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

Academic Year 2020-2021

Dedication...

Todo list

Abstract

queste tabelle ovviamente non vanno qui

Contingency table per le segmentazioni sostituibili

	ICU Admission	
	0	1
Death		
0	7	2
1	8	1

Contingency table per le segmentazioni brutte o dubbie

	ICU Admission	
	0	1
Death		
0	32	4
1	11	2

bisogna dividere i pagrafi, 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.

ti consiglio di mettere la parola covid come macro per scriverla sempre in modo consistente

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

Following a first introductiory chapter in the second 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

Chapter 1

Introduction

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

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

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

li sposterei in capitoli successivi

1. **Ground Glass Opacity(GGO)**:
   Small diffused changes in density of the lung structure cause a hazy look in the affected region. This complicates the individuation of pulmonary vessels.

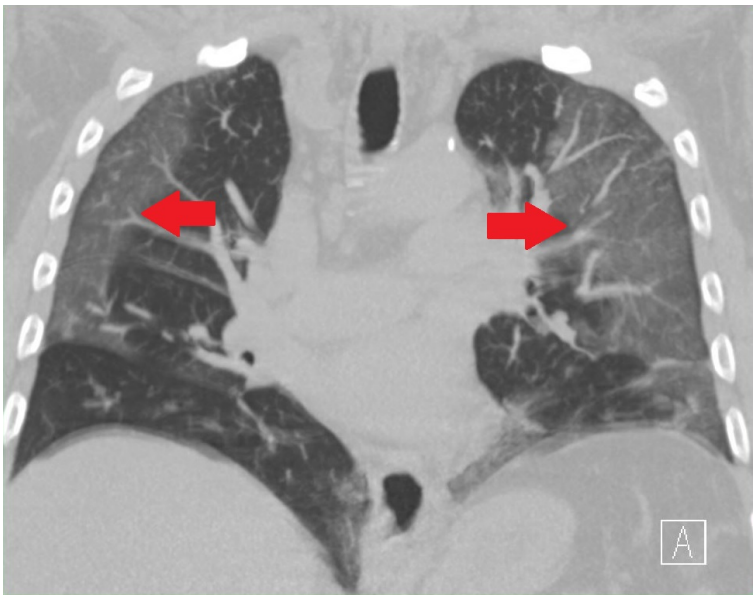
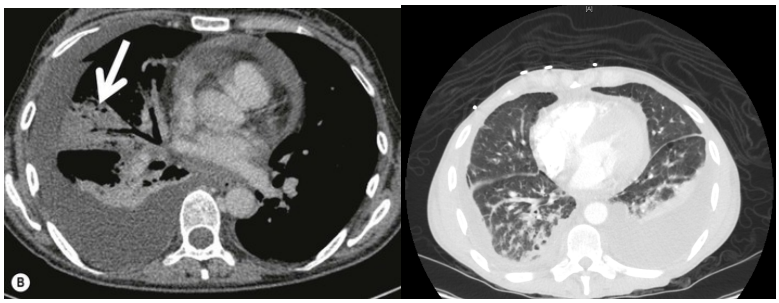


Figure 1.1: Example of GGO

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

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



(a) a **(b)** a

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

4. **Collapsed Lungs and Pleural Effusion**:
   Both of these manifest themself as regions of the lungs that take the same

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

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

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

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

Chapter 2

Materials and methodologies

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

2.1 Theoretical background

questo titolo è superfluo, come abbiamo discusso, questa roba va spostata in introduzione, quelle che chiami sottosezioni dovrebbero essere sezioni

2.1.1 Medical Images

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

Definition 2.1.1 (Digital Image)**.** A numerical representation of an object; more specifically an ordered array of values representing the amount of radiation emitted (or reflected) by the object itself. The values of the array are associated to the intensity of the radiation coming from the physical object; to represent the image these values need to be associated to a scale and then placed on a discrete 2D grid. To store these intensities the physical image is divided into regular rectangular spacings, each of which is called pixel[1], to form a 2D grid; inside every spacing is then stored a number (or set thereof) which measures the intensity of light, or color, coming from the physical space corresponding to that grid-spacing. The term digital refers to the discretization process that inherently happens in storage of the values,

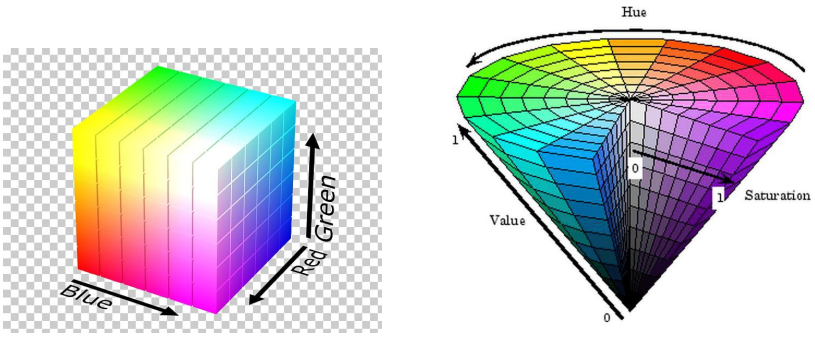
[1]The term pixel seems to originate from a shortening of the expression Picture's (pics=pix) Element(el). The same hold for voxel which stands for Volume Element

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.

CYMK si usa per la stampa (rappresenta lo spettro RGB in sottrazione), non per le immagini scientifiche

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



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

Figure 2.1: Examples of color spaces

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

Definition 2.1.2 (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 [11]

$$HU = 1000 * \frac{\mu - \mu_{H_2O}}{\mu_{H_2O} - \mu_{Air}} \tag{2.1}$$

[2]The three values correspond each to the intensity of a single color, the most commonly used set of colors is the RGB-scale (Red, Green, Blue). Further information can be found by looking into Tristimulus theory[23]

[3]Same as RGB but with four colors, the most common scale is CMYK (Cyan, Magenta, Yellow, blacK)

[4]Besides RGB and CMYK 2.1a the most common color spaces are CIE (Commision Internationale d'Eclairage) and HSV fig:2.1b (Hue,Saturation and Value). Refer to [9] for further details

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

- Spatial Resolution: A measure of how many pixel are in the image or, equivalently, how small each pixel is; a larger resolution implies that smaller details can be seen better fig:2.2. Can be measured as the number of pixel measured over a distance of an inch ppi(Pixel Per Inch) or as number of line pairs that can be distinguished in a mm of image lp/mm (line pair per millimeter).

- Gray-level Resolution: The range of the pixel values, a classic example is an 8-bit resolution which yields 256 levels of gray. A better resolution allows a better distinction of colors within the image fig:2.2.

  di solito si parla di color quantization

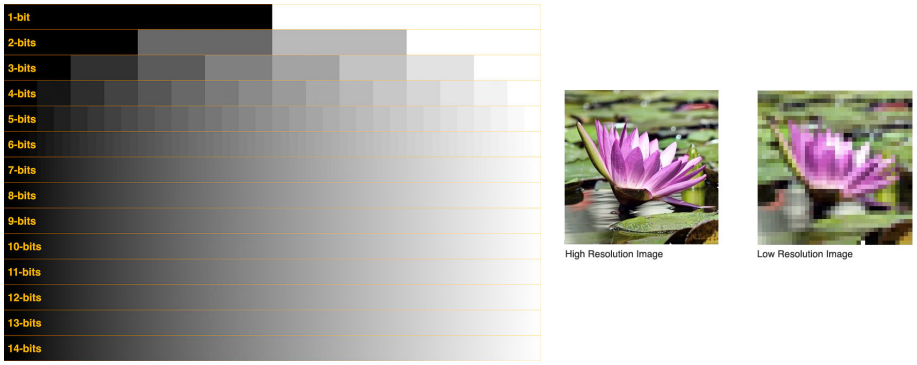


Figure 2.2: Example of visual differences in Gray-level (left) and spatial (right) resolution

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

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

[5]In 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 field[6].

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

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

[6] A detailed description of compression algorithms is beyond the scopes of this thesis, for this reason please refer to [25] for more information

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

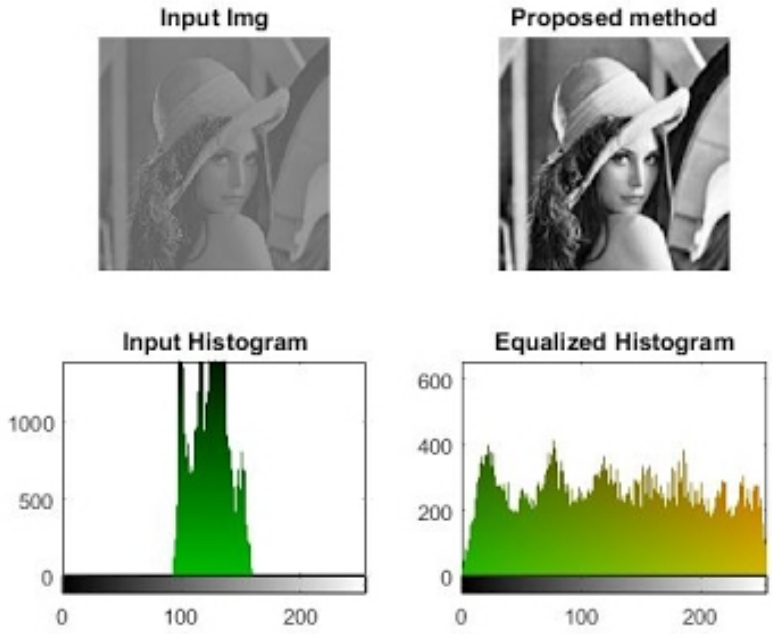


Figure 2.3: Example of differences in contrast due to histogram equalization
cancella la scritta "proposed method"

This leads us to the concept of *Contrast* which is a quantification of how well different intensities can be distinguished. If all the pixel values are bundled in a small range leaving most of the histogram empty then it's difficult to pick up the differences because they are small however, if the histogram has no preferentially populated ranges then the differences in values are being showed in the best possible way fig:2.3. Note also that if looking at the histogram there are two(or more) well separated distributions it's possible that these also identify different objects in the image, which will for example allow for some basic background-foreground distinction.

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

1. Geometric Transformations involve the following steps:

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

   (b) Interpolation: Since the coordinates of the transformed pixel might not fall exactly on the grid it might become necessary to compute a kind of average contribution of the pixel around the destination coordinate

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

to find a most believable value. Examples of such methods are linear, nearest neighbour and bicubic.

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

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

   $$q = p^{\gamma} \tag{2.2}$$

   (b) Local: The output value at specific coordinates depends on a combination of the original values in a neighbourhood around that same coordinates. Some examples are all filtering operation such as edge enhancement, dilation and erosion. These filtering methods are based on performing convolutions in which the output value at each pixel is given bu the sum of pixel-wise multiplication between the starting matrix and a smaller (usually 3x3 or 5x5) matrix called kernel. The output image is obtained by moving the kernel along the starting matrix following a predefined stride, when moving near the borders the behavour is defined by the padding of the image. Stride, kernel shape and padding determine the shape of the output matrix

   un po' di esempi sarebbero utili

   .

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

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

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

perché hai scelto queste specifiche tecniche? a random?

1. Magnetic Resonance Imaging (MRI): This technique is based on the phenomenon of Nuclear Magnetic Resonance(NMR) which is what happens when diamagnetic atoms are placed inside a very strong uniform magnetic field are

subject to Radio Frequency (RF) stimulus. These atoms absorb and re-emit the RF and supposing this behaviour can somehow be encoded with a positional dependence then it's possible to locate the resonant atoms given the response frequency measured. Suffices to say that this encoding is possible however the setup is very complex and the possible images obtainable with this method are very different and can emphasize very different tissue/material properties. Nothing more will be said on the topic since no data obtained with this methodology will be used. More details can be found in [5]

2. Ultra-Sound (US): The images are obtained by sending waves of frequency higher to those audible by humans and recording how they reflect back. This technique is used mainly in imaging soft peripheral tissues and the contrast between tissues is given by their different responses to sound and how they generate echo. The main advantages such as low cost, portability and harmlessness come at the expense of explorable depth, viewable tissues, need for a skilled professional and dependence on patient bodily composition as well as cooperation.

3. Positron Emission Tomography (PET): In this case the images are obtained thanks to the phenomenon of annihilation of particle-antiparticle, specifically of electron-positron pairs. The positrons come from the β^+ decay of a radionucleide bound to a macromolecule, which is preferentially absorbed by the site of interest [9]. Once the annihilation happens a pair of (almost) co-linear photons having (almost) the same energy of 511 keV is emitted, the detection of this pair is what allows the reconstruction of the image representing the pharmaceutical distribution within the body. Once again there are a lot of subtleties that are beyond the scopes of this thesis, suffices to say that: firstly the exam is primarily used in oncology given the greater energy consumption, hence nutrients absorption, of cancerous tissue and secondly this technique can be combined with CT scans to obtain a more detailed representation of the internal environment of the patient

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

X-ray imaging and Computed Tomography (CT)

It's well known that the term x-rays is used to characterize a family electromagnetic radiation defined by their high energy and penetrative properties. Radiation of this kind is created in various processes such as characteristic emission of atoms, also referred to as x-ray fluorescence, and Bremsstrahlung, braking radiation[10]. The discovery that "A new kind of ray"[26] with such properties existed was carried out

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

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

by W.C.Roentgen in 1895, which allowed him to win the first Nobel prize in physics in the same year. Clearly the first imaging techniques that involved this radiation were much simpler than their modern counterpart, first of all they were planar and analog in nature, as well as not as refined in image quality. The first CT image was obtained in 1968 in Atkinson Morley's Hospital in Wimbledon. Tomography indicates a set of techniques[11] that originate as an advancement of planar x-ray imaging; these techniques share most of the physical principles with planar imaging while overcoming some of it's major limitations, main of which being the lack of depth information. X-ray imaging, both planar and tomographic, involves seeing how a beam of photons changes after traversing a target, the process amounts to a kind of average of all the effects occurred over the whole depth travelled.

The way in which slices are obtained is called focal plane tomography and, as the name suggests, the basic idea is to focus in the image only the desired depth leaving the unwanted regions out of focus. This selective focusing can be obtained either by taking geometrical precautions while using analog detectors, such as screen-film cassettes, or by feeding the digital images to reconstruction algorithms to perform digitally the required operations[12].

In both planar and tomographic setting the rough description of the data acquisition process can be summarized as follows: First x-rays are somehow generated by the machine, the quality of these x-rays is optimized with the use of filters then focused and positioned such that they mostly hit the region that needs imaging. The beam then exits the machine and starts interacting with the imaged object[13], this process causes an attenuation in the beam which depends on the materials composing the object itself. Having then travelled across the whole object it interacts with a sensor, be it film, semiconductor or other, which stores the data that will then constitute the final image. In a digital setting this final step has to be performed following a (tomographic) reconstruction algorithm which given a set of 2D projections returns a single 3D image.

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

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

It would also be interesting to talk about reconstruction algorithms however since it's beyond the scopes of this work

lascerei stare la parte in cui dici che è interessante, direi solo che non è in scope

, refer to [13] and [30].

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

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

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

Generation and management of radiation: digital CT scanners

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

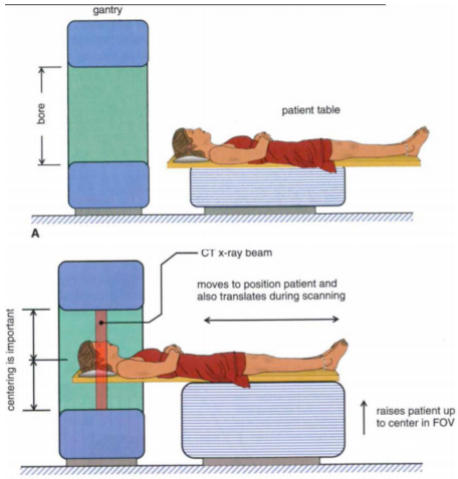


Figure 2.4: General set-up of a CT machine

The beam is generally created by the interaction of high energy particles with some kind of material, so that the particle's kinetic energy can be converted into radiation. In practice this means that an x-ray tube is encapsulated in the machine. Inside this vacuum tube charged particles[14] are emitted from the cathode, accelerated by a voltage differential and shot onto a solid anode[15]. This creation process implies that the spectrum of the produced x-rays is composed of the almost discrete peaks of characteristic emission, due to the atoms composing the target, superimposed with the continuum Bremsstrahlung radiation.

