

**Alma Mater Studiorum · University of Bologna**

---

---

Department of Physics and Astronomy  
Master Degree in Physics

## **THESIS TITLE**

**Supervisor:**

**Prof. Enrico Giampieri**

**Co-Supervisor:**

**Dott.sa Lidia Strigari**

**Submitted by:**

**Lorenzo Spagnoli**

*Dedication...*

# Todo list

bisogna dividere i paragrafi, e consiglio una sola frase per riga. Dividere i paragrafi aiuta anche con il layout di tabelle e figure. FATTO COSÌ VA BENE O È MEGLIO DIVIDERE DI PIÙ' . . . . .	1
li sposterei in capitoli successivi . . . . .	1
perché hai scelto queste specifiche tecniche? a random? Spiegato così va bene? questa tabella richiede più dettagli di spiegazione. che cosa sono i numeri inseriti? la tabella dovrebbe poter essere compresa in modo completo senza troppi riferimenti al testo. . . . .	9
visto che usi la regolarizzazione in modo consistente spiegherei meglio i vantaggi dei vari approcci uno rispetto all'altro Lorenzo: Aggiunto un po' di discussione . . . . .	15
la spiegherei meglio. Lorenzo:Così va meglio? . . . . .	21
queste figure devono avere i numeri ed il testo più grandi, sono illeggibili	22
altrimenti; inoltre i risultati delle divisioni vanno inseriti dopo, non qui. . . . .	24
usare excel come immagine non è il massimo . . . . .	38
non mi piace la forma, troppo personale. Corretta così va meglio? . . . . .	38
Per dott.ssa Strigari: Metto il nome del programma? Va citato in qualche modo particolare? . . . . .	39
Ad esempio la questione della divisione in ondate . . . . .	45
A seconda di come viene decisa la struttura finale questa frase va cambiata e aggiunta la reference al punto preciso . . . . .	45
Non ricordo come mai avevamo fatto la distinzione in ondate . . . . .	66
Forse già questo può andare bene come conclusione? . . . . .	67
A questo punto magari cito l'appendice con la parte di clustering quando sarà pronta . . . . .	69

## Abstract

bisogna dividere i paragrafi, e consiglio una sola frase per riga. Dividere i paragrafi aiuta anche con il layout di tabelle e figure. FATTO COSÌ VA BENE O È MEGLIO DIVIDERE DI PIÙ'

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

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

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

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Medical Images . . . . .	3
1.1.1	X-ray imaging and Computed Tomography (CT) . . . . .	10
1.1.2	Generation and management of radiation: digital CT scanners . . . . .	11
1.1.3	Radiation-matter interaction: Attenuation in body and measurement . . . . .	15
1.2	Artificial Intelligence (AI) and Machine Learning(ML) . . . . .	17
1.2.1	Regression, Classification and Penalization . . . . .	19
1.2.2	Decision Trees and Random Forest . . . . .	22
1.2.3	Synthetic Minority Oversampling TEchnique (SMOTE) . . . . .	23
1.2.4	Dimensionality reduction and clustering . . . . .	25
1.3	Combining radiological images with AI: Image segmentation and Radiomics . . . . .	27
1.3.1	Image Segmentation . . . . .	28
1.3.2	Radiomics . . . . .	30
1.4	Survival Analysis . . . . .	33
1.4.1	Kaplan-Meier(KM) curves and log-rank test . . . . .	34
1.4.2	Cox Proportional-Hazard (CoxPH) model . . . . .	35
<b>2</b>	<b>Materials and methodologies</b>	<b>37</b>
2.1	Data and objective . . . . .	37
2.2	Preprocessing and data analysis . . . . .	41
<b>3</b>	<b>Results</b>	<b>47</b>
3.1	Feature selection through Lasso regularization and clinical outcomes prediction using regression . . . . .	47
3.1.1	Using DEATH as label . . . . .	48
3.1.2	Using ICU ADMISSION as label . . . . .	54
3.2	Classification of patients using Random forests . . . . .	58
3.2.1	Classifying the oucome Death . . . . .	58
3.2.2	Classifying the outcome ICU ADMISSION . . . . .	61
3.2.3	Using survival analysis . . . . .	64
3.3	Summarizing the results and further analysis . . . . .	67
<b>4</b>	<b>Conclusion</b>	<b>72</b>

<b>5 Appendix</b>	<b>73</b>
5.1 Additional Results and complete tables relative to Random Forest . . .	73
5.2 Using Dimensionality reduction to further investigate the dataset . . .	82
5.2.1 Explaining total variance using PCA . . . . .	83
5.2.2 Exploring data structure with UMAP . . . . .	83
5.2.3 Predicting clinical outcome using PLS-DA . . . . .	85
<b>Bibliography</b>	<b>92</b>

# <sup>1</sup> Chapter 1

## <sup>2</sup> Introduction

<sup>3</sup> Nowadays everybody knows of *Sars-COVID19* which, since the start of 2020, has  
<sup>4</sup> made necessary a few world-wide quarantines forcing everybody in self-isolation. It  
<sup>5</sup> is also well known that, among the main complications and features of this virus,  
<sup>6</sup> symptoms gravity as well as the rate of deterioration of the conditions are some  
<sup>7</sup> of the most relevant and problematic. In some cases asymptomatic or near to  
<sup>8</sup> asymptomatic people may, in the span of a week, get to conditions that require  
<sup>9</sup> hospital admission. This peculiarity is also what heavily complicates the triage  
<sup>10</sup> process, since trying to predict with some degree of accuracy the prognosis of the  
<sup>11</sup> patient at admission is a thoroughly complex task.

<sup>12</sup> In this thesis the aim will be to use data, specifically including data that cannot  
<sup>13</sup> be easily interpreted by humans, to try various methods to predict a couple of clinical  
<sup>14</sup> outcomes, namely the death of the patient or the admission in the Intensive Care  
<sup>15</sup> Unit (ICU), while assessing their performance.

<sup>16</sup> These analyses will be carried out on a dataset of 434 patients with different  
<sup>17</sup> variables associated to every person. A part of the variables, which will be called  
<sup>18</sup> clinical and radiological, are defined by humans and are generally discrete in nature  
<sup>19</sup> but mostly boolean. The most part of the available variables, however, will be image-  
<sup>20</sup> derived following the approaches used in the field of radiomics. While the utility of  
<sup>21</sup> clinical variables, such as age, obesity and history of smoking, is very straightforward  
<sup>22</sup> it's interesting and helpful to understand the basis behind the utility of radiomic  
<sup>23</sup> and radiological features.

<sup>24</sup> Generally speaking it's clear that images have the ability to convey a slew of use-  
<sup>25</sup> ful images, this is especially true in the medical field where digital images are used  
<sup>26</sup> to inspect also the internal state of the patient giving far more detailed informa-  
<sup>27</sup> tion than that obtainable by visual inspection at the hand of medical professionals.  
<sup>28</sup> Among the ways in which *Sars-COVID19* can manifest himself the one that is most  
<sup>29</sup> relevant to the scopes of this thesis pneumonia and the complications that stem  
<sup>30</sup> from it. Some of these complications, which are not specific of *Sars-COVID19* but  
<sup>31</sup> can happen in any pneumonia case, display very peculiar patterns when visualizing  
<sup>32</sup> the lungs through CT exams.

<sup>33</sup> These patterns are due to the pulmonary response to inflammation which may  
<sup>34</sup> lead to thickening of the bronchial and alveolar structures up to pleural effusions  
<sup>35</sup> and collapsed lungs. Without going too much in clinical detail what is of interest is  
<sup>36</sup> how these condition manifest themselves in the CT exams:

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

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

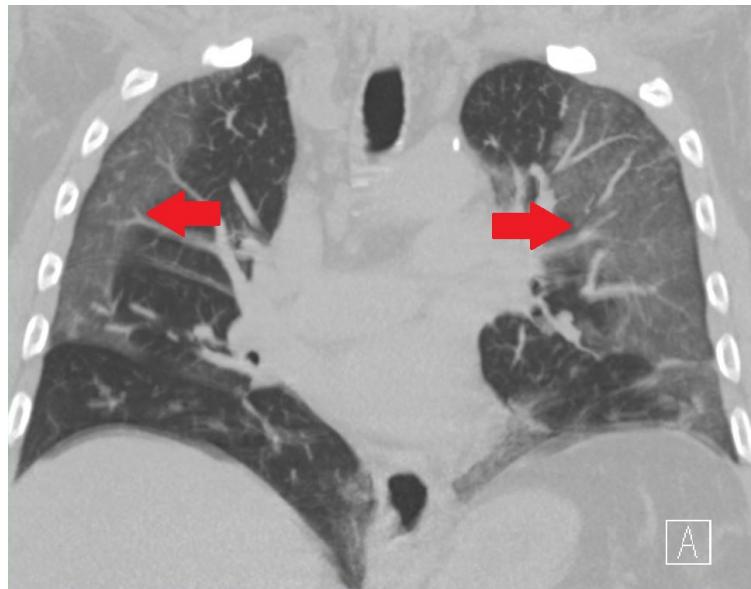


Figure 1.1: Example of GGO

41 2. **Lung Consolidations:**

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

46 3. **Crazy paving:**

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

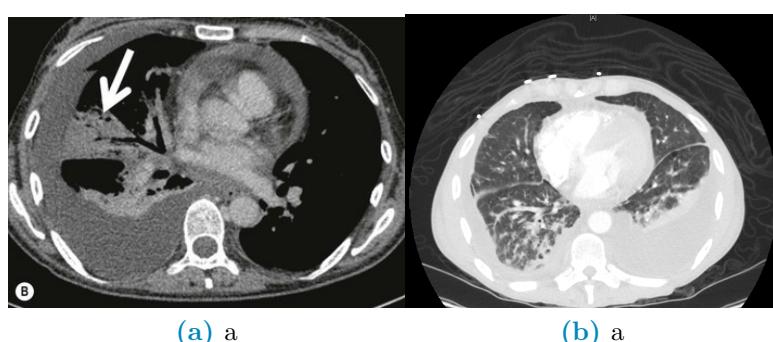


Figure 1.2: Differences between a collapsed lung (a) and pleural effusion(b)

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

Both of these manifest themself as regions of the lungs that take the same

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

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

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

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

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

## 1.1 Medical Images

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

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

---

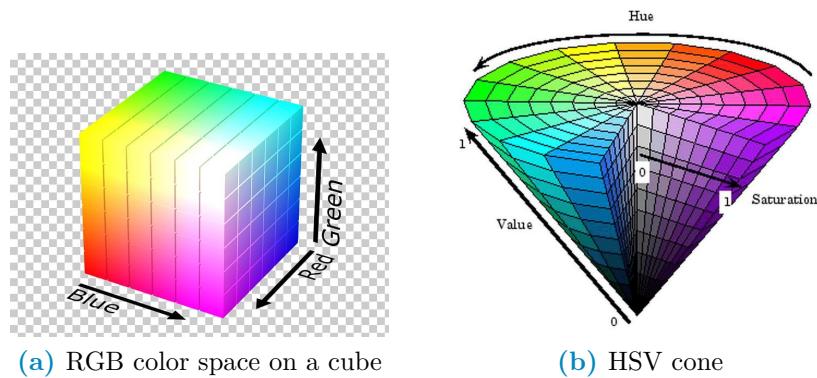
<sup>1</sup>The term pixel seems to originate from a shortening of the expression Picture's (pics=pix) Element(el). The same hold for voxel which stands for Volume Element

91 then stored a number (or set thereof) which measures the intensity of light, or color,  
92 coming from the physical space corresponding to that grid-spacing.

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

99 Generally pixel values stored as integers  $p \in [0, 2^n - 1]$  with  $p, n \in \mathbb{N}$  or as  $p \in [0, 1]$   
100 with  $p \in \mathbb{R}$ , the type of value stored within each pixel changes the nature of the  
101 image itself.

102 A single value is to be intended as the overall intensity of light coming from  
103 the part of the object contained corresponding to the gridspace and is used for a  
104 gray-scale representation, a set of three<sup>2</sup> or four<sup>3</sup> values can be intended as a color  
105 image.



**Figure 1.3:** Examples of color spaces

106 There are a lot of possible scales for representation<sup>4</sup>, which are sometimes called  
107 color-spaces, however the most noteworthy in the scope of this work is the Hounsfield  
108 unit (HU) scale.

109 **Definition 1.1.2** (Hounsfield unit (HU)). A scale used specifically to describe ra-  
110 diodensity, frequently used in the context of CT (Computed Tomography) exams.  
111 The values are obtained as a transformation of the linear attenuation coefficient ??  
112 of the material being imaged and, since the scale is supposed to be used on humans,  
113 it's defined such that water has value zero and air has the most negative value -1000.  
114 For a more in depth discussion refer to [13]

<sup>2</sup>The three values correspond each to the intensity of a single color, the most commonly used set of colors is the RGB-scale (Red, Green, Blue). Further information can be found by looking into Tristimulus theory[26]

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

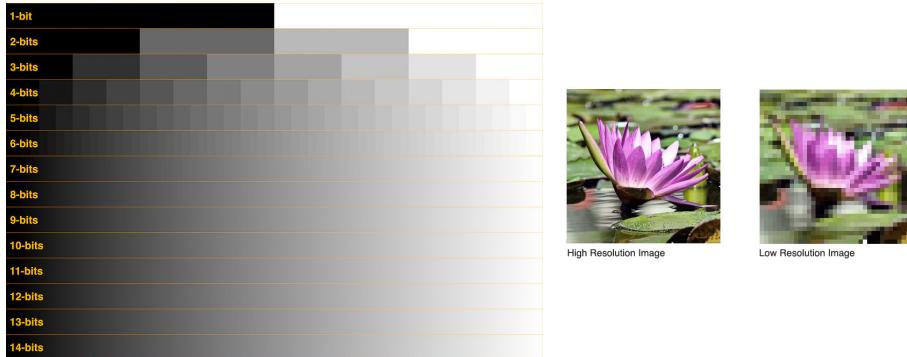
<sup>4</sup>Besides RGB and CMYK 1.3a the most common color spaces are CIE (Commision Internationale d'Eclairage) and HSV fig:1.3b (Hue,Saturation and Value). Refer to [10] for further details

$$HU = 1000 * \frac{\mu - \mu_{H_2O}}{\mu_{H_2O} - \mu_{Air}} \quad (1.1)$$

115 The utility of this scale is in its definition, since the pixel value depends on  
 116 the attenuation coefficient it's possible to individuate a set of ranges that identify,  
 117 within good reason, the various tissues in the human body: for example lungs are  
 118 [-700, -600] while bone can be in the [500, 1900] range.

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

- 122 • Spatial Resolution: A measure of how many pixels are in the image or, equivalently,  
 123 how small each pixel is; a larger resolution implies that smaller details  
 124 can be seen better fig:1.4. Can be measured as the number of pixels measured  
 125 over a distance of an inch ppi(Pixel Per Inch) or as number of line pairs that  
 126 can be distinguished in a mm of image lp/mm (line pair per millimeter).
- 127 • Color quantization: The range of the pixel values, a classic example is an 8-bit  
 128 resolution which yields 256 levels of gray. A better resolution allows a better  
 129 distinction of colors within the image fig:1.4.



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

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

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

136 The most commonly used formats are .PNG and .JPG however there are a lot  
 137 of other formats. In the context of this work, which is going to be centered on  
 138 medical images, the most interesting formats are going to be the nii.gz (Nifti)  
 139 and the .dcm (DICOM). The first contains only the pixel value information  
 140 hence it's a lighter format, it originates in the field of Neuroimaging<sup>5</sup>, it is used

<sup>5</sup>In fact Nifti stands for Neuroimaging Informatics Technology Initiative (NIfTI)

mainly in Magnetic resonance images of the brain but also for CT scans and, since it contains only numeric information, it's the less memory consuming option out of the two. The second contains not only the image data but also some data on the patient, such as name and age, and details on how the exam was carried out, such as machine used and specifics of the acquisition routine. This format is heavier than the previous one and, for privacy purposes, is much more delicate to handle which is why anonymization of the data needs to be taken in consideration.

For a thorough description of the DICOM standard refer to [1].

The format in which the image is saved depends on the compression algorithm used to store the information within the file. These algorithms can be lossy, in which case some of the information is lost to reduce the memory needed for storage, or lossless which means that all the information is kept at the expense of memory space. The first set of methods is preferred for storage of natural images, these are cases in which details have no importance, whereas the second set of methods is used where minute details can make a considerable difference such as in the medical field<sup>6</sup>.

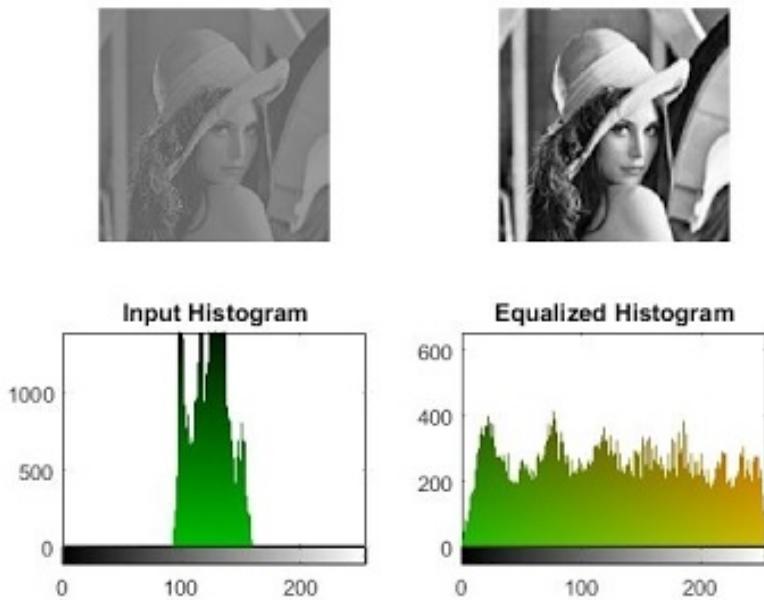
Given this set of characteristics it should now be clear that images can be thought of as array of numbers, for this reason they are often treated as matrices and, as such, there is a well defined set of valid operations and transformations that can be performed on them. All these operations and transformations,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

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

---

<sup>6</sup>A detailed description of compression algorithms is beyond the scopes of this thesis, for this reason please refer to [28] for more information

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



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

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

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

184 1. Geometric Transformations involve the following steps:

- 185     (a) Affine transformations: Transformations that can be performed via  
 186         matrix multiplication such as rotations, scaling, reflections and transla-  
 187         tions. This step basically involves computing where each original pixel  
 188         will fall in the transformed image
- 189     (b) Interpolation: Since the coordinates of the transformed pixel might not  
 190         fall exactly on the grid it might become necessary to compute a kind  
 191         of average contribution of the pixel around the destination coordinate  
 192         to find a most believable value. Examples of such methods are linear,  
 193         nearest neighbour and bicubic.

---

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

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

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

$$q = p^\gamma \quad (1.2)$$

202        (b) Local: The output value at specific coordinates depends on a combination  
203        of the original values in a neighbourhood around that same coordinates.  
204        Some examples are all filtering operation such as edge enhancement, di-  
205        lation and erosion. These filtering methods are based on performing con-  
206        volutions in which the output value at each pixel is given by the sum of  
207        pixel-wise multiplication between the starting matrix and a smaller (usu-  
208        ally 3x3 or 5x5) matrix called kernel. The output image is obtained by  
209        moving the kernel along the starting matrix following a predefined stride  
210        for example a (2,2) stride will move the kernel 2 pixels to the right and  
211        2 down. When moving near the borders the behaviour is defined by the  
212        padding of the image, most common choice for padding is zero padding,  
213        in which the image is considered to have only zeros outside of it, or no  
214        padding at all. Stride  $S$ , kernel shape  $K$ <sup>9</sup> and padding  $P$  determine  
215        the shape of the output matrix given the input dimension  $W$  via the  
216        following formula:

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

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

223        Having seen what constitutes an image and what can be done with one it  
224        becomes interesting to explore how images are obtained. The following discussion  
225        is going to introduce briefly some of the methods used to obtain medical images,  
226        getting more in depth only on the modality used to obtain all the images used in this  
227        thesis which is Computed Tomography. This technique was used because it's the  
228        only one that can provide information on the internal structure of an organ which  
229        is very low in density and that has parts that are deep in the patients body. Since  
230        no metabolic process of interest is in play PET is not advisable, Ultra Sounds are  
231        used for superficial soft tissue which is not the case of the lungs and MRI, despite

---

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

232 it's clear advantage in avoiding ionizing radiation, still has close to no acquisition  
233 protocol dedicated to lungs.

234 perché hai scelto queste specifiche tecniche? a random? Spiegato così va bene?

- 235 1. Magnetic Resonance Imaging (MRI): This technique is based on the phe-  
236 nomenon of Nuclear Magnetic Resonance(NMR) which is what happens when  
237 diamagnetic atoms are placed inside a very strong uniform magnetic field are  
238 subject to Radio Frequency (RF) stimulus. These atoms absorb and re-emit  
239 the RF and supposing this behaviour can somehow be encoded with a pos-  
240 tional dependence then it's possible to locate the resonant atoms given the  
241 response frequency measured. Suffices to say that this encoding is possible  
242 however the setup is very complex and the possible images obtainable with  
243 this method are very different and can emphasize very different tissue/material  
244 properties. Nothing more will be said on the topic since no data obtained with  
245 this methodology will be used. More details can be found in [5]
- 246 2. Ultra-Sound (US): The images are obtained by sending waves of frequency  
247 higher to those audible by humans and recording how they reflect back. This  
248 technique is used mainly in imaging soft peripheral tissues and the contrast  
249 between tissues is given by their different responses to sound and how they  
250 generate echo. The main advantages such as low cost, portability and harm-  
251 lessness come at the expense of exploratory depth, viewable tissues, need for a  
252 skilled professional and dependence on patient bodily composition as well as  
253 cooperation.
- 254 3. Positron Emission Tomography (PET): In this case the images are obtained  
255 thanks to the phenomenon of annihilation of particle-antiparticle, specifically  
256 of electron-positron pairs. The positrons come from the  $\beta^+$  decay of a radio-  
257 nucleide bound to a macromolecule, which is preferentially absorbed by the  
258 site of interest <sup>10</sup>. Once the annihilation happens a pair of (almost) co-linear  
259 photons having (almost) the same energy of 511 keV is emitted, the detection  
260 of this pair is what allows the reconstruction of the image representing the  
261 pharmaceutical distribution within the body. Once again there are a lot of  
262 subtleties that are beyond the scopes of this thesis, suffices to say that: firstly  
263 the exam is primarily used in oncology given the greater energy consumption,  
264 hence nutrients absorption, of cancerous tissue and secondly this technique  
265 can be combined with CT scans to obtain a more detailed representation of  
266 the internal environment of the patient

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

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

## <sup>270</sup> 1.1.1 X-ray imaging and Computed Tomography <sup>271</sup> (CT)

<sup>272</sup> It's well known that the term x-rays is used to characterize a family electromagnetic  
<sup>273</sup> radiation defined by their high energy and penetrative properties. Radiation of this  
<sup>274</sup> kind is created in various processes such as characteristic emission of atoms, also  
<sup>275</sup> referred to as x-ray fluorescence, and Bremsstrahlung, braking radiation<sup>11</sup>.

<sup>276</sup> The discovery that "A new kind of ray"<sup>[29]</sup> with such properties existed was  
<sup>277</sup> carried out by W.C.Roentgen in 1895, which allowed him to win the first Nobel  
<sup>278</sup> prize in physics in the same year. Clearly the first imaging techniques that involved  
<sup>279</sup> this radiation were much simpler than their modern counterpart, first of all they  
<sup>280</sup> were planar and analog in nature, as well as not as refined in image quality. The  
<sup>281</sup> first CT image was obtained in 1968 in Atkinson Morley's Hospital in Wimbledon.

<sup>282</sup> Tomography indicates a set of techniques<sup>12</sup> that originate as an advancement of  
<sup>283</sup> planar x-ray imaging; these techniques share most of the physical principles with  
<sup>284</sup> planar imaging while overcoming some of it's major limitations, main of which being  
<sup>285</sup> the lack of depth information. X-ray imaging, both planar and tomographic, involves  
<sup>286</sup> seeing how a beam of photons changes after traversing a target, the process amounts  
<sup>287</sup> to a kind of average of all the effects occurred over the whole depth travelled.

<sup>288</sup> The way in which slices are obtained is called focal plane tomography and, as the  
<sup>289</sup> name suggests, the basic idea is to focus in the image only the desired depth leaving  
<sup>290</sup> the unwanted regions out of focus. This selective focusing can be obtained either  
<sup>291</sup> by taking geometrical precautions while using analog detectors, such as screen-film  
<sup>292</sup> cassettes, or by feeding the digital images to reconstruction algorithms to perform  
<sup>293</sup> digitally the required operations<sup>13</sup>.

