1372 almost the same performance. The combined results can be seen in Figure 3.6 and Table

1373 3.11.

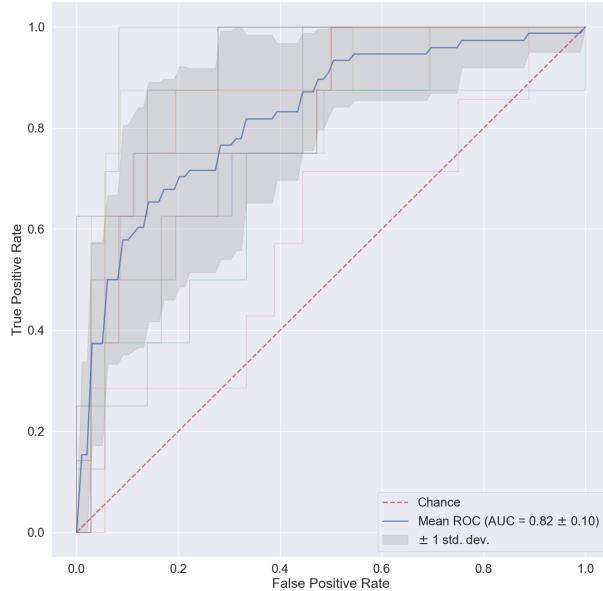


Figure 3.6: ROC curves obtained with crossvalidation procedure using all the available features. Model obtained with Lasso regularization of a linear regression modelling death

1374 In spite of what the expectations were, the performance of the combined case seems

1375 comparable if not equal to that obtained with clinical variables. To be sure of this

1376 claim a Delong test was used to compare pairwise the receiver operator curves and their

1377 respective AUCs.

1378 The null hypothesis of this test is that the two models are the same, hence a p-value

1379 smaller than 0.05 means that the curves and their AUCs are statistically different. The

1380 results from this analysis can be seen in 3.7

Table 3.5: 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

Table 3.6: 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

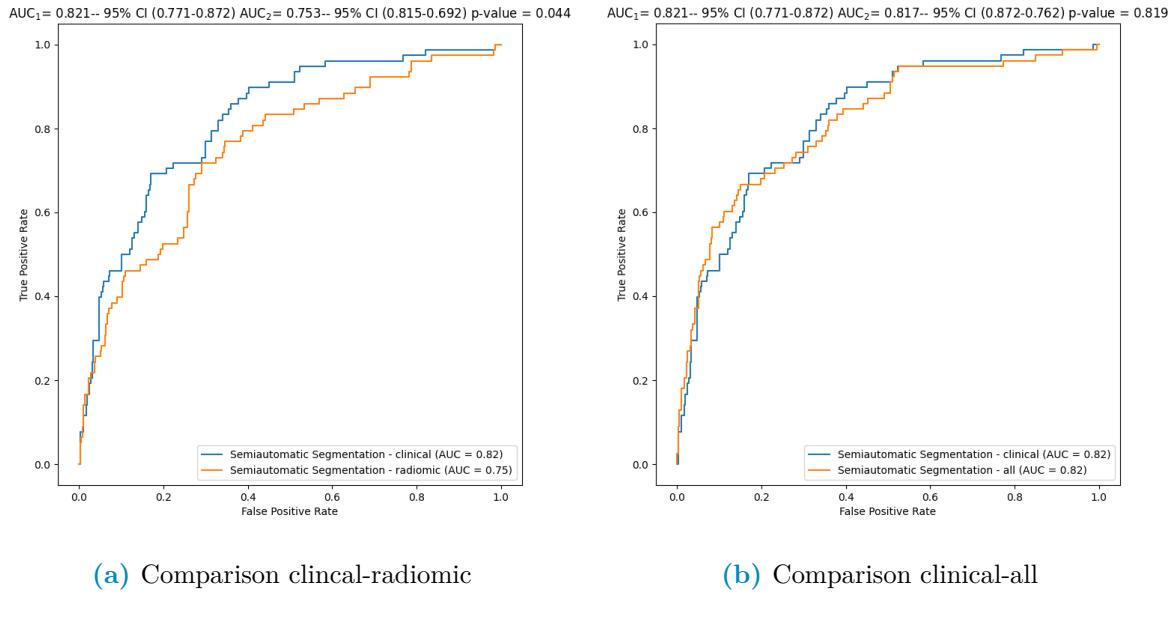


Figure 3.7: Comparison between ROC curves for clinical vs radiomic curves (a) and clinical vs all (b). The p-values are obtained with a Delong Test

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 Classification of patients using Random forests

For the sake of brevity all of the results will be reported and then discussed. Since RF classifiers use all of the available 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 ??.

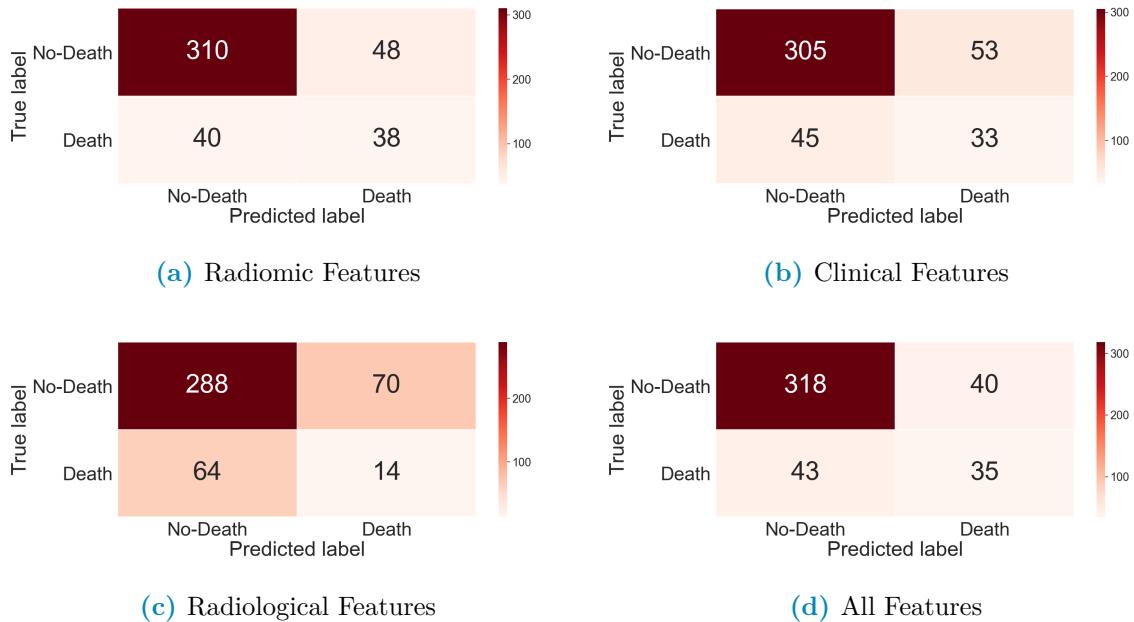


Figure 3.8: Confusion matrices for Random Forest cross-validated predictions after training on Synthetically oversampled data to predict DEATH . All of the available feature families are reported

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 3.9 and Table 3.7 summarises the results.

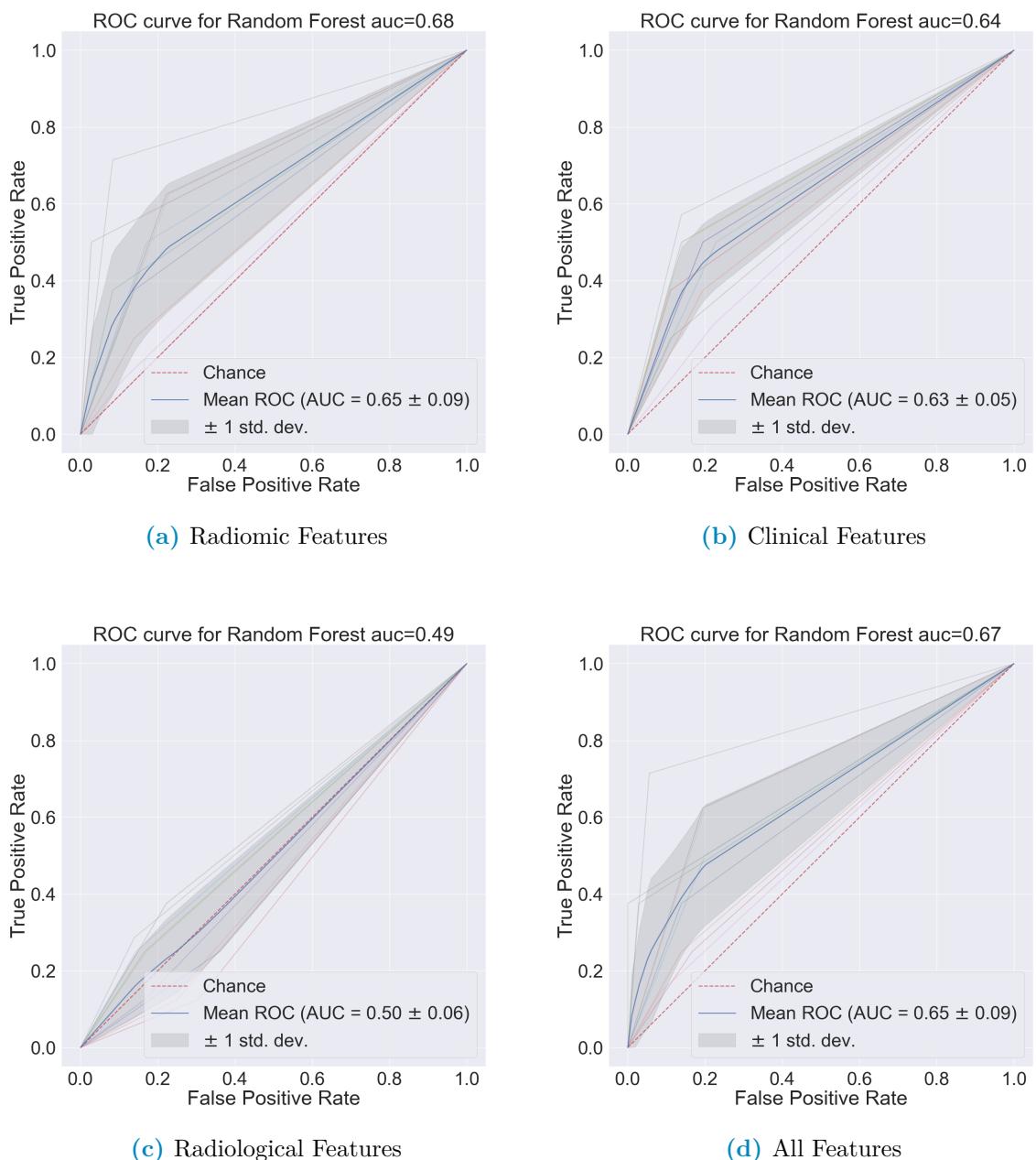


Figure 3.9: Cross-validated ROC curves built with Random forest classifier predictions of DEATH . Performances of allvariable families are reported

