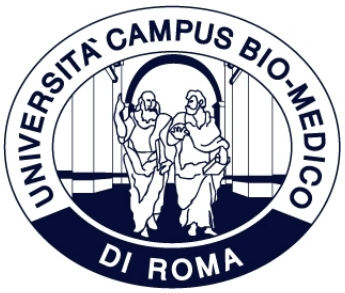


Università Campus Bio-Medico di Roma

Doctor of Philosophy in
Scienze e Ingegneria per l'Uomo e l'Ambiente/
Science and Engineering for Humans and the Environment
XXXIV cycle a.a. 2018-2019.

Computational Methods to Boost Radiomics

Natascha Claudia D'Amico

Academic Supervisor:
Prof. Paolo Soda

Company Supervisor:
Dr. Sergio Papa

February 28^{th}, 2022

PhD developed at Centro Diagnostico Italiano S.p.A.

My sincere thanks to Centro Diagnostico Italiano, particularly Dr Sergio Papa, for the opportunity to do my PhD in-house. I am also extremely grateful to my supervisor Prof. Paolo Soda for the valuable guidance and Prof. Giulio Iannello for his support. I would like to thank my colleagues Giovanni, Deborah, Isa, Patrizia, Ermanno, Rosa, Valerio and Lorenzo, who have motivated and helped me along the way.

Contents

Abstract

In the field of medicine, *radiomics* is a method that extracts a large number of quantitative features from medical images with the ultimate goal to enhance the prognosis of patients. Since the term radiomics was coined in 2012, its research efforts has been growing exponentially, fuelled by the ambition to move towards more personalised medicine and thanks to technological development in the hardware and software of medical scanners, as well as to advances in artificial intelligence. This thesis explores different aspects of the radiomics workflow with the aim of finding techniques that improve the results and stability of this method. In details, we investigated here: the development and introduction of new features, available solutions to cope with imbalanced learning, the combination of deep learning and machine learning techniques, and the influence of segmentation on model performance. The results shed light on ways to improve the standard radiomics workflow, by modifying the standard procedures.

Keywords: Radiomics, Machine Learning, Deep Learning, Imbalance Learning, Local Binary Patterns, multi-VOI analysis

List of Abbreviations

18F-FDG 18F-fluorodeoxyglucose
ADASYN Adaptive synthetic sampling approach
CAD Computer-aided detection
CBCT Cone bBam Computed Tomography
CNN Convolutional Neural Network
CNS Central Nervous System
CT Computed Tomography
CTV Clinical Target Volume
CXR Chest x-ray
DBN Deep Belief Networks
DCE Dynamic Contrast-Enhanced
DFS Disease Free Survival
FCN Fully Convolutional Network
GLCM Grey-Level Cooccurence Matrix
GLRLM Grey-Level Run-length Matrix
GLSZM Gray-Level Size Zone Matrix
GTV Gross Tumour Volume
HCC Hepatocellualr Carcimoma
ICC Intrahepatic cholangiocarcinoma
LBP Local Binary Pattern
ML Machine Learning
mpMRI Multiparametric Magnetic Resonance Imaging
MRI Magnetic Resonance Imaging
MVI Microvascular Invasion
NGTDM Neighborhood Grey-Tone Difference Matrix
NSCLC Non-Small-Cell Lung Cancer
OS Overall Survival
PCA Principal Component Analysis
PET Positron Emission Tomography
PFS Progression Free Survival
PTV Planning Target Volume
RFM Radiomic Feature Maps
ROI Region of Interest
RQS Radiomics Quality Score
SMOTE Synthetic Minority Over-Sampling Technique
VLBP Volume Local Binary Pattern
VOI Volume Of Interest

List of Figures

List of Tables

Chapter 1

Introduction

The last decade has been characterised by a major revolution in medicine and its applications, manifested in a shift from a reactive to a proactive approach. While in the past the focus was on treating diseases, it is now on maximising individuals' health. This new form of medicine is predictive, personalised, preventive and participatory (P4 medicine), as Hood and Friend describe [3]. Traditionally, most medical treatments were designed for the average patient, with a "one-size-fits-all" approach, even if it was apparent that patients respond differently to the same treatment. In response to this, medicine is moving towards a more predictive approach, in which patients are stratified based on their individual differences, e.g., genes, environment or family history. This permits to develop tailored therapeutic strategies for each individual patient. After the completion of the Human Genome Project (HGP) in 2003 [4], scientists have started to use *genomic methods* in their research [5]. Subsequently, researchers have developed several other '-omic' disciplines, such as proteomics (analysing information about produced proteins), phenomics (analysing information of mutational phenotypes) and epigenomics (analysing the complete set of methylation alterations in the genome) [6].

Following the development of the '-omic' sciences, in 2012 Lambin et al. [7] introduced the concept of *radiomics*. Radiomics is based on the idea that medical images contain more information than physicians can detect, so that medical images have started to play a greater role in the personalisation of therapeutic treatments. Since the acquisition of medical images is a non-invasive procedure, there is the possibility to use them not only assess the clinical condition of patients, but also to discover information used to guide the treatment and, in general, to predict the prognonis. While radiomics was initially based on Machine Learning methods, it later also made use of Deep Learning methods. Both Learning techniques are Artificial Intelligence methods, developed to simulate human intelligence by machines. After Aerts et al. [8] published the first article on using for prognosis purposes, many researchers followed the lead and focused on applying radiomics to clinical problems. Nowadays, radiomics applications are developed especially in the field of oncology, to predict

the evolution or treatment response of specific tumours, but applications have also been developed in non-oncological fields. Furthermore, initially radiomics was centered on prognosis problems as introduced by Aerts et al., but due to its popularity in the medical society, some research groups applied radiomics methods to diagnosis problems replacing the concept of Computer Aided Diagnosis which was used until then. The analysis of the literature presented in section 2.2 and section 2.3 shows that, even if many studies focused on different radiomics applications, the technical methodologies used during the radiomic workflow were similar. In general the radiomic workflow constist of the following steps: the segmentation of the area of interest, the quantitiative features extraction and selection and, finally, the analysis, which usually consists in developing a classification or regression model. As reported in chapter 2, since the introduction of radiomics, the literature follows the techniques introduced by the first published paper, without evaluating substantial innovations.

To address this issue, the aim of this thesis is to offer several improvements to the radiomics workflow. Firstly, the thesis analyses quantitative features that are commonly used in radiomic applications. Two alternatives to the standard radiomics features are proposed; the first analyses the value of features extracted from feature maps, while the second introduces the *Local Binary Patterns*, wellknown in computer science [9], but never used for medical applications. Secondly, to tackle the undesirable effect that imbalanced class distribution can have on classification results, the thesis analyses two approaches to learn under this condition. The first consists in an over-sampling technique, which generates synthetic samples belonging to the minority class. The second biases the learning process by using an error-cost approach to guide the classifier. Thirdly, this thesis discusses how to combine automatic features computed by deep network with shallow learners, showing that in the case of datasets with reduced size this could be beneficial with respect to use a deep network only or handcrafted features. Finally, the thesis analyses two ways in which segmentation can influence model performance. The first compares the result of an automatic segmentation and a manual segmentation of chest X-Ray images. The second analyses the differences between three segmentation methods commonly used in radiotherapy (Gross Tumour Volume, Clinical Target Volume and Planning Target Volume).

The research carried out used datasets of patients with Non-small-cell lung cancer studied in chest Computed Tomography images, patients with acoustic neuroma studied with Brain Magnetic Resonance Imaging and patients with COVID-19 studied with Chest X-Ray images.

This work is organised as follows. Chapter 2 consists of three parts: section 2.1 defines the concept of radiomics and its workflow, section 2.2 provides an overview of the literature on clinical applications, and section 2.3 provides an overview of

the technical aspects used in the radiomic literature. Chapter 3 describes the aforementioned four contributions to boost the radiomic workflow: section 3.1 describes the development and introduction of new features; section 3.2 describes the two approaches to address imbalanced learning; section 3.3 introduces the combination of Deep Learning and Machine Learning algorithms, whilst section 3.4 analyses how segmentation influences model performance. Finally, Chapter 4 concludes and offers future perspectives, while Appendix A summarises of all the other papers published during the course of this PhD not already mentioned in this thesis.

Chapter 2

State of the Art

The term *radiomics* has been introduced by Lambin et al. in 2012 in *Radiomics: Extracting more information from medical images using advanced feature analysis* [7], inspired by the already known term *radiogenomics*[1]. The authors highlight that future medical treatments will be directed toward personalised medicine using different data sources, such as demographics, pathology, radiology and genomics data. The idea of the authors was that imaging data as a source could contain complementary and interchangeable information to determine personalised treatments [10]. Moreover, medical images, such as CT, MRI and PET were routinely used during medical practice for cancer management, prediction, screening or follow-up during screening and are therefore available for patients. Furthermore, in recent years the information content of medical images has improved thanks to innovations in different fields. Indeed, hardware and software have improved due to the innovation and refinement of imaging contrast agents, the standardisation of acquisition protocols, and the innovation in imaging analysis techniques. All these innovations lead to the introduction of the *radiomics* analysis.

Radiomics was defined as the automatic high-throughput extraction of large amounts (200+) of quantitative features of medical images to support the prognosis of patients. The hypothesis is that quantitative analysis of medical image data through automatic or semi-automatic software of a given imaging modality can provide more and better information than a physician. This is supported by the fact that patients exhibit differences in tumour shape and texture measurable by different imaging modalities [7].

Even though the usage of radiomics has gained increasing popularity in the last years, the quantitative features used were first proposed and defined in 1973 by Haralick et al. [11].

Radiomics and artificial intelligence applications applied to medical images have increased in the last decade due to the non-invasive characterisation of diseases

[1]The first article published and available on Scopus containing the keyword *radiogenomics* has been published in 2003 [5]

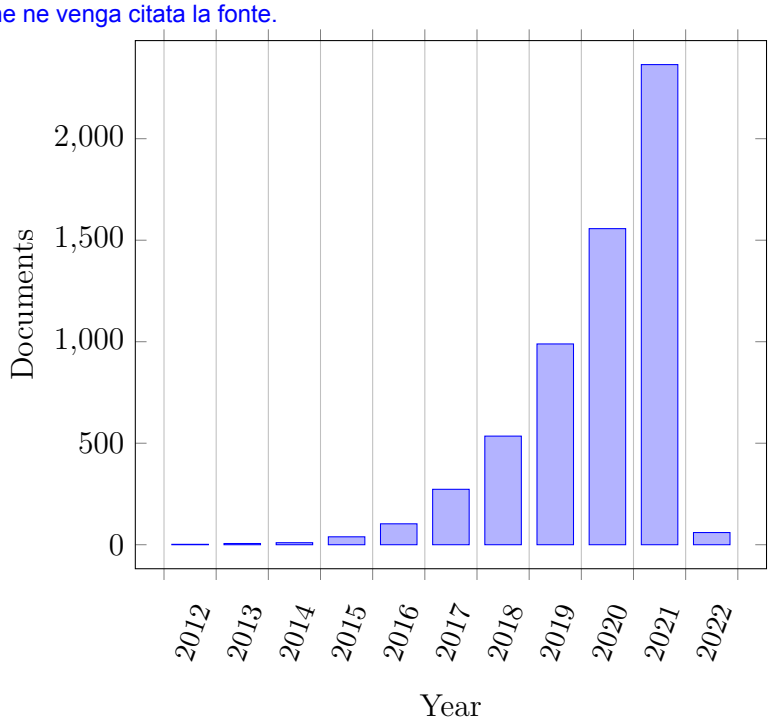


Figure 2.1: Amout of published articles for each year
Figure obtained from "Scopus" using as keyword "radiomics" updated on the 23 December 2021.

and tissues as described by Sollini et al. [12] and as shown in Figure 2.1. Furthermore, Sollini et al. analysed the clinical issue most commonly addressed with radiomics or artificial intelligence methods. Of the analysed 171 articles, 147 addressed oncological problems while 24 addressed non-oncological problems. The cancers most frequently analysed were Brain, Lung, Gastrointestinal and Breast tumours.

In this chapter, the first section (2.1) will focus on radiomics with its definition and workflow. The second section (2.2) will deepen the radiomics literature distinguishing applications to different anatomical districts. Finally, the last section (2.3) will handle and resume the methodological aspects developed in the radiomics literature.

2.1 Radiomics

2.1.1 Definition of Radiomics

As introduced at the beginning of the current chapter, radiomics is defined as the high-throughput extraction of a large number of imaging features extracted from medical images [7]. The central hypothesis is that quantitative analysis of medical images using semi-automatic or automatic software can obtain more information

than visual inspections performed by physicians. The hypothesis is based on the idea that these imaging features have the potential to capture distinct phenotypic differences of tumours and have great prognostic power, thus improving clinical significance across different diseases [13]. Therefore, patients with differences in the tumour shape and texture can be more easily recognised by the quantitative features than by the radiologist. Radiomics focuses on optimising unsupervised quantitative imaging feature extraction through a mathematical algorithm based on intensity, grey-level intensities and texture-based features, followed by developing decision support systems to estimate patient risk and improve individualised treatment selection and monitoring accurately. Figure 2.2 highlighted in yellow some examples wherein the clinical pathway routine, radiomics could be used, and its potentiality in clinical practice.

As shown in Figure 2.2 clinician could be supported by a radiomics approach in detecting and characterising suspected lesions. As a second step, the decision making step could be supported by suggesting an observation or treatment path. Furthermore, radiomics analysis could help in the treatment planning, and the outcome prediction could be indicated when integrated with data from different sources such as molecular or histological data. Finally, radiomics optimisation can support clinicians to diagnose or predicting the patients' outcomes correctly.

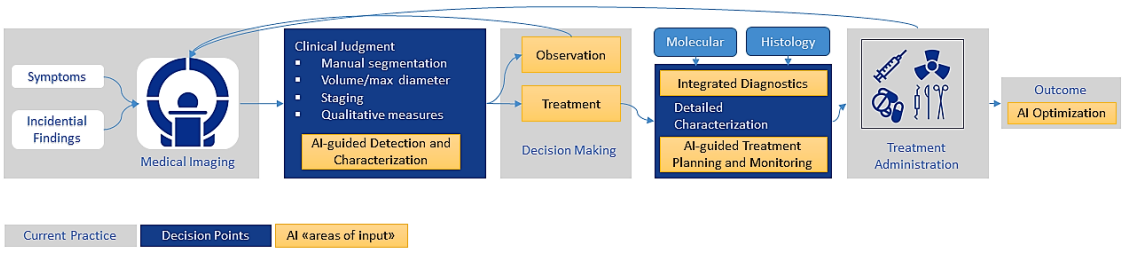


Figure 2.2: Current Practice and Decision Points of the current clinical pathway. In yellow possible radiomics input are evidenced to show how radiomics can help in the different clinical pathway steps. Figure taken from Bi et al.[14].

2.1.2 Radiomics Workflow

The radiomics workflow consists of the steps shown in Figure 2.3. The first step is the database acquisition, collection and reconstruction. The collected images were segmented to define the Region or Volume of interest, where usually this area coincides with the tumour regions. Finally, quantitative features are extracted from the segmented regions and then used to obtain a diagnosis or patient prognosis.

Figure 2.3: The radiomics workflow.

2.1.2.1 Image acquisition and reconstruction

The imaging acquisition step is the first step of the radiomics workflow. Even if the following manuscript handles only MRI and chest X-ray images, for the sake of completeness, a global overview of all commonly used image modalities for radiomic applications was made:

- *CT*: As it will be shown in section 2.3.1, CT images is the modality most commonly used for radiomics purposes [15]. CT images are widely used since tissue density, shape and texture of tumour and lymph-nodes are assessed. Although, images quality depend on the parameters used during acquisition and on the reconstruction algorithms as described by Kumar et al. [10]. Parameters that affect the image quality are the slice thickness, the axial field of view and the reconstruction matrix, or the Hounsfield units. For example, decreasing the slice thickness reduces the slice's photon statistics, increasing the image noise. Moreover, the axial field of view and the reconstruction matrix size determines the pixel size and the spatial sampling having an impact on the heterogeneity of the image [16]. Despite the differences, vendors develop similar algorithms to have comparable image qualities. A possible solution to compare features extracted from CT images that are used for radiomics purposes is the usage of a phantom to test the effect of different scanners and acquisition parameters [16]. Alternatively, radiomics studies investigated the robustness and stability of features extracted from CT images applying test-retest studies [17].

- *MRI*: MRI images intensity values do not depend directly on the tissue density, as in CT images, but depends on the tissue properties, such as the relaxation times and their response in certain conditions. Therefore, MRI image quality depends on the acquisition parameters such as the field of view, the field strength and the slice thickness. The Radiological Society of North America defined a *Quantitative Imaging Biomarkers Alliance* who defined a standard protocol for acquisitions defining the acquisition parameters [18]. However, MRI sequences such as *Diffusion-weighted imaging (DWI)* and *Dynamic contrast-enhanced (DCE) MRI*, allow the assessment of physiological tissues

properties. The first shows the apparent water diffusion coefficient that is related to the tissue cellularity, while the DCE shows the vascular flow, the permeability and volume fractions due to the contrast agent [10]. Although both sequences represent quantitative information, their reproducibility is still dependent on the acquisition parameter.

- *PET/CT*: The last commonly used imaging technique is the PET/CT image. PET images used for radiomics purposes are usually acquired with the radiotracer 18F-fluorodeoxylglucose (18F-FDG) since most of the malignant tumour types exhibit a high glycolytic rate [19]. PET and CT images are usually combined because CT images give information about the patients' anatomy, while PET images give information about the functionality of the organs. Features extracted from PET images are the most difficult features to compare since they depend on the calibration of the scanner and the acquisition protocol and the patient condition (blood glucose level, uptake period, breathing, and inflammation). Inter-institution cross-calibrations are necessary before quantitative analysis of the images.

- *Other image modalities*: Besides the three imaging modalities most frequently used for radiomics applications, the following imaging modalities were also used for radiomics applications.

  - *PET/MRI*: PET/MRI technology is a quite new technology not commonly installed and available compared to PET/CT [20]. Due to the small numbers of worldwide available PET/MRI machines, collecting databases that can be used for radiomics applications is still an issue. Similarly to PET/CT, PET/MRI combines the anatomical information obtained with the MRI with the functional information of the PET image. Even if the imaging technology is quite new, radiomics applications have already been developed for example by Antunes et al. and by Chen [21, 22].
  - *Ultrasound*: Ultrasound imaging is the second most commonly used diagnosis modality [23], due to the ability in distinguishing between benign and malignant lesions, using safe and non-ionising technologies [24]. Although, ultrasound images are highly affected by acquisition variability caused by the operator experience and the used technology [24]. Consequently, the repeatability of the images is affected and hence the extracted features, reducing the number of radiomics applications using this imaging methodology.
  - *X-Ray*: X-Ray imaging is the most common medical imaging acquired in medical practice [23]. Despite the usage of this imaging technique, ra-

diomics applications are not very common since X-ray images were usually not used for oncological issues. In the literature, the most common applications concerns mammography imaging for breast cancer detection and lung x-ray imaging for COVID-19 diagnosis.

– *Photon-counting CT (PCCT)*: Photon-counting CT, introduced in clinical routine in 2021 [25], represents a great innovation for the CT scanners due to the single one-step conversion of X-ray photons into an electrical current that generates the medical image instead of the two steps, necessary in the standard CT images. Given the novelty, radiomics applications have been developed on this new imaging modality to study the differences from the original CT images [26].