<sup>294</sup> In both planar and tomographic setting the rough description of the data acqui-  
<sup>295</sup> sition process can be summarized as follows: First x-rays are somehow generated  
<sup>296</sup> by the machine, the quality of these x-rays is optimized with the use of filters then  
<sup>297</sup> focused and positioned such that they mostly hit the region that needs imaging.  
<sup>298</sup> The beam then exits the machine and starts interacting with the imaged object<sup>14</sup>,  
<sup>299</sup> this process causes an attenuation in the beam which depends on the materials com-  
<sup>300</sup> posing the object itself. Having then travelled across the whole object it interacts  
<sup>301</sup> with a sensor, be it film, semiconductor or other, which stores the data that will  
<sup>302</sup> then constitute the final image. In a digital setting this final step has to be per-  
<sup>303</sup> formed following a (tomographic) reconstruction algorithm which given a set of 2D  
<sup>304</sup> projections returns a single 3D image.

<sup>305</sup> In this light the interesting processes are how the radiation is created and shaped  
<sup>306</sup> before hitting the patient and how said radiation then interacts with the matter of  
<sup>307</sup> both the patient's body and the sensor beyond it. To explore these topics it's  
<sup>308</sup> necessary to see:

---

<sup>11</sup>From the German terms *Bremsen* "to brake" and *Strahlung* "radiation"

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

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

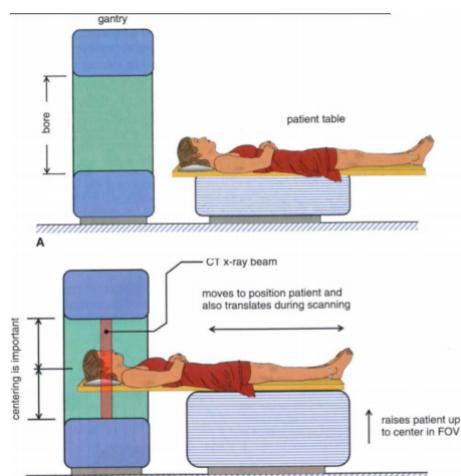
<sup>14</sup>In this work it's always going to be a patient, however this process is general and is also used in industry to investigate object construction

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

311     For more information on reconstruction algorithms refer to [15] and [33].

### 312    1.1.2 Generation and management of radiation: 313           digital CT scanners

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



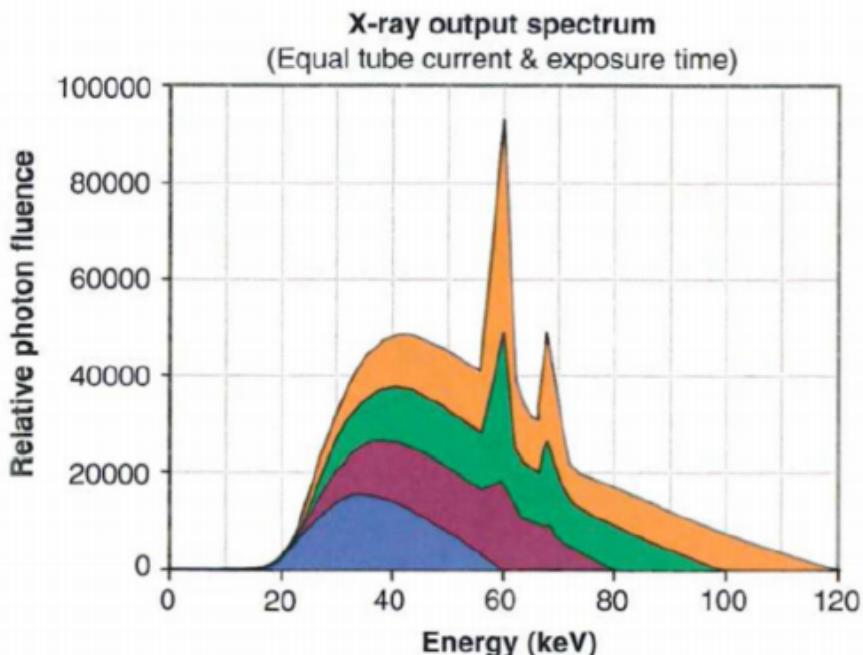
**Figure 1.6:** General set-up of a CT machine

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

---

<sup>15</sup>Most commonly electrons

<sup>16</sup>Typical materials can be Tungsten, Molybdenum



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

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

333 Other relevant characteristics in the tube are the anode material, which changes  
 334 the peaks in the x-ray spectrum and time duration of the emission, which is called  
 335 exposure time and influences dose as well as exposure<sup>17</sup>.

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

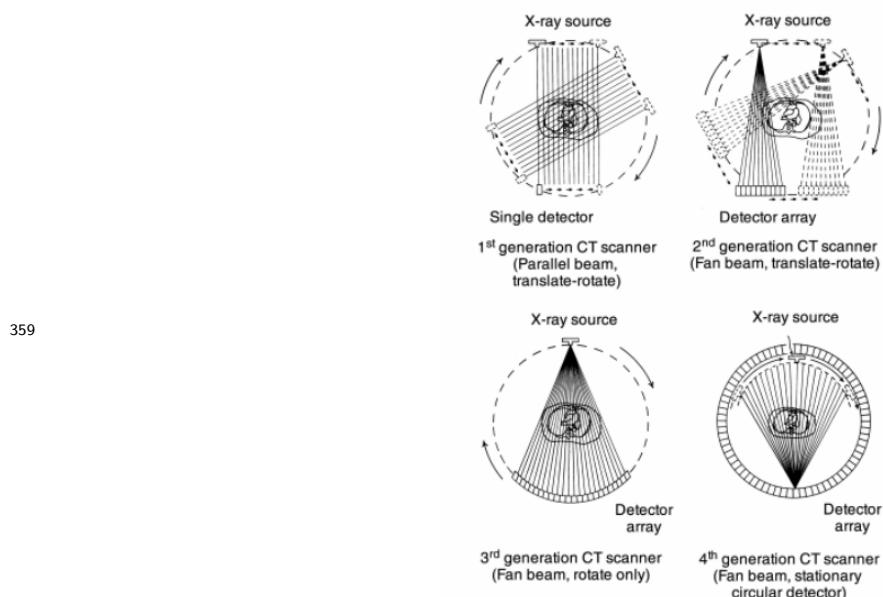
341 Exploiting the phenomenon of beam hardening a filter, usually of the same ma-  
 342 terial as the anode, is interposed between the beam and the patient to block lower  
 343 energy photons from passing through thereby reducing the dose conveyed to the  
 344 patient. At this point there may also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

---

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

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

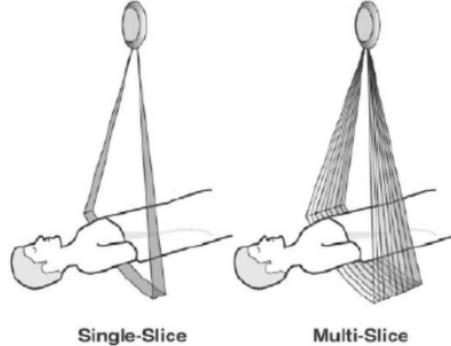
- 350 • 1<sup>st</sup> generation-Pencil Beam: A single beam is shot onto a single sensor, both  
 351 sensor and beam are translated across the body of the patient and then rotated  
 352 of some angle. The process is repeated for various angles. Main advantages  
 353 are scattering rejection and no need for relative calibration, main disadvantage  
 354 is time of the exam
- 355 • 2<sup>nd</sup> generation-fan Beam: Following the same process as the previous genera-  
 356 tion the main advantage is the reduction of the time of acquisition by intro-  
 357 ducing N beam and N sensors which don't wholly cover the patient's body so  
 358 still need to translate.



**Figure 1.8:** First four generation of CT scanners

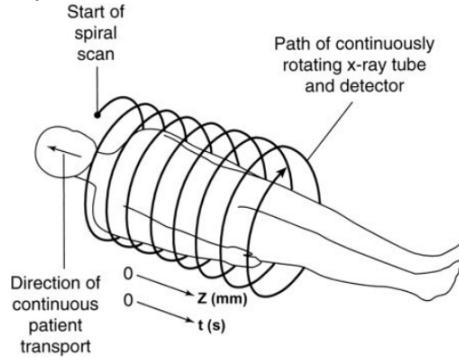
- 360 • 3<sup>rd</sup> generation-Rotate Rotate Geometry: Enlarging the span of the fan of  
 361 beams and using a curved array of sensor a single emission of the N beams  
 362 engulfs the whole body so the only motion necessary is rotation of the couple  
 363 beam-sensor array around the patient.
- 364 • 4<sup>th</sup> generation-Rotate Stationary Geometry: The sensors are now built to com-  
 365 pletely be around the patient so that only the beam generator has to rotate  
 366 around the body
- 367
- 368 • 5<sup>th</sup> generation-Stationary Stationary Geometry: The x-ray tube is now a large  
 369 circle that is completely around the patient. This is only used in cardiac  
 370 tomography and as such will not be described further [14]
- 371 • 6<sup>th</sup> generation-Spiral CT: Supposing the patient is laying parallel to the axis  
 372 of rotation, all previous generations acquired, along the height of the patient,  
 373 a single slice at a time. In this generation as the tube rotates around the

patients the bed on which they're laying moves along the rotation axis so that the acquisition is continuous and not start-and-stop. This further reduces the acquisition time while significantly complicating the mathematical aspect of the reconstruction. It's necessary to add another important parameter which is the pitch of the detector<sup>18</sup>. This quantifies how much the bed moves along the axis at each turn the tube makes around the patient. Pitches smaller than one indicate oversampling at the cost of longer acquisition times, pitches greater than one indicate shorter acquisition times at the expense of a sparser depth resolution.



**Figure 1.9:** 7<sup>th</sup> generation setup

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



**Figure 1.10:** 7<sup>th</sup> generation setup

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

<sup>18</sup>Once again the important parameters, such as this, can be accessed in the DICOM file resulting from the exam,

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

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

questa tabella richiede più dettagli di spiegazione. che cosa sono i numeri inseriti? la tabella dovrebbe poter essere compresa in modo completo senza troppi riferimenti al testo.

- 395 1. Ingenuity CT (Philips Medical Systems Cleveland):  $\sim 56\%$  of the exams were  
396 obtained with this machine
- 397 2. Lightspeed VCT (General Electric Healthcare, Chicago-Illinois):  $\sim 41\%$  of the  
398 exams in study come from this machine
- 399 3. ICT SP (Philips Medical Systems Cleveland):  $\sim 3\%$  of the exams were per-  
400 formed with this machine

### 401 1.1.3 Radiation-matter interaction: Attenuation 402 in body and measurement

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

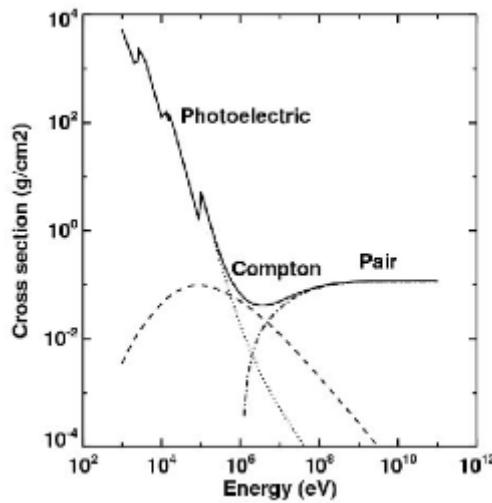
405 Let's first consider how a monochromatic beam of x-rays would interact with an  
406 object while passing through it. All materials can be characterized by a quantity  
407 called attenuation coefficient  $\mu$  which quantifies how waves are attenuated traversing  
408 them, this energy dependent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

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

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

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

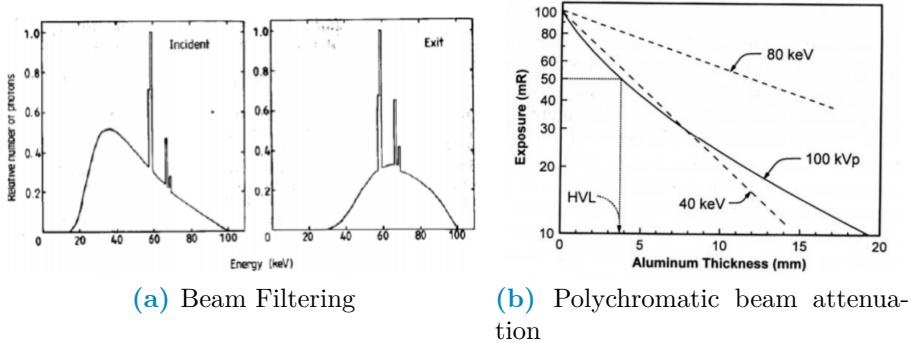
414 Where  $\rho$  is material density,  $N_A$  is Avogadro's number,  $A$  is the atomic weight  
 415 in grams and the distinction among the various possible interaction processes for a  
 416 generic photon of high energy  $E$  is made explicit. Overall the behaviour of the cross  
 417 section is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



**Figure 1.12:** Photon cross-section in Pb

418 Overall, given eq:1.4, it's clear to see that the attenuation behaviour of a monochro-  
 419 matic beam would be linear in semi-logarithmic scale hence the name for  $\mu$  "*Linear*  
 420 *Attenuation Coefficient*".

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



**Figure 1.13:** Polychromatc beam behaviour

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

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

This formula, that connects the beam to the image, together with eq: 1.4, which connects the beam property to the patient's composition, make clear the processes by which the beam carries patient information. Another complication arises in the context of this last equation due to the phenomenon of scattering which reduces the contrast by changing the direction of the bean 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 biological effects of radiation won't be treated in this thesis. Suffices to say that damage can be classified as primary, due to ionization events within the nucleus of the cell, or secondary, due to chemical changes in the cell environment. The energy deposited per unit mass is called dose and is measured in Gy(Gray) and, as said before, depends on exposure time, current and kVP of the tube. Most of contemporary machines for CT self-regulate exposure time during the acquisition automatically using Automatic Exposure Control(AEC). Having the dose it's possible to estimate the fraction of surviving cells and, to do so, various models are used. In the clinical practice it's common to find, still within the DICOM image metadata, the information regarding Dose delivered such as CTDI (Computed Tomography Dose Index) from which it's possible to obtain the DLP (Dose Length Product) taking into consideration the total length of irradiated body. For an introduction to one of these models, the Linear Quadratic (LQ) refer to [23].

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

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

457 Starting from the definition given by John McCarthy in [21] "[AI] is the science  
458 and engineering of making intelligent machines, especially intelligent computer pro-  
459 grams. It is related to the similar task of using computers to understand human  
460 intelligence, but AI does not have to confine itself to methods that are biologically  
461 observable.". Machine Learning (ML) is a sub-branch of AI and contains all tech-  
462 niques that make the computer improve performances via experience in the form  
463 of exposure to data, practically speaking this finds it's application in classification  
464 problems, image/speech/pattern recognition, clustering, autoencoding and others.

465 The general workflow of Machine Learning is the following: given a dataset, the  
466 objective is to define a model or function which depends on some parameters which  
467 is able to manipulate the data in order to obtain as output something that can be  
468 evaluated via a predefined performance metric. The parameters of the model are  
469 then automatically adjusted in steps to minimize or maximize this performance met-  
470 ric until a stable point at which the model with the current parameters is considered  
471 finalized; one of the main problems in this procedure is being sure that the stable  
472 point found is global and not local. The whole procedure is carried out keeping in  
473 mind that the resulting model needs to be able to generalize it's performance on  
474 data that it has never seen before, for this reason usually ML is divided in a training  
475 phase and a testing phase.

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

- 480 1. Supervised Learning: In this type of ML the model is provided with the input  
481 data as well as the correct expected output, which is hence called *label*. The  
482 objective of the model is to obtain an output as similar to the labels as possible,  
483 while also retaining the best possible generalization ability in predicting never  
484 seen before data. Some problems that benefit from the use of these techniques  
485 are regression and classification problems.
- 486 2. Unsupervised Learning: As the name suggests this category of models trains on  
487 the data alone, without having the labels available by minimizing some metric  
488 defined from the data. For example clustering techniques try to find a set of  
489 groups in the data such that the difference within each group is minimal while  
490 the difference among groups is maximal, ideally producing dense groups, called  
491 clusters, that are each well separated from all the others. Other techniques in  
492 this family are Principal Component Analysis (PCA) and autoencoding but  
493 the general objective is to infer some kind of structure within the data and the  
494 relation between data points.
- 495 3. Reinforcement Learning: This kind of ML is well suited for data which has  
496 a clear sequential structure in which the required task is to develop good  
497 long term planning. Broadly speaking the general set-up is that given a set of

---

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

(state<sub>t</sub>, action, reward, state<sub>t+1</sub>) these techniques try to maximize the cumulative reward<sup>20</sup>. The main applications of these techniques are in Autonomous driving and learning how to play games

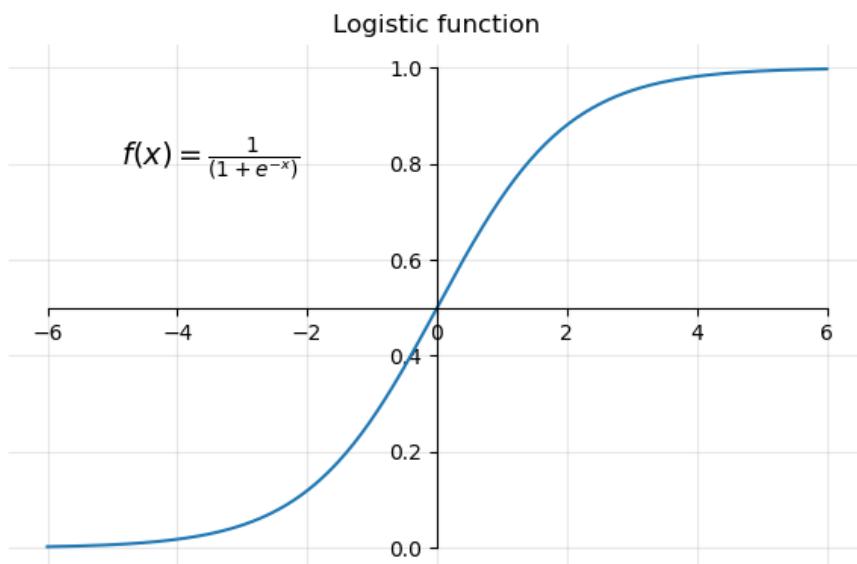
The following methods are those directly involved in this work.

### 1.2.1 Regression, Classification and Penalization

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

The most basic example of regression is linear regression which consists in finding the slope and intercept of a line passing through a set of points. A branch of classification that's similar to linear regression is logistic regression, this translates in finding out whether or not each point belongs to a certain category given it's properties and can be practically thought of, for a binary classification, as a fitting procedure such that the output can be either 0 for one category and 1 for the other, this is usually done using the logistic function, also called sigmoid, which compresses  $\mathbb{R}$  in  $[0, L]$  as seen in 1.14. L represents the maximum value desired,  $x_0$  is the midpoint and k is the steepness. Note that for multiple categories it would conceptually suffice to add sigmoids with different centers to obtain a step function in which each step corresponds to a category, in this case the procedure is usually called multinomial regression.

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



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

---

<sup>20</sup>i.e. the sum of the rewards obtained at all previous time steps

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

523 When each datapoint is characterized by a lot of different features the procedure  
 524 is called multiple linear regression, the task becomes finding out how much each  
 525 feature contributes in predicting the output within a weighted linear combination of  
 526 features. Practically speaking this can be done in matricial form, supposing that  
 527 each of the m datapoints has n features associated let  $\mathbf{X}$  be a matrix with  $n+1^{21}$   
 528 columns and m rows and let  $\mathbf{Y}$  be a vector with m entries. Let also  $\theta$  be a vector of  
 529  $n+1$  entries, one for each feature plus the intercept  $\theta_0$ , then we are supposing that:

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

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

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

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

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

534 It's evident that to obtain a result from the previous operation it's necessary that  
 535  $(\mathbf{X}^T * \mathbf{X})$  be invertible, which in turn requires that there be no correlated features  
 536 and that the features be less than the datapoints. To solve this problem the first  
 537 step is being careful in choosing the data that goes through the regression, which  
 538 may even involve some preprocessing. The second step is called Regularization, it  
 539 involves adding a penalty to the cost function by adding a small quantity along the  
 540 diagonal of the matrix. The nature of this small quantity changes the properties  
 541 of the regularization procedure, the most famous penalties are Lasso, Ridge and  
 542 ElasticNet. In practical terms the shape of the penalty determines how much and  
 543 how fast the slopes relative to the features can be shrunk.

- 544 1. Ridge: Adds  $\delta^2 * \sum_{j=1}^{n+1} \theta_j^2$ , is called also  $L^2$  regularization since it adds the  
 545  $L^2$  norm of the parameter vector. This penalty can only shrink parameters  
 546 asymptotically to zero but never exactly, which means that all features will  
 547 always be used, even with very small contributions
- 548 2. Lasso: Adds  $\frac{1}{b} * \sum_{j=1}^{n+1} |\theta_j|$ , is called also  $L^1$  regularization since it adds the  $L^1$   
 549 norm of the parameter vector. This penalty can shrink parameters to exactly  
 550 zero, getting rid of the useless variables within the model.
- 551 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  
 552 between the Lasso and Ridge methods, where the balance is dictated by the  
 553 value of  $\alpha$ .

---

<sup>21</sup> $n+1$  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

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

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

562 Lasso regularization has the ability to shrink parameters exactly to zero and  
563 does so in cases in which variables are correlated with one another. However the  
564 choice on which feature to keep is random if the variables are highly correlated. The  
565 advantages of Lasso, namely the feature selection it produces while also improving  
566 the prediction on the data, come at the cost of difficult to interpret results in some  
567 cases as well as a limit in the maximum number of features that can survive the  
568 procedure.<sup>22</sup>. This means that practically one would choose Lasso regression in  
569 cases in which the expectation is that only a few variables drive the behaviour of  
570 the response.

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

577 For more in depth information refer to [12].

visto che usi la regolarizzazione in modo consistente spiegherei meglio i vantaggi  
dei vari approcci uno rispetto all'altro Lorenzo: Aggiunto un po' di discussione

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

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

- 587 1. Standard Scaling: This amounts to subtracting the mean of the distribution  
588 to the feature and dividing by the standard deviation so that the resulting distribution  
589 is somewhat centered around zero and has close to unitary standard deviation
- 591 2. Boxcox transform: When distributions are heavy-tailed, like a gamma distribution  
592 would be, this transform is used to find the optimal power-law that  
593 more closely turns the data in a normal distribution. More specifically the  
594 data is transformed according to:

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

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

595         $\lambda$  is varied from -5 to 5, The best value is chosen so that the transformed  
 596        data approximates a normal distribution as closely as possible Being that the  
 597        exponent can be positive or negative this transform cannot handle distributions  
 598        with negative values. For more information refer to the [7]

599        la spiegherei meglio. Lorenzo:Così va meglio?

600        Practical application of these procedures can be found in cap:2. These consider-  
 601        ations conclude the theoretical background on regression, classification and penal-  
 602        ization thereof.

### 603        1.2.2 Decision Trees and Random Forest

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

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

610        Since it's a classification method let's consider the simple case of binary classi-  
 611        fication with categorical<sup>23</sup> features. The task of the decision tree is to approximate  
 612        to the best of it's abilities the labels contained in the training set using all the fea-  
 613        tures associated with each datapoint, this is done by building a graph-like structure  
 614        in which the nodes represent the features and the links departing from it are the  
 615        possible values the feature takes. This graph is built in a top-down approach by  
 616        choosing at every step the feature that best separates the data in the label cate-  
 617        gories, this process is done along each branch until a node in which the separation  
 618        of the preceding feature is better then that provided by all remaining features or all  
 619        features have been considered. The first node is called root node, while the nodes  
 620        that have no branches going out of them are called leaves, the graph represents a  
 621        tree hence the name of the method. Note that at every node only the subset of data  
 622        corresponding to all previous feature categories is used. At each node the separation  
 623        between the two label categories  $c_1, c_2$  due to the feature is commonly measured with  
 624        Gini impurity coefficient, which is computed as:

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

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

---

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

627 the nodes. This method can be obviously generalized to cases in which features are  
628 continuous by thresholding the features choosing the value that best improves the  
629 separation of the deriving node, a way to choose possible thresholds is to take all  
630 the means computed with all adjacent measurements. It's important to note firstly  
631 that there is no restriction on using, along different branches, different thresholds for  
632 the same feature and, secondly, that the same feature can end up at different depths  
633 along different branches. The strength of this method is its performance on data  
634 it has seen however it has very poor generalization abilities [12].