Table 3.7: 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	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
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

(a) Radiological Features

(b) Clinical Features

Figure 3.10: Importances estimated by random forest

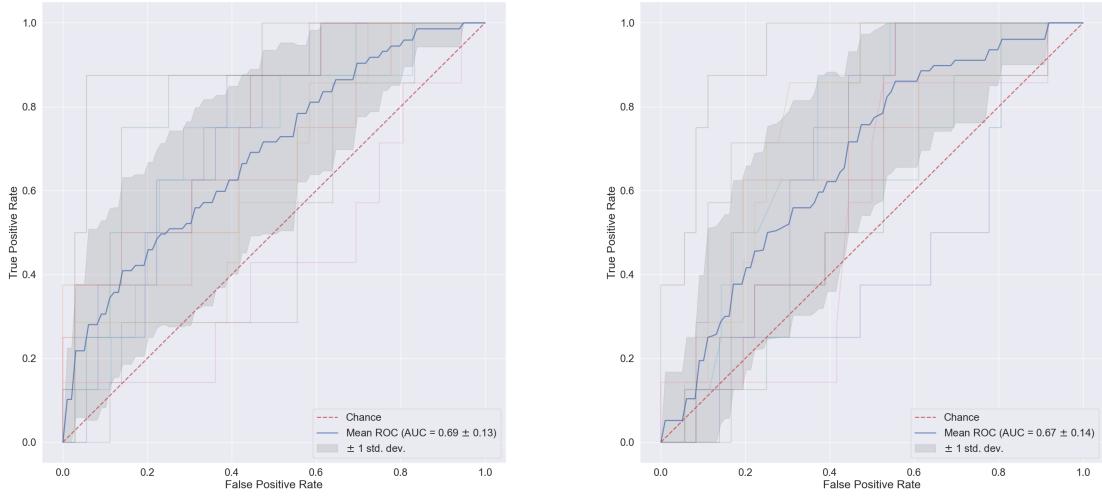
1404 Once again the curves are evidently not statistically different. This counterintuitive
 1405 behaviour seems to be constant across the two implemented methods and also across
 1406 different labels tried. An attempt to explain this phenomenon will be postponed to the
 1407 next chapter

1408 3.2 Predicting and classifying the outcome ICU Ad- 1409 mission

1410 The second step in reporting the results is the comparison of the methods used to predict
 1411 ICU ADMISSION as clinical outcome.

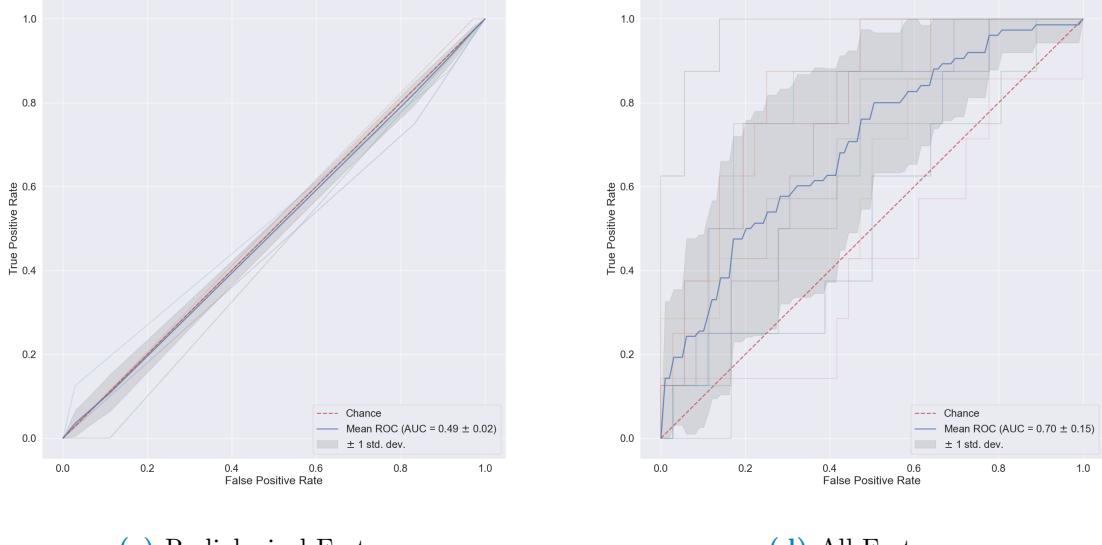
1412 3.2.1 Feature selection through Lasso regularization 1413 and clinical outcome prediction using regres- 1414 sion

1415 In trying to predict if the patient will be admitted in the Intensive Care Unit of the
 1416 hospital the same procedure as before has been used. The ROC curves obtained with
 1417 the various features are reported in Figure 3.11. Just like before the curve in bold is an
 1418 average curve obtained by aggregating the ten curves relative to each of the folds used in
 1419 testing and the gray band represents a ± 1 standard deviation. Following the blueprint
 1420 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 3.11: 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

1421 Compared to before the performance is definitely worse. The radiological features
 1422 have the same performance of a random variable, which is not that concerning given their
 1423 expected impact on gravity of the clinical picture of the patient. Even if superfluous a
 1424 Delong test was used to confirm that the hypothesis of the curves being equal could not
 1425 be rejected. When it comes to the relevant features in each model the following can be
 1426 deduced:

- 1427 • For the clinical features all of them have a role in the prediction. The only surprising

Table 3.8: 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 3.9: 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 3.10: 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 3.11: 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 3.12: Recap table with the performance of the various models built for different families of features when predicting ICU ADMISSION

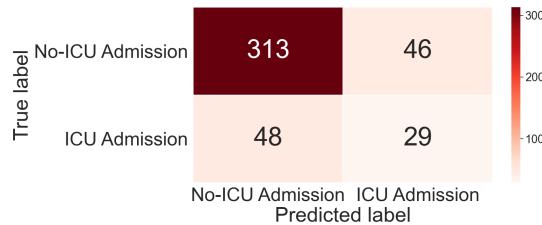
Features used	mean AUC ± std
Radiomic	0.69 ± 0.13
Clinical	0.67 ± 0.14
Radiological	0.49 ± 0.02
All	0.70 ± 0.15

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 .

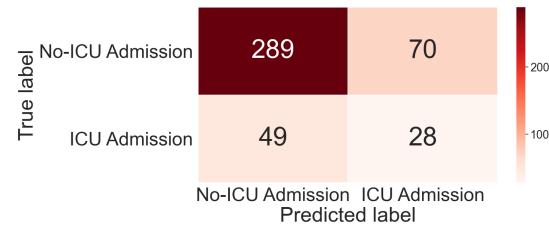
- None of the radiological features have virtually any impact
- The radiomic features still value intensity measurements and disorder in the image as primary origins of information. However it seems that shape of the lung now has more relevance in the whole model.

3.2.2 Classification of patients using Random forests

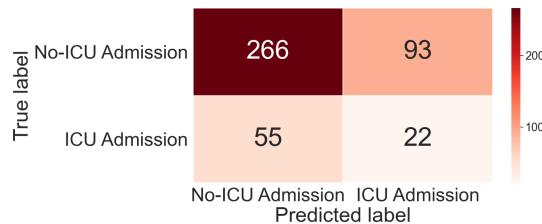
Even when using the admission in the ICU the performance of random forests remains pretty much the same when compared to the outcome DEATH , so the comments would still be the same as before.



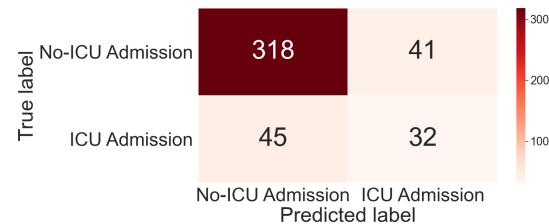
(a) Radiomic Features



(b) Clinical Features



(c) Radiological Features

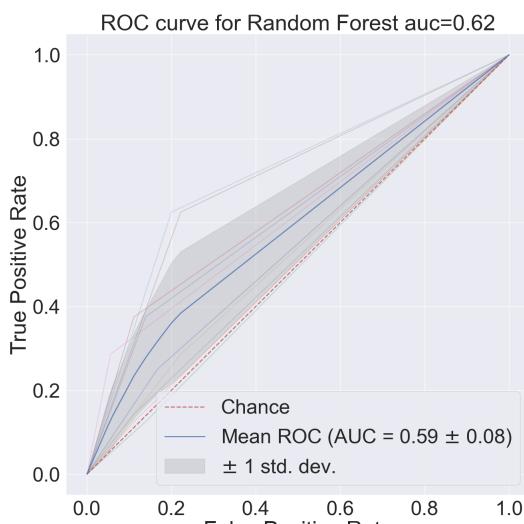


(d) All Features

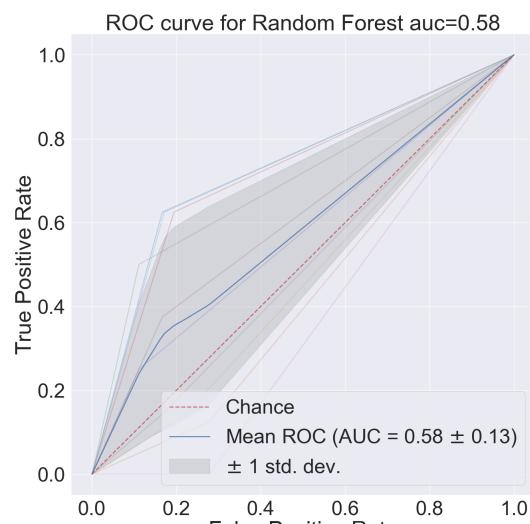
Figure 3.12: Confusion matrices for Random Forest cross-validated predictions after training on Synthetically oversampled data predicting ICU ADMISSION . All of the available feature families are the reported

Table 3.13: Recap table with the performance of the various families of features

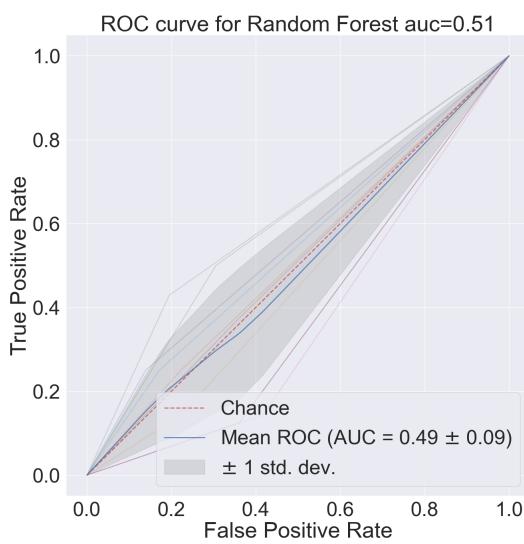
Features used	mean AUC \pm std
Radiomic	0.62 ± 0.08
Clinical	0.56 ± 0.08
Radiological	0.51 ± 0.11
All	0.64 ± 0.09



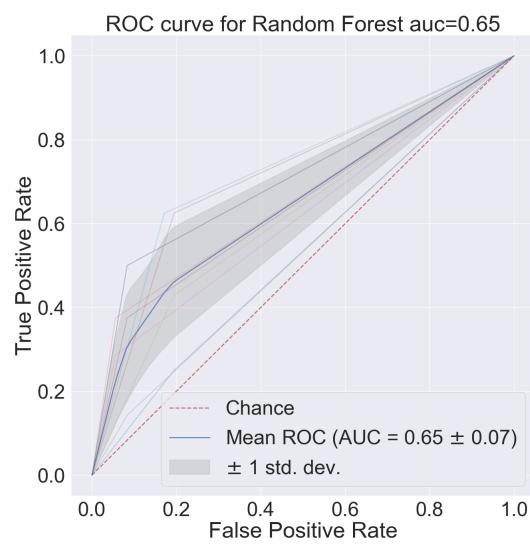
(a) Radiomic Features



(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 3.14: 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

1438 3.3 Using survival analysis

1439 Following the preprocessing steps delineated in Materials and Methodologies, a Cox
 1440 Proportional-Hazard produces the results presented in Table 3.14.