[14] Most commonly electrons

[15] Typical materials can be Tungsten, Molybdenum

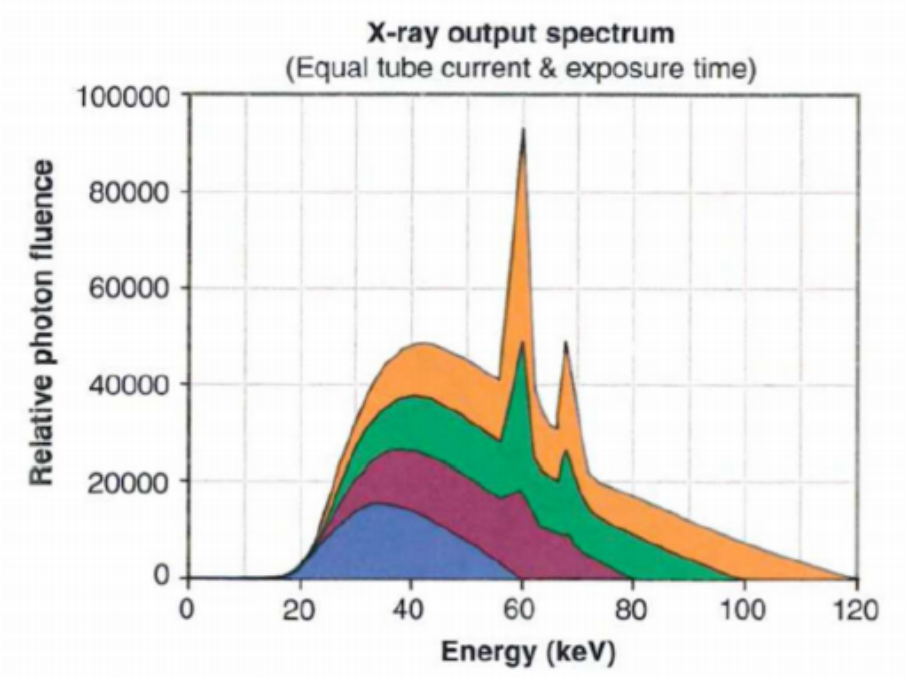


Figure 2.5: X-ray spectrum, composed of characteristic peaks and Bremsstrahlung continuum, computed at various tube voltages

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

Other relevant characteristics in the tube are the anode material, which changes the peaks in the x-ray spectrum and time duration of the emission, which is called exposure time and influences dose as well as exposure[16].

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

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

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

Naturally the description of these technologies can be done on a much finer level, however for the scopes of this work this description is deemed sufficient

non stare sempre a giustificare il livello di dettaglio che hai inserito. ovviamente in tutti i topic si può dire di più.

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

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

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

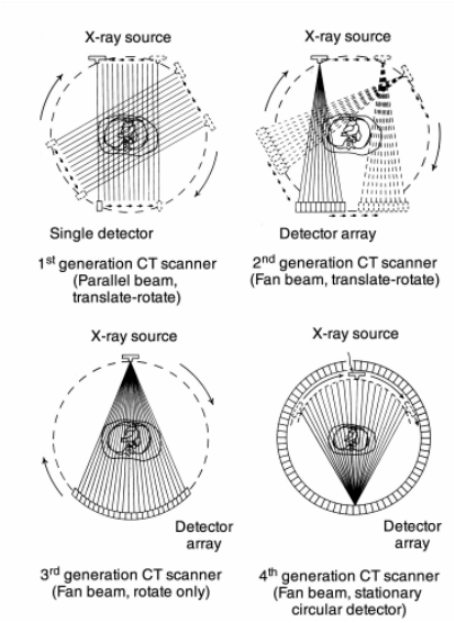


Figure 2.6: First four generation of CT scanners

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

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

- 5^{th} generation-Stationary Stationary Geometry: The x-ray tube is now a large circle that is completely around the patient. This is only used in cardiac tomography and as such will not be described further [12]

- 6^{th} generation-Spiral CT: Supposing the patient is laying parallel to the axis of rotation, all previous generations acquired, along the height of the patient, a single slice at a time. In this generation as the tube rotates around the patients the bed on which they're laying moves along the rotation axis so that the acquisition is continuous and not start-and-stop. This further reduces the acquisition time while significantly complicating the mathematical aspect of the reconstruction. It's necessary to add another important parameter which is the pitch of the detector[17]. 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.

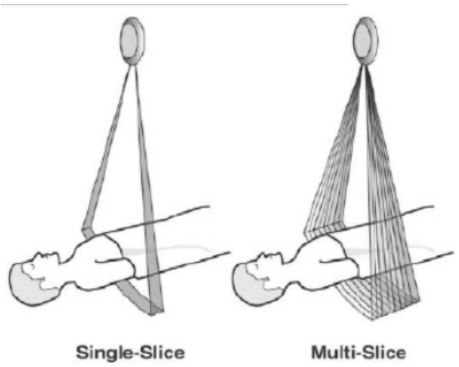


Figure 2.7: 7^{th} *generationsetup*

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

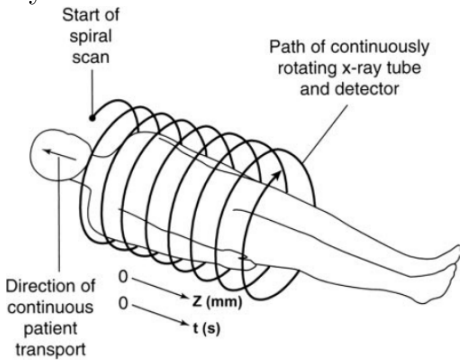


Figure 2.8: 7^{th} *generationsetup*

The machines used to obtain the images used in this thesis, all belonging to the Ospedale S. Orsola, were distributed as follows shown in 2.9:

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

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

Figure 2.9: Acquisition parameter and machine distribution

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.

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

Radiation-matter interaction: Attenuation in body and measurement

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

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

$$N(x) = N_0 e^{-\mu(E)*x} \tag{2.3}$$

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

$$\mu(E) = \frac{\rho * N_A}{A} * (\sigma_{Photoelectric}(E) + \sigma_{Compton}(E) + \sigma_{PairProduction}(E)) \tag{2.4}$$

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

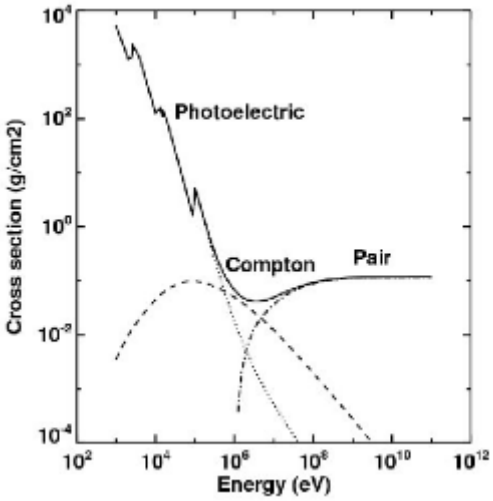
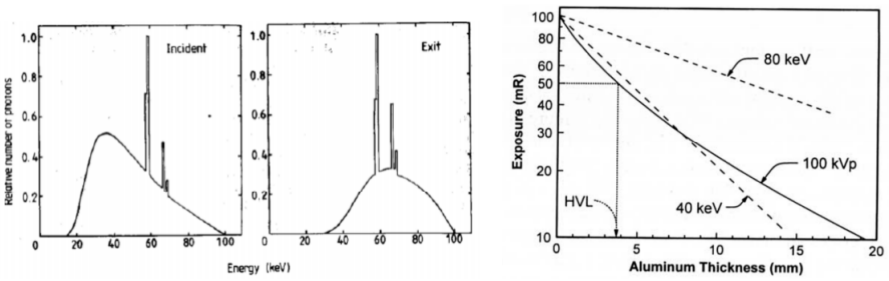


Figure 2.10: Photon cross-section in Pb

Overall, given eq:2.3, it's clear to see that the attenuation behaviour of a monochromatic beam would be linear in semi-logarithmic scale hence the name for μ "*Linear Attenuation Coefficient*". The first complication comes from the fact that, given their generation method, the x-rays are not monochromatic but rather polychromatic. This introduces a further complication which is the phenomenon of beam hardening: lower energy x-rays interact much more likely than those at higher energies which implies that as it crosses some material the mean energy of the whole beam increases. This behaviour is exploited still within the machine, filters are interposed between anode and patient to reduce the useless part of the spectrum as shown in fig: 2.11a:



(a) Beam Filtering

(b) Polychromatic beam attenuation

Figure 2.11: 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: 2.11b. Since the image brightness is related to the number of photons that get on the sensor it's still possible to define the contrast between two pixel p_1, p_2 as:

$$C(p_1, p_2) = \frac{N_{\gamma,p_2} - N_{\gamma,p_1}}{N_{\gamma,p_2}} \tag{2.5}$$

This formula, that connects the beam to the image, together with eq: 2.3, which connects the beam property to the patient's composition, make clear the processes by which the beam carries patient information. Another complication arises in the context of this last equation due to the phenomenon of scattering which reduces the contrast by changing the direction of the bean and introducing an element of noise. Anti-scattering grids are positioned right before the sensors to reduce this effect by allowing to reach the sensor to only the photons with the correct direction.

The biological effects of radiation won't be treated in this thesis. Suffices to say that damage can be classified as primary, due to ionization events within the nucleus of the cell, or secondary, due to chemical changes in the cell environment. The energy deposited per unit mass is called dose and is measured in Gy(Gray) and, as said before, depends on exposure time, current and kVP of the tube. Most of contemporary machines for CT self-regulate exposure time during the acquisition automatically using Automatic Exposure Control(AEC). Having the dose it's possible to estimate the fraction of surviving cells and, to do so, various models are used. In the clinical practice it's common to find, still within the DICOM image metadata, the information regarding Dose delivered such as CTDI (Computed Tomography Dose Index) from which it's possible to obtain the DLP (Dose Length Product) taking into consideration the total length of irradiated body. For an introduction to one of these models, the Linear Quadratic (LQ) refer to [20].

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

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

Starting from the definition given by John McCarthy in [18] "[AI] is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.". Machine Learning (ML) is a sub-branch of AI and contains all techniques that make the computer improve performances via experience in the form of exposure to data, practically speaking this finds it's application in classification problems, image/speech/pattern recognition, clustering, autoencoding and others. The general workflow of Machine Learning is the following: given a dataset, the objective is to define a model or function which depends on some parameters which is able to manipulate the data in order to obtain as output something that can be

evaluated via a predefined performance metric. The parameters of the model are then automatically adjusted in steps to minimize or maximize this performance metric until a stable point at which the model with the current parameters is considered finalized; one of the main problems in this procedure is being sure that the stable point found is global and not local. The whole procedure is carried out keeping in mind that the resulting model needs to be able to generalize it's performance on data that it has never seen before, for this reason usually ML is divided in a training phase and a testing phase. The training phase involves looking at the data and improving the performance of the model on a specific dataset[18], the testing phase involves using brand new data to evaluate the performance of the model obtained in the preceding phase. Machine Learning techniques can be further grouped into the following categories:

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

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

3. Reinforcement Learning: This kind of ML is well suited for data which has a clear sequential structure in which the requiered task is to develop good long term planning. Broadly speaking the general set-up is that given a set of (state_t, action, reward, state_{t+1}) these techniques try to maximize the cumulative reward[19]. The main applications of these techniques are in Autonomous driving and learning how to play games

The following methods are those directly involved in this work.

Regression, Classification and Penalization

Regression and Classification are methods used to make predictions via supervised learning by understanding the input-output relation in continuous and discrete cases respectively. The most basic example of regression is linear regression which consists

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

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

in finding the slope and intercept of a line passing through a set of points. A branch of classification thats similar to linear regression is logistic regression, this translates in finding out wether or not each point belongs to a certain category given it's properties and can be practically thought of, for a binary classification,as a fitting procedure such that the output can be either 0 for one category and 1 for the other, this is usually done using the logistic funtion, also called sigmoid, which compresses $\mathbb{R}$ in [0,L] as seen in 2.12. L represents the maximum value desidered, x_0 is the midpoint and k is the steepness. Note that for multiple categories it would conceptually suffice to add sigmoids with different centers to obtain a step function in which each step corresponds to a category, in this case the procedure is usually called multinomial regression.

$$\sigma_{(x)} = \frac{L}{1 + e^{-k(x-x_0)}} \tag{2.6}$$

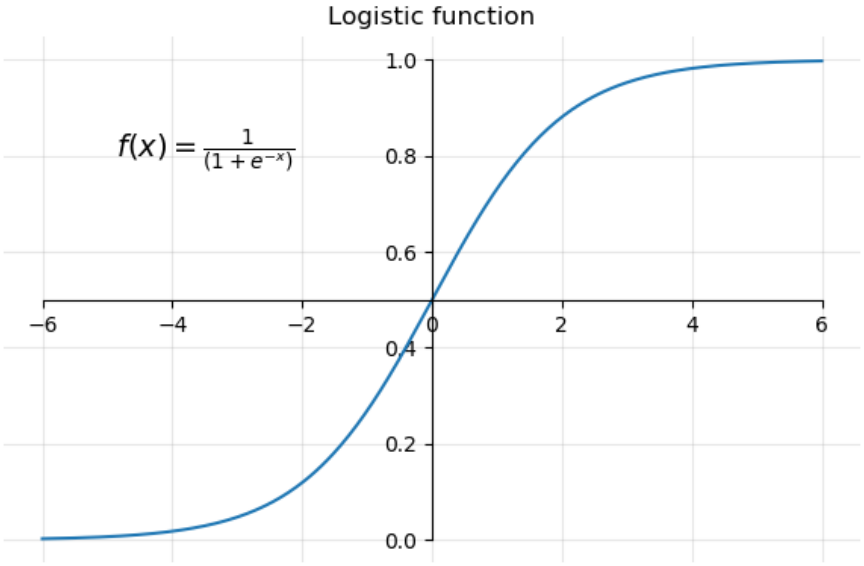


Figure 2.12: sigmoid function with L=1, x_0=0, k=1

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

When each datapoint is characterized by a lot of different features the procedure is called multiple linear regression, the task becomes finding out how much each feature contributes in predicting the output within a weighted linear combinaion of features. Practically speaking this can be done in matricial form, supposing that each of the m datapoints has n features associated let $\mathbf{X}$ be a matrix with n+1[20]

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

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

$$y_i = \Sigma_{j=0}^{n+1} x_{i,j} * \theta_i + \epsilon_i => \mathbf{Y} = X * \theta + \epsilon \tag{2.7}$$

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

$$J_\theta = (\mathbf{Y} - X * \theta)^T * (\mathbf{Y} - X * \theta) \tag{2.8}$$

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

$$\theta^* = (\mathbf{X}^T * \mathbf{X})^{-1} * \mathbf{X}^T \mathbf{Y} \tag{2.9}$$

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

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

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

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

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.

