

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 par-	
grafi aiuta anche con il layout di tabelle e figure	1
li sposterei in capitoli successivi	1
di solito si parla di color quantization	5
perché hai scelto queste specifiche tecniche? a random? Spiegato così va bene?	9
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.	15
visto che usi la regolarizzazione in modo consistente spiegherei meglio i van-	
taggi dei vari approcci uno rispetto all'altro	21
la spiegherei meglio	21
queste figure devono avere i numeri ed il testo più grandi, sono illeggibili	
altrimenti; inoltre i risultati delle divisioni vanno inseriti dopo, non qui. .	24
usare excel come immagine non è il massimo	36
non mi piace la forma, troppo personale. Corretta così va meglio?	36
non qui, magari nella discussione	38
forse usare i smallcaps per i nomi delle variabili aiuta a separarle dal resto del	
testo	38
inserirrei un flowchart della procedura, aiuta a capire	39
Ad esempio la questione della divisione in ondate	42
A seconda di come viene decisa la struttura finale questa frase va cambiata e	
aggiunta la reference al punto preciso	42
Non ricordo come mai avevamo fatto la distinzione in ondate	62
Forse già questo può andare bene come conclusione?	63
A questo punto magari cito l'appendice con la parte di clustering quando sarà	
pronta	65

Abstract

bisogna dividere i paragrafi, e consiglio una sola frase per riga. Dividere i paragrafi aiuta anche con il layout di tabelle e figure

. 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

1	Introduction	1
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	21
1.2.3	Synthetic Minority Oversampling TEchnique (SMOTE)	22
1.2.4	Dimensionality reduction and clustering	24
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	32
2	Materials and methodologies	35
2.1	Data and objective	35
2.2	Preprocessing and data analysis	39
3	Results	43
3.1	Feature selection through Lasso regularization and clinical outcomes prediction using regression	43
3.1.1	Using DEATH as label	44
3.1.2	Using ICU ADMISSION as label	50
3.2	Classification of patients using Random forests	54
3.2.1	Classifying the oucome death	54
3.2.2	Classifying the outcome ICU ADMISSION	57
3.2.3	Using survival analysis	60
3.3	Summarizing the results and further analysis	63
4	Conclusion	68
5	Appendix	69
5.1	Additional Results and complete tables relative to Random Forest	69
5.2	Using Dimensionality reduction to further investigate the dataset	78
5.2.1	Explaining total variance using PCA	79

5.2.2	Exploring data structure with UMAP	79
5.2.3	Predicting clinical outcome using PLS-DA	81
Bibliography		88

¹ Chapter 1

² Introduction

³ Nowadays everybody knows of *Sars-COVID19* which, since the start of 2020, has
⁴ made necessary a few world-wide quarantines forcing everybody in self-isolation. It
⁵ is also well known that, among the main complications and features of this virus,
⁶ symptoms gravity as well as the rate of deterioration of the conditions are some
⁷ of the most relevant and problematic. In some cases asymptomatic or near to
⁸ asymptomatic people may, in the span of a week, get to conditions that require
⁹ hospital admission. This peculiarity is also what heavily complicates the triage
¹⁰ process, since trying to predict with some degree of accuracy the prognosis of the
¹¹ patient at admission is a thoroughly complex task. In this thesis the aim will be to
¹² use data, specifically including data that cannot be easily interpreted by humans,
¹³ to try various methods to predict a couple of clinical outcomes, namely the death of
¹⁴ the patient or the admission in the Intensive Care Unit (ICU), while assessing their
¹⁵ performance. These analyses will be carried out on a dataset of 434 patients with
¹⁶ different variables associated to every person. A part of the variables, which will
¹⁷ be called clinical and radiological, are defined by humans and are generally discrete
¹⁸ in nature but mostly boolean. The most part of the available variables, however,
¹⁹ will be image-derived following the approaches used in the field of radiomics. While
²⁰ the utility of clinical variables, such as age, obesity and history of smoking, is very
²¹ straightforward it's interesting and helpful to understand the basis behind the utility
²² of radiomic and radiological features.

²³ Generally speaking it's clear that images have the ability to convey a slew of useful images, this is especially true in the medical field where digital images are used
²⁴ to inspect also the internal state of the patient giving far more detailed information than that obtainable by visual inspection at the hand of medical professionals.
²⁵ Among the ways in which *Sars-COVID19* can manifest himself the one that is most
²⁶ relevant to the scopes of this thesis pneumonia and the complications that stem
²⁷ from it. Some of these complications, which are not specific of *Sars-COVID19* but
²⁸ can happen in any pneumonia case, display very peculiar patterns when visualizing
²⁹ the lungs through CT exams.

³⁰ These patterns are due to the pulmonary response to inflammation which may
³¹ lead to thickening of the bronchial and alveolar structures up to pleural effusions
³² and collapsed lungs. Without going too much in clinical detail what is of interest is
³³ how these condition manifest themselves in the CT exams:

37 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.

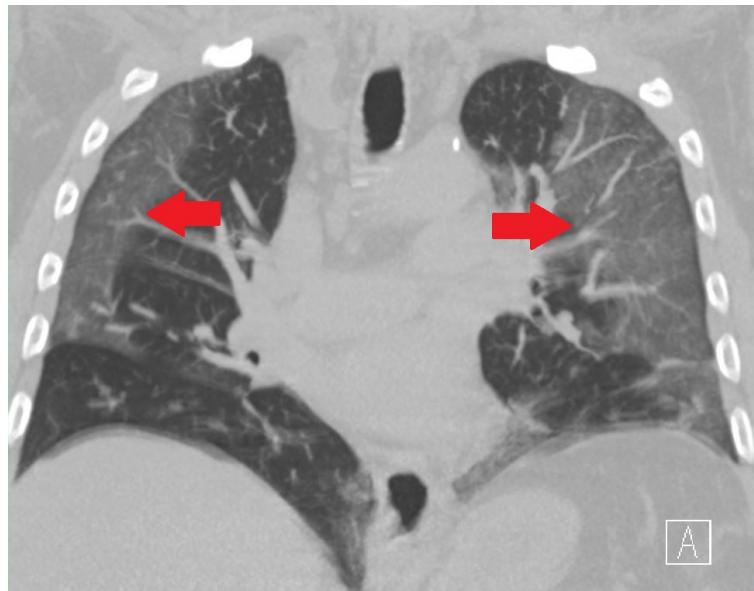


Figure 1.1: Example of GGO

40 2. **Lung Consolidations:**

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

45 3. **Crazy paving:**

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

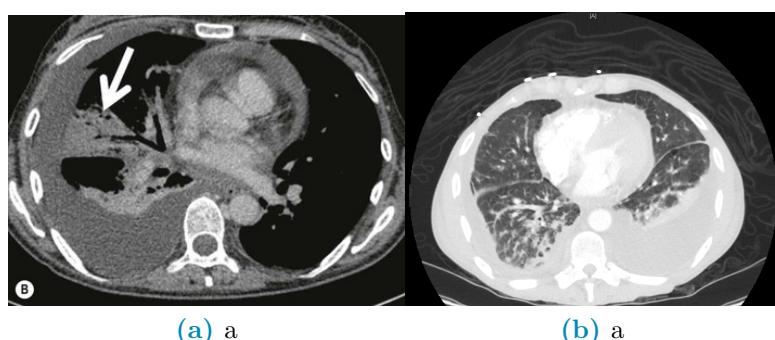


Figure 1.2: Differences between a collapsed lung (a) and pleural effusion(b)

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

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

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

56 Having these manifestation it's clear that they are mainly textural and intensity-
57 like changes in the normal appearance of the lungs. However, whereas these proper-
58 ties can be easily described in a qualitative and subjective way, it's rather complex
59 to describe them in a quantitative and objective way. The field of radiomics, when
60 coupled with digital images and preprocessing steps which must include image seg-
61 mentation, is exactly what undertakes this daunting task. Radiomics comes from
62 the combination of radiology and the suffix *-omics*, which is characteristic of
63 high-throughput methods that aim to generate a large number of numbers, called
64 biomarkers or features, as such it uses very precise and strict mathematical defini-
65 tions to quantify in various ways either shape, textural or intensity based properties
66 of the radiological image under analysis.

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

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

77 1.1 Medical Images

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

83 **Definition 1.1.1** (Digital Image). A numerical representation of an object; more
84 specifically an ordered array of values representing the amount of radiation emitted
85 (or reflected) by the object itself. The values of the array are associated to the
86 intensity of the radiation coming from the physical object; to represent the image
87 these values need to be associated to a scale and then placed on a discrete 2D grid.
88 To store these intensities the physical image is divided into regular rectangular
89 spacings, each of which is called pixel¹, to form a 2D grid; inside every spacing is

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

then stored a number (or set thereof) which measures the intensity of light, or color, coming from the physical space corresponding to that grid-spacing. The term digital refers to the discretization process that inherently happens in storage of the values, called pixel values, as well as in arranging them within the grid. It's possible to generalize from 2D images to 3D volumes, simply by stacking images of the same object obtained at different depths. In this context, the term pixel is substituted by voxel, however since they are used interchangeably in literature they will, from now on, be considered equivalent.

Generally pixel values stored as integers $p \in [0, 2^n - 1]$ with $p, n \in \mathbb{N}$ or as $p \in [0, 1]$ with $p \in \mathbb{R}$, the type of value stored within each pixel changes the nature of the image itself. A single value is to be intended as the overall intensity of light coming from the part of the object contained corresponding to the gridspace and is used for a gray-scale representation, a set of three² or four³ values can be intended as a color image.

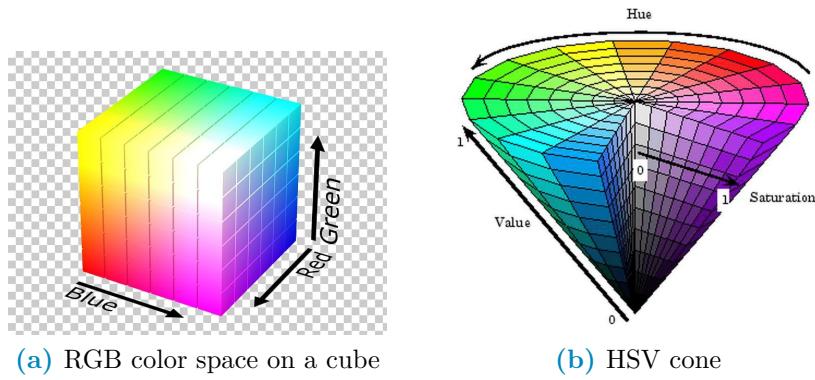


Figure 1.3: Examples of color spaces

There are a lot of possible scales for representation⁴, which are sometimes called color-spaces, however the most noteworthy in the scope of this work is the Hounsfield unit (HU) scale.

Definition 1.1.2 (Hounsfield unit (HU)). A scale used specifically to describe radiodensity, frequently used in the context of CT (Computed Tomography) exams. The values are obtained as a transformation of the linear attenuation coefficient ?? of the material being imaged and, since the scale is supposed to be used on humans, it's defined such that water has value zero and air has the most negative value -1000. For a more in depth discussion refer to [12]

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

²The three values correspond each to the intensity of a single color, the most commonly used set of colors is the RGB-scale (Red, Green, Blue). Further information can be found by looking into Tristimulus theory[25]

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

⁴Besides RGB and CMYK 1.3a the most common color spaces are CIE (Commission Internationale d'Eclairage) and HSV fig:1.3b (Hue,Saturation and Value). Refer to [9] for further details

113 The utility of this scale is in its definition, since the pixel value depends on
114 the attenuation coefficient it's possible to individuate a set of ranges that identify,
115 within good reason, the various tissues in the human body: for example lungs are
116 [-700, -600] while bone can be in the [500, 1900] range. A more in depth discussion
117 of the topics relative to Hounsfield units is going to be carried out at a later point
118 throughout this chapter, in the meantime it's necessary to clarify what are the most
119 important characteristics of an image:

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

128 di solito si parla di color quantization



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

- 129 • Size: Refers to the number of pixels per side of the image, for example in CT-
130 derived images the coronal slices are usually 512x512. These numbers depend
131 on the acquisition process and instrument but in all cases these refer to the
132 number of rows and columns in the sampling grid as well as in the matrix
133 representing the image.
- 134 • Data-Format: How the pixel values are stored in the file of the image. The
135 most commonly used formats are .PNG and .JPG however there are a lot of
136 other formats. In the context of this work, which is going to be centered on
137 medical images, the most interesting formats are going to be the nii.gz (Nifti)
138 and the .dcm (DICOM). The first contains only the pixel value information
139 hence it's a lighter format, it originates in the field of Neuroimaging⁵, it is used
140 mainly in Magnetic resonance images of the brain but also for CT scans and,
141 since it contains only numeric information, it's the less memory consuming

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

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 field⁶.

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 context⁷, 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.

⁶A detailed description of compression algorithms is beyond the scopes of this thesis, for this reason please refer to [27] for more information

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

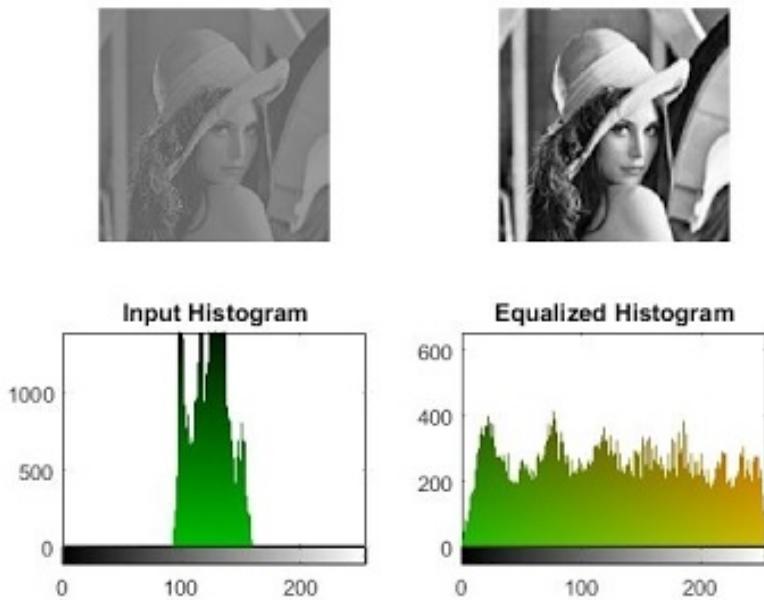


Figure 1.5: Example of differences in contrast due to histogram equalization

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

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

183 1. Geometric Transformations involve the following steps:

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

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

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

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

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

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

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

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

222 The aforementioned is surely not a comprehensive list of all that can be said on
223 images however it should be enough for the scopes of this work.

224 Having seen what constitutes an image and what can be done with one it
225 becomes interesting to explore how images are obtained. The following discussion
226 is going to introduce briefly some of the methods used to obtain medical images,
227 getting more in depth only on the modality used to obtain all the images used in this
228 thesis which is Computed Tomography. This technique was used because it's the
229 only one that can provide information on the internal structure of an organ which
230 is very low in density and that has parts that are deep in the patient's body. Since

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

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

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

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

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

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

²⁷¹ 1.1.1 X-ray imaging and Computed Tomography ²⁷² (CT)

²⁷³ It's well known that the term x-rays is used to characterize a family electromagnetic
²⁷⁴ radiation defined by their high energy and penetrative properties. Radiation of this
²⁷⁵ kind is created in various processes such as characteristic emission of atoms, also
²⁷⁶ referred to as x-ray fluorescence, and Bremsstrahlung, braking radiation¹¹. The
²⁷⁷ discovery that "A new kind of ray"^[28] with such properties existed was carried out
²⁷⁸ by W.C.Roentgen in 1895, which allowed him to win the first Nobel prize in physics
²⁷⁹ in the same year. Clearly the first imaging techniques that involved this radiation
²⁸⁰ were much simpler than their modern counterpart, first of all they were planar and
²⁸¹ analog in nature, as well as not as refined in image quality. The first CT image
²⁸² was obtained in 1968 in Atkinson Morley's Hospital in Wimbledon. Tomography
²⁸³ indicates a set of techniques¹² that originate as an advancement of planar x-ray
²⁸⁴ imaging; these techniques share most of the physical principles with planar imaging
²⁸⁵ while overcoming some of it's major limitations, main of which being the lack of
²⁸⁶ depth information. X-ray imaging, both planar and tomographic, involves seeing
²⁸⁷ how a beam of photons changes after traversing a target, the process amounts to a
²⁸⁸ kind of average of all the effects occurred over the whole depth travelled.

²⁸⁹ The way in which slices are obtained is called focal plane tomography and, as the
²⁹⁰ name suggests, the basic idea is to focus in the image only the desired depth leaving
²⁹¹ the unwanted regions out of focus. This selective focusing can be obtained either
²⁹² by taking geometrical precautions while using analog detectors, such as screen-film
²⁹³ cassettes, or by feeding the digital images to reconstruction algorithms to perform
²⁹⁴ digitally the required operations¹³.

²⁹⁵ In both planar and tomographic setting the rough description of the data acqui-
²⁹⁶ sition process can be summarized as follows: First x-rays are somehow generated
²⁹⁷ by the machine, the quality of these x-rays is optimized with the use of filters then
²⁹⁸ focused and positioned such that they mostly hit the region that needs imaging.
²⁹⁹ The beam then exits the machine and starts interacting with the imaged object¹⁴,
³⁰⁰ this process causes an attenuation in the beam which depends on the materials com-
³⁰¹ posing the object itself. Having then travelled across the whole object it interacts
³⁰² with a sensor, be it film, semiconductor or other, which stores the data that will
³⁰³ then constitute the final image. In a digital setting this final step has to be per-
³⁰⁴ formed following a (tomographic) reconstruction algorithm which given a set of 2D
³⁰⁵ projections returns a single 3D image.

³⁰⁶ In this light the interesting processes are how the radiation is created and shaped
³⁰⁷ before hitting the patient and how said radiation then interacts with the matter of
³⁰⁸ both the patient's body and the sensor beyond it. To explore these topics it's
³⁰⁹ necessary to see:

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

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

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

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

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

312 For more information on reconstruction algorithms refer to [14] and [32].

313 1.1.2 Generation and management of radiation: 314 digital CT scanners

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

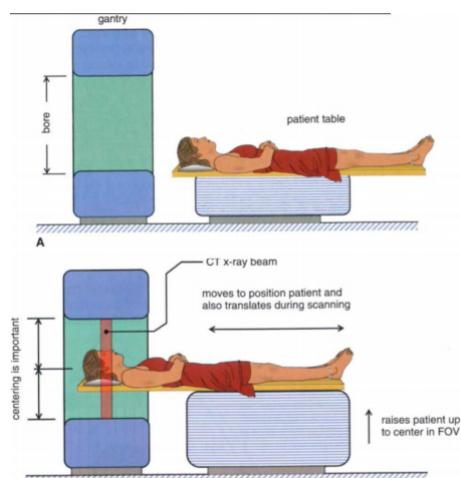


Figure 1.6: General set-up of a CT machine

320 The beam is generally created by the interaction of high energy particles with
 321 some kind of material, so that the particle's kinetic energy can be converted into
 322 radiation. In practice this means that an x-ray tube is encapsulated in the ma-
 323 chine. Inside this vacuum tube charged particles¹⁵ are emitted from the cathode,
 324 accelerated by a voltage differential and shot onto a solid anode¹⁶. This creation
 325 process implies that the spectrum of the produced x-rays is composed of the almost
 326 discrete peaks of characteristic emission, due to the atoms composing the target,
 327 superimposed with the continuum Bremsstrahlung radiation.