1441 As explained in section 1.5 the relevant columns are the coeff column, that expressed
 1442 percentual difference of survival, and the p column, that indicate the significance of the
 1443 first value. It turns out that, out of the reduced variables fed to the Cox model, the
 1444 most relevant are: SEX, ASPHERICITY, FATSURFACE, NORMALIZED ZONE DISTANCE
 1445 NON-UNIFORMITY AND KVP.

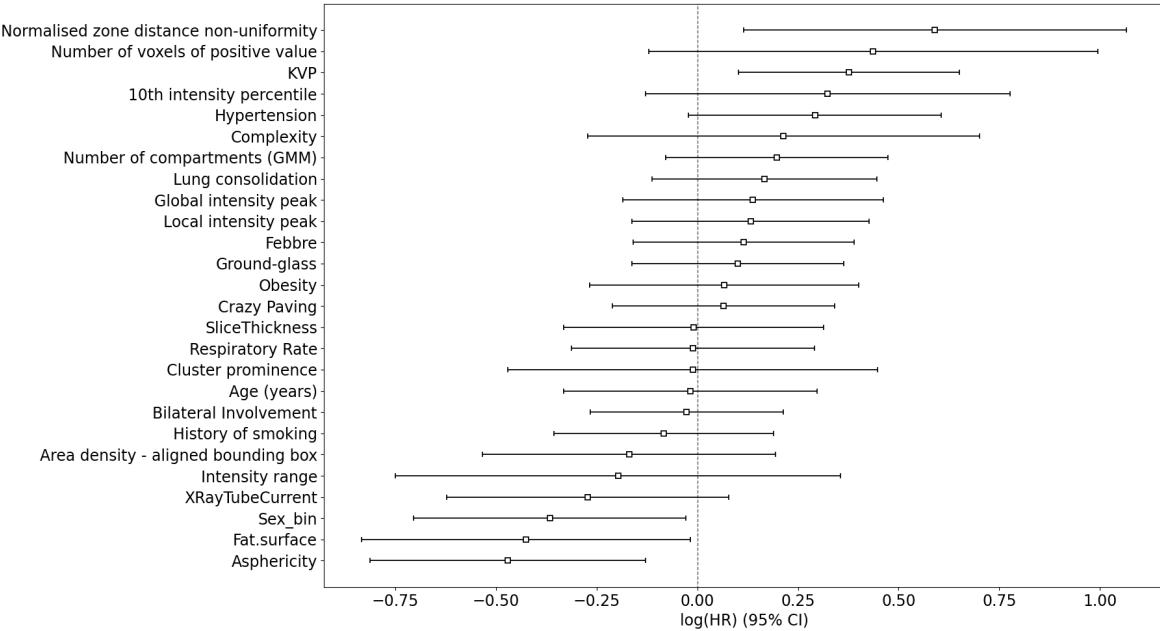
1446 ZONE DISTANCE NON UNIFORMITY measures distribution of zone counts over the dif-
 1447 ferent zone distances, it is low when the count relative to the zones are equally distributed
 1448 along zone distances

1449 ASPHERICITY quantifies how much the segmented region deviates from a sphere.

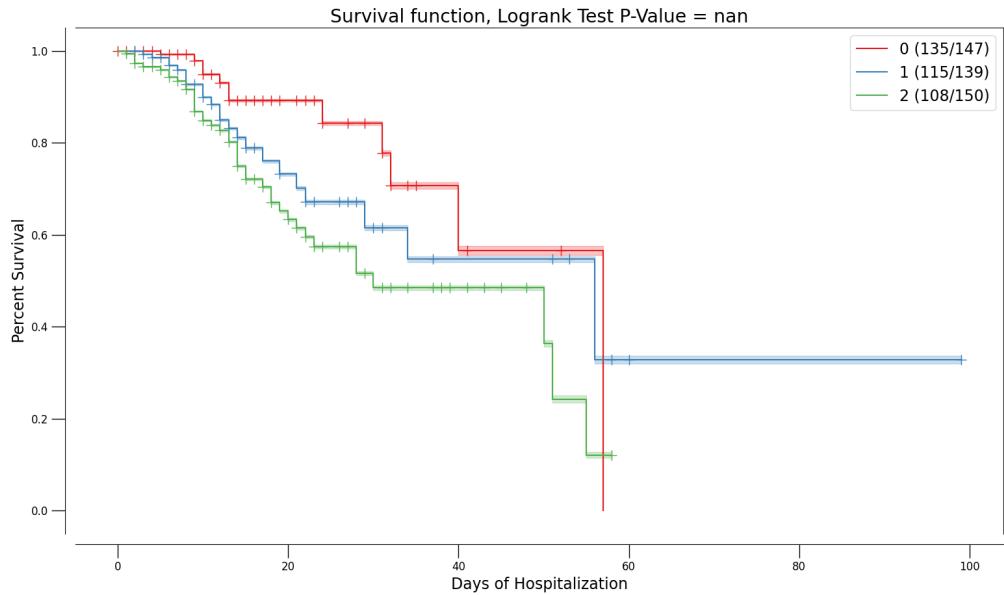
1450 In this caseSEX and FATSURFACE and NORMALIZED ZONE DISTANCE NON-UNIFORMITY
 1451 can be reasonable variables to expect, however KVP, ASPHERICITY seem quite strange.

1452 A score was built automatically using the predict method of the CoxPH fitter and
 1453 assigned to each patient of the dataset using the previously described cross-validated
 1454 prediction procedure. To see if the prediction was representative of differences in the
 1455 individuated populations first the Kaplan-Meier curves according to thirds in the score

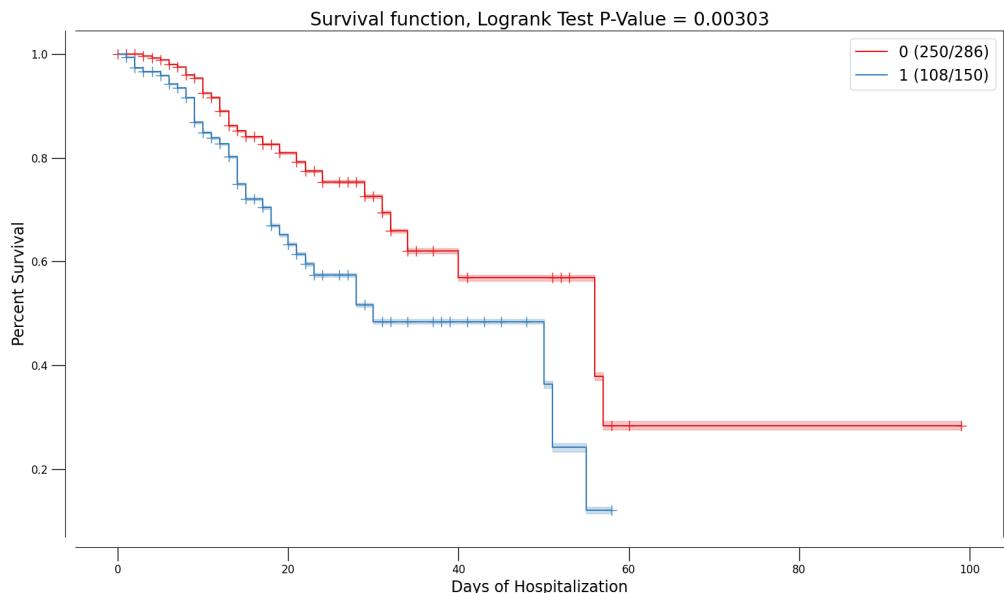
Figure 3.14: Graph that represents the coefficient values estimated by the CoxPH model with their respective 95% confidence intervals



1456 distribution were used and then the score was binarized using the 66th percentile in the
 1457 score distribution as threshold. The results of this procedures are reported in Figure
 1458 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 populations 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

1459 It can be seen that the groups built in this ways can be used to drive some differences
 1460 in survival, when binarizing the score obtained with Cox the curves also turn out to be
 1461 significantly different.

1462 Finally, in order to see if there were differences in treatment or in survival between
 1463 the two waves of admission, the population was divided in two subgroups according to

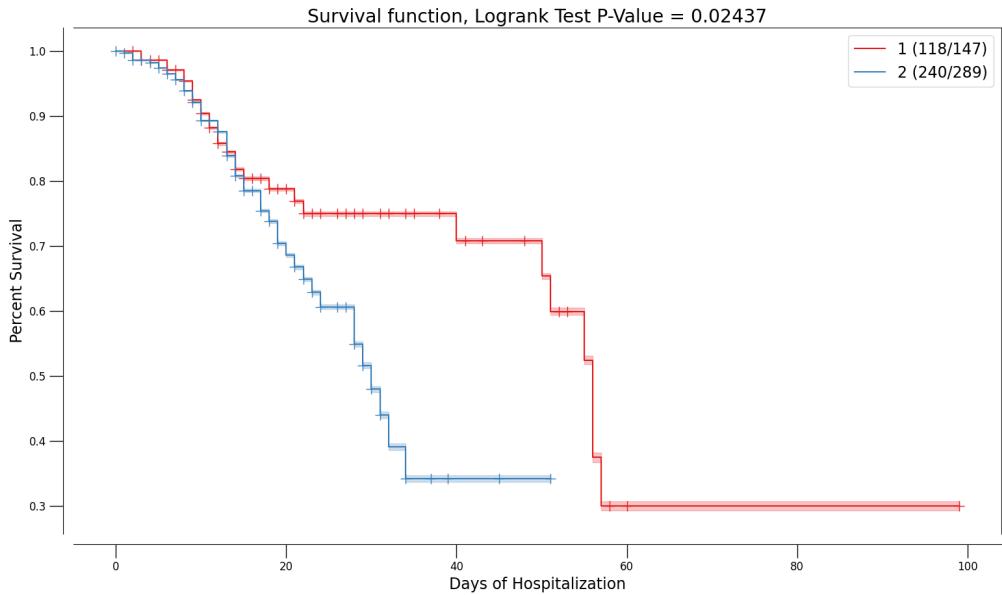


Figure 3.16: Kaplan-Meier curves for patient admitted before (red curve) and after (blue curve) 20/07/2020

1464 the date of admission. The two groups were representatively called 1st and 2nd wave and
 1465 the division was drawn on the 20th of July 2020 and the results can be seen in Figure
 1466 3.16.