2.1.2.2 Segmentation

Radiomics features are usually extracted from a *segmented* area within the collected dataset. Segmentation is defined as the contouring of the region of interest (ROI) or volume of interest (VOI) such as a tumour, an anatomical structure or normal tissue. There are three main segmentation methods: manual, semi-automatic, and automatic. The first, the manual segmentation, is obtained manually by expert readers and is often treated as ground truth [10]. Even if this segmentation is considered as the gold standard, this method commonly suffers from high inter-variability and is time-consuming and labour-intensive [27]. To limit the variability, different physicians could perform different segmentations, which resolves the inter-variability problem but increases the dedicated time. Moreover, the segmentation variability introduces a bias in the radiomic feature extraction and evaluation since they highly depend on the used segmentation. Balagurunathan et al. [28] compared the segmentation obtained manually and by an automatic software, analysing the feature stability and the similarity of the segmentation. The authors showed that in large tumours, the variability decreases compared to small tumours. Moreover, in section 3.4.2, the segmentations' influence on the results has been studied, highlighting the dependence between segmentation, features, and results.

Since in radiomics, large datasets are necessary, causing a high number of necessary segmentations, an alternative segmentation method must be found. The alternatives were the automatic or semi-automatic segmentation methods developed for different imaging modalities and anatomical regions. These segmentation methods can be distinguished into four techniques, based on the used methodology to distinguish areas within an image [29].

- *Threshold-based technique*: The Threshold-based technique is the simplest method because it relies on the pixel/voxel intensity value distinguishing pixels using one or more threshold values. However, due to the algorithm's simplicity,

this method does not work well with complex images, especially with images without enhanced tumour areas.

- *Region-based technique*: The Region-based segmentation method groups pixels/voxels with homogeneous properties according to a predefined criterion [30]. The most common segmentation method belonging to this technique is the region-growing method. This method starts from a pixel chosen by the reader and adds automatically neighbouring pixels that respect a similarity criterion [15]. Unfortunately, this method suffers if the image contains too much noise, leading to a wrong segmentation [10].

- *Model-based technique*: This third introduced method builds a model based on parameters or on geometry. Based on different rules, the model can improve the segmentation border of the wished area. An example is the level-set method. Based on a level-set function, this method ideally evolves by finding the contour equation.

- *Pixel/Voxel Classification Techniques*: The last segmentation method classifies each pixel/voxel to a specific class, classifying the different areas of the image. This type of classification is usually fully automatic and two examples will be discussed. The first is the semantic segmentation, where the model assigns each voxel to a specific class of what is being represented, for example: people, building, street. Two examples are the fully convolutional network (FCN) model first applied by Long et al. [31] and the U-Net [32] . The second method is the instance segmentation, where the model needs to classify correctly each voxel to a class, but it also assigns the voxel also to the specific instance, for example person 1, person 2 or person 3. Instance segmentation is preceded by a object detection step due to the necessity to recognise all objects before classifying them. An example is the well known Mask R-CNN [33] which classifies and localised objects using a bounding box followed by a semantic segmentation that classifies each pixel into a set of categories. Applications of this method in medical imaging have been developed by Anantharaman et al. [34] to segment oral lesions.

Semi-segmentation methods are based on user input, while the utterly automatic segmentation method is based on Deep Learning algorithms, for example, the U-Net [35], that automatically segments the ROI or VOI. Despite the differences, all algorithms should be as automatic as possible, time-efficient and accurate.

2.1.2.3 Feature extraction

After the segmentation, the following step in the radiomics workflow is the feature extraction and qualification. These features describe the tumour characteristics such

as the tumour intensity histogram, the tumour shape, and the texture. The following items describe the most commonly used feature classes in a radiomics workflow:

- *Shape-based features*: Features belonging to this class describe the 2D or 3D shape of the segmented area. For example, features can describe the total area or volume of the ROI/VOI, the global shape of the area, if roundish or speculated, and the compactness or shape of the area.

- *First-order histogram*: Represents the distribution of intensities of the data within the segmented area in a single histogram. The histogram describes the range of voxel values that can be Hounsfield units in CT images, SUV values in PET images or signal intensity in MRI resulting from the signal equation used. Different statistics can be obtained from the histogram, such as mean, max, median, min values and range, skewness, and kurtosis values. The latter represents the degree of histogram asymmetry and sharpness. Moreover, the uniformity and the entropy are extracted and represent the inhomogeneity of the selected area. More complex values can be extracted if needed.

- *Texture based features*: Texture based features, also referred to as second-order histogram features, refers to the spatial variation of the intensity level within the segmented area and were introduced in 1973 by Haralick et al. [11]. These features are based on different support matrices obtained from the original image and deepened in the following overview[10, 15]:

  - *Grey-Level Cooccurrence Matrix (GLCM)*: The most common feature group are the Grey-Level Cooccurrence Matrix (GLCM) features. These features are based on a joint conditional probability function obtained counting the number of times a specific pixel, or voxel if 3D, with value x is close to a pixel with value y separated by a distance d in direction a. The output matrix size depends on the intensity levels within the original image. Consequently, from these new functions, features can be extracted describing characteristics such as the entropy, related to the heterogeneity, the energy, describing the homogeneity of the image, the contrast, which measured the local variation or the correlation, the cluster prominence, the cluster shade or the cluster tendency.

  - *Gray Level Run-Length Matrix (GLRLM)*: The GLRLM is a 2D matrix, that describes the times each element of intensity j of the original matrix appears consecutively to the element of intensity i in a specified direction as defined by Galloway[36]. Moreover, the matrix is based on the Grey level run defined as the length of consecutive voxels having the same intensity value. Features as Short Run Emphasis, Long Run Emphasis,

Short Run Low/High Gray Level Intensity and Long Run High/Low Gray Level Intensity are calculated based on this matrix.

- *Gray Level Size Zone Matrix (GLSZM)*: The GLSZM is a 2D matrix [37] where in column j and row i the number of areas of voxel with the same grey level i with size j are stored, resulting in wide and flat matrices if the ROI/VOI is homogeneous. Features such as Small/Large Zone Emphasis and Low/High Grey Level Zone Emphasis resulted from this matrix.
- *Neighborhood Gray-tone Difference Matrix (NGTDM)*: The last commonly used matrix is the NGTDM, which contains the summation of the differences between all pixels with the considered grey tone and the average value of their surrounding neighbours for each entry. Features obtained from this matrix include coarseness, contrast, busyness, complexity and texture strength.

- *Other Features*: Further features can be extracted using a filtered image as the input image. The most common filters used for radiomic application were the wavelet filter or the gaussian filter, both implemented in the library *pyradiomics* widely used for radiomics applications. The first yields eight decompositions per level (low and high pass filter), while the second is an edge enhancement filter. The advantage of applying filters on images before extracting the features is that filters, such as wavelet or gaussian filter, highlight information hidden in the images that could provide valuable information.

- *Automatic Feature*: The last type of features used for radiomics applications are the features automatically extracted by Deep Learning algorithms. The most common and efficient deep learning algorithm applied to images is the convolutional Neural Network (CNN) [38]. These Networks are based on different layers: Convolutional Layer, Pooling Layer, and Fully-Connected Layer. The Networks can have different combinations of these layers, but all of these models reduce the entire image into a single vector of class scores. The obtained single vector can be seen as the list of automated features since it comprises quantitative values describing imaging characteristics. Briefly, the convolutional layer consists of several filters, for example with size $5 \times 5 \times 3$ that are slide across the entire input volume computing a dot product between the filter and the input image at the specific position. As a result, an activation map is obtained, containing the response of the image to that specific filter. The filters usually focus on visual features such as edge detection or orientation detection. For each convolutional layer, a set of filters is used (for example, 12), and each analyse a different aspect of the original image, resulting in 12 different activation maps obtained from the original image stacked and passed to the next layer as a 3D volume [39]. After the convolutional layer,

the pooling layer has the function to reduce the dimensionality of the stacked images. After repeating this process many times, a last flatten layer reduces the obtained data into one vector representing the automated features.

2.1.2.4 Feature selection

As introduced before, many quantitative features can be extracted, resulting in thousands of descriptors for each image, and usually, their number exceeds the dimension of the total dataset. Since features can be redundant and correlated, a high number of features can lead to overfitting problems [2]. Therefore, a feature selection step is necessary to reduce the number of features keeping the significant and relevant features for the given task. Feature selection usually increases classification performances and lowers computational costs deleting irrelevant features, those who cannot help with the classification, redundant features, those who give the same information as other features, and noisy features, those who may be relevant but due to the introduced noise they may be less useful than others.

Feature selection can be distinguished into three categories of label availability: the unsupervised, the supervised and the semi-supervised features selection [41].

The unsupervised features selection is generally used for clustering tasks since the data label is unknown, and feature selection cannot be based on the class distribution. The supervised feature selection method is mainly used for classification tasks. The availability of the labels allows the algorithm to identify and select irrelevant, redundant and noisy features effectively. Finally, semi-supervised feature selection methods were used if only part of the dataset was labelled. In this condition, an unsupervised method is not recommended since it does not use important information such as the label, while the supervised method necessarily needs to exclude all unlabelled data losing information. On the other hand, feature selection can be made through three different search strategies, the filter selection, the wrapper, and the embedded feature selection [42]. Note that each of this category contains unsupervised, supervised or semi-supervised methods.

Filter, wrapper and embedded feature selection are now briefly summarised following three reviews [43, 44, 45]. Nevertheless there are other feature selection methods, such as clustering or Principal Component Analysis (PCA), described by Rizzo et al. [16] and which will not be deepened here.

- *Filter Feature selection*: This feature selection approach selects features based on their characteristics without using any learning algorithm. Features are ranked based on specific characteristics, and subsequently, features with the

[2] **Overfitting definition**: "A model overfits the training data when it describes features that arise from noise or variance in the data, rather than the underlying distribution from which the data were drawn. Overfitting usually leads to loss of accuracy on out-of-sample data".[40]

highest ranking are chosen. This method is very efficient but has the disadvantage of losing important features since no learning methods are used. The most common filter feature selection approach used is the minimum redundancy maximum relevance (mRMR), where mutual information (MI) compares the outcome and the single features. Features with the maximum MI are selected [15, 46]. However, the analysis of single features with the outcome is defined as univariate selection and has the disadvantage of not considering the relationships within the features. Another commonly used feature selection method is the *Relevance in estimating features* (RELIEF) method which analyses the inter-dependent features. The core idea was to rank the features according to how well attributes can distinguish data that are next to each other [15, 47]. Different variations of this method were developed, such as the ReliefF, which deals with multiclass problems, or the RReliefF algorithm developed for regression problems. However, even considering the advantages of this method such as the time reduction or the complexity reduction, the method has shown to be unable to detect redundant features [15].

- *Wrapper Feature selection*: The wrapper feature selection approach, differently from the filter selection approach, uses a learning algorithm to detect the ideal feature subset. The chosen learning method analyses different subsets of features, adjusting the included features based on the performance evaluation to find the best combination that maximises the model performance [48]. The feature set with the best performances will be chosen as the final feature set. Wrapper feature selection methods can be distinguished into three types, the forward selection method, which iteratively adds features to the final feature set, the backward elimination, which started including all features all iteratively excluded features and the step-wise selection, which is based on a combination of the forward selection and the backward elimination methods. An example of wrapper feature selection method belonging to the backward elimination method is the *Recursive Feature Elimination*. This method introduced by Guyon et al. [49], begins using all available features and finding an importance score for each feature. In the second step, the worse feature is excluded, the model is built again, and the importance score is calculated for all remaining features. This procedure is repeated until the model performance is maximised.

- *Embedded Feature selection*: The last introduced feature selection method includes the feature selection into the model construction. This helps to include the advantages of the two previously introduced methods, the filter method's computational efficiency, and the interaction with the learning model of the wrapper approach. Unlike the wrapper approach, in the embedded feature

selection, the selection is made directly during the training, without further feature evaluation.

2.1.2.5 Analysis

Since a radiomics analysis aims to develop models to diagnose or prognosis patient outcome, the development of machine learning or deep learning models is necessary. There are many machine and deep learning methods, but all learning models require observations or data, known as features, to analyse the dataset's hidden pattern. Different machine learning perspectives will be introduced in this section, followed by a deep learning algorithms introduction.

Machine Learning Machine learning algorithm can be subdivided into four main categories, the supervised, the unsupervised, the semi-supervised and the reinforcement learning.

- *Supervised learning*: As already introduced in the feature extraction paragraph, supervised learning needs to have a labelled dataset describing the label, for example, overall survival, or the class of the data, for example benign/malignant. Within the supervised learning models, two different groups can be distinguished, the classification task, which needs a discrete outcome such as categories, and the regression task, where a continuous outcome, for example, Overall Survival is needed. Both methods use a training set to learn the data and the corresponding outcome and produce an inferred function to predict the test set where the label is unknown. Therefore, the model assigns a new target to the data in the test set and the comparison between the real label and the assigned ones determine its performance. The most common used classification method is the *Logistic Regression* (LR) [15]. LR is a classification method used with binary categorical variables, for example malignant or benign tumours. LR uses a logistic function to get a binary output variable. Furthermore, LR's range is limited to values between 0 and 1, meaning that all input values are transferred using the LR function to values between 0 and 1 using the following equation: $Logistic function = \frac{1}{1+e^{(-x)}}$. Finally, all values under 0.5 are assigned to class 0, while all values above belong to class 1 [50]. The second commonly used classification method is the *Random Forest* (RF). This method is based on multiple independent *Decision Trees*, a classification methods based on *if-then* reasoning, that are trained on a random subset of the data [51]. The subsets of the dataset are obtained using a bootstrap method. Each DT in the RF works in parallel on a different dataset and uses different subsets of available features, leading to unique DTs' within the RF. Finally, the RF aggregates the decisions obtained by the single DT models using a

majority decision. *Support Vector Machine* (SVM) is the third most common used classification method [15]. The main idea behind this technique is to find a hyperplane in an n-dimensional space that classifies the data contained in the dataset. Data that falls on a different side of the hyperplane belong to different classes. Furthermore, the hyperplane needs to have the maximum margin of all points of the classes, because this can increase the confidence that future points will be classified correctly. The hyperplane has many dimensions as the number of features and can be represented by linear, gaussian or other functions. Besides the three introduced ML classifiers, summarised in the review written by Avanzo et al. [15] and commomly used for radiomics purposes, many other algorithms are available and can be summarised in the following groups [52], (i) *Regression Algorithms*, which models the relationship between variables, iteratively refined using a measure of error and the models prediction, (ii) *Instance-based Algorithms*, that are based on specific instances that are deemed important and generalised based on similarity measure, (iii) *Regularisation Algorithms*, who are an extension of other methods and penalise complex models, favoring simple and good generalising models, (iv) *Decision Tree Algorithms*, that construct models based on decisions made on values of the data, (v) *Bayesian Algorithms*, who apply the Bayes' Theorem and (vi) *Ensemble Algorithms*, that are made of multiple weaker models that are independently trained followed by a combination of the single predictions.

- *Unsupervised learning*: Oppositely to supervised learning, no label is furnished in unsupervised learning. This method is often used for data clustering since it analyses the dataset's relationship and hidden structure. Also, unsupervised learning can be subdivided into two tasks, the clustering and the association task. In the clustering task, the aim is to divide the dataset into groups based on specific feature characteristics, while in the association task, the aim is to find association rules within the dataset. Some examples of models that belong to this family are K-means and DBSCAN, used for clustering purposes, and apriori used for association tasks. K-mean, the fastest and easiest clustering model, starts with k randomly positioned centroids and recursively adjust their position to have the smallest incluster sum of squares [53]. The number of k is usually furnished by the user in charge to find the correct number of clusters within the dataset. The second mentioned model, DBSCAN searches for each data, how many instances are located within a small distance ϵ. If at least n samples are in the ϵ-neighbourhood, the region can be considered a core instance. All data within the same core instance belong to the same cluster, and more than one neighbouring core instance can also belong to the same cluster. Since association tasks are barely or never used for radiomics purposes, the apriori method will not be deepened. The models can be summarised

in the following groups [52]: (i) *Clustering Algorithms* which describe the class of problem, (ii) *Association Rule Learning Algorithms* that defines rules describing the relation between variables in data.

- *Semi-supervised learning*: The semi-supervised learning, like the semi-supervised feature selection, has both labelled and unlabelled data. As described by Van Engelen and Hoos [54], semi-supervised learning uses the unlabelled data as a support for classification task, or uses the available data as a support during the clustering process.

- *Reinforcement learning*: The Reinforcement learning, not commonly used for radiomics purposes, learns how to behave through trial-and-error interactions with a dynamic environment, using estimated errors as rewards or penalty during the classification task [55]. The higher the error, the bigger the penalty and the smaller the rewards and reverse. This algorithms uses the errors and rewards to automatically determine the behaviour and maximise the performance.

Deep Learning Recently, Deep Learning (DL) methods applied to medical images have increased due to the availability of larger datasets and the increased computational speed [56]. DL applications obtained good results in object detection and recognition tasks, recognising imaging patterns and performing accurate segmentations [57]. Moreover, DL methods have the advantage of not needing preprocessing steps, such as tumour segmentation and feature extraction, but it takes as input the whole image as it is. The depth of DL models is not given by a *deeper understanding* of the image, but is referred to the model's depth, determined by the number of layers within the model [58]. A DL model can be represented by Figure 2.4 which shows the input image, the output, and the hidden layers. There are many different types of DL models, but all have in common the main structure shown in Figure 2.4, where the number and complexity of the hidden layers distinguish the deep learning models.

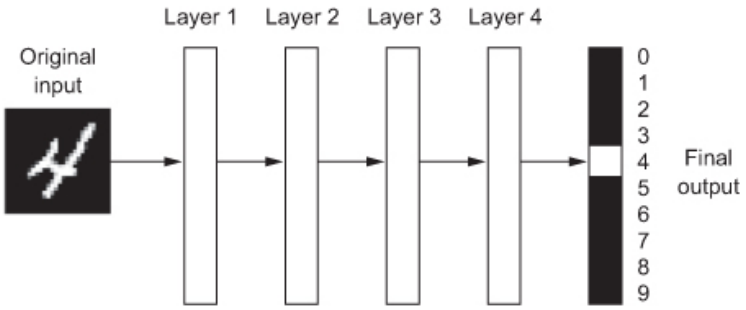


Figure 2.4: Overview of a DL Model
Figure taken from Chollet [58].

As introduced in section 2.1.2.3, the most common deep learning models are the convolutional Neural Network (CNN) due to their ability to handle 2D and 3D images. In Figure 2.5 an overview of the model construction is shown, showing that a CNN model is made by a first part, made of different types of layers, and a second part, represented by green and red dots, with a fully connected neural network.

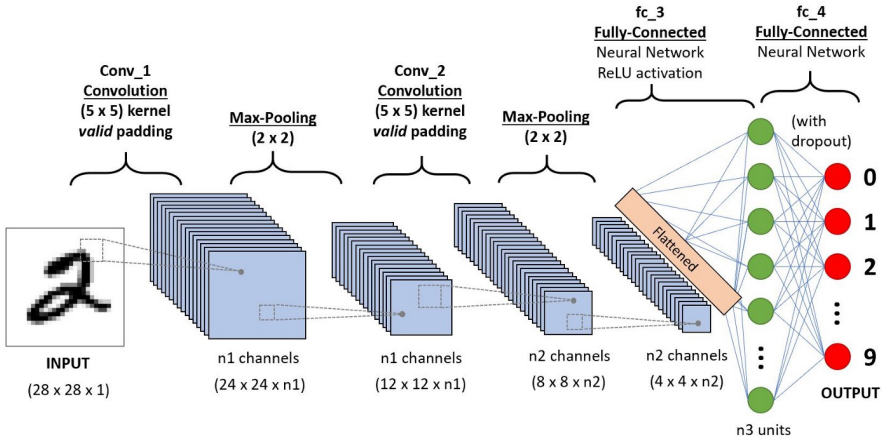


Figure 2.5: Overview of a CNN model structure
Figure taken from Saha [59].