Decision Trees and Random Forest

Apart from logistic and multinomial regression there are various other supervised classifiying methods such as Support Vector Machines (SVM), Neural Networks and Decision Trees. In this thesis, among the aforementioned algorithms, the chosen one was a particular evolution of DecisionTrees called RandomForest (RF). To provide some insight in the method it's necessary to first explain how decision trees work, specifying what problems they face and what are their strong points. Since it's a classification method let's consider the simple case of binary classification with categorical[21] features. The task of the decision tree is to approximate to the best of it's abilities the labels contained in the training set using all the features associated with each datapoint, this is done by building a graph-like structure in which the nodes represent the features and the links departing from 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 unitl a node in which the separation of the preceding feature is better then that provided by all remaining features or all features have been considered. The first node is called root node, while the nodes that have no branches going out of them are called leaves, the graph represents a tree hence the name of the method. Note that at every node only the subset of data corresponding to all previous feature categories is used. At each node the separation between the two label categories c_1, c_2 due to the feature is commonly measured with Gini impurity coefficient, which is computed as:

$$\begin{aligned} G &= 1 - p_1^2 - p_2^2 \\ &= 1 - \left[\frac{N_{c_1 \in node}}{N_{samples \in node}}\right]^2 - \left[\frac{N_{c_2 \in node}}{N_{samples \in node}}\right]^2 \end{aligned} \tag{2.10}$$

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

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

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

non farti mai scrupoli a citare testi, le citazioni non sono mai troppe

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

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

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

Synthetic Minority Oversampling TEchnique (SMOTE)

This technique considers a user-defined number of nearest neighbours of randomly chosen points in the minority class and populates the feature space by generating samples on the lines that connect the chosen sample with a random neighbour[22].

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

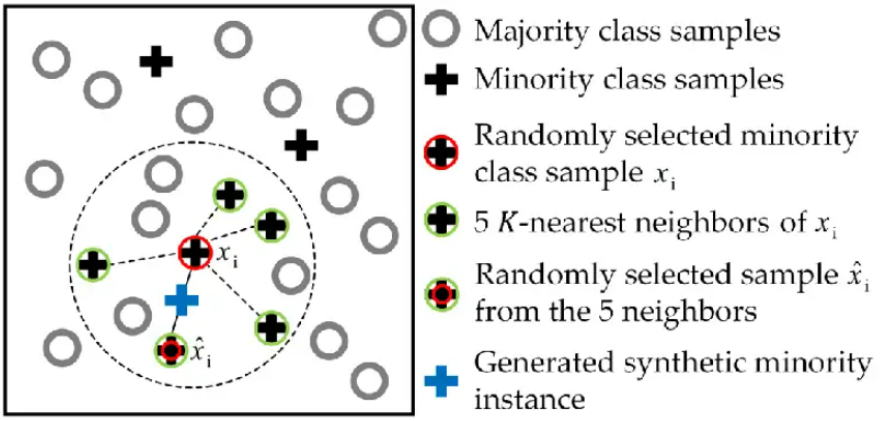
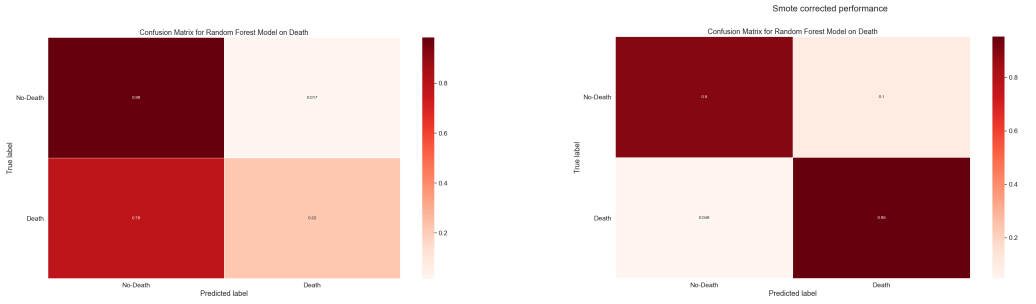


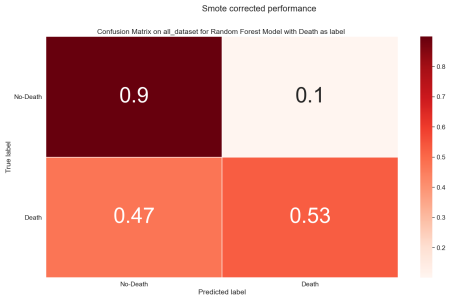
Figure 2.13: Example of SMOTE with 5 nearest neighbours

Worth noting that this has been done using the library imblearn in python [17]. To preview some of the data, looking at the performance of a vanilla random forest implementation with all preset parameters, the effect of the position of oversamplling is the following.



(a) without smote

(b) SMOTE-before train-test split



(c) SMOTE after train-test split

Figure 2.14: 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 contiains information regarding the training data and viceversa. In practice, especially because the points are created as convex combination of existing data[23], when the synthetically generated points end up in the testing they depend, at least partially, on the data used in the training procedure.

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

1. **Unsupervised methods**

    - Principal Component Analysis (PCA): Linearly combines the pre-existing features to obtain new ones and orders them by decreasing ability to explain total variance of data. The first principal component is the combination of features that most explains the variance in the original dataset. Given it's nature it focuses on the global characteristics of the dataset.
    - t-distributed Stochastic Neighbor Embedding (t-SNE): Keeps a mixture of local and global information by using a distance metric in the full high dimensional space and trying to reproduce the distances measured in the lower dimensional space which can be 2- or 3-dimensional. NON SO SE VA SPIEGATO MEGLIO O PIU' IN DETTAGLIO
      Let's define similarity between two points as the value taken by a normal distribution in a point away from the centre, corresponding to the first point of interest, the same distance as the two points taken in consideration. Fixing the first point the similarities with all other points can be computed and normalized. Doing this procedure for all points it's poissible to build a normalized similarity matrix. Then the whole dataset

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

is projected in the desired space, usually $\mathbb{R}^i$ with i=1,2 or 3, in which a second similarity matrix is built using a t-distribution instead of a normal distribution [24]. At this point, in iterative manner, the points are moved in small steps in directions such that the second matrix becomes more similar to the one computed in the full space.

2. **Supervised methods**

   - Partial Least Squared Discriminant Analysis (PLS-DA): This technique is the classification version of Partial Least Squares (PLS) regression. Much like PCA the idea is to find a set of orthonormal vectors as linear combination of the original features in the dataset, however in PLS and PLS-DA is necessary to add the constraint that the new component, besides being perpendicular to all previous ones, explains the most variability in a given target variable, or set thereof. For an in depth description refer to [4] while for a more modern review refer to [16] 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 space[25], 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 [19] 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

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

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

their distance from the closest center. The main technique in this category is k-means clustering, and some, if not most, other techniques include it as step in the processing pipeline[26]. The main problems are firstly that these techniques require prior knowledge, or at the very least a good intuition, on the number of clusters in the dataset while also assuming that the clusters are distributed in spherical gaussian distribution, which is not always the case.

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

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

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

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

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

2.1.3 Combining radiological images with AI: Image segmentation and Radiomics

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

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

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

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

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

Image Segmentation

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

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

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

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

On a practical level these techniques can be used in various fields, as illustrated by the variety of aforementioned tasks, but the one that interests this work the most is the medical field. Nowadays in medicine, where most of imaging exams are stored in digital form, the ability to automatically discern specific structures within

the images can provide a way to aid clinical professionals in their everyday decision making their workload lighter and helping in otherwise difficult cases. A staggering example connected to this thesis is the process of organ segmentation in CT scans: usually these scans are n[27] stacked images in 512x512 resolution. To have a contour of the lungs in a chest CT a radiologist would need to draw by hand the contour of the lung in each slice of the scan. Even if some shorcuts exist to reduce the number of slices to draw on this very boring, time consuming and repetitive task can occupy hours if not days of work to a human while a machine can take minutes to complete a whole scan. Then considering lesion detection in medical exams having a machine that consistently finds lesions that would otherwise be difficult to discern by a human eye can be of paramount importance in diagnosis as well as treatment. In the medical field the difficulty comes from the fact that different exams have different types of image formats which means that an algorithm that works well on CT may not work as intended on MRI or other procedures. Automatic image segmentation can be performed in various ways:

1. Artificial Neural Network (ANN): These techniques belong to a sub-field in ML called Deep Learning, they involve building network structures with more layers in which each node is a processing unit that takes a combination of the input data and gives an output according to a certain activation function. These structures are called Neural Networks because they resemble and are modeled after the workings of neuron-dendrite structures while the deep in deep learning refers to the fact that various layers composed of various neurons are stacked one after the other to complete the structure. Learning is obtained by changing how each neuron combines the inputs it recieves. In the case of images these structures are called Convolutional Neural Networks because the first layers, which are intended to extract the latent features or structures in the images, perform convolution operation in which the parameters are the values pixel values of convolutional kernels

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

3. Deformable models and Region Growing: Both these technique involve setting a starting seed within the image, in the first case the seed is a closed surface which is deformed by forces bound to the region of interest, such as the desired edge. In the second case the seed is a single point within the region of interest, step by step more points are added to a set which started as the seed alone according to a similarity rule or a predefined criteria.

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

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

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

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

For a more in depth review of the main methods used in medical image segmentation refer to [24]. Worth noting, at this point, that the semi-automatic segmentation software used in this work uses probably a mixture of region growing and thresholding methods, maybe guided by an atlas. The segmentation process generally produces a boolean mask which can be used to select, using pixel-wise multiplication, the region of interest in the image. The next step in image analysis would be to derive information from the region defined during segmentation, in general this step is called feature extraction[28] and it's objective is to find non-redundant quantities that meaningfully summarize as much properties of the original data as possible.

Radiomics

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

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

2. First order features: These features depend strictly on the gray levels within the region of interest since they evaluate the distribution of these values, as

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

such they need that the boolean mask of the segmentation be multiplied pixel-wise with the origian image to obtain a new image with only the interesting part in it. Most of these features are commonly used quantities, such as Energy, Entropy, Minimum and Maximum value which have been adapted to the imaging context using the histogram of the original image or by considering intensities within an enclosed region.

3. Higher order features: All the other fatures fall in this macro-category which can be clearly subdivided in other smaller catgories in which the features are obtained following the same guiding principle or starting point. Generally these describe more texture-like properties of the image and, to do so, use particular matrices derived from the original image which contain specific information regarding order and relationships in pixel value positioning within the image. These matrices have very precise definition, as such only the general idea behind them will be reported here redirecting to [32] for a more strict and in detail description. The matrices from which the features are computed also give name to the smaller categories in this family, these categories are:

   - Gray Level Co-occurence Matrix (GLCM) features: This matrix expresses how combination of pixel values are distributed in a 2D or 3D region by considering connected all neighbouring pixel in a certain direction with respects to the one in consideration. Using all possible directions for a set distance, usually δ=1 or δ=2, various matrices are obtained and from these a probability distribution can be built and evaluated. It should be noted that before computing these matrices the intensities in the image are discretized.
   - Gray Level Run Length Matrix (GLRLM) features: Much like before the task of these features is to quantify the distribution of relative values in gray levels trhoughout the image, as the name suggests what this matrix quantifies is how long a path can be built by connecting pixel of the same value along a single direction. This time information from the matrices computed by considering different directions are aggregated in different ways to improve rotational invariance of the final features
   - Gray Level Size Zone Matrix (GLSZM) features: This matrix counts the number of zones in which voxel have the same discretized gray level. The zones are defined by a notion of connectedness most commonly first neighbouring voxel are considered as connectable if they have the same value, this leads to a 26 neighbouring voxel in 3 dimensions and to 8 connected pixels in 2 dimensions[29]. 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.

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

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

In talking about the previous feature groups the concept of discretization of data which is already digital, and hence a discretized, has emerged. This is often a required step to make the computations of the matrices tractable and consists in further binning together the pixel values, which is commonly done in two main ways: either the number of bins is fixed or the width of the bins is fixed preceeding the discretization process. It should be evident that the results of the feature extraction procedure is heavily dependent on the choices made in all the steps that preceed it, main of which being the segmentation, the eventual re-discretization of the image leading to the algorithm used to compute the feature themselves. For this reason recently the International Biomarker Standardization Initiative (IBSI [32]) wrote a *"reference manual"* which details in depth the definitons of the features, description of data-processing procedures as well as a set of guidelines for reporting results. This was done in an attempt to reduce as much as possible the variability and lack of reproducibility of radiomic studies. In the past the main attention in radiological research was focused on improving machine performances and evaluating acquisition sequence technologies, however the great developements in artificial intelligence and performance of computers have brought a lot of attention to the field. Various papers, such as [15] and [3] have been written with the objective of presenting the general workflow. By design the topics in this thesis have been presented to resemble the general order of the pipeline, which can be summarised as:

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

Another interesting possibility offered by the biomarkers computed following radiomics is the ability to quatify the differences between successive exams of the same patient. This specific branch of radiomics is called Delta-radiomics, referencing to the time differential that become the main focus of the analysis. NON SAPREI SE VA BENE COSI' O BISOGNA AGGIUNGERE ALTRI DETTAGLI

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

2.2 Data and objective

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

- All patients that had undergone a CT exam which was retrievable via the PACS (Picture Archiving and Communication System) of *Azienda ospedaliero-universitaria di Bologna - Policlinico Sant'Orsola-Malpighi*
- All patients that had a all of the clinical and laboratory features, listed in fig:2.15, suggested by Lidia Strigari

  questa nota sembra abbastanza strana...

  30

- Since all patient had at least 2 CT exams only the closest date to the hospital admission date was taken. When more exams were performed on the same date all of them were initially taken. At first only chest or abdomen CTs were taken regardless of the acquisition protocol used.

[30] She is, as of the writing of this thesis, the head of the Medical Physics department in the S. Orsola hospital. These suggestions were given to her by clinical professionals.

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
qSOFA	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-intesive care unit admission	336	77%	0	MulBSTA score total	70	16%	13
SOFA score	7	2%	6	Sub-intesive 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
CURB65	30	7%	3	Sex	284	65%	Male	Lung consolidation	225	52%	1
CURB65	4	1%	4	Sex	1	0%		Hypertension	194	45%	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 2.15: Description of clinical label dataset, in sets of four columns there's the clinical feature, the total count of occurences, 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, for those that were obscure to me as an outsider and not otherwise explained in **??**, I'm going to provide a very simple explanation:

non mi piace la forma, troppo personale

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

(a) MulBSTA: This score acconts for **Mul**tilobe lung involvement, absolute **L**ymphocyte count, **B**acterial coinfection, history of **S**moking, history of hyper**T**ension and **A**ge over 60 yrs. [10]

(b) MEWS: Modified Early Warning Score for clinical deterioration. Computed considering systolic blood pressure, heart rate, respiratory rate, temperature and AVPU(Alert Voice Pain Unresponsive) score. [28]

(c) CURB65: **C**onfusion, blood **U**rea Nitrogen or Urea level, **R**espiratory Rate, **B**lood pressure, age over **65** years. This score is specific for pneumonia severity [31]

(d) SOFA: **S**equential **O**rgan **F**ailure **A**ssessment score. Considers various quantities from all systems to assess the overall state of the patient, PaO_2/FiO_2[31] for respiratory system, Glasgow Coma scale[32] for nervous, mean pressure for cardiovascular, Bilirubin levels for liver, platelets for coagulation and creatine for kidneys [2]

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

This procedure produced a starting cohort of ∼700 patients which, having all various images available, created a huge set of ∼2200 CT scans. Since this analysis is focused on radiomics there is an evident need for as much consistency as possible in the images analysed. For this reason all CTs taken with medium of contrast were excluded, since they would have brightnesses not indicative of the disease, and for every patient only images with thin slice reconstruction were considered. More specifically only images with slice thickness of 1 o 1.25 mm[33] along the z-axis were taken into consideration, which meant excluding all the 1.5,2,2.5 and 5 mm slice thicknesses. Since all these images were segmented by me and two other students using a semi-automatic segmentation tool provided by the hospital it sometimes happened that manual corrections were necessary, these were performed only in very obvious and simple cases while, in all remaining cases, the patient was dropped out of the study. Overall this left the final study cohort to be composed of 435 patients, all descriptions and analyses are related to this cohort. The same software used for segmentation allowed the extraction of the radiomic features form the segmented volumes, even

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

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

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

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 one were the aforementioned scores and the boolean ones were sex of the patient, obesity status, if the patient had a fever[34] as of hospital admission and whether or not the patient suffered of Hypertension
   - The remaining features, namely those in 2.2 as well as the death status of the patient, were wither 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 [32] with the addition of fat and muscle surface, computed as cm^2 by counting pixel identified via threshold as fat or muscle tissue in thoracic slices taken at height of vertebra T-12

3. Radiological: These features are those that can be derived from CT exams by humas. Namely acquisition parameters, such as KVP and Current, were used to search for eventual correlations between image quality and predictive power of the feature derived from the image while boolean features, such as Bilaterality of lung damage, presence of Groun Glass Opacities (GGO), lung consolidations as well as crazy paving were used to see if they were sufficient in determining outcome.