¹⁵Most commonly electrons

¹⁶Typical materials can be Tungsten, Molybdenum

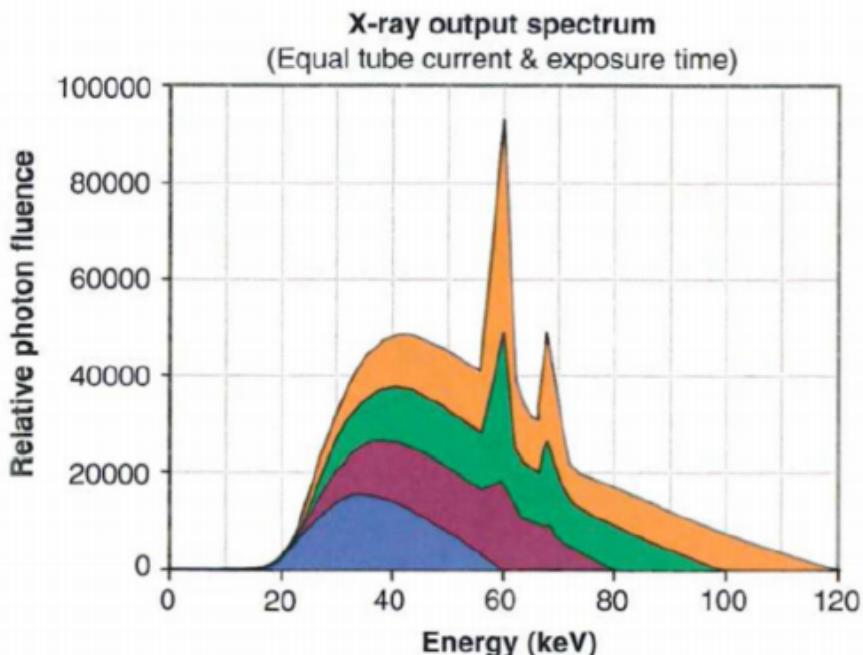


Figure 1.7: X-ray spectrum, composed of characteristic peaks and Bremsstrahlung continuum, computed at various tube voltages

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

334 Other relevant characteristics in the tube are the anode material, which changes
 335 the peaks in the x-ray spectrum and time duration of the emission, which is called
 336 exposure time and influences dose as well as exposure¹⁷.

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

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

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

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

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

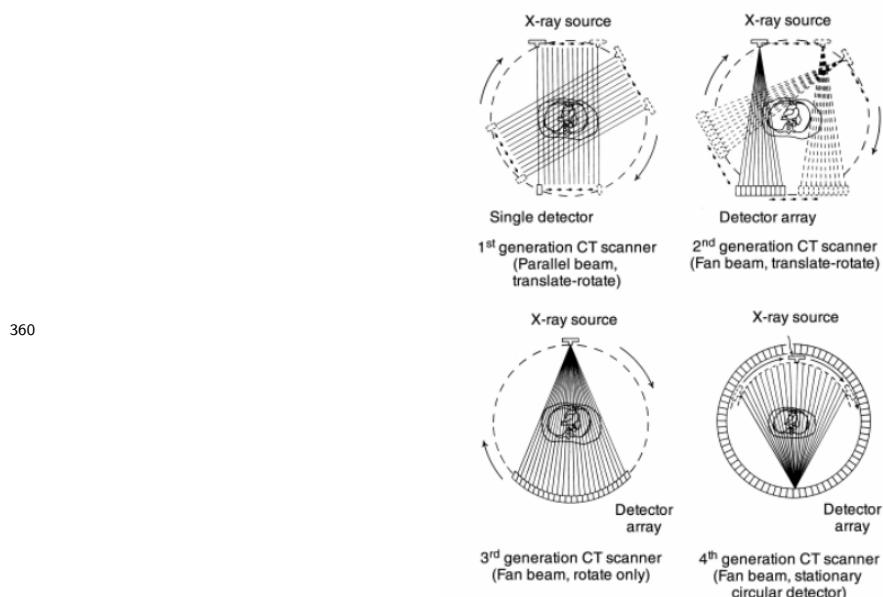


Figure 1.8: First four generation of CT scanners

- 361 • 3rd generation-Rotate Rotate Geometry: Enlarging the span of the fan of
 362 beams and using a curved array of sensor a single emission of the N beams
 363 engulfs the whole body so the only motion necessary is rotation of the couple
 364 beam-sensor array around the patient.
- 365 • 4th generation-Rotate Stationary Geometry: The sensors are now built to com-
 366 pletely be around the patient so that only the beam generator has to rotate
 367 around the body
- 368
- 369 • 5th generation-Stationary Stationary Geometry: The x-ray tube is now a large
 370 circle that is completely around the patient. This is only used in cardiac
 371 tomography and as such will not be described further [13]
- 372 • 6th generation-Spiral CT: Supposing the patient is laying parallel to the axis
 373 of rotation, all previous generations acquired, along the height of the patient,
 374 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 detector¹⁸. 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.

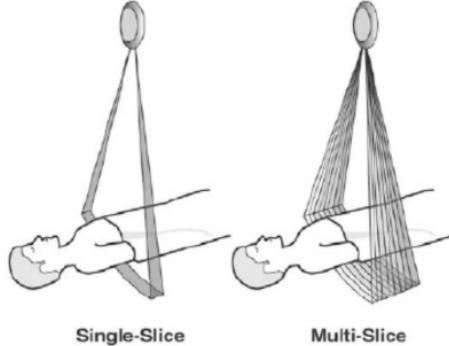


Figure 1.9: 7th generation setup

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

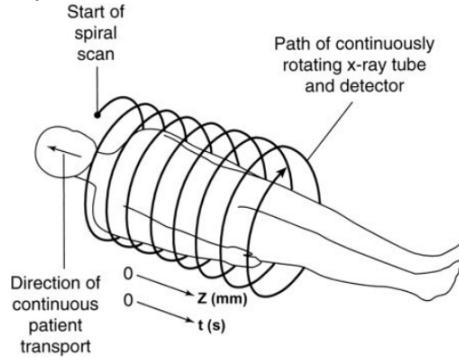


Figure 1.10: 7th generation setup

The machines used to obtain the images used in this thesis, all belonging to the Ospedale S. Orsola, were distributed as follows shown in 1.11:

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

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

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

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 μ 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 μ :

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

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 σ as:

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

414 Where ρ 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

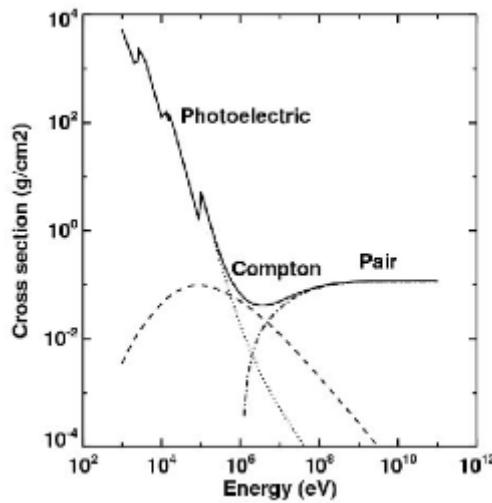


Figure 1.12: Photon cross-section in Pb

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

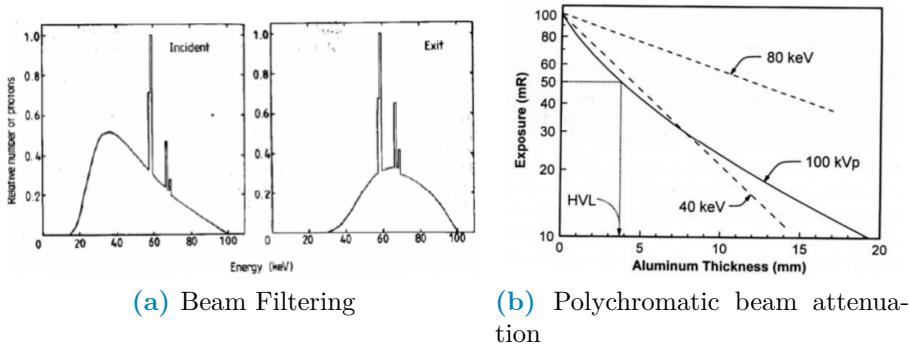


Figure 1.13: Polychromatic beam behaviour

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

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

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 beam and introducing an element of noise. Anti-scattering grids are positioned right before the sensors to reduce this effect by allowing to reach the sensor to only the photons with the correct direction.

The 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 [22].

452

453 1.2 Artificial Intelligence (AI) and Machine Learning(ML)

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

457 techniques will be used to analyze it.

458 Starting from the definition given by John McCarthy in [20] "[AI] is the science
459 and engineering of making intelligent machines, especially intelligent computer pro-
460 grams. It is related to the similar task of using computers to understand human
461 intelligence, but AI does not have to confine itself to methods that are biologically
462 observable.". Machine Learning (ML) is a sub-branch of AI and contains all tech-
463 niques that make the computer improve performances via experience in the form
464 of exposure to data, practically speaking this finds it's application in classification
465 problems, image/speech/pattern recognition, clustering, autoencoding and others.
466 The general workflow of Machine Learning is the following: given a dataset, the
467 objective is to define a model or function which depends on some parameters which
468 is able to manipulate the data in order to obtain as output something that can be
469 evaluated via a predefined performance metric. The parameters of the model are
470 then automatically adjusted in steps to minimize or maximize this performance met-
471 ric until a stable point at which the model with the current parameters is considered
472 finalized; one of the main problems in this procedure is being sure that the stable
473 point found is global and not local. The whole procedure is carried out keeping in
474 mind that the resulting model needs to be able to generalize it's performance on
475 data that it has never seen before, for this reason usually ML is divided in a train-
476 ing phase and a testing phase. The training phase involves looking at the data and
477 improving the performance of the model on a specific dataset¹⁹, the testing phase
478 involves using brand new data to evaluate the performance of the model obtained
479 in the preceding phase. Machine Learning techniques can be further grouped into
480 the following categories:

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

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

499 (state_t, action, reward, state_{t+1}) these techniques try to maximize the cumulative
500 reward²⁰. The main applications of these techniques are in Autonomous
501 driving and learning how to play games

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

503 1.2.1 Regression, Classification and Penalization

504 Regression and Classification are methods used to make predictions via supervised
505 learning by understanding the input-output relation in continuous and discrete cases
506 respectively. The most basic example of regression is linear regression which consists
507 in finding the slope and intercept of a line passing through a set of points. A
508 branch of classification that's similar to linear regression is logistic regression, this
509 translates in finding out whether or not each point belongs to a certain category
510 given its properties and can be practically thought of, for a binary classification, as
511 a fitting procedure such that the output can be either 0 for one category and 1 for
512 the other, this is usually done using the logistic function, also called sigmoid, which
513 compresses \mathbb{R} in $[0, L]$ as seen in 1.14. L represents the maximum value desired,
514 x_0 is the midpoint and k is the steepness. Note that for multiple categories it would
515 conceptually suffice to add sigmoids with different centers to obtain a step function
516 in which each step corresponds to a category, in this case the procedure is usually
517 called multinomial regression.

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

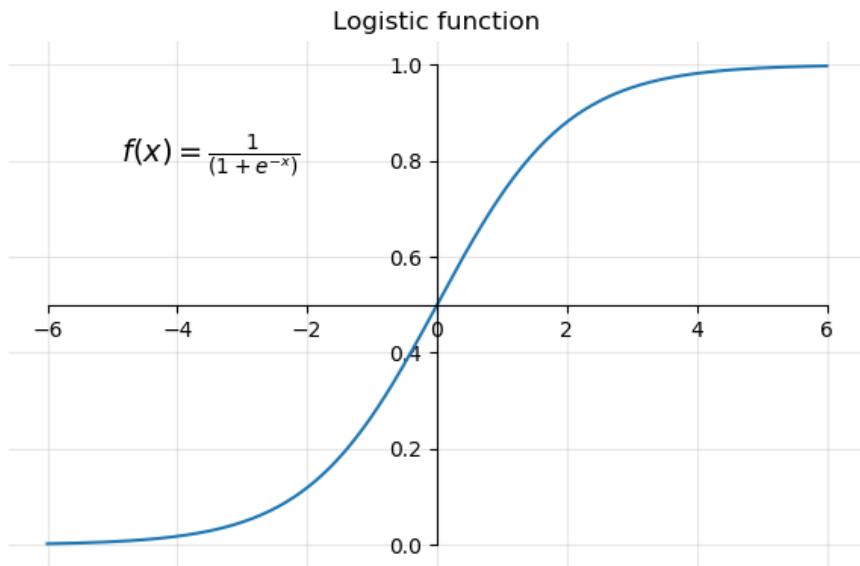


Figure 1.14: sigmoid function with L=1, $x_0=0$, $k=1$

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

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

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 combinaion 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 θ be a vector of
 529 $n+1$ entries, one for each feature plus the intercept θ_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 ϵ 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 θ^* 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 α .

²¹n+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

visto che usi la regolarizzazione in modo consistente spiegherei meglio i vantaggi dei vari approcci uno rispetto all'altro

Following regularization procedures, specifically Lasso regularization, as a byproduct of avoiding divergences a selection and ranking of the important features is obtained however it's still important to preprocess the data to simplify the job of the regularization.

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

1. Standard Scaling: This amounts to subtracting the mean of the distribution to the feature and dividing by the standard deviation so that the resulting distribution is somewhat centered around zero and has close to unitary standard deviation
2. Boxcox transform: When distributions are heavy-tailed, like a gamma distribution would be, this finds the best exponent to transform via a power-law the data. Since the exponent $\lambda \in \mathbb{R}$ the input data needs to be strictly positive and not constant.

la spiegherei meglio

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

1.2.2 Decision Trees and Random Forest

Apart from logistic and multinomial regression there are various other supervised classifying methods such as Support Vector Machines (SVM), Neural Networks and Decision Trees. In this thesis, among the aforementioned algorithms, the chosen one was a particular evolution of DecisionTrees called RandomForest (RF). To provide some insight in the method it's necessary to first explain how decision trees work, specifying what problems they face and what are their strong points. Since it's a classification method let's consider the simple case of binary classification with categorical²² features. The task of the decision tree is to approximate to the best of its abilities the labels contained in the training set using all the features associated with each datapoint, this is done by building a graph-like structure in which the nodes represent the features and the links departing from it are the possible values the feature takes. This graph is built in a top-down approach by choosing at every step the feature that best separates the data in the label categories, this process is done along each branch until a node in which the separation of the preceding feature is better than that provided by all remaining features or all features have

²²Categorical is to be intended as features with discrete value, opposed to continuous variables which can potentially take any value in \mathbb{R}

592 been considered. The first node is called root node, while the nodes that have no
 593 branches going out of them are called leaves, the graph represents a tree hence the
 594 name of the method. Note that at every node only the subset of data corresponding
 595 to all previous feature categories is used. At each node the separation between
 596 the two label categories c_1, c_2 due to the feature is commonly measured with Gini
 597 impurity coefficient, which is computed as:

$$\begin{aligned} G &= 1 - p_1^2 - p_2^2 \\ &= 1 - \left[\frac{N_{c_1 \in \text{node}}}{N_{\text{samples} \in \text{node}}} \right]^2 - \left[\frac{N_{c_2 \in \text{node}}}{N_{\text{samples} \in \text{node}}} \right]^2 \end{aligned} \quad (1.11)$$

598 The Gini impurity for a feature is then computed as an average of the gini
 599 coefficients of all the deriving nodes weighted by the number of samples in each of
 600 the nodes. This method can be obviously generalized to cases in which features are
 601 continuous by thresholding the features choosing the value that best improves the
 602 separation of the deriving node, a way to choose possible thresholds is to take all
 603 the means computed with all adjacent measurements. It's important to note firstly
 604 that there is no restriction on using, along different branches, different thresholds for
 605 the same feature and, secondly, that the same feature can end up at different depths
 606 along different branches. The strength of this method is its performance on data
 607 it has seen however it has very poor generalization abilities [11].

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

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

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

625 1.2.3 Synthetic Minority Oversampling TEchnique 626 (SMOTE)

627 This technique considers a user-defined number of nearest neighbours of randomly
 628 chosen points in the minority class and populates the feature space by generating

629 samples on the lines that connect the chosen sample with a random neighbour²³.

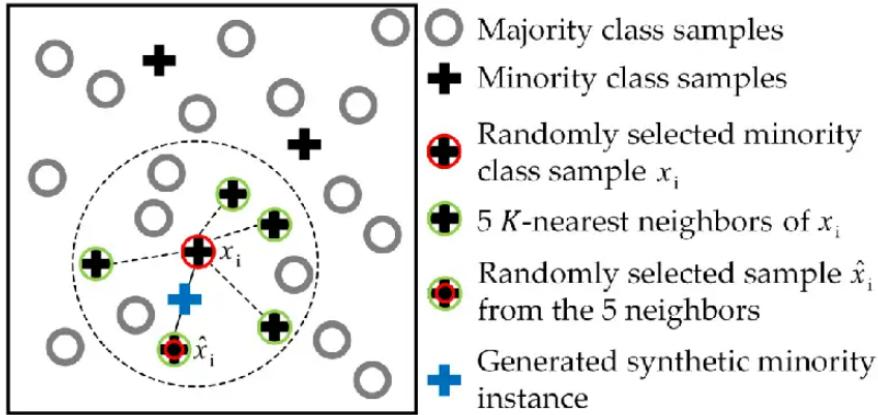


Figure 1.15: Example of SMOTE with 5 nearest neighbours

630 Worth noting that this has been done using the library imblearn in python [19].
631 To preview some of the data, looking at the performance of a vanilla random forest
632 implementation with all preset parameters, the effect of the position of oversampling
633 is the following.