The first part can be summarised as the feature extraction part, while the second is the classification part. The feature extraction part of the CNN model contains at least two types of layers, the convolutional layer already introduced in section 2.1.2.3 and the pooling layer, which is necessary since it reduces the feature map dimension according to specific rules. A third layer, commonly part of a CNN model, is the activation layer. This layer is necessary to add some non-linearity to the model and solve the Vanishing Gradient problem for profound models. The Vanishing Gradient problem is caused in profound learning models when the value of the gradient of the initial layers cannot learn correctly. The most common activation functions used in the activation layers were the Rectified Linear Activation (ReLU), the Logistic (Sigmoid) function and the Hyperbolic Tangent (Tanh) [60]. Finally, the second part of the model, made by a fully connected neural network, analyses the output of the feature extraction part, to obtain the data classification. The data is flattened before the fully connected neural network analysis. Some of the most recognised architectures, which belong to the CNN and are commonly used for medical applications, are Alexnet, VGGNet, ResNet, and InceptionNet [57].

VALIDATION METHODS Performance of the models is mainly represented by accuracy, sensitivity and specificity or by the recall, precision, F1-score and AUC values. The choice of the performance metric is based on the author and the data distribution. For imbalance datasets, as shown in section 3.2 the accuracy is not the

correct metric since it highly depends on the dataset distribution, while the balanced accuracy or the F1-score are appropriate metrics in these conditions. Moreover, performance is evaluated in different conditions, the validation performance and the test performance, obtained on the validation set and the test set, respectively [61].

To train and evaluate the model correctly, a validation process is necessary to split the dataset into training, validation and test. The most common internal validation method used, is the *leave-one-out* validation, where the whole dataset is used for training except a single element that is used for validation [15]. The procedure is repeated until every element has been in the validation set at least one time. Other commonly used validation methods are the bootstrap method and the k-fold validation method. The Bootstrap method consists in the extraction with replacement of the samples contained in the dataset to generate the training set, in which elements could be represented more than once. Data never chosen were assigned to the validation set. The second method splits the entire dataset into k parts, using one part as validation and the others as the training set repeating this procedure until all parts have been used as the validation set.

2.2 Clinical Applications

After the first radiomics application proposed by Aerts et al. [8] many articles analysed radiomics applications using Machine Learning or Deep Learning methods. Reviews of the radiomics literature can be subdivided into two types. The first analyses the literature from a clinical point of view, while the second analyses the literature deepening the technical aspects used. In this section, the first type of reviews will be reported and summed up, while section 2.3 reports the reviews deepening the technical issues.

Yip and Aerts [62], Bi et al. [14], Lee and Lee [63] and Sollini et al. [12] focused their reviews on the clinical applications of radiomics works using machine learning methods. In Table 2.1 an overview of the cited articles is shown, comparing the found potential applications of radiomics, the clinical areas of applications, the used images, factors that affect radiomic feature quantification, false positive discovery rate and proper study design and finally conclusion /limitation of the analysed studies. Yip and Aerts and Bi et al. introduced the potential applications of radiomics in a clinical environment summarising that radiomics has a great potential in detecting abnormalities, predicting the treatment response and the patients outcome, in tumour staging and identification and in the assessment of cancer genetics. All included reviews, except for Lee and Lee includes articles using CT, MRI and PET images while Lee and Lee deepen radiomics application only using PET/CT images. Sollini et al. reports that from the 171 considered articles, 86% of the works handled with oncology problems including brain (31%), Lung (25%), Breast (15%),

Gastrointestinal (14%), Urology (11%), Musculoskeletal (2%) and Skin (1%). The other 14% of the articles focused on not oncological problems such as infections or ophthalmological problems.

Clinical applications summed up by the included reviews showed that articles analysing the Central nervous system (CNS) focused on an accurate diagnosis and an extent of the disease. Another focus of radiomics applications is analysing the effect after treatments since the surrounding neural tissue may be affected by the treatment.

Radiomics applications on brain tumours considering PET/CT images are limited since the necessity to use adequate radiopharmacy that are not commonly used in clinical practice[3]. As on the other imaging techniques, these studies deepen the patient survival, the diagnosis value of the signature, and the tumour progression after radiotherapy. The second most studied topic is the Lung tumour. These works aim to improve the early detection, characterisation and staging of lung tumours, specifically the lung cancer screening, lung cancer characterisation, the assessment of the intra-tumour heterogeneity, the cancer subtypes and the response to target therapies.

Radiomics studies focusing on breast cancer aim to differentiate the breast cancer subtypes and predict the treatment response, such as the staging and prognosis prediction.

Yip and Aerts summed up the factors that affect the radiomics features quantification: the acquisition modes, the reconstruction parameters, the smoothing and segmentation thresholds, the reproducibility of the radiomics features, the image discretisation schemes, the Respiratory motion, and the Tumour size and intratumoral heterogeneity.

Finally, articles conclude that radiomics has a high potential but that not all features are recommended to be used due to the instability of the features. Furthermore, authors conclude that the dataset size has to be at least ten to fifteen patients per feature [62], that the database should be published to improve collaborations [14] and that PET/CT images have a lower spatial resolution compared to other imaging methodologies. Finally, that especially for small tumours, the reliability of heterogeneity of the parameters is limited [63]. In the following subsection, 2.2.1 to 2.2.7 reviews focusing on specific radiomics applications in clinical areas are deepened and summarised. As described previously, in literature, the most common radiomics applications were studied for the following tumours: brain, lung, oesophagus and gastric, hepatic, kidney and rectal, breast and prostate. For completeness, all applications have been summarised, even if the manuscript addressed only lung and brain applications. For each subparagraph a table summarising the technical

[3]The most common used radiopharmacy in clinical practice is FDG which has string limitations when used for brain oncological issues [64, 65].

evaluation and summary and the clinical applications of each article considered in the reviews is reported.

2.2.1 Brain Cancer

Radiomics and Machine Learning applications on brain tumours have been widely analysed and summed up by different revisions, such as the review written by Zhou et al. [66], who analysed radiomics application on general brain tumours, and Chaddad et al. [67], Lotan et al. [68] and Soni et al. [69] who studied radiomics application on glioblastomas. Focusing on the reviews written by Zhou et al. and Chaddad et al., the authors analysed and summarised the different steps of a radiomics workflow. They started from the role of clinical imaging as a prerequisite for radiomics models to the outcome of predictive analysis with machine learning techniques, deepening the main clinical applications: the survival time prediction, the classification of tumour subtypes and the tumour tissue discriminative analysis for general brain tumours. Moreover, Chaddad et al. focused on the prediction of clinical, proteomic (e.g., Ki-67 expression), genomic (e.g., IDH1 status) and transcriptomic characteristics, or the development of personalised treatments. Furthermore, the review summarised that different articles studied the discrimination between pseudoprogression and progressive disease. Soni et al. deepened the literature of radiomics applied to gliomas focusing on texture analysis. The authors compared the concepts and methodologies of texture analysis through various MR imaging texture analysis applications. Finally, Lotan et al. discussed the used resources, the segmentation methods and the machine learning methods and results of recent articles that focused on machine learning applications on gliomas. The first reported result was that since 2014, the number of papers that addresses the segmentation using CNN-based models has continued to improve every year. Comparing 33 articles, the authors found that about 30% analyses the IDH mutation, 21% of the articles analysed the survival time, and 18% analysed the histological grade of the tumours. 70% of the analysed articles used Machine Learning methods while 30% used Deep Learning methods, and the overall results obtained were in the range of 0.66- 0.96 the AUC value and 67% - 98% the accuracy value. The authors conclude by highlighting the challenge of ML and DL applied to gliomas.

Table 2.2: Overview of cited reviews focusing on radiomics applications on brain cancer

	Technical evaluation and summary	**Clinical Applications**
Zhou et al. [66]	Features Extraction: Quantitative Features, biologically inspired Features, Machine Learning Applications: Supervised, Unsupervised, Semisupervised applications, Deep Learning	Survival Time Prediction, Classification of Glioblastoma Subtypes , Tumour–Tissue Discriminative Analysis, Applications of Imaging and genomics
Chaddad et al. [67]	Image acquisition, Standardisation, Segmentation, Feature extraction and Analysis, Model Building	Radiomics applications on GBM, Intratumoural heterogeneity and radiogenomics, Prediction of clinical, proteomic (e.g., Ki-67 expression), genomic (e.g., IDH1 status) and transcriptomic characteristics, or the Development of Personalised Treatments.
Lotan et al. [68]	Dataset availability, Segmentation	Machine learning and Deep learning applications on Gliomas
Soni et al. [69]	Texture analysis and Feature definition and description	Texture analysis for Gliomas: grading, survival analysis, radiogenomics, miscellaneous applications

2.2.2 Lung Cancer

The first clinical application of a radiomics method was developed on lung images by Aerts et al. [8] in 2014. The application developed a radiomics algorithm applied to 788 patients with non-small-cell lung cancer and 231 with other head and neck cancers. For each image, radiomics features were extracted, and the patients' clinical and gene information were added to the imaging features. Finally, to prognose the overall survival, a radiomics heat map and a prognostic signature were developed. Since this first application, thousands of articles have been published regarding radiomics applications on lung images.

The following seven reviews summarised and analysed the vast literature of radiomics and machine learning methods applied to lung images. Ather et al. [70] described the need for the development of automatic nodule identification methods to support the radiologist in their activity due to the substantial variability between radiologists. Furthermore, besides the nodule identification task, the authors discussed the nodule segmentation task, widely studied by published articles, and the promising results mainly obtained by deep learning algorithms. The authors distinguished between two main targets, radiomics used for lung cancer detection, segmentation and characterisation and radiomics used for the treatment outcome prediction. The first target can be further exploited into the following tasks: Cancer detection, contouring, characterisation, segmentation of regions of interest in lung cancer, prediction of histology and tumour stage and tumour genotype classification. The second target can be exploited as complete response and local control prognosis, distant metastases prognosis, survival prognosis, response to chemo- and targeted molecular therapy, side effects prediction, and post-treatment recurrence identification.

Similar main applications have been observed also by Avanzo et al. [71], Chen et al. [72], Constanzo et al. [73] and Rabbani et al. [74]. Chen et al. discussed the most commonly used imaging techniques for radiomics application, concluding that the most common imaging modality for radiomics analyses is the CT. In contrast to CT, MRI and PET are used less frequently because even if they could give complementary information to CT images, standardisation is needed, or the analysis is affected by the different acquisition parameters.

Thawani et al. [75] decided to deepen the limitations of radiomics applications on lung cancer tasks and concluded that the most critical limitation is the lack of reproducibility of the biomarkers. Furthermore, the variability of acquisition parameters, for example, the contrast enhancement and convolution kernel, affect the diagnostic performance of the found radiomics signature. The authors described that to overcome these limitations, standard imaging protocol should be developed. Besides the limitations, Thawani et al. summarised a few results of the evaluated articles showing that the AUC value obtained with a patient survival analysis task using size, intensity, shape, texture and wavelet features is 0.6. In contrast, the best reported AUC result was 0.85 obtained with a recurrence-free analysis task using texture features.

The last reported review was written by Scrivener et al. [76] who compared 11 papers that used computed tomography images, 3 papers that used positron emission tomography and 8 papers that used PET/CT images. Again, the authors summed up the main radiomics tasks in classifying lung nodules and the prognosis of established lung cancer. Since the compared methodological issues were different in the reviewed papers, a direct comparison could not be made, but the authors identified only five studies out of the 22 analysed that were externally validated. Results reported in articles with classification tasks showed AUC values ranging from 0.56 to 0.981. Articles facing the Overall Survival issue reported AUC results in the range of 0.62 to 0.82. In contrast, the only reported article regarding recurrence analysis had an AUC value of 0.85 and the article that faced a staging task obtained an AUC value of 0.56. The authors conclude that radiomics analysis has great potential in improving diagnosis and prognosis tasks.

Table 2.3: Overview of cited reviews focusing on radiomics applications on lung cancer

	Technical evaluation and summary	**Clinical Applications**
Ather et al. [70]		Nodule detection, nodule segmentation, nodule classification, follow-up
Chen et al. [72]	Used imaging, segmentation, feature extraction: shape, intensity, texture, wavelet,	precision diagnosis and treatment: differentiate between cancer and nodules, pathological and molecular classification, treatment response and prognosis indication
Constanzo et al. [73]	segmentation, features extraction: SUV, IVH, texture, dynamic image features , model development	patient survival, tumour response, radiogenomics example.
Rabbani et al. [74]		diagnosis, genomic classification, prognosis, treatment in radiation oncology, immunotherapy, treatment selection and outcome prediction,
Avanzo et al. [71]	segmentation	cancer detection and characterisation: Prediction of histology and tumor stage, Tumour genotype,Prediction of treatment outcome: complete response and local control, distant metastases, survival , response to treatment
Thawani et al. [75]	segmentation, Feature analysis and their categories,	lung cancer diagnosis and prognosis , radiogenomics and prediction of treatment
Scrivener et al. [76]	Image acquisition, segmentation, feature extraction, model development and validation,	tumour classification, tumour prognosis,

2.2.3 Esophageal and Gastric cancer

The first reviews who handled the radiomics literature applied on esophageal cancer were written by van Rossum et al. [77] and Sah et al. [78], whereas Sah et al. discussed also the gastric literature. The images commonly used to analyse oesophagal cancer were CT and PET, while MRI was not routinely performed in the clinical pathway. From these images, both reviews found that the same features, mainly texture features, were used for radiomics analysis. van Rossum et al. deepened the reproducibility of the features and suggested using only a limited number of reproducible features. Even if the definition of reproducible features is still an issue, the authors conclude that GLCM features are the most reproducible. As for the other anatomical regions, the main radiomics applications on oesophagal cancer are overall survival prediction and Pathologic response analysis. The second analysed review compared articles separately using PET/CT and CT images of oesophagal and gastric cancer. No studies using PET/CT studies have been performed on gastric cancer. Analysed papers regarding oesophagal cancer using PET/CT images focused on the treatment response, distinguishing responders or non-responders, or

the outcome prognosis. The authors subsequently analysed papers that focused on using CT images for both problems, oesophagal and gastric cancer. For both clinical problems, papers addressed three main problems: classification, response prediction, and overall survival prediction. For oesophagal cancer, studies have focused on the response prediction or prognosis and found a higher heterogeneity in non-responders. Regarding gastric cancer, three of the analysed papers focused on the classification problem, finding that first and second-order features could help in the tumour classification. None of the included papers addressed the patient prognosis or the response to therapy for gastric cancer.

Table 2.4: Overview of cited reviews focusing on radiomics applications on esophageal and gastric cancer

	Technical evaluation and summary	**Clinical Applications**
van Rossum et al. [77]	Texture feature analysis: Reproducibility, Influence of smoothing, Quantisation and Segmentation	Tumour staging, Prediction of treatment response, Prediction of survival
Sah et al. [78]	Imaging	Staging, Treatment response, Outcome prognosis

2.2.4 Hepatic cancer

Focusing on radiomics application on hepatic cancer, four reviews were included and considered for this section. The first written by Jeong et al. [79] analysed papers that focused on radiomics and radiogenomic applications on primary liver cancers; hepatocellular carcinoma (HCC), and intrahepatic cholangiocarcinoma (ICC). The authors found that most of the studies have been based on CT images and that the usually addressed outcomes regard the diagnosis, the prognosis and the treatment response assessment. All included articles found valid radiomics signatures to solve the addressed problem. The authors conclude that non-invasive diagnostic tests should be further investigated and used due to the correlation between imaging features and molecular genomic data. The second introduced review was written by Wakabayashi et al. [80]. The authors included 23 studies in the review. Fourteen articles used CT images, 7 used MRI images, and two PET/CT images. The included articles focused on prediction problems (pathological grading or microvascular invasion), Overall Survival and Progression or Disease-Free Survival, Diagnosis and recurrence prediction. Finally, the authors evaluated all articles based on the Radiomics Quality Score (RQS) introduced by Lambin et al. [27] finding that the overall scores are relatively low.

The third introduced review written by Fiz et al. [81], focused on radiomics applications on liver metastases. The authors included 32 studies in the qualitative analysis. Considering the included articles, the authors found that 60% of the articles analysed CT images, 25% used MRI images, 9% used PET/CT images, and the

others used two multiple imaging modalities. The most common scope analysed by the selected studies was the analysis of technical aspects, such as the influence of acquisition or reconstruction parameters on the values of texture analysis indices. The second most common scopes was the prognosis and the therapy response assessment of metastasis given by colorectal cancer. The authors conclude that radiomics allows, in general, the non-invasive differential diagnosis, for example, between metastases from benign lesions and primary tumours. Moreover, the authors reported that liver metastases with higher entropy and lower homogeneity at diagnosis had been associated with a better prognosis and response to therapy.

The last review that focused on radiomics applications on hepatic cancer was written by Park et al. [82]. Park et al. initially introduced articles that used radiomics methods describing the main clinical scopes and results, for example, the chronic liver disease analysis or the malignant tumour prognosis. After the clinical applications, the authors address the radiomics pitfalls, focusing on the standardisation problems of the imaging protocol, the VOI selection and the feature extraction methods. The result, in their opinion, is to use deep learning methods instead of radiomics. The authors included nine studies that used deep learning methods on liver diseases. The included tasks were tumour segmentation, liver fibrosis staging, classification of liver tumours, MRI reconstruction and motion artefact reduction. All included articles obtained good and promising results. The authors conclude that both radiomics and deep learning are good liver disease assessment techniques.

Table 2.5: Overview of cited reviews focusing on radiomics applications on liver cancer

	Technical evaluation and summary	**Clinical Applications**
Jeong et al. [79]	Comparison between quantitative and qualitative analysis	Survival, Recurrence, and Treatment response after chemotherapy, Relationship between the genomic signatures and imaging findings, HCC, intrahepatic cholangiocarcinoma,
Wakabayashi et al. [80]	Radiomics workflow analysis	Tumour characterisation, Radiomics applications on HCC, prediction problems, Overall Survival and Progression or Disease-Free Survival, Diagnosis and Recurrence prediction
Fiz et al. [81]	Influence of acquisition or reconstruction parameters on the values of texture analysis indices, Influence of features on radiomics analysis	Liver metastasis from colorectal cancer: Survival prediction, Chemotherapy response prediction, Pathology data prediction, Radiomics applications on non colorectal cancer metastasis
Park et al. [82]	Radiomics features, Feature selection, Model development, Deep learning models: CNN, training of DL models	chronic liver disease, prognosis of malignant liver tumours, DL applications: segmentation using DL, liver fibrosis staging, diagnosis of fatty liver, detection and classification of tumours, image quality improvement

2.2.5 Kidney and Rectal cancer

An attempt to resume the literature of radiomics application on kidney cancer was made by de Leon et al. [83]. The authors found articles focused on the radiomics analysis of pre-and post-treatment assessment of renal masses. The analysed tasks were the renal masses subtyping and the tumour biology prediction regarding the pretreatment assessment. Most articles focused on the response prediction after treatments for the post-treatment task.

Regarding the Rectal Cancer a review was written by Dinapoli et al. [84]. The authors found that different radiomics studies focusing on the pathological characterisation, the primary tumour characterisation, the prediction of histopathological tumour response could be found. Furthermore, studies focusing on lymph nodes and distant metastases were summarised.