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

642 This method vastly improves the performance and robustness of the final predic-  
643 tion while retaining the simplicity and ease of interpretation of the decision trees,  
644 naturally there are methods, such as AdaBoost, to deploy and precautions, such as  
645 having balanced dataset, to take to further improve the performance of RF classi-  
646 fiers.

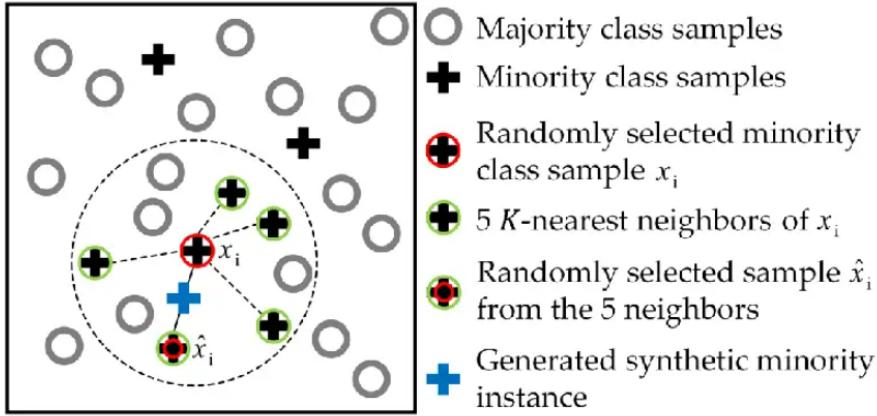
647 In the context of this thesis it will become necessary to take care of balancing the  
648 input dataset, to do so one could randomly oversample, by duplicating instances in  
649 the minority class, or undersample, by removing instances within the majority class.  
650 The choice that was made was to use Synthetic Minority Oversampling TEchnique  
651 (SMOTE)[6] to rebalance the dataset.

### 652 1.2.3 Synthetic Minority Oversampling TEchnique 653 (SMOTE)

654 This technique considers a user-defined number of nearest neighbours of randomly  
655 chosen points in the minority class and populates the feature space by generating  
656 samples on the lines that connect the chosen sample with a random neighbour<sup>24</sup>.

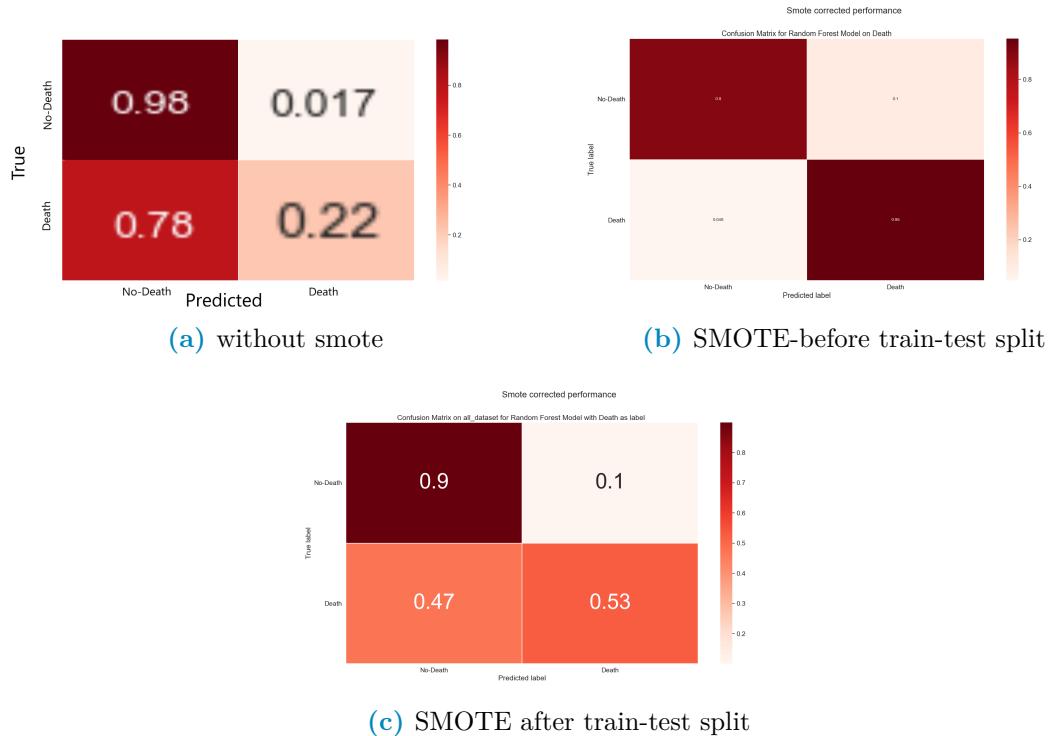
---

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



**Figure 1.15:** Example of SMOTE with 5 nearest neighbours

Worth noting that this has been done using the library imblearn in python [20]. To preview some of the data, looking at the performance of a vanilla random forest implementation with all preset parameters, the effect of the position of oversampling is the following.



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

queste figure devono avere i numeri ed il testo più grandi, sono illeggibili altrimenti; inoltre i risultati delle divisioni vanno inseriti dopo, non qui.

It's clear to see that in unbalanced case, like the one analysed in this thesis in which the label classes are 15%-85%, oversampling the data always determines an

663 improvement in performance. However performing it in the wrong place it clearly  
664 makes a far too optimistic evaluation of the performance. This highlights the point  
665 that all preprocessing should be done with care when train-test splitting the data,  
666 specifically it's very important that the oversampling, as well as all other data  
667 handling, be performed after the train-test split of the data on the train data alone.  
668 It's clear to see that what's being observed is a leakage phenomenon, in which the  
669 testing data contains information regarding the training data and viceversa. In  
670 practice, especially because the points are created as convex combination of existing  
671 data<sup>25</sup>, when the synthetically generated points end up in the testing they depend,  
672 at least partially, on the data used in the training procedure.

### 673 1.2.4 Dimensionality reduction and clustering

674 When dataset are composed of many features, namely more than three, it becomes  
675 difficult if not impossible to visualize the distribution of data. There are various  
676 techniques that can mitigate or solve this problem, they are grouped under the  
677 umbrella term of "Dimensionality reduction techniques" and they are generally based  
678 on ML learning methods. As such , following the same reasoning and definitions  
679 given in regard to ML, it's possible to introduce a sub-categorization of the whole  
680 family of techniques in supervised and unsupervised techniques. Dimensionality  
681 reduction, beyond providing useful insight in the general structure of the data, can  
682 also be used as a preprocessing step in some analysis pipelines.

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

688 The most common dimensionality reduction techniques are:

#### 689 1. Unsupervised methods

- 690 • Principal Component Analysis (PCA): Linearly combines the pre-existing  
691 features to obtain new ones and orders them by decreasing ability to ex-  
692 plain total variance of data. The first principal component is the combi-  
693 nation of features that most explains the variance in the original dataset.  
694 Given it's nature it focuses on the global characteristics of the dataset.
- 695 • t-distributed Stochastic Neighbor Embedding (t-SNE): Keeps a mixture  
696 of local and global information by using a distance metric in the full high  
697 dimensional space and trying to reproduce the distances measured in the  
698 lower dimensional space which can be 2- or 3-dimensional. NON SO  
699 SE VA SPIEGATO MEGLIO O PIU' IN DETTAGLIO  
700 Let's define similarity between two points as the value taken by a normal  
701 distribution in a point away from the centre, corresponding to the first  
702 point of interest, the same distance as the two points taken in consider-  
703 ation. Fixing the first point the similarities with all other points can be

---

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

704 computed and normalized. Doing this procedure for all points it's possible  
705 to build a normalized similarity matrix. Then the whole dataset  
706 is projected in the desired space, usually  $\mathbb{R}^i$  with  $i=1,2$  or 3, in which a  
707 second similarity matrix is built using a t-distribution instead of a normal  
708 distribution <sup>26</sup>. At this point, in iterative manner, the points are moved  
709 in small steps in directions such that the second matrix becomes more  
710 similar to the one computed in the full space.

## 711 2. Supervised methods

- 712 • Partial Least Squared Discriminant Analysis (PLS-DA): This technique is  
713 the classification version of Partial Least Squares (PLS) regression. Much  
714 like PCA the idea is to find a set of orthonormal vectors as linear combina-  
715 tion of the original features in the dataset, however in PLS and PLS-DA  
716 is necessary to add the constraint that the new component, besides be-  
717 ing perpendicular to all previous ones, explains the most variability in a  
718 given target variable, or set thereof. For an in depth description refer to  
719 [4] while for a more modern review refer to [19] Hanno senso queste  
720 due in bibliografia?

## 721 3. Mixed techniques

- 722 • Uniform Manifold Approximation and Projection (UMAP): Builds a net-  
723 work using a variable distance definition on the manifold on which the  
724 data is distributed then uses cross-entropy as a metric to reproduce a  
725 network with the same structure in the space with lower dimension. This  
726 technique maintains very local information on the data and allows a com-  
727 plete freedom in the choice of the final embedding space as well as the  
728 definition of distance metrics in the feature space<sup>27</sup>, it's also implemented  
729 to work an a generic pandas dataframe in python so it can take in im-  
730 put a vast range of datatypes. The math behind this method is much  
731 beyond the scopes of this thesis, as such refer to [22] for more in depth  
732 information.

733 It should be noted that dimensionality reduction is not to be taken as a necessary  
734 step, however it can reduce the noise in the data by extrapolating the most infor-  
735 mative features while easing in visualization and reducing computational costs of  
736 subsequent data analysis. These upsides become particularly relevant in the field of  
737 clustering, where the objective is to group data in sets with similar features by min-  
738 imizing the differences within each group while maximizing the differences between  
739 different groups. Clustering techniques can be roughly divided in:

---

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

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

- 740        1. Centroid based techniques: A user defined number of points is randomly lo-  
 741        cated in the data space, datapoints are then assigned to groups according to  
 742        their distance from the closest center. The main technique in this category  
 743        is k-means clustering, and some, if not most, other techniques include it as  
 744        step in the processing pipeline<sup>28</sup>. The main problems are firstly that these  
 745        techniques require prior knowledge, or at the very least a good intuition, on  
 746        the number of clusters in the dataset while also assuming that the clusters are  
 747        distributed in spherical gaussian distribution, which is not always the case.
- 748        2. Hierarchical clustering: The idea is to find the hierarchical structure in the  
 749        data using a bottom-up or a top-down approach, as such this category further  
 750        subdivides in agglomerative and divisive methods. In the first each point  
 751        starts by itself and then points are agglomerated using a similarity or distance  
 752        metric, this build a dendrogram in which the  $k^{th}$  level roughly corresponds  
 753        to a  $k$ -centroid clustering. In the latter the idea is to start with a unique  
 754        category and then divide it in subgroups.
- 755        3. Density based techniques: These methods use data density to define the groups  
 756        by looking for regions with larger and lower density as clusters and separations.

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

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

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

### 774        1.3 Combining radiological images with AI: Im- 775        age segmentation and Radiomics

776        Having seen the kind of data that will be of interest throughout this thesis, and  
 777        having a set of techniques used describe and make prediction on the data at hand,  
 778        the final step in this theoretical background chapter will be to combine these two  
 779        notion in describing first how images can be treated in general terms and then, more

---

<sup>28</sup>An example of such techniques is: affinity propagation

780 specifically, how medical images can be analysed to exploit as much as possible the  
781 vast range of information they contain.

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

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

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

### 803 1.3.1 Image Segmentation

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

- 810 • Manual segmentation: The regions of interest are manually defined usually by  
811 a trained individual. The main advantage of this it's the versatility, on the  
812 other hand this process can be very time consuming
- 813 • Semi-automatic segmentation: A machine defines as best as it can the shape  
814 of the region of interest, however the process is then thought to receive inter-  
815 vention of an expert to correct the eventual mistakes or refine the necessary  
816 details. This provides the best compromise between time needed and accuracy  
817 obtained and becomes of interest in fields in which finer details are important.
- 818 • Automatic segmentation: A machine performs the whole segmentation without  
819 requiring human intervention

820 On a practical level these techniques can be used in various fields, as illustrated  
821 by the variety of aforementioned tasks, but the one that interests this work the

most is the medical field. Nowadays in medicine, where most of imaging exams are stored in digital form, the ability to automatically discern specific structures within the images can provide a way to aid clinical professionals in their everyday decision making their workload lighter and helping in otherwise difficult cases. A staggering example connected to this thesis is the process of organ segmentation in CT scans: usually these scans are  $n^{29}$  stacked images in 512x512 resolution. To have a contour of the lungs in a chest CT a radiologist would need to draw by hand the contour of the lung in each slice of the scan. Even if some shorcuts exist to reduce the number of slices to draw on this very boring, time consuming and repetitive task can occupy hours if not days of work to a human while a machine can take minutes to complete a whole scan. Then considering lesion detection in medical exams having a machine that consistently finds lesions that would otherwise be difficult to discern by a human eye can be of paramount importance in diagnosis as well as treatment. In the medical field the difficulty comes from the fact that different exams have different types of image formats which means that an algorithm that works well on CT may not work as intended on MRI or other procedures. Automatic image segmentation can be performed in various ways:

1. Artificial Neural Network (ANN): These techniques belong to a sub-field in ML called Deep Learning, they involve building network structures with more layers in which each node is a processing unit that takes a combination of the input data and gives an output according to a certain activation function. These structures are called Neural Networks because they resemble and are modeled after the workings of neuron-dendrite structures while the deep in deep learning refers to the fact that various layers composed of various neurons are stacked one after the other to complete the structure. Learning is obtained by changing how each neuron combines the inputs it recieves. In the case of images these structures are called Convolutional Neural Networks because the first layers, which are intended to extract the latent features or structures in the images, perform convolution operation in which the parameters are the values pixel values of convolutional kernels
2. Thresholding: These techinques involve using the histogram of the image to identify two or more groups of pixel values that correspond to specific parts/objects within the image. An obvious case would be a bi-modal distribution in which the two sets can be clearly identified but there are no requirements on the histogram shape. In the case of CT this has a very simple and clear interpretation since HU depend on tissue type it's reasonable to expect that some tissues can be differentiated with good approximation by pixel value alone
3. Deformable models and Region Growing: Both these technique involve setting a starting seed within the image, in the first case the seed is a closed surface which is deformed by forces bound to the region of interest, such as the desired edge. In the second case the seed is a single point within the region of interest,

---

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

step by step more points are added to a set which started as the seed alone according to a similarity rule or a predefined criteria.

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

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

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

For a more in depth review of the main methods used in medical image segmentation refer to [27]. Worth noting, at this point, that the semi-automatic segmentation software used in this work uses probably a mixture of region growing and thresholding methods, maybe guided by an atlas. The segmentation process generally produces a boolean mask which can be used to select, using pixel-wise multiplication, the region of interest in the image. The next step in image analysis would be to derive information from the region defined during segmentation, in general this step is called feature extraction<sup>30</sup> and its objective is to find non-redundant quantities that meaningfully summarize as much properties of the original data as possible.

### 1.3.2 Radiomics

When the images are medical in nature and when referring to high-throughput quantitative analysis the task of finding these features fall in the realm of radiomics, which uses mathematical tools to describe properties of the images that would otherwise be unquantifiable to the human eye. The features that can be computed from images are of various types, some of them can be understood somewhat easily through intuition since they are close to what humans generally use to describe images, others are much more complex in definition and quantify more difficult perceived properties of the image. Features can be then roughly classified in different families:

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

---

<sup>30</sup>Even if the term is used in pattern recognition as well as machine learning in general

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

912        3. Higher order features: All the other fatures fall in this macro-category which  
913        can be clearly subdivided in other smaller catgories in which the features are  
914        obtained following the same guiding principle or starting point. Generally  
915        these describe more texture-like properties of the image and, to do so, use  
916        particular matrices derived from the original image which contain specific in-  
917        formation regarding order and relationships in pixel value positioning within  
918        the image. These matrices have very precise definition, as such only the gen-  
919        eral idea behind them will be reported here redirecting to [35] for a more strict  
920        and in detail description. The matrices from which the features are computed  
921        also give name to the smaller categories in this family, these categories are:

- 922        • Gray Level Co-occurrence Matrix (GLCM) features: This matrix expresses  
923        how combination of pixel values are distributed in a 2D or 3D region by  
924        considering connected all neighbouring pixel in a certain direction with  
925        respects to the one in consideration. Using all possible directions for a  
926        set distance, usually  $\delta=1$  or  $\delta=2$ , various matrices are obtained and from  
927        these a probability distribution can be built and evaluated. It should be  
928        noted that before computing these matrices the intensities in the image  
929        are discretized.
- 930        • Gray Level Run Length Matrix (GLRLM) features: Much like before the  
931        task of these features is to quantify the distribution of relative values in  
932        gray levels trthroughout the image, as the name suggests what this matrix  
933        quantifies is how long a path can be built by connecting pixel of the same  
934        value along a single direction. This time information from the matrices  
935        computed by considering different directions are aggregated in different  
936        ways to improve rotational invariance of the final features
- 937        • Gray Level Size Zone Matrix (GLSZM) features: This matrix counts the  
938        number of zones in which voxel have the same discretized gray level.  
939        The zones are defined by a notion of connectedness most commonly first  
940        neighbouring voxel are considered as connectable if they have the same  
941        value, this leads to a 26 neighbouring voxel in 3 dimensions and to 8  
942        connected pixels in 2 dimensions<sup>31</sup>. The matrix contains in position (i,j)  
943        the number of zones of size j in which pixel have value i.
- 944        • Neighbouring Gray Tone Difference Matrix (NGTDM) features: Born as  
945        an alternative to GLCM these features rely on a matrix that contains

---

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

946 the sum of differences between all pixel with a given pixel value and the  
947 average of the gray levels in a neighbourhood around them.

- 948 • Gray Level Dependence Matrix (GLDM) features: The aim of these  
949 features is to capture in a rotationally invariant way the texture and  
950 coarseness of the image. This matrix requires the already seen concept of  
951 connectedness with a given distance as well as dependence among pixel.  
952 Two voxel in a neighbourhood are dependent if the absolute value of the  
953 difference between their discretized value is less than a certain thresh-  
954 old. The number of dependent voxel is then counted with a particular  
955 approach to guarantee that the value be at least one

956 In talking about the previous feature groups the concept of discretization of  
957 data which is already digital, and hence a discretized, has emerged. This is often  
958 a required step to make the computations of the matrices tractable and consists in  
959 further binning together the pixel values, which is commonly done in two main ways:  
960 either the number of bins is fixed or the width of the bins is fixed preceding the  
961 discretization process. It should be evident that the results of the feature extraction  
962 procedure is heavily dependent on the choices made in all the steps that precede it,  
963 main of which being the segmentation, the eventual re-discretization of the image  
964 leading to the algorithm used to compute the feature themselves. For this reason  
965 recently the International Biomarker Standardization Initiative (IBSI [35]) wrote a  
966 "*reference manual*" which details in depth the definitions of the features, description  
967 of data-processing procedures as well as a set of guidelines for reporting results.  
968 This was done in an attempt to reduce as much as possible the variability and lack  
969 of reproducibility of radiomic studies.

970 In the past the main attention in radiological research was focused on improving  
971 machine performances and evaluating acquisition sequence technologies, however  
972 the great developments in artificial intelligence and performance of computers have  
973 brought a lot of attention to the field. Various papers, such as [18] and [3] have been  
974 written with the objective of presenting the general workflow.

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

- 977 • Data acquisition  
978 • Definition of the Region Of Interest (ROI)  
979 • Pre-processing  
980 • Feature extraction  
981 • Feature selection  
982 • Classification

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

987 With this the theoretical background has been laid out and the thesis can finally  
988 proceed to practically and more specifically describing the data at hand as well as  
989 the specific techniques used to elaborate it.

## 990 1.4 Survival Analysis

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

997 The main concepts and terminologies in this field are:

998 1. Event: This is the phenomenon under analysis, generally it could be death,  
999 remission from recovery, recovery and so on. It is common to refer to the event  
1000 as failure since usually the event is a negative event.

1001 2. Censoring: When collecting data to develop a model that describes survival it  
1002 may happen that the individual drops out of the study without incurring in  
1003 the event under analysis. For example when looking at effectiveness of a drug  
1004 the patient develops an adverse reaction to the drug and needs to stop using  
1005 it, hence falling out of the study.

1006 In the context of this thesis all individuals that got sent home from the hospital  
1007 are censored since their survival is known only up until the dropout and not  
1008 after. Cases like this when the start of the followup is well known but dropout  
1009 happens are called right-censored, because only the right side of the timeline  
1010 is abruptly interrupted. Cases can also be left-censored, e.g. a patients with  
1011 unknown time of contraction of a disease, and interval-censored, e.g. when the  
1012 contraction of the disease can be restricted to an interval as it may happen  
1013 when a negative and positive test happen at successive times.

1014 Usually this variable is called **d** and is a binary variable where 1 indicates that  
1015 the event occurred and 0 indicates all possible censoring causes.

1016 3. Time: This is to be intended as time, be it days, weeks, months or years, since  
1017 the start of the followup to either the event or the censoring. It usually is  
1018 indicated with **T**, is referred to as survival time and, being a random variable  
1019 that indicates time, it cannot be negative.

1020 4. Survivor function **S(t)**: This is to be intended as the probability that the  
1021 subject in the study survives a time t before incurring in the event. In theory  
1022 this function is smooth from zero to infinity, it's strictly non-increasing, it  
1023 starts at  $S(0)=1$  and ends up at  $S(\infty)=0$ .

1024 These assumptions are all resonable since at no point the survival probability  
1025 can increase, since everybody is alive at the start of observing them and since  
1026 nobody can live to infinity. However, when it comes to practice, these prop-  
1027 erties are not necessarily verified. Since no study can continue to infinity the

1028 last value need not be zero and since the timesteps at which it's possible to  
1029 perform a checkup are discrete the curve is actually a step function.

- 1030 5. Hazard Function  $h(t)$ : This function is difficult to explain practically, citing [17] "The hazard function  $h(t)$  gives the instantaneous potential  
1031 per unit time for the event to occur, given that the individual has  
1032 survived up to time  $t$ ". The mathematical definition is:

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

1034 Dividing by a time interval the hazard function can also be intended as conditional  
1035 failure rate, where the conditional refers to the "given that the individual  
1036 has survived up to time  $t$ ". Being a rate this quantity need not be bound in  
1037  $[0,1]$  but ranges in  $[0, \infty]$  and depends on the time unit used. This will be  
1038 interpreted as  $h(t)$  events per unit time.

1039 The hazard function is non-negative and has no upper bounds, since it can be  
1040 related to the survival function <sup>32</sup> the possible shapes give different names to  
1041 the final model. These could be increasing or decreasing Weibull, exponential  
1042 or lognormal survival models

1043 A further step in the analysis could be to try to measure the differences in survival  
1044 between two groups, i.e. using a single variable expected to drive the differences in  
1045 survival. The most basic example would be looking at the effectiveness of a drug by  
1046 dividing in group A and B the patients given the actual medicine and the placebo  
1047 respectively, however this could be done even for males vs females or, in the case  
1048 of continuous variables, for patients with above or below threshold values <sup>33</sup>. When  
1049 the analysis is univariate in nature then the Kaplan-Meier survival curves are used  
1050 in conjunction with the log-rank test, when the analysis is multivariate then the Cox  
1051 Proportional-Hazard model is used. The Cox model is very similar to linear and  
1052 logistic regression.

### 1053 1.4.1 Kaplan-Meier(KM) curves and log-rank test

1054 Kaplan-Meier curves can be built following a well defined procedure: The data is  
1055 separated in sets using the label of the groups that need to be used then the patients  
1056 in each set are ordered in ascending order of permanence in the study, which is the  
1057 time variable. At each failure time  $T_f$  the conditional probability of surviving past  
1058  $T_f$  given availability is computed as ratio of subjects left in the study right after  $T_f$   
1059 divided by the number of available people at that same time <sup>34</sup> Note that this takes  
1060 in account also the possible censoring by reducing the number of available people

---

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

<sup>33</sup>Generally, when no obvious threshold is available, the median of the variable is chosen.

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

1061 at times of event, these censoring event will be drawn on the curve using ticks at  
 1062 the corresponding time. The survival probability is computed at each event time as  
 1063 the product of the probability at the previous failure time with the one computed  
 1064 as explained before in at the time of the current event.

1065 Let's consider an imaginary study with 4 patients with one failing at week one,  
 1066 one dropping out at week 2, one failing at week 3 and one surviving after 4. The  
 1067 KM curve will start at 1, at week one it will drop at  $\frac{3}{4}$  since 3 people will be alive  
 1068 out of 4 available, at week two no drop will happen but a tick will be put on the  
 1069 curve and, finally at week 3 it will drop at value  $\frac{3}{4} * \frac{1}{2}$  since only one out of the two  
 1070 available will survive.