²³This 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

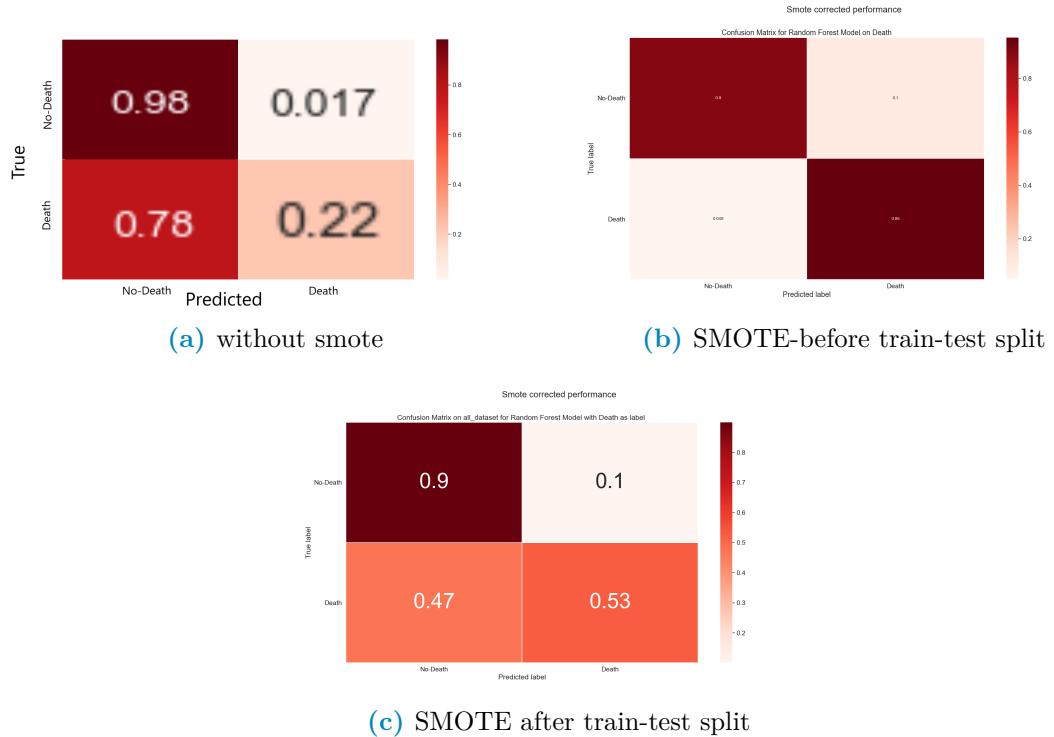


Figure 1.16: Confusion matrix used to evaluate performance done using datasets with various combinations of SMOTE position

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 improvement in performance. However performing it in the wrong place it clearly makes a far too optimistic evaluation of the performance. This highlights the point that all preprocessing should be done with care when train-test splitting the data, specifically it's very important that the oversampling, as well as all other data handling, be performed after the train-test split of the data on the train data alone. It's clear to see that what's being observed is a leakage phenomenon, in which the testing data contains information regarding the training data and viceversa. In practice, especially because the points are created as convex combination of existing data²⁴, when the synthetically generated points end up in the testing they depend, at least partially, on the data used in the training procedure.

1.2.4 Dimensionality reduction and clustering

When dataset are composed of many features, namely more than three, it becomes difficult if not impossible to visualize the distribution of data. There are various techniques that can mitigate or solve this problem, they are grouped under the umbrella term of "Dimensionality reduction techniques" and they are generally based

²⁴Note that this would happen with any other over/under-sampling methods, such as random oversampling which randomly duplicates data points

651 on ML learning methods. As such , following the same reasoning and definitions
652 given in regard to ML, it's possible to introduce a sub-categorization of the whole
653 family of techniques in supervised and unsupervised techniques. Dimensionality
654 reduction, beyond providing useful insight in the general structure of the data, can
655 also be used as a preprocessing step in some analysis pipelines. A first example
656 could be to find more meaningful features before analyzing with methods such as
657 Random Forests or Regression techniques, a second example could be using the
658 reduced representation that keeps the most information possible to ease in clustering
659 analysis by highlighting the differences in subgroups within the dataset. The most
660 common dimensionality reduction techniques are:

661 **1. Unsupervised methods**

- 662 • Principal Component Analysis (PCA): Linearly combines the pre-existing
663 features to obtain new ones and orders them by decreasing ability to ex-
664 plain total variance of data. The first principal component is the combi-
665 nation of features that most explains the variance in the original dataset.
666 Given it's nature it focuses on the global characteristics of the dataset.
- 667 • t-distributed Stochastic Neighbor Embedding (t-SNE): Keeps a mixture
668 of local and global information by using a distance metric in the full high
669 dimensional space and trying to reproduce the distances measured in the
670 lower dimensional space which can be 2- or 3-dimensional. NON SO
671 SE VA SPIEGATO MEGLIO O PIU' IN DETTAGLIO
672 Let's define similarity between two points as the value taken by a normal
673 distribution in a point away from the centre, corresponding to the first
674 point of interest, the same distance as the two points taken in consider-
675 ation. Fixing the first point the similarities with all other points can be
676 computed and normalized. Doing this procedure for all points it's pois-
677 sible to build a normalized similarity matrix. Then the whole dataset
678 is projected in the desired space, usually \mathbb{R}^i with $i=1,2$ or 3, in which a
679 second similarity matrix is built using a t-distribution instead of a normal
680 distribution ²⁵. At this point, in iterative manner, the points are moved
681 in small steps in directions such that the second matrix becomes more
682 similar to the one computed in the full space.

683 **2. Supervised methods**

- 684 • Partial Least Squared Discriminant Analysis (PLS-DA): This technique is
685 the classification version of Partial Least Squares (PLS) regression. Much
686 like PCA the idea is to find a set of orthonormal vectors as linear combina-
687 tion of the original features in the dataset, however in PLS and PLS-DA
688 is necessary to add the constraint that the new component, besides be-
689 ing perpendicular to all previous ones, explains the most variability in a

²⁵The 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 [?]

given target variable, or set thereof. For an in depth description refer to [4] while for a more modern review refer to [18] Hanno senso queste due in bibliografia?

3. Mixed techniques

- Uniform Manifold Approximation and Projection (UMAP): Builds a network using a variable distance definition on the manifold on which the data is distributed then uses cross-entropy as a metric to reproduce a network with the same structure in the space with lower dimension. This technique maintains very local information on the data and allows a complete freedom in the choice of the final embedding space as well as the definition of distance metrics in the feature space²⁶, it's also implemented to work an a generic pandas dataframe in python so it can take in input a vast range of datatypes. The math behind this method is much beyond the scopes of this thesis, as such refer to [21] for more in depth information.

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

1. Centroid based techniques: A user defined number of points is randomly located in the data space, datapoints are then assigned to groups according to their distance from the closest center. The main technique in this category is k-means clustering, and some, if not most, other techniques include it as step in the processing pipeline²⁷. The main problems are firstly that these techniques require prior knowledge, or at the very least a good intuition, on the number of clusters in the dataset while also assuming that the clusters are distributed in spherical gaussian distribution, which is not always the case.
2. Hierarchical clustering: The idea is to find the hierarchical structure in the data using a bottom-up or a top-down approach, as such this category further subdivides in agglomerative and divisive methods. In the first each point starts by itself and then points are agglomerated using a similarity or distance metric, this build a dendrogram in which the k^{th} level roughly corresponds to a k-centroid clustering. In the latter the idea is to start with a unique category and then divide it in subgroups.
3. Density based techniques: These methods use data density to define the groups by looking for regions with larger and lower density as clusters and separations.

²⁶Actually to keep the speed in performance the distance function needs to be Numba-jet compilable

²⁷An example of such techniques is: affinity propagation

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

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

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

746 1.3 Combining radiological images with AI: Im- 747 age segmentation and Radiomics

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

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

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

768 The next step would be to quantify the characteristics of the region identified, as
769 such it should be clear that finding ways to derive objective information from images
770 it's of paramount importance, especially when this information can aid in describing
771 the health of a patient. It should also be clear that medical images are a kind of
772 high dimensional data as such, as it's fashion with fields that occupy themselves with

773 big biological data, the field that studies driving quantitative information out of
774 radiological images is called radiomics.

775 1.3.1 Image Segmentation

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

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

792 On a practical level these techniques can be used in various fields, as illustrated
793 by the variety of aforementioned tasks, but the one that interests this work the
794 most is the medical field. Nowadays in medicine, where most of imaging exams are
795 stored in digital form, the ability to automatically discern specific structures within
796 the images can provide a way to aid clinical professionals in their everyday decision
797 making their workload lighter and helping in otherwise difficult cases. A staggering
798 example connected to this thesis is the process of organ segmentation in CT scans:
799 usually these scans are n²⁸ stacked images in 512x512 resolution. To have a contour
800 of the lungs in a chest CT a radiologist would need to draw by hand the contour
801 of the lung in each slice of the scan. Even if some shorcuts exist to reduce the
802 number of slices to draw on this very boring, time consuming and repetitive task
803 can occupy hours if not days of work to a human while a machine can take minutes to
804 complete a whole scan. Then considering lesion detection in medical exams having
805 a machine that consistently finds lesions that would otherwise be difficult to discern
806 by a human eye can be of paramount importance in diagnosis as well as treatment.
807 In the medical field the difficulty comes from the fact that different exams have
808 different types of image formats which means that an algorithm that works well
809 on CT may not work as intended on MRI or other procedures. Automatic image
810 segmentation can be performed in various ways:

²⁸n 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

- 811 1. Artificial Neural Network (ANN): These techniques belong to a sub-field in
812 ML called Deep Learning, they involve building network structures with more
813 layers in which each node is a processing unit that takes a combination of
814 the input data and gives an output according to a certain activation function.
815 These structures are called Neural Networks because they resemble and are
816 modeled after the workings of neuron-dendrite structures while the deep in
817 deep learning refers to the fact that various layers composed of various neurons
818 are stacked one after the other to complete the structure. Learning is obtained
819 by changing how each neuron combines the inputs it receives. In the case of
820 images these structures are called Convolutional Neural Networks because the
821 first layers, which are intended to extract the latent features or structures in
822 the images, perform convolution operation in which the parameters are the
823 values pixel values of convolutional kernels
- 824 2. Thresholding: These techniques involve using the histogram of the image
825 to identify two or more groups of pixel values that correspond to specific
826 parts/objects within the image. An obvious case would be a bi-modal distri-
827 bution in which the two sets can be clearly identified but there are no require-
828 ments on the histogram shape. In the case of CT this has a very simple and
829 clear interpretation since HU depend on tissue type it's reasonable to expect
830 that some tissues can be differentiated with good approximation by pixel value
831 alone
- 832 3. Deformable models and Region Growing: Both these techniques involve setting
833 a starting seed within the image, in the first case the seed is a closed surface
834 which is deformed by forces bound to the region of interest, such as the desired
835 edge. In the second case the seed is a single point within the region of interest,
836 step by step more points are added to a set which started as the seed alone
837 according to a similarity rule or a predefined criteria.
- 838 4. Atlas-guided: By collecting and summarizing the properties of the object that
839 needs to be segmented it's possible to compare the image at hand with these
840 properties to identify the object within the image itself.
- 841 5. Classifiers: These are supervised methods that focus on classifying via by
842 focusing on a feature space of the image. A feature space can be obtained
843 by applying a function to the image, an example of feature space could be
844 the histogram. The main distinction from other methods is the supervised
845 approach
- 846 6. Clustering: Having a starting set of random clusters the procedure computes
847 the centroids of these clusters, assigns each point to the closest cluster and re-
848 computes centroids. This is done iteratively until either the point distribution
849 or the centroid position doesn't change significantly between iterations.

850 For a more in depth review of the main methods used in medical image segmenta-
851 tion refer to [26]. Worth noting, at this point, that the semi-automatic segmentation
852 software used in this work uses probably a mixture of region growing and thresh-
853 olding methods, maybe guided by an atlas. The segmentation process generally

854 produces a boolean mask which can be used to select, using pixel-wise multiplication,
855 the region of interest in the image. The next step in image analysis would be to
856 derive information from the region defined during segmentation, in general this step
857 is called feature extraction²⁹ and it's objective is to find non-redundant quantities
858 that meaningfully summarize as much properties of the original data as possible.

859 1.3.2 Radiomics

860 When the images are medical in nature and when referring to high-throughput
861 quantitative analysis the task of finding these features fall in the realm of radiomics,
862 which uses mathematical tools to describe properties of the images that would oth-
863 erwise be unquantifiable to the human eye. The features that can be computed
864 from images are of various types, some of them can be understood somewhat eas-
865 ily through intuition since they are close to what humans generally use to describe
866 images, others are much more complex in definition and quantify more difficulty per-
867 cieved properties of the image. Features can be then roughly classified in different
868 families:

- 869 1. Morphological features: These features describe only the shape of the region of
870 interest, as such they are independent of the pixel values inside the region and
871 hence, to be computed, require only a boolean mask of the segmented region.
872 These features can be further subcategorized as two or three dimensional fea-
873 tures based on whether they focus on single slices or whole volumes. Most of
874 these features compute volumes, lengths, surfaces and shape properties such
875 as sphericity, compactness, flatness and so on.
- 876 2. First order features: These features depend strictly on the gray levels within
877 the region of interest since they evaluate the distribution of these values, as
878 such they need that the boolean mask of the segmentation be multiplied pixel-
879 wise with the origian image to obtain a new image with only the interesting
880 part in it. Most of these features are commonly used quantities, such as
881 Energy, Entropy, Minimum and Maximum value which have been adapted to
882 the imaging context using the histogram of the original image or by considering
883 intensities within an enclosed region.
- 884 3. Higher order features: All the other fatures fall in this macro-category which
885 can be clearly subdivided in other smaller catgories in which the features are
886 obtained following the same guiding principle or starting point. Generally
887 these describe more texture-like properties of the image and, to do so, use
888 particular matrices derived from the original image which contain specific in-
889 formation regarding order and relationships in pixel value positioning within
890 the image. These matrices have very precise definition, as such only the gen-
891 eral idea behind them will be reported here redirecting to [34] for a more strict
892 and in detail description. The matrices from which the features are computed
893 also give name to the smaller categories in this family, these categories are:

²⁹Even if the term is used in pattern recognition as well as machine learning in general

- Gray Level Co-occurrence Matrix (GLCM) features: This matrix expresses how combination of pixel values are distributed in a 2D or 3D region by considering connected all neighbouring pixel in a certain direction with respects to the one in consideration. Using all possible directions for a set distance, usually $\delta=1$ or $\delta=2$, various matrices are obtained and from these a probability distribution can be built and evaluated. It should be noted that before computing these matrices the intensities in the image are discretized.
- Gray Level Run Length Matrix (GLRLM) features: Much like before the task of these features is to quantify the distribution of relative values in gray levels throughout the image, as the name suggests what this matrix quantifies is how long a path can be built by connecting pixel of the same value along a single direction. This time information from the matrices computed by considering different directions are aggregated in different ways to improve rotational invariance of the final features
- Gray Level Size Zone Matrix (GLSZM) features: This matrix counts the number of zones in which voxel have the same discretized gray level. The zones are defined by a notion of connectedness most commonly first neighbouring voxel are considered as connectable if they have the same value, this leads to a 26 neighbouring voxel in 3 dimensions and to 8 connected pixels in 2 dimensions³⁰. The matrix contains in position (i,j) the number of zones of size j in which pixel have value i.
- Neighbouring Gray Tone Difference Matrix (NGTDM) features: Born as an alternative to GLCM these features rely on a matrix that contains the sum of differences between all pixel with a given pixel value and the average of the gray levels in a neighbourhood around them.
- Gray Level Dependence Matrix (GLDM) features: The aim of these features is to capture in a rotationally invariant way the texture and coarseness of the image. This matrix requires the already seen concept of connectedness with a given distance as well as dependence among pixel. Two voxel in a neighbourhood are dependent if the absolute value of the difference between their discretized value is less than a certain threshold. The number of dependent voxel is then counted with a particular approach to guarantee that the value be at least one

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

³⁰To visualize, imagine a 2D grid: the 8 pixel are the four at the sides of each square and the four at the corners.

937 recently the International Biomarker Standardization Initiative (IBSI [34]) wrote a
938 "reference manual" which details in depth the definitions of the features, description
939 of data-processing procedures as well as a set of guidelines for reporting results.
940 This was done in an attempt to reduce as much as possible the variability and lack
941 of reproducibility of radiomic studies. In the past the main attention in radiological
942 research was focused on improving machine performances and evaluating acquisition
943 sequence technologies, however the great developments in artificial intelligence and
944 performance of computers have brought a lot of attention to the field. Various
945 papers, such as [17] and [3] have been written with the objective of presenting the
946 general workflow. By design the topics in this thesis have been presented to resemble
947 the general order of the pipeline, which can be summarised as:

- 948 • Data acquisition
- 949 • Definition of the Region Of Interest (ROI)
- 950 • Pre-processing
- 951 • Feature extraction
- 952 • Feature selection
- 953 • Classification

954 Another interesting possibility offered by the biomarkers computed following ra-
955 diomics is the ability to quantify the differences between successive exams of the
956 same patient. This specific branch of radiomics is called Delta-radiomics, refer-
957 encing to the time differential that become the main focus of the analysis. **NON**
958 **SAPREI SE VA BENE COSÌ' O BISOGNA AGGIUNGERE ALTRI**
959 **DETTAGLI**

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

963 1.4 Survival Analysis

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

970 The main concepts and terminologies in this field are:

- 971 1. Event: This is the phenomenon under analysis, generally it could be death,
972 remission from recovery, recovery It is common to refer to the event as
973 failure since usually the event is a negative event.

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

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

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

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

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

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

1003 5. Hazard Function **h(t)**: This function is difficult to explain practically, cit-
1004 ing [16] "**The hazard function $h(t)$ gives the instantaneous potential**
1005 **per unit time for the event to occur, given that the individual has**
1006 **survived up to time t** ". The mathematical definition is:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (1.12)$$

1007 Dividing by a time interval the hazard function can also be intended as condi-
1008 tional failure rate, where the conditional refers to the "given that the individual
1009 has survived up to time t ". Being a rate this quantity need not be bound in
1010 $[0,1]$ but ranges in $[0, \infty]$ and depends on the time unit used. This will be
1011 interpreted as $h(t)$ events per unit time.

1012 The hazard function is non-negative and has no upper bounds, since it can be

related to the survival function ³¹ the possible shapes give different names to the final model. These could be increasing or decreasing Weibull, exponential or lognormal survival models

A further step in the analysis could be to try to measure the differences in survival between two groups, i.e. using a single variable expected to drive the differences in survival. The most basic example would be looking at the effectiveness of a drug by dividing in group A and B the patients given the actual medicine and the placebo respectively, however this could be done even for males vs females or, in the case of continuous variables, for patients with above or below threshold values ³². When the analysis is univariate in nature then the Kaplan-Meier survival curves are used in conjunction with the log-rank test, when the analysis is multivariate then the Cox Proportional-Hazard model is used. The Cox model is very similar to linear and logistic regression.

Explaining these methods is beyond the scopes of this thesis, all details can be found in [16]. Suffices to say that, as rule of thumb, two Kaplan-Meier curves that do not intersect at any point indicate good separation among the groups which can then be evaluated using the p-value from the log-rank test. When it comes to Cox models they can be used to create a new variable, much like the linear combination of variables in linear regression, which in turn can be used to identify groups in the patient cohort. This division in groups can be obtained using the median, or the percentiles if one desires more than two groups. The division in the resulting groups can be used as proxy for the quality of the score built using the Cox model.