Table 2.6: Overview of cited reviews focusing on radiomics applications on kidney and rectal cancer

	Technical evaluation and summary	**Clinical Applications**
de Leon et al. [83]	Radiomics overview: Segmentation, Feature extraction	Subtyping of renal masses, Radiomics for prediction of tumour biology, Posttreatment assessment of renal cell carcinoma
Dinapoli et al. [84]		Primary tumour and Treatment monitoring, Definition of primary lymph nodes and Distant metastases

2.2.6 Breast Cancer

As shown by Sollini et al. another clinical area usually studied is the breast. Four recent and often cited reviews will be discussed, to sum up, the literature. The first reviews done by Valdora et al. [85] and Crivelli et al. [86] summarised the dedicated literature comparing 17 studies and 19 studies, respectively. Both reviews showed that almost all compared papers, except for one, were retrospective papers, and in most of the analysed papers of both reviews, the diagnostic modality used was the MRI followed by Mammography, ultrasound and FDG PET/CT. The most common aim of the summarised papers was for both reviews the prognosis, the molecular subtype distinction and the malignancy detection. In all papers, the best AUC value found was 0.87, while the worst was 0.56. Reig et al. [87] deepened the comparison of the segmentation method, finding that the most common segmentation method was to include the anatomic region or the fibroglandular tissue. Only one of the considered papers segmented the background parenchymal enhancement. Furthermore, the review shows that commonly radiomics paper focused on breast cancer studies the prediction of occult invasive Cancer in ductal carcinoma in situ, the lymph node status and the marker of aggressiveness and the prognosis prediction and the recurrence likelihood. Also, this review found results of AUC going from 0.78 to 0.96 the best. The last found review written by Chitalia and Kontos [88] analysed the literature from a texture analysis point of view, comparing the features extracted from the literature for the different clinical needs. The review showed that the most common extracted features are the grey level histogram features and the grey level co-occurrence matrix features. These features were commonly used for diagnostic applications, histopathological and molecular subtype classification and breast cancer prognosis.

Table 2.7: Overview of cited reviews focusing on radiomics applications on breast cancer

	Technical evaluation and summary	**Clinical Applications**
Valdora et al. [85]		Prognosis, the molecular sub-type distinction and the malignancy detection
Crivelli et al. [86]		Classifications task using MRI or US, Predict treatment response, Prognostic factors: lymph node metastasis, peritumoural fat, Ki67. Predict breast cancers molecular profile, Cancer recurrence prediction
Reig et al. [87]	Machine learning methods: Supervised, Unsupervised learning, Deep learning methods, Breast segmentation, Lesion segmentation, Texture analysis	Lesion classification, Predicting Occult Invasive Cancer in DCIS, Lymph Node Status and Markers of Aggressiveness, Predicting Prognosis and Likelihood of Recurrence. Radiogenomics: molecular subtypes analysis, genomic predictor recurrence, chemotherapy response prediction
Chitalia and Kontos [88]	Feature analysis	Applications in Breast Computer-Aided Diagnosis, Histopatologic and Molecular subtype classification, Breast cancer prognosis, Therapy response prediction

2.2.7 Prostate cancer

Prostate cancer is the most frequent male tumour and, therefore, one of the most analysed cancers with a radiomics approach. Four reviews that summarised the existing literature were found and introduced here. The first review, written by Cuocolo et al. [89] focused on the radiomics and machine learning applications for gland segmentation, prostate cancer detection, local staging, lesion aggressiveness assessment and the pretreatment assessment and follow-up. The authors conclude that machine learning applications can expand the role of prostate MRI and improve diagnostic performance. Moreover, Cuocolo et al. highlighted that the dynamic contrast-enhanced sequence was not always used. Thereby machine learning methods could avoid the systematic usage of contrast agents since good results have been obtained with and without the sequence. Finally, the authors underlined that further work is necessary to validate the process.

Both reviews, written by Sun et al. [90] and by Stoyanova et al. [91] focused on similar aspects. Sun et al. [90] initially reviewed multiparametric MRI and its role in Prostate Cancer detection, staging and management and Stoyanova et al. included papers that analysed prostate cancer aggressiveness. Both reviews described that the majority of the articles focused on the task of cancer detection, and Sun et al. classified radiomics as an evolution of a computer-aided detection system.

Subsequently, both reviews summarised the steps of the radiomics workflow,

such as segmentation, registration, feature extraction and selection, and the most frequently used classifiers, focusing on the advantages and disadvantages of each step. Finally, the authors describe the future perspectives, especially the correlation between imaging and genetic data, defined as *radiogenomics*, and its possible applications.

The last included review was written by Stanzione et al. [92]. These authors included 73 studies in this review regarding radiomics application on prostate cancer. All studies were classified according to the Radiomics Quality Score [27]. The authors found that the mean Radiomics Quality Score was 7.93 ± 5.13 on a maximum of 36 points finding the post-critical points as the lack of feature robustness and the missing external validation. The authors conclude that the lack of quality, as described by the Radiomics Quality score, should be addressed in future.

Table 2.8: Overview of cited reviews focusing on radiomics applications on prostate cancer

	Technical evaluation and summary	**Clinical Applications**
Cuocolo et al. [89]	types of ML algorithms: supervised, unsupervised, reinforcement learning, segmentation analysis,	cancer detection, Assessment of lesion aggressiveness,Local staging and pre-treatment assessment, Biochemical recurrence
Sun et al. [90]	multiparametric MRI: t2, DWI, dynamic MRI, MR spectroscopy, Blood oxygen level dependent MRI, feature extraction and selection, classifier training	detection of prostate cancer, segmentation and registration, Assessment of aggressiveness and staging,
Stoyanova et al. [91]	radiomics pipeline analysis,	prostate cancer diagnosis, Prostate cancer aggressiveness, Radiogenomics: MRI-US guided fusion biopsy, RNA extraction and microarray hybridisation, radiomics features,
Stanzione et al. [92]		RQS

2.3 Learning methods

2.3.1 Machine Learning

In this section review analysing the literature deepening the technical aspects were reported and summed up.

The reviews written by Avanzo et al. [15], Larue et al. [93], Cook et al. [94], Lambin et al. [27], Rizzo et al. [16] and Aerts [95] focused their reviews on the technical side of the literature that concerns radiomics applications using machine learning methods (Table 2.9, Table 2.10, Table 2.11). Avanzo et al. introduced the

review with an overview of the clinical areas of application, confirming that the most common studies organs were the lung, especially regarding the NSCLC, the breast, the prostate, the renal cell carcinoma and the head and neck cancer. All considered reviews, except for Cook et al. who deepened PET applications, considered articles that analysed CT, MRI, PET and ultrasound images. Avanzo et al. focused on the advantages of every single imaging technique, summarising that CT images are most commonly used for radiomics applications due to the ability to describe tissue density, shape and texture of tumours. On the other side, MRI provides structural and functional information of the soft tissues and, when combined with contrast agents, characterises the concentration of an injected gadolinium contrast agent over time. Finally, PET images give functional information about the studied area and are usually acquired using the radiotracer 18F-fluorodeoxyglucose (18F-FDG).

Segmentation methods In the literature, three different types of segmentation were applied to medical images; manual segmentation, semi-automatic and automatic segmentation. Lambin et al. evidenced In general, a key condition of the segmentation method choice is how the segmentation is performed and how sensitive the features and the following analysis are to the different segmentations [27]. The first introduced segmentation is manual segmentation, defined as a straightforward solution and as the ground truth when performed by expert readers. Despite that, they suffer from high inter-reader variability and are time-consuming [15, 16]. Semi-automatic and automatic segmentations are being studied and investigated to minimise the manual input and the inter-reader variability and to fasten the process [15]. The most commonly used semi-automatic method in clinical practice is the region-growing method. However, it works very good only for homogeneous regions and worse for inhomogeneous regions [16, 93]. Moreover, Aerts [95] reported that automatic segmentations were already introduced especially for breast cancer, where computer-aided detection systems are reliable for identifying tumours or nodular lesions. Finally, Vial et al. [56] reported that deep-learning methods as a segmentation tool are being developed, but further researches are necessary.

Feature extraction and classification methods After the description of the segmentation methods used, all reviews focused on the extracted features and classification methods. Aerts introduced that initially, semantic annotations were obtained from the expert radiologist who used their experience to determine non-standardised lexicon. However, these semantic features suffer from intra- and inter-reader variability, and therefore quantitative features have been introduced. Extracted features can be classified into four main groups: shape/size, first and second order histogram and texture features [15, 16, 27, 93, 94]. Besides the standard used features, when using different imaging techniques, new features can be included,

such as the SUV metric when using PET images or the wash-in, wash-out curve in dynamic MRI images [15]. Image pre-processing or reconstruction affect features values and therefore must be considered for future evaluations [27]. The most common used pre-processing is the wavelet decomposition of the original image, which has been widely used to extract textures from different frequency bands [93]. Moreover, Cook et al. evidenced that varying the acquisition methods or the matrix size changes the values of the features.

The first step before the classification algorithm is implementing a feature selection step as mentioned by all included reviews. Avanzo et al. defined feature selection as an algorithm that selects relevant features for a given task. Features can be selected either by grouping highly correlated features obtained by clustering [16, 27] or using Principal Component Analysis (PCA) [16, 93] that can highlight outliers or features that were stable during a test/retest analysis [15]. Besides these feature selection techniques, alternative methods used in radiomics studies are the filter based selection of univariate type or the wrapper selection [93]. Finally, the last step analysed by the reviews was the classification task. Larue et al. summed up in their review that machine learning methods can be distinguished into unsupervised and supervised algorithms and that the choice of the classification is the dominant source of performance variation. Avanzo et al. sums up that in radiomics the most used classifiers are the logistic regression and the random forest, which is based on the decision trees. Finally, SVM is also frequently used and mainly used for CAD systems since it is a discriminative supervised machine learning technique.

All reviews concluded reporting similar limitations found in the radiomics literature. The lack of reproducibility, the variability of medical scanners and the used parameters, the small sample size and the missing independent validation cohort are important limitations of the analysed studies since they make it difficult or impossible to reproduce the result with external datasets. Finally, Lambin et al. [27] proposed a radiomics Quality Score (RQS) to assess the quality of the radiomics literature. The authors believed that the quality of the reported radiomics study is poor and that clear and complete reporting is required to enhance the usefulness of the models. To classify the whole literature, the score can be applied to both past and future works. Following this score, publications should report the study design, the protocols, the process and the standard operating procedures to make external validations possible. Moreover, publications should clearly describe the identified exigent unmet need. Totally, the score that can be assigned to each article is 36 points analysing 16 criteria.

2.3.2 Deep Learning

The use of Deep Learning applications for radiomics studies has started a few years after the radiomics introduction with the first work[4] written by Huynh et al. [96]. Since this first article, many more works have been published focusing on radiomics applications using deep learning techniques. These articles where summarised by the four following reviews written by Parekh and Jacobs [57], Boldrini et al. [97], Suzuki [98] and Vial et al. [56].

The first introduced review was written by Parekh and Jacobs [57]. The authors summarised three types of deep neural networks usually used for medical applications; Discriminative deep learning models, Generative deep learning models and Deep reinforcement learning. Moreover, the authors described some attempts to *open black box* such as the activation maximisation or the deconvolutional network. Following, the authors analysed the concept of *multiparametric radiomics* where radiomics applications were applied to multiple different image sequences or images type obtaining, for example, better tissue segmentations. Moreover, the interpretability of radiomics features was analysed, finding that these were not standardised, and describing the difficulty in finding a relationship between the features and the underlying biology.

Boldrini et al. [97] and Suzuki [98] both focused their reviews on the clinical applications which were focused on using radiomics with DL. Both resumed that most articles focused on lesion segmentation and detection and lesion classification. Boldrini et al. additionally found that part of the included 43 studies focused on the clinical outcome prediction, on the images dose quantification and the dose-response modelling and adaptation. Furthermore, Suzuki focused also on the task of the bone and soft tissue separation in CXR images, since studies evidenced that around 90% of missed lung cancers were partly obscured by ribs or clavicle. The last included review, written by Vial et al. [56] focused on the main used DL techniques such as CNN, Deep Belief Networks (DBN) and Deep Autoencoders, which are commonly used due to their ability in the image texture detection.

Boldrini et al., Suzuki and Vial et al. reported that the usage of small datasets is an important limitation of the literature since it causes overfitting problems. Moreover, Suzuki reported that a high computational cost for training is necessary. Vial et al. underlined that to apply deep learning systems to medical images, expertise in biology and computer science is necessary to avoid the *black box* effect that could lead to high accuracy results without any medical reason.

Vial et al. and Parekh and Jacobs both conclude that the combination of radiomics and deep learning has the potential to strengthen the role of radiology and personalised medicine. Furthermore, Parekh and Jacobs reports that especially

[4]First published article found with scopus using as keyword *radiomics* and *deep learning*

CNN methods can capture textural information contained in medical images in the initial convolutional layers. Boldrini et al. [97] concluded that promising results have shown how deep learning systems could help clinicians in the daily practice, supporting in the segmentation or the prediction of treatment outcomes. Finally, Suzuki [98] concluded that they expect that machine learning and deep learning systems will be the mainstream technology in medical imaging in the future.

Table 2.1: Overview and summary of references cited in section 2.2

reference	Potential applications of radiomics	Clinical Areas of Applications	Images	Factors that affect radiomic features quantification	False positive discovery rate and proper study design	Conclusion /limitation
Yip and Aerts[62]	Prediction of treatment response and outcomes, Tumour staging, Tissue identification, Assessment of cancer genetics		MR , CT, PET	Acquisition modes, reconstruction parameters, smoothing, and segmentation thresholds, Reproducibility of radiomic features, Image discretization (resampling) schemes, Computation of radiomic features, Respiratory motion, Tumour size and intratumoural heterogeneity	Many studies examined the prognostic value of radiomic features, However, it is not uncommon that the number of examined radiomic features is much greater than the number of patients which can lead to feature selection bias and false positive results	not all radiomics features are recommended for use due to their sensitivity to acquisition modes and reconstruction parameters. To examine the prognostic power of radiomic features, datasets consisting of ten to fifteen patients per feature evaluated has been recommended.
Bi et al.[14]	3 broad categories of image-based clinical tasks in oncology: 1) detection of abnormalities; 2) characterization of a suspected lesion by defining ts shape or volume, histopathologic diagnosis, stage of disease, or molecular profile; and 3) determination of prognosis or response to treatment over time during monitoring. 2D	Lung cancer: 1) early detection and characterisation 2) staging and characterisation, generally: lung cancer screening , lung cancer characterisation, assessing intratumour heterogeneity, assessing response to target therapies, CNS TUMOUR : 1) accurate diagnosis of the type and extent of disease; 2) reliable tracking of neoplastic disease over time, 3) the ability to extract genotype signatures from the phenotypic manifestation of tumours on imaging, BREAST CANCER, PROSTATE CANCER: 1) overdiagnosis and overtreatment 2)inadequate targeted biopsy sampling, leading to misdiagnosis and to disease progression in men with seemingly low-risk prostate cancer.				The curation of medical data represents a major obstacle in developing automated clinical solutions.Furthermore, access to available data sets should be improved to promote intellectual collaboration.Another limitation includes the interpretability of AI and the ability to interrogate such methods for reasons behind a specific outcome, as well as the anticipation of failures.
Lee and Lee[63]		Brain, FET and FLT are used form glioma evaluation. 1) predictive value for survival, 2) assessed the diagnostic value , HEAD AND NECK: 1) predicting disease-free survival and OS 2) intratumoral heterogeneity in patients with HPV-positive oropharyngeal cancer. oral cavity cancer, 1)predictive value for OS:, NON SMALL CELL LUNG CANCER: 1) intratumoral heterogeneity on PET/CT 2) relationships between textural features and tumor characteristics 3) clinical significance for planning radiotherapy of intratumoral heterogeneity . 4) predicting survival in diverse clinical NSCLC settings BREAST 1)differentiating subtypes and predicting treatment response to neoadjuvant treatment as well as staging and predicting prognosis	PET/CT		PET/CT has lower spatial resolution than that of anatomical imaging modalities such as CT and MRI, the reliability of heterogeneity parameters in tumors with small volume is limit
Sollini et al.[12]		86% of works handle with oncology problems, brain, urology, skin, muskolosche, lung, Gastrointesinal , breast the others not oncological problems	ultrasound, radiography, mammography, endoscopy, skin pictures, ocular fundus pictures,CT, MRI, scintigraphy and PET , PET/CT			

Table 2.9: Overview and summary of references cited in section 2.3.1 (Clinical Areas of Applications, Images, Segmentation)

Reference	Clinical Areas of Applications	Images	Segmentation
Avanzo et al.[15]	Lung: Lung NSCLC is the tumor which has been most extensively studied and characterized , Breast, Prostate: Radiomic features, derived primarily from T2-w and ADC MRI scanning, correlate with Gleason score, which is probably the most powerful prognostic factor for prostate cancer, GBM, Renal cell carcinoma, Head and Neck, soft tissue carcinoma, rectal	CT The most widely used imaging modality in radiomics studies is CT, which assesses tissue density, shape and texture of tumor and lymph-nodes PET is used for detecting and staging cancer, most commonly with the radiotracer 18F-fluorodeoxyglucose (18F-FDG), MRI provides high-contrast structural and functional information to characterize soft tissue Dynamic contrast-enhanced (DCE) MRI characterizes the concentration of an injected gadolinium contrast agent over time	automatic or semi-automatic segmentation. Region growing is a semiautomatic method often applied to the segmentation of masses, 2)watershed method Automati c segmentation has been used in breast, brain has been developed
Larue et al.[93]		widely used CT, MRI and PET, also used : ultrasound, Quantitative features retrieved from ultrasound (US) images mainly have shown to be useful to discriminate between normal, malignant and benign tissue	manual segmentation: Manual delineation is a straightforward solution, but can also be very time-consuming and is susceptible to inter-observer variability, Automatic or semi-automatic segmentation methods currently are investigated extensively to minimize manual input and increase consistency in delineating the regions of interest
Cook et al.[94]	Characterisation and segmentation, Prediction and prognosis	PET	
Lambin et al.[27]			VOIs are segmented manually or (semi-) automatically, Key considerations are how the segmentation was performed, and how sensitive the radiomics analysis is to different segmentation methods
Rizzo et al.[16]		computed tomography (CT), positron emission tomography (PET), and magnetic resonance imaging (MRI)	Indeed, many authors consider manual segmentation by expert readers the ground truth despite high inter-reader variability. Automatic and semi -automatic segmentation methods have been developed across imaging modalities and different anatomical regions: Some segmentation algorithms rely on region-growing methods that require an operator to select a seed point within the volume of interest (works well for homegeous regions and not well for inhomogeneous regions
Aerts[95]		The first step of image-based phenotyping involves data acquisition. The quality of the imaging data depends on the reliability of the acquisition protocols used in clinical centers	Automated tumour detection and segmentation methods have also been introduced into clinical practice. Computer-aided detection systems are reliable for identifying tumors or nodular lesions. The largest successes have been observed in breast cancer

Table 2.10: Overview and summary of references cited in section 2.3.1(Features, Classification)