1071 As a rule of thumb, two Kaplan-Meier curves that do not intersect at any point  
 1072 indicate good separation among the groups, this can then be formally evaluated  
 1073 performing a log-rank test on the data used to build the curves. The null hypothesis  
 1074 of this specific test is that "...there is no overall difference between the two survival  
 1075 curves" which can be tested with the log-rank test. This basically consist in performing  
 1076 a large-sample  $\chi^2$  test that uses the ordered failure times for the entire dataset  
 1077 as expected values vs those observed in the subsets. From this testing procedure a  
 1078 p-value can be obtained to reject the null hypothesis, for more in depth information  
 1079 refer to [17].

## 1080 1.4.2 Cox Proportional-Hazard (CoxPH) model

1081 As it was mentioned before the shape of the hazard curve, an hence of the survival  
 1082 curve, implies a specific shape for the model used to describe it. Some of the possible  
 1083 models are increasing Weibull, decreasing Weibull, Exponential and log-normal. All  
 1084 of these models are parametric because known the parameters the distribution of the  
 1085 outcome can be known, when it comes to Cox PH model the distribution cannot be  
 1086 known because part of the model, namely the baseline hazard, remains unestimated.

1087 Generally the data has an optimal parametric model that describes it and, if this  
 1088 information were known, it would make perfect sense to use said model. However  
 1089 this knowledge is not always obtainable, which give Cox PH model an occasion to  
 1090 shine. This model can be described as robust in the sense that the results that  
 1091 it gives will approximate those obtained by the correct model, without needing to  
 1092 know which of the model needs to be used. For a more in depth description refer to  
 1093 [17]

1094 Cox PH models rely on the use of the formula in Equation 1.14 to express the  
 1095 risk of a patient with a set of k characteristic variables  $\mathbf{X}$  at time t.

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

1096 The quantity  $h_0(t)$  is called baseline hazard and it hides one of the main hy-  
 1097 potheses behind this method which is the *Proportional Hazard* assumption. This  
 1098 assumption is that the baseline hazard  $h_0(t)$  depends on time alone and not on all  
 1099 the other variables  $\mathbf{X}$ , note also that the exponent is time-independent since the  $\mathbf{X}$ s  
 1100 are supposed to be constant in time <sup>35</sup>.

---

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

1101        The  $\beta$  in the exponent represents the weights assigned to each variable and, very  
1102 much like what happened in linear regression, these coefficients can be a proxy of  
1103 importance in the model: coefficients close to zero signify no particular importance  
1104 of the variable whereas the more the value is distant from zero the more important  
1105 the variable can be considered. The reason why it's common practice to report  
1106 the exponentiated coefficient for features is that  $1-\exp[\beta]$  represents the difference in  
1107 likelihood to die of individuals separated using the feature relative to  $\beta$ .

1108        For example considering a binary feature AGE OVER 50 with  $\exp[\beta]=1.6$  would  
1109 indicate that people over 50 are 60% more likely to die when adjusting for all other  
1110 variables. It is also common procedure to provide, for each coefficient, the 95%  
1111 confidence interval of the parameter and a p-value relative to the hypothesis that  
1112 the parameter be equal to zero.

1113        Cox models provide a way to predict hazard for patients, this predict method  
1114 can be used to evaluate the model built by identifying groups in the patient cohort  
1115 using hazard quantiles. The division in the resulting groups can be used as proxy for  
1116 the quality of the score built using the Cox model.

<sub>1117</sub> **Chapter 2**

<sub>1118</sub> **Materials and methodologies**

<sub>1119</sub> In this section there's going to be an explanation of the dataset as well as instruments  
<sub>1120</sub> and methodologies used to analyze it's properties, as such the first step is going to  
<sub>1121</sub> be an in depth discussion of the data available and a general overview of the final  
<sub>1122</sub> use. The following step is going to be a description of the preliminary work done to  
<sub>1123</sub> the data itself and to the results of this preliminary analysis in order to select the  
<sub>1124</sub> important features. The final step of this chapter is going to be an explanation of  
<sub>1125</sub> the methods used to derive the final results and to evaluate them.

<sub>1126</sub> **2.1 Data and objective**

<sub>1127</sub> The objective of this thesis will be to compare how different methods perform in  
<sub>1128</sub> predicting clinical outcomes in covid patients, while also determining if different  
<sub>1129</sub> kind of input data imply different performances of the same methods. All images are  
<sub>1130</sub> non segmented, as such all of them are going to be semi-automatically segmented  
<sub>1131</sub> via a new software being tested in the medical physics department called *Sophia*  
<sub>1132</sub> *Radiomics*, which will be briefly explained shortly. Statistical analysis are going to be  
<sub>1133</sub> performed mainly in python using libraries such as scikit-learn, pandas, numpy, scipy  
<sub>1134</sub> while the graphical part of the analysis is done with either seaborn or matplotlib.  
<sub>1135</sub> The starting dataset was a list of all the patients that, from 02/2020 to 05/2021,  
<sub>1136</sub> were hospitalized as COVID-19 positive inside the facilities of *Azienda ospedaliero-*  
<sub>1137</sub> *universitaria di Bologna - Policlinico Sant'Orsola-Malpighi*. As far as exclusion  
<sub>1138</sub> criteria go the main exclusion criteria, except unavailability of the feature related to  
<sub>1139</sub> the patient, was visibly damaged or lower quality images, for example images with  
<sub>1140</sub> cropped lungs. The first set of selection criteria were:

- <sub>1141</sub> • All patients that had undergone a CT exam which was retrievable via the  
<sub>1142</sub> PACS (Picture Archiving and Communication System) of *Azienda ospedaliero-*  
<sub>1143</sub> *universitaria di Bologna - Policlinico Sant'Orsola-Malpighi*
- <sub>1144</sub> • All patients that had a all of the clinical and laboratory features, listed in  
<sub>1145</sub> fig:2.1, suggested by Lidia Strigari and medical professionals.
- <sub>1146</sub> • Since all patient had at least 2 CT exams only the closest date to the hospital  
<sub>1147</sub> admission date was taken. When more exams were performed on the same

date all of them were initially taken. At first only chest or abdomen CTs were taken regardless of the acquisition protocol used.

Feature	counts	freqs	categories	Feature2	counts3	freqs4	categories5	Feature7	counts8	freqs9	categories10
Obesity	363	83%	0	HRCT performed	479	100%	1	MulBSTA score total	6	1%	0
Obesity	73	17%	1	High Flow Nasal Cannulae	382	88%	0	MulBSTA score total	7	2%	2
qSOFA	226	52%	0	High Flow Nasal Cannulae	54	12%	1	MulBSTA score total	7	2%	4
qSOFA	178	41%	1	Bilateral Involvement	33	8%	0	MulBSTA score total	50	11%	5
gSOFA	29	7%	2	Bilateral Involvement	403	92%	1	MulBSTA score total	5	1%	6
qSOFA	3	1%	3	Respiratory Failure	231	53%	0	MulBSTA score total	72	17%	7
SOFA score	28	6%	0	Respiratory Failure	205	47%	1	MulBSTA score total	9	2%	8
SOFA score	114	26%	1	DNR	413	95%	0	MulBSTA score total	111	25%	9
SOFA score	144	33%	2	DNR	23	5%	1	MulBSTA score total	1	0%	10
SOFA score	72	17%	3	ICU Admission	359	82%	0	MulBSTA score total	61	14%	11
SOFA score	42	10%	4	ICU Admission	77	18%	1	MulBSTA score total	7	2%	12
SOFA score	23	5%	5	Sub-Intensive care unit admission	336	77%	0	MulBSTA score total	70	16%	13
SOFA score	7	2%	6	Sub-Intensive care unit admission	100	23%	1	MulBSTA score total	17	4%	15
SOFA score	3	1%	7	Death	358	82%	0	MulBSTA score total	2	0%	16
SOFA score	3	1%	8	Death	78	18%	1	MulBSTA score total	8	2%	17
CURB65	141	32%	0	Febbre	186	43%	0	MulBSTA score total	1	0%	18
CURB65	141	32%	1	Febbre	250	57%	1	MulBSTA score total	2	0%	19
CURB65	120	28%	2	Sex	151	35% Female	Lung consolidation	211	48%	0	20
CURB65	30	7%	3	Sex	284	65% Male	Lung consolidation	225	52%	1	21
CURB65	4	1%	4	Sex	1	0%	Hypertension	194	45%	0	22
MEWS score	27	6%	0	Ground-glass	54	12%	0	Hypertension	242	56%	1
MEWS score	143	33%	1	Ground-glass	382	88%	1	History of smoking	348	80%	0
MEWS score	127	29%	2	Crazy Paving	337	77%	0	History of smoking	88	20%	1
MEWS score	82	19%	3	Crazy Paving	99	23%	1	NIV	371	85%	0
MEWS score	33	8%	4	O2-therapy	66	15%	0	NIV	65	15%	1
MEWS score	13	3%	5	O2-therapy	370	85%	1				2
MEWS score	10	2%	6	cPAP	368	84%	0				3
MEWS score	1	0%	7	cPAP	68	16%	1				4

**Figure 2.1:** Description of clinical label dataset, in sets of four columns there's the clinical feature, the total count of occurrences, the percentage over the final dataset and the possible values the feature could take

usare excel come immagine non è il massimo

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

non mi piace la forma, troppo personale. Corretta così va meglio?

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

1170 in how much they add for each condition and the set of symptoms the check  
1171 for.

- 1172 (a) **MulBSTA**: This score accounts for **M**ultilobe lung involvement, absolute  
1173 **L**ymphocyte count, **B**acterial coinfection, history of **S**moking, history of  
1174 **h**yper**T**ension and **A**ge over 60 yrs. [11]
- 1175 (b) **MEWS**: Modified Early Warning Score for clinical deterioration. Com-  
1176 puted considering systolic blood pressure, heart rate, respiratory rate,  
1177 temperature and AVPU(Alert Voice Pain Unresponsive) score. [31]
- 1178 (c) **CURB65**: **C**onfusion, blood **U**rea Nitrogen or Urea level, **R**espiratory  
1179 **B**lood pressure, age over **65** years. This score is specific for pneu-  
1180 monia severity [34]
- 1181 (d) **SOFA**: **S**equential **O**rgan **F**ailure **A**sessment score. Considers various  
1182 quantities from all systems to assess the overall state of the patient,  
1183  $\text{PaO}_2/\text{FiO}_2$ <sup>1</sup> for respiratory system, Glasgow Coma scale<sup>2</sup> for ner-  
1184 vous, mean pressure for cardiovascular, Bilirubin levels for liver, platelets  
1185 for coagulation and creatine for kidneys [2]
- 1186 (e) **qSOFA**: quick SOFA. Only considers pressure, high respiratory rate and  
1187 the low values in the Glasgow scale.

1188 This procedure produced a starting cohort of  $\sim$ 700 patients which, having all  
1189 various images available, created a huge set of  $\sim$ 2200 CT scans. Since this analysis  
1190 is focused on radiomics there is an evident need for as much consistency as possible  
1191 in the images analysed. For this reason all CTs taken with medium of contrast were  
1192 excluded, since they would have brightnesses not indicative of the disease, and for  
1193 every patient only images with thin slice reconstruction were considered.

1194 More specifically only images with slice thickness of 1 o 1.25 mm<sup>3</sup> along the  
1195 z-axis were taken into consideration, which meant excluding all the 1.5,2,2.5 and 5  
1196 mm slice thicknesses.

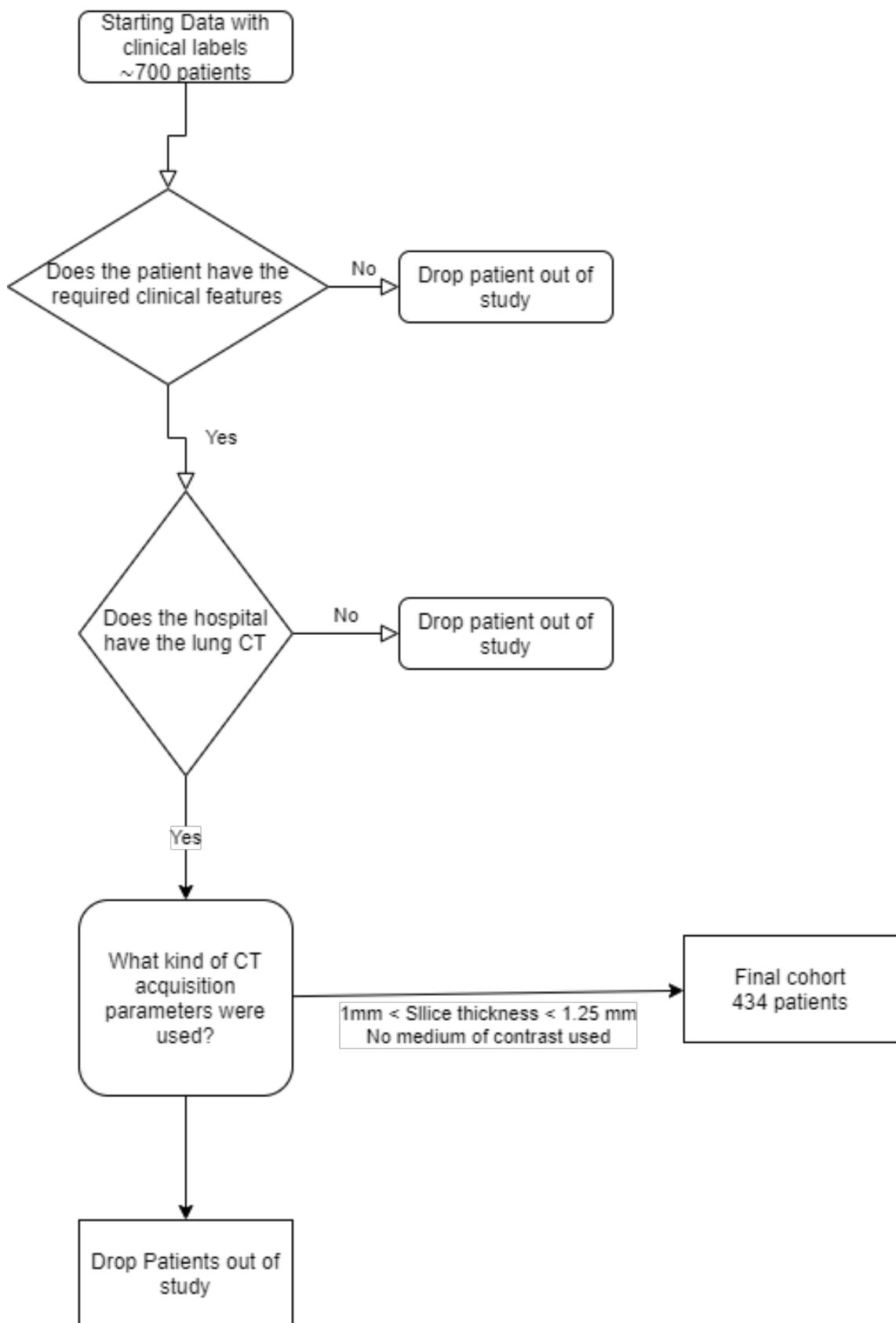
1197 All images were segmented using a tool provided by *Azienda ospedaliero-universitaria*  
1198 *di Bologna - Policlinico Sant'Orsola-Malpighi*.

1199 Per dott.ssa Strigari: Metto il nome del programma? Va citato in qualche modo  
particolare?

<sup>1</sup>Very unrefined yet widely used indicator for lung dysfunction

<sup>2</sup>GCS 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

<sup>3</sup>This meant that only exams called 'Parenchima' or 'HRCT' were included. Throughout the internship 'parenchima' has always appeared in contrast with 'mediastino'. These two keywords are used in the phase of reconstruction of the raw data to identify reconstructions with specific properties. Parenchima is used for finer reconstruction of lung specifically, the requiring professional uses these images to look for small nodules with very high contrast and, to do so, the reconstruction allows some noise to achieve the best resolution possible. Mediastino is used in the lung, as well as other regions, to look for bigger lesions but with low contrast. As such the 'mediastino' reconstruction compromises a worse spatial resolution for a better display of contrast, visually speaking the first images are more coarse and noisy while the second are smoother. It should be noted that even with the same identifier, be it HRCT parenchima or others, the machines on which the exams were made were different and had different proprietary convolutional kernels used for reconstruction.



**Figure 2.2:** Flowchart of the patient selection procedure

Overall this left the final study cohort to be composed of 435 patients, all descriptions and analyses are related to this cohort.

The same software used for segmentation allowed the extraction of the radiomic features from the segmented volumes, even if it did not allow the extraction of the segmentation masks nor any changes in the segmentation parameters, which seems a region growth algorithm mixed with thresholding. 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  $\sim 200$  features, which were 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 ones were the aforementioned scores and the boolean ones were SEX OF THE PATIENT, OBESITY STATUS, IF THE PATIENT HAD A FEVER<sup>4</sup> as of hospital admission and whether or not the patient suffered of HYPERTENSION
- The remaining features, namely those in 2.1 as well as the DEATH status of the patient, were either used as labels or not used at all because they refer to treatments used and not characteristics of the patient. As such these features, while plausibly correlated to the clinical outcome, are not really descriptive of the patient as of admission and are not information that can be used to aid professionals at admission to assess the situation

2. Radiomic: These features were all the ones supported by the segmentation software and are pretty much most of those described in [35] with the addition of fat and muscle surface, computed as  $\text{cm}^2$  by counting pixels identified via threshold as fat or muscle tissue in thoracic slices taken at height of vertebra T-12

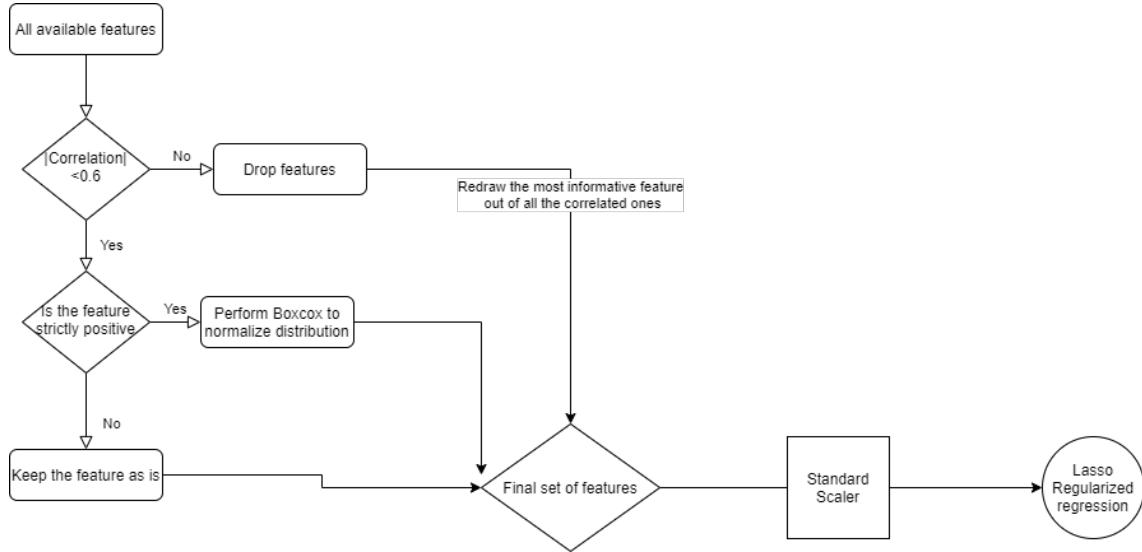
3. Radiological: These features are those that can be derived from CT exams by humans. Namely acquisition parameters, such as KVP and CURRENT, were used to search for eventual correlations between image quality and predictive power of the feature derived from the image while boolean features, such as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

## 2.2 Preprocessing and data analysis

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

---

<sup>4</sup>Defined as body temperature  $> 38^\circ$



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

1241     The first step in the analysis of this data is going to be a lasso regularized  
 1242     regression using either death or ICU admission as target. As mentioned before  
 1243     when operating with regressions it's a necessity that the residuals be normally  
 1244     distributed, a common way to get as close as possible to this hypothesis is to boxcox  
 1245     transform the data. Apart from the data which contained negative values, which  
 1246     cannot be fed into the boxcox transform, all variables have been transformed using  
 1247     this method.

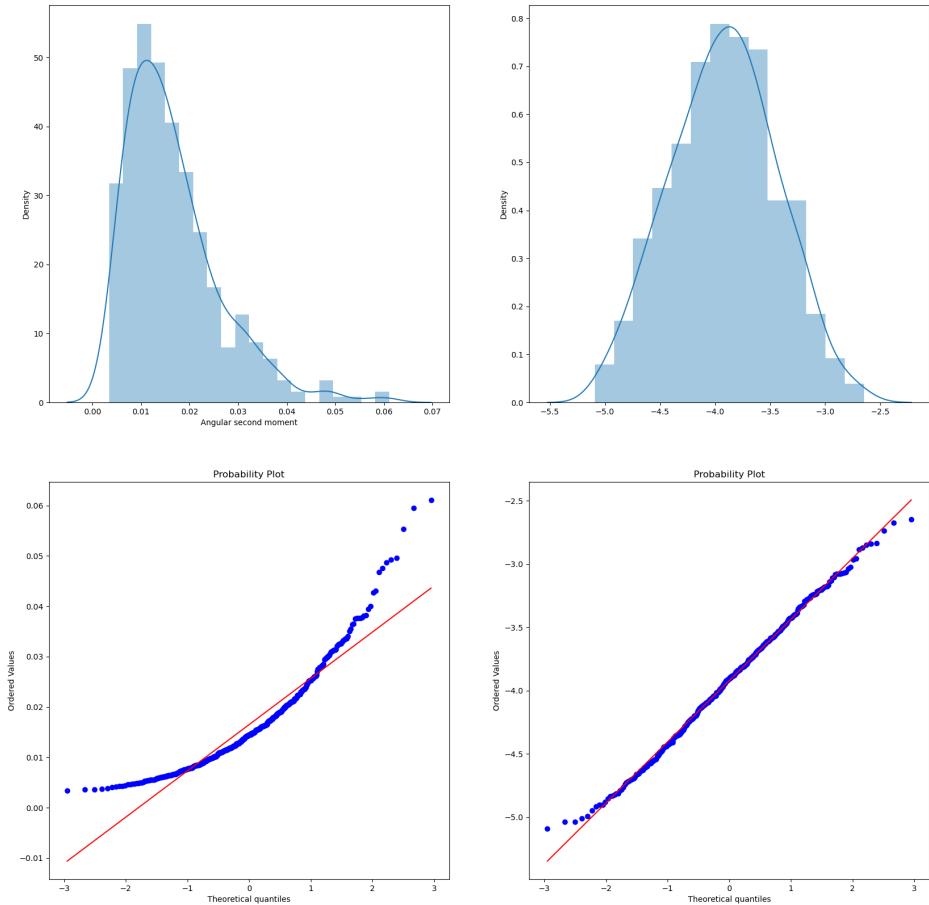
1248     The quantile-quantile plots have been used for a visual check of normality, normally  
 1249     distributed data will populate the bisector of the graph while deviations are  
 1250     symptoms of non-normality. Heavy and light tails are visible as deviations respectively  
 1251     above and below the bisector.