1467 It can be seen that there is a statistical difference in survival between the patients
 1468 admitted in the first wave vs those admitted in the second. Furthermore this difference
 1469 is quite perplexing as it seems to indicate that people in the second wave died more
 1470 than people in the first wave, which seems counter-intuitive given that one would expect
 1471 the experience from the previous wave to improve performance. The most reasonable
 1472 explanation for this fact is the change in admission policy as time advanced. Probably in
 1473 the first wave, when still little was known on *Sars-COVID19* patients, more people were
 1474 addmitted in less problematic condition whereas in the second wave, having understood
 1475 better what were the most dangerous cases as well as in an attempt to admitt only those
 1476 strictly in need, most of the admitted patients were in more critical condition.

1477 It's also possible that this result that has been obtained could be a symptom of subtle
 1478 differences in the two *Sars-COVID19* manifestations, as if to indicate different variants.
 1479 Further analysis in this direction could be a followup work of this thesis.

₁₄₈₀ **Chapter 4**

₁₄₈₁ **Discussion**

₁₄₈₂ 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.
- The performance of the models is the same for both clinical and radiomic features, with no statistical difference between the two. Since combining them does not provide any added value, at least in the context of this thesis, the two sets of variables could be considered almost equivalent. This result is quite perplexing and it could have multiple causes

₁₄₉₄ While it's an interesting result that radiomic features and clinical variables provide ₁₄₉₅ the same information, it's quite perplexing that combining does not produce improvements ₁₄₉₆ in the performance.

₁₄₉₇ The perplexity seem to arise from the tacit assumption of all 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 conducive to predicting the prognosis of the patient and is different from that conveyed by clinical variables.

₁₅₀₃ The first three hypotheses are verified with a caveat, the performance of any radiomic ₁₅₀₄ pipeline hinges on the quality of the images and of the segmentation procedure performed ₁₅₀₅ on them. Since the images used in this thesis were, are and will be used by the hospital ₁₅₀₆ in routine processes it's close to impossible that all of them have problems, especially ₁₅₀₇ because of the preliminary screening done before segmentation.

1508 That radiomics can extract useful quantities and that this information can be used
1509 for prognosis, specifically in a *Sars-COVID19* context, has been discussed in various
1510 papers such as [42], [22], [59], [60], [48], [50], [44] and [31].

1511 A first possibility is that when trying to segment lungs affected by *Sars-COVID19* the
1512 peculiar patterns developed in some way reduce the quality and quantity of information
1513 obtainable. In fact considering that the segmentation method used for this thesis relies
1514 on region growth and thresholding methods it's possible that the worst cases end up
1515 with unrepresentative segmentations.

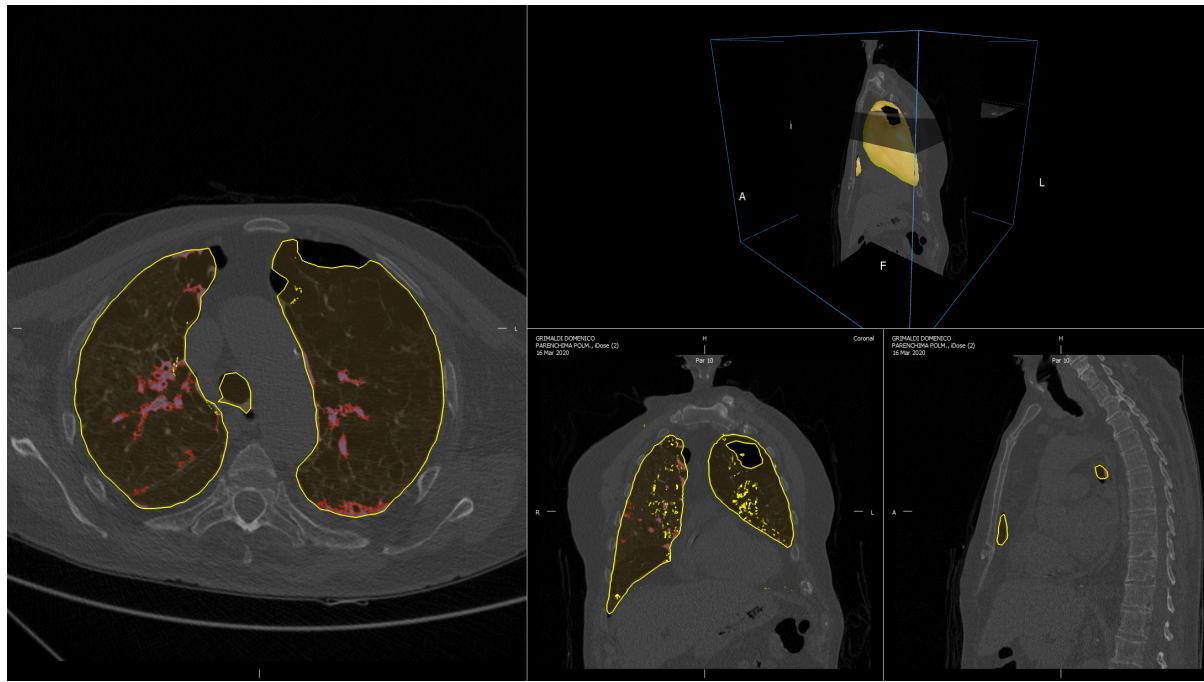
1516 Another possibility is that the images are not really representative of the situation
1517 of the patient. Since one of the prevailing properties of *Sars-COVID19* is the speed
1518 with which the clinical picture of the patient can change it's possible that images taken
1519 at admission are not as informative of the final prognosis. This problem is, however,
1520 unavoidable in the setting of this thesis which has as aim the construction of a model
1521 that, exactly at admission, can discriminate between serious and easier cases.

1522 Another possibility is that there are two or more subgroups in the patient cohort
1523 and the performance on these is widely different, determining an average performance
1524 below the expectations. To diagnose if this is the case a few dimensionality reduction
1525 techniques have been used to visualize the data and to prepare for clustering in case of
1526 need, all of the results of these procedure will be presented in the Appendix.

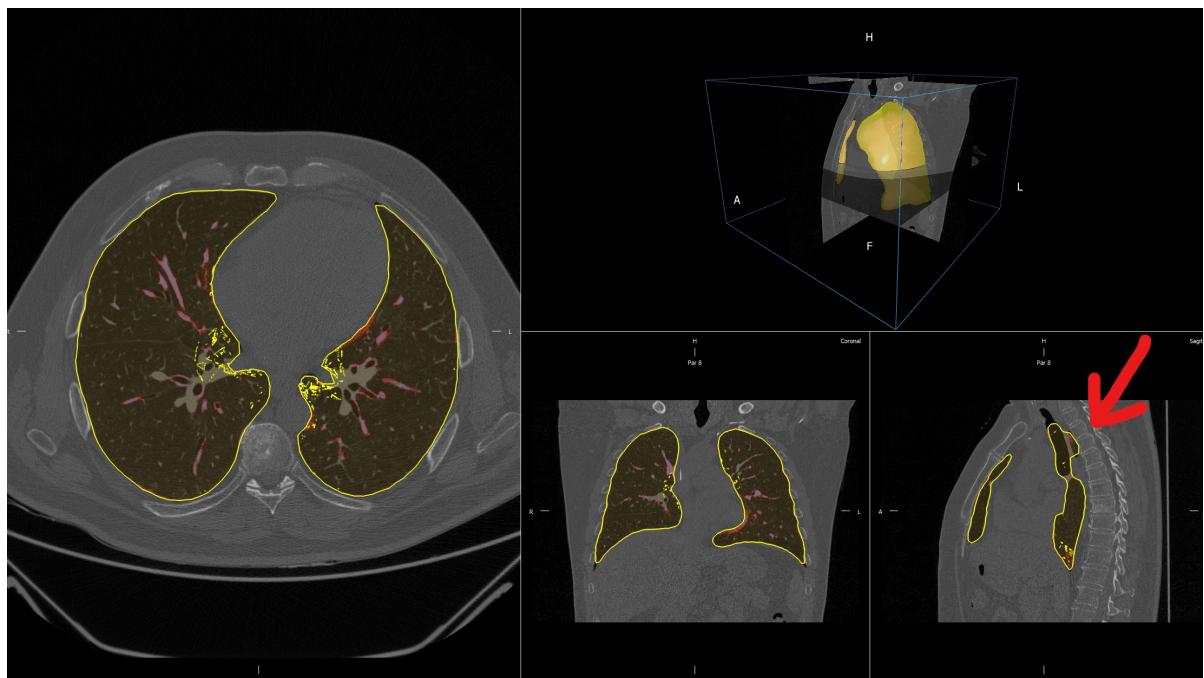
1527 To give a qualitative idea of the situation regarding the segmentations ~ 70 were
1528 looked at and evaluated as good, unsure or bad.¹

1529 Good segmentations are those in which an untrained professional could not see any
1530 problems, unsure are those in which there are small inaccuracies, such as lungs that
1531 connect in some points and the inclusion of the trachea. Finally segmentation were
1532 classified as bad in cases with obvious errors, such as part of the intestine being labelled
1533 as lung, damages in the lung being labelled as outside tissue or holes in what is supposed
1534 to be lung.

¹This 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.

1535 The result of this qualitative analysis can be seen in Table 4.1 in which the incidence
 1536 of both DEATH and ICU ADMISSION labels is computed in all possible segmentation
 1537 categories.

1538 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	Death	Segmentation Status	Subject		
			Bad	Good	Unsure
0	0		8	11	24
	1		5	0	6
1	0		2	0	2
	1		1	1	1

1539 end up being not perfectly segmented. It's possible, even if unlikely, that the sample of
 1540 analysed segmentations is biased in some way. However, given the numbers in Table 4.1,
 1541 it can be reasonable to think that in some of the most important cases, which means
 1542 those that correspond to 1 labels in DEATH or ICU ADMISSION , the values of the
 1543 radiomic features used in this thesis were not the best possible.

1544 Another thing that can be noted is that in all models that included radiomics feature
 1545 there are, among the important variables, features that quantify volumetric or shape
 1546 information regarding the lung.

1547 It's possible that some of the most severe cases, which may need a more delicate and
 1548 ad hoc handling during segmentation, end up being associated to lung shapes that are
 1549 somewhat different from almost healthy patients. In this light it's possible that the model
 1550 is using these shape information, which is in some way altered with respect to reality,
 1551 to infer the gravity of the situation and is finding something because the information is
 1552 somewhat unrealistic in the most severe cases.

1553 Chapter 5

1554 Conclusion

1555 In this thesis various methods were used in an attempt to predict the prognosis of *Sars-COVID19* patients.

1557 Regularized regression was used to predict clinical outcome using various families of
1558 variables to compare the information hidden in each of them and to evaluate if CT exams
1559 add any value to a small set of clinical variables.