[34] Defined as body temperature>38°

2.3 Preprocessing and data analysis

inserirei un flowchart della procedura, aiuta a capire

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

The quatile-quantile plots have been used for a visual check of normality, normally distributed data will populate the bisector of the graph while deviations are symptoms of non-normality. Heavy and light tails are visible as deviations respectively above and belox the bisector. It should be noted that when the data is normally distributed then the transform doesn't change much the distribution while, in cases with more heavy tailed distributions, the improvement is clear to see as it can be seen in Figures 2.16 and 2.17 The next preprocessing step has been to apply a *StandardScaler* to all of the features, which corresponds to subtracting the mean and dividing by the standard deviation, in order to center all the features around zero. The final step in preprocessing has been to reduce the features by using a correlation threshold which means that all variables that correlate with another more, in absolute value, than a certain threshold, which has been set to 0.6 in this work, are dropped a priori. A rather important thing to notice is that correlation has been computed using Spearman correlation and not Pearson since the first is invariant under monotone transformations, such as boxcox, and the second is not. Given the large number of features it's very plausible that at least one of the eliminated features is correlated with all other dropped features but with none of the remaing ones, since a Lasso regularization will be used introducing a few redundant features is not too damaging and the possible benefits outweight the risks. For this reason a redrawing method has been implemented to add one of the dropped features, this has been done by choosing the one that most correlates with the label being used. A pivotal point in all of this analysis is that, to obtain reasonable values in the cross-validation procedures and to avoid leakage[35] problems, all of the preprocessing steps have been done after train-test splitting the data on the train set and then applied as defined during training on the test dataset. When it comes to cross-validation procedure the choice was made to use a stratified k-fold approach with k=10, the data is split in 10 parts with the same percentages of labels[36] then a model is built by training on 9 of the folds and it's performance is then tested on the remain-

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

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

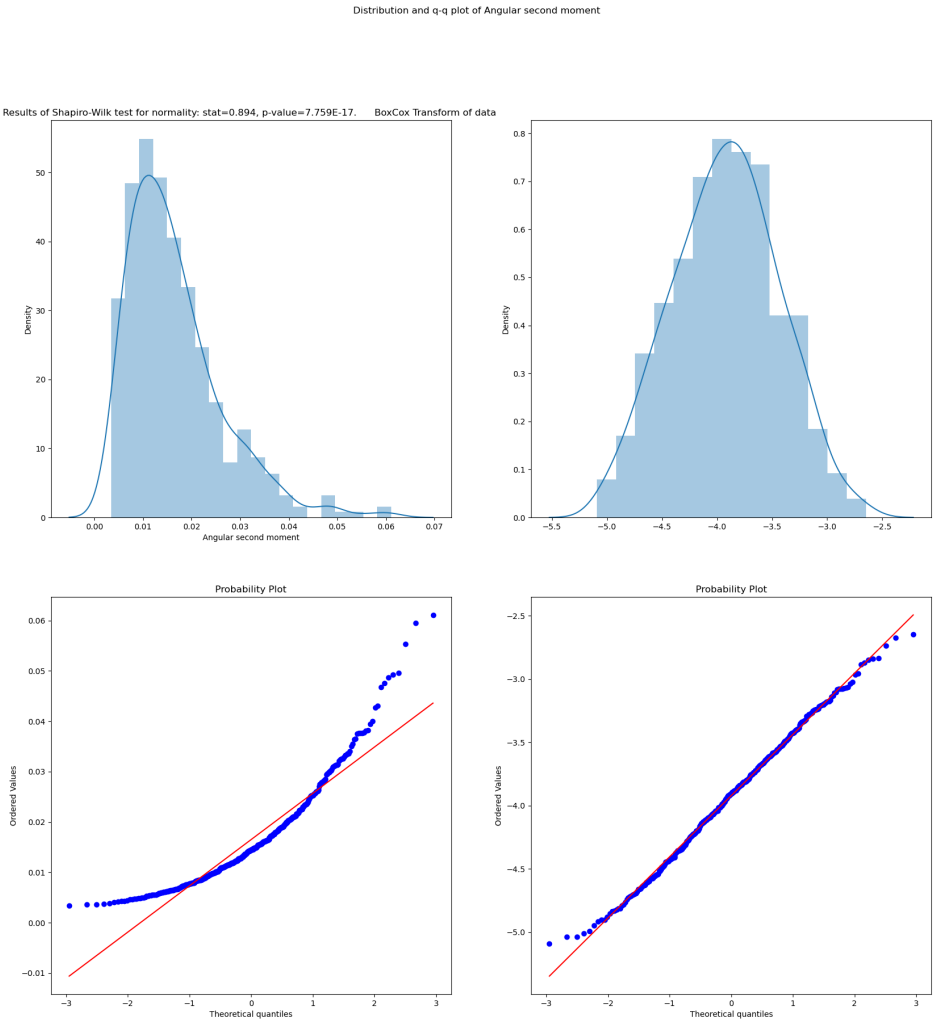


Figure 2.16: Example of boxcox applied to the radiomic feature *Angular second moment* which has a heavy tailed distribution. The graphs contain the original distribution (top-left) the transformed distribution (top-right) and the two respective quantile-quantile plots

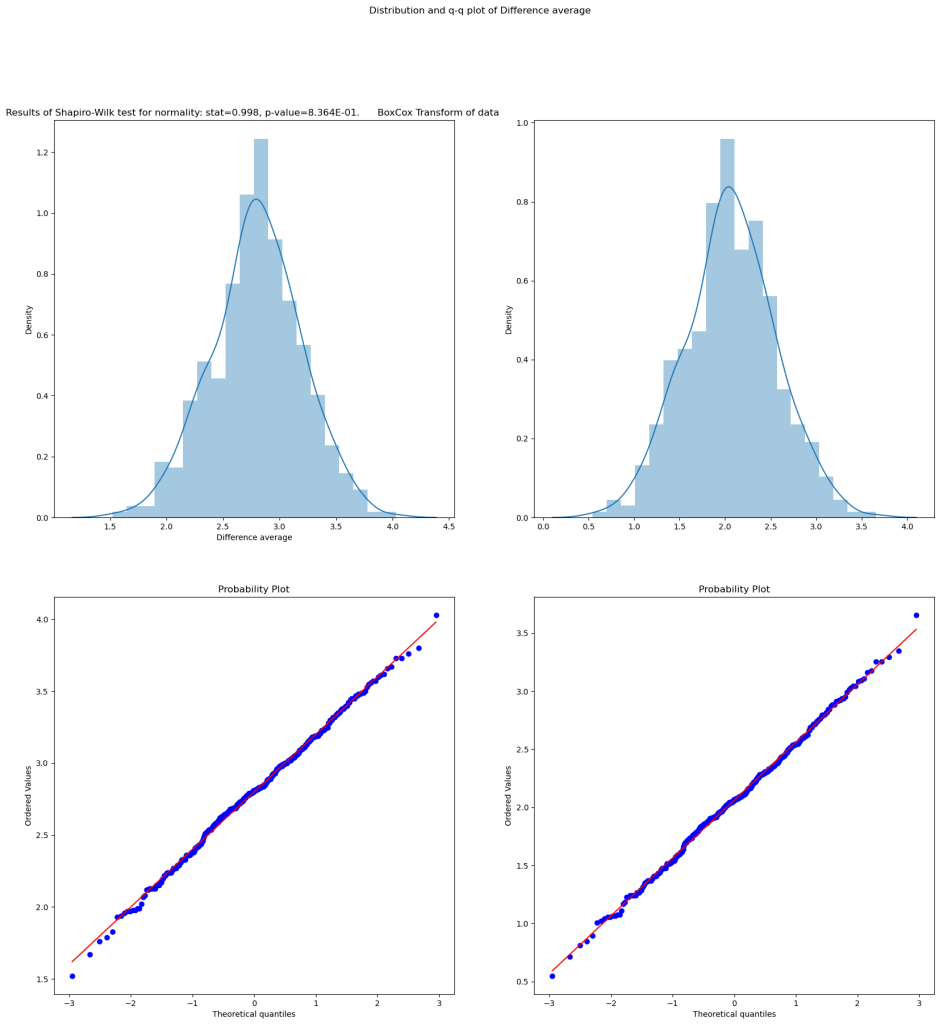


Figure 2.17: Example of boxcox applied to the radiomic feature *Difference Average* which has a close to normal distribution. The graphs contain the original distribution (top-left) the transformed distribution (top-right) and the two respective quantile-quantile plots

ing fold. To use the whole dataset for testing a prediction of it has been built by combinig the predictions on the 10^{th}" fold for ten different models trained on the respective 9 remaining folds. The lasso model from training on the 9 folds is actually chosen as the model with the hyperparameters that give the best performance with another 10-fold crossvalidation. This has been obtained by using *cross_val_predict* on a model obtained by including a *LassoCV* step inside a *pipeline* from scikit-learn library in python. The performance of the cross-validated predictions that, when built this way, is much more representative of the real-world performance of the model, has been evaluated using ROC curves and AUC. Different models have been compared with a Delong test[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.

Finally a few dimensionality reduction techniques, namely the unsupervised PCA[8] and Umap [19] 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.

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

Chapter 3

Results

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

3.1 LASSO

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

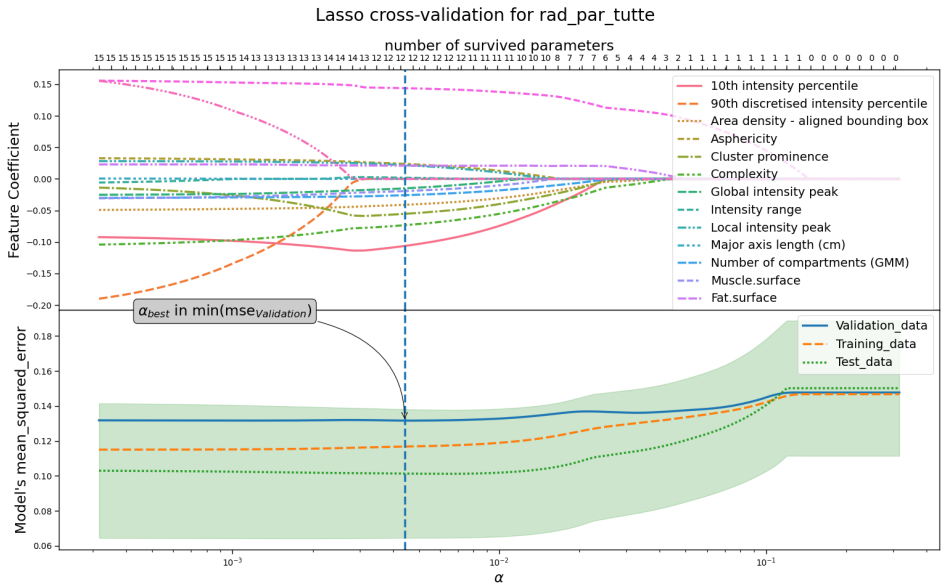


Figure 3.1: Example of graph representing the convergence of the coefficients in a lasso procedure. The final model is chosen by looking at the lowest value in mean squared error computed on the testing dataset. This graph is relative to only radiomic features with the model using death as label.

All of the graphs with respective captions will be, maybe, put in the final appendix. Finally it seems useful to report the following contingency table that gives

an idea on how superimposed the labels on death and ICU admission are.

	ICU Admission	
	0	1
Death		
0	311	47
1	48	30

3.1.1 Death

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

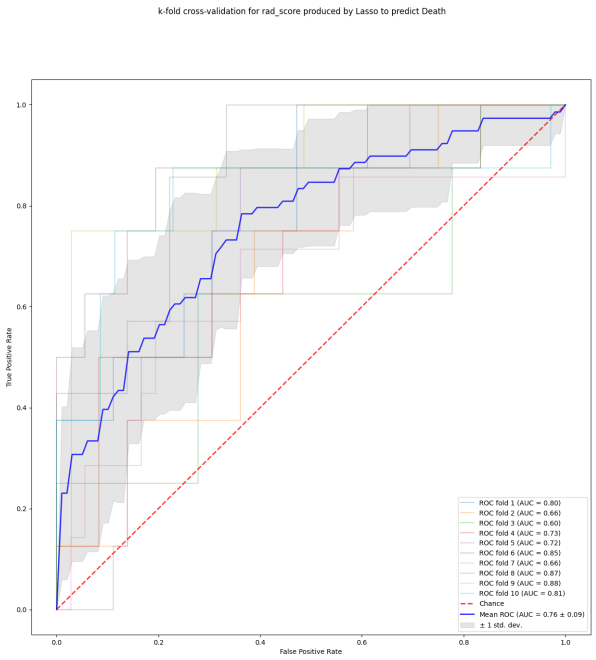


Figure 3.2: ROC curves obtained with crossvalidation procedure using the radiomic features alone.

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

Before commenting more in depth the performance of this model it seems appropriate to see at least the other models built with singular features groups, so, when it comes to the clinical features, the results reported in Figure3.3 and Table3.7 have been obtained. All the results will be put together for ease in Table 3.5.

Table 3.1: Coefficients used in the linear combination estima ted by a Lasso regularization relative to the radiomic features

	lassocvL1 penalty Importance
Intercept	0.178899
10th intensity percentile	-0.125094
Intensity-based interquartile range	0.103349
Complexity	-0.102924
Cluster prominence	-0.064690
Area density - aligned bounding box	-0.039374
Entropy	0.033002
Number of compartments (GMM)	-0.032441
Asphericity	0.028517
Local intensity peak	0.028478
Global intensity peak	-0.024832
Intensity range	0.012509
Fat.surface	0.007267
Major axis length (cm)	0.000000
Number of voxels of positive value	0.000000

Table 3.2: Coefficients used in the linear combination estimated by a Lasso regularization relative to the clinical features

	lassocvL1 penalty 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

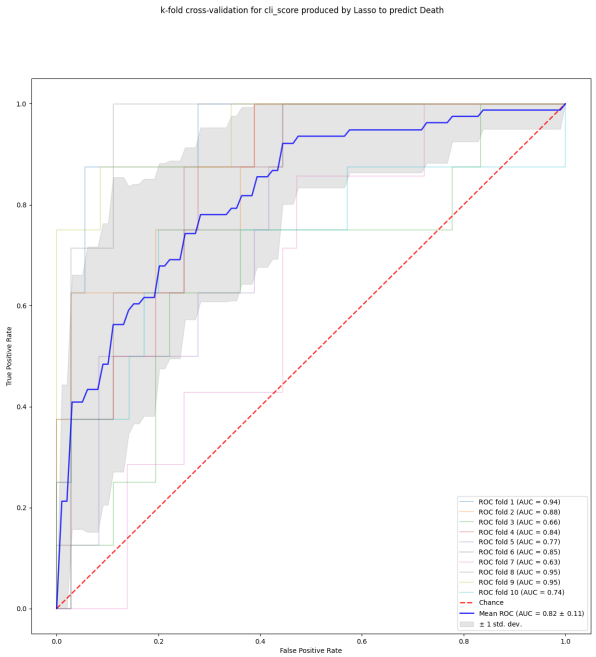


Figure 3.3: ROC curves obtained with crossvalidation procedure using the clinical features alone

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

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

Table 3.3: Coefficients used in the linear combination estimated by a Lasso regularization relative to the radiological features

	lassocvL1 penalty 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

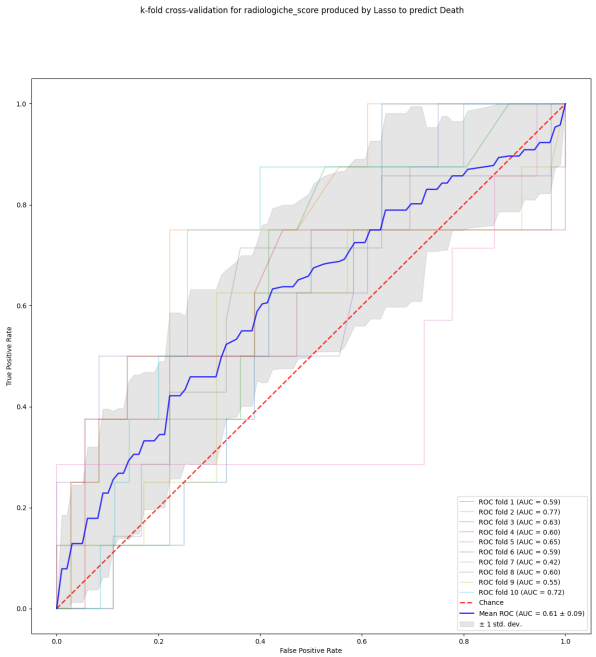


Figure 3.4: ROC curves obtained with crossvalidation procedure using the radiolocical features alone

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

- 10^{th} 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.