1252     It should be noted that when the data is normally distributed then the trans-  
 1253     form doesn't change much the distribution while, in cases with more heavy tailed  
 1254     distributions, the improvement is clear to see as it can be seen in Figures 2.4 and  
 1255     2.5

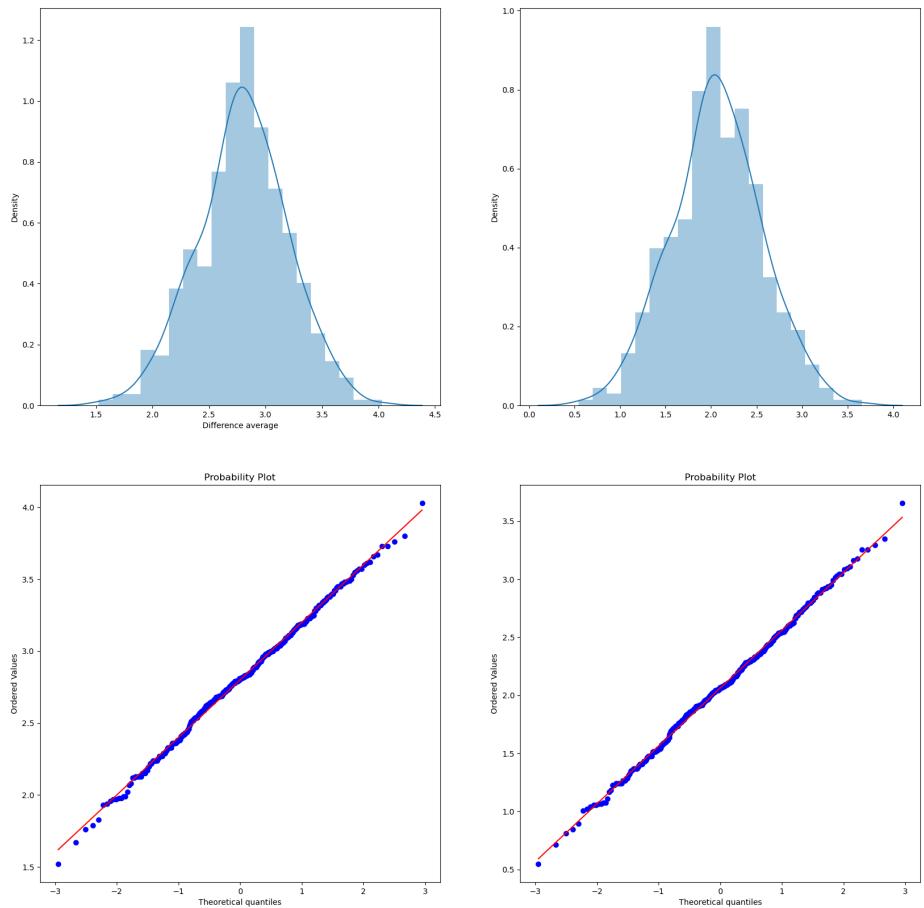
1256     The next preprocessing step has been to apply a *StandardScaler* to all of the  
 1257     features, which corresponds to subtracting the mean and dividing by the standard  
 1258     deviation, in order to center all the features around zero.

1259     The final step in preprocessing has been to reduce the features by using a corre-  
 1260     lation threshold which means that all variables that correlate with another more, in  
 1261     absolute value, than a certain threshold, which has been set to 0.6 in this work, are  
 1262     dropped a priori. A rather important thing to notice is that correlation has been  
 1263     computed using Spearman correlation and not Pearson since the first is invariant  
 1264     under monotone transformations, such as boxcox, and the second is not.

1265     Given the large number of features it's very plausible that at least one of the  
 1266     eliminated features is correlated with all other dropped features but with none of the  
 1267     remaining ones, since a Lasso regularization will be used introducing a few redundant  
 1268     features is not too damaging and the possible benefits outweigh the risks. For  
 1269     this reason a redrawing method has been implemented to add one of the dropped



**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



**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

1270 features, this has been done by choosing the one that most correlates with the label  
1271 being used.

1272 A pivotal point in all of this analysis is that, to obtain reasonable values in the  
1273 cross-validation procedures and to avoid leakage<sup>5</sup> problems, all of the preprocessing  
1274 steps have been done after train-test splitting the data on the train set and then  
1275 applied as defined during training on the test dataset.

1276 When it comes to cross-validation procedure the choice was made to use a stratified  
1277 k-fold approach with k=10, the data is split in 10 parts with the same percentages  
1278 of labels<sup>6</sup> then a model is built by training on 9 of the folds and its performance  
1279 is then tested on the remaining fold. To use the whole dataset for testing a pre-  
1280 diction of it has been built by combinig the predictions on the 10<sup>th</sup> fold for ten  
1281 different models trained on the respective 9 remaining folds.

1282 The lasso model from training on the 9 folds is actually chosen as the model with  
1283 the hyperparameters that give the best performance with another 10-fold crossvali-  
1284 dation. This has been obtained by using *cross\_val\_predict* on a model obtained by  
1285 including a *LassoCV* step inside a *pipeline* from scikit-learn library in python.

1286 The performance of the cross-validated predictions that, when built this way,  
1287 is much more representative of the real-world performance of the model, has been  
1288 evaluated using ROC curves and AUC. Different models have been compared with  
1289 a Delong test[8] for the significance of difference in the ROC curves.

1290 The second analysis method used was RandomForest. As said before in this case  
1291 no preprocessing was needed nor has been done, however particular care was taken in  
1292 handling the imbalances in the dataset by using SMOTE [6] once again being careful  
1293 to avoid leakage. The performance of this model was evaluated using confusion  
1294 matrices which, at a glance, provide very much information on the situation of the  
1295 data. To facilitates the comparison of the results of the Random Forests with those  
1296 obtained using Lasso regularized regression ROC curves were also made.

1297 As a standalone method a Cox Proportional-Hazard model was used on the  
1298 standard scaled variables remaining after the feature reduction performed through  
1299 correlation thresholding. The score obtained with this procedure was then divided  
1300 using different percentiles and tested using the log-rank test on Kaplan-Meier curves  
1301 relative to the groups built. In the case of this thesis the time variable was repre-  
1302 sented by the days of hospitalization computed using the dates of admission and  
1303 discharge from the hospital provided by the hospital itself. The idea of resorting to  
1304 Kaplan-Meier curves was also used with other single variables to see if they had any  
1305 effect.

### Ad esempio la questione della divisione in ondate

1306  
1307 Finally a few dimensionality reduction techniques, namely the unsupervised  
1308 PCA[9] and Umap [22] and the supervised PLS-DA[4], have been used to under-  
1309 stand better the state of the data and further explain some of the obtained results.  
1310 These peculiar analyses will be reported in the Appendix

---

<sup>5</sup>This term is used in the field of Machine Learning. It refers to models being created on information that comes from outside the training data.

<sup>6</sup>The stratified in the name refers to this property. This method is useful when dealing with unbalanced datasets, such the one under analysis, in which the label has an uneven 15-85% frequency of occurences of the two labels

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

1311

1312 This concludes the discussion on the methodologies used and leads perfectly in  
1313 the discussion of the results.

1314

## Chapter 3

1315

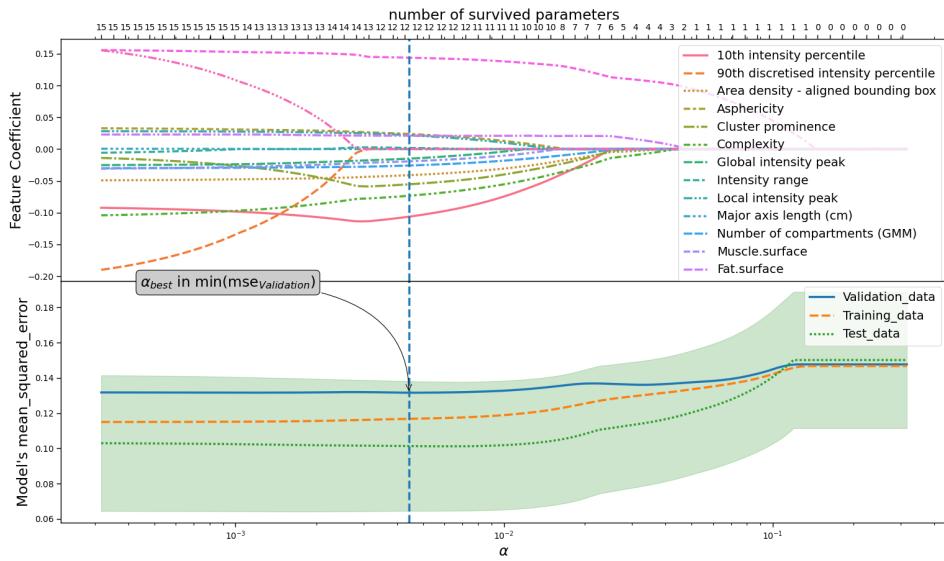
# Results

1316 In this chapter the result obtained with the methods explained in the previous  
1317 chapters will be briefly presented to ease in the discussion in the final chapter.

1318

### 3.1 Feature selection through Lasso regularization and clinical outcomes prediction using 1319 regression

1321 When it comes to lasso regression usually graphs are reported that show the convergence  
1322 of the parameters to the final value. Since the real information of this process  
1323 is the value to which the coefficients converge only these values will be presented in  
1324 tables and the ROC curves, with respective AUCs, will be provided. An example  
1325 of the graph i'm talking about is the one visible in Figure 3.1.



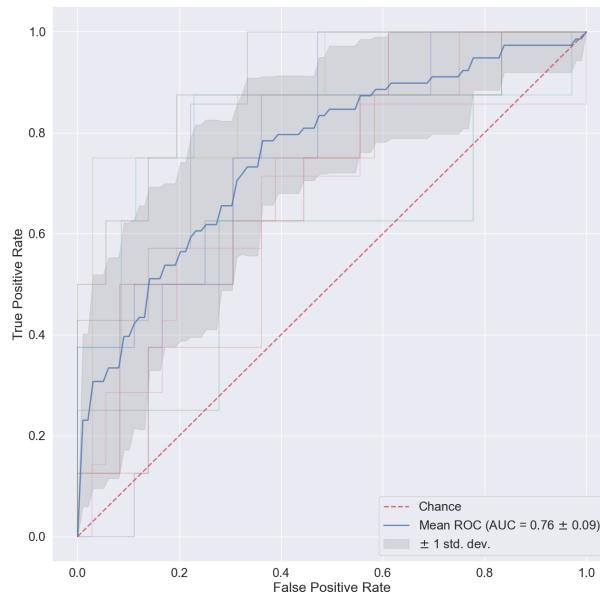
**Figure 3.1:** Example of graph representing the convergence of the coefficients in a lasso procedure. The final model is chosen by looking at the lowest value in mean squared error computed on the testing dataset. This graph is relative to only radiomic features with the model using DEATH as label.

1326 All of the graphs with respective captions will be, maybe, put in the final ap-  
 1327 pendix. Finally it seems useful to report the following contingency table that gives  
 1328 an idea on how superimposed the labels on death and ICU ADMISSION are.

		ICU ADMISSION	
		0	1
DEATH	0	311	47
	1	48	30

### 1330 3.1.1 Using Death as label

1331 Starting to predict the death outcome of the patient different groups of features  
 1332 have been used, first of all only the radiomic features have been used. The ROC  
 1333 curve obtained with the radiomic features is the one in Figure 3.2. The curve in  
 1334 bold is an average curve obtained by aggregating the ten curves relative to each of  
 1335 the folds used in testing and the gray band represents a  $\pm 1$  standard deviation.



**Figure 3.2:** 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.

1336 The model was built using the coefficients reported in 3.1 reaches a  $AUC =$   
 1337  $0.76 \pm 0.09$ . The features inside the tabular, as well as those in the tabulars that  
 1338 follow, will have the coefficients in descending order by absolute value so that the  
 1339 top features are the most relevant within it's relative model.

1340 Before commenting more in depth the performance of this model it seems appropriate to see at least the other models built with singular features groups, so, when

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

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

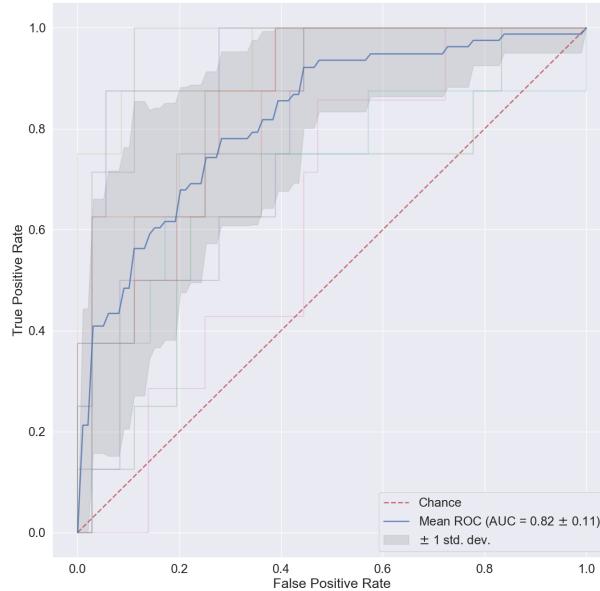
1342 it comes to the clinical features, the results reported in Figure3.3 and Table3.7 have  
 1343 been obtained. All the results will be put together for ease in Table 3.5.

1344 And finally, considering the radiological features Figure3.4 and Table 3.8 are  
 1345 obtained.

1346 The first thing to notice is that the radiological features, when considered alone,  
 1347 have close to null predictive power. This is reasonable for at least part of the  
 1348 features because there is no reason for acquisition parameter to actually influence the  
 1349 outcome of the patient. When it comes to the radiologically determined quantities,  
 1350 such as GGO, Crazy paving, lung consolidation and bilaterlaity even if one would  
 1351 expect these to be relevant their distribution across the dataset is not condusive the

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

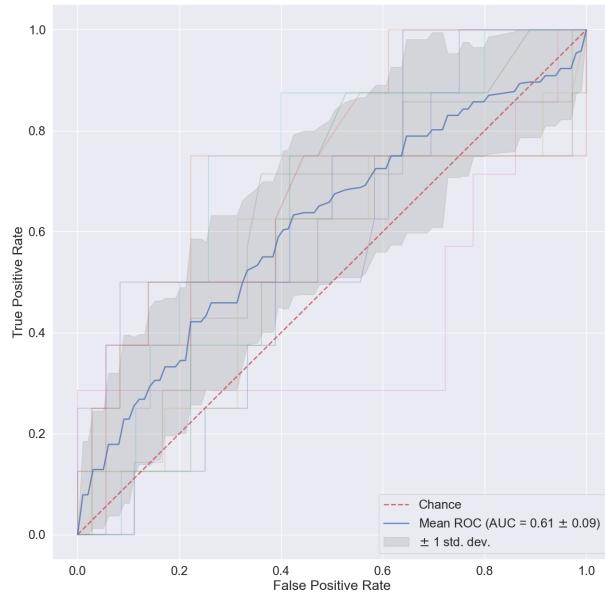
Feature Name	Importance
Intercept	0.178899
Age (years)	0.116771
Respiratory Rate	0.082292
Sex	-0.037591
Febbre	-0.022923
Hypertension	-0.000000
History of smoking	-0.000000
Obesity	0.000000



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

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

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



**Figure 3.4:** 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.

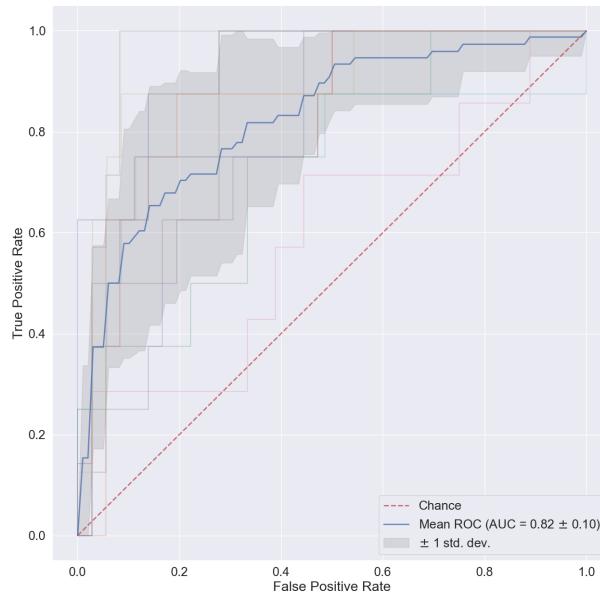
good predictions. In fact 88% of patients had GGO, 50% of all patient had Lung consolidation, 77% of all patients did not have Crazy paving and 92% had bilateral involvement.

hen it comes to clinical features, categories that performs better when considered singularly, nothing ground breaking has been obtained. Age, Respiratory rate and sex are the most relevant features and are all very much in concordance with what is expected. Finally radiomic features preform slightly worse than the clinical features. To see if the feature obtained are, at least, reasonable a quick explanation is needed. This will be done only for the top performing features:

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

1372 To summarize, since all of the features are computed on the whole lung segmen-  
1373 tation, it seems that the some information on the distribution of gray levels as well  
1374 as some textural information inside the whole lung are important. It also seems that  
1375 some information on the shape of the organ itself is also relevant.

1376 One would expect that when combining all of the available features, i.e. by  
1377 building a model using the previous clinical, radiomic and radiological features,  
1378 the performance should somewhat rise especially given the fact that clinical and  
1379 radiomic features have almost the same performance. The combined results can be  
1380 seen in Figure 3.5 and Table 3.9.



**Figure 3.5:** ROC curves obtained with crossvalidation procedure using all the available features. Model obtained with Lasso regularization of a linear regression modelling death

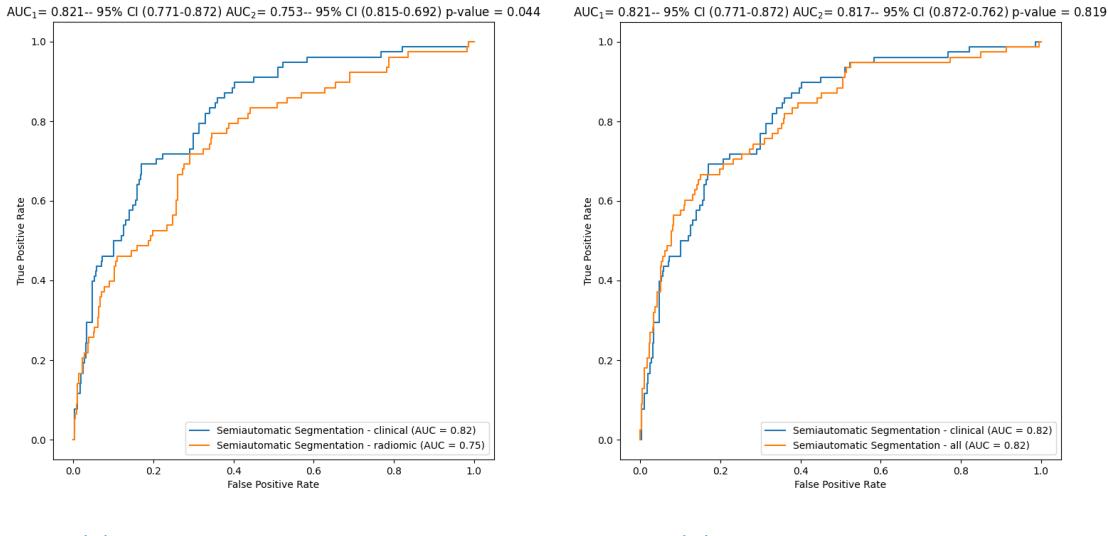
1381 In spite of what the expectations were the performance not only doesn't improve  
1382 but it doesn't even change. To be sure of this claim a Delong test was used to  
1383 compare pairwise the receiver operator curves and their respective AUCs. The null  
1384 hypothesis of this test is that the two models are the same, hence a p-value smaller  
1385 than 0.05 means that the curves and their aucs are statistically different. The results  
1386 from this analysis can be seen in 3.6

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

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

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

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



(a) Comparison clinical-radiomic

(b) Comparison clinical-all

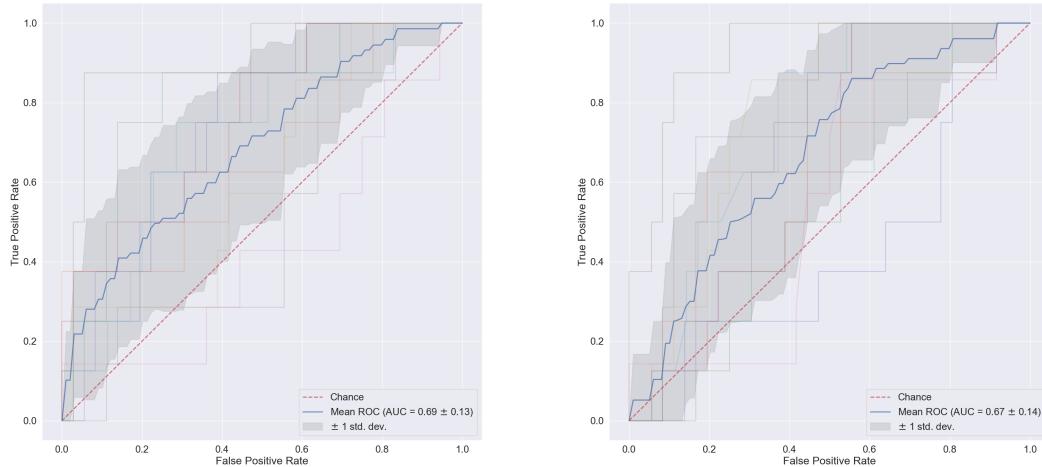
**Figure 3.6:** Comparison between ROC curves for clinical vs radiomic curves (a) and clinical vs all (b). The p-values are obtained with a Delong Test

1387 As one would have expected the models from radiomic and clinical features are  
 1388 different while the ones built using clinical and all features are not statistically  
 1389 different. Inspecting the coefficient of the parameters there are a few perplexing  
 1390 things to notice and a few reassuring ones. First of the reassuring facts is that the  
 1391 most relevant clinical features are still relevant in this combined model. Then, as one  
 1392 would expect, the radiological features retain some importance when combined with  
 1393 the others. However, when it comes to perplexing behaviours, the most concerning  
 1394 fact is that the radiomic features have mostly lost all relevance in the model which  
 1395 is surely unexpected.

1396 Some possible explanations will be given in the concluding remarks at the end  
 1397 of this subsection. as well as in the final chapter of the thesis.

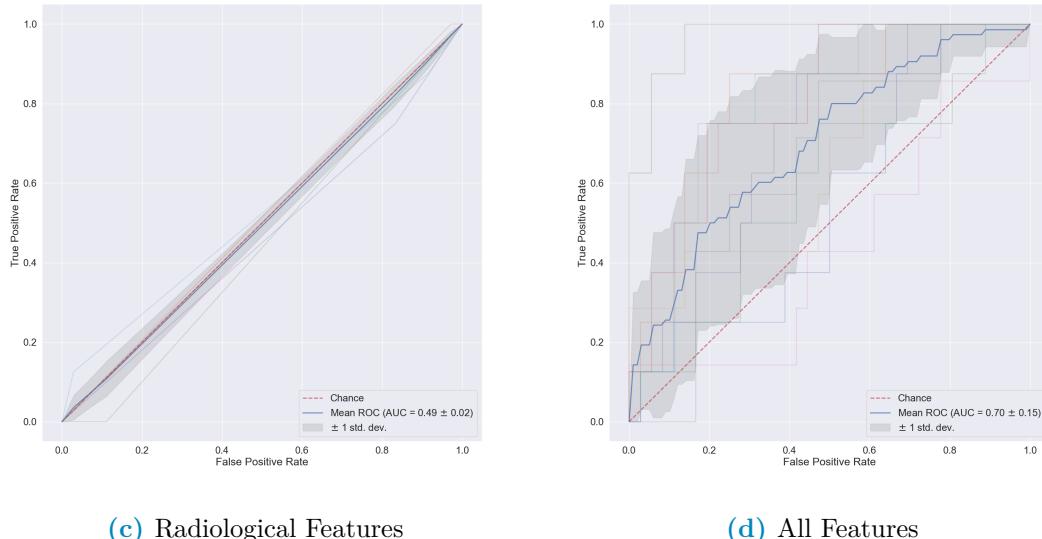
### 1398 3.1.2 Using ICU Admission as label

1399 In trying to predict if the patient will be admitted in the Intensive Care Unit of the  
 1400 hospital the same procedure as before has been used. The ROC curves obtained  
 1401 with the various features are reported in Figure 3.7. Just like before the curve in  
 1402 bold is an average curve obtained by aggregating the ten curves relative to each of  
 1403 the folds used in testing and the gray band represents a  $\pm 1$  standard deviation.  
 1404 Following the blueprint of the previous subsection, all of the results will be presented  
 1405 and then briefly discussed



(a) Radiomic Features

(b) Clinical Features



(c) Radiological Features

(d) All Features

**Figure 3.7:** 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

1406       Compared to before the performance is definitely worse. The radiological features  
 1407       have the same performance of a random variable, which is not that concerning given  
 1408       their expected impact on gravity of the clinical picture of the patient. Even if  
 1409       superfluous a Delong test was used to confirm that the hypothesis of the curves  
 1410       being equal could not be rejected. When it comes to the relevant features in each  
 1411       model the following can be deduced:

- 1412       • For the clinical features all of them have a role in the prediction. The only  
 1413       surprising fact, even if it keeps a certain degree of plausibility, is that age is the  
 1414       less relevant out of the available features when it comes to ICU ADMISSION .