³¹The 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.

³²Generally, when no obvious threshold is available, the median of the variable is chosen.

¹⁰³⁵ **Chapter 2**

¹⁰³⁶ **Materials and methodologies**

¹⁰³⁷ In this section there's going to be an explanation of the dataset as well as instruments
¹⁰³⁸ and methodologies used to analyze it's properties, as such the first step is going to
¹⁰³⁹ be an in depth discussion of the data available and a general overview of the final
¹⁰⁴⁰ use. The following step is going to be a description of the preliminary work done to
¹⁰⁴¹ the data itself and to the results of this preliminary analysis in order to select the
¹⁰⁴² important features. The final step of this chapter is going to be an explanation of
¹⁰⁴³ the methods used to derive the final results and to evaluate them.

¹⁰⁴⁴ **2.1 Data and objective**

¹⁰⁴⁵ The objective of this thesis will be to compare how different methods perform in
¹⁰⁴⁶ predicting clinical outcomes in covid patients, while also determining if different
¹⁰⁴⁷ kind of input data imply different performances of the same methods. All images are
¹⁰⁴⁸ non segmented, as such all of them are going to be semi-automatically segmented
¹⁰⁴⁹ via a new software being tested in the medical physics department called *Sophia*
¹⁰⁵⁰ *Radiomics*, which will be briefly explained shortly. Statistical analysis are going to be
¹⁰⁵¹ performed mainly in python using libraries such as scikit-learn, pandas, numpy, scipy
¹⁰⁵² while the graphical part of the analysis is done with either seaborn or matplotlib.
¹⁰⁵³ The starting dataset was a list of all the patients that, from 02/2020 to 05/2021,
¹⁰⁵⁴ were hospitalized as COVID-19 positive inside the facilities of *Azienda ospedaliero-*
¹⁰⁵⁵ *universitaria di Bologna - Policlinico Sant'Orsola-Malpighi*. As far as exclusion
¹⁰⁵⁶ criteria go the main exclusion criteria, except unavailability of the feature related to
¹⁰⁵⁷ the patient, was visibly damaged or lower quality images, for example images with
¹⁰⁵⁸ cropped lungs. The first set of selection criteria were:

- ¹⁰⁵⁹ • All patients that had undergone a CT exam which was retrievable via the
¹⁰⁶⁰ PACS (Picture Archiving and Communication System) of *Azienda ospedaliero-*
¹⁰⁶¹ *universitaria di Bologna - Policlinico Sant'Orsola-Malpighi*
- ¹⁰⁶² • All patients that had all of the clinical and laboratory features, listed in
¹⁰⁶³ fig:2.1, suggested by Lidia Strigari and medical professionals.
- ¹⁰⁶⁴ • Since all patient had at least 2 CT exams only the closest date to the hospital
¹⁰⁶⁵ 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	0
CURB65	30	7%	3	Sex	284	65% Male	Lung consolidation	225	52%	1	1
CURB65	4	1%	4	Sex	1	0%	Hypertension	194	45%	0	0
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				
MEWS score	10	2%	6	cPAP	368	84%	0				
MEWS score	1	0%	7	cPAP	68	16%	1				

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 ??.

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

1088 in how much they add for each condition and the set of symptoms the check
1089 for.

- 1090 (a) MulBSTA: This score accounts for **M**ultilobe lung involvement, absolute
1091 **L**ymphocyte count, **B**acterial coinfection, history of **S**moking, history of
1092 **h**yper**T**ension and **A**ge over 60 yrs. [10]
- 1093 (b) MEWS: Modified Early Warning Score for clinical deterioration. Com-
1094 puted considering systolic blood pressure, heart rate, respiratory rate,
1095 temperature and AVPU(Alert Voice Pain Unresponsive) score. [30]
- 1096 (c) CURB65: **C**onfusion, blood **U**rea Nitrogen or Urea level, **R**espiratory
1097 **B**lood pressure, age over **65** years. This score is specific for pneu-
1098 monia severity [33]
- 1099 (d) SOFA: **S**equential **O**rgan **F**ailure **A**sessment score. Considers various
1100 quantities from all systems to assess the overall state of the patient,
1101 $\text{PaO}_2/\text{FiO}_2$ ¹ for respiratory system, Glasgow Coma scale² for ner-
1102 vous, mean pressure for cardiovascular, Bilirubin levels for liver, platelets
1103 for coagulation and creatine for kidneys [2]
- 1104 (e) qSOFA: **q**uick SOFA. Only considers pressure, high respiratory rate and
1105 the low values in the Glasgow scale.

1106 This procedure produced a starting cohort of \sim 700 patients which, having all
1107 various images available, created a huge set of \sim 2200 CT scans. Since this analysis is
1108 focused on radiomics there is an evident need for as much consistency as possible in
1109 the images analysed. For this reason all CTs taken with medium of contrast were ex-
1110 cluded, since they would have brightnesses not indicative of the disease, and for every
1111 patient only images with thin slice reconstruction were considered. More specifically
1112 only images with slice thickness of 1 or 1.25 mm³ along the z-axis were taken into
1113 consideration, which meant excluding all the 1.5, 2, 2.5 and 5 mm slice thicknesses.
1114 Since all these images were segmented by me and two other students using a semi-
1115 automatic segmentation tool provided by the hospital it sometimes happened that
1116 manual corrections were necessary, these were performed only in very obvious and
1117 simple cases while, in all remaining cases, the patient was dropped out of the study.
1118 Overall this left the final study cohort to be composed of 435 patients, all descriptions

¹Very unrefined yet widely used indicator for lung dysfunction

²GCS 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

³This 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.

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. NON SO SE POSSO SCRIVERE CHE IL PROGRAMMA ERA DELUDENTE NE' QUANTO POSSO SBILANCIARMI A DARE INFO CHE NON HO SULLA SEGMENTAZIONE, SE CON LA STRIGARI HANNO PUBBLICATO L'ARTICOLO SULLA IBSI-COMPLIANCE SI POTREBBE CITARE QUELLO

non qui, magari nella discussione

So, having segmented all the images and extracted all the features supported in the software, the next step is the definition of the actual analysis pipeline. The whole dataset was comprised of ~200 features, 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.

forse usare i smallcaps per i nomi delle variabili aiuta a separarle dal resto del testo

- The discrete ones were the aforementioned scores and the boolean ones were sex of the patient, obesity status, if the patient had a fever⁴ 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 [34] with the addition of fat and muscle surface, computed as cm² 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

⁴Defined as body temperature > 38°

1159 Bilaterality of lung damage, presence of Groun Glass Opacities (GGO), lung
1160 consolidations as well as crazy paving were used to see if they were sufficient
1161 in determining outcome.

1162 2.2 Preprocessing and data analysis

1163 inserirei un flowchart della procedura, aiuta a capire

1164 Before any preprocessing a choice was made to exclude all of the clinical scores,
1165 this was done to avoid having them mask other, more straightforward variables.
1166 **QUESTO ANDRA MESSO IN UNA FOOTNOTEIn APPENDICE SE**
1167 **MAI RIESCO A SCRIVERLA** an alternative choice will be made, namely to
1168 keep only the most strongly predicting one out of all of them. The first step in the
1169 analysis of this data is going to be a lasso regularized regression using either death or
1170 ICU admission as target. As mentioned before when operationg with regressions it's
1171 a necessity that the residuals be normally distributed, a common way to get as close
1172 as possible to this hypothesis is to boxcox transform the data. Apart from the data
1173 which contained negative values, which cannot be fed into the boxcox transform, all
1174 variables have been transformed using this method.

1175 The quatile-quantile plots have been used for a visual check of normality, nor-
1176 mally distributed data will populate the bisector of the graph while deviations are
1177 symptoms of non-normality. Heavy and light tails are visible as deviations respec-
1178 tively above and belox the bisector. It should be noted that when the data is nor-
1179 mally distributed then the transform doesn't change much the distribution while,
1180 in cases with more heavy tailed distributions, the improvement is clear to see as it
1181 can be seen in Figures 2.2 and 2.3 The next preprocessing step has been to apply
1182 a *StandardScaler* to all of the features, which corresponds to subtracting the mean
1183 and dividing by the standard deviation, in order to center all the features around
1184 zero. The final step in preprocessing has been to reduce the features by using a cor-
1185 relation threshold which means that all variables that correlate with another more,
1186 in absolute value, than a certain threshold, which has been set to 0.6 in this work,
1187 are dropped a priori. A rather important thing to notice is that correlation has been
1188 computed using Spearman correlation and not Pearson since the first is invariant
1189 under monotone transformations, such as boxcox, and the second is not. Given
1190 the large number of features it's very plausible that at least one of the eliminated
1191 features is correlated with all other dropped features but with none of the remaing
1192 ones, since a Lasso regularization will be used introducing a few redundant features
1193 is not too damaging and the possible benefits outweigh the risks. For this reason
1194 a redrawing method has been implemented to add one of the dropped features, this
1195 has been done by choosing the one that most correlates with the label being used. A
1196 pivotal point in all of this analysis is that, to obtain reasonable values in the cross-
1197 validation procedures and to avoid leakage⁵ problems, all of the preprocessing steps
1198 have been done after train-test splitting the data on the train set and then applied
1199 as defined during training on the test dataset. When it comes to cross-validation

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

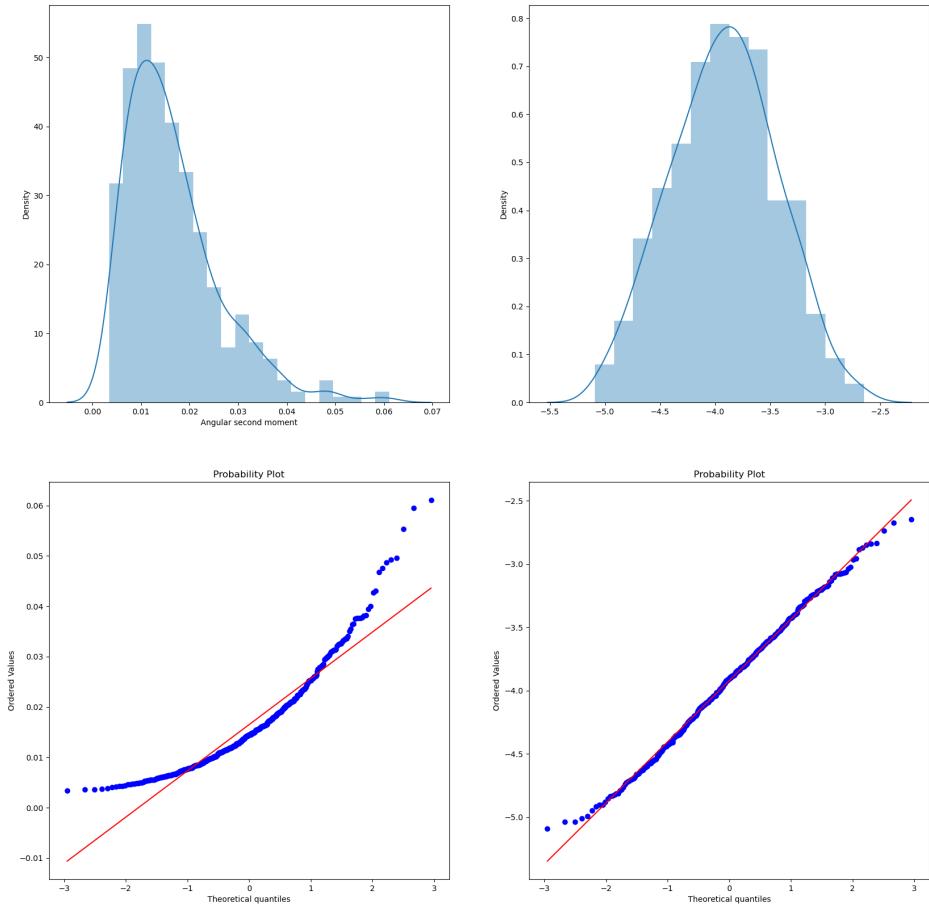


Figure 2.2: Example of boxcox applied to the radiomic feature *Angular second moment* which has a heavy tailed distribution. The graphs contain the original distribution (top-left) the transformed distribution (top-right) and the two respective quantile-quantile plots

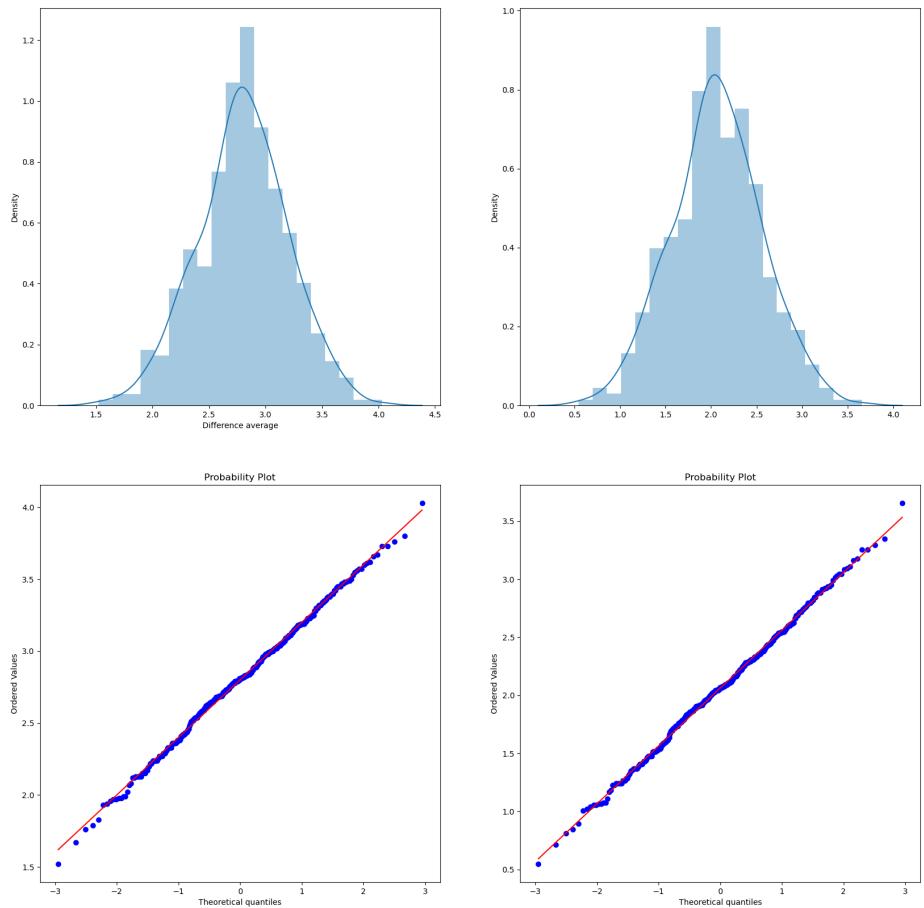


Figure 2.3: Example of boxcox applied to the radiomic feature *Difference Average* which has a close to normal distribution. The graphs contain the original distribution (top-left) the transformed distribution (top-right) and the two respective quantile-quantile plots

procedure the choice was made to use a stratified k-fold approach with k=10, the data is split in 10 parts with the same percentages of labels⁶ then a model is built by training on 9 of the folds and it's performance is then tested on the remaining fold. To use the whole dataset for testing a prediction of it has been built by combinig the predictions on the 10th" fold for ten different models trained on the respective 9 remaining folds. The lasso model from training on the 9 folds is actually chosen as the model with the hyperparameters that give the best performance with another 10-fold crossvalidation. This has been obtained by using *cross_val_predict* on a model obtained by including a *LassoCV* step inside a *pipeline* from scikit-learn library in python. The performance of the cross-validated predictions that, when built this way, is much more representative of the real-world performance of the model, has been evaluated using ROC curves and AUC. Different models have been compared with a Delong test[7] for the significance of difference in the ROC curves.

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

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

Ad esempio la questione della divisione in ondate

Finally a few dimensionality reduction techniques, namely the unsupervised PCA[8] and Umap [21] and the supervised PLS-DA[4], have been used to understand better the state of the data and further explain some of the obtained results. These peculiar analyses will be reported in the Appendix

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

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

⁶The 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

1237

Chapter 3

1238

Results

1239 In this chapter the result obtained with the methods explained in the previous
1240 chapters will be briefly presented to ease in the discussion in the final chapter.

1241

3.1 Feature selection through Lasso regularization and clinical outcomes prediction using 1242 regression

1244 When it comes to lasso regression usually graphs are reported that show the convergence
1245 of the parameters to the final value. Since the real information of this process
1246 is the value to which the coefficients converge only these values will be presented in
1247 tables and the ROC curves, with respective AUCs, will be provided. An example
1248 of the graph i'm talking about is the one visible in Figure 3.1.

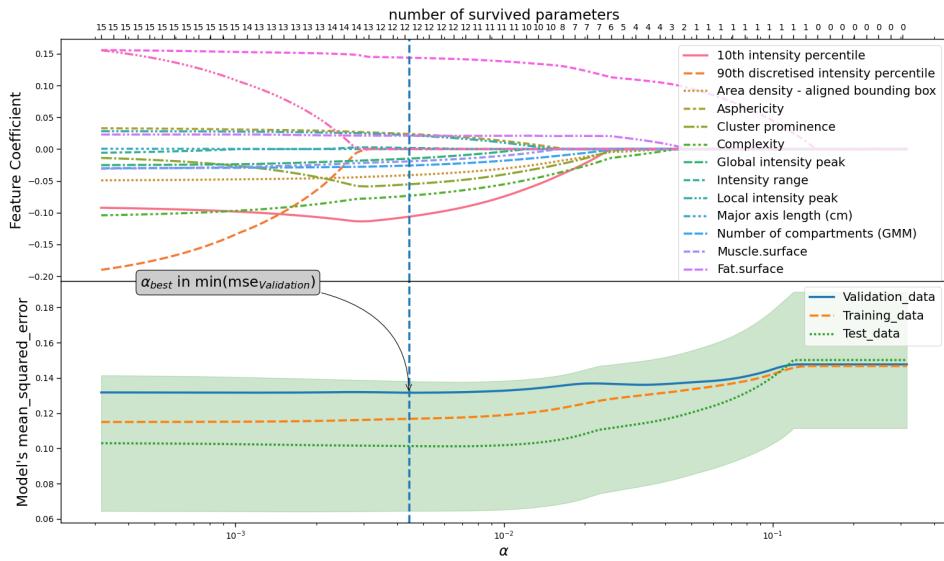


Figure 3.1: Example of graph representing the convergence of the coefficients in a lasso procedure. The final model is chosen by looking at the lowest value in mean squared error computed on the testing dataset. This graph is relative to only radiomic features with the model using DEATH as label.

1249 All of the graphs with respective captions will be, maybe, put in the final app-
 1250 pendix. Finally it seems useful to report the following contingency table that gives
 1251 an idea on how superimposed the labels on death and ICU ADMISSION are.