1560 Random forest classifiers were used with the same aim. As a by-product these two
1561 methods can be compared to see which, given the same data, can extract the most
1562 information.

1563 Survival analysis was used, mostly by itself, to see if the data could be divided in
1564 smaller groups with different survival functions.

1565 In the first two lines of developement it was found that, in the specific case of data
1566 available for this thesis, models built on radiomic features perform in a statistically
1567 equivalent way to models built from clinical variables.

1568 This means that, especially in times of system overload due to increased accesses
1569 during a pandemic, this system could be integrated in PACS systems of the hospital to
1570 bring to attention some of the patients in worst condition. This, especially thanks to the
1571 objectivity of radiomics, the user-independence and, most of all, to the semi-automatic
1572 nature of the pipeline could be done on a large scale to aid in hospitals that don't have
1573 the facilities , or that lack the personnel, to analyze in detail all cases. It has also been
1574 hypothesised that a more careful handling of images during the segmentation phase,
1575 perhaps obtainable with instruments specifically developed for *Sars-COVID19* , could
1576 lead to imporvements in performances of the model.

1577 Regarding survival analysis it was found that the cohort can be divided in parts using
1578 the hazard predicted by a Cox Proportional Hazard model, and it was also found that
1579 a time-division driven by the "wave" definition of the pandemic produces signiicanly
1580 different survival curves.

1581 All of that said, some directions in which future works could start from this thesis
1582 are:

- 1583 • The implementation of a *Sars-COVID19* specific segmentation method using the
1584 vast amount of available data, similar to what has been done in [7].
- 1585 • It would be interesting to test the pipeline, as well as the deriving models, on
1586 manually segmented CT scans to compare the results and performances.

- 1587 • Given the statistical difference found in the survival of patients in the first and sec-
1588 ond waves, defined here as before and after 20/07/2020, it might be very interesting
1589 to investigate the causes of this finding and to prospectively continue this analysis
1590 with the data from the third wave and eventual next ones that might occur.
- 1591 • Using CT scans acquired at different points in the course of the illness it would
1592 be very interesting to implement a variation of the pipeline along the direction of
1593 delta-radiomics.

₁₅₉₄ **Chapter 6**

₁₅₉₅ **Appendix**

₁₅₉₆ **6.1 Additional Results and complete tables relative
to Random Forest**

₁₅₉₈ Here are the tables with all of the features from random forest. There is no real utility
₁₅₉₉ in providing them for all possible feature combination, hence only the importances for
₁₆₀₀ all features will be given. This will be done for both labels used, i.e. Death and ICU
₁₆₀₁ 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

1604 For these analyses the data was always fed into a standard scaler before applying the
1605 technique of choice, furthermore a custom gravity score by classifying as 4 the dead
1606 individuals and then by assigning a progressive score from 1 to 3 by looking at the time
1607 of permanence was built as follows:

- 1608 1. Gravity 1: Survived individuals with permanence from 0th percentile to 25th per-

1609 centile

1610 2. Gravity 2:Survived individuals with permanence from 25th percentile to 75th per-

1611 centile

1612 3. Gravity 3:Survived individuals with permanence from 75th percentile to 100th per-

1613 centile

1614 4. Gravity 4: Dead individuals without regard for permanence in the hospital

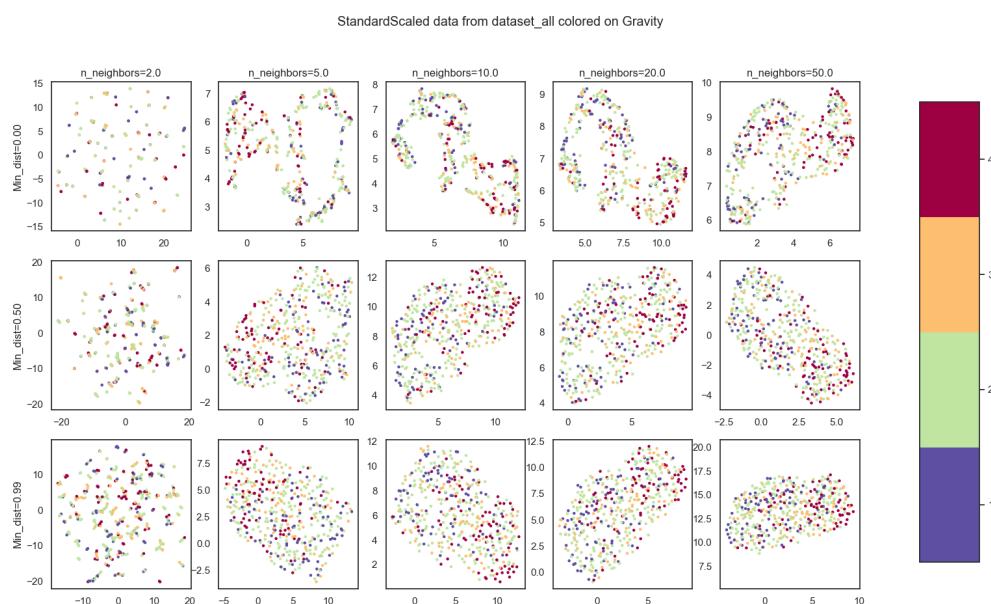


Figure 6.1: Possible combination for umap hyperparameters "number of neighbours" and "minimum distance". Color coding is done with aforementioned gravity score and all the features, i.e. clinical radiomic and radiological, were used.

Also, before proceeding, 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

1621 6.2.1 Explaining total variance using PCA

1622 Starting from PCA, the data was reduced to either two or three dimensions considering
1623 clinical and radiomic features, both separated and together. In this first example there
1624 seems to be a kind of left leaning polarization of the dead individuals, however there
1625 are no clear separations in the data.

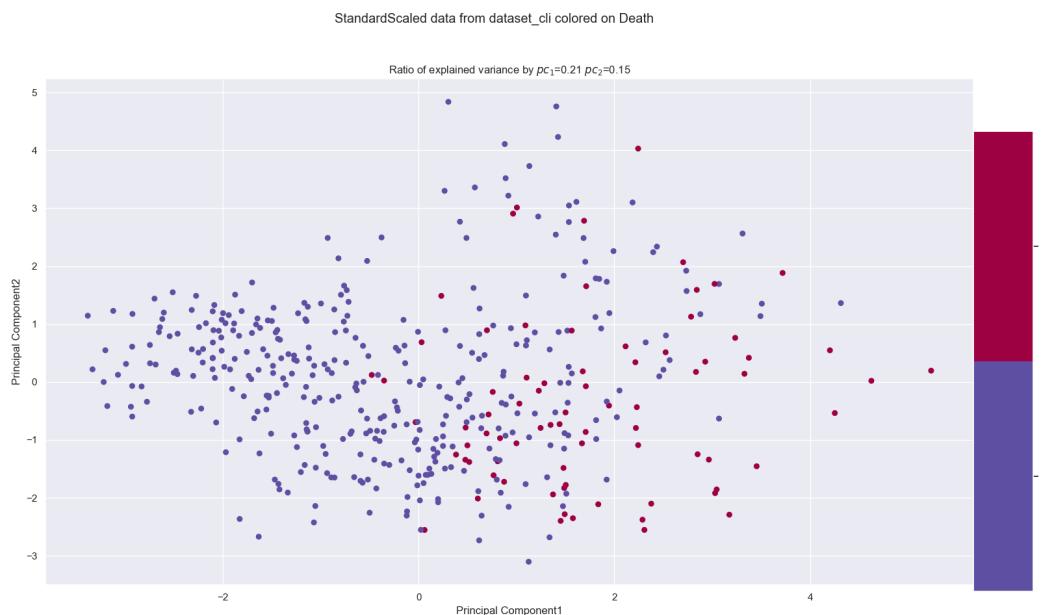


Figure 6.2: 2-Principal Component on clinical features.

1626 Working on the clinical dataset it can also be noted that the first two components of
1627 the PCA explain only 36% of the total variance. This leads to the conclusion that
1628 changes in the data cannot be explained by a single, nor a few, features or linear
1629 combination thereof. The next approach was using the first three principal components
1630 using various labels available, most relevant of which being ICU admission, Death and
1631 Gravity score.

1632 In all cases it seems like introducing the radiomic features causes the loss of the
1633 polarization structure that could be seen in the PCA on the clinical dataset alone in
1634 fig6.2.

1635 Since there are no visible clusters proceeding with cluster analysis would mean
1636 incurring in the risk of finding non meaningful results so it seemed appropriate to try
1637 other dimensionality reduction techniques.

1638 6.2.2 Exploring data structure with UMAP

1639 The next technique tried was unsupervised Umap. Following the conclusions derived
1640 from fig:6.1 the number of neighbours was set to 10 and the minimum distance was set
1641 to 0. Once again the comparison were made between clinical and radiomic dataset as
1642 well as different possible labellings. Starting from the clinical dataset, without
1643 reporting all labels used, it's clear to see that the dataset seems to indicate very local
1644 well separated structures which don't seem correlated to gravity outcome

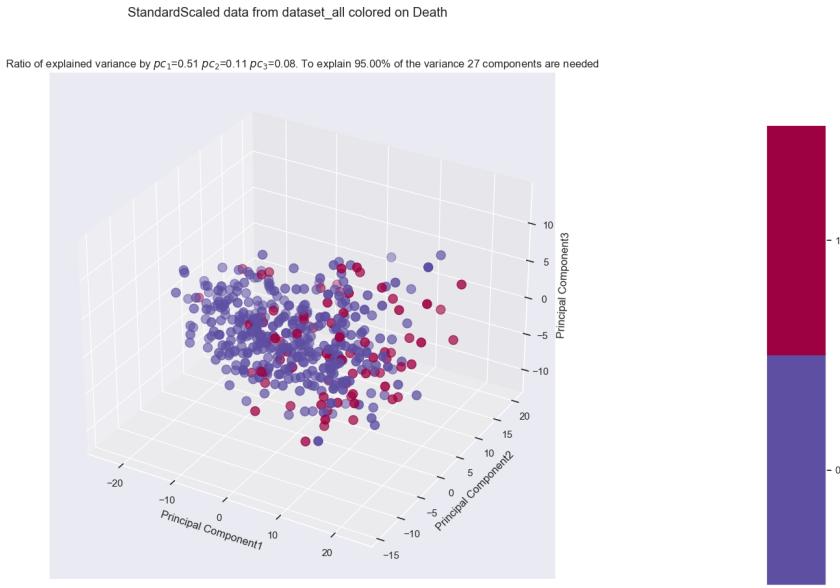


Figure 6.3: 3D PCA of whole dataset, colored with death label

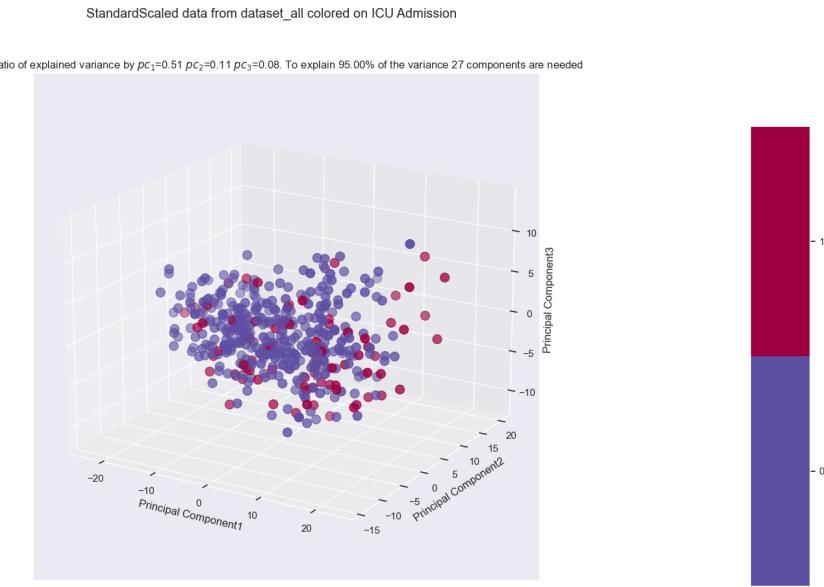


Figure 6.4: 3D PCA of whole dataset, colored with ICU Admission label

1645 There are 9 well defined groups which don't seem to be correlated to any of the
 1646 available labels. The dimension of these group is also very prohibitive if thinking of
 1647 further analyses since groups of 35-50 people in a dataset with 15% mortality rate
 1648 would mostly be very unbalanced if they were to be used for classification. However if
 1649 the introduction of radiomic features were to unite some of these groups then this
 1650 embedding could be meaningfully used for analysis. Looking at the 3D embedding for
 1651 the whole dataset, the results are:

1652 Once again the introduction of the radiomic feature seems to be a confounding factor

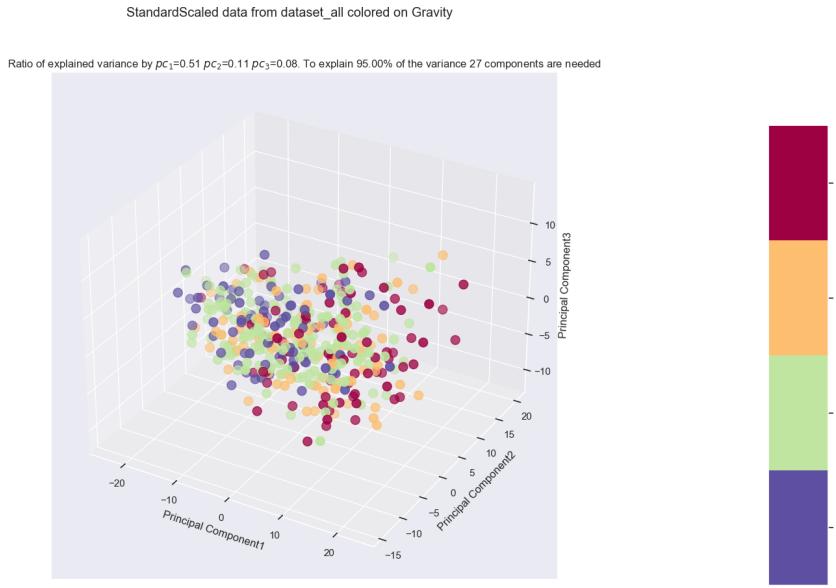


Figure 6.5: Comparison between various colour labels of the top 3 principal components for the entire dataset. Note that to explain 95% of the variance 27 components would be needed

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

StandardScaled data from dataset_cli colored on Gravity
Umap parameters: N_Neighbours = 10, Min_distance =0, Distance metric=euclidean

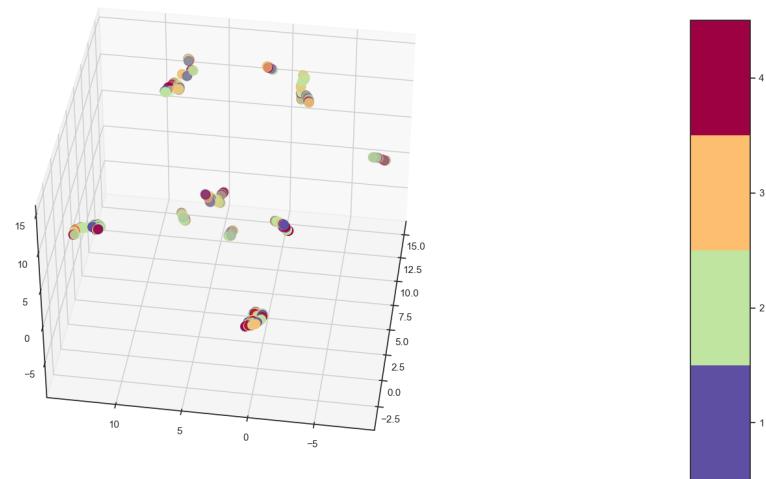


Figure 6.6: 3D umap of clinical dataset, colored based on gravity

StandardScaled data from dataset_all colored on Death
Umap parameters: N_Neighbours = 10, Min_distance =0, Distance metric=euclidean