Reference	Features	Classification
Avanzo et al.[15]	Radiomic features can be divided into: shape/size, first order histograms or global statistics, second order histograms or textural (see below).SUV metrics in PET. dynamic MRI: Its peak value or "maximum uptake", the time frame index at which the maximum, gray-level cooccurrence matrix (GLCM), gray level run-length matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), neighborhood gray-tone difference matrix (NGTDM), uptake occurs or "peak location", uptake rate (maximum uptake/- peak location), and washout rate can then be used as descriptors	By feature selection we intend an algorithm used to select "effective" features for a given task, i.e. those features who are relevant to explain a given output as a function of a group of features, univariate predictors vs multivaraite, most stable radiomic features based on test/retest analysis [56] and choosing the single feature with the highest performance for each category CLASSIFIER : most used logistic regression, Random forest [27,34,45,53,54,78] is based on decision trees, a popular concept in machine learning especially in the field of medicine,SVM is a discriminative supervised machine learning technique previously used for CAD
Larue et al.[93]	First-order statistics , Texture or gray scale variation features, Wavelet decomposition of the original image has been employed to extract intensity and texture features from different frequency bands, Shape-based features	Feature selection methods:Filter-based selection techniques of the univariate type, Wrapper selection techniques, Principal component analysis performs a transformation for dimensionality reduction and can highlight outliers ML: Unsupervised and supervised ML: the choice of classification method was found to be the dominant source of performance variation.
Cook et al.[94]		The methods for measuring image spatial heterogeneity can be divided into global, regional or local parameters representing the relationships between voxel intensities.The most commonly used statistical methods include firstorder (one voxel) , second-order (two voxels) and highorder (three of more voxels) parameters, but other methods exist, such as model-based, e.g. fractal analysis, or transform- based ones
Lambin et al.[27]	Feature values are dependent upon factors that can include image pre-processing (for example, filtering, or intensity discretization) and reconstruction	Groups of highly correlated radiomics features can be identified via clustering, and these features can be reduced to sing. Radiomic modelling involves three major aspects: feature selection, modelling methodology, and validationle archetypal features per cluster
Rizzo et al.[16]	Quantitative features are descriptors extracted from the images by software implementing mathematical algorithms [4]. They exhibit different levels of complexity and express properties firstly of the lesion shape and the voxel intensity histogram, secondarily of the spatial arrangement of the intensity values at voxel level	methods for data analysis strictly depend on the number of input variables, possibly affecting the final result. One possible approach is to start from all the features provided by the calculation tool, and to perform a preliminary analysis to select the most repeatable and reproducible parameters; to subsequently reduce them by correlation and redundancy analysis. Radiomics' analysis usually includes two main steps: 1. Dimensionality reduction and feature selection, usually obtained via unsupervised approaches; and 2. Association analysis with one or more specific outcome(s) via supervised approaches.Different methods of dimensionality reduction/feature selection and model classification have been compared [13, 14]. The two most commonly used unsupervised approaches are cluster analysis [7, 14, 15] and principal component analysis (PCA). Supervised multivariate analysis consists of building a mathematical model to predict an outcome or response
Aerts[95]		Semantic annotation refers to the manual assessment of the tumor phenotype by an expert radiologist. In current clinical practice, this assessment is often made in a qualitative manner using a nonstandardized lexicon. However, there are also disadvantages with semantic annotations. Large intrareader (same reader) and interreader (different reader) variability exists. The process of automated phenotype quantification is also referred to as radiomics. Semantic and radiomic feature representations often provide complementary information about the tumor phenotype. To take advantage of this scenario, the radiomic workflow includes an interactive component in the quantification phase

Table 2.11: Overview and summary of references cited in section 2.3.1(Conclusion/Limitations, RQS)

Reference	Conclusion /Limitations	RQS
Avanzo et al.[15]	Reproducibility: Reproducibility or robustness, in contrast, is measured when measuring system or parameters differ. The major sources for variability of radiomic features are the imaging scanners, the parameters of acquisition and reconstruction of the image, and delineation of ROI. Sample size and statistical power: Most radiomics studies do not report sufficient validations in independent cohorts, thereby limiting generalizability to additional patient populations, Standardization and benchmarking: standardized acquisition and reconstruction protocol will be needed to smooth out input data variability PITFALS. The main criticisms to radiomics is that the link between the imaged properties of tumors and tumor biology is not straightforward.	
Larue et al.[93]	Standardization or calibration: of standardization/harmonization or at least a correlation between radiomic features acquired in different settings (e.g. scanner type, hospital, radiomics software) makes it difficult to directly compare different studies and extracted feature	
Cook et al.[94]	Indeed, different textural features have been found to show different variability when varying the acquisition method (2D vs. 3D), matrix size reconstruction algorithm and post-reconstruction filter	
Lambin et al.[27]	Validation is the first step towards a model being accepted in both the scientific and clinical communities. Independent verification of the results is a necessary additional step. Replication means independent verification of the results by independent researchers repeating the analysis using the same technique and different (but appropriately selected) datasets	We propose the radiomics quality score (RQS) to aid assessment of both past and future radiomic studies.Publications should extensively report study-design ,protocols, detailed quality assurance processes, and standard operating procedures.Overwhelming evidence shows that the quality of reporting of prediction model studies is currently poor
Rizzo et al.[16]	As shown, there is still no universal segmentation algorithm for all image applications, and new algorithms are under evaluation to overcome these limitations [56–58]. Indeed, some features may show stability and reproducibility using one segmentation method, but not another.	
Aerts[95]		

Chapter 3

Insights into radiomics workflow steps

This thesis aims to analyse different aspects of the radiomics workflow to find new techniques that could improve the results and stability of this approach. Different aspects were investigated: the development and introduction of new features (section 3.1), available solution to cope with imbalanced learning (section 3.2), the combination of deep learning and machine learning techniques (section 3.3) and finally, how the segmentation influence model performance (section 3.4).

3.1 Development and evaluation of new features

3.1.1 Feature maps

An alternative to the conventional features extracted during a radiomics workflow is the generation of feature maps, also referred to as parametric maps. Conventional features represent a mean value for the whole segmented area and produce a single value for each included feature. Conversely, the feature maps represent the value of the chosen feature for each pixel of the ROI/VOI. Parekh and Jacobs [99] published an article where they presented a radiomics feature mapping framework to generate radiomics MRI texture image representations called the radiomics feature maps (RFM). These maps correlated with quantitative texture values extracted from MRI and breast tissue biology to classify benign from malignant tumours. Moreover, Gonzalez and Alberich Bayarri [100] published on the 19^{th} of February 2020, a blog where the process behind a COVID-19 detection and follow-up tool using Artificial Intelligence and radiomics applied to X-Ray and Computed Tomography was described. This tool calculated features maps of lungs and extracted seven statistical values of each map.

In our application, parametric maps, as an alternative to standard radiomics features, were applied on a COVID-19 positive X-Ray dataset. Parametric maps

were used for this study because since the segmented areas were extensive and nonhomogeneous, standard radiomics features would not be able to detect relevant features.

This work investigated whether artificial intelligence working with chest X-Ray scans and clinical data could be used as a possible tool for the early identification of patients at risk of severe outcome. The developed learning approaches were specifically designed to use image-based features together with clinical data.

To obtain this goal, three learning approaches were developed and compared to predict clinical outcome. Indeed, in addition to clinical information consisting of general information, laboratory data and comorbidities, such approaches use quantitative information extracted from the CXR images, which are also referred to as image features or quantitative biomarkers in the following. The first approach, referred to as *Handcrafted approach*, computes handcrafted texture features used by a shallow classifier; together with clinical data, the second approach, called the *Hybrid Approach*, automatically extracts image descriptors by using a CNN, that, again, are used by a shallow learner together with clinical data. The third approach, referred to as *End-To-End Approach*, is fully based on DNNs, processing both clinical and image data (Figure 3.1).

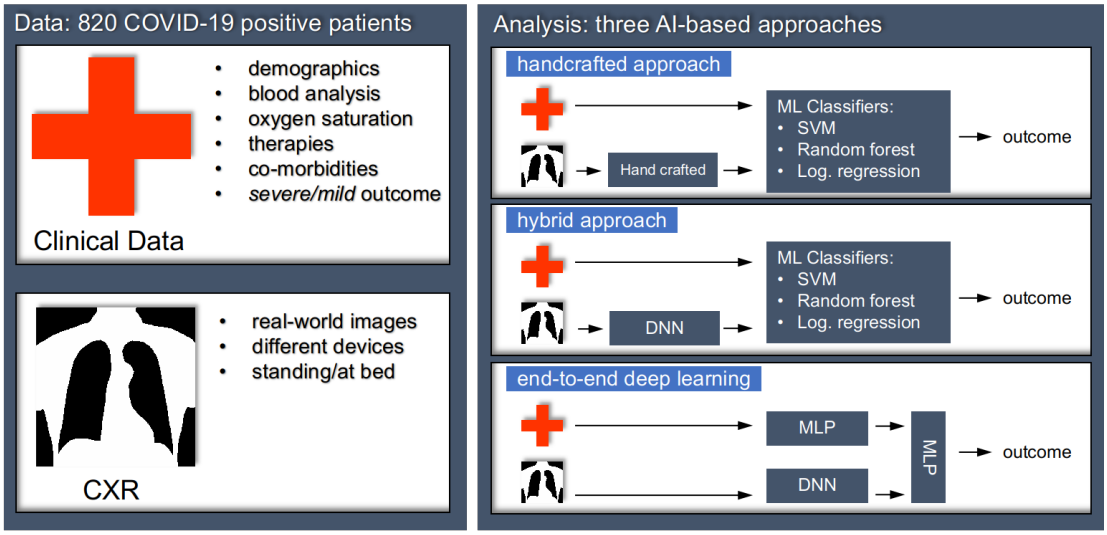


Figure 3.1: Overview of the method for automatic prognosis of COVID-19 in two classes, namely mild and severe.
Our works includes data collected in 6 independent cohorts, resulting in 820 COVID-19 patients. For each, we collected several clinical attributes, combined with quantitative imaging biomarkers computed by handcrafted features or automatically computed by CNNs.

This section presents, the dataset, the preprocessing steps and the learning algorithms.

In particular, we deepen now, the *Handcrafted Approach*, while the *Hybrid Ap-*

Table 3.1: Patient distribution across the hospitals where the data were collected.

Hospital	Number of patients	Mild class prior probability	Anterior Posterior (AP) projection prior probability
A	120	29.2%	81.67%
B	104	56.7 %	97.12%
C	31	25.8 %	90.32%
D	139	54.7 %	38.85%
E	101	54.5%	87.13%
F	325	46.5%	98.46%
Total	**820**	**46.8%**	**83.72%**

proach will be deepened in section 3.3. The *End-to-End Approach* is not presented in here as I parially contributed to its development.

3.1.1.1 The AIforCOVID dataset

The AIforCovid dataset includes the images and clinical data collected in six Italian hospitals during the hospitalization of symptomatic patients with COVID-19 during the first wave of emergency in the country (March-June 2020). Such data was generated during the clinical activity with the primary purpose of managing COVID-19 patients within the daily practice, and they were retrospectively reviewed and collected after patients' anonymization. Ethics Committee approval was obtained (Trial-ID: 1507; Approval date: April 7th, 2020), and all data were managed following the GDPR. Furthermore, we randomly assigned to each centre a symbolic label, from A up to F.

The 820 CXR examinations reviewed in this study were performed in COVID-19-positive adult patients at the time of hospital admission (Table 3.1): all the patients resulted positive for SARS-CoV-2 infection at the RT-PCR test [101]. In 5% of such cases, the positivity to the swab was obtained only at the second RT-PCR examination. In the different centres, CXR examinations were performed using different analogue and digital units. Furthermore, the execution parameters were settled according to the patient conditions. Paired with CXR examinations, we also collected relevant clinical parameters listed in Table 3.2.

According to the clinical outcome, each patient was assigned to the *mild* or the *severe* group. The former contains the patients sent back to domiciliary isolation or hospitalised without ventilatory support, whereas the latter is composed of patients who required non-invasive ventilation support, intensive care unit (ICU) and deceased patients. Figure 3.2 shows four difficult examples of CXR images within the dataset: indeed, panels A and B show two images of patients with severe outcome whilst the radiological visual inspection may suggest severe and mild prognoses, respectively. Similarly, panels C and D show two images of patients with mild outcome whilst a radiologist may report severe and mild prognoses, respectively.

During an initial data quality cleaning, we double-checked with the clinical part-

Table 3.2: Description of the clinical data available within the repository. First and second columns report variables label and description. Summary statistics for the overall population and for the two patients groups are reported in the following columns. For continuous variables median and interquartile range are reported, for categorical variables proportions are reported. Feature names followed by $'+'$ were not used for the analysis described in this work.P−values lower than 0.05 were considered significant. ∗Mann−Whitney U test. † z−test for proportions with Yates continuity correction. † Fisher exact test.

Name	Description	Overall−population	Mild−group (A)	Severe−group (B)	A vs B p−value	Missing data (%)
Active cancer in the last 5 years	Patient had active cancer in the last 5 years (% reported)	7%	5%	8%	<0.05†	1.4
Age	Patient's age (years)	64; 54 – 77	60; 49 – 72	70; 60 – 79	<0.001∗	0
Atrial Fibrillation	Patient had atrial fibrillaton (% reported)	9%	5%	11%	<0.01†	2.2
Body temperature (°C)	Patients temperature at admission (in °C)	38; 37 – 38	38; 37 – 38	38; 37 – 38	0.171	8.8
Cardiovascular Disease	Patient had cardiovascular diseases (% reported)	35%	23%	40%	<0.001†	1.7
Chronic Kidney disease	Patient had chronic kidney disease (% reported)	6%	4%	9%	<0.01†	1.4
COPD	Chronic obstructive pulmonary disease (% reported)	7%	4%	10%	<0.01†	1.4
Cough	Cought (%yes)	54%	59%	50%	<0.05†	0.5
CRP	C−reactive protein concentration (mg/dL)	57; 24 – 119	42; 17 – 75	103; 48 – 163	<0.001∗	3.5
Days Fever	Days of fever up to admission (days)	3; 2-4	3; 2 – 4	3; 2 – 3	0.289	10.96
D−dimer	D−dimer amount in blood	632; 352−1287	549; 262 – 909	820; 438 – 2056	<0.001∗	77.6
Death+	Death of patient occurred during hospitalization for any cause	168	0	168	–	–
Dementia	Patient had dementia (% reported)	4%	3%	6%	0.087	1.8
Diabetes	Patient had diabetes (% reported)	16%	10%	21%	<0.001†	1.4
Dyspnea	Patient had intense tightening in the chest, air hunger, difficulty breathing, breathlessness or a feeling of suffocation (%yes)	50%	37%	62%	<0.001†	0.4
Fibrinogen	Fibrinogen concentration in blood (mg/dL)	607; 513 – 700	550; 473 – 658	615; 549−700	<0.001∗	73.6
Glucose	Glucose concentration in blood (mg/dL)	110; 96 – 130	104; 93 – 121	114; 101 – 139	<0.001∗	20.6
Heart Failure	Patient had heart failure (% reported)	2%	1%	3%	0.157	2.3
Hypertension	Patient had high blood pressure (% reported)	46%	38%	54%	<0.001†	1.4
INR	International Normalized Ratio	1.13; 1.07 – 1.25	1.11; 1.06 – 1.20	1.15; 1.08 – 1.28	0.004∗	28.8
Ischemic Heart Disease	Patient had ischemic heart disease (% reported)	15%	11%	18%	<0.01†	18.3
LDH	Lactate dehydrogenase concentration in blood (U/L)	320; 249 – 431	271; 214 – 323	405; 310 – 527	<0.001∗	24.6
O_2 (%)	Oxygen percentage in blood (%)	95; 90 – 97	96; 94 – 98	92; 87 – 96	<0.001∗	16.5
Obesity	Patient had obesity (% reported)	9%	6%	11%	0.058	36.1
$PaCO_2$	Partial pressure of carbon dioxide in arterial blood (mmHg)	33;30 – 36	34; 30 – 37	33; 30 – 35	0.116	15.4
PaO_2	Partial pressure of oxygen in arterial blood (mmHg)	69; 59 – 80	73; 67 – 81	64; 54 – 76	<0.001∗	15.3
PCT	Platelet count (ng/mL)	0.19; 0.09 – 0.56	0.09; 0.05 – 0.26	0.28; 0.13 – 0.72	<0.001∗	71.8
pH	Blood pH	7; 7 – 7	7; 7 – 7	7; 7 – 7	<0.001∗	17.3
Position+	Patient position during chest x−ray (%supine)	78%	68%	87%	<0.001†	0
Positivity at admission	Positivity to the SARS−CoV−2 swab at the admission time (% positive)	95%	94%	96%	0.142	4.7
Prognosis	Patient outcome (% cases)	–	46.8%	53.2%	0.468†	0.0
RBC	Red blood cells count ($10^3$9/L)	4.65; 4.26 – 5.07	4.70; 4.34 – 5.11	4.59;4.13 – 5.03	<0.001∗	3.0
Respiratory Failure	Patient had respiratory failure (% reported)	1%	100%	2%	0.131	19.0
SaO_2	arterial oxygen saturation (%)	95; 91 – 97	96; 94 – 98	92;87 – 96	<0.001∗	59.2
Sex	Patient's sex (% males)	68%	59%	75%	<0.001†	0
Stroke	Patient had stroke (% reported)	4%	3%	4%	0.447	2.3
Therapy Anakinra+	Patient was treated with Anakinra (%yes)	100%	0%	0%	–	10.8
Therapy anti−inflammatory+	Patient was treated with anti−inflammatory drugs therapy (%yes)	55%	53%	57%	0.243	13.5
Therapy antiviral+	Patient was treated with antiviral drugs (%yes)	47%	44%	50%	0.129	10.7
Therapy Eparine+	Patient was treated with eparine (no; yes; prophylactic treatment; therapeutic treatment)	56.6%; 11.5%; 28%; 3.9%	73.3%; 8.3%; 17.2%; 1.1%	39.9%; 14.7%; 38.8%; 6.6%	<0.001†	13.4
Therapy hydroxychloroquine+	Patient was treated with hydroxychloroquine (%yes)	59%	56%	62%	0.118	11.6
Therapy Tocilizumab+	Patient was treated with Tocilizumab (%yes)	9%	2%	15%	<0.001†	12.4
WBC	White blood cells count ($10^3$9/L)	6.30; 4.73 – 8.42	5.58; 4.32 – 7.17	7.10; 5.25 – 9.80	0.012∗	0.7

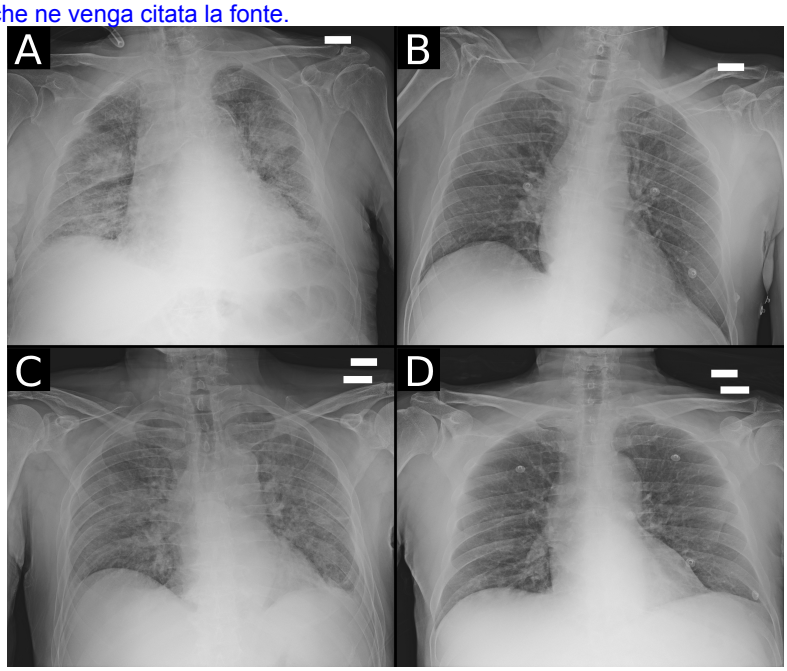


Figure 3.2: Examples of CRX images of patients with COVID-19 available within the dataset.
Panels A and B show two images of patients with severe outcome whilst the radiological visual inspection may suggest severe and mild prognoses, respectively. Similarly, panels C and D show two images of patients with mild outcome whilst a radiologist may suggest severe and mild prognosis, respectively, based on the visual interpretation.

ners the anomalous data and the outliers, i.e. those values lying outside the expected clinical range or identified applying the interquartile range method, which were then corrected when needed. Categorical variables values were homogenized to a coherent coding, such as 0 and 1 values for binary variables like comorbidities and sex, and we adopted the string "NaN" to denote missing data. No exclusion rules were applied for images based on device type or brand (e.g. digital or analogue devices) or patient positions (standing or at bed), whereas X-ray images taken with lateral projection were excluded because they were not available for patients whose images were acquired in the lying position. In the case of multiple CXR images delivered for the same patient, the dataset contains only the first one. It is worth noting that the presence of missing entries in the clinical data mostly depends upon the procedures carried out in the individual hospitals and the pressure due to the overwhelming number of patients hospitalized during the COVID-19 emergency. For the sake of completeness, the rate of missing data is reported in the last column of Table 3.2.