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

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

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

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

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

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

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

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

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

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

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

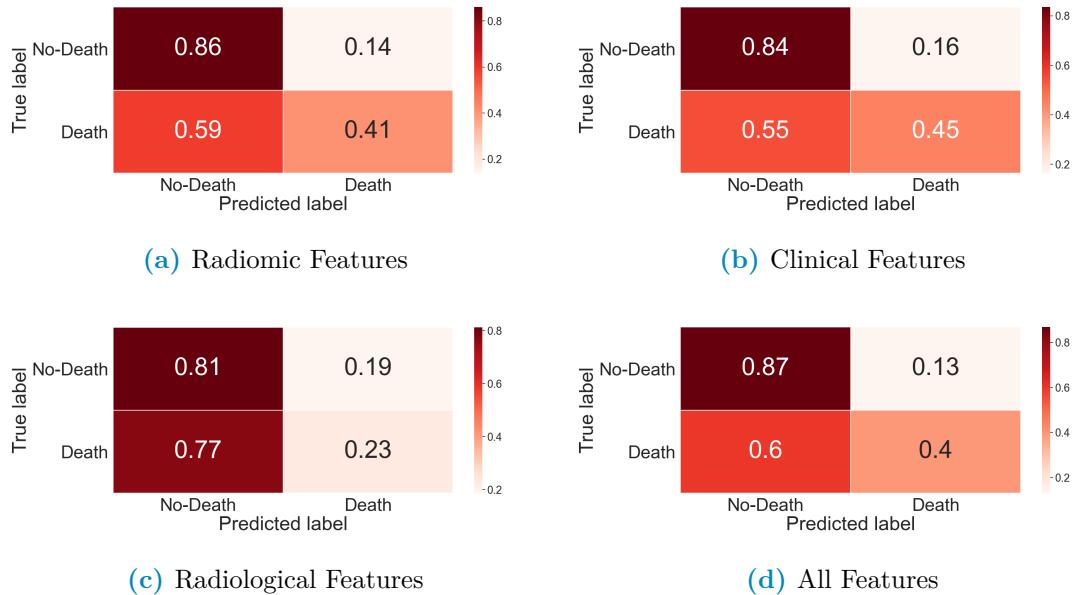
## 3.2 Classification of patients using Random forests

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

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

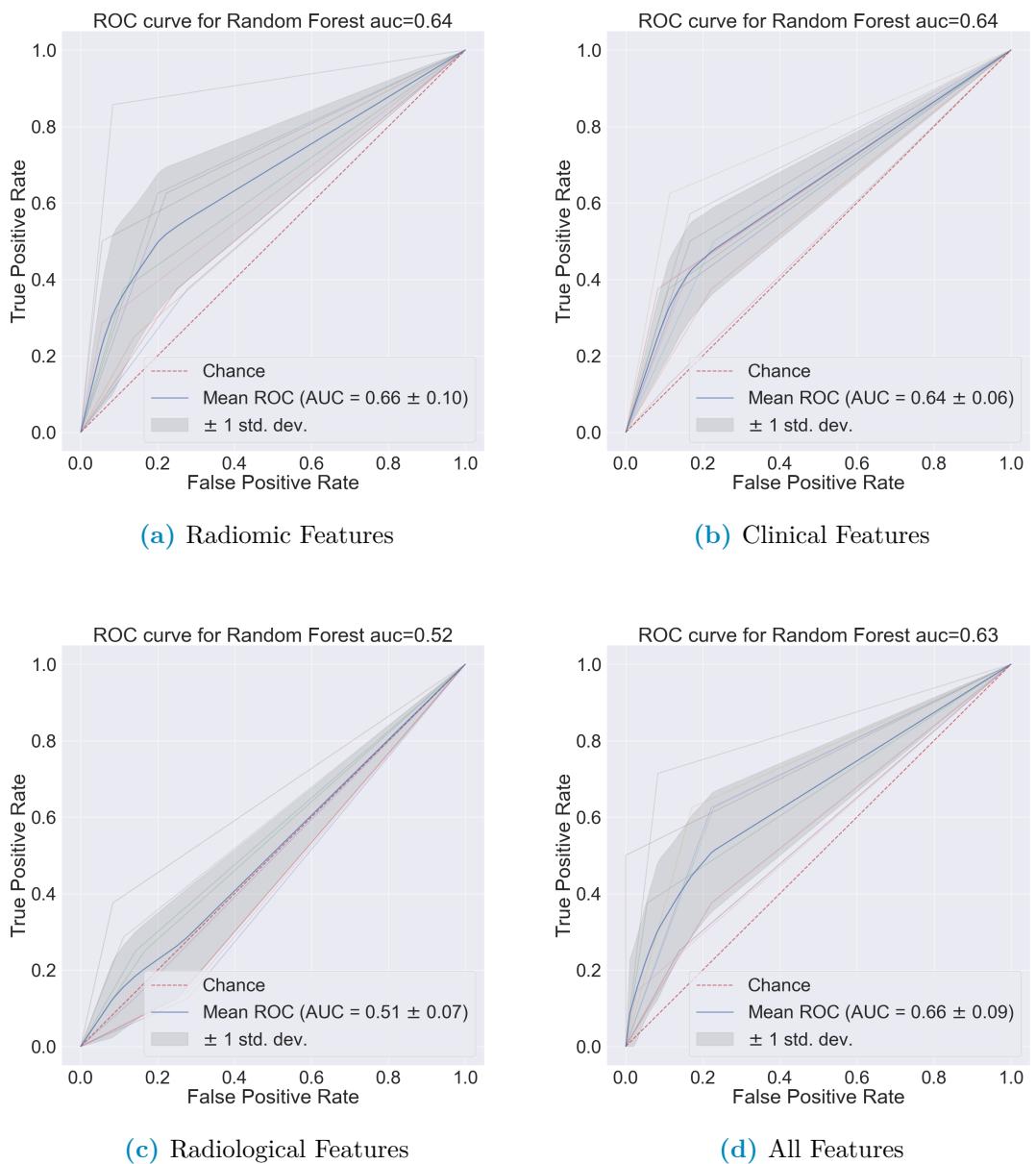
### 3.2.1 Classifying the oucome Death

For the sake of brevity all of the results will be reported and then discussed. Since RF classifiers use all of the availalble features it is very space consuming to report a table with all of the importances for the radiomic features as well as those used in model with all the features. These two will be found in the appendix 5.1 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 the reported

1435 Even without looking at the ROC curves it's plain to see that the data at hand  
 1436 is proving to be difficult for this model. Much like before radiological features alone  
 1437 are useless. In evaluating these confusion matrices it should be kept in mind that the  
 1438 data is heavily unbalanced, since only  $\sim 15\%$  of the patient died or were admitted  
 1439 in the ICU. Even when using SMOTE in the training phase to correct this probelm  
 1440 it seems that the classifiers learns that it's optimal to guess that someone is alive.  
 1441 When it comes to the ROC curves Figure 3.9 and Table 3.11 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 3.11:** Recap table with the performance of the various families of features

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

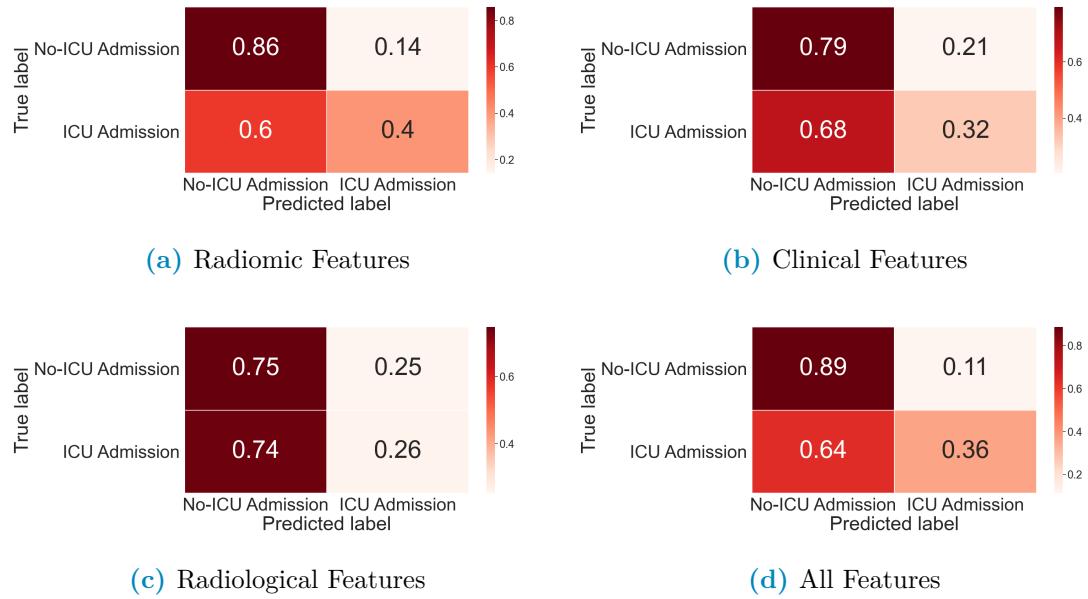
RF_importances		RF_importances
(a) Radiological Features		XRayTubeCurrent
Age (years)	0.463821	0.800101
Respiratory Rate	0.287735	0.043942
Febbre	0.084276	0.039768
Sex_bin	0.079571	0.032511
Hypertension	0.033435	0.031362
History of smoking	0.029404	0.030423
Obesity	0.021758	0.021893
		HRCT performed
(b) Clinical Features		0.000000

**Figure 3.10:** Importances estimated by random forest

1442 Once again the curves are evidently not statistically different. This counterintuitive behaviour seems to be constant across the two implemented methods and  
 1443 also across different labels tried. An attempt to explain this phenomenon will be  
 1444 postponed to the end of the next subsection  
 1445

### 1446 3.2.2 Classifying the outcome ICU Admission

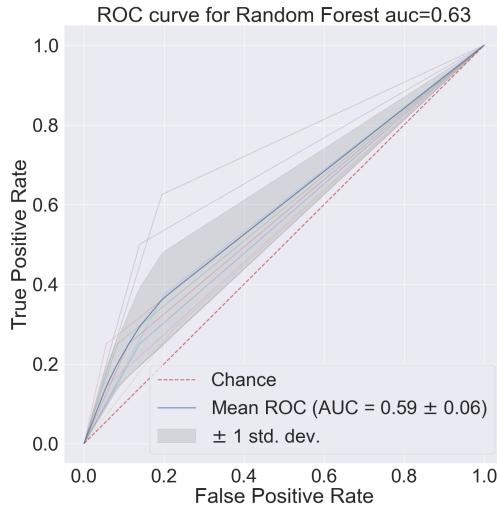
1447 Even when using the admission in the ICU the performance remains pretty much  
 1448 the same, so the comments would still be the same as before



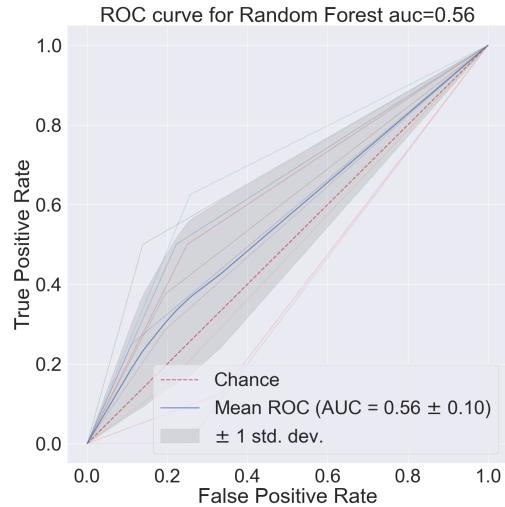
**Figure 3.11:** Confusion matrices for Random Forest cross-validated predictions after training on Synthetically oversampled data predicting ICU ADMISSION . All of the available feature families are the reported

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

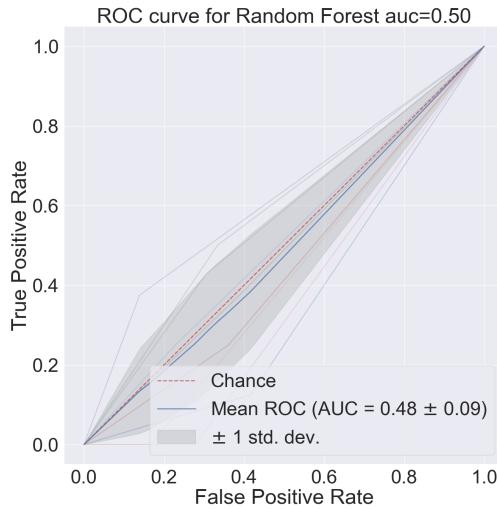
Features used	mean AUC $\pm$ std
Radiomic	$0.62 \pm 0.08$
Clinical	$0.56 \pm 0.08$
Radiological	$0.51 \pm 0.11$
All	$0.64 \pm 0.09$



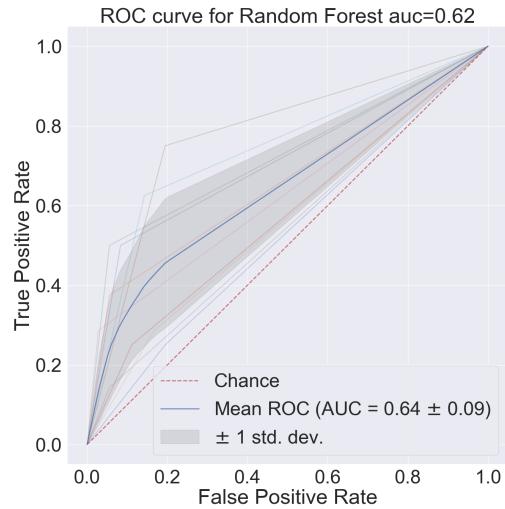
(a) Radiomic Features



(b) Clinical Features



(c) Radiological Features



(d) All Features

**Figure 3.12:** Cross-validated ROC curves built with Random forest classifier predictions of DEATH . Performances of all variable families are reported

### <sup>1449</sup> 3.2.3 Using survival analysis

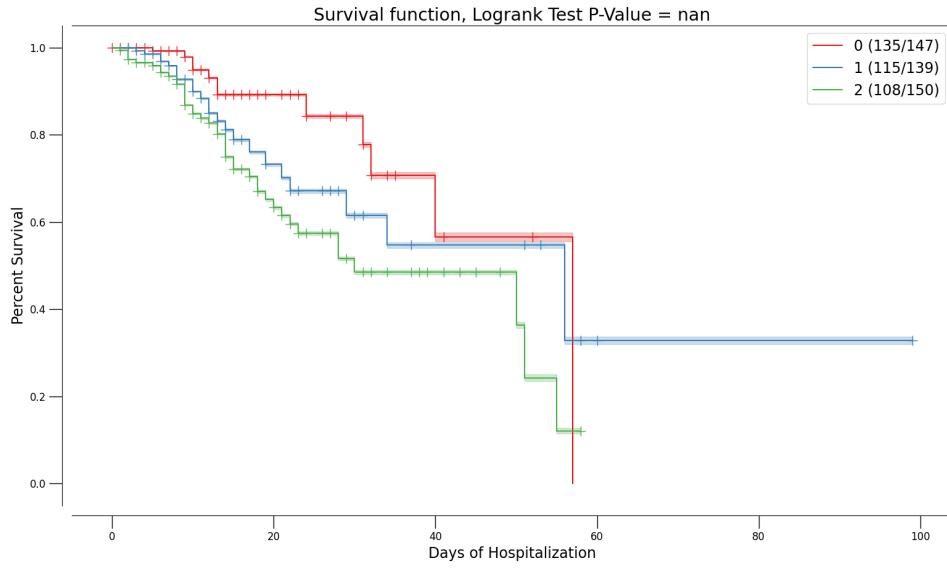
<sup>1450</sup> Following the preprocessing steps delineated before a Cox Proportional-Hazard pro-  
<sup>1451</sup> duces the results presented in Table 3.13.

<sup>1452</sup> As explained in section 1.4 the relevant columns are the coeff column, that ex-  
<sup>1453</sup> pressed percentual difference of survival, and the p column, that indicate the sig-  
<sup>1454</sup> nificance of the first value. It turns out that, out of the reduced variables fed to  
<sup>1455</sup> the Cox model, the most relevant are: SEX, ASPHERICITY, FATSURFACE, NOR-  
<sup>1456</sup> MALIZED ZONE DISTANCE NON-UNIFORMITY AND KVP. ZONE DISTANCE NON  
<sup>1457</sup> UNIFORMITY measures distribution of zone counts over the different zone distances,  
<sup>1458</sup> it is low when the count relative to the zones are equally distributed along zone  
<sup>1459</sup> distances, ASPHERICITY quantifies how much the segmented region deviates from a  
<sup>1460</sup> sphere. In this case SEX and FATSURFACE and NORMALIZED ZONE DISTANCE NON-  
<sup>1461</sup> UNIFORMITY can be reasonable variables to expect, however KVP, ASPHERICITY  
<sup>1462</sup> seem quite strange.

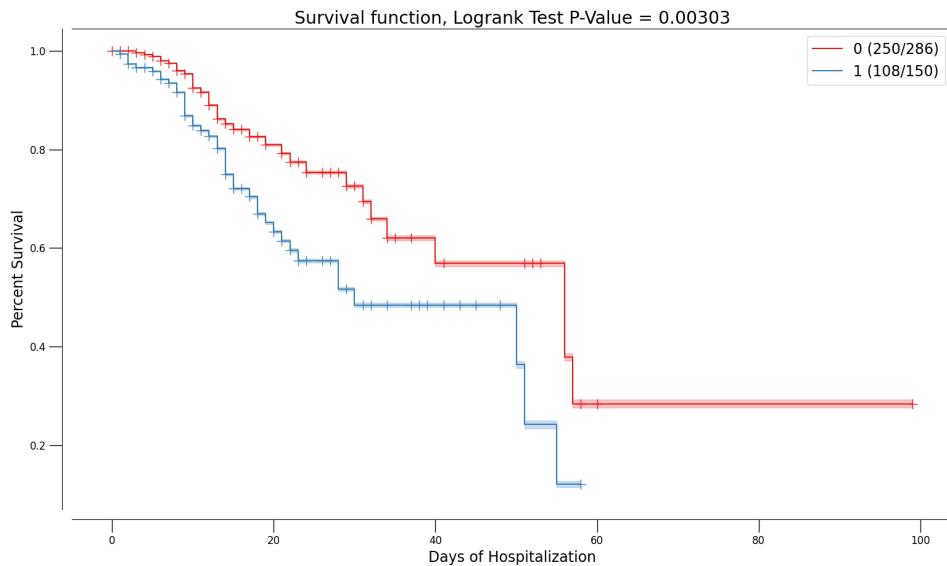
<sup>1463</sup> A score was built automatically using the predict method of the CoxPH fitter and  
<sup>1464</sup> assigned to each patient of the dataset using the previously described cross-validated  
<sup>1465</sup> prediction procedure. To see if the prediction was representative of differences in  
<sup>1466</sup> the individuated populations first the Kaplan-Meier curves according to thirds in  
<sup>1467</sup> the score distribution were used and then the score was binarized using the 66<sup>th</sup>  
<sup>1468</sup> percentile in the score distribution as threshold. The results of this procedures are  
<sup>1469</sup> reported in Figure 3.13

**Table 3.13:** Results obtained with CoxPH fitter from lifelines library

covariate	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	z	p	-log2(p)
Lung consolidation	0.166411	1.181058	0.142506	-0.112895	0.445717	0.893244	1.561610	1.167749	0.242908	2.041517
Ground-glass	0.109946	1.106217	0.134109	-0.161902	0.363794	0.850524	1.438778	0.752719	0.451619	1.146822
Crazy Paving	0.064744	1.060886	0.140817	-0.211253	0.340740	0.809569	1.405088	0.459771	0.645680	0.631108
Bilateral Involvement	-0.026048	0.974288	0.121990	-0.265144	0.213048	0.767096	1.237444	-0.213524	0.830918	0.267222
SliceThickness	-0.008439	0.991597	0.164518	-0.330888	0.314010	0.718286	1.368904	-0.051293	0.959092	0.060259
KVP	0.376184	1.456715	0.139983	0.1010821	0.650546	1.107185	1.916587	0.007202	7.117333	
XRayTubeCurrent	-0.272076	0.761796	0.178915	-0.022743	0.078591	0.536471	1.081762	-1.520693	0.128355	2.962010
Age (years)	-0.016550	0.983587	0.160470	-0.331065	0.297966	0.718158	1.347116	-0.103132	0.917858	0.123657
Hypertension	0.292450	1.339705	0.160211	-0.021557	0.606457	0.978673	1.833822	1.825407	0.067940	3.879603
History of smoking	-0.083840	0.919578	0.139121	-0.356512	0.188882	0.700114	1.207838	-0.602041	0.546747	0.871054
Obesity	0.066777	1.069057	0.170581	-0.267556	0.401110	0.765247	1.493482	0.391469	0.695451	0.523979
Respiratory Rate	-0.010716	0.989341	0.154003	-0.312957	0.291125	0.731574	1.337932	-0.060583	0.944525	0.082338
Sex_bin	-0.366086	0.693443	0.172651	-0.704476	-0.027695	0.494367	0.972685	-2.120377	0.033974	4.879413
Febbre	0.115458	1.122387	0.139878	-0.158698	0.389614	0.853254	1.476411	0.825418	0.409134	1.289354
10th intensity percentile	0.324188	1.382908	0.230611	-0.127801	0.776178	0.880028	2.173151	1.405779	0.159790	2.645753
Area density - aligned bounding box	-0.169228	0.844316	0.185602	-0.533001	0.194544	0.586842	1.214758	-0.911781	0.361884	1.466401
Asphericity	-0.470777	0.624517	0.174449	-0.812691	-0.128863	0.443662	0.870995	-2.698646	0.006962	7.166236
Cluster prominence	-0.011242	0.988821	0.234222	-0.470309	0.447825	0.624809	1.564905	-0.047996	0.961720	0.056312
Complexity	0.214330	1.239032	0.248380	-0.272486	0.701147	0.761484	2.016063	0.862911	0.388186	1.365179
Global intensity peak	0.138264	1.148279	0.165033	-0.185195	0.461723	0.830942	1.586806	0.837795	0.402146	1.314210
Intensity range	-0.196751	0.821395	0.281701	-0.748875	0.355372	0.472898	1.426711	-0.698441	0.484901	1.044237
Local intensity peak	0.132819	1.142044	0.150161	-0.161491	0.427129	0.850875	1.532851	0.884514	0.376419	1.409589
Number of compartments (GMM)	0.197584	1.218821	0.140554	-0.077597	0.473365	0.925338	1.605387	1.407887	0.159164	2.651410
Number of voxels of positive value	0.436757	1.547680	0.284518	-0.120887	0.886134	2.703107	1.535079	0.124764	3.002721	
Fat.surface	-0.425033	0.653748	0.208060	-0.832823	-0.017242	0.434820	0.982906	-2.042835	0.041069	4.605815
Normalised zone distance non-uniformity	0.589917	1.803838	0.242437	0.114750	1.065084	2.901082	2.433282	0.014963	6.062489	



(a) Population divided according to score tertiles

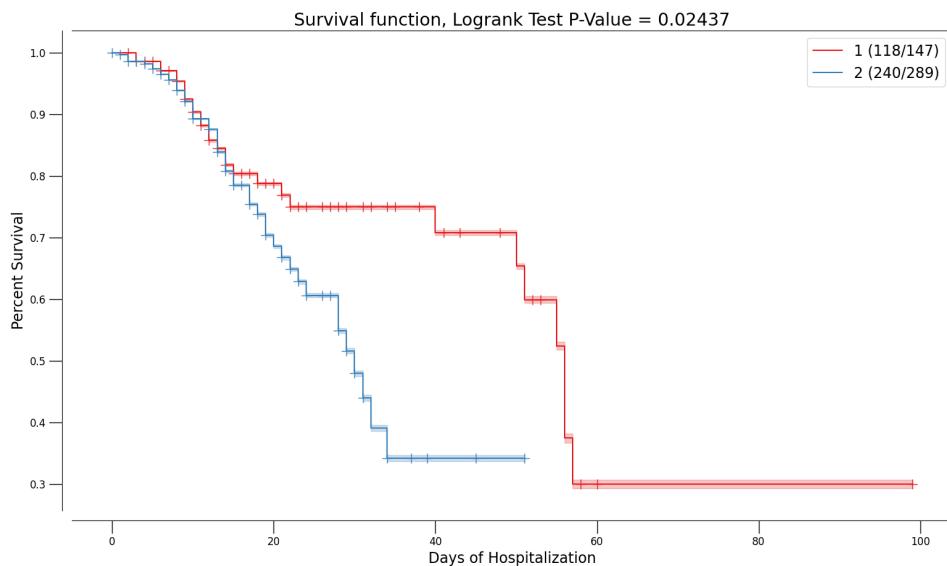


(b) Population divided in two groups 0-66<sup>th</sup> percentile and 66<sup>th</sup> to 100<sup>th</sup>

**Figure 3.13:** Kaplan-Meier curves for populations divided using either tertiles (a) or binarized using 66<sup>th</sup> percentile as threshold (b). The prediction on the whole database is obtained with the aforementioned cross-validation procedure

1470 It can be seen that the groups built in this ways can be used to drive some  
 1471 differences in survival, when binarizing the score obtained with Cox the curves also  
 1472 turn out to be significantly different.