		ICU ADMISSION	
		0	1
DEATH	0	311	47
	1	48	30

1253 3.1.1 Using Death as label

1254 Starting to predict the death outcome of the patient different groups of features
 1255 have been used, first of all only the radiomic features have been used. The ROC
 1256 curve obtained with the radiomic features is the one in Figure 3.2. The curve in
 1257 bold is an average curve obtained by aggregating the ten curves relative to each of
 1258 the folds used in testing and the gray band represents a ± 1 standard deviation.

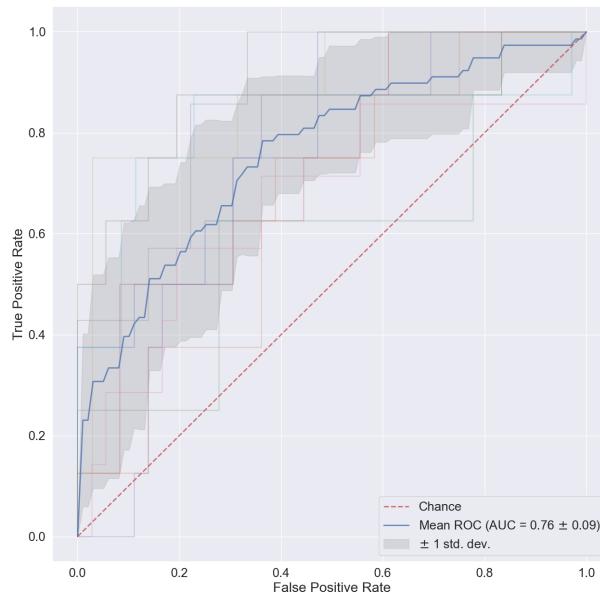


Figure 3.2: ROC curves obtained with crossvalidation procedure using the radiomic features alone. In bold is the mean ROC with gray bands of width equal to the standard deviation.

1259 The model was built using the coefficients reported in 3.1 reaches a $AUC =$
 1260 0.76 ± 0.09 . The features inside the tabular, as well as those in the tabulars that
 1261 follow, will have the coefficients in descending order by absolute value so that the
 1262 top features are the most relevant within its relative model.

1263 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

1265 it comes to the clinical features, the results reported in Figure3.3 and Table3.7 have
 1266 been obtained. All the results will be put together for ease in Table 3.5.

1267 And finally, considering the radiological features Figure3.4 and Table 3.8 are
 1268 obtained.

1269 The first thing to notice is that the radiological features, when considered alone,
 1270 have close to null predictive power. This is reasonable for at least part of the
 1271 features because there is no reason for acquisition parameter to actually influence the
 1272 outcome of the patient. When it comes to the radiologically determined quantities,
 1273 such as GGO, Crazy paving, lung consolidation and bilaterality even if one would
 1274 expect these to be relevant their distribution across the dataset is not conducive 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

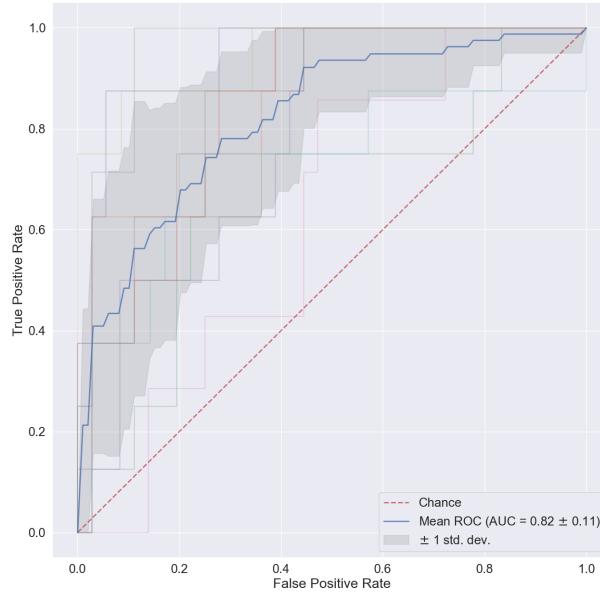


Figure 3.3: ROC curves obtained with crossvalidation procedure using the clinical features alone. In bold is the mean ROC with gray bands of width equal to the standard deviation.

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

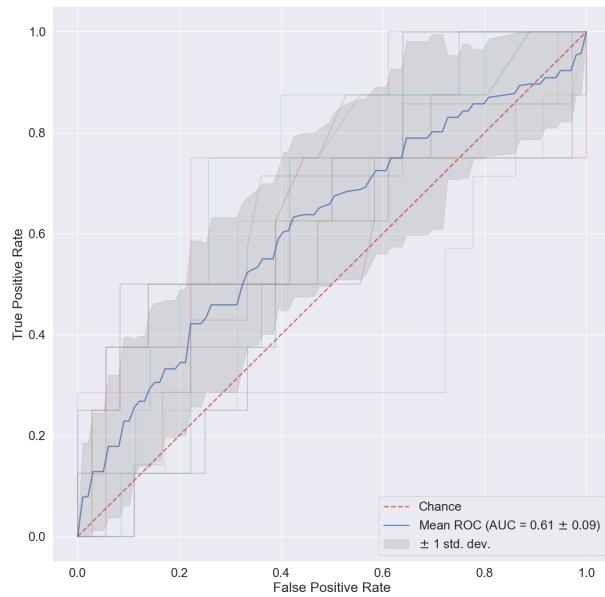


Figure 3.4: ROC curves obtained with crossvalidation procedure using the radiological features alone. As before in bold is the mean ROC with bands of width equal to the standard deviation.

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

- 10th 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.

1294 To summarize, since all of the features are computed on the whole lung segmen-
 1295 tation, it seems that some information on the distribution of gray levels as well
 1296 as some textural information inside the whole lung are important. It also seems that
 1297 some information on the shape of the organ itself is also relevant. One would expect
 1298 that when combining all of the available features, i.e. by building a model using the
 1299 previous clinical, radiomic and radiological features, the performance should some-
 1300 what rise especially given the fact that clinical and radiomic features have almost
 1301 the same performance. The combined results can be seen in Figure 3.5 and Table
 1302 3.9.

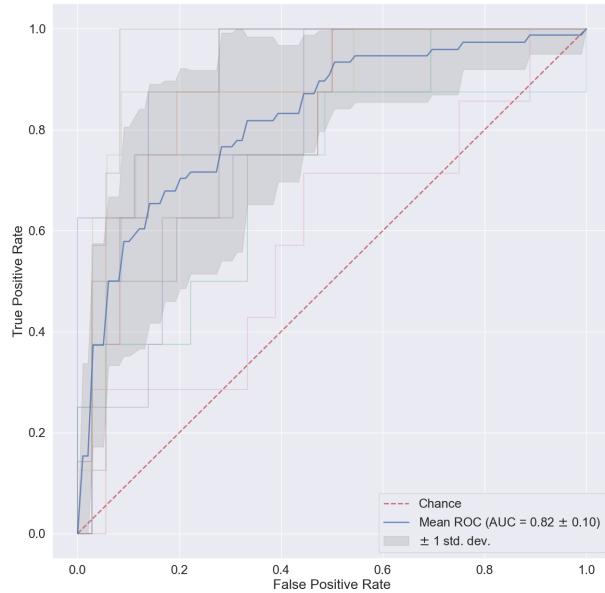


Figure 3.5: ROC curves obtained with crossvalidation procedure using all the available features. Model obtained with Lasso regularization of a linear regression modelling death

1303 In spite of what the expectations were the performance not only doesn't improve
 1304 but it doesn't even change. To be sure of this claim a Delong test was used to
 1305 compare pairwise the receiver operator curves and their respective AUCs. The null
 1306 hypothesis of this test is that the two models are the same, hence a p-value smaller
 1307 than 0.05 means that the curves and their aucs are statistically different. The results
 1308 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

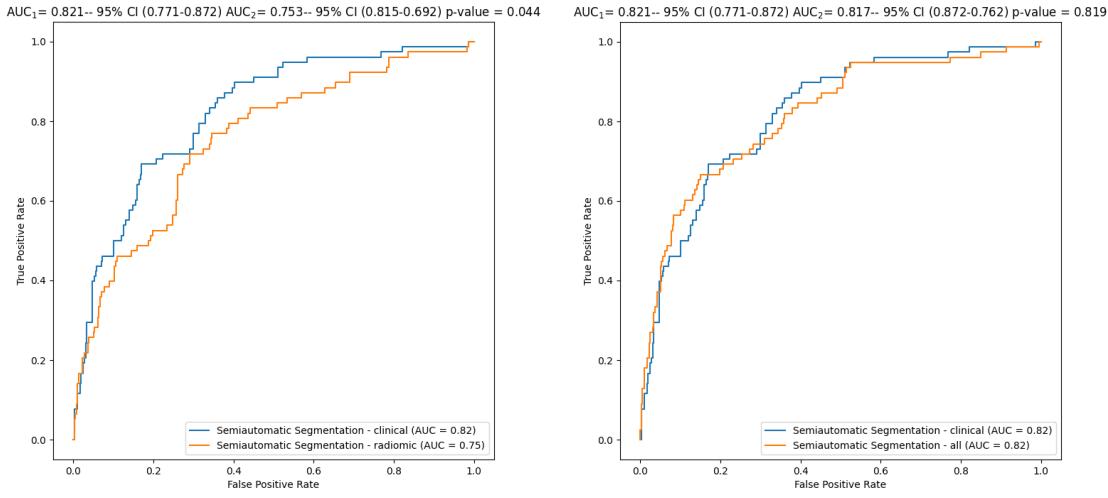
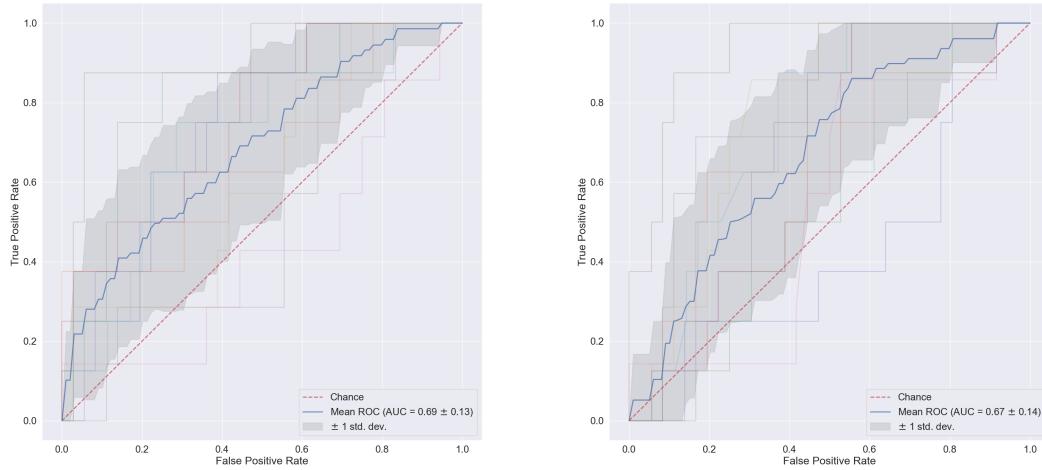


Figure 3.6: Comparison between ROC curves for clinical vs radiomic curves (a) and clinical vs all (b). The p-values are obtained with a Delong Test

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

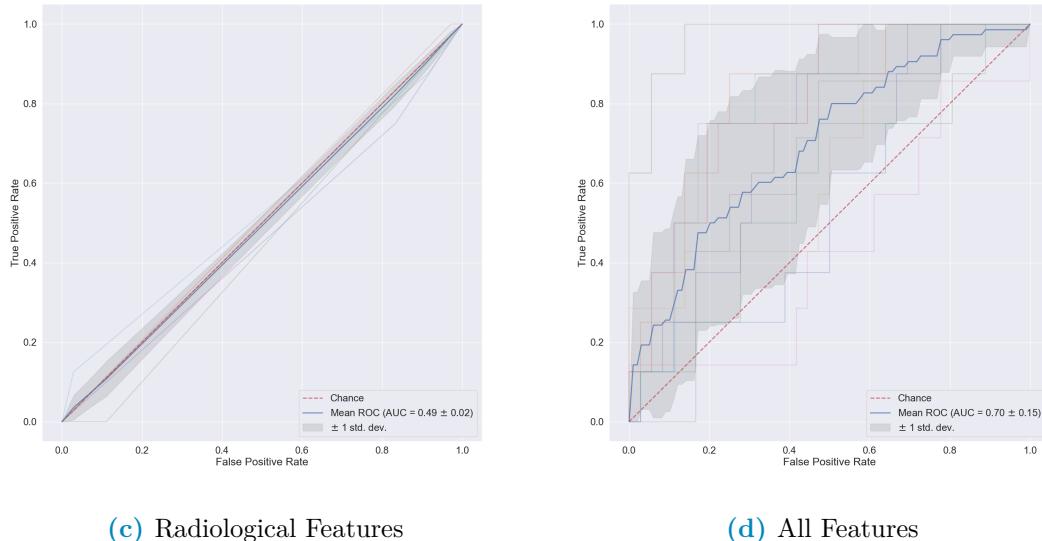
3.1.2 Using ICU Admission as label

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



(a) Radiomic Features

(b) Clinical Features



(c) Radiological Features

(d) All Features

Figure 3.7: 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

1327 Compared to before the performance is definitely worse. The radiological features
 1328 have the same performance of a random variable, which is not that concerning given
 1329 their expected impact on gravity of the clinical picture of the patient. Even if
 1330 superfluous a Delong test was used to confirm that the hypothesis of the curves
 1331 being equal could not be rejected. When it comes to the relevant features in each
 1332 model the following can be deduced:

- 1333 • For the clinical features all of them have a role in the prediction. The only
 1334 surprising fact, even if it keeps a certain degree of plausibility, is that age is the
 1335 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 previously described Lasso regression. Once again, following the description scheme used in the section dedicated to Lasso, the results will be divided using the two labels ICU ADMISSION and Death.

3.2.1 Classifying the outcome death

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

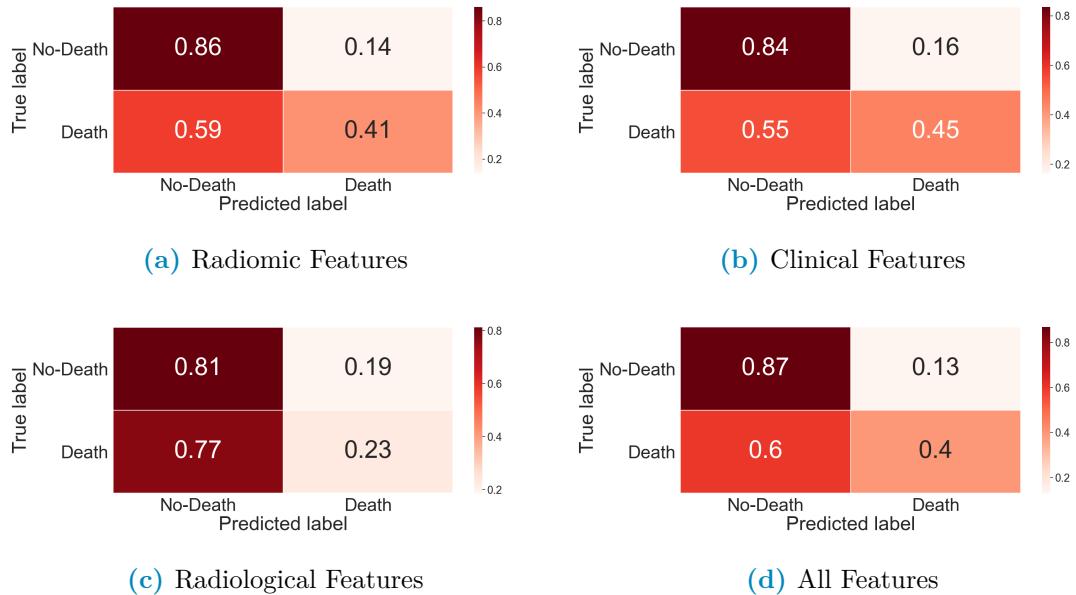


Figure 3.8: Confusion matrices for Random Forest cross-validated predictions after training on Synthetically oversampled data to predict DEATH . All of the available feature families are the reported

1355 Even without looking at the ROC curves it's plain to see that the data at hand
 1356 is proving to be difficult for this model. Much like before radiological features alone
 1357 are useless. In evaluating these confusion matrices it should be kept in mind that the
 1358 data is heavily unbalanced, since only ~15% of the patient died or were admitted
 1359 in the ICU. Even when using SMOTE in the training phase to correct this probelm
 1360 it seems that the classifiers learns that it's optimal to guess that someone is alive.
 1361 When it comes to the ROC curves Figure 3.9 and Table 3.11 summarises the results

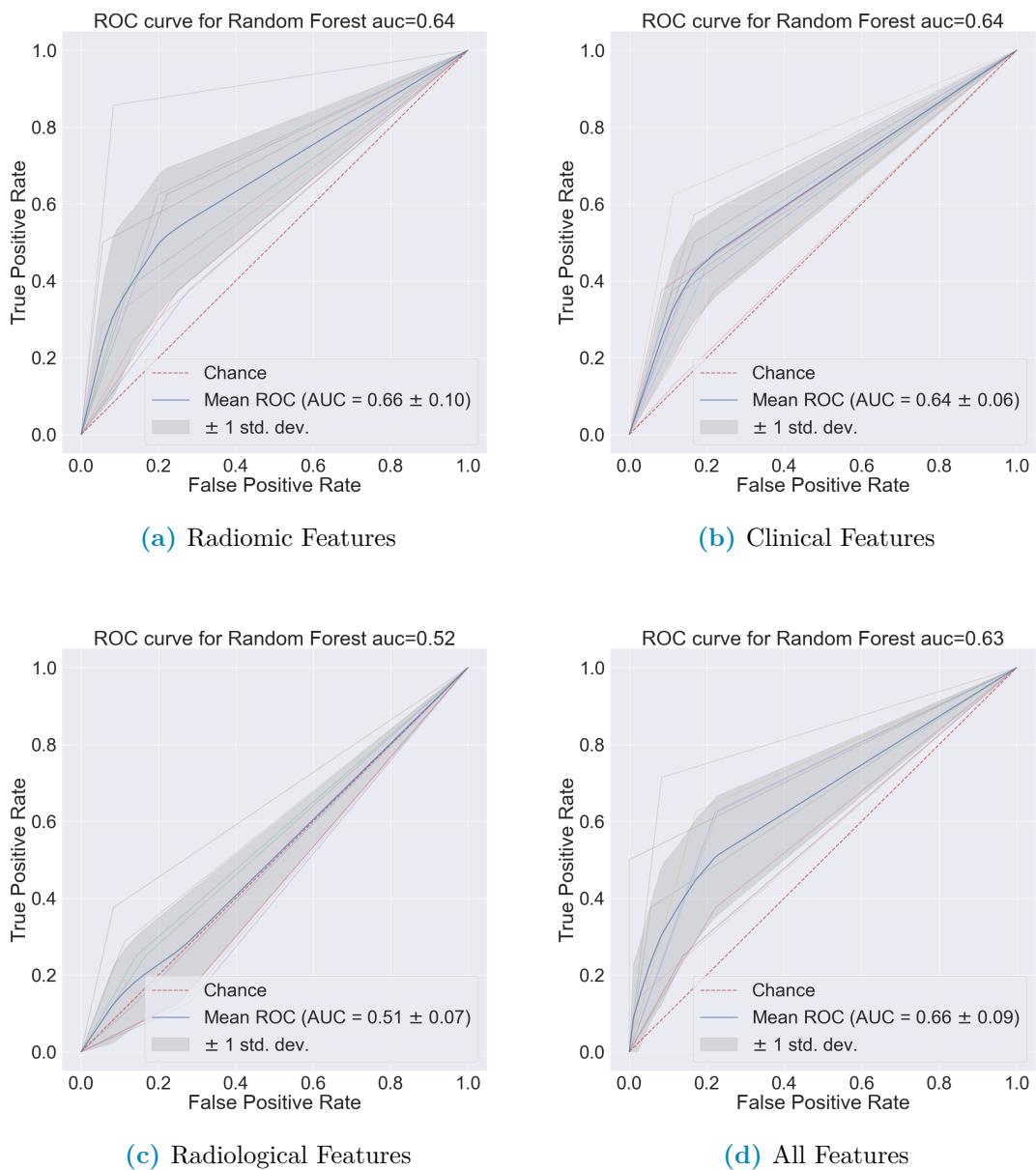


Figure 3.9: Cross-validated ROC curves built with Random forest classifier predictions of DEATH . Performances of allvariable families are reported