CXR images were collected in DICOM format and, for anonymization constraints, all the fields but a set of selected metadata related to acquisition parameters were blanked in the DICOM header (e.g. image modality, allocated bits, pixel spacing, etc.).

All the images in the repository are currently stored using 16 bits, while acqui-

sition precision varies: 13.5% were acquired at 10 bits precision, 35.4% at 12 bits, 46.6% at 14 bits and 4.5% using the full 16 bits precision. Furthermore, all the images were acquired with isotropic pixel spacing ranging from 0.1 mm to 0.2 mm. The most common pixel spacing is 0.15 mm, 0.1 mm and 0.16 mm for 43.9%, 13.7% and 13.6% of images, respectively. Image sizes, in pixels, are distributed as follows: 33.4% of the images have 2336×2836 pixels, 13.5% of images have 3520×4280 pixels, and 10.1% of the images have 3480×4240 pixels. The other images had a number of rows ranging from 1396 up to 4280, whilst the number of columns ranges from 1676 up to 4280.

3.1.1.2 Methods

The Handcrafted approach deepened in this section employs first order and texture features computed from images, which were mined together with the clinical data feeding a supervised learner. Moreover, it first computes parametric maps of the lungs segmented in the CXR image; second, it extracts several features that were then provided with the clinical data to a supervised learner.

3.1.1.2.1 Images standardisation

CXR images collected for this study were acquired with different devices and acquisition conditions, as mentioned in section 3.1.1.1. For this reason, we applied image normalisation that, to a large extent, was the same for all three methods. Indeed, for the handcrafted approach, pixels values were normalised to have zero mean and unit standard deviation, whilst the images were resized to 1024×1024 pixels using bilinear interpolation.

3.1.1.2.2 Lung segmentation

When needed, to segment the lung, we apply a semi-automatic approach that initially delineates the lung borders using a U-Net, which is a convolutional neural network architecture for fast and precise segmentation of images. In this respect, it is well known that the semantic segmentation provided by this deep network has proven to have very satisfactory performance when using medical images [35, 102, 103]. The implementation of the segmentation with U-Net is described in section 3.4.1.

3.1.1.2.3 Feature selection and classifiers

In general, we had a large number of descriptors that suggested applying a feature selection stage, which was set up in two steps. The first is a coarse step that runs univariate filtering based on mutual information as a score function to preselect a reduced set of image descriptors, whatever the approach used for their computation. The calculation of mutual information between continuous features

with the discrete class variable was addressed by estimating the entropy from the k-nearest neighbours' distances [104]. The second feature selection step merges the pre-selected imaging features with the clinical data. To this end, we applied a wrapper approach, namely the Recursive Feature Elimination and Cross-Validated selection (RFECV) method [49], which receives as input the pre-selected imaging descriptors and the 34 clinical features. Indeed, the RFECV is fed by an increasing number of pre-selected imaging descriptors (D_{pr}): fine-grained sampling was carried out for $D_{pr} \leq 10$ applying a step of 2; for $10 < D_{pr} \leq 50$, D_{pr} was sampled with step of 5; finally, RFECV was fed with all the image features. RFECV applies a pruning procedure that starts considering all features in the dataset and recursively eliminates the less important according to a feature's importance score calculated using a classifier. Note that the optimal number of features is selected by RFECV using nested 5-fold cross-validation on the training set.

Regarding the base learner, we evaluated three different computational paradigms: Logistic Regression (LGR); Support Vector Machines with a linear kernel (SVM); and Random Forests (RF). For all parameters in the adopted models, we used the default values provided by the libraries without any fine-tuning. Indeed, we were not interested in the best absolute performance. Moreover, Arcuri and Fraser [105] empirically observed that in many cases the use of tuned parameters cannot significantly outperform the default values of a classifier suggested in the literature.

3.1.1.2.4 Models validation

Model validation for the three tested methods consists of k-fold and leave-one-centre-out cross validation. For each cross-validation run, the training fold was used for data normalization, parameters' estimation and/or features ' selection depending on the applied method. Classification performance assessment was carried out using testing fold data only; k-fold cross-validation was repeated with k equal to 3 and 10 with 20 repetitions. In leave-one-centre-out (LOCO) cross-validation, in each run, the test set is composed of all the samples belonging to one centre only, while the others were assigned to the training set. When needed, the validation set was extracted from the training set using any policies (such as random selection, hold-out, nested cross-validation, etc.), and considering also the computational burden.

Performance of the learning models was measured in terms of accuracy, sensitivity and specificity, reporting the average and standard deviation of each experiment. When needed, we ran the pairwise two-sided Mann Whitney U test to compare the results provided by two methods, whereas we performed the Kruskal-Wallis test followed by the Dunn's test with Bonferroni correction for multiple comparisons. In the rest of the manuscript, we assume that the pairwise two-sided Mann Whitney U test was performed by default, otherwise, we will specify the test used.

3.1.1.2.5 Handcrafted Approach

As introduced in section 3.1.1.2 in the current section, the developed approach is presented. The first step was the lung segmentation, which was done applying the approach presented in section 3.1.1.2.2 but, as mentioned there, we deem that the segmentation performance is not satisfactory for exact lung delineation. For this reason, the lung masks are then reviewed by expert radiologists and then used to compute the handcrafted features as follows.

From the segmented lungs, we computed the parametric maps using a pixel-based approach as proposed by Penny et al. [106]. Pixels values of the parametric maps were obtained by computing first- and second-order radiomics features on a 21x21 sliding window running over each pixel of the entire lung region. First-order measures describe the statistical distribution of tissue density inside the kernel; from its grey levels' histogram, we extracted 18 descriptors: Energy, Total Energy, Entropy, Minimum, Maximum, Mean, Median, Interquartile Range, Range, Mean Absolute Deviation, Robust Mean Absolute Deviation, Root Mean Squared, Skewness, Kurtosis, Variance and Uniformity. Second-order descriptors are based on the Grey Level Co-occurrence Matrix (GLCM): at each location, we got a GLCM image, where we computed 24 Haralick descriptors [11]: Sum Squares, Sum Entropy, Sum Average, MCC (Maximal Correlation Coefficient), Maximum Probability, Joint Entropy, Joint Energy, Joint Average, Inverse Variance, IMC (Informational Measure of Correlation) 2, IMC (Informational Measure of Correlation) 1, IDN (Inverse Difference Normalized), IDM (Inverse Difference Moment), ID (Inverse Difference), Difference Variance, Difference Entropy, Difference Average, Correlation, Contrast, Cluster Tendency, Cluster Shade and Cluster Prominence. All features equations were reported in Table 3.3 and Table 3.4, respectively. This procedure returned 42 parametric images (18 First-order + 24 GLCM) for each CXR image, where we finally computed seven statistics, namely: mean, median, variance, skewness, kurtosis, energy and entropy. This resulted in 294 image features (i.e. 7 statistics by 42 parametric maps).

Table 3.3: Definition of the first-order statistical measures. Notation: X is a set of N_p pixel in a ROI; $S(i)$ is the first order histogram of the ROI using N_g discrete intensity levels, equally spaced from 0 with a defined width of 0.1; $s(i) = \frac{S(i)}{N_p}$ is the normalized first order histogram; V_{pixel} is the volume of a pixel in mm; X_{10} is the 10^{th} percentile of X; X_{90} is the 90^{th} percentile of X; X_{10-90} is the image array with gray levels in between, or equal to the 10^{th} and 90^{th} percentile of X; $\overline{X}$ is the mean value of the image array.

Feature	**Definition**
Energy	$\sum_{i=0}^{N_p} X(i)^2$
Total Energy	$V_{pixel} \cdot \sum_{i=0}^{N_p} X(i)^2$
Entropy	$-\sum_{i=1}^{N_g} s(i) \cdot \log[s(i)]$, for $s(i) > 0$
Minimum	$min(X)$
Maximum	$max(X)$
Mean	$\frac{1}{N_p} \sum_{i=1}^{N_p} X(i)$
Median	median grey level intensity
Interquartile Range	$X_{75} - X_{25}$
Range	$max(X) - min(X)$
Mean Absolute Deviation	$\frac{1}{N_p} \cdot \sum_{i=1}^{N_p} \lvert X(i) - \overline{X} \rvert$
Robust Mean Absolute Deviation	$\frac{1}{N_{10-90}} \cdot \sum_{i=1}^{N_{10-90}} \lvert X_{10-90}(i) - \overline{X}_{10-90} \rvert$
Root Mean Squared	$\sqrt{(\frac{1}{N_p} \cdot \sum_{i=1}^{N_p} X(i)^2)}$
Skewness	$\frac{\frac{1}{N_p} \cdot \sum_{i=1}^{N_p} (X(i) - \overline{X})^3}{(\sqrt{\frac{1}{N_p} \cdot \sum_{i=1}^{N_p} (X(i) - \overline{X})^2})^3}$
Kurtosis	$\frac{\frac{1}{N_p} \cdot \sum_{i=1}^{N_p} (X(i) - \overline{X})^4}{(\frac{1}{N_p} \cdot \sum_{i=1}^{N_p} (X(i) - \overline{X})^2)^2}$
Variance	$\frac{1}{N_p} \cdot \sum_{i=1}^{N_p} (X(i) - \overline{X})^2$
Uniformity	$\sum_{i=1}^{N_g} s(i)^2$

Table 3.4: Definition of the second-order statistical measures. Notation: $P(i,j)$ co-occurence matrix with a defined distance (δ= 1) and angle (θ=0); $p(i,j) = \frac{P(i,j)}{\sum P(i,j)}$ is the normalized co-occurence matrix; $p_x(i) = \sum_{j=1}^{N_g} P(i,j)$ and $p_y(i) = \sum_{i=1}^{N_g} P(i,j)$ are the marginal probabilities per row and per column, respectively; μ_x and μ_y are the mean grey level intensities, defined as Joined Average, of p_x and p_y respectively. If $P(i,j)$ is symmetrical $p_x = p_y$; σ_x and σ_y are the standard deviations of p_x and p_y respectively; $p_{x+y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)$, where $i + j = k$, and $k = 2, 3, .., 2N_g$; $p_{x-y}(k) = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)$, where $|i - j| = k$, and $k = 0, 1, .., N_g - 1$; HX, HY and HXY are the entropy of p_x, p_y and $p(i,j)$, respectively. $HXY1 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \cdot \log[p_x(i)p_y(j)]$ is an auxiliary quantity; $HXY2 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p_x(i)p_y(j) \cdot \log[p_x(i)p_y(j)]$ is an auxiliary quantity; DA is the Difference Average used to obtain the Difference Variance.

Feature	**Definition**
Sum Squares	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i - \mu_x)^2 \cdot p(i,j)$
Sum Entropy	$\sum_{k=2}^{2N_g} p_{x+y}(k) \cdot \log[p_{x+y}(k)]$, for $p_{x+y}(k) > 0$
Sum Average	$\sum_{k=2}^{2N_g} p_{x+y}(k)k$
MCC (Maximal Correlation Coefficient)	$\sqrt{\text{second largest eigenvalue of Q}}$, where $Q(i,j) = \sum_{k=0}^{N_g} \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}$
Maximum Probability	$\max(p(i,j))$
Joint Entropy	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \log[p(i,j)]$, for $p(i,j) > 0$
Joint Energy	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (p(i,j))^2$
Joint Average	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j)i$
Inverse Variance	$\sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{k^2}$
IMC (Informational Measure of Correlation) 2	$\sqrt{1 - e^{-2(HXY2 - HXY)}}$
IMC (Informational Measure of Correlation) 1	$\frac{HXY - HXY1}{max\{HX, HY\}}$
IDN (Inverse Difference Normalized)	$\sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+(\frac{k}{N_g})}$
IDM (Inverse Difference Moment)	$\sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+k^2}$
ID (Inverse Difference)	$\sum_{k=0}^{N_g-1} \frac{p_{x-y}(k)}{1+k}$
Difference Variance	$\sum_{k=0}^{N_g-1} (k - DA)^2 \cdot p_{x-y}(k)$
Difference Entropy	$\sum_{k=0}^{N_g-1} k \cdot p_{x-y}(k) \log[p_{x-y}(k)]$, for $p_{x-y}(k) > 0$
Difference Average	$\sum_{k=0}^{N_g-1} k \cdot p_{x-y}(k)$
Correlation	$\frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j) \cdot i \cdot j - \mu_x \mu_y}{\sigma_x(i)\sigma_y(j)}$
Contrast	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i - j)^2 \cdot p(i,j)$
Cluster Tendency	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i + j - \mu_x - \mu_y)^2 \cdot p(i,j)$
Cluster Shade	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i + j - \mu_x - \mu_y)^3 \cdot p(i,j)$
Cluster Prominence	$\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} (i + j - \mu_x - \mu_y)^4 \cdot p(i,j)$

To cope with the large number of descriptors, we proceeded as described in section 3.1.1.2.3, adopting the base learners already described there. Then, for each tested classifier, given the set and number of descriptors selected by the wrapper approach in the nested cross-validation fashion, we trained the same classifier on

the whole training fold and measured recognition performance on the test fold.

3.1.1.3 Results

This section reports the results attained using the described approach in staging the patients with COVID-19 in severe and mild classes. The goal is to provide a baseline characterisation of the performance achieved by integrating quantitative image data with clinical information using state-of-the-art approaches.

Table 3.5 and Table 3.6 presents all recognition performance attained by the learning methods when the experiments were executed according to the 10-fold and LOCO cross validation, respectively (see section 3.1.1.2.4 for further details). In the former case, the results are averaged over the 20 repetitions.

Furthermore, to see if there exists a statistically significant difference between the various performance, we ran the Kruskal-Wallis and the Dunn test with Bonferroni correction for multiple comparisons ($p < 0.05$): the results were reported in Figure 3.3. The statistical analysis shows that in almost all the experiments, the results achieved by the three learners (LGR, SVM and RF) are not different at the given significance level.

Table 3.5: Recognition performance of the handcrafted approach achieved by all the learners considered, and specified in the last column of the table, when the experiments were executed according to the 10-fold cross-validation (20 repetitions).

Input Data	**Accuracy**	**Sensitivity**	**Specificity**	**Learner**
CXR images	$.658 \pm .06$	$.676 \pm .077$	$.639 \pm .088$	LGR
	$.653 \pm .053$	$.658 \pm .073$	$.648 \pm .083$	SVM
	$.653 \pm .054$	$.667 \pm .082$	$.64 \pm .081$	RF
Clinical data and CXR images	$.755 \pm .047$	$.761 \pm .064$	$.749 \pm .069$	LGR
	$.756 \pm .045$	$.759 \pm .069$	$.753 \pm .067$	SVM
	$.751 \pm .047$	$.75 \pm .067$	$.755 \pm .075$	RF

Table 3.6: Recognition performance of the handcrafted approach achieved by all the learners considered, and specified in the last column of the table, when the experiments were executed according to the LOCO cross-validation.

Input Data	**Accuracy**	**Sensitivity**	**Specificity**	**Learner**
CXR images	$.619 \pm .07$	$.644 \pm .142$	$.62 \pm .191$	LGR
	$.625 \pm .083$	$.641 \pm .159$	$.644 \pm .12$	SVM
	$.622 \pm .066$	$.619 \pm .107$	$.652 \pm .139$	RF
Clinical data and CXR images	$.752 \pm .067$	$.711 \pm .165$	$.824 \pm .154$	LGR
	$.746 \pm .03$	$.741 \pm .122$	$.757 \pm .129$	SVM
	$.691 \pm .056$	$.705 \pm .159$	$.754 \pm .196$	RF

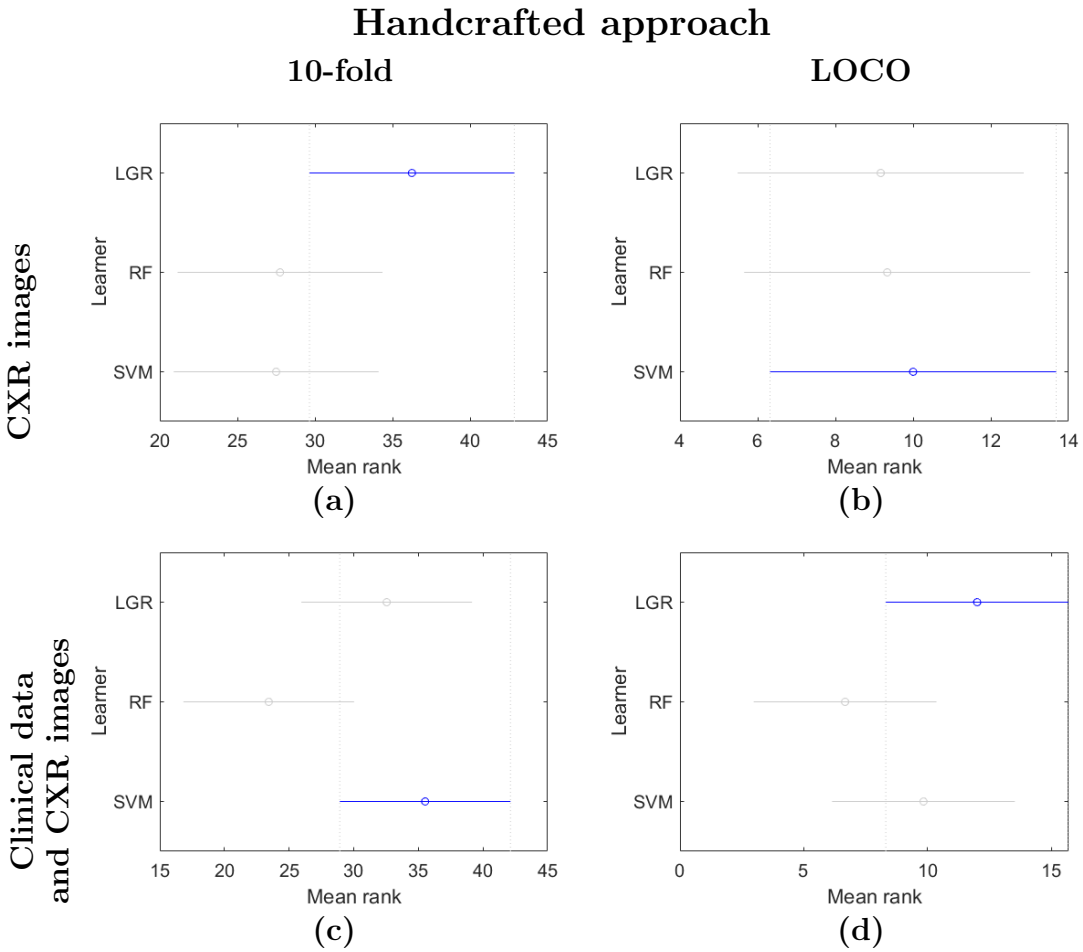


Figure 3.3: Mean rank results of the learners used in the the handcrafted approach.
In blue the learner which gave the best result reported in Tables 3.5 and 3.6.