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

Table 3.4: Coefficients used in the linear combination estimated by a Lasso regularization relative to all available features features

	lassocvL1 penalty 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

some information on the shape of the organ itself is also relevant. One would expect that when combining all of the available features, i.e. by building a model using the previous clinical, radiomic and radiological features, the performance should somewhat rise especially given the fact that clinical and radiomic features have almost the same performance. The combined results can be seen in Figure 3.5 and Table 3.9.

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

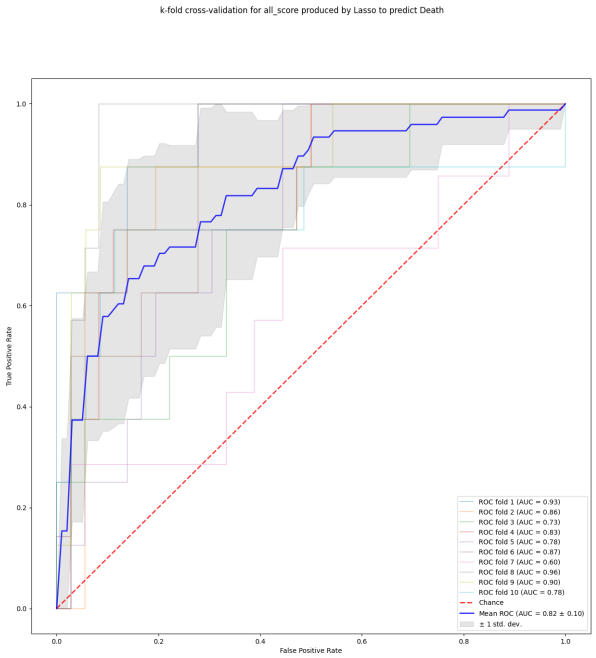
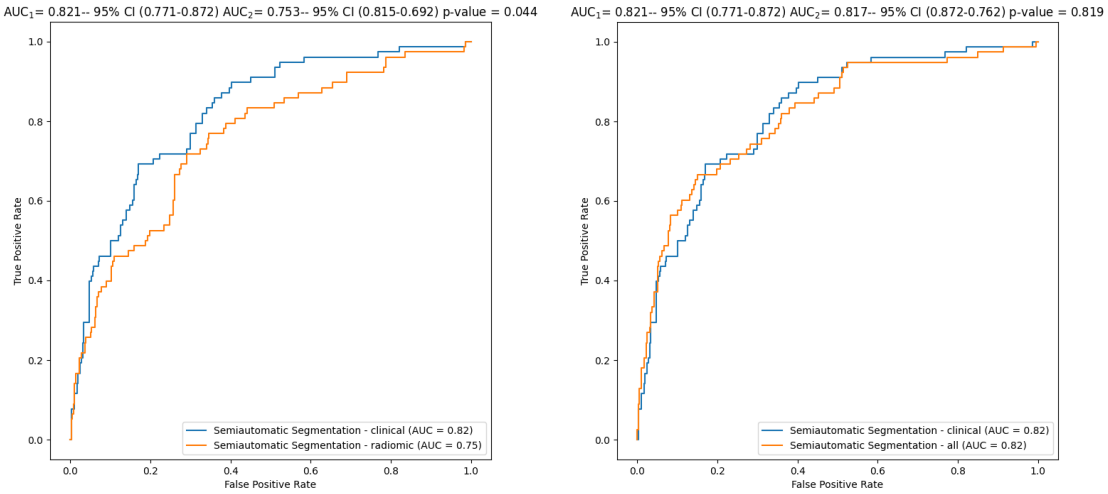


Figure 3.5: ROC curves obtained with crossvalidation procedure using all the available features

Table 3.5: Recap table with the performance of the various families of features

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



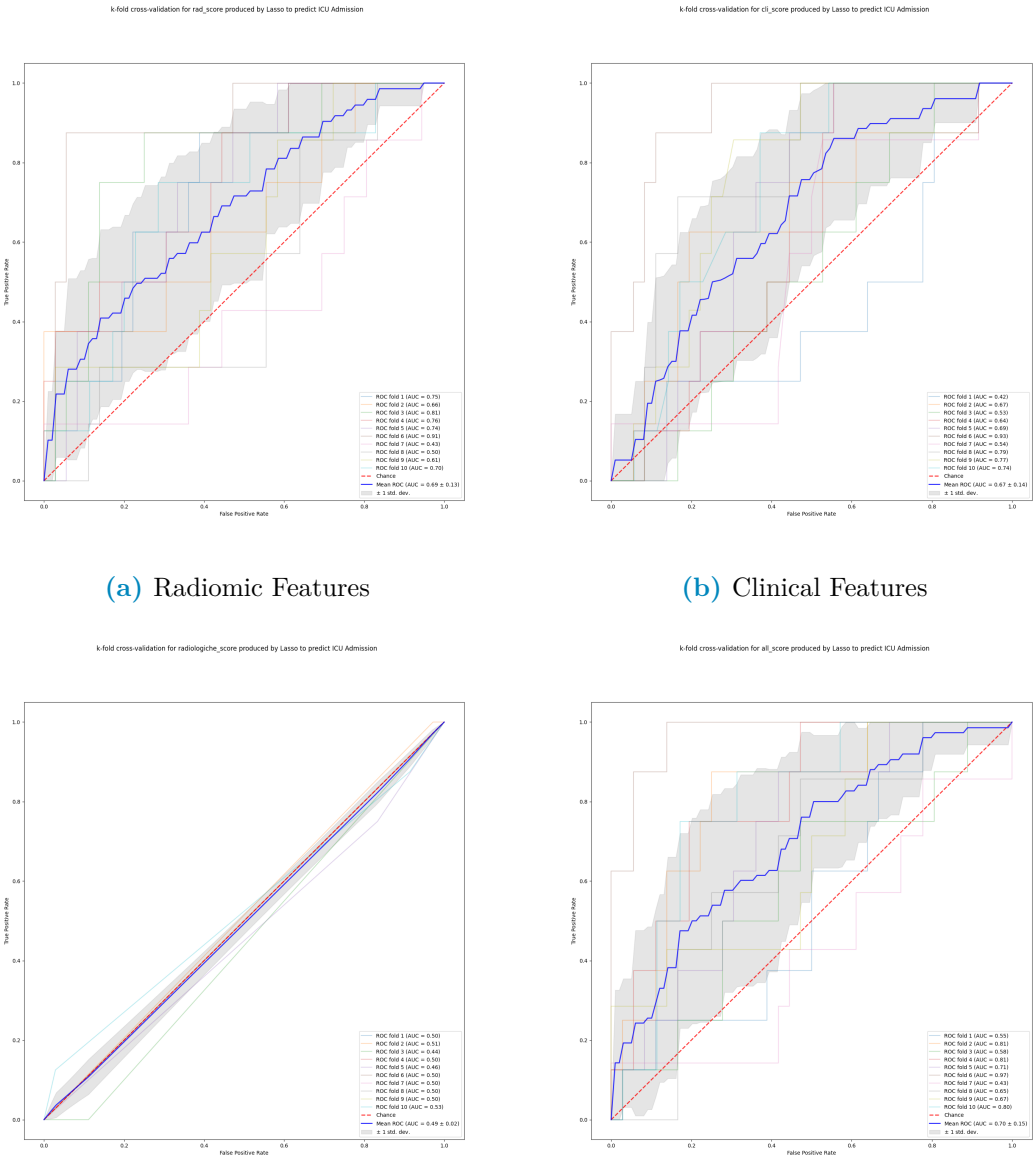
(a) Comparison clincal-radiomic **(b)** Comparison clinical-all

Figure 3.6: Comparison for the curves that are not evidently different

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 ICU Admission

In trying to predic 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 varous 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 $\pm$ 1 standard deviation. Following the blueprint of the previous subsection, all of the results will be presented and then briefly discussed



(a) Radiomic Features

(b) Clinical Features

(c) Radiological Features

(d) All Features

Figure 3.7: Performances of all the models

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

- For the clinical features all of them have a role in the predicition. The only suprising fact, even if it keeps a certain degree of plausibility, is that age is the less relevant out of the available features when it comes to ICU admission.
- None of the radiological features have virtually any impact

Table 3.6: Coefficients used in the linear combination estimated by a Lasso regularization relative to the radiomic features

	lassocvL1 penalty 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 relative to the clinical features

	lassocvL1 penalty 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 relative to the radiological features

	lassocvL1 penalty Importance
Intercept	1.766055e-01
XRayTubeCurrent	6.111319e-18
Lung consolidation	0.000000e+00
Ground-glass	0.000000e+00
Crazy Paving	0.000000e+00
Bilateral Involvement	0.000000e+00
SliceThickness	-0.000000e+00
KVP	0.000000e+00

Table 3.9: Coefficients used in the linear combination estimated by a Lasso regularization relative to all available features features

	lassocvL1 penalty 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 families of features

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

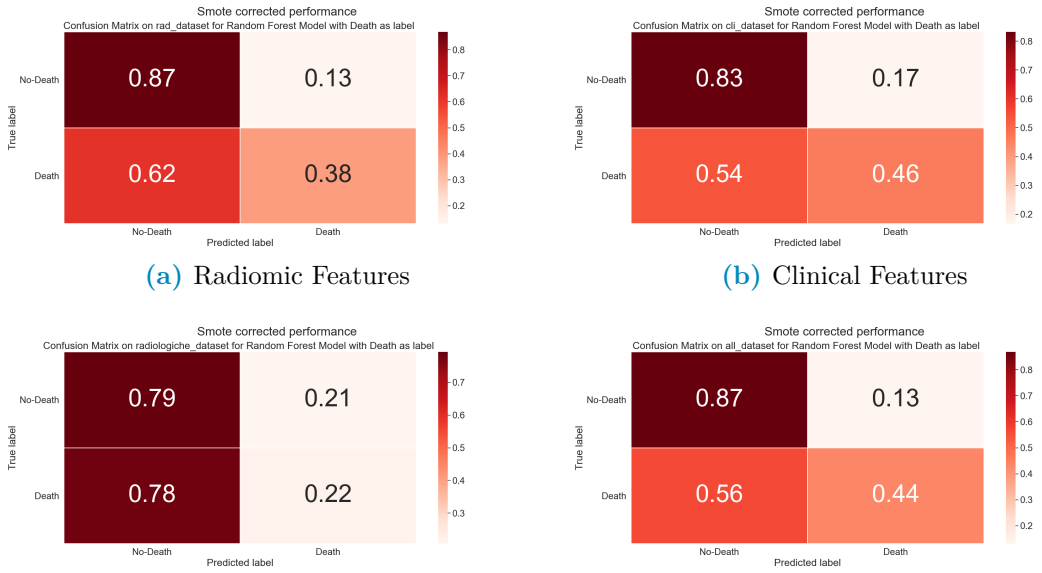
- The radiomic features still value intensity measurements and disorder in the image as primary origins of information. However it seems that shape of the lung now has more relevance in the whole model.

3.2 Ramdom forest

When it comes to random forests the approach followed was similar, in the sense that the various combinations of features were tried, yet different, because the pre-processing step consisted in simply applying a standard scaler to the data. The performance of the models has been looked at using both confusion matrices and ROC curves, the last of which has the only objective of comparing RF classifiers to the perviously described Lasso regression. Once again, following the description scheme used in the section dedicated to Lasso, the results will be divided using the two labels ICU Admission and Death.

3.2.1 Death

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

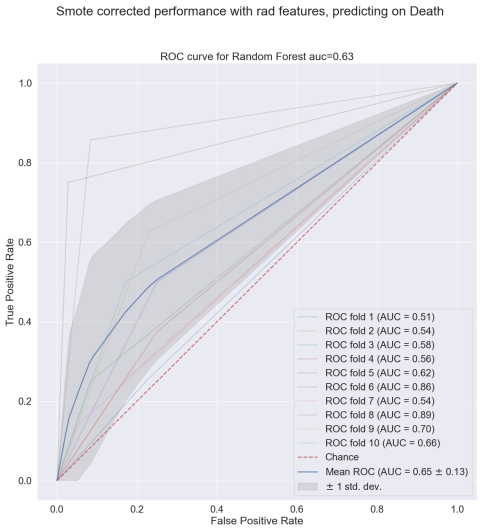


(a) Radiomic Features **(b)** Clinical Features

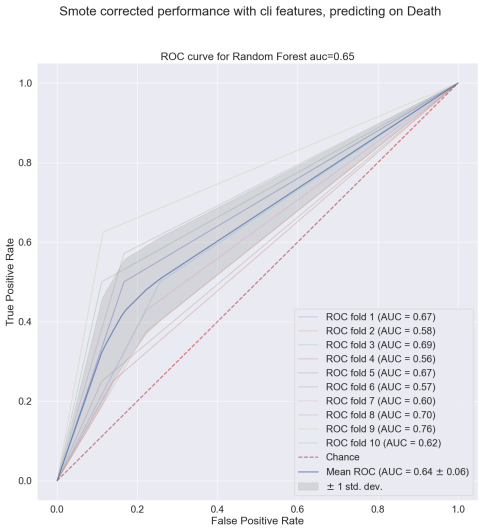
(c) Radiological Features **(d)** All Features

Figure 3.8: Performances of all the models

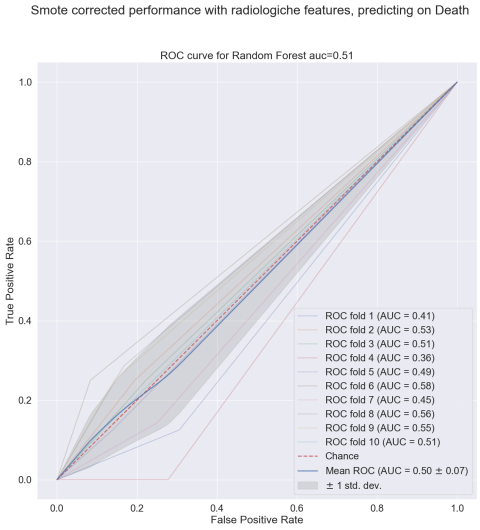
Even without looking at the ROC curves it's plain to see that the data at hand is proving to be difficult for this model. Much like before radiological features alone are useless. In evaluating these confusion matrices it should be kept in mind that the data is heavily unbalanced, since only ∼15% of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