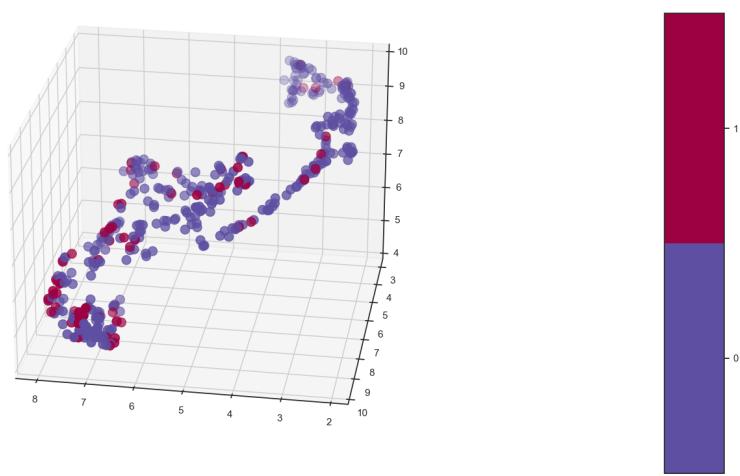


Figure 6.7: 3D umap of whole dataset, colored based on death

StandardScaled data from dataset_all colored on ICU Admission

Umap parameters: N_Neighbours = 10, Min_distance =0, Distance metric=euclidean

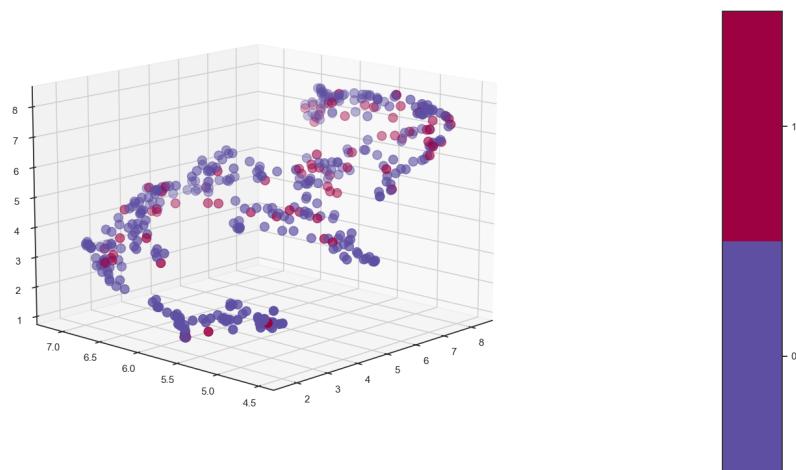


Figure 6.8: 3D umap of whole dataset, colored based on ICU admission

StandardScaled data from dataset_all colored on Gravity

Umap parameters: N_Neighbours = 10, Min_distance =0, Distance metric=euclidean

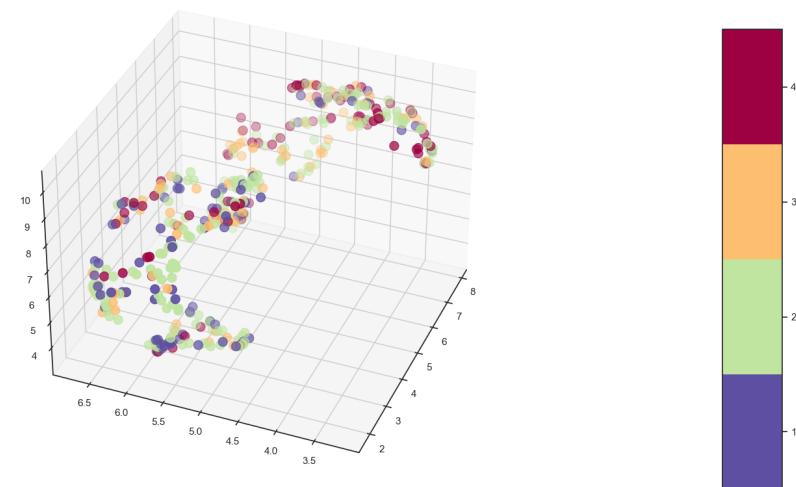


Figure 6.9: 3D umap of whole dataset, colored based on gravity

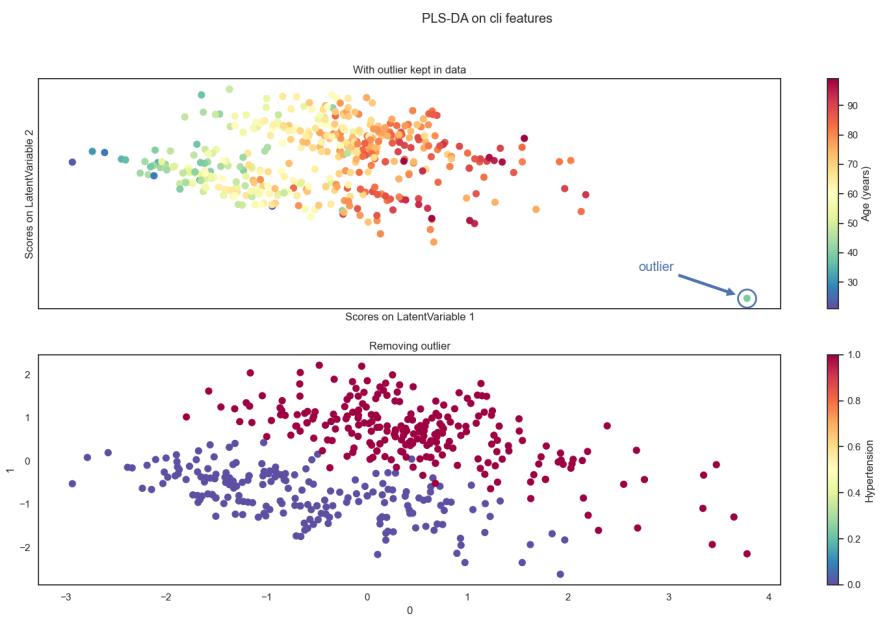


Figure 6.10: PLS-DA predicting on death coloured with age and hypertension

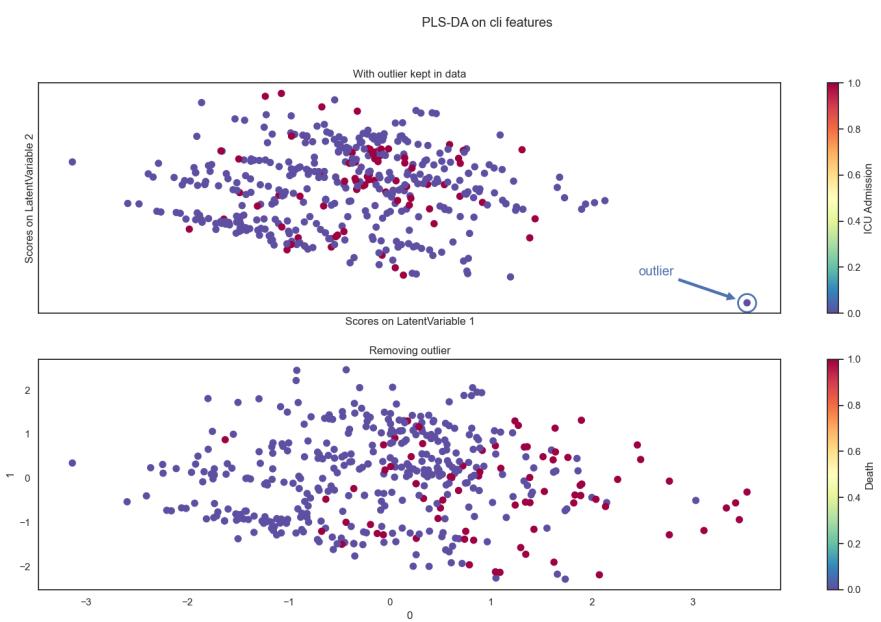


Figure 6.11: PLS-DA predicting on death coloured with death(bottom) and ICU Admission(top)

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

1676 by looking at how the ages are distributed in the first plot of fig: 6.10. From this it's
 1677 possible to deduce that older individuals tend to die more and that hypertension does
 1678 not seem to be relevant when considering death as a clinical outcome. If necessary the
 1679 PLS-DA algorithm allows also to see the weights given to the features in predicting the
 1680 label. At least for the clinical dataset, which has a reasonable number of features, it's
 1681 interesting to report it ordering the coefficients by descending absolute value:
 1682 Doing the exact same procedure on the whole dataset, which means by including the
 1683 radiomic features, Figure 6.12 can be obtained.

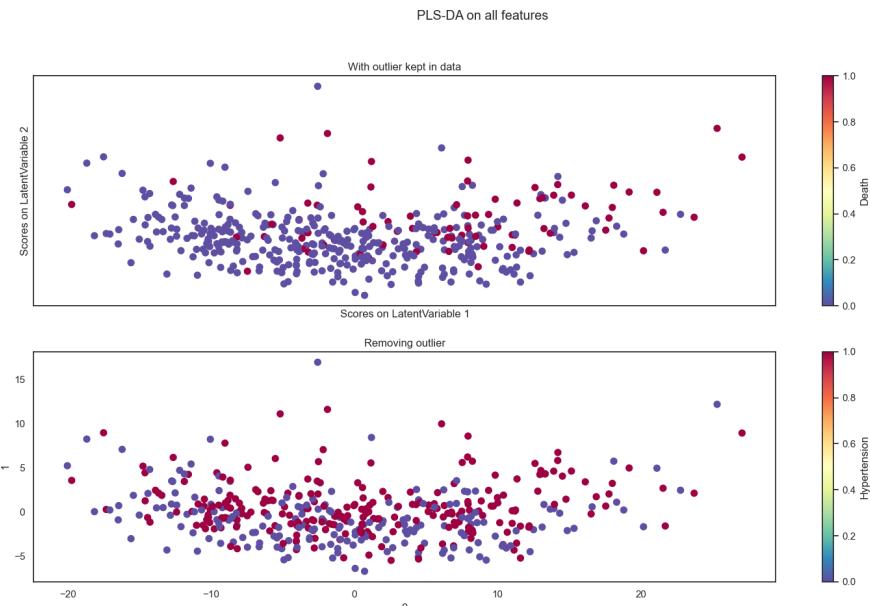


Figure 6.12: PLS-DA predicting on death coloured with death (top) and hypertension (bottom) on whole dataset

1684 Once again adding the radiomic features has evidently introduced noise in the system,
 1685 which no longer displays any kind of behaviour, pattern nor separation. Doing the
 1686 same analysis but using ICU Admission as a label Figure 6.13 can be obtained.
 1687 Now the colors have been chosen to highlight that respiratory rate and age have the
 1688 main role in determining the latent variables. However, in this case, there doesn't
 1689 appear to be a clear cut distinction as it happened before with hypertension. Looking
 1690 at how the points scatter by coloring them according to the two interesting clinical

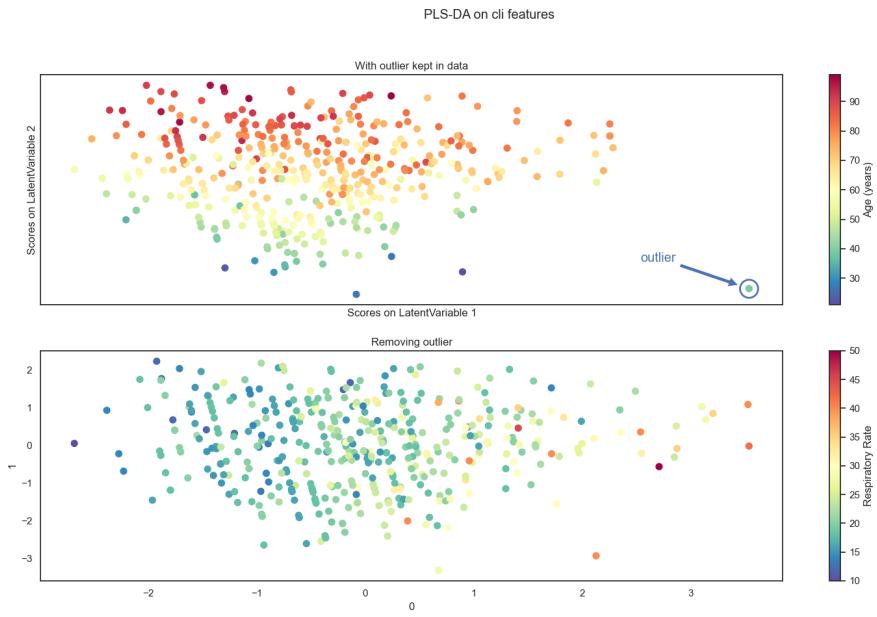


Figure 6.13: PLS-DA predicting on death coloured with Age and respiratory rate on clinical features

1691

labels the following figure can be obtained:

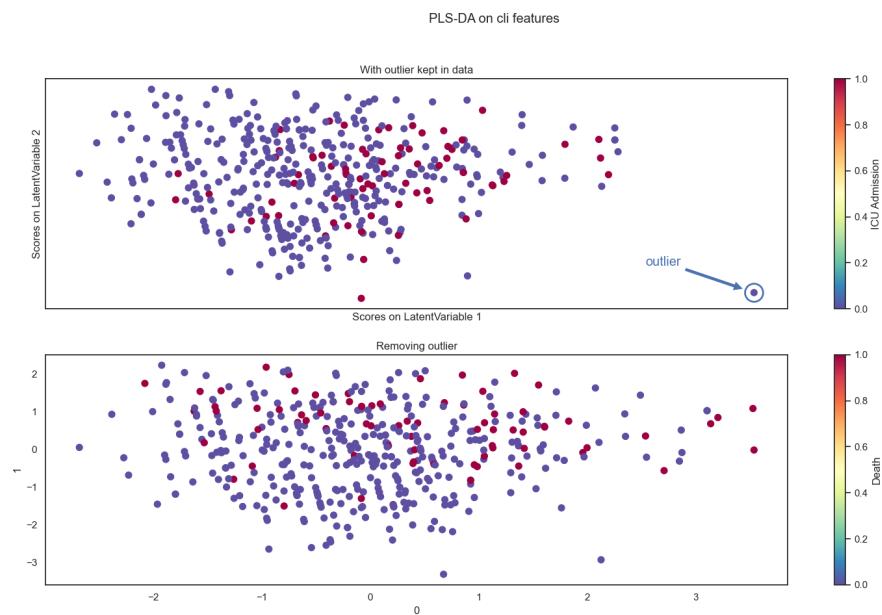


Figure 6.14: PLS-DA predicting on death(bottom) coloured with death and ICU admission(top) on clinical features

1692

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

1693

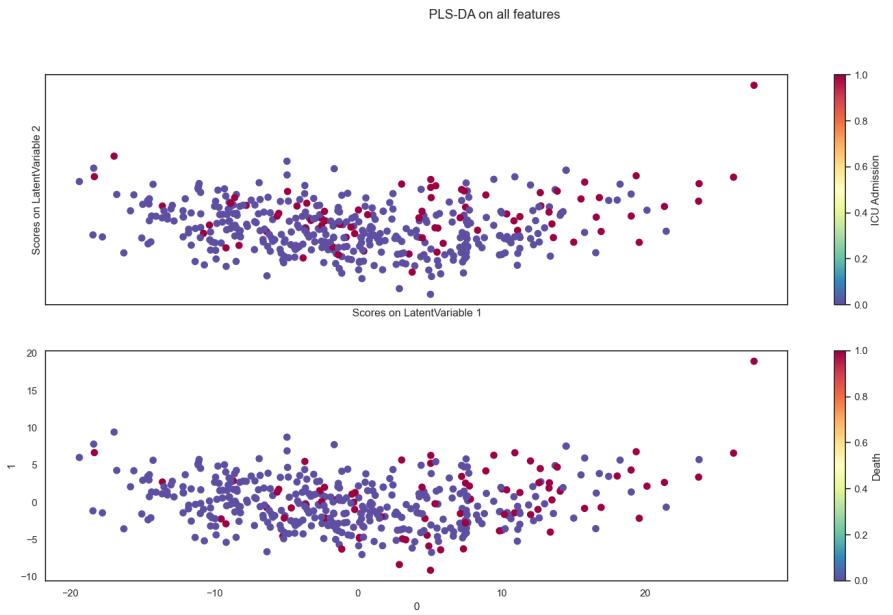


Figure 6.15: PLS-DA predicting on death coloured with death(bottom) and ICU admission(top) on all available features