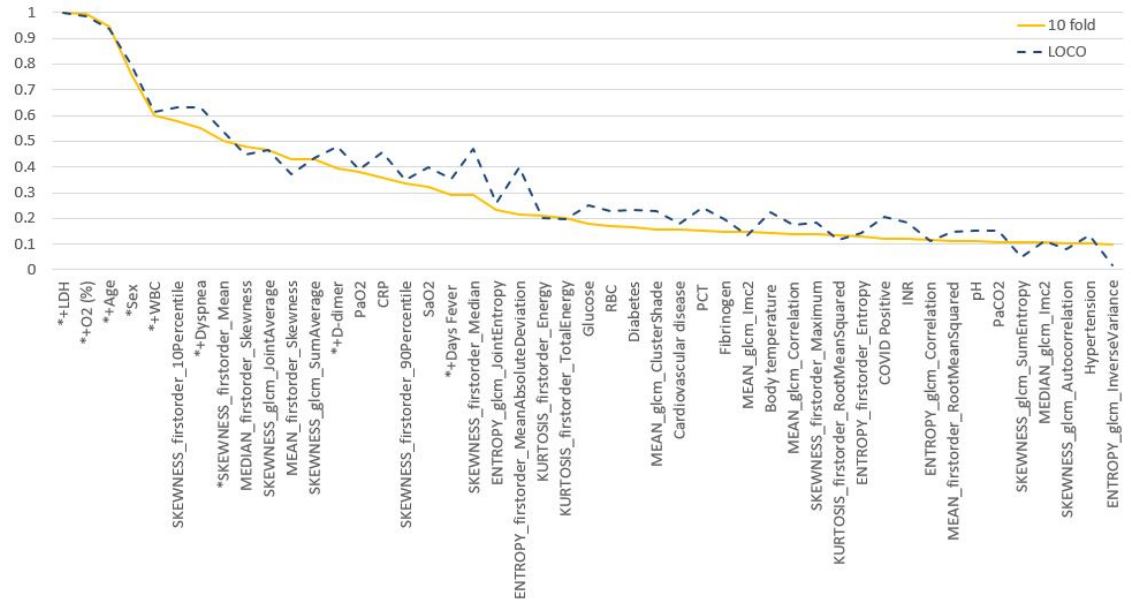


Figure 3.4: Importance of clinical and handcrafted measured as the rate each descriptor was selected by the RFECV wrapper
during the 10-fold and LOCO cross-validation experiments considering all the three classifiers employed. The y axis scale is normalized to one. Moreover, we add a "*" or a "+" before each feature name if it is included in the feature set used to get the best handcrafted results reported in the last section of Table 3.5 and 3.6, respectively.

Furthermore, Figure 3.4 shows the feature importance of the 40 most selected handcrafted descriptors by the RFECV wrapper during the experiments in 10-fold and LOCO cross-validation. The feature relevance is computed as the number of times a feature is included in the selected subset during all the experiments performed using all learners, and for the sake of clarity, values are normalized in [0,1]. The plot shows that the top-five descriptors most frequently detected as discriminative are clinical measures, followed by several texture measures almost equally distributed between the first- and second-order measures. For the sake of completeness, in this figure on the x-axis we add a "*" or a "+" before each feature name when it is included in the feature set used to get the best results by combining handcrafted measures from CXR images and clinical data, reported in Table 3.5 and 3.6, respectively.

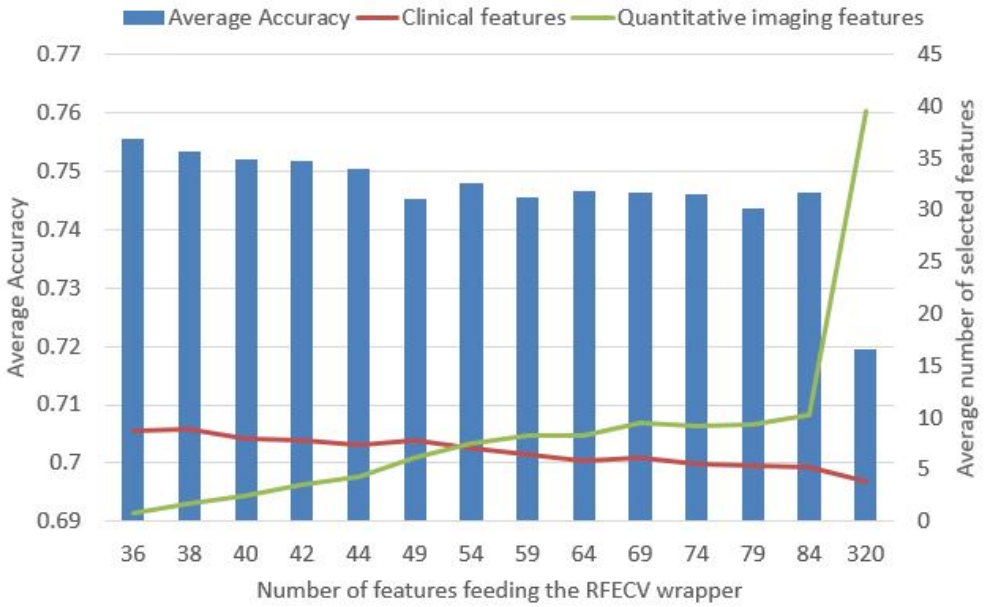


Figure 3.5: Variation of the average classification accuracy (blue bars) with the number of features feeding the RFECV wrapper.
The red and green curves show the number of clinical and texture features selected by the RFECV wrapper, respectively. The experiments plotted here refer to the best results shown in Table 3.5 integrating clinical and imaging features for the handcrafted approach.

Table 3.7: Recognition performance of the handcrafted approach achieved by all the learners considered, and specified in the last column of the table, when the experiments were executed using only clinical data and according to the 10-fold cross-validation (20 repetitions) and the LOCO cross-validation.

Validation	**Accuracy**	**Sensitivity**	**Specificity**	**Learner**
	$.755 \pm .049$	$.757 \pm .063$	$.755 \pm .076$	LGR
10-fold	$.757 \pm .045$	$.761 \pm .063$	$.755 \pm .073$	SVM
	$.750 \pm .041$	$.755 \pm .061$	$.748 \pm .068$	RF
	$.707 \pm .045$	$.677 \pm .172$	$.778 \pm .150$	LGR
LOCO	$.734 \pm .044$	$.699 \pm .158$	$.795 \pm .136$	SVM
	$.656 \pm .073$	$.666 \pm .226$	$.739 \pm .231$	RF

We now analyse how the performance of the handcrafted approach varies with the number of features selected by the coarse step, which fed the fine selection based on the RFECV method, as described in section 3.1.1.2.5. To this end, Figure 3.5 reports on the x-axis the number of features in input to the RFECV, which ranges from 36 (i.e. 34 clinical plus 2 texture measures) up to 84 (i.e. 34 clinical plus 50 texture measures), plus the last value where the RFECV received all the clinical and

all the image features[1]. The bars show the average classification accuracy (y-axis, left side), while the curves in red and green show the average number of clinical and handcrafted texture features selected by the RFECV, respectively (y-axis, right side).

Furthermore, to analyse the added value of the handcrafted features, performance in discriminating between patients with mild and severe prognosis attained using clinical data only was analysed and reported in Table 3.7. In this respect, the table shows the best performance achieved by the RFECV and by the learners described in the last part of section 3.1.1.2.5. In the case of experiments performed in 10-fold cross-validation (Table 3.7), the best accuracy is up to 75.7%, it is attained by an SVM retaining on average 11 clinical features, and the sensitivity and the specificity are almost balanced. This latter observation can be expected since the a-priori class distribution is not skewed. The same observations also hold in the case of the experiments performed in a LOCO modality, and it is worth noticing the performance drops. This can be due to the variation of data distribution among the centres, limiting the generalisation capability of the learners.

Finally, Figure 3.6 shows the rate each clinical descriptor was included in the selected feature subset by the RFECV wrapper, distinguishing also per classifier used, using a normalised unitary scale. The figure shows the cumulative results observed running both the 10-fold and LOCO cross validation experiments. We opted for this cumulative representation since the trend is very similar in both the experiments. Furthermore, also the set of biomarkers providing the best performance shown in the first section of Table 3.5 and 3.6, which are denoted by reporting before an "*" or a "+" for 10-fold and LOCO cross validation experiments are reported, respectively. Interestingly, Figure 3.6 shows that age, LDH and O_2, were chosen in every fold for all the classifiers. If we used only such three descriptors, the average classification accuracy attained by learners in 10-fold and LOCO cross validation is equal to 0.74 ± 0.05 and to 0.70 ± 0.10, respectively. Moreover, sex, dyspnoea and WBC were always selected by the wrapper with the SVM and RF, whereas the D-dimer was always selected by the logistic regressor and by SVM. Oppositely, heart failure and cough were scarcely selected. Notably, some features such as LDH, D-dimer and SaO_2 were selected very frequently despite a high fraction of data was obtained by imputation (see Table 3.2). We deem that is mostly related to the strong differences in the distributions of these features between the two classes.

As already noticed in Table 3.5, the use of texture measures do not improve the performance attained using the clinical descriptors reported in Table 3.7; this is also confirmed by observing that, as the number of input features increases, the wrapper tends to select more imaging biomarkers than clinical ones, dropping the

[1]The experiments plotted in Figure 3.5 refer to the best results shown in Table 3.5 integrating clinical and imaging features by the handcrafted approach.

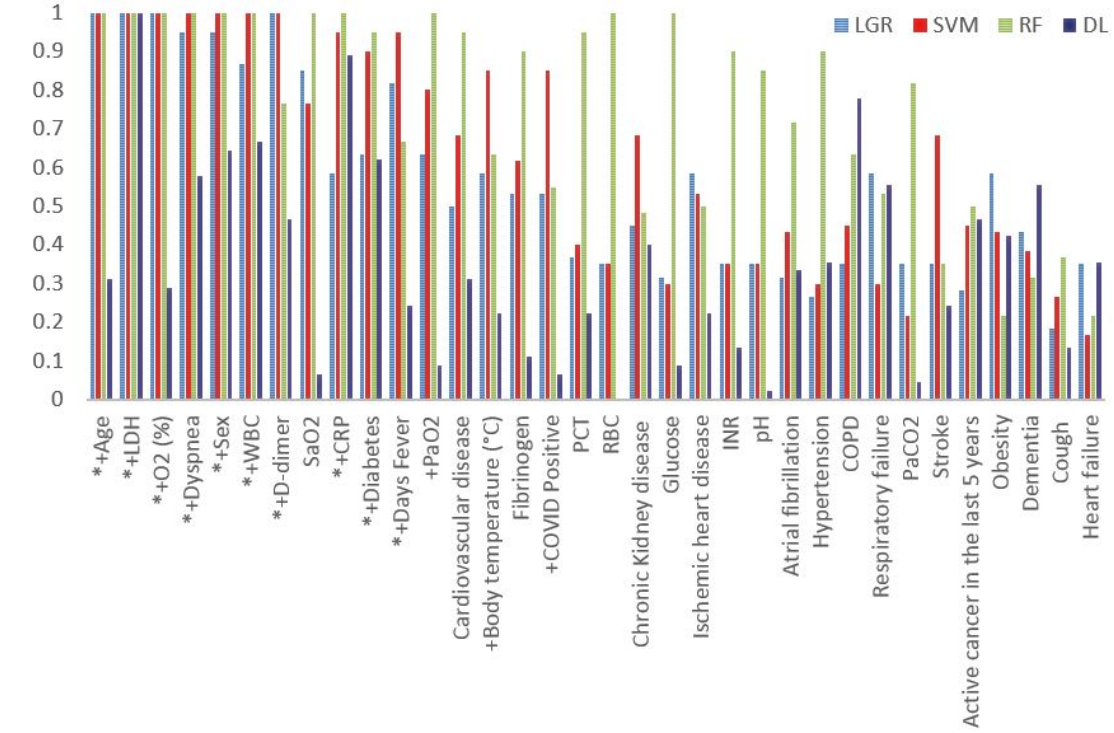


Figure 3.6: Clinical feature importance represented by the rate each descriptor was selected by the RFECV wrapper during both the 10-fold and LOCO cross validation experiments using the three classifiers (LGR, SVM and RF series). The DL series represents feature importance estimated as the maximum absolute value of weights in the first layer of the perceptron of the DL network, after averaging over folds and repetitions and rescaling in the [0,1] interval. Moreover, the "*" or a "+" reported before each feature name means that it is included in the feature set used to get the best handcrafted results reported in Table 3.7.

performance. This may remark the importance of using both clinical and imaging biomarkers since they may provide complementary information: while the former, and especially comorbidities, refers to the functional reserve of the patient, the latter may quantify the actual impact on the lungs. Indeed, fit patients with severe infection and damage are as likely as unfit patients with less severe infections to have a poor prognosis. Although not reported, similar considerations can be derived in the case of LOCO cross-validation, where we noticed that the best performances are attained by an almost balanced number of clinical and imaging features.

The approach that computes handcrafted features from the images also unfavourably compares with those using CNNs. Indeed, comparing with the hybrid and the end-to-end DL approaches, we found that the performances are statistically different in both the 10-fold and LOCO cross-validation tests, as we always got $p < 0.05$.

Finally, a comparison of the handcrafted approach using Clinical data and CXR images was done comparing the following validations methods: 10-fold cross vali-

dation, 3-fold cross validation, 10 fold Shufflesplit validation and 3-fold Shufflesplit validation. Results are shown in the following table, were the best mean accuracy and mean Balanced accuracy were found using the 10-fold cross validation, confirming the choice of the manuscipt.

Table 3.8: Comparison of four different validation methods using both Clinical data and CXR images analysed with the handcrafted method.

		3 Fold		10 Fold	
typefold	**Classifier**	**Average of Accuracy**	**Average of Balanced Accuracy**	**Average of Accuracy**	**Average of Balanced Accuracy**
RepeatedKFold	LGR	0.7405	0.7400	0.7497	0.7495
	RF	0.7285	0.7299	0.7349	0.7364
	SVM	0.7314	0.7316	0.7409	0.7410
Shufflesplit	LGR	0.7320	0.7324	0.7385	0.7390
	RF	0.7444	0.7469	0.7355	0.7382
	SVM	0.7311	0.7311	0.7361	0.7366

3.1.2 Local Binary Patterns (LBP)

In general, the introduction and implementation of new 2D texture features could give further information than the features usually implemented for radiomics applications developed by Haralick et al. [11]. In training and testing approaches, it is essential to consider that textures in the test set could be different in respect to the training set under various aspects, the spatial scale, the orientation or the grey-scale properties as introduced by Ojala et al. [107]. This has inspired the development of a new category of features referred to as *Local Binary Patterns* (LBP) and developed by Ojala et al.. These features are supposed to be grey-scale and rotation invariant and computationally simple to achieve.

3.1.2.1 Local Binary Patterns

The LBP features are developed to find features capable of describing the local properties of an image to identify the local pattern of each part of the image. The idea behind the features is to use binary codes to represent the local pattern of the image. Due to this idea, LBP features can detect microstructures such as spots, lines and edges.

3.1.2.1.1 LBP construction

For the LBP construction, considering a bi-dimensional image I, the intensity I_p of each pixel p is compared with the intensity I_j of the neighbourhood of the given pixel in the pixel connectivity theory where the distance D between the given pixel and the neighbours is given by r. In the example of Figure 3.7 the parameters used where

$r = 1$ and $p = 8$. If $I_j > I_p$, the jth pixel is set to 1, 0 otherwise. Next, it is possible to process all p's neighbours in a circular direction, interpreting the sequence of 0s and 1s as a binary string and setting the value of p to the equivalent decimal value. This procedure was then repeated until the corresponding decimal value replaced all original image values. The number of patterns for this 2D implementation is 2^P, with P denoting the number of neighbouring local points around the central point.

Proceeding as described through all of the pixels of I, we obtain a new image encoding the intensity distribution of each pixel with respect to its neighbours that can describe the texture of the original image.

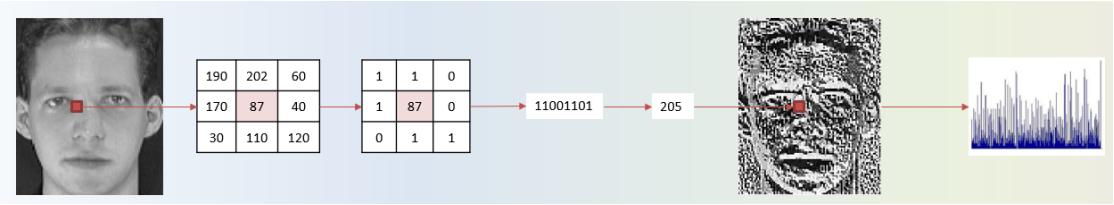


Figure 3.7: LBP construction

Starting from the obtained LBP image all values were then reported on a histogram containing the distribution of the values in the image. From this histogram, features were extracted.

Based on the standard LBP construction, a few variations have been developed to obtain rotation invariant and grey-scale invariant features.

Grey-Scale Invariance The first step to achieve grey-scale invariance is to subtract the value of the central pixel from all surrounding pixels. A grey-scale invariance can be obtained by considering only the sign of the difference instead of the exact values. The second step is to replace all negative values with 0 and positive values and zeros to 1. The obtained LBP value, given by the binary value obtained from the string of zeros and ones, is no longer affected by the grey-scale.

Rotation Invariance The rotation invariance is a second import aspect of the LBP construction since the same pixel distribution rotated along the perimeter of a circle around the central point gives a different LBP value. The pattern where the surrounding points only give zeros or ones were not affected by the rotation. The solution was to rotate all pixels clockwise until the maximal number of most relevant bits were 0.

3.1.2.1.2 Volume Local Binary Patterns (VLBP)

To extend a 2D LBP to the 3D environment, a first solution consists in considering a helix neighbourhood of each voxel [108]. However, to avoid the high increase of the computational burden, associated with the very large number of patterns for volume

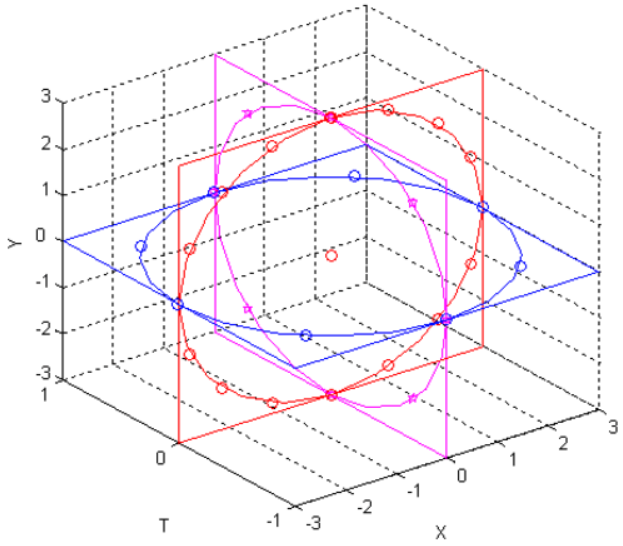


Figure 3.8: Visual representation of a TOP-LBP example
In this Figure each ellipsoid shows a planar LBP calculated on a specific orthogonal plane.

LBP, when P increases, as 2^{3P+2}, another 3D implementation of LBP transformation that was also shown to work fine [108] was developed and referred to as TOP-LBP.

TOP-LBP The first idea was to apply the 2D LBP to the three orthogonal planes which go through the central point. The single histograms obtained for the three orthogonal planes were then concatenated, and subsequently, features were extracted from the entire histogram. It considers the co-occurrence on three orthogonal planes crossing the centre of the analysed volume, as depicted in Figure 3.8.

VLBP Zhao and Pietikainen [108] developed the idea to use instead of a single slice, a small volume around the central point with a number of slices depending on the value r. As represented in Figure 3.9 points around the central point were chosen and a spiral, crossing all points was created. Considering $p = 4$, in total 14 points were chosen, 5 in the two external slices and 4 for the middle slice. After the comparison of the considered values with the central point, a string containing all binary values was finally obtained and a decimal value was calculated.