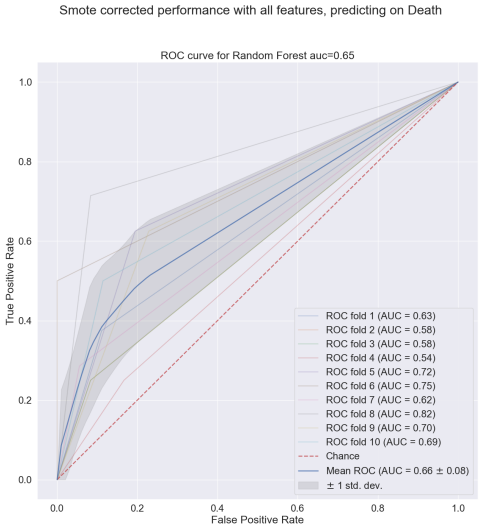
(a) Radiomic Features



(b) Clinical Features



(c) Radiological Features



(d) All Features

Figure 3.9: Performances of all the models

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

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

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

(a) Radiological Features

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

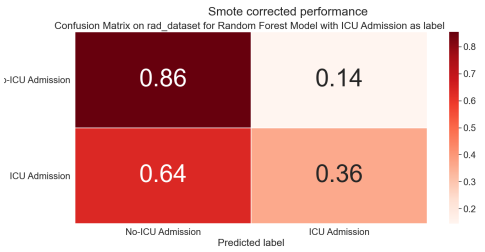
(b) Clinical Features

Figure 3.10: Importances estimated by random forest

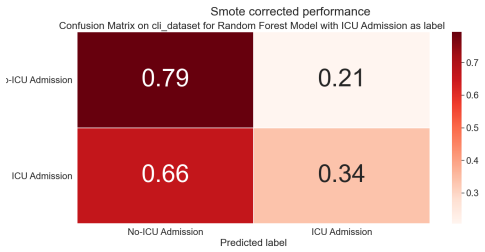
Once again the curves are evidently not statistically different. This counterintuitive behaviour seems to be constant across the two implemented methods and also across different labels tried. An attempt to explain this phenomenon will be postponed to the end of the next subsection

3.2.2 ICU Admission

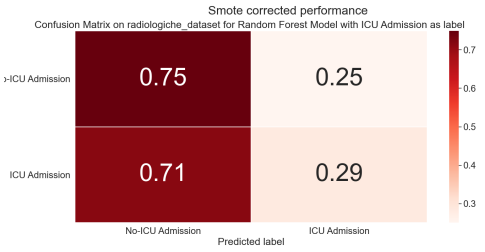
Even when using the admission in the ICU the performance remains pretty much the same, so the comments would still be the same as before



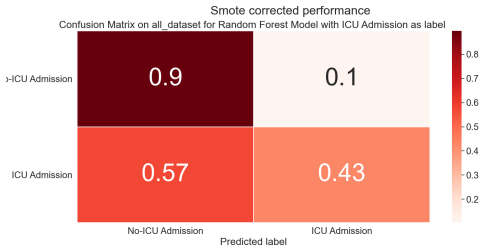
(a) Radiomic Features



(b) Clinical Features



(c) Radiological Features

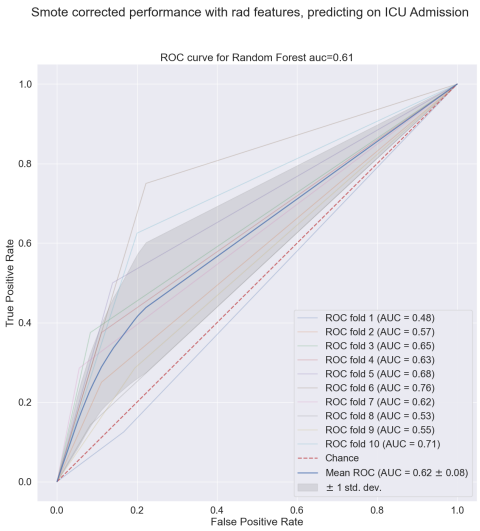


(d) All Features

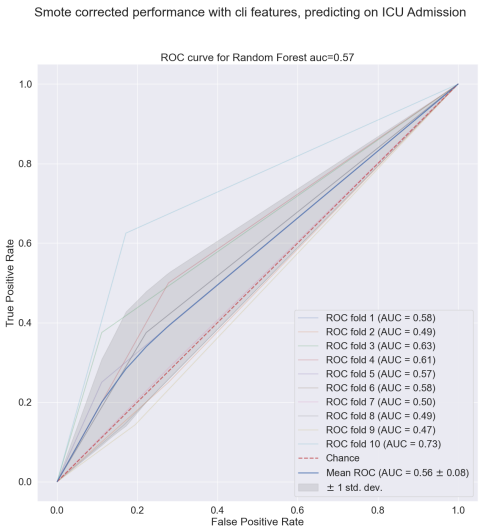
Figure 3.11: Performances of all the models

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

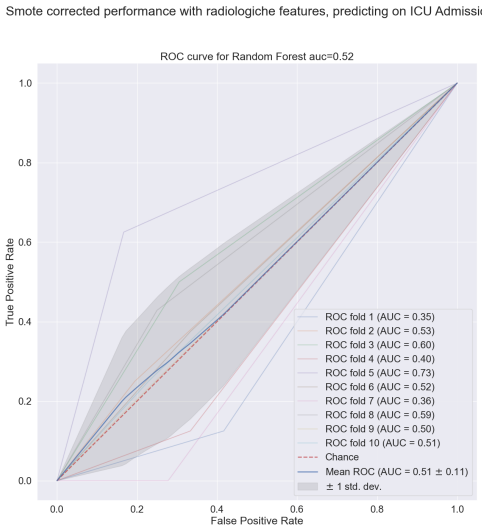
Features used	mean AUC ±std
Radiomic	0.62 ± 0.08
Clinical	0.56 ± 0.08
Radiological	0.51 ± 0.11
All	0.64 ± 0.09



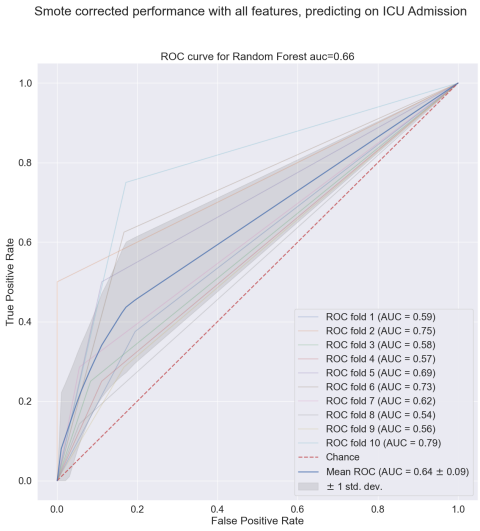
(a) Radiomic Features



(b) Clinical Features



(c) Radiological Features



(d) All Features

Figure 3.12: Performances of all the models

It is very difficult to quantify what the expectations for each model were a priori, this transposes to difficulties in trying to diagnose problems in the models when considered singularly. However when combining all of the available features it's very

strange that no improvement in performance can be achieved. This would mean that 8 clinical variables provide the exact same amount of information as the total $\sim$200 features most of which derived from images which, at least ideally, contain much more accurate information. One could surmise that the analysis methods have been implemented in an incorrect way, yet when two methods implemented differently and separately obtain the same result it reinforces the idea that the problem is somewhere before the analysis. All of these analses started from the following hypotheses:

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

It is very clear that the first three hypotheses are verified with a caveat, the performance of the radiomic pipeline hinges on the quality of the images and of the segmentation procedure performed on them. Since the images used in this thesis were, are and will be used by the hospital in routine processes it's close to impossible that all of them have problems, expecially because of the preliminary screening done before segmentation. Another possibility is that the images are not really really representative of the situation of the patient. Since one of the prevailing properties of *Sars-COVID19* is the speed with which the clinical picture of the patient can change it's possible that images taken at admission are not as informative of the final prognosis. This problem is, however, unavoidable in the setting of this thesis which has as aim the construction of a model that, exactly at admission, can discriminate between serious and easier cases. Another possibility is that there are two or more subgroups in the patient cohort and the performance on these is widely different, determiniing an average performance below the expectations. To diagnose if this is the case a few dimensionality reduction techniques have been used to visualize the data and to prepare for clustering in case of need, all of the results of these procedure will be presented in the followig section.

Chapter 4

Conclusding remarks

Chapter 5

Appendix

5.1 Random Forest

Here are the tables with all of the features from random forest.

	RF_importances
Intensity histogram quartile coefficient of dis...	0.029548
Intensity-based interquartile range	0.019703
Quartile coefficient of dispersion	0.018654
Intensity-based median absolute deviation	0.017953
Discretised interquartile range	0.017409
Entropy	0.015616
Maximum histogram gradient intensity	0.014988
Small zone emphasis	0.014766
Normalised zone size non-uniformity	0.014295
Uniformity	0.013181
Intensity-based robust mean absolute deviation	0.012899
Information correlation 2	0.011514
Information correlation 1	0.010328
Intensity histogram mode	0.010108
Grey level variance (GLSZM)	0.009572
Variance	0.009489
Volume density - enclosing ellipsoid	0.009423
Small zone low grey level emphasis	0.009294
High dependence low grey level emphasis	0.009282
Centre of mass shift (cm)	0.008994
RECIST (cm)	0.008748
Fat.surface	0.008673
Zone size entropy	0.008618
Normalised grey level non-uniformity (GLSZM)	0.008317
Skewness	0.008129
Normalised grey level non-uniformity (NGLDM)	0.008092
Intensity-based mean absolute deviation	0.008066
Run entropy	0.007956
Cluster shade	0.007879

Area density - enclosing ellipsoid	0.007732
Strength	0.007714
Grey level variance (GLDZM)	0.007624
Number of voxels of positive value	0.007468
Small distance high grey level emphasis	0.007429
Discretised intensity skewness	0.007271
Thresholded area intensity peak (50%)	0.007256
Cluster prominence	0.007156
Max value	0.007093
Intensity fraction difference between volume fr...	0.007064
Grey level non-uniformity (NGLDM)	0.006823
Number of grey levels	0.006765
Small distance low grey level emphasis	0.006687
Normalised zone distance non-uniformity	0.006680
Dependence count energy	0.006624
10th intensity percentile	0.006534
Short run low grey level emphasis	0.006477
Low grey level count emphasis	0.006429
Low dependence emphasis	0.006421
Dependence count entropy	0.006393
Normalised grey level non-uniformity (GLDZM)	0.006317
Minor axis length (cm)	0.006285
Intensity histogram median absolute deviation	0.006284
Local intensity peak	0.006278
Large distance high grey level emphasis	0.006204
Volume at intensity fraction 10%	0.006153
Zone distance non-uniformity	0.006063
Intensity range	0.006049
Intensity histogram robust mean absolute deviation	0.006030
Thresholded area intensity peak (75%)	0.005936
Zone distance variance	0.005884
Number of compartments (GMM)	0.005813
Small distance emphasis	0.005732
Volume density - convex hull	0.005725
Discretised intensity uniformity	0.005615
Dependence count variance	0.005582
Area density - convex hull	0.005566
Inverse elongation	0.005563
Integrated intensity	0.005497
Low grey level zone emphasis	0.005257
Long run low grey level emphasis	0.005238
Zone distance entropy	0.005171
Maximum histogram gradient	0.005151
Flatness	0.005066
Large zone high grey level emphasis	0.005056
Volume density - oriented bounding box	0.005050
Grey level non-uniformity (GLRLM)	0.005039

Low grey level run emphasis	0.005011
Kurtosis	0.004993
Difference variance	0.004958
Difference entropy	0.004955
Minimum histogram gradient	0.004873
Volume density - aligned bounding box	0.004822
Correlation	0.004817
Low grey level zone emphasis.1	0.004795
Intensity histogram coefficient of variation	0.004748
Sum entropy	0.004744
Major axis length (cm)	0.004657
Low dependence low grey level emphasis	0.004641
Muscle.surface	0.004627
Global intensity peak	0.004566
Large distance emphasis	0.004547
Sphericity	0.004515
Discretised intensity variance	0.004511
Normalised homogeneity	0.004504
Surface to volume ratio	0.004483
Area density - aligned bounding box	0.004475
Volume fraction difference between intensity fr...	0.004464
Intensity-based coefficient of variation	0.004421
Zone percentage (GLDZM)	0.004409
Intensity at volume fraction 90%	0.004392
Intensity histogram mean absolute deviation	0.004347
Large distance low grey level emphasis	0.004275
Intensity-based energy	0.004272
Run length variance	0.004266
Normalised grey level non-uniformity (GLRLM)	0.004243
Sum variance	0.004203
Area density - oriented bounding box	0.004180
Maximum 3D diameter (cm)	0.004140
Grey level variance (GLRLM)	0.004103
Spherical disproportion	0.004082
Minimum histogram gradient intensity	0.004025
Large zone low grey level emphasis	0.004004
Grey level non-uniformity (GLDZM)	0.003987
Joint maximum	0.003907
Short run emphasis	0.003906
Volume at intensity fraction 90%	0.003904
Inverse variance	0.003899
Low dependence high grey level emphasis	0.003874
Dependence count non-uniformity	0.003858
Homogeneity	0.003804
Zone percentage (GLSZM)	0.003709
Discretised intensity kurtosis	0.003649
Contrast (GLCM)	0.003622

90th discretised intensity percentile	0.003614
Run length non-uniformity	0.003599
Normalized dependence count non-uniformity	0.003581
Angular second moment	0.003577
High dependence high grey level emphasis	0.003552
Compactness 2	0.003532
Short run high grey level emphasis	0.003411
Energy	0.003383
Discretised intensity standard deviation	0.003333
Cluster tendency	0.003298
Run percentage	0.003295
Asphericity	0.003269
Grey level non-uniformity (GLSZM)	0.003266
Sum average	0.003255
Standard deviation	0.003245
Normalised run length non-uniformity	0.003236
Compactness 1	0.003224
Least axis length (cm)	0.003223
Area under the IVH curve	0.003208
High grey level zone emphasis.1	0.003195
Complexity	0.003127
Joint Entropy	0.003124
Joint variance	0.003100
High grey level run emphasis	0.003088
Number of voxels	0.003059
Difference average	0.003040
Autocorrelation	0.003015
Small zone high grey level emphasis	0.002987
High dependence emphasis	0.002985
Discretised intensity entropy	0.002928
Intensity mean value	0.002916
Grey level variance (NGLDM)	0.002913
Mean discretised intensity	0.002898
Coarseness	0.002887
Busyness	0.002865
Long run high grey level emphasis	0.002830
Large zone emphasis	0.002829
Dissimilarity	0.002744
Intensity at volume fraction 10%	0.002709
Normalised inverse difference	0.002642
High grey level zone emphasis	0.002615
Quadratic mean	0.002601
Contrast (NGTDM)	0.002524
Joint average	0.002438
High grey level count emphasis	0.002437
Intensity median value	0.002347
Min value	0.002292

Inverse difference	0.002250
Long run emphasis	0.002236
90th intensity percentile	0.001716
Median discretised intensity	0.001637
Number of grey levels after quantization	0.000000
Discretized intensity range	0.000000
Discretised min value	0.000000
Discretised max value	0.000000
Dependence count percentage	0.000000

Table 5.1: Importances determined by RandomForest predicting death

	RF_importances
Age (years)	0.071763
Febbre	0.027207
Quartile coefficient of dispersion	0.019169
Intensity histogram quartile coefficient of dis...	0.018356
Intensity-based interquartile range	0.015819
Normalised zone size non-uniformity	0.014963
Small zone emphasis	0.014390
Maximum histogram gradient intensity	0.013045
Respiratory Rate	0.012050
Information correlation 2	0.011174
Uniformity	0.011141
Run entropy	0.011108
Dependence count entropy	0.011059
Intensity-based median absolute deviation	0.010593
Zone size entropy	0.010560
Discretised interquartile range	0.010542
Information correlation 1	0.010528
Dependence count energy	0.009528
Ground-glass	0.009497
XRayTubeCurrent	0.008914
Centre of mass shift (cm)	0.008845
Small zone low grey level emphasis	0.008554
Entropy	0.008272
Fat.surface	0.008249
Normalised zone distance non-uniformity	0.007995
Volume density - enclosing ellipsoid	0.007707
Correlation	0.007107
Intensity histogram mode	0.006950
Local intensity peak	0.006785
RECIST (cm)	0.006770
Volume at intensity fraction 90%	0.006717
Intensity-based robust mean absolute deviation	0.006703
Integrated intensity	0.006683