¹⁶⁹⁴ Bibliography

- ¹⁶⁹⁵ [1] Sophia ddm for radiomics. <https://www.sophiagenetics.com/technology/sophia-ddm-for-radiomics/>. Accessed: 26/08/2021.
- ¹⁶⁹⁶
- ¹⁶⁹⁷ [2] Nema ps3 / iso 12052, digital imaging and communications in medicine (dicom) standard, national electrical manufacturers association, rosslyn, va, usa. available free at, 2021.
- ¹⁶⁹⁸
- ¹⁶⁹⁹
- ¹⁷⁰⁰ [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.
- ¹⁷⁰¹
- ¹⁷⁰²
- ¹⁷⁰³
- ¹⁷⁰⁴
- ¹⁷⁰⁵ [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.
- ¹⁷⁰⁶
- ¹⁷⁰⁷
- ¹⁷⁰⁸ [5] M. Barker and W. Rayens. Partial least squares for discrimination, journal of chemometrics. *Journal of Chemometrics*, 17:166 – 173, 03 2003.
- ¹⁷⁰⁹
- ¹⁷¹⁰ [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.
- ¹⁷¹¹
- ¹⁷¹²
- ¹⁷¹³
- ¹⁷¹⁴ [7] R. Biondi, N. Curti, F. Coppola, E. Giampieri, G. Vara, M. Bartoletti, A. Catatabriga, 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.
- ¹⁷¹⁵
- ¹⁷¹⁶
- ¹⁷¹⁷
- ¹⁷¹⁸
- ¹⁷¹⁹ [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.
- ¹⁷²⁰
- ¹⁷²¹
- ¹⁷²² [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.
- ¹⁷²³
- ¹⁷²⁴
- ¹⁷²⁵ [10] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- ¹⁷²⁶

- 1727 [11] T. Daimon. *Box-Cox Transformation*, pages 176–178. Springer Berlin Heidelberg,
 1728 Berlin, Heidelberg, 2011.
- 1729 [12] C. Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source
 1730 Software*, 4(40):1317, 2019.
- 1731 [13] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under
 1732 two or more correlated receiver operating characteristic curves: A nonparametric
 1733 approach. *Biometrics*, 44(3):837–845, 1988.
- 1734 [14] K. P. F.R.S. Liii. on lines and planes of closest fit to systems of points in space.
 1735 *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*,
 1736 2(11):559–572, 1901.
- 1737 [15] D. G. George H. Joblove. Color spaces for computer graphics. *ACM SIGGRAPH
 1738 Computer Graphics*, (12(3), 20–25), 1978.
- 1739 [16] L. Guo. Clinical features predicting mortality risk in patients with viral pneumonia:
 1740 The mulbst score. *Frontiers in Microbiology*, 2019.
- 1741 [17] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cour-
 1742 napeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer,
 1743 M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson,
 1744 P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke,
 1745 and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362,
 1746 Sept. 2020.
- 1747 [18] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*.
 1748 Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.
- 1749 [19] T. K. Ho. Random decision forests. In *Proceedings of 3rd international conference
 1750 on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.
- 1751 [20] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthog-
 1752 onal problems. *Technometrics*, 42(1):80–86, 2000.
- 1753 [21] G. N. Hounsfield. Computed medical imaging. nobel lecture. *Journal of Computer
 1754 Assisted Tomography*, (4(5):665-74), 1980.
- 1755 [22] Y. Huang, Z. Zhang, S. Liu, X. Li, Y. Yang, J. Ma, Z. Li, J. Zhou, Y. Jiang, and
 1756 B. He. Ct-based radiomics combined with signs: a valuable tool to help radiologist
 1757 discriminate covid-19 and influenza pneumonia. *BMC Medical Imaging*, 21, 02 2021.
- 1758 [23] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science &
 1759 Engineering*, 9(3):90–95, 2007.
- 1760 [24] L. M. J. Cine computerized tomography. *The International Journal of Cardiac
 1761 Imaging*, 1987.
- 1762 [25] A. Kaka and M. Slaney. *Principles of Computerized Tomographic Imaging*. Society
 1763 of Industrial and Applied Mathematics, 2001.

- 1764 [26] J. Kirby. Mosmeddata: dataset, 09/2021.
- 1765 [27] D. Kleinbaum and M. Klein. *Survival Analysis: A Self-Learning Text*. Statistics for
1766 Biology and Health. Springer New York, 2006.
- 1767 [28] P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. de Jong, J. van
1768 Timmeren, S. Sanduleanu, R. T. H. M. Larue, A. J. G. Even, A. Jochems, Y. van
1769 Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker,
1770 F. M. Mottaghy, J. E. Wildberger, and S. Walsh. Radiomics: the bridge between
1771 medical imaging and personalized medicine. *Nature reviews. Clinical oncology*,
1772 14(12):749—762, December 2017.
- 1773 [29] L. Lee and C.-Y. Liong. Partial least squares-discriminant analysis (pls-da) for classi-
1774 fication of high-dimensional (hd) data: a review of contemporary practice strategies
1775 and knowledge gaps. *The Analyst*, 143, 06 2018.
- 1776 [30] G. Lemaître, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A python toolbox
1777 to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine
1778 Learning Research*, 18(17):1–5, 2017.
- 1779 [31] H. Liu, H. Ren, Z. Wu, H. Xu, S. Zhang, J. Li, L. Hou, R. Chi, H. Zheng, Y. Chen,
1780 S. Duan, H. Li, Z. Xie, and D. Wang. Ct radiomics facilitates more accurate di-
1781 agnosis of covid-19 pneumonia: compared with co-rads. *Journal of Translational
1782 Medicine*, 19, 01 2021.
- 1783 [32] J. McCarthy. What is artificial intelligence? 01 2004.
- 1784 [33] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous
1785 activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.
- 1786 [34] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and
1787 projection for dimension reduction, 2020.
- 1788 [35] W. McKinney et al. Data structures for statistical computing in python. In *Pro-
1789 ceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin,
1790 TX, 2010.
- 1791 [36] S. J. McMahon. The linear quadratic model: usage, interpretation and challenges.
1792 *Physics in medicine and biology*, 2018.
- 1793 [37] S. Morozov. Mosmeddata: Chest ct scans with covid-19 related findings dataset.
1794 *IAU Symp.*, (S227), 2020.
- 1795 [38] MosMed. Mosmeddata: dataset, 28/04/2020.
- 1796 [39] Y. Ohno. Cie fundamentals for color measurements. *International Conference on
1797 Digital Printing Technologies*, 01 2000.
- 1798 [40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel,
1799 M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos,
1800 D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine
1801 learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

- 1802 [41] D. L. Pham, C. Xu, and J. L. Prince. Current methods in medical image seg-
1803mentation. *Annual Review of Biomedical Engineering*, 2(1):315–337, 2000. PMID:
1804 11701515.
- 1805 [42] J. Qiu, S. Peng, J. Yin, J. Wang, J. Jiang, Z. Li, H. Song, and W. Zhang. A
1806 radiomics signature to quantitatively analyze covid-19-infected pulmonary lesions.
1807 *Interdisciplinary Sciences: Computational Life Sciences*, 13, 01 2021.
- 1808 [43] P. C. Rajandeep Kaur. A review of image compression techniques. *International
1809 Journal of Computer Applications*, 2016.
- 1810 [44] S. Rezaeijo, R. Abedi-Firouzjah, M. Ghorvei, and S. Sarnameh. Screening of covid-
1811 19 based on the extracted radiomics features from chest ct images. *Journal of X-Ray
1812 Science and Technology*, 29:1–15, 02 2021.
- 1813 [45] W. C. Roentgen. On a new kind of rays. *Science*, 1986.
- 1814 [46] A. Saltelli. *Global Sensitivity Analysis. The Primer*. John Wiley and Sons, Ltd,
1815 2007.
- 1816 [47] S. Seabold and J. Perktold. statsmodels: Econometric and statistical modeling with
1817 python. In *9th Python in Science Conference*, 2010.
- 1818 [48] I. Shiri, M. Sorouri, P. Geramifar, M. Nazari, M. Abdollahi, Y. Salimi, B. Khosravi,
1819 D. Askari, L. Aghaghazvini, G. Hajianfar, A. Kasaeian, H. Abdollahi, H. Arabi,
1820 A. Rahmim, A. Radmard, and H. Zaidi. Machine learning-based prognostic mod-
1821eling using clinical data and quantitative radiomic features from chest ct images in
1822 covid-19 patients. *Computers in Biology and Medicine*, 132:104304, 03 2021.
- 1823 [49] C. P. Subbe. Validation of a modified early warning score in medical admissions.
1824 *QJM: An international journal of medicine*, 2001.
- 1825 [50] Z. Tang, W. Zhao, X. Xie, Z. Zhong, F. Shi, and J. Liu. Severity assessment of
1826 covid-19 using ct image features and laboratory indices. *Physics in medicine and
1827 biology*, 66, 10 2020.
- 1828 [51] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal
1829 Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
- 1830 [52] L. Tommy. Gray-level invariant haralick texture features. *PLOS ONE*, 2019.
- 1831 [53] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine
1832 Learning Research*, 9:2579–2605, 2008.
- 1833 [54] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau,
1834 E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett,
1835 J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson,
1836 C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perk-
1837 told, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald,
1838 A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy
1839 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*,
1840 17:261–272, 2020.

- 1841 [55] M. Waskom, O. Botvinnik, D. O’Kane, P. Hobson, S. Lukauskas, D. C. Gemperline,
1842 T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, J. de Ruiter, C. Pye,
1843 S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin,
1844 K. Meyer, A. Miles, Y. Ram, T. Yarkoni, M. L. Williams, C. Evans, C. Fitzgerald,
1845 Brian, C. Fonnesbeck, A. Lee, and A. Qalieh. mwaskom/seaborn: v0.8.1 (september
1846 2017), Sept. 2017.
- 1847 [56] S. Webb. The physics of medical imaging. 1988.
- 1848 [57] L. WS, van der Eerden MM, and e. a. Laing R. Defining community acquired
1849 pneumonia severity on presentation to hospital: an international derivation and
1850 validation study. *Thorax* 58(5):377-382, 2003.
- 1851 [58] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan,
1852 A. Ng, B. Liu, S. Y. Philip, et al. Top 10 algorithms in data mining. *Knowledge
1853 and information systems*, 14(1):1–37, 2008.
- 1854 [59] Z. Xie, H. Sun, J. Wang, H. Xu, S. Li, C. Zhao, Y. Gao, X. Wang, T. Zhao, S. Duan,
1855 C. Hu, and W. Ao. A novel ct-based radiomics in the distinction of severity of
1856 coronavirus disease 2019 (covid-19) pneumonia. *BMC Infectious Diseases*, 21, 06
1857 2021.
- 1858 [60] H. Yue, Q. Yu, C. Liu, Y. Huang, Z. Jiang, C. Shao, H. Zhang, B. Ma, Y.-C. Wang,
1859 G. Xie, H. Zhang, X. Li, N. Kang, X. Meng, S. Huang, D. Xu, J. Lei, H. Huang,
1860 J. Yang, and X. Qi. Machine learning-based ct radiomics method for predicting
1861 hospital stay in patients with pneumonia associated with sars-cov-2 infection: a
1862 multicenter study. *Annals of Translational Medicine*, 8:859–859, 07 2020.
- 1863 [61] H. Zou and T. Hastie. Regularization and variable selection via the elastic net.
1864 *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.
- 1865 [62] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. W. L. Aerts, V. Andreadczyk,
1866 A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, and et al. The image
1867 biomarker standardization initiative: Standardized quantitative radiomics for high-
1868 throughput image-based phenotyping. *Radiology*, 295(2):328–338, May 2020.