3.1.2.2 Application of LBP to clinical problem

LBP is a texture measure successfully used in computer vision but was barely used for medical applications. To evidence the value of these features also in medical applications, we extended the set of texture features typically used in the radiomics literature, which are based on the two-dimensional and 3D second-order joint probability functions and their derivatives (introduced in section 2.1.2.3) with 3D LBP

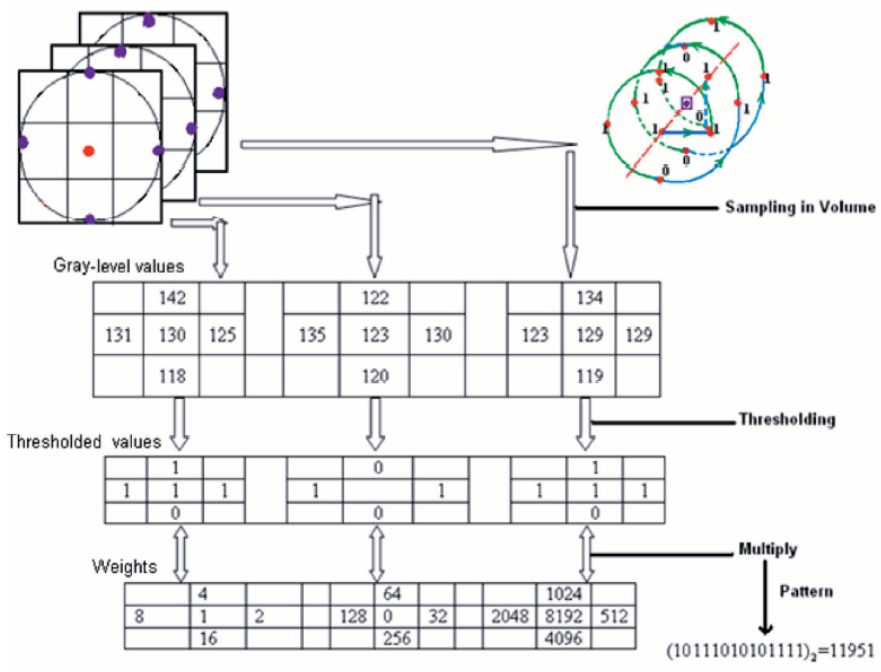


Figure 3.9: VLBP procedure, developed by Zhao and Pietikainen[108] for volume LBP.

features. Furthermore, to deepen the analysis, the approach developed was compared to a deep approach, where features were automatically computed, as well as to the signature proposed by Aerts et al. [8], which is the most cited at the state-of-the-art. The study aimed to predict the overall survival (OS) time of the included patients affected by Non-small cell lung cancer (NSCLC).

3.1.2.2.1 Material and Methods

DATASET The developed work studied 97 patients with NSCLC stage III treated with definitive concurrent chemoradiotherapy. The enrollment protocol was approved by the Ethical Committee of Campus Bio-Medico University on 30 October 2012 and registered at ClinicalTrials.gov on 12 July 2018 with Identifier NCT03583723 after an initial exploratory phase. The Institutional review board approved this analysis and written informed consent was collected from all the patients. For each patient, the simulation CT images were acquired before the treatment using a Siemens Somatom Emotion, with 140 Kv, 80 mAs, and 3 mm for slice thickness. Subsequently, all images were preprocessed using a lung filter (kernel B70) and a mediastinum filter (kernel B31). The patients were then clinically followed, for a median follow-up of 18.55 months after that, it was possible to divide patients into two classes: 53 dead (class 0) and 44 alive (class 1). The patients had a mean Overall Survival (OS) time of 28.7 ± 26.4 months (min 4.8 months, max 142.6 months).

In this study, we considered three different VOIs for each patient, the Gross Tumour Volume (GTV), the Clinical Target Volume (CTV), and the Planning Target Volume (PTV) segmented by radiation oncologists. The results obtained for the different used segmentation will be deepened in section 3.4.

FEATURE COMPUTATION Besides the LBP features, the first-order statistical features and the 3D Grey level co-occurrence matrix features were obtained. The features were computed using an in-house software tool coded in MATLAB R2019a (The MathWorks, Inc., Natick, Massachusetts, United States) that calculates the first-order statistical features and two families of textural features: 3D Grey Level Co-occurrence Matrix (GLCM) and Three Orthogonal Planes-Local Binary Patterns (TOP-LBP) (section 3.1.2.1.2)

The First-order features were computed from the grey levels' histogram of an ROI and therefore described the statistical distribution of tissue's density inside the volume. From such histograms, we extracted 12 descriptors, which were the moments from first to fourth-order, namely the mean, the standard deviation, the skewness and the kurtosis, the histogram width, the energy, the entropy, the value of the histogram absolute maximum and the corresponding grey-level value, the energy around such maximum, and the number of relative maxima in the histogram and their energy. The healthy and the cancerous tissues usually present different structural patterns and for this reason, we go beyond the distribution of the grey levels of the VOIs voxels performing additional textural analysis. In this respect, we computed the 3D Grey Level Co-occurrence Matrix, which generalises the 2D GLCM to the third dimension, catching the differences of the tissues at the micro-scale. Finally, from each GLCM3, we extract seven second-order statistical features, referred to as Haralick descriptors [11]. Namely, the autocorrelation, the homogeneity, the entropy, the energy, the covariance, the inertia, and the absolute contrast. Concatenating such measures for each GLCM3 we get $13 \times 7 = 91$ textural descriptors per patient.

As previously introduced, the TOP-LBP features were implemented. These features are based on the computation of three 2D LBPs, each derived from each of the three planes; their histograms are then computed and concatenated to obtain a unique histogram for the specific volume. This conspicuously alleviates the computational load since the number of patterns for TOP-LBP is 3×2^P. Furthermore, in our LBP implementation, we consider two more variants to cope with the other two issues of 2D LBP definition. First, we computed rotation invariant LBP, i.e., all of the binary strings obtained as the circular shift of a fundamental string are considered the same. Second, we implemented a uniform version of LBP, i.e., all binary strings containing more than two crossings from 0 to 1 or from 1 to 0 are considered not uniform and coded with a specific string. In this case, setting $P = 8$,

we get 48 features by computing first-order measures from each of the three 2D LBP.

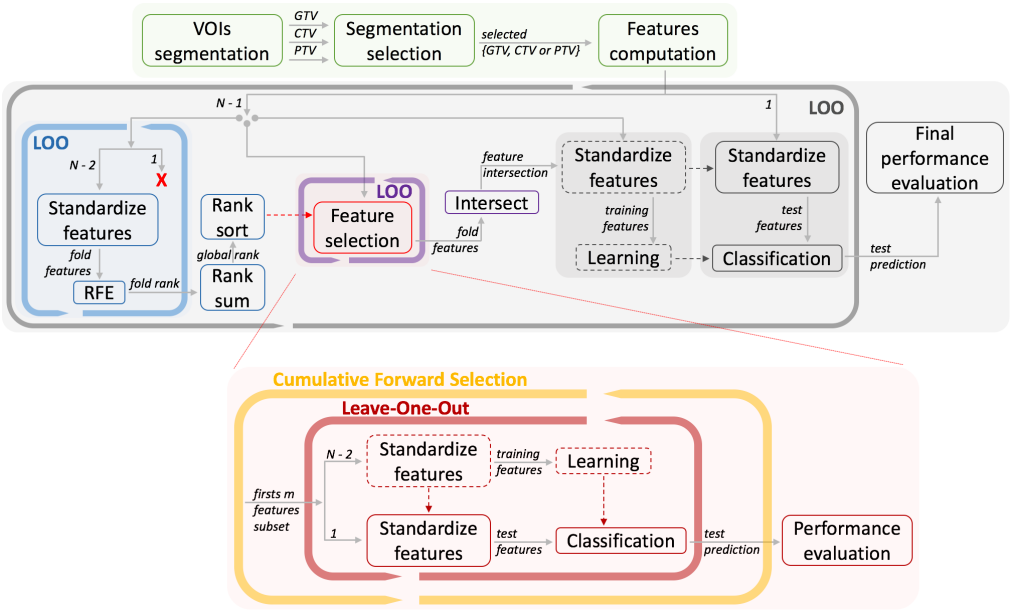


Figure 3.10: Graphical representation of the method, where the different colours refer to different steps of the pipeline.
List of abbreviations: VOI: Volume of Interest, LOO: Leave-One-Out, RFE: Recursive Feature Elimination, N: number of patients, GTV: Gross Tumour Volume, CTV: Clinical Target Volume, PTV: Planning Target Volume.

Features Selection and Classification An important step during a radiomics approach is the features selection step: it helps in reducing the dimensionality of the problem, lowering the curse of dimensionality and the risk of overfitting, and it helps finding the most informative set of descriptors for the problem at hand.

The feature selection pipeline, represented in the blue, purple and red blocks of Figure 3.10, consists of a wrapper-based approach, which searches and evaluates the best subset of features maximising the performance of a given classification algorithm. To avoid any bias, features are normalised before each step, using a standard scaler, as represented in Figure 3.10. This approach scales features using the following equation: $z = (x - \mu)/\sigma$, with μ mean value, σ standard deviation, and x and z as the original feature and the scaled feature, respectively.

Starting from the blue box in Figure 3.10, the feature set was analysed by the Recursive Feature Elimination (RFE) approach [49], making use of four different classification algorithms, namely AdaBoost, CART Decision Tree (DT), Random Forest (RF), and XGBoost (XGB) [109].

The goal of the RFE is to select a feature subgroup by initially considering all of the features and recursively examining a smaller and smaller dataset according to an assigned feature importance attribute. During this process, the selected classifier

is trained on the standardised descriptor set to compute the feature importance attribute based on the accuracy level of the model. The feature with the lowest attribute is excluded from further analysis. This recursive analysis is performed until the best feature is found, resulting in a final descriptor ranking.

To avoid any bias during the features selection approach, the RFE step was applied according to a nested leave-one-out (LOO) procedure: indeed, referring to Figure 3.10, there is an outer grey loop for final performance computation and an inner blue loop for RFE application. This means that this analysis was repeated $N-1$ times, to obtain a global view of the feature importance. Hence, all of the $N-1$ rankings obtained were summed up to get a final rank of each descriptor.

From this step on, we considered two alternative paths referred to as single ranking and total ranking, where the ranking procedure was done for the single segmentation or global for the three segmentation. The differences of these paths were deepened in section 3.4.2. Turning our attention to the purple and red boxes in Figure 3.10, these represent the feature evaluation workflow to find the best subset. We tested different combinations of features subsets whose performance was evaluated according to a given metric to find out the best descriptors. To this goal, all features were initially sorted by their rank and, then, the subsets were sequentially inspected using a cumulative forward selection approach (yellow loop). The cumulative forward selection search procedure starts from a subset with one feature with the highest rank position and then incrementally adds descriptors with lower ranks until the set contains all the descriptors.

To avoid any bias, each subset was then evaluated using a nested LOO configuration, where the inner loop (red LOO) evaluates the best subset maximising the Area Under the ROC Curve (AUC) metric, whereas the outer loop is again the grey LOO for final performance computation. Hence, the output of the purple box is a list of $N-1$ best subsets found with the nested procedure.

Finally, with this set of descriptors, the outer grey leave-one-out computes the ultimate performance of the system.

We measured the following performance metrics: $accuracy = \frac{TP+TN}{P+N}$, the *area under the ROC curve* (*AUC*), the $precision = \frac{TP}{TP+FP}$ and the $recall = \frac{TP}{TP+FN}$. Straightforwardly, TP, TN, P, and N stands for the true positive, true negative, the total number of positive samples, and the total number of negative samples, respectively. Hereinafter, we also use FP and FN to denote the false positive and false negative, respectively. Let us recall that the positive and negative classes correspond dead and alive patients, respectively.

Comparative Analysis As reported in section 3.1.2.2, we deepen the analysis comparing the proposed approach with the state-of-the-art and with an approach that leverages on image features automatically computed.

Comparison with the State-Of-The-Art As a first point, we compare our approach with the most cited one at the state-of-the-art, represented by the radiomics signature that was proposed by Aerts et al. [8]. This signature includes four features: (i) the "statistics energy" that describes the overall density of the tumour volume, (ii) the "shape compactness" quantifying how compact the tumour shape is, (iii) the "grey level nonuniformity" that measures for intratumour heterogeneity, and (iv) the "grey level nonuniformity HLH" measuring intratumour heterogeneity after decomposing the image in mid-frequencies through wavelet analysis. The formal definition of such measures in the supplementary material of [8], while their implementation is available in Python [110]. We compute this signature on our dataset and, for a fair comparison, all tests are performed in LOO fashion for all of the classifier-VOI combinations.

Comparison with a Deep Learning Approach Besides the comparison with the handcrafted signature at the state-of-the-art, we investigate the possibility to use automatic feature extractors. To this goal, we consider two well-established deep networks, i.e., the AlexNet [111] and the ResNet50 [112], and we train them from scratch on our dataset considering only the CTVs, since they yield the largest performance in the machine learning pipeline proposed here, as shown in section 3.1.2.2.2. On the one hand, the AlexNet has a feature extraction step composed of three consecutive blocks containing two convolutional layers and one max pooling layer. After each convolutional layer, we included a batch normalization layer and a dropout layer, to prevent exploding gradient values and overfitting, respectively. Subsequently, the classification step is designed with two dense layers, one with 256 neurons and the other with one neuron, as it is the output layer. All of the layers implement a LeakyRelu activation function, except for the output that uses a sigmoid for the final classification. On the other hand, the ResNet50 was designed, as reported by He et al. [112], except for the first max-pooling layer, which was eliminated to maximally preserve the information contained in the images, and the output layer that is a dense layer with one neuron with a sigmoid activation function. In both cases, due to the large computational efforts needed, the test procedure was conducted in 10-fold cross-validation, with eight folds as training, one as validation, and one as a test. Here, cross-validation is preferred to LOO to alleviate the computational burden, while having enough samples in the training and validation sets. In this case, the CNNs classify each slice and then a patient label is given by majority voting.

As a further direction of the investigation, we also study what happens using the CNNs only as feature extractors that feed the same four classifiers that were used in our approach. In practice, from both networks pretrained on our dataset, we use as feature set the last layer in the feature extraction blocks: this yields 256

descriptors from the dense layer of AlexNet and 8192 descriptors from the average pooling layer of ResNet50. These two feature sets separately fed the four classifiers used in the proposed pipeline, which were trained to classify each 2D slice. Again, the patient label is given by majority voting and the performance are computed in 10-fold cross validation

3.1.2.2.2 Results

The main experiment described in the previous section was applied to all possible combinations of classifiers and VOIs where the features were computed. Furthermore, the experiments were run also using both the single ranking and the total ranking methods for feature selection. The use of four classifiers, three VOIs, and two feature selection ranking approaches results in twenty-four experiments, whose results are presented in Table 3.9.

Table 3.9: Performance in all possible combinations with and without LBP
The upper panel refers to the single ranking strategy for feature selection, whereas the lower panel shows the results considering the total ranking option. System performance in all possible combinations including and excluding LBP descriptors from the final feature subset in all possible combinations for single ranking (upper panel) and total ranking (lower panel). The notation '-' indicates that no feature was left in the final set after excluding LBP descriptors. The largest performance is highlighted in bold.

			AdaBoost			**Decision Tree**			**Random Forest**			**XGBoost**		
			GTV	**CTV**	**PTV**	**GTV**	**CTV**	**PTV**	**GTV**	**CTV**	**PTV**	**GTV**	**CTV**	**PTV**
Single Ranking	All features	Accuracy	70.10	**83.51**	71.13	70.10	60.82	63.92	73.20	78.35	77.32	68.04	71.13	76.29
		AUC	70.13	**82.78**	70.11	70.13	61.06	64.09	71.81	77.29	76.54	66.70	70.11	75.02
		Precision	69.81	**90.57**	81.13	69.81	58.49	62.26	86.79	88.68	84.91	67.19	70.49	73.44
		Recall	74.00	**81.36**	70.49	74.00	65.96	68.75	70.77	75.81	76.27	81.13	81.13	88.68
	without LBP	Accuracy	**76.29**	69.07	62.89	71.13	48.45	72.16	70.10	69.07	69.07	55.67	69.07	71.13
		AUC	**75.99**	68.22	62.75	70.50	47.81	72.02	68.01	66.87	68.42	51.71	68.22	69.14
		Precision	**77.78**	69.49	66.67	71.93	52.73	75.00	66.67	65.75	70.18	55.56	69.49	67.61
		Recall	**79.25**	77.36	64.15	77.36	54.72	73.58	90.57	90.57	75.47	94.34	77.36	90.57
Total Ranking	All features	Accuracy	64.95	72.16	68.04	44.33	58.76	65.98	75.26	**77.32**	74.23	64.95	72.16	74.23
		AUC	64.84	71.83	67.47	44.43	59.18	65.59	73.89	**75.96**	73.33	62.52	70.09	72.94
		Precision	66.04	75.47	73.58	43.40	54.72	69.81	88.68	**90.57**	83.02	62.67	68.06	71.88
		Recall	68.63	74.07	69.64	48.94	64.44	68.52	72.31	**73.85**	73.33	88.68	92.45	86.79
	without LBP	Accuracy	74.23	73.20	52.58	61.86	-	-	**75.26**	72.16	71.13	56.70	32.99	67.01
		AUC	73.52	72.77	52.17	61.23	-	-	**74.08**	70.28	68.95	54.97	36.36	64.02
		Precision	74.14	74.55	56.60	64.29	-	-	**73.02**	68.57	67.12	58.21	0	62.96
		Recall	81.13	77.36	56.60	67.92	-	-	**86.79**	90.57	92.45	73.58	0	96.23

Turning our attention to the newly introduced set of features, i.e., 3D LBPs, and still focusing on the single ranking approach, these descriptors were always chosen in the best feature subsets, regardless of the classifier and the VOIs under consideration. For the sake of brevity, we hereby do not report the full list of features chosen in each of the 24 experiments. However, to provide a synthetic analysis of

the best feature sets, the second and the third columns of Table 3.10 show how many times each feature in the first column (i.e., the radiomics signature of the best case) was included in the best feature set using the Adaboost with the GTV and PTV (second column) and the DT, RF, and XGBoost with the CTV (third column). Note that finding zeros in the occurrences of Table 3.10 does not exclude the presence of other LBP features in the best subsets besides those chosen in the experiment with largest performance.

Table 3.10: Best radiomics signatures.
The fist column lists the feature selected in the best performance. The other two columns indicate the number of times that each descriptor was selected in the best feature set. Abbreviations: U—uniform, RI—rotation invariant.

Features Selected for AdaBoost CTV	#occurrences in AdaBoost with GTV and PTV	#occurrences in DT, RF and XGBoost with CTV
U 3D LBP kurtosis	0	3
3D LBP energy	0	1
RI 3D LBP maxAss	0	0
3D LBP energy around maxAss	0	1
RI 3D LBP energy	0	1
U 3D LBP energy around maxRel	0	1
U 3D LBP entropy	1	3
U 3D LBP skewness	0	3
inverse GLCM (−1,−1,0)	0	1

To deepen the discussion on the effectiveness of 3D LBPs, we excluded them from each signature selected by our approach and train again each learner. The results are reported in Table 3.9, where "-" indicates that the classifier was not trained, since the best set included only 3D LBP descriptors. When comparing these results with the relative panel in Table 3.9 we notice that there is not an evident pattern in the performance differences between the two experiments. Despite that, the use of 3D LBPs improves the largest accuracy obtained: indeed, without them, the overall best accuracy score decreases from 83.51% to 76.29% and from 77.32% to 75.26% for the single ranking and the total ranking, respectively. This trend is also manifested by the other performance measures reported for the best cases, indicating that LBPs give an important contribution to the classification task at hand. Furthermore, we statistically pairwise compared the results reported in Table 3.9 using the Wilcoxon signed-rank test. In the case of single ranking, for AdaBoost+CTV, DT+CTV, RF+CTV, RF+PTV, and XGBoost+GTV the use of