Intensity histogram robust mean absolute deviation	0.006672
Inverse elongation	0.006464
Difference variance	0.006397
Normalised grey level non-uniformity (NGLDM)	0.006304
Low dependence emphasis	0.006206
Variance	0.006167
Muscle.surface	0.006136
Number of grey levels	0.006086
Large distance high grey level emphasis	0.005992
Zone distance non-uniformity	0.005963
Volume density - oriented bounding box	0.005962
Large distance emphasis	0.005933
Intensity-based energy	0.005833
Normalised grey level non-uniformity (GLSZM)	0.005791
Cluster shade	0.005748
Intensity-based coefficient of variation	0.005698
Thresholded area intensity peak (75%)	0.005593
Minor axis length (cm)	0.005591
Skewness	0.005528
Intensity fraction difference between volume fr...	0.005457
Area density - enclosing ellipsoid	0.005426
Low dependence low grey level emphasis	0.005393
Small distance low grey level emphasis	0.005372
Intensity-based mean absolute deviation	0.005342
Grey level variance (GLSZM)	0.005326
Sex_bin	0.005325
Low grey level zone emphasis.1	0.005256
Volume at intensity fraction 10%	0.005220
Discretised intensity uniformity	0.005215
Zone distance entropy	0.005162
Intensity at volume fraction 90%	0.005113
Small distance high grey level emphasis	0.005086
Max value	0.005068
Grey level non-uniformity (GLSZM)	0.004997
Thresholded area intensity peak (50%)	0.004991
High dependence low grey level emphasis	0.004936
Number of voxels of positive value	0.004893
Inverse variance	0.004862
Normalised homogeneity	0.004862
Low grey level run emphasis	0.004837
Low grey level zone emphasis	0.004794
Area density - oriented bounding box	0.004768
Cluster prominence	0.004755
Surface to volume ratio	0.004729
Long run low grey level emphasis	0.004692
Grey level variance (GLDZM)	0.004681
Large distance low grey level emphasis	0.004671

Large zone high grey level emphasis	0.004635
Normalised grey level non-uniformity (GLDZM)	0.004585
Small distance emphasis	0.004565
Dissimilarity	0.004557
Intensity range	0.004506
Sum entropy	0.004495
Difference entropy	0.004477
Grey level non-uniformity (GLDZM)	0.004368
Global intensity peak	0.004363
Strength	0.004354
Area density - aligned bounding box	0.004321
Contrast (GLCM)	0.004280
Zone percentage (GLSZM)	0.004247
Volume density - convex hull	0.004244
Grey level non-uniformity (GLRLM)	0.004229
Discretised intensity kurtosis	0.004193
Number of compartments (GMM)	0.004189
Normalised inverse difference	0.004163
Major axis length (cm)	0.004148
Volume density - aligned bounding box	0.004119
Small zone high grey level emphasis	0.004046
Run length non-uniformity	0.004043
Dependence count non-uniformity	0.003985
Low grey level count emphasis	0.003960
Grey level variance (GLRLM)	0.003893
High dependence emphasis	0.003838
Min value	0.003808
Least axis length (cm)	0.003785
Difference average	0.003785
Zone percentage (GLDZM)	0.003784
Volume fraction difference between intensity fr...	0.003777
Large zone emphasis	0.003710
Discretised intensity variance	0.003684
Normalised grey level non-uniformity (GLRLM)	0.003645
Maximum histogram gradient	0.003605
Energy	0.003602
Compactness 1	0.003541
Number of voxels	0.003514
Intensity histogram coefficient of variation	0.003501
Compactness 2	0.003442
Short run emphasis	0.003425
High grey level zone emphasis.1	0.003424
Short run low grey level emphasis	0.003422
10th intensity percentile	0.003370
Discretised intensity skewness	0.003370
Autocorrelation	0.003328
Run percentage	0.003320

90th intensity percentile	0.003256
Normalised run length non-uniformity	0.003242
Zone distance variance	0.003207
Large zone low grey level emphasis	0.003180
Intensity histogram mean absolute deviation	0.003099
Grey level variance (NGLDM)	0.003099
Run length variance	0.003081
Sphericity	0.002994
Area density - convex hull	0.002974
Maximum 3D diameter (cm)	0.002954
Normalized dependence count non-uniformity	0.002895
High grey level zone emphasis	0.002887
Angular second moment	0.002855
Intensity mean value	0.002833
Asphericity	0.002802
Dependence count variance	0.002793
Flatness	0.002756
Busyness	0.002713
Obesity	0.002699
Coarseness	0.002674
Kurtosis	0.002651
High grey level run emphasis	0.002601
SliceThickness	0.002589
Contrast (NGTDM)	0.002581
Complexity	0.002575
High dependence high grey level emphasis	0.002566
Homogeneity	0.002557
Minimum histogram gradient	0.002495
Inverse difference	0.002483
Joint variance	0.002408
Grey level non-uniformity (NGLDM)	0.002400
Low dependence high grey level emphasis	0.002332
Joint average	0.002325
Discretised intensity standard deviation	0.002309
KVP	0.002279
Long run high grey level emphasis	0.002207
Long run emphasis	0.002203
Mean discretised intensity	0.002189
90th discretised intensity percentile	0.002176
Intensity median value	0.002146
Cluster tendency	0.002107
Crazy Paving	0.002097
Intensity at volume fraction 10%	0.002050
Quadratic mean	0.001997
Discretised intensity entropy	0.001957
Spherical disproportion	0.001938
Short run high grey level emphasis	0.001915

Sum average	0.001910
Sum variance	0.001795
Standard deviation	0.001791
Intensity histogram median absolute deviation	0.001672
Area under the IVH curve	0.001653
Median discretised intensity	0.001640
Joint Entropy	0.001634
Joint maximum	0.001573
Minimum histogram gradient intensity	0.001321
High grey level count emphasis	0.001244
History of smoking	0.001035
Bilateral Involvement	0.000801
Hypertension	0.000749
Lung consolidation	0.000381
HRCT performed	0.000000
Discretised max value	0.000000
Discretised min value	0.000000
Dependence count percentage	0.000000
Discretized intensity range	0.000000
Number of grey levels after quantization	0.000000

Table 5.2: Importances determined by RandomForest predicting ICU Admission

5.2 Dimensionality reduction

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

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

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

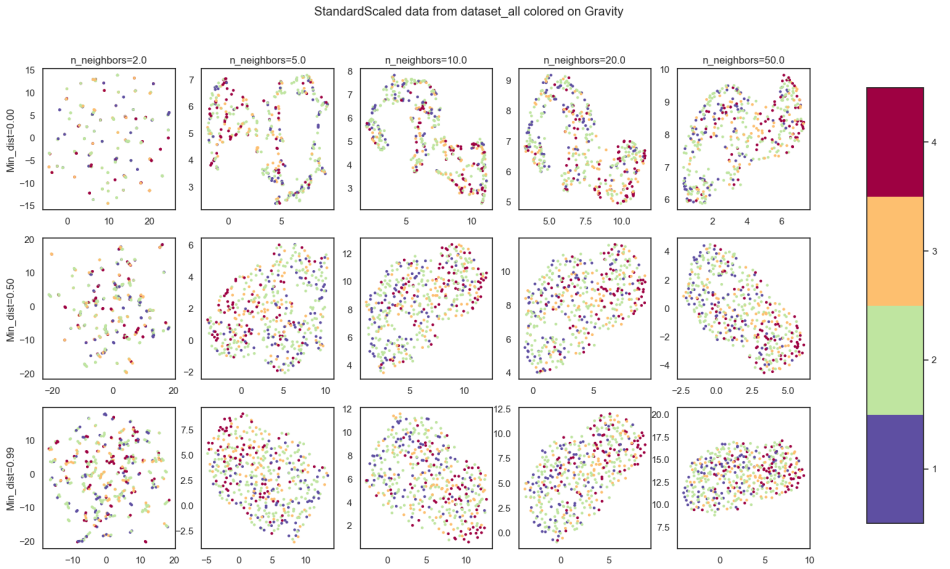


Figure 5.1: Possible combination for umap hyperparameters "number of neighbours" and "minimum distance". Color coding is done with aforementioned gravity score and all the features, i.e. clinical radiomic and radiological, were used.

5.2.1 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.

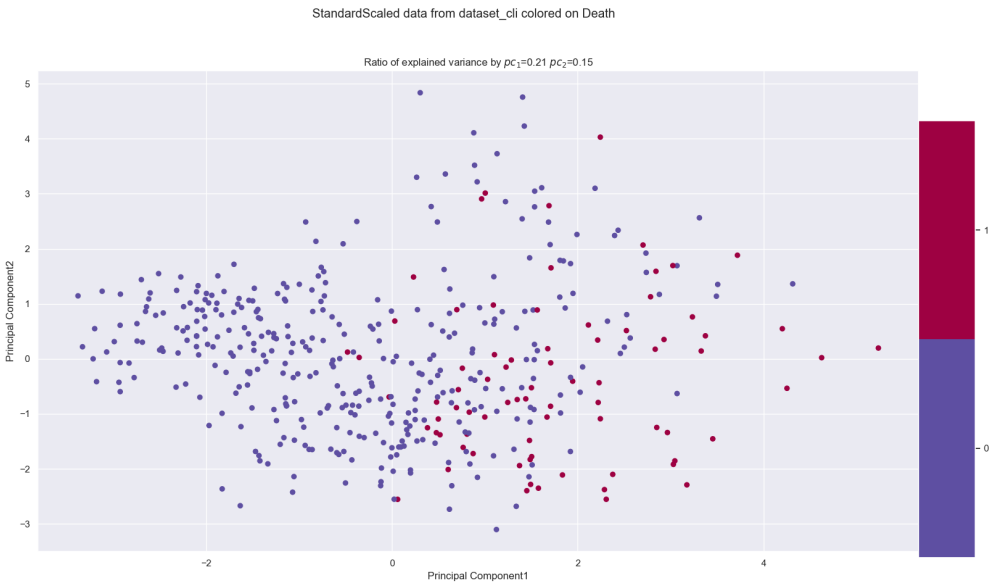


Figure 5.2: 2-Principal Component on clinical features.

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.

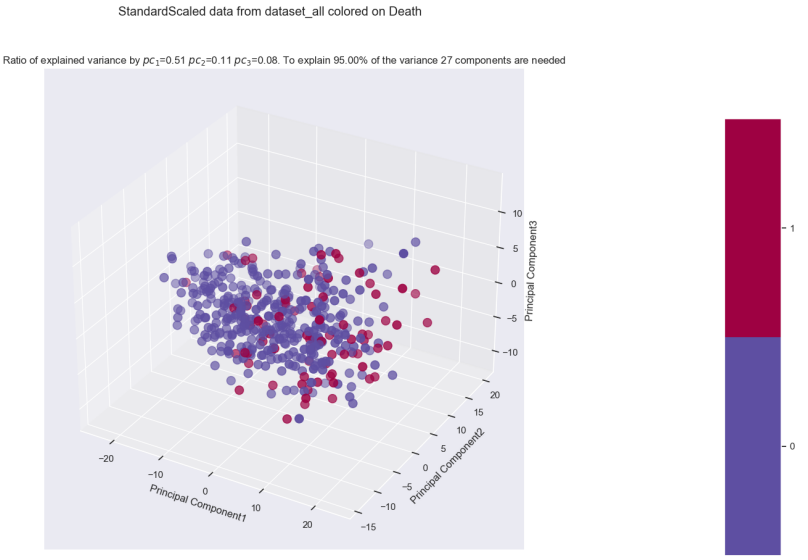


Figure 5.3: 3D PCA of whole dataset, colored with death label

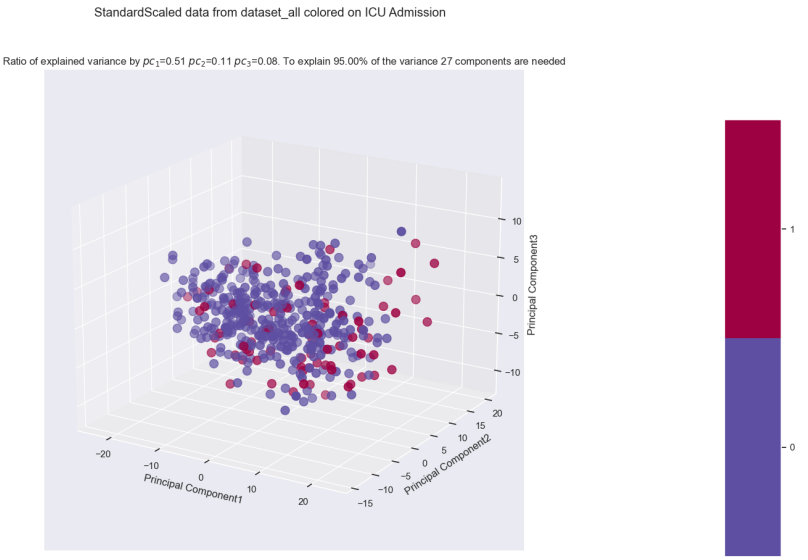


Figure 5.4: 3D PCA of whole dataset, colored with ICU Admission label

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.

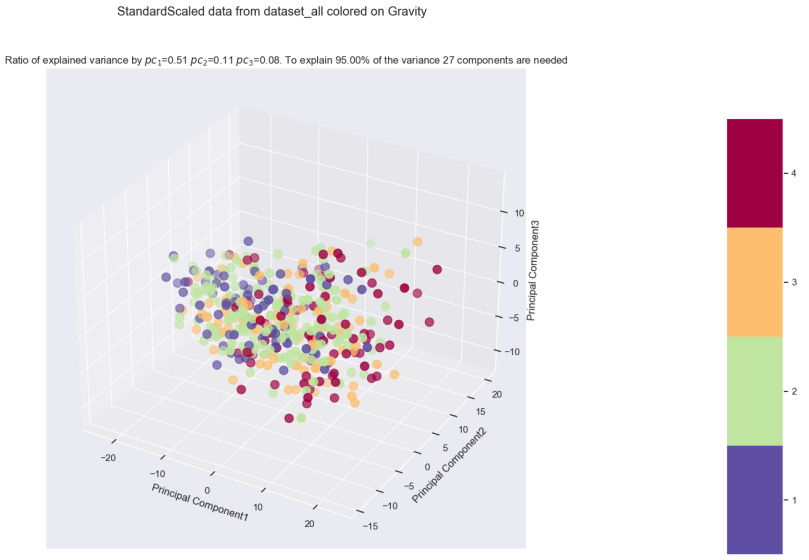


Figure 5.5: Comparison between various colour labels of the top 3 principal components for the entire dataset. Note that to explain 95% of the variance 27 components would be needed

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.