1473 Finally, to better understand the data as well as the situation in the hospital,  
 1474 the population was divided in two subgroups according to the date of admission.  
 1475 The two groups were representatively called 1<sup>st</sup> and 2<sup>nd</sup> wave and the division was  
 1476 drawn on the 20<sup>th</sup> of July 2020 and the results can be seen in Figure 3.14.



**Figure 3.14:** Kaplan-Meyer curves for patient admitted before (red curve) and after (blue curve) 20/07/2020

1477 Non ricordo come mai avevamo fatto la distinzione in ondate

1478 It can be seen that there is a statistical difference in survival between the pa-  
 1479 tients admitted in the first wave vs those admitted in the second. Furthermore this  
 1480 difference is quite perplexing as it seems to indicate that people in the second wave  
 1481 died more than people in the first wave, which seems counter-intuitive given that one  
 1482 would expect the experience from the previous wave to improve performance. The  
 1483 most reasonable explanation for this fact is the change in admission policy as time  
 1484 advanced. Probably in the first wave, when still little was known on *Sars-COVID19*  
 1485 patients, more people were addmitted in less problematic condition whereas in the  
 1486 second wave, having understood better what were the most dangerous cases as well  
 1487 as in an attempt to admitt only those strictly in need, most of the admitted patients  
 1488 were in more critical condition.

1489 It's also possible that this result that has been obtained could be a symptom of  
 1490 subtle differences in the two *Sars-COVID19* manifestations, as if to indicate different  
 1491 variants. Further analysis in this direction could be a followup work of this thesis.

### 1492 3.3 Summarizing the results and further analysis

1493 Forse già questo può andare bene come conclusione?

1494 Having presented all of the results obtained in this thesis it has been seen that:

- 1495 • The chosen preprocessing method followed by Lasso regularized regressions  
 1496 perform overall well when it comes to predicting either DEATH or ICU AD-  
 1497 MISSION

- Random Forest classifiers are consistently worse than Lasso regularized regressions, probably mainly due to the unbalancedness of the dataset at hand.
- Cox proportional Hazard allows us to distinguish at least two groups with statistically different Survival curves.
- Combining clinical, radiomic and radiological information does not determine an improvement when compared to clinical features alone. This result is quite perplexing and it could have multiple causes

It is very difficult to quantify what the expectations for each model were a priori, this transposes to difficulties in trying to diagnose problems in the models when considered singularly. However when combining all of the available features it's very strange that no improvement in performance can be achieved. This would mean that 8 clinical variables provide the exact same amount of information as the total  $\sim 200$  features most of which derived from images which, at least ideally, contain much more accurate information. One could surmise that the analysis methods have been implemented in an incorrect way, yet when two methods implemented differently and separately obtain the same result it reinforces the idea that the problem lies somewhere before the analysis. All of these analyses started from the following hypotheses:

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

It is very clear that the first three hypotheses are verified with a caveat, the performance of the radiomic pipeline hinges on the quality of the images and of the segmentation procedure performed on them. Since the images used in this thesis were, are and will be used by the hospital in routine processes it's close to impossible that all of them have problems, especially because of the preliminary screening done before segmentation. A first possibility is that when trying to segment lungs affected by *Sars-COVID19* the peculiar patterns developed make the task much harder than expected. In fact considering that the method used for this thesis relies on region growth and thresholding methods it's possible that the worst cases end up with unrepresentative segmentations. Another possibility is that the images are not really representative of the situation of the patient. Since one of the prevailing properties of *Sars-COVID19* is the speed with which the clinical picture of the patient can change it's possible that images taken at admission are not as informative of the final prognosis. This problem is, however, unavoidable in the setting of this thesis which has as aim the construction of a model that, exactly at admission, can discriminate between serious and easier cases. Another possibility is that there are two or more subgroups in the patient cohort and the performance on these is widely different, determining an average performance below the expectations. To diagnose if this is the case a few dimensionality reduction techniques have been used

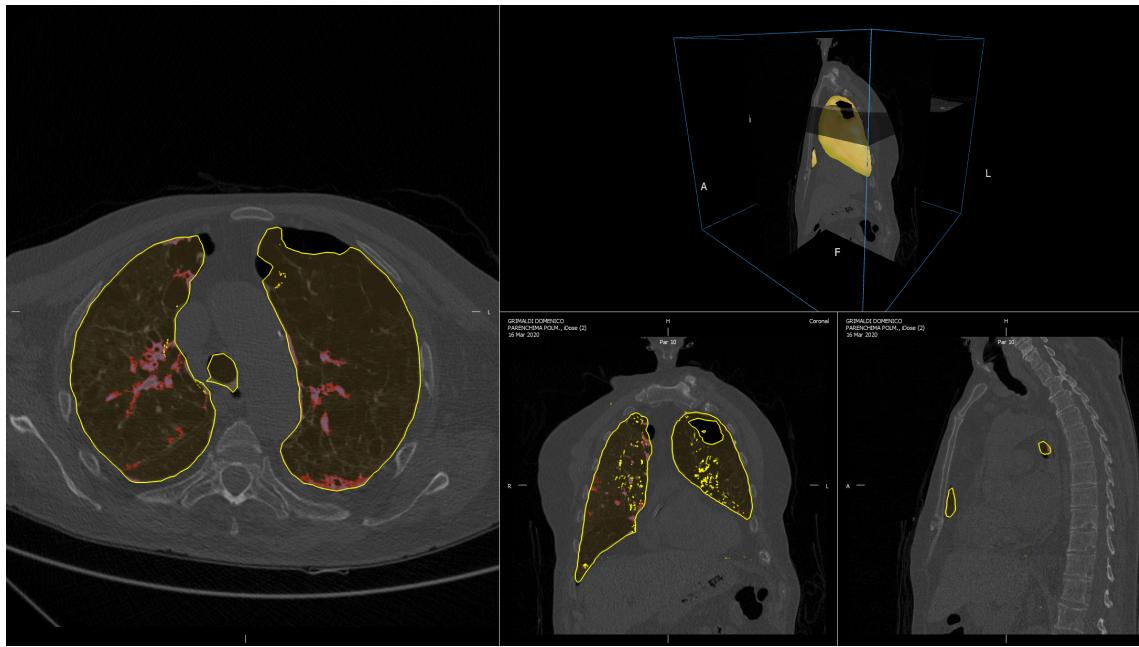
1539 to visualize the data and to prepare for clustering in case of need, all of the results  
1540 of these procedure will be presented in the followig section.

1541 A questo punto magari cito l'appendice con la parte di clustering quando sarà  
pronta

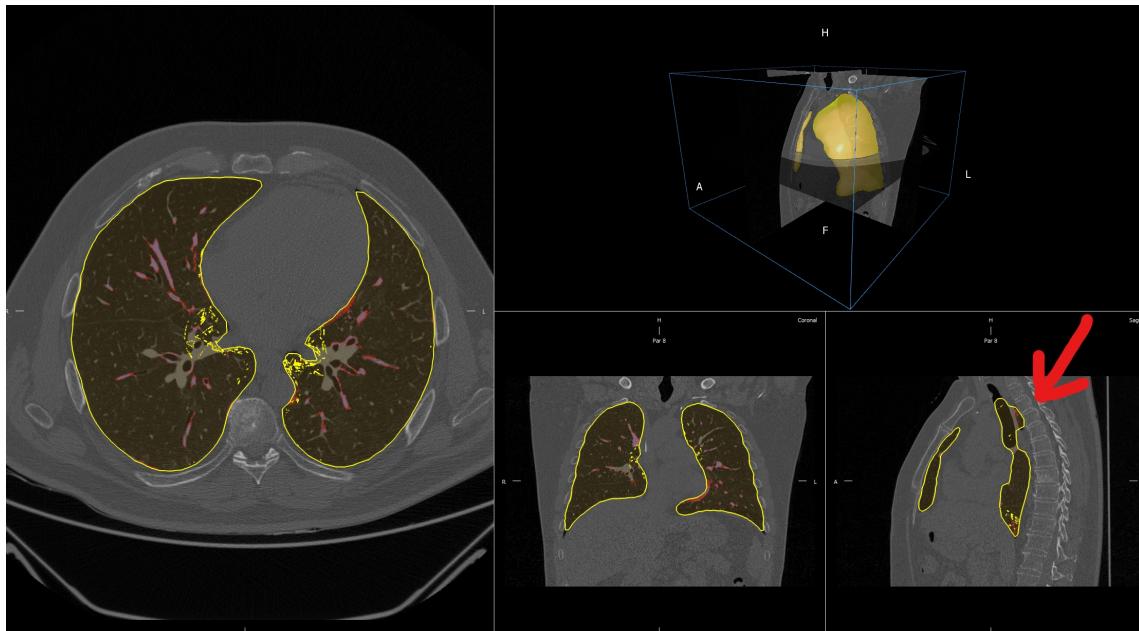
1542 To give a qualitative idea of the situation regarding the segmentations  $\sim 70$  were  
1543 looked at and evaluated as good, unsure or bad.<sup>1</sup> Good segmentations are those  
1544 in which an untrained professional could not see any problems, unsure are those in  
1545 which there are small inaccuracies, such as lungs that connect in some points and  
1546 the inclusion of the trachea. Finally segmentation were classified as bad in cases  
1547 with obvious errors, such as part of the intestine being labelled as lung, damages in  
1548 the lung being labelled as outside tissue or holes in what is supposed to be lung.

---

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



(a) Example of segmentation classified as bad



(b) Example of segmentation classified as dubious

**Figure 3.15:** 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 correct, in the second (b) there are small portion of outside tissue being labelled as lung as well as the whole trachea.

1549 The result of this qualitative analysis can be seen in Table 3.14 in which the  
 1550 incidence of both DEATH and ICU ADMISSION labels is computed in all possible  
 1551 segmenation categories.

1552 This table suggests that there is large difference of very severe individuals which  
 1553 end up being not perfectly segmented. It's possible, even if unlikely, that the sample  
 1554 of analysed semgentations is biased in some way however, given the numbers in Table

**Table 3.14:** 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

1555 3.14, it's reasonable to fear that in the most important cases, which means those that  
 1556 correspond to 1 labels in DEATH or ICU ADMISSION , the values of the radiomic  
 1557 features used in this thesis were not the best possible.

1558 Another thing that can be noted is that in all models that included radiomics  
 1559 feature there are, among the important variables, features that quantify volumetric  
 1560 or shape information regarding the lung. A priori it's very difficult to imagine how  
 1561 lung size by itself could determine the prognosis of the patient. However it's possible  
 1562 that most of the severe cases, corresponding to the most difficult segmentation to  
 1563 perform, end up being associated to lung shapes that are significantly different  
 1564 from almost healthy patients. In this light it's possible that the model is using these  
 1565 shape information, which is naturally unrepresentative of reality, to infer the gravity  
 1566 of the situation and is finding something simply because the information is the most  
 1567 unrealistic in the most severe cases.

1568

## Chapter 4

1569

### Conclusion

1570 In this thesis varoius methods were used in an attempt to predict the prognosis of  
1571 *Sars-COVID19* patients.

1572 Regularized regression was used to predict clinical outcome using various families  
1573 of variables to compare the information hidden in each of them and to evaluate if  
1574 CT exams add any value to a small set of clinical variables.

1575 Random forest classifiers were used with the same aim. As a by-product these  
1576 two methods can be compared to see which, given the same data, can extract the  
1577 most information.

1578 Survival analysis was used, mostly by itself, to see if the data could be divided  
1579 in smaller groups with different survival functions.

1580 In the first two lines of developement it was found that, in the specific case  
1581 of data available for this thesis, radiomic features added no information to the  
1582 clinical variables. This was taken as a (grim) reminder that the quality of the  
1583 segmentations is very relevant in determining the results obtained by radiomics  
1584 analysis and that current commercial segmentation techniques based on thresholding  
1585 and region growing are not yet ready to face the problem posed by *Sars-COVID19*  
1586 pneumonia lungs. Some directions in which future works could start from this thesis  
1587 are:

- 1588 • The implementation of a *Sars-COVID19* specific segmentation method using  
1589 the vast amount of available data
- 1590 • Given the statistical difference found in the survival of patients in the first  
1591 and second waves, defined here as before and after 20/07/2020, it might be  
1592 very interesting to investigate the causes of this finding and to prospectively  
1593 continue this analysis with the data from the third wave and eventual next  
1594 ones that might occur.
- 1595 • Eventualmente se esce roba dal clustering

<sub>1596</sub> **Chapter 5**

<sub>1597</sub> **Appendix**

<sub>1598</sub> **5.1 Additional Results and complete tables relative to Random Forest**

<sub>1600</sub> Here are the tables with all of the features from random forest. There is no real utility  
<sub>1601</sub> in providing them for all possible feature combination, hence only the importances  
<sub>1602</sub> for all features will be given. This will be done for both labels used, i.e. Death and  
<sub>1603</sub> ICU Admission.

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

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

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

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

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

**Table 5.1:** 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 (GLSJM)	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

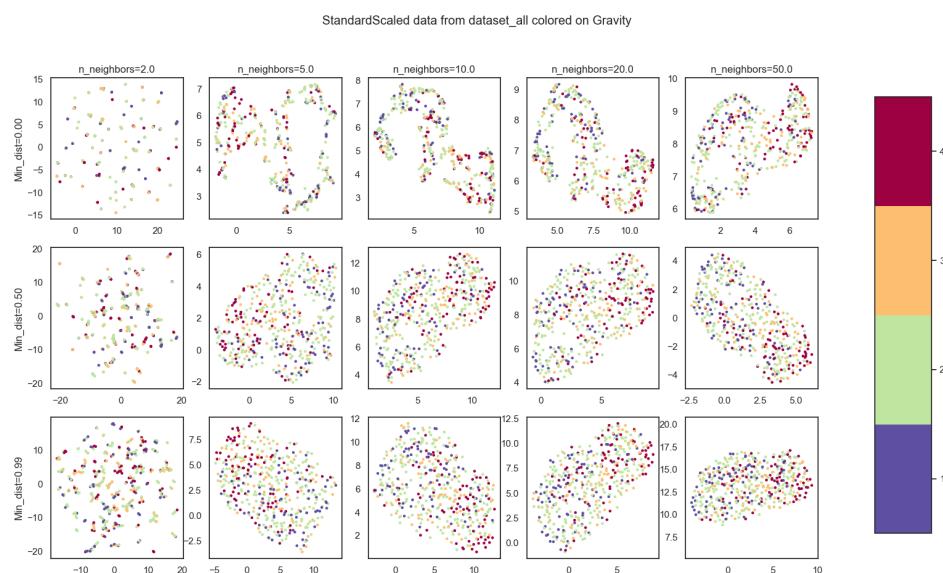
Intensity-based median absolute deviation	0.003112
90th discretised intensity percentile	0.003094
Large zone emphasis	0.003052
90th intensity percentile	0.003030
10th intensity percentile	0.003021
Low grey level run emphasis	0.002930
Kurtosis	0.002860
Cluster tendency	0.002743
Intensity at volume fraction 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 5.2:** Importances determined by RandomForest predicting ICU Admission using all available features. The values are in descending order.

## 1604 5.2 Using Dimensionality reduction to further in- 1605 vestigate the dataset

1606 For these analyses the data was always fed into a standard scaler before applying  
1607 the technique of choice, furthermore a custom gravity score by classifying as 4 the  
1608 dead individuals and then by assigning a progressive score from 1 to 3 by looking  
1609 at the time of permanence was built as follows:

- 1610 1. Gravity 1: Survived individuals with permanence from 0<sup>th</sup> percentile to 25<sup>th</sup>  
1611 percentile
- 1612 2. Gravity 2: Survived individuals with permanence from 25<sup>th</sup> percentile to 75<sup>th</sup>  
1613 percentile
- 1614 3. Gravity 3: Survived individuals with permanence from 75<sup>th</sup> percentile to 100<sup>th</sup>  
1615 percentile
- 1616 4. Gravity 4: Dead individuals without regard for permanence in the hospital

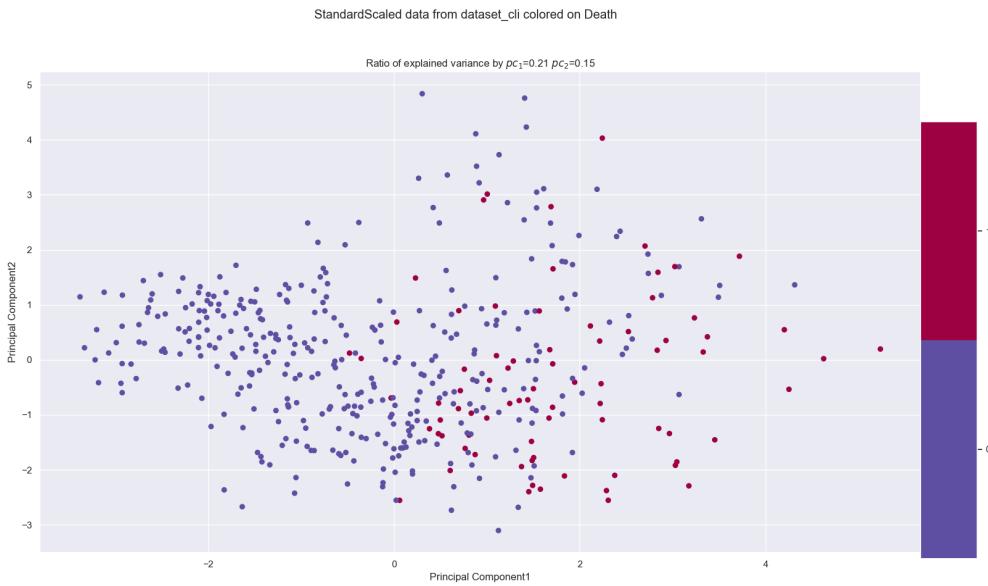


**Figure 5.1:** Possible combination for umap hyperparameters "number of neighbours" and "minimum distance". Color coding is done with aforementioned gravity score and all the features, i.e. clinical radiomic and radiological, were used.

1617 Also, before proceeding, the hyperparameter space for umap was explored since it's  
1618 the method that allows the most control over rather intuitive parameters.  
1619 Changing the value of the minimum distance of points in the final space from 0 to  
1620 0.99 changes how the structure is projected, while changing the number of  
1621 neighbours changes how much the local or global structure of the data influences  
1622 the final projection. Some of the combinations of these parameters can be seen in  
1623 Figure 5.1

## 5.2.1 Explaining total variance using PCA

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



**Figure 5.2:** 2-Principal Component on clinical features.

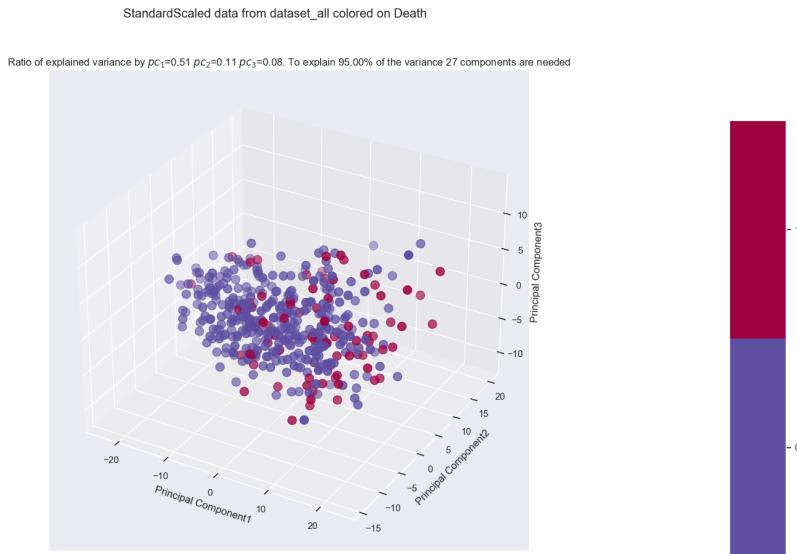
Working on the clinical dataset it can also be noted that the first two components of the PCA explain only 36% of the total variance. This leads to the conclusion that changes in the data cannot be explained by a single, nor a few, features or linear combination thereof. The next approach was using the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

In all cases it seems like introducing the radiomic features causes the loss of the polarization structure that could be seen in the PCA on the clinical dataset alone in fig5.2.

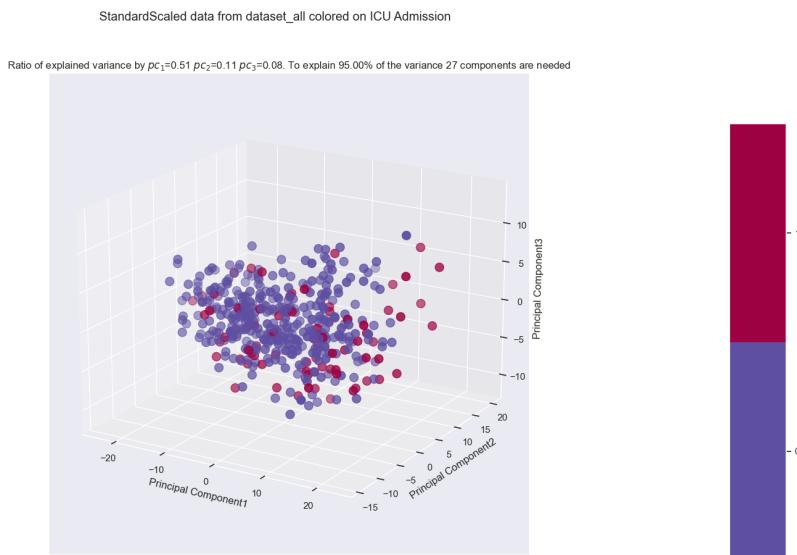
Since there are no visible clusters proceeding with cluster analysis would mean incurring in the risk of finding non meaningful results so it seemed appropriate to try other dimensionality reduction techniques.

## 5.2.2 Exploring data structure with UMAP

The next technique tried was unsupervised Umap. Following the conclusions derived from fig:5.1 the number of neighbours was set to 10 and the minimum distance was set to 0. Once again the comparison were made between clinical and radiomic dataset as well as different possible labellings. Starting from the clinical dataset, without reporting all labels used, it's clear to see that the dataset seems to indicate very local well separated structures which don't seem correlated to gravity outcome



**Figure 5.3:** 3D PCA of whole dataset, colored with death label

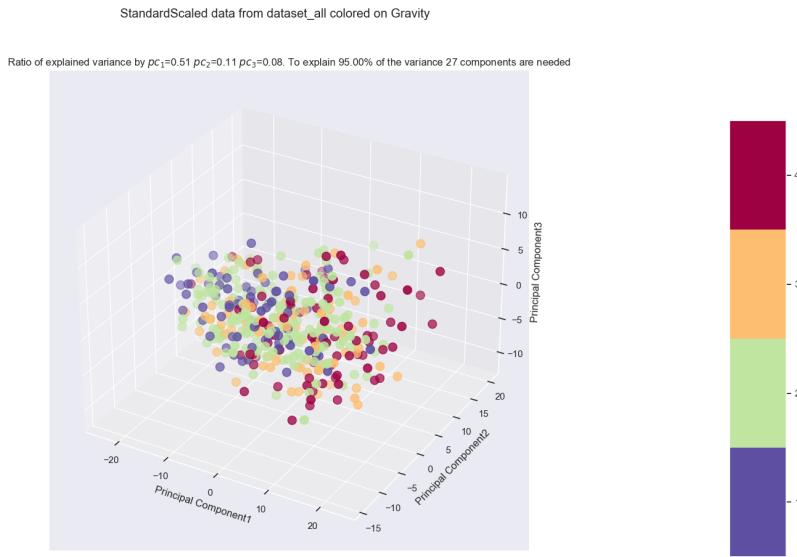


**Figure 5.4:** 3D PCA of whole dataset, colored with ICU Admission label

There are 9 well defined groups which don't seem to be correlated to any of the available labels. The dimension of these group is also very prohibitive if thinking of further analyses since groups of 35-50 people in a dataset with 15% mortality rate would mostly be very unbalanced if they were to be used for classification.

However if the introduction of radiomic features were to unite some of these groups then this embedding could be meaningfully used for analysis. Looking at the 3D embedding for the whole dataset, the results are:

Once again the introduction of the radiomic feature seems to be a confounding factor in the seemingly clear-cut order present in the clinical dataset alone. There



**Figure 5.5:** 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

1658 seems to be a well connected structure, which makes sense because umap sets out  
 1659 with the objective of preserving said structure. However since the variables of  
 1660 interest as label are Death, ICU admission or some kind of combination of them  
 1661 with hospital permanence there seems to be no visual correlation between  
 1662 structure and label. As such the next dimensionality reduction was tried to see if  
 1663 it yielded better results.