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

1362 Once again the curves are evidently not statistically different. This counterintuitive behaviour seems to be constant across the two implemented methods and
 1363 also across different labels tried. An attempt to explain this phenomenon will be
 1364 postponed to the end of the next subsection
 1365

1366 3.2.2 Classifying the outcome ICU Admission

1367 Even when using the admission in the ICU the performance remains pretty much
 1368 the same, so the comments would still be the same as before

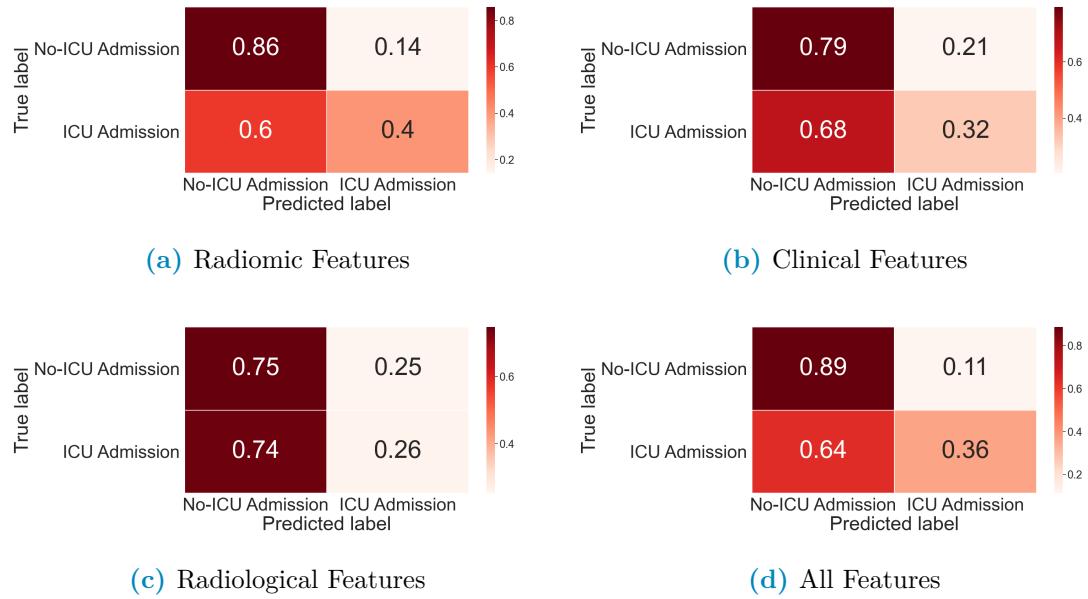
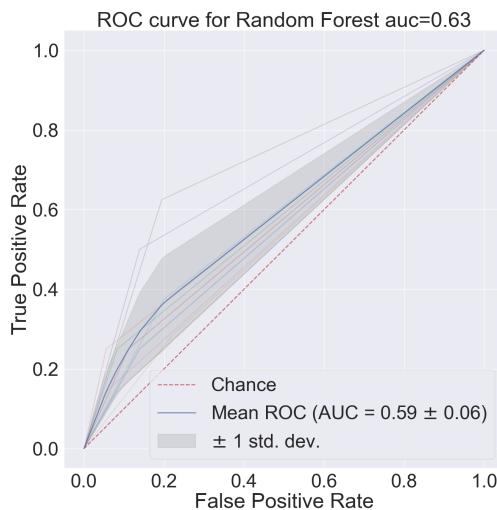


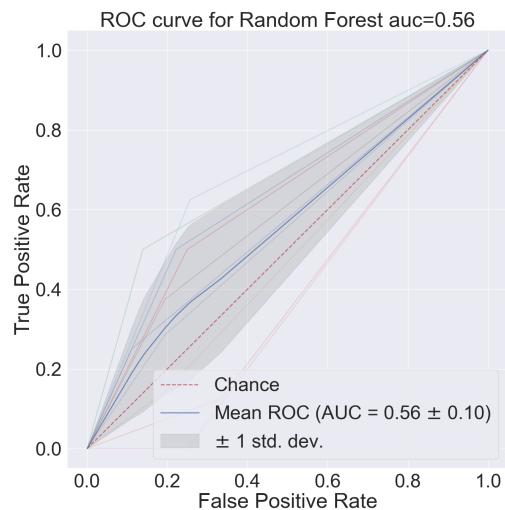
Figure 3.11: Confusion matrices for Random Forest cross-validated predictions after training on Synthetically oversampled data predicting ICU ADMISSION . All of the available feature families are the reported

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

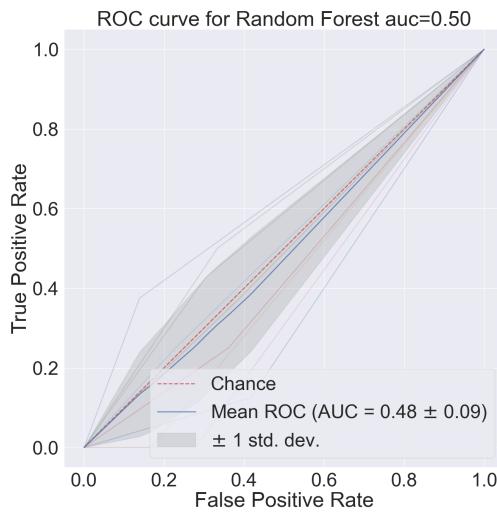
Features used	mean AUC \pm std
Radiomic	0.62 ± 0.08
Clinical	0.56 ± 0.08
Radiological	0.51 ± 0.11
All	0.64 ± 0.09



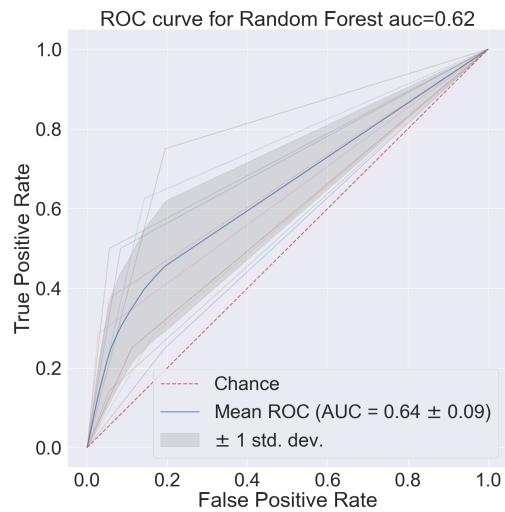
(a) Radiomic Features



(b) Clinical Features



(c) Radiological Features



(d) All Features

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

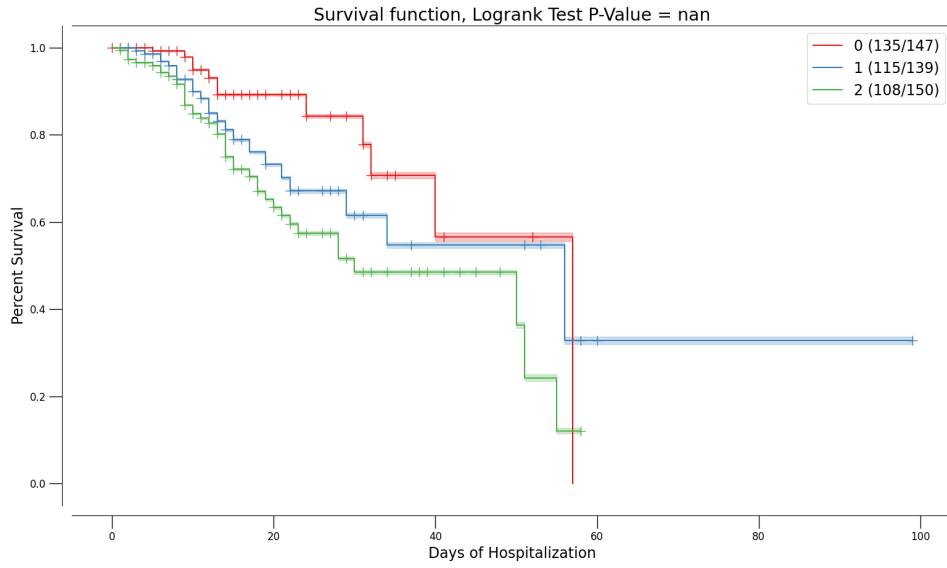
3.2.3 Using survival analysis

¹³⁶⁹ Following the preprocessing steps delineated before a Cox Proportional-Hazard pro-
¹³⁷⁰ duces the results presented in Table 3.13

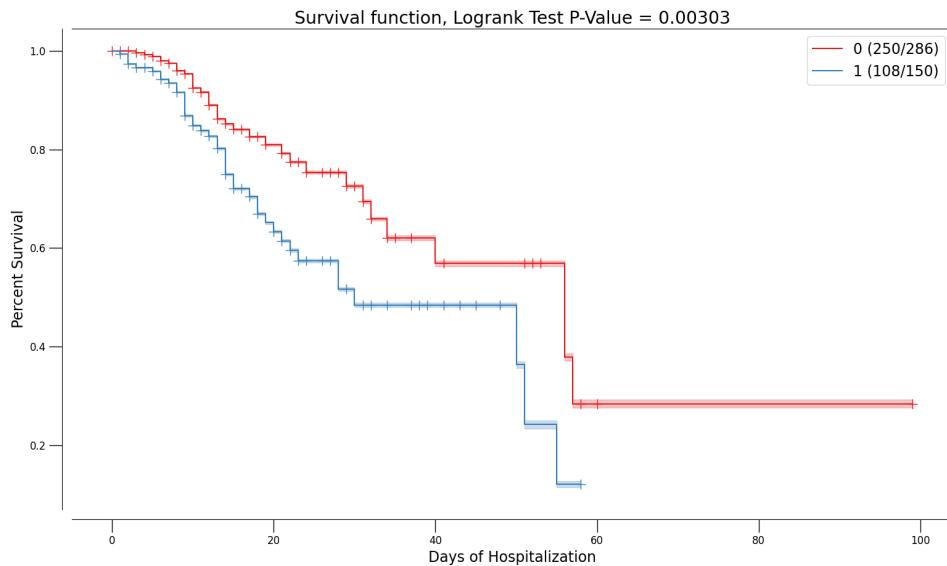
¹³⁷¹ A score was built automatically using the predict method of the CoxPH fitter and
¹³⁷² assigned to each patient of the dataset using the previously described cross-validated
¹³⁷³ prediction procedure. To see if the prediction was representative of differences in
¹³⁷⁴ the individuated populations first the Kaplan-Meier curves according to thirds in
¹³⁷⁵ the score distribution were used and then the score was binarized using the 66th
¹³⁷⁶ percentile in the score distribution as threshold. The results of this procedures are
¹³⁷⁷ 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	0.188832	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	0.944225	
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 0-66th percentile and 66th to 100th

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

1379 It can be seen that the groups built in this ways can be used to drive some
 1380 differences in survival, when binarizing the score obtained with Cox the curves also
 1381 turn out to be significantly different.

1382 Finally, to better understand the data as well as the situation in the hospital,
 1383 the population was divided in two subgroups according to the date of admission.
 1384 The two groups were representatively called 1st and 2nd wave and the division was
 1385 drawn on the 20th of July 2020 and the results can be seen in Figure 3.14.

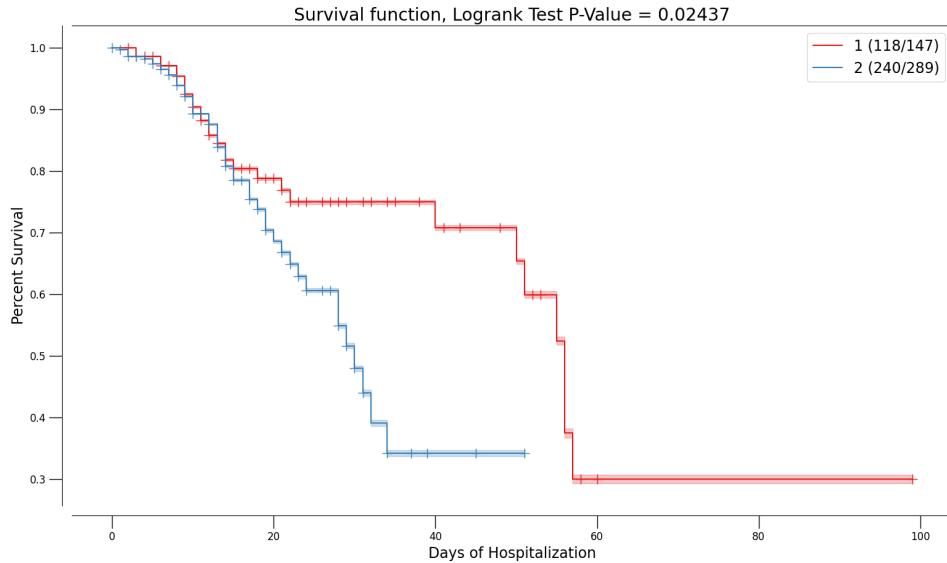


Figure 3.14: Kaplan-Meyer curves for patient admitted before (red curve) and after (blue curve) 20/07/2020

1386 Non ricordo come mai avevamo fatto la distinzione in ondate

1387 It can be seen that there is a statistical difference in survival between the pa-
 1388 tients admitted in the first wave vs those admitted in the second. Furthermore this
 1389 difference is quite perplexing as it seems to indicate that people in the second wave
 1390 died more than people in the first wave, which seems counter-intuitive given that one
 1391 would expect the experience from the previous wave to improve performance. The
 1392 most reasonable explanation for this fact is the change in admission policy as time
 1393 advanced. Probably in the first wave, when still little was known on *Sars-COVID19*
 1394 patients, more people were addmitted in less problematic condition whereas in the
 1395 second wave, having understood better what were the most dangerous cases as well
 1396 as in an attempt to admitt only those strictly in need, most of the admitted patients
 1397 were in more critical condition.

1398 It's also possible that this result that has been obtained could be a symptom of
 1399 subtle differences in the two *Sars-COVID19* manifestations, as if to indicate different
 1400 variants. Further analysis in this direction could be a followup work of this thesis.

1401 3.3 Summarizing the results and further analysis

1402 Forse già questo può andare bene come conclusione?

1403 Having presented all of the results obtained in this thesis it has been seen that:

- 1404 • The chosen preprocessing method followed by Lasso regularized regressions
 1405 perform overall well when it comes to predicting either DEATH or ICU AD-
 1406 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 ~ 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

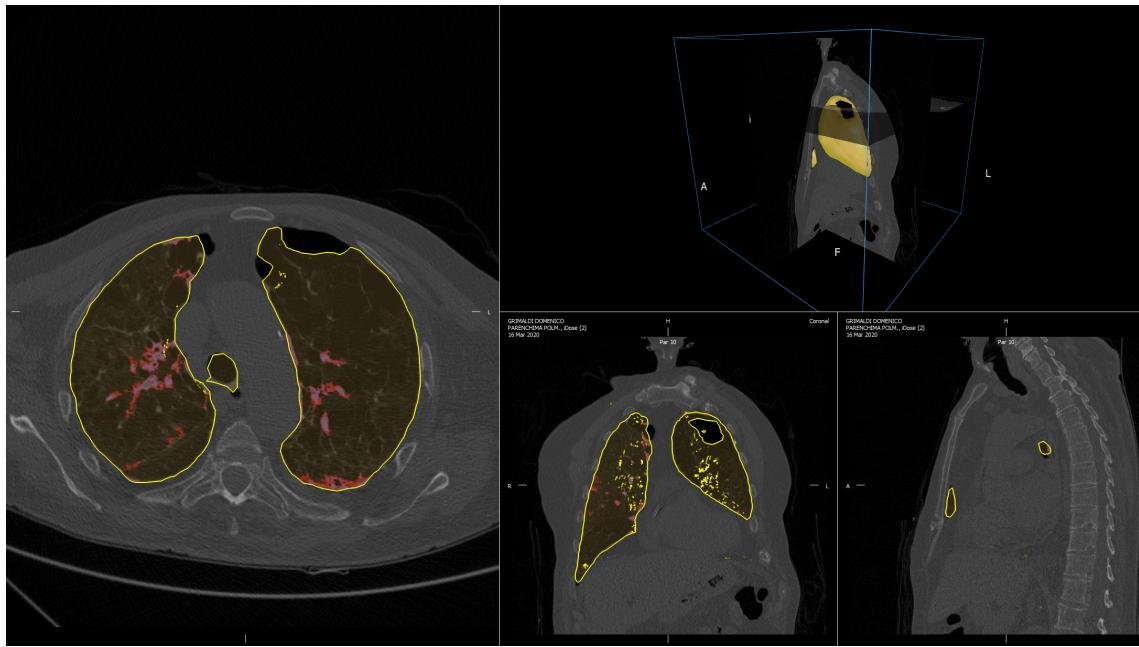
It is very clear that the first three hypotheses are verified with a caveat, the performance of the radiomic pipeline hinges on the quality of the images and of the segmentation procedure performed on them. Since the images used in this thesis were, are and will be used by the hospital in routine processes 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

1448 to visualize the data and to prepare for clustering in case of need, all of the results
1449 of these procedure will be presented in the followig section.