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

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

Moving on from unsupervised methods to a supervised one, PLS-DA was used

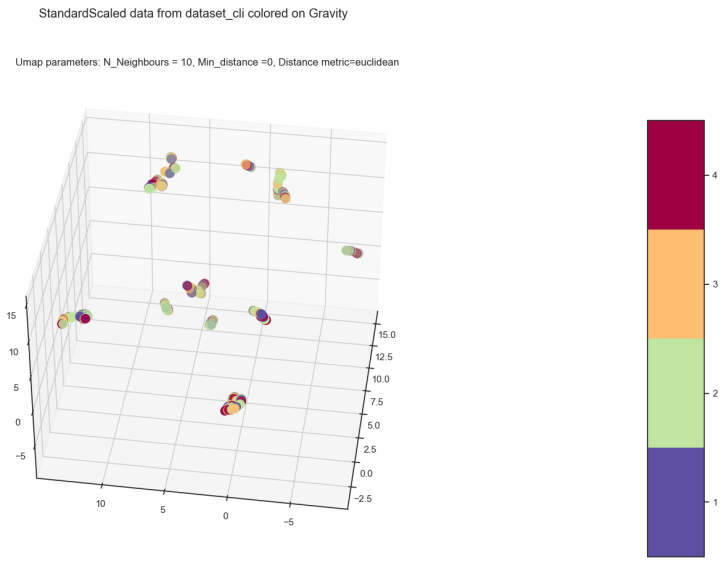


Figure 5.6: 3D umap of clinical dataset, colored based on gravity

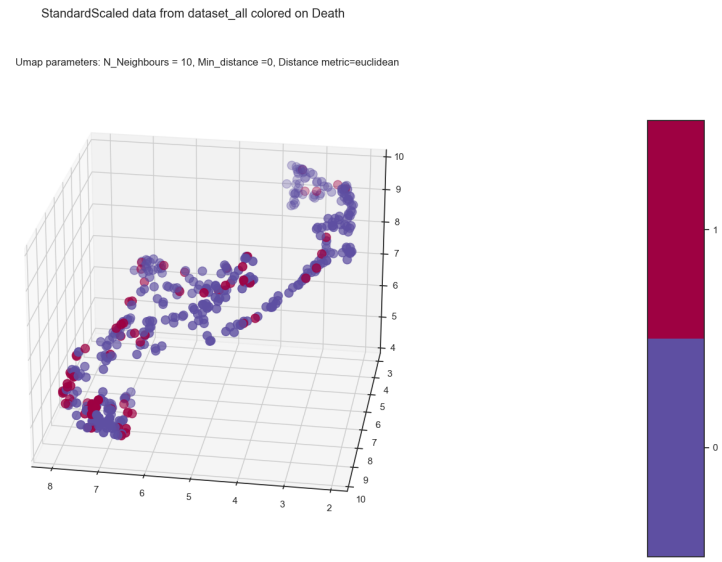


Figure 5.7: 3D umap of whole dataset, colored based on death

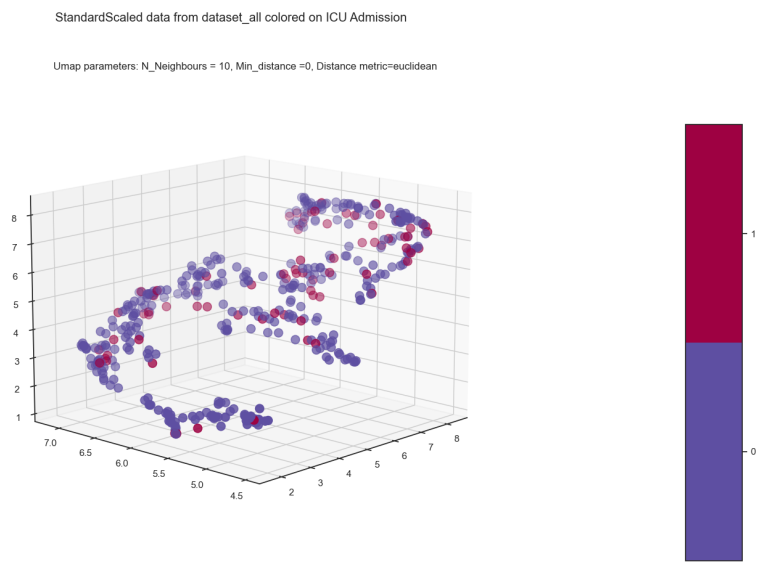


Figure 5.8: 3D umap of whole dataset, colored based on ICU admission

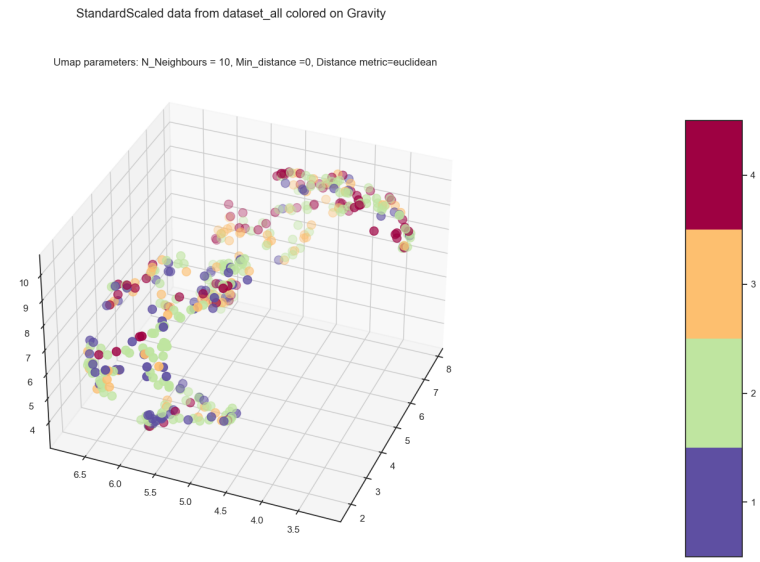


Figure 5.9: 3D umap of whole dataset, colored based on gravity

giving as label both death and ICU using both whole dataset, and singularly radiomic or clinical features. Starting from the clinical features alone, predicting on death Figure 5.10 can be obtained.

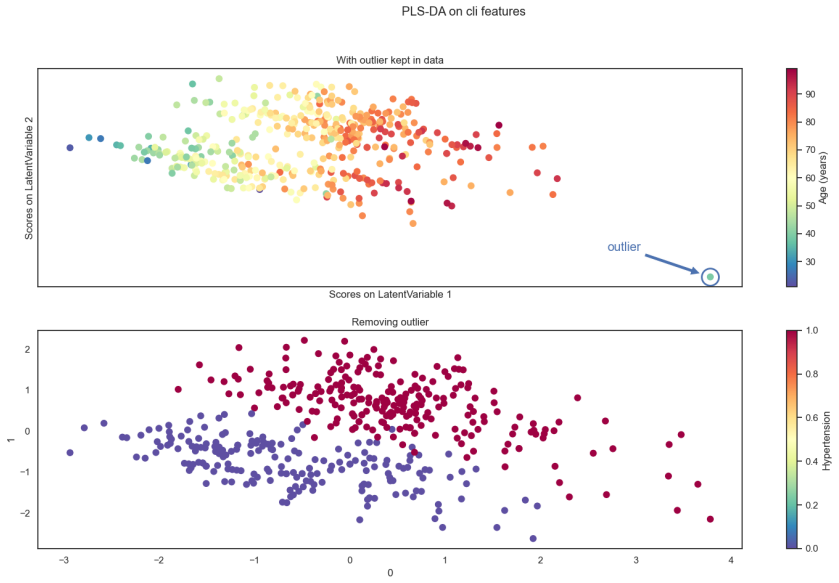


Figure 5.10: PLS-DA predicting on death coloured with age and hypertension

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

It's clear to see that, when predicting on death, the PLS-DA algorithm doesn't find any behaviour relevant for ICU admission. It's also clear that there is at least a pattern of points labeled as dead being towards the right of the image, this can be easily explained by looking at how the ages are distributed in the first plot of fig: 5.10. From this it's possible to deduce that older individuals tend to die more and that hypertension does not seem to be relevant when considering death as a clinical outcome. If necessary the PLS-DA algorithm allows also to see the weights given to the features in predicting the label. At least for the clinical dataset, which has a reasonable number of features, it's interesting to report it ordering the coefficients by descending absolute value:

Doing the exact same procedure on the whole dataset, which means by including the radiomic features, Figure 5.12 can be obtained.

Once again adding the radiomic features has evidently introduced noise in the system, which no longer displays any kind of behaviour, pattern nor separation.

Doing the same analysis but using ICU Admission as a label Figure 5.13 can be

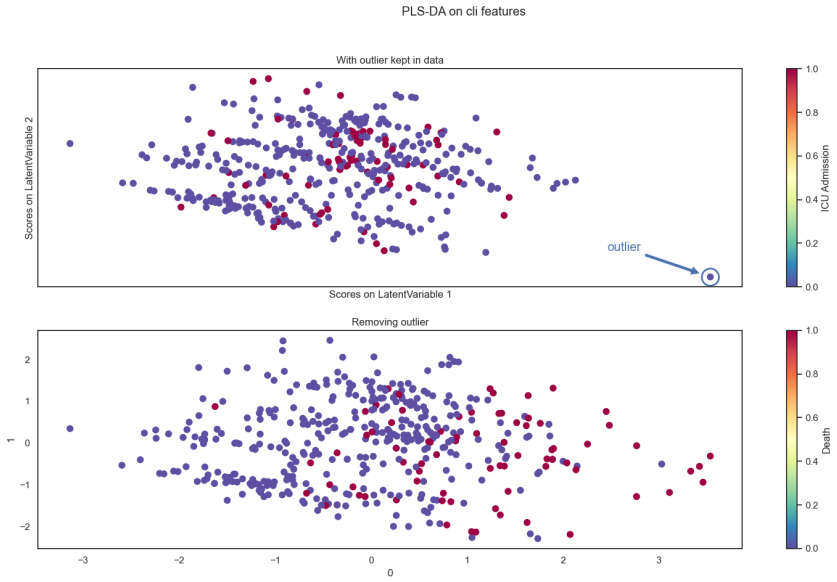


Figure 5.11: PLS-DA predicting on death coloured with death(bottom) and ICU Admission(top)

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

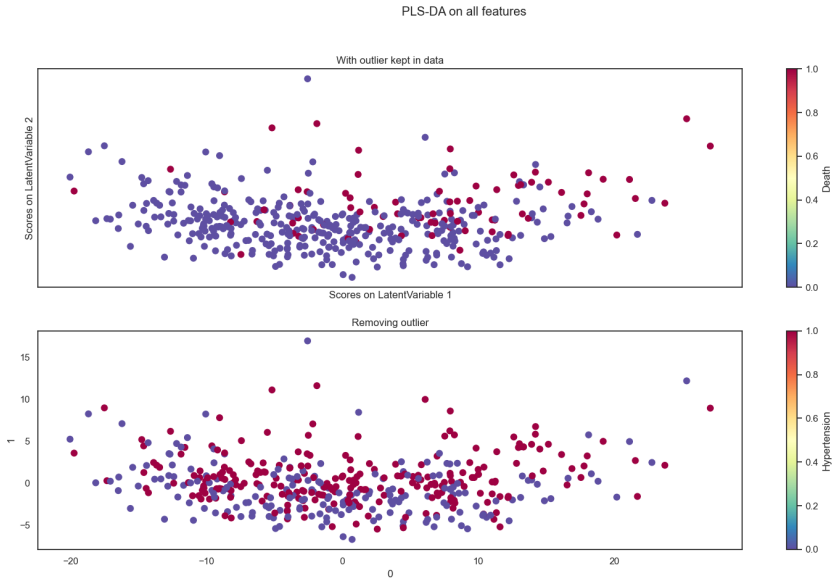


Figure 5.12: PLS-DA predicting on death coloured with death(top) and hypertension(bottom) on whole dataset

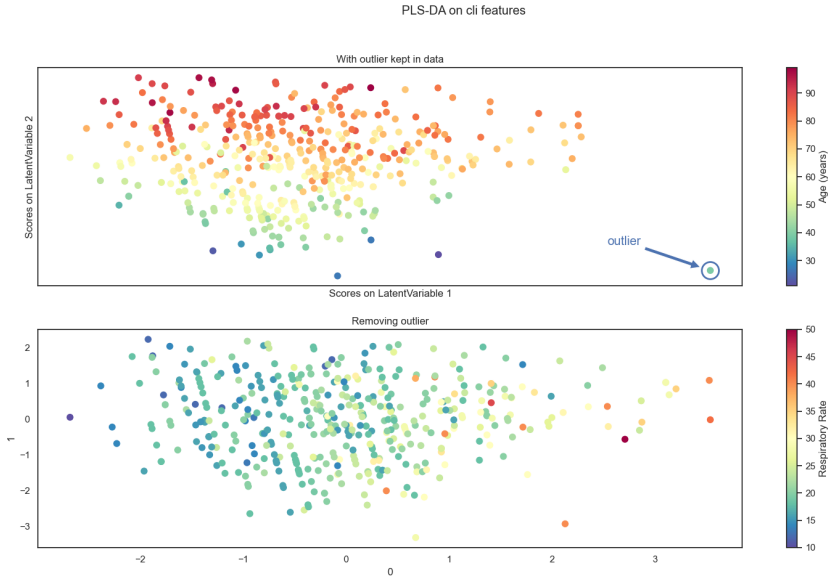


Figure 5.13: PLS-DA predicting on death coloured with Age and respiratory rate on clinical features

obtained.

Now the colors have been chosen to highlight that respiratory rate and age have the main role in determining the latent variables. However, in this case, there doesn't appear to be a clear cut distinction as it happened before with hypertension. Looking at how the points scatter by coloring them according to the two interesting clinical labels the following figure can be obtained:

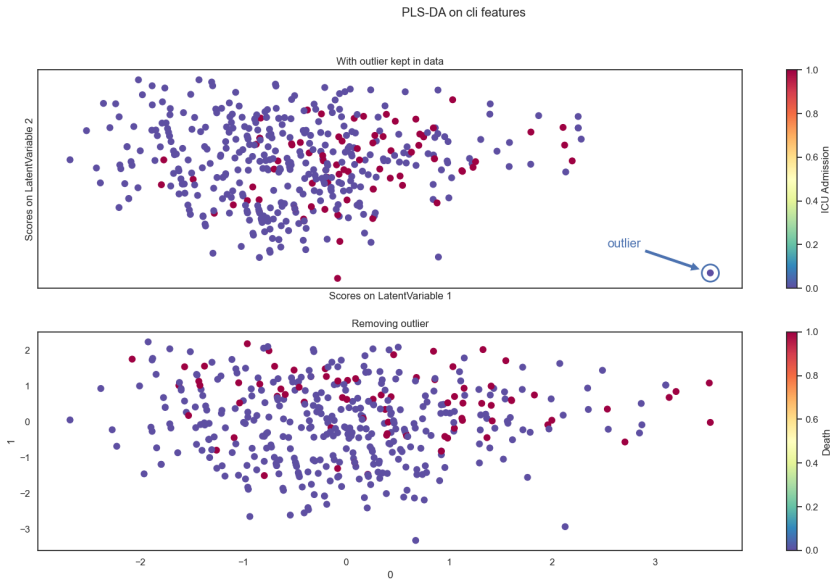


Figure 5.14: PLS-DA predicting on death(bottom) coloured with death and ICU admission(top) on clinical features

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

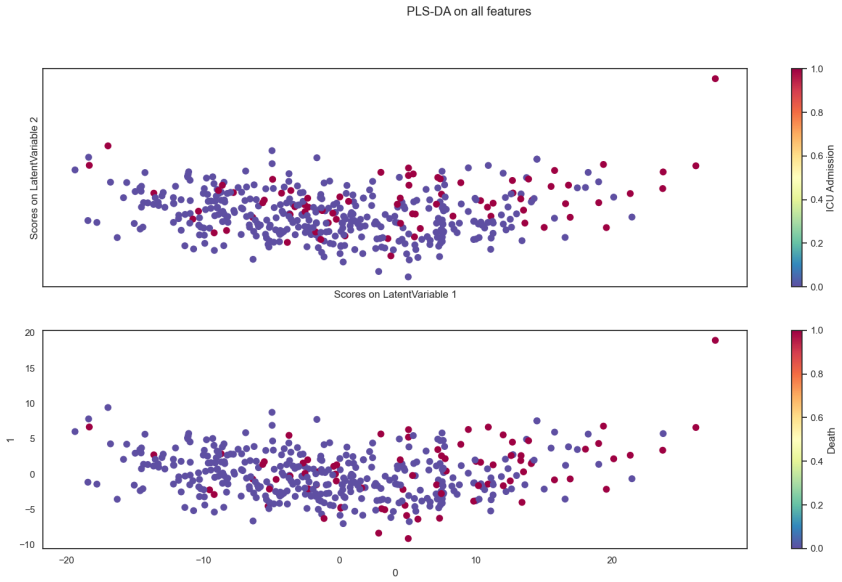


Figure 5.15: PLS-DA predicting on death coloured with death(bottom) and ICU admission(top) on all available features

Chapter 6

Bibliography

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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