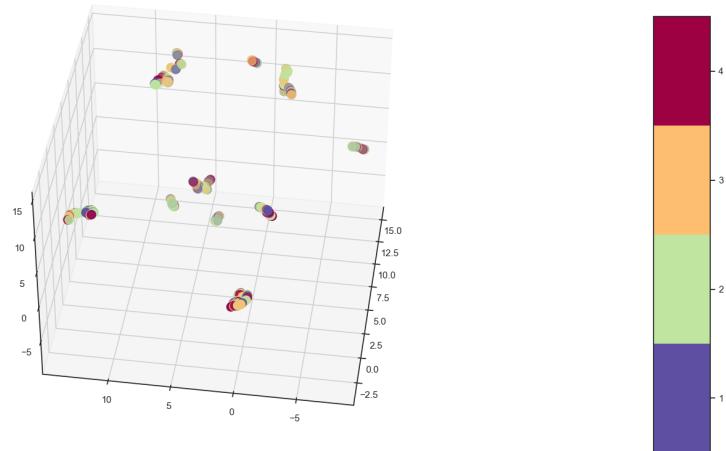
### 1664 5.2.3 Predicting clinical outcome using PLS-DA

1665 Moving on from unsupervised methods to a supervised one, PLS-DA was used  
 1666 giving as label both death and ICU using both whole dataset, and singularly  
 1667 radiomic or clinical features. Starting from the clinical features alone, predicting  
 1668 on death Figure 5.10 can be obtained.

1669 In Figure 5.10 there are a few things to note. The first is the presence, in the top  
 1670 plot, of an outlier which, since PLS-DA is based on minimization of least squares,  
 1671 can ruin a lot the performance of the procedure. For this reason in the second plot  
 1672 the outlier was removed and the algorithm was run again on the cleaned data. The  
 1673 second thing to notice is the coloring used which, in the first plot, was used to  
 1674 highlight that along the one of the two latent variables the data is roughly  
 1675 distributed depending on age while, in the second plot, was used to highlight that  
 1676 the algorithm is able to perfectly separate the subjects with hypertension from  
 1677 those without it. However, by looking at the same embedding labelled with death  
 1678 and ICU admission fig 5.11 can be obtained

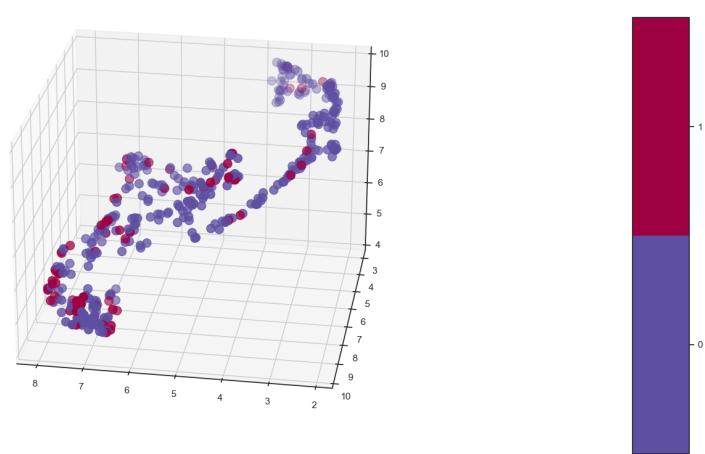
1679 It's clear to see that, when predicting on death, the PLS-DA algorithm doesn't  
 1680 find any behaviour relevant for ICU admission. It's also clear that there is at least  
 1681 a pattern of points labeled as dead being towards the right of the image, this can

StandardScaled data from dataset\_cli colored on Gravity  
Umap parameters: N\_Neighbours = 10, Min\_distance =0, Distance metric=euclidean



**Figure 5.6:** 3D umap of clinical dataset, colored based on gravity

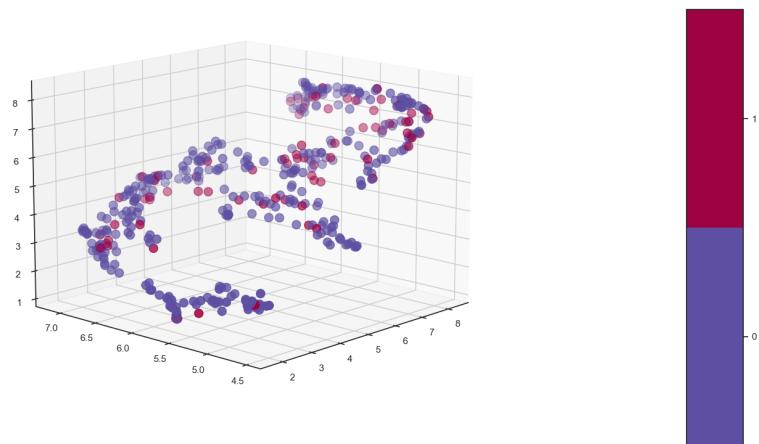
StandardScaled data from dataset\_all colored on Death  
Umap parameters: N\_Neighbours = 10, Min\_distance =0, Distance metric=euclidean



**Figure 5.7:** 3D umap of whole dataset, colored based on death

StandardScaled data from dataset\_all colored on ICU Admission

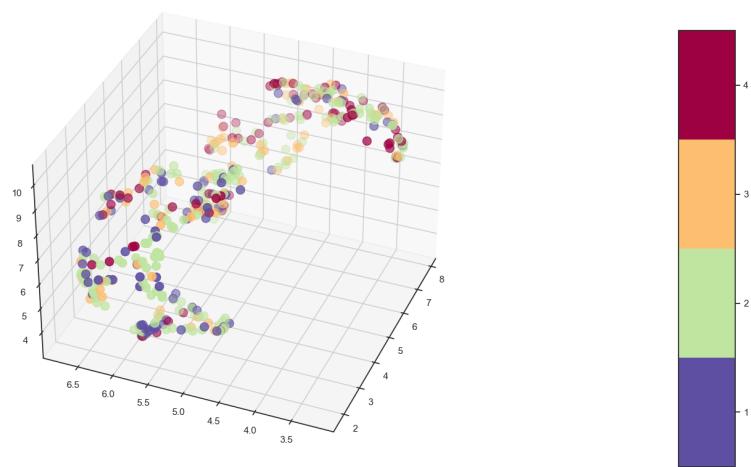
Umap parameters: N\_Neighbours = 10, Min\_distance =0, Distance metric=euclidean



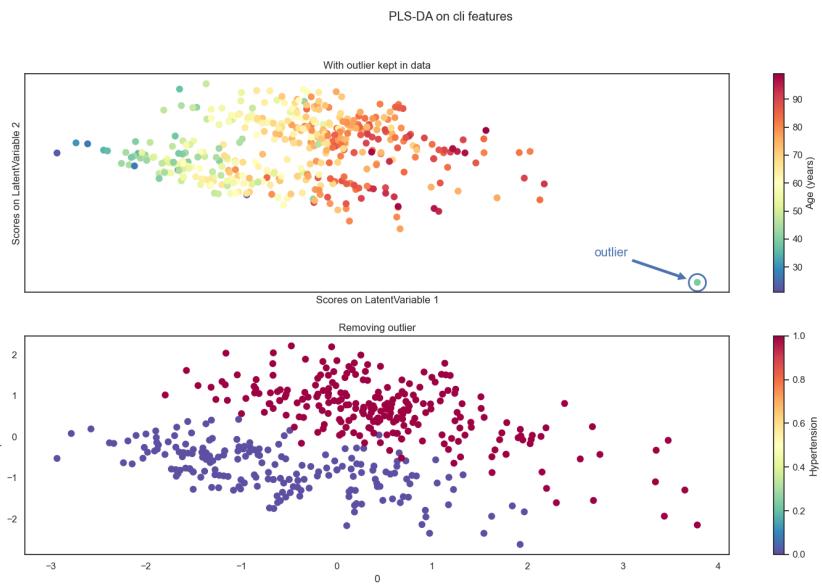
**Figure 5.8:** 3D umap of whole dataset, colored based on ICU admission

StandardScaled data from dataset\_all colored on Gravity

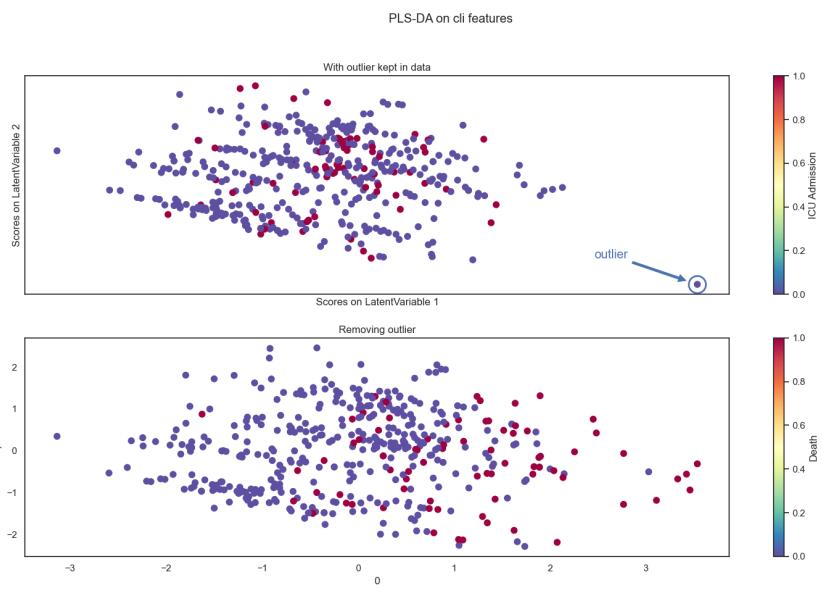
Umap parameters: N\_Neighbours = 10, Min\_distance =0, Distance metric=euclidean



**Figure 5.9:** 3D umap of whole dataset, colored based on gravity



**Figure 5.10:** PLS-DA predicting on death coloured with age and hypertension

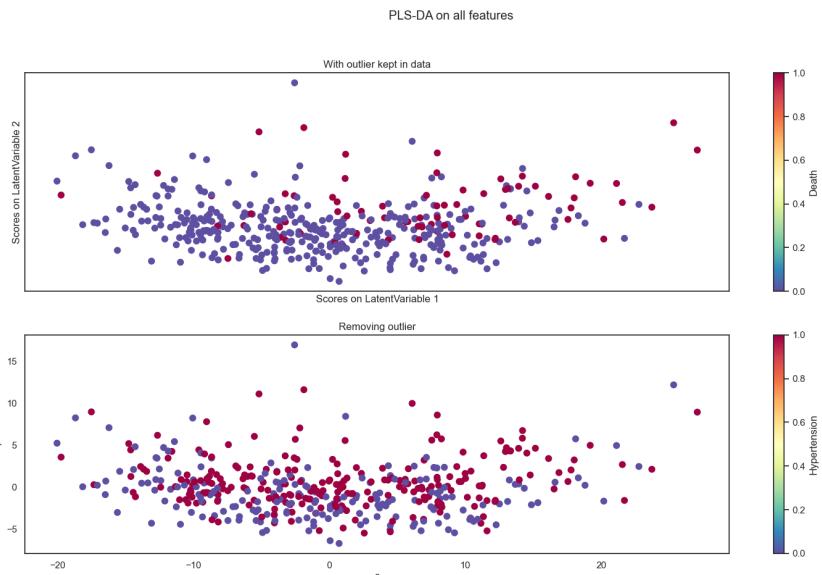


**Figure 5.11:** PLS-DA predicting on death coloured with death(bottom) and ICU Admission(top)

**Table 5.3:** 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

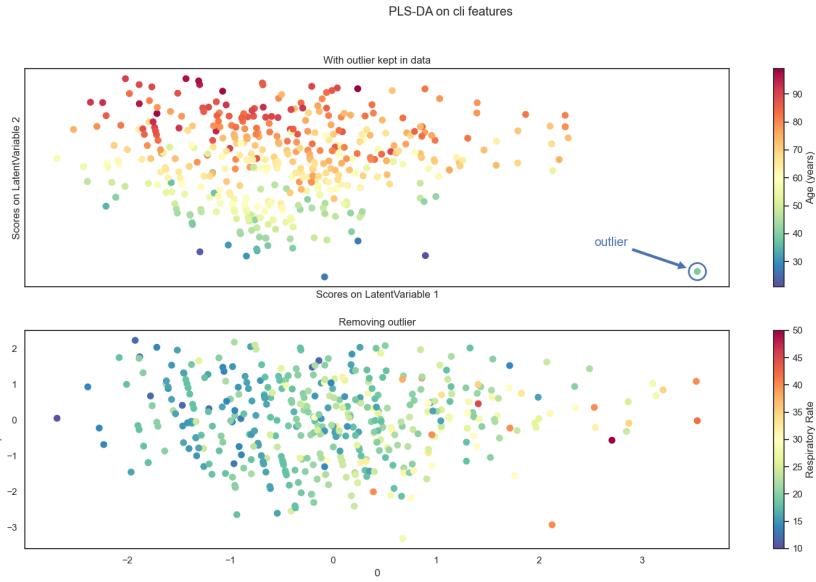
be easily explained by looking at how the ages are distributed in the first plot of fig: 5.10. From this it's possible to deduce that older individuals tend to die more and that hypertension does not seem to be relevant when considering death as a clinical outcome. If necessary the PLS-DA algorithm allows also to see the weights given to the features in predicting the label. At least for the clinical dataset, which has a reasonable number of features, it's interesting to report it ordering the coefficients by descending absolute value:  
Doing the exact same procedure on the whole dataset, which means by including the radiomic features, Figure 5.12 can be obtained.



**Figure 5.12:** PLS-DA predicting on death coloured with death (top) and hypertension (bottom) on whole dataset

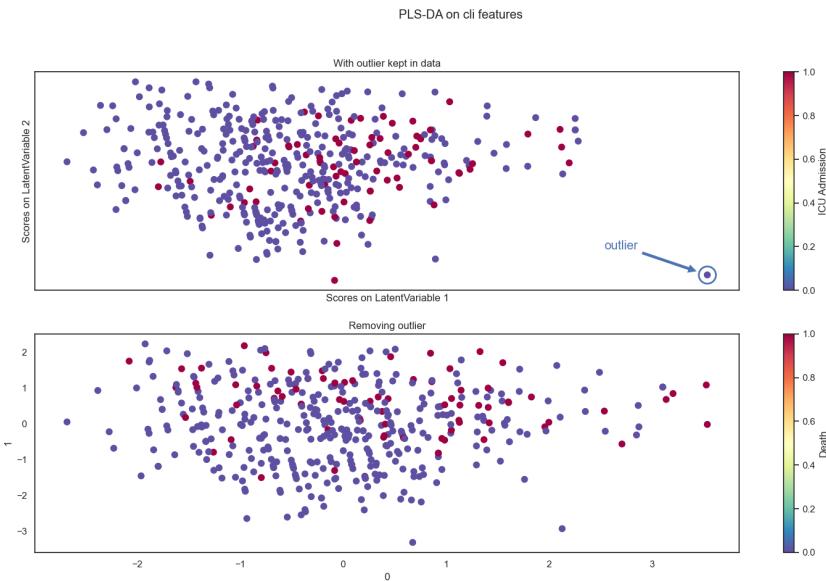
Once again adding the radiomic features has evidently introduced noise in the system, which no longer displays any kind of behaviour, pattern nor separation. Doing the same analysis but using ICU Admission as a label Figure 5.13 can be obtained.

Now the colors have been chosen to highlight that respiratory rate and age have the main role in determining the latent variables. However, in this case, there doesn't appear to be a clear cut distinction as it happened before with



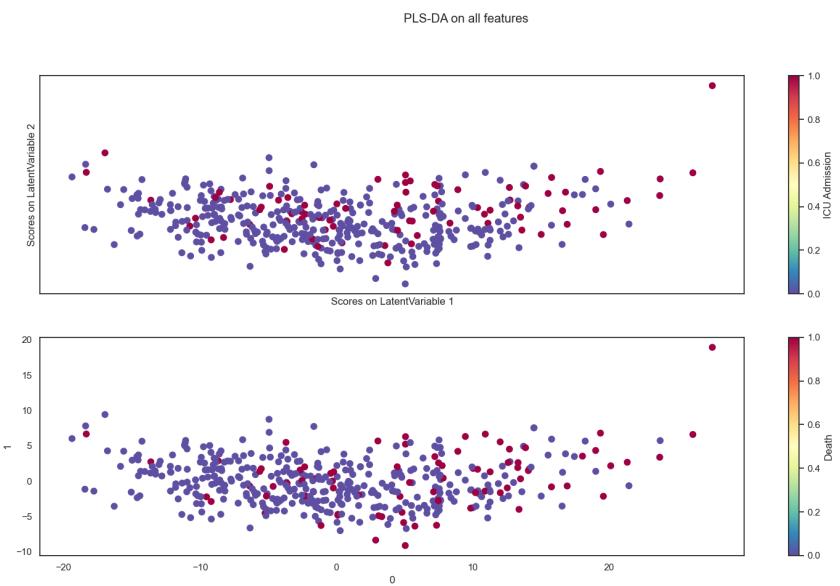
**Figure 5.13:** PLS-DA predicting on death coloured with Age and respiratory rate on clinical features

1698 hypertension. Looking at how the points scatter by coloring them according to the  
 1699 two interesting clinical labels the following figure can be obtained:



**Figure 5.14:** PLS-DA predicting on death(bottom) coloured with death and ICU admission(top) on clinical features

1700 Finally, introducing the radiomic features in the analysis the usual effect of  
 1701 reducing separation can be seen can be seen in the figure below:



**Figure 5.15:** PLS-DA predicting on death coloured with death(bottom) and ICU admission(top) on all available features

# <sup>1702</sup> Bibliography

- <sup>1703</sup> [1] Nema ps3 / iso 12052, digital imaging and communications in medicine (di-  
<sup>1704</sup> com) standard, national electrical manufacturers association, rosslyn, va, usa.  
<sup>1705</sup> available free at, 2021.
- <sup>1706</sup> [2] J. L. V. 1, R. Moreno, J. Takala, S. Willatts, A. D. Mendonça, H. Bruining,  
<sup>1707</sup> C. K. Reinhart, P. M. Suter, and L. G. Thijs. The sofa (sepsis-related organ  
<sup>1708</sup> failure assessment) score to describe organ dysfunction/failure. on behalf of the  
<sup>1709</sup> working group on sepsis-related problems of the european society of intensive  
<sup>1710</sup> care medicine. *Intensive Care Medicine*, 1996.
- <sup>1711</sup> [3] U. Attenberger and G. Langs. How does radiomics actually work? – review.  
<sup>1712</sup> *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden*  
<sup>1713</sup> *Verfahren*, 193, 12 2020.
- <sup>1714</sup> [4] M. Barker and W. Rayens. Partial least squares for discrimination, journal of  
<sup>1715</sup> chemometrics. *Journal of Chemometrics*, 17:166 – 173, 03 2003.
- <sup>1716</sup> [5] R. W. Brown, Y.-C. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan.  
<sup>1717</sup> *Magnetic Resonance Imaging: Physical Principles and Sequence Design*,  
<sup>1718</sup> 2nd Edition. Wiley-Blackwell, 2014.
- <sup>1719</sup> [6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote:  
<sup>1720</sup> Synthetic minority over-sampling technique. *Journal of Artificial Intelligence*  
<sup>1721</sup> *Research*, 16:321–357, Jun 2002.
- <sup>1722</sup> [7] T. Daimon. *Box-Cox Transformation*, pages 176–178. Springer Berlin Heidel-  
<sup>1723</sup> berg, Berlin, Heidelberg, 2011.
- <sup>1724</sup> [8] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas  
<sup>1725</sup> under two or more correlated receiver operating characteristic curves: A non-  
<sup>1726</sup> parametric approach. *Biometrics*, 44(3):837–845, 1988.
- <sup>1727</sup> [9] K. P. F.R.S. Liii. on lines and planes of closest fit to systems of points in space.  
<sup>1728</sup> *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of*  
<sup>1729</sup> *Science*, 2(11):559–572, 1901.
- <sup>1730</sup> [10] D. G. George H. Joblove. Color spaces for computer graphics. *ACM SIG-*  
<sup>1731</sup> *GRAPH Computer Graphics*, (12(3), 20–25), 1978.
- <sup>1732</sup> [11] L. Guo. Clinical features predicting mortality risk in patients with viral pneu-  
<sup>1733</sup> monia: The mulbsta score. *Frontiers in Microbiology*, 2019.

- 1734 [12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*.  
1735 Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.
- 1736 [13] G. N. Hounsfield. Computed medical imaging. nobel lecture. *Journal of Com-*  
1737 *puter Assisted Tomography*, (4(5):665-74), 1980.
- 1738 [14] L. M. J. Cine computerized tomography. *The International Journal of Cardiac*  
1739 *Imaging*, 1987.
- 1740 [15] A. Kaka and M. Slaney. *Principles of Computerized Tomographic Imaging*.  
1741 Society of Industrial and Applied Mathematics, 2001.
- 1742 [16] J. Kirby. Mosmeddata: dataset, 09/2021.
- 1743 [17] D. Kleinbaum and M. Klein. *Survival Analysis: A Self-Learning Text*. Statistics  
1744 for Biology and Health. Springer New York, 2006.
- 1745 [18] P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. de Jong,  
1746 J. van Timmeren, S. Sanduleanu, R. T. H. M. Larue, A. J. G. Even, A. Jochems,  
1747 Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt,  
1748 A. Dekker, F. M. Mottaghy, J. E. Wildberger, and S. Walsh. Radiomics: the  
1749 bridge between medical imaging and personalized medicine. *Nature reviews.*  
1750 *Clinical oncology*, 14(12):749—762, December 2017.
- 1751 [19] L. Lee and C.-Y. Lioung. Partial least squares-discriminant analysis (pls-da) for  
1752 classification of high-dimensional (hd) data: a review of contemporary practice  
1753 strategies and knowledge gaps. *The Analyst*, 143, 06 2018.
- 1754 [20] G. Lemaître, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A python  
1755 toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal*  
1756 *of Machine Learning Research*, 18(17):1–5, 2017.
- 1757 [21] J. McCarthy. What is artificial intelligence? 01 2004.
- 1758 [22] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation  
1759 and projection for dimension reduction, 2020.
- 1760 [23] S. J. McMahon. The linear quadratic model: usage, interpretation and chal-  
1761 lenges. *Physics in medicine and biology*, 2018.
- 1762 [24] S. Morozov. Mosmeddata: Chest ct scans with covid-19 related findings dataset.  
1763 *IAU Symp.*, (S227), 2020.
- 1764 [25] MosMed. Mosmeddata: dataset, 28/04/2020.
- 1765 [26] Y. Ohno. Cie fundamentals for color measurements. *International Conference*  
1766 *on Digital Printing Technologies*, 01 2000.
- 1767 [27] D. L. Pham, C. Xu, and J. L. Prince. Current methods in medical image  
1768 segmentation. *Annual Review of Biomedical Engineering*, 2(1):315–337, 2000.  
1769 PMID: 11701515.

- <sub>1770</sub> [28] P. C. Rajandep Kaur. A review of image compression techniques. *International*  
<sub>1771</sub> *Journal of Computer Applications*, 2016.
- <sub>1772</sub> [29] W. C. Roentgen. On a new kind of rays. *Science*, 1986.
- <sub>1773</sub> [30] A. Saltelli. *Global Sensitivity Analysis. The Primer*. John Wiley and Sons, Ltd,  
<sub>1774</sub> 2007.
- <sub>1775</sub> [31] C. P. Subbe. Validation of a modified early warning score in medical admissions.  
<sub>1776</sub> *QJM: An international journal of medicine*, 2001.
- <sub>1777</sub> [32] L. Tommy. Gray-level invariant haralick texture features. *PLOS ONE*, 2019.
- <sub>1778</sub> [33] S. Webb. The physics of medical imaging. 1988.
- <sub>1779</sub> [34] L. WS, van der Eerden MM, and e. a. Laing R. Defining community acquired  
<sub>1780</sub> pneumonia severity on presentation to hospital: an international derivation and  
<sub>1781</sub> validation study. *Thorax* 58(5):377-382, 2003.
- <sub>1782</sub> [35] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. W. L. Aerts, V. Andread-  
<sub>1783</sub> rczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, and  
<sub>1784</sub> et al. The image biomarker standardization initiative: Standardized quan-  
<sub>1785</sub> titative radiomics for high-throughput image-based phenotyping. *Radiology*,  
<sub>1786</sub> 295(2):328–338, May 2020.