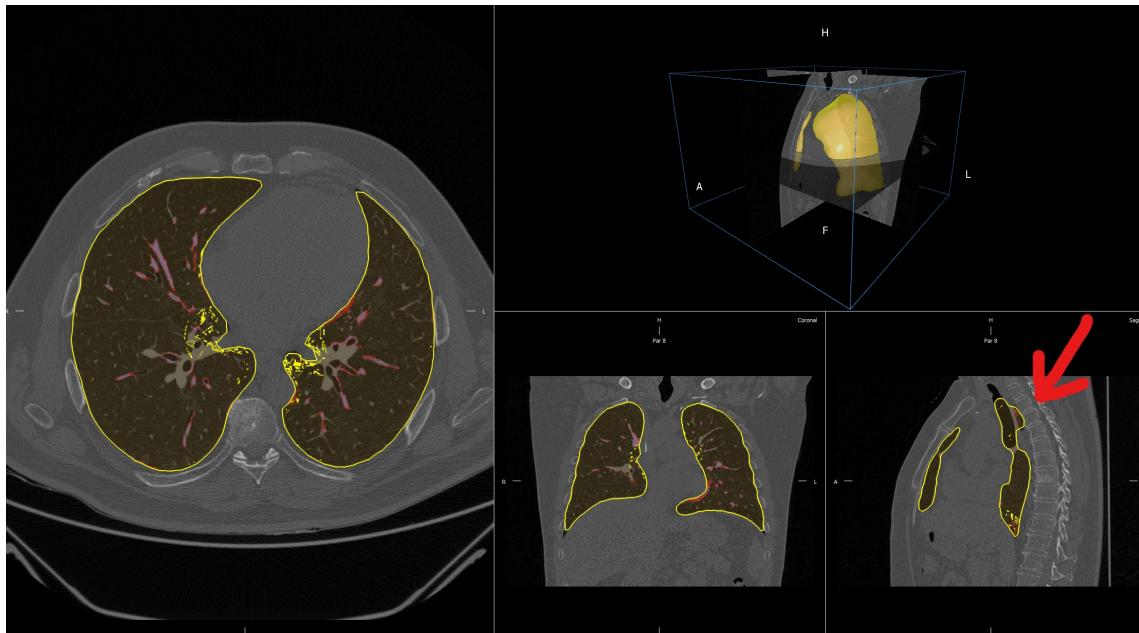
1450 A questo punto magari cito l'appendice con la parte di clustering quando sarà
pronta

1451 To give a qualitative idea of the situation regarding the segmentations ~ 70 were
1452 looked at and evaluated as good, unsure or bad.¹ Good segmentations are those
1453 in which an untrained professional could not see any problems, unsure are those in
1454 which there are small inaccuracies, such as lungs that connect in some points and
1455 the inclusion of the trachea. Finally segmentation were classified as bad in cases
1456 with obvious errors, such as part of the intestine being labelled as lung, or damages
1457 in the lung being labelled as outside tissue.

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



(a) Example of segmentation classified as bad



(b) Example of segmentation classified as dubious

Figure 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.

1458 The result of this qualitative analysis can be seen in Table 3.14 in which the
 1459 incidence of both DEATH and ICU ADMISSION labels is computed in all possible
 1460 segmenation categories.

1461 This table suggests that there is large difference of very severe individuals which
 1462 end up being not perfectly segmented. It's possible, even if unlikely, that the sample
 1463 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

¹⁴⁶⁴ 3.14, it's reasonable to fear that in the most important cases, which means those that
¹⁴⁶⁵ correspond to 1 labels in DEATH or ICU ADMISSION , the values of the radiomic
¹⁴⁶⁶ features used in this thesis were not the best possible.

¹⁴⁶⁷

Chapter 4

¹⁴⁶⁸

Conclusion

¹⁴⁶⁹ In this thesis varoius methods were used in an attempt to predict the prognosis of
¹⁴⁷⁰ *Sars-COVID19* patients.

¹⁴⁷¹ Regularized regression was used to predict clinical outcome using various families
¹⁴⁷² of variables to compare the information hidden in each of them and to evaluate if
¹⁴⁷³ CT exams add any value to a small set of clinical variables.

¹⁴⁷⁴ Random forest classifiers were used with the same aim. As a by-product these
¹⁴⁷⁵ two methods can be compared to see which, given the same data, can extract the
¹⁴⁷⁶ most information.

¹⁴⁷⁷ Survival analysis was used, mostly by itself, to see if the data could be divided
¹⁴⁷⁸ in smaller groups with different survival functions.

¹⁴⁷⁹ In the first two lines developement it was found that, in the specific case of
¹⁴⁸⁰ data available for this thesis, radiomic features added no information to the clinical
¹⁴⁸¹ variables. This was taken as a (grim) reminder that the quality of the segmen-
¹⁴⁸² tations is very relevant in determining the results obtained by radiomics analysis and
¹⁴⁸³ that current commercial segmentation techniques based on thresholding and region
¹⁴⁸⁴ growing are not yet ready to face the problem posed by *Sars-COVID19* pneumonia
¹⁴⁸⁵ lungs. Some directions in which future works could start from this thesis are:

- ¹⁴⁸⁶ • The implementation of a *Sars-COVID19* specific segmentation method using
¹⁴⁸⁷ the vast amount of available data
- ¹⁴⁸⁸ • Given the statistical difference found in the survival of patients in the first
¹⁴⁸⁹ and second waves, defined here as before and after 20/07/2020, it might be
¹⁴⁹⁰ very interesting to investigate the causes of this finding and to prospectively
¹⁴⁹¹ continue this analysis with the data from the third wave and eventual next
¹⁴⁹² ones that might occur.
- ¹⁴⁹³ • Eventualmente se esce roba dal clustering

1494 **Chapter 5**

1495 **Appendix**

1496 **5.1 Additional Results and complete tables relative to Random Forest**

1498 Here are the tables with all of the features from random forest. There is no real utility
1499 in providing them for all possible feature combination, hence only the importances
1500 for all features will be given. This will be done for both labels used, i.e. Death and
1501 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.

5.2 Using Dimensionality reduction to further investigate the dataset

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

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

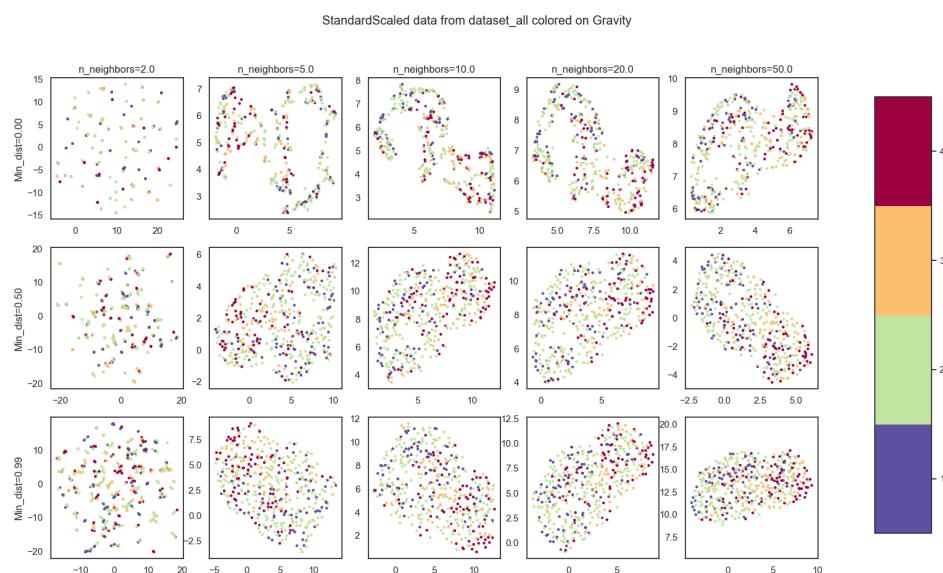


Figure 5.1: Possible combination for umap hyperparameters "number of neighbours" and "minimum distance". Color coding is done with aforementioned gravity score and all the features, i.e. clinical radiomic and radiological, were used.

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

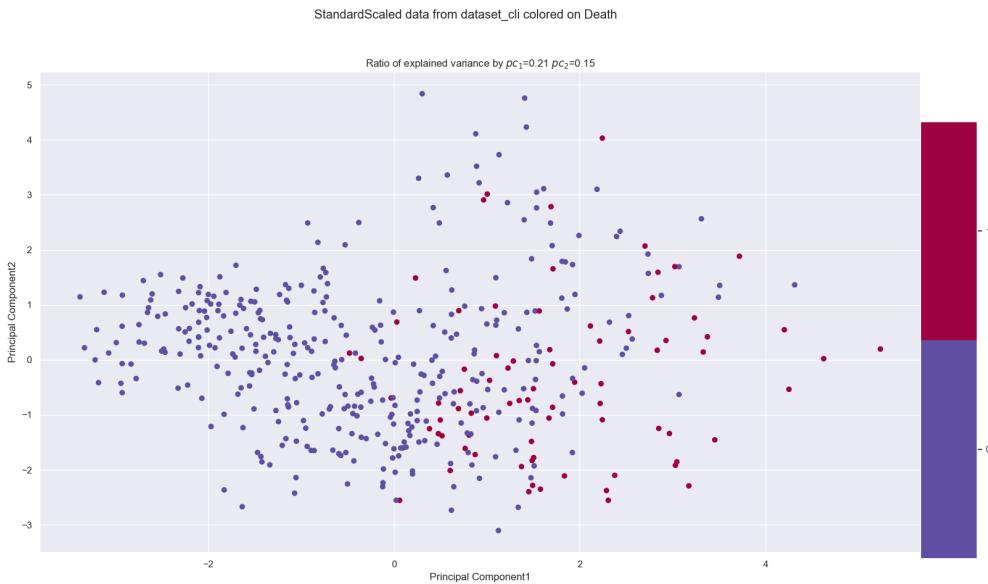


Figure 5.2: 2-Principal Component on clinical features.

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

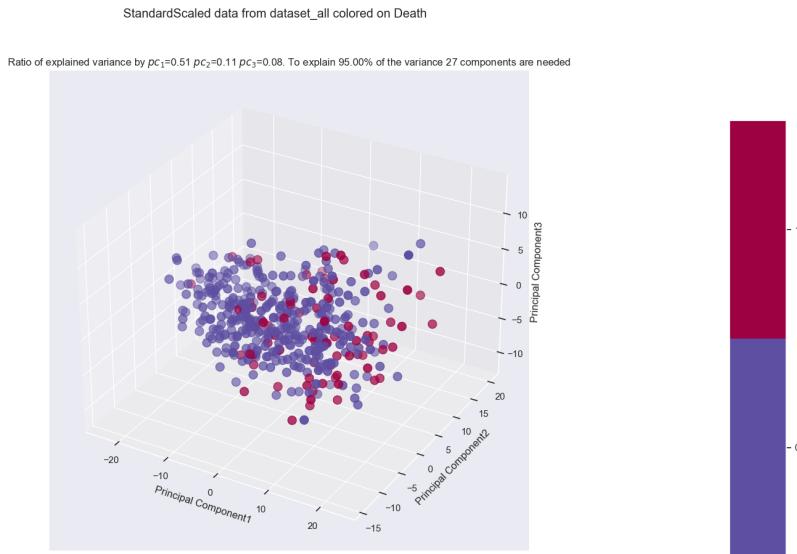


Figure 5.3: 3D PCA of whole dataset, colored with death label

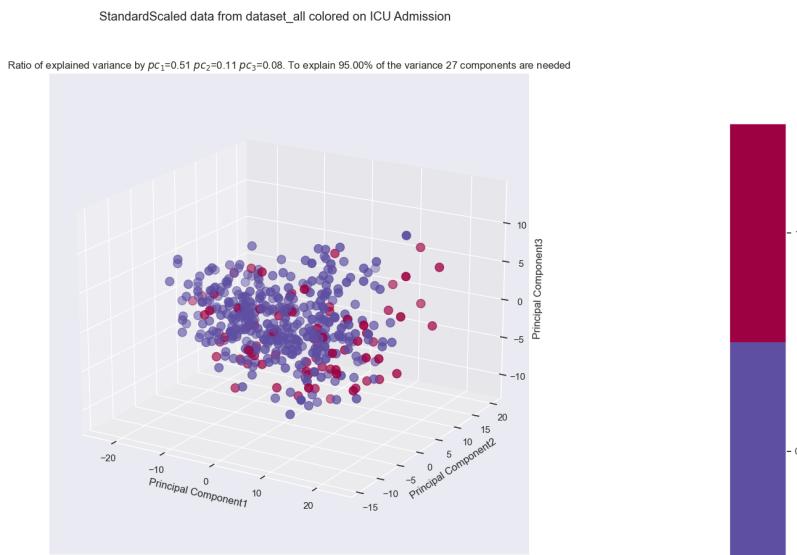


Figure 5.4: 3D PCA of whole dataset, colored with ICU Admission label

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

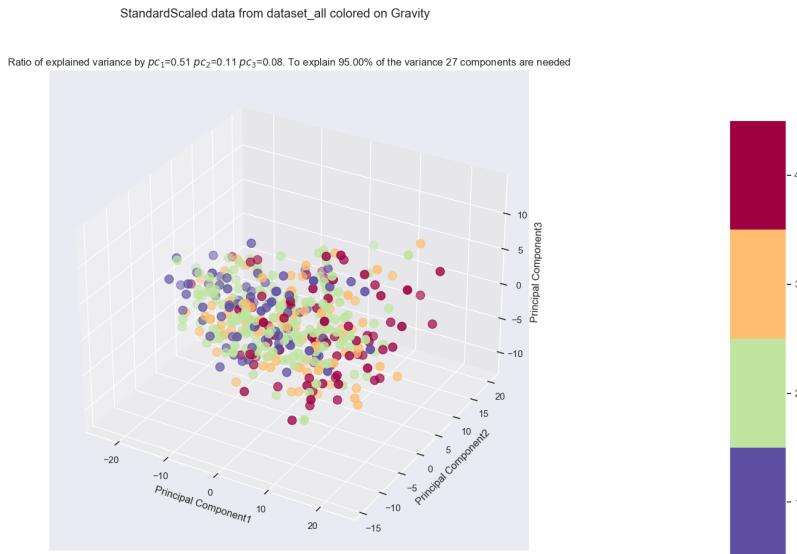


Figure 5.5: Comparison between various colour labels of the top 3 principal components for the entire dataset. Note that to explain 95% of the variance 27 components would be needed

1556 seems to be a well connected structure, which makes sense because umap sets out
 1557 with the objective of preserving said structure. However since the variables of
 1558 interest as label are Death, ICU admission or some kind of combination of them
 1559 with hospital permanence there seems to be no visual correlation between
 1560 structure and label. As such the next dimensionality reduction was tried to see if
 1561 it yielded better results.

1562 5.2.3 Predicting clinical outcome using PLS-DA

1563 Moving on from unsupervised methods to a supervised one, PLS-DA was used
 1564 giving as label both death and ICU using both whole dataset, and singularly
 1565 radiomic or clinical features. Starting from the clinical features alone, predicting
 1566 on death Figure 5.10 can be obtained.

1567 In Figure 5.10 there are a few things to note. The first is the presence, in the top
 1568 plot, of an outlier which, since PLS-DA is based on minimization of least squares,
 1569 can ruin a lot the performance of the procedure. For this reason in the second plot
 1570 the outlier was removed and the algorithm was run again on the cleaned data. The
 1571 second thing to notice is the coloring used which, in the first plot, was used to
 1572 highlight that along the one of the two latent variables the data is roughly
 1573 distributed depending on age while, in the second plot, was used to highlight that
 1574 the algorithm is able to perfectly separate the subjects with hypertension from
 1575 those without it. However, by looking at the same embedding labelled with death
 1576 and ICU admission fig 5.11 can be obtained

1577 It's clear to see that, when predicting on death, the PLS-DA algorithm doesn't
 1578 find any behaviour relevant for ICU admission. It's also clear that there is at least
 1579 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

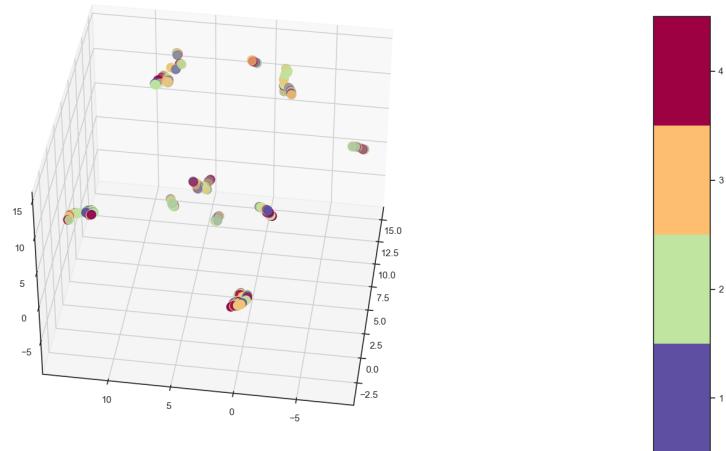


Figure 5.6: 3D umap of clinical dataset, colored based on gravity

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

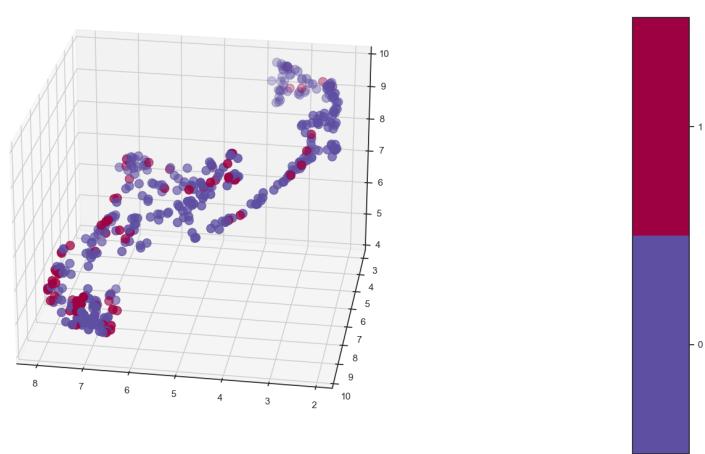


Figure 5.7: 3D umap of whole dataset, colored based on death

StandardScaled data from dataset_all colored on ICU Admission

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

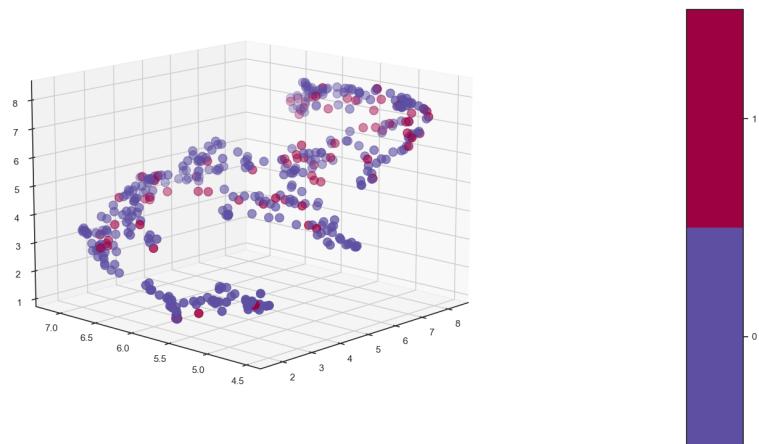


Figure 5.8: 3D umap of whole dataset, colored based on ICU admission

StandardScaled data from dataset_all colored on Gravity

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

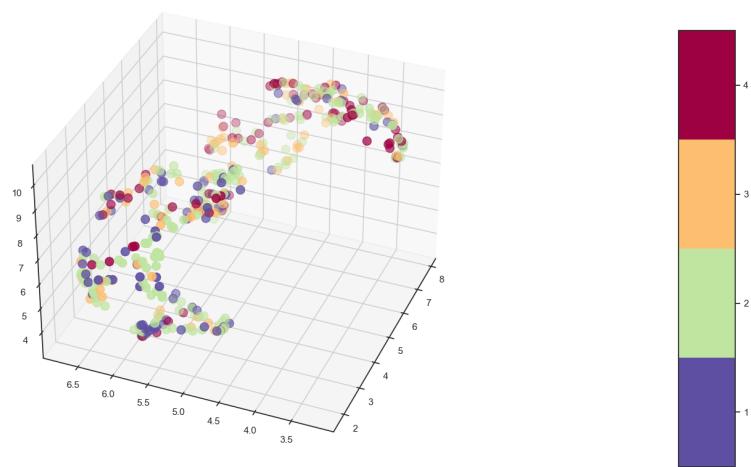


Figure 5.9: 3D umap of whole dataset, colored based on gravity

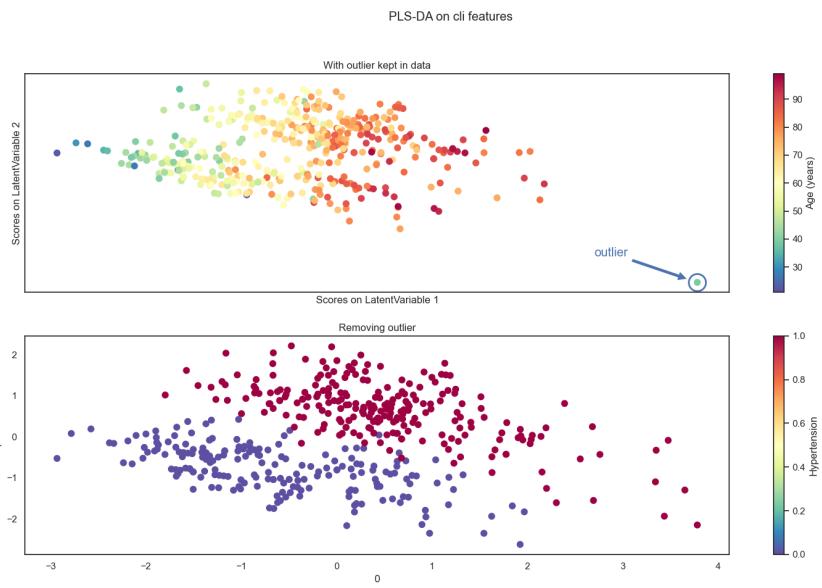


Figure 5.10: PLS-DA predicting on death coloured with age and hypertension

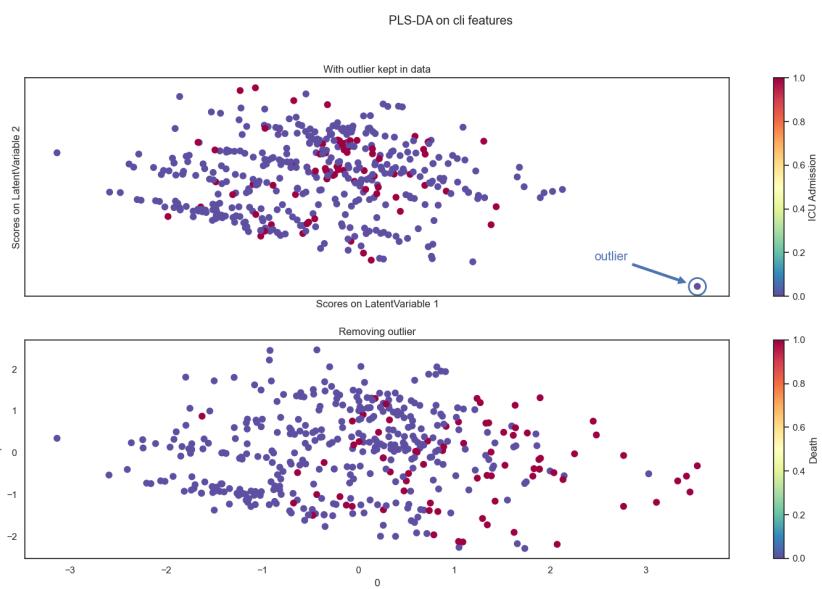


Figure 5.11: PLS-DA predicting on death coloured with death(bottom) and ICU Admission(top)

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

1580 be easily explained by looking at how the ages are distributed in the first plot of
 1581 fig: 5.10. From this it's possible to deduce that older individuals tend to die more
 1582 and that hypertension does not seem to be relevant when considering death as a
 1583 clinical outcome. If necessary the PLS-DA algorithm allows also to see the weights
 1584 given to the features in predicting the label. At least for the clinical dataset, which
 1585 has a reasonable number of features, it's interesting to report it ordering the
 1586 coefficients by descending absolute value:
 1587 Doing the exact same procedure on the whole dataset, which means by including
 1588 the radiomic features, Figure 5.12 can be obtained.

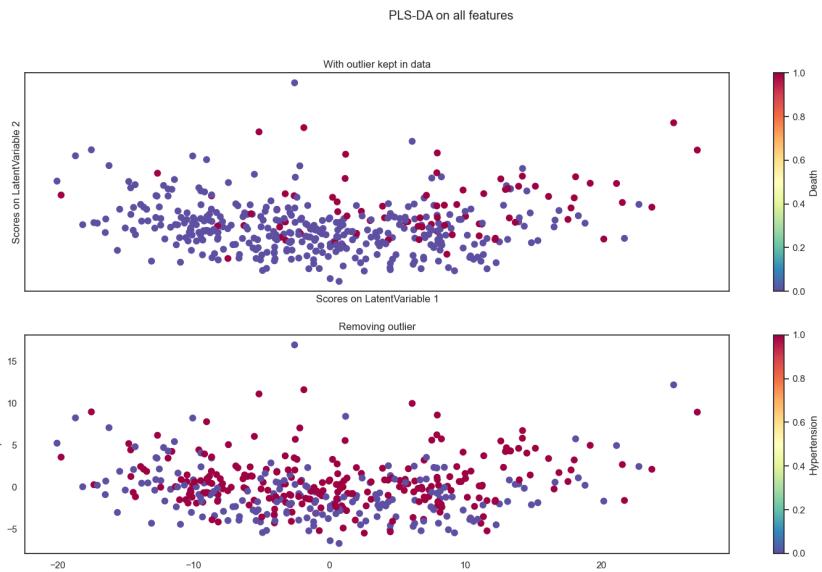


Figure 5.12: PLS-DA predicting on death coloured with death (top) and hypertension (bottom) on whole dataset

1589 Once again adding the radiomic features has evidently introduced noise in the
 1590 system, which no longer displays any kind of behaviour, pattern nor separation.
 1591 Doing the same analysis but using ICU Admission as a label Figure 5.13 can be
 1592 obtained.
 1593 Now the colors have been chosen to highlight that respiratory rate and age have
 1594 the main role in determining the latent variables. However, in this case, there
 1595 doesn't appear to be a clear cut distinction as it happened before with

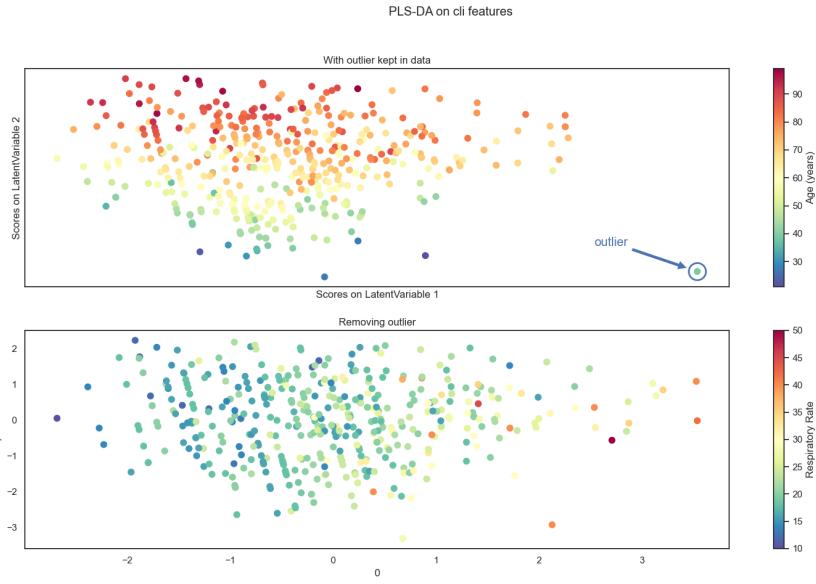


Figure 5.13: PLS-DA predicting on death coloured with Age and respiratory rate on clinical features

1596 hypertension. Looking at how the points scatter by coloring them according to the
 1597 two interesting clinical labels the following figure can be obtained:

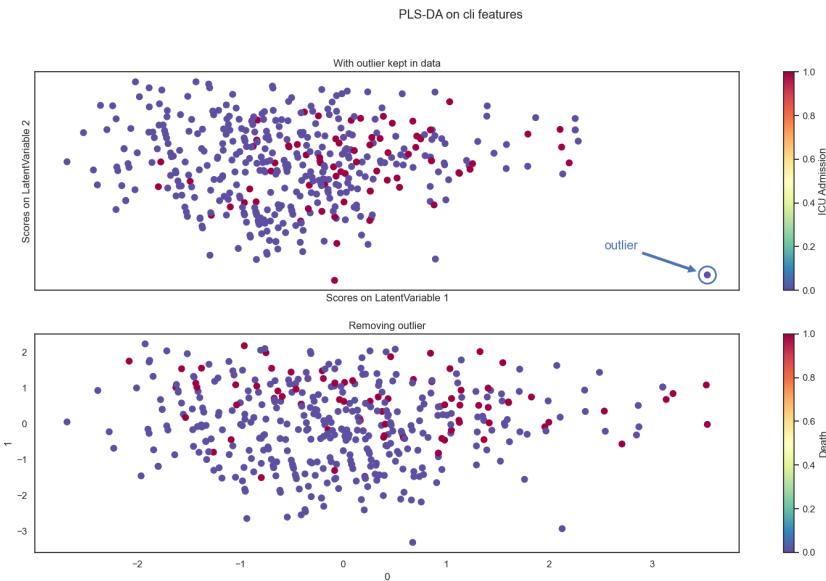


Figure 5.14: PLS-DA predicting on death(bottom) coloured with death and ICU admission(top) on clinical features

1598 Finally, introducing the radiomic features in the analysis the usual effect of
 1599 reducing separation can be seen can be seen in the figure below:

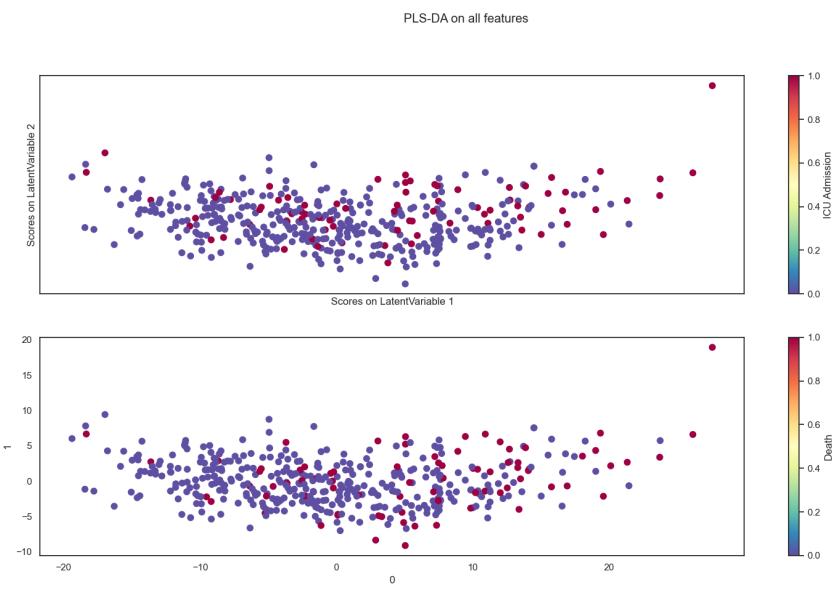


Figure 5.15: PLS-DA predicting on death coloured with death(bottom) and ICU admission(top) on all available features

¹⁶⁰⁰ Bibliography

- ¹⁶⁰¹ [1] Nema ps3 / iso 12052, digital imaging and communications in medicine (di-
¹⁶⁰² com) standard, national electrical manufacturers association, rosslyn, va, usa.
¹⁶⁰³ available free at, 2021.
- ¹⁶⁰⁴ [2] J. L. V. 1, R. Moreno, J. Takala, S. Willatts, A. D. Mendonça, H. Bruining,
¹⁶⁰⁵ C. K. Reinhart, P. M. Suter, and L. G. Thijs. The sofa (sepsis-related organ
¹⁶⁰⁶ failure assessment) score to describe organ dysfunction/failure. on behalf of the
¹⁶⁰⁷ working group on sepsis-related problems of the european society of intensive
¹⁶⁰⁸ care medicine. *Intensive Care Medicine*, 1996.
- ¹⁶⁰⁹ [3] U. Attenberger and G. Langs. How does radiomics actually work? – review.
¹⁶¹⁰ *RöFo - Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden*
¹⁶¹¹ *Verfahren*, 193, 12 2020.
- ¹⁶¹² [4] M. Barker and W. Rayens. Partial least squares for discrimination, journal of
¹⁶¹³ chemometrics. *Journal of Chemometrics*, 17:166 – 173, 03 2003.
- ¹⁶¹⁴ [5] R. W. Brown, Y.-C. N. Cheng, E. M. Haacke, M. R. Thompson, and R. Venkatesan.
¹⁶¹⁵ *Magnetic Resonance Imaging: Physical Principles and Sequence Design*,
¹⁶¹⁶ 2nd Edition. Wiley-Blackwell, 2014.
- ¹⁶¹⁷ [6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote:
¹⁶¹⁸ Synthetic minority over-sampling technique. *Journal of Artificial Intelligence*
¹⁶¹⁹ *Research*, 16:321–357, Jun 2002.
- ¹⁶²⁰ [7] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas
¹⁶²¹ under two or more correlated receiver operating characteristic curves: A non-
¹⁶²² parametric approach. *Biometrics*, 44(3):837–845, 1988.
- ¹⁶²³ [8] K. P. F.R.S. Liii. on lines and planes of closest fit to systems of points in space.
¹⁶²⁴ *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of*
¹⁶²⁵ *Science*, 2(11):559–572, 1901.
- ¹⁶²⁶ [9] D. G. George H. Joblove. Color spaces for computer graphics. *ACM SIG-*
¹⁶²⁷ *GRAPH Computer Graphics*, (12(3)), 20–25), 1978.
- ¹⁶²⁸ [10] L. Guo. Clinical features predicting mortality risk in patients with viral pneu-
¹⁶²⁹ monia: The mulbsta score. *Frontiers in Microbiology*, 2019.
- ¹⁶³⁰ [11] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*.
¹⁶³¹ Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

- 1632 [12] G. N. Hounsfield. Computed medical imaging. nobel lecture. *Journal of Computer Assisted Tomography*, (4(5):665-74), 1980.
- 1633
- 1634 [13] L. M. J. Cine computerized tomography. *The International Journal of Cardiac Imaging*, 1987.
- 1635
- 1636 [14] A. Kaka and M. Slaney. *Principles of Computerized Tomographic Imaging*.
1637 Society of Industrial and Applied Mathematics, 2001.
- 1638 [15] J. Kirby. Mosmeddata: dataset, 09/2021.
- 1639 [16] D. Kleinbaum and M. Klein. *Survival Analysis: A Self-Learning Text*. Statistics
1640 for Biology and Health. Springer New York, 2006.
- 1641 [17] P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. de Jong,
1642 J. van Timmeren, S. Sanduleanu, R. T. H. M. Larue, A. J. G. Even, A. Jochems,
1643 Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt,
1644 A. Dekker, F. M. Mottaghay, J. E. Wildberger, and S. Walsh. Radiomics: the
1645 bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology*, 14(12):749—762, December 2017.
- 1646
- 1647 [18] L. Lee and C.-Y. Liong. Partial least squares-discriminant analysis (pls-da) for
1648 classification of high-dimensional (hd) data: a review of contemporary practice
1649 strategies and knowledge gaps. *The Analyst*, 143, 06 2018.
- 1650 [19] G. Lemaître, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A python
1651 toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.
- 1652
- 1653 [20] J. McCarthy. What is artificial intelligence? 01 2004.
- 1654 [21] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation
1655 and projection for dimension reduction, 2020.
- 1656 [22] S. J. McMahon. The linear quadratic model: usage, interpretation and chal-
1657 lenges. *Physics in medicine and biology*, 2018.
- 1658 [23] S. Morozov. Mosmeddata: Chest ct scans with covid-19 related findings dataset.
1659 *IAU Symp.*, (S227), 2020.
- 1660 [24] MosMed. Mosmeddata: dataset, 28/04/2020.
- 1661 [25] Y. Ohno. Cie fundamentals for color measurements. *International Conference
1662 on Digital Printing Technologies*, 01 2000.
- 1663 [26] D. L. Pham, C. Xu, and J. L. Prince. Current methods in medical image
1664 segmentation. *Annual Review of Biomedical Engineering*, 2(1):315–337, 2000.
1665 PMID: 11701515.
- 1666 [27] P. C. Rajandep Kaur. A review of image compression techniques. *International
1667 Journal of Computer Applications*, 2016.
- 1668 [28] W. C. Roentgen. On a new kind of rays. *Science*, 1986.

- 1669 [29] A. Saltelli. *Global Sensitivity Analysis. The Primer*. John Wiley and Sons, Ltd,
1670 2007.
- 1671 [30] C. P. Subbe. Validation of a modified early warning score in medical admissions.
1672 *QJM: An international journal of medicine*, 2001.
- 1673 [31] L. Tommy. Gray-level invariant haralick texture features. *PLOS ONE*, 2019.
- 1674 [32] S. Webb. The physics of medical imaging. 1988.
- 1675 [33] L. WS, van der Eerden MM, and e. a. Laing R. Defining community acquired
1676 pneumonia severity on presentation to hospital: an international derivation and
1677 validation study. *Thorax* 58(5):377-382, 2003.
- 1678 [34] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. W. L. Aerts, V. Andreadczyk,
1679 A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, and
1680 et al. The image biomarker standardization initiative: Standardized quantitative
1681 radiomics for high-throughput image-based phenotyping. *Radiology*,
1682 295(2):328–338, May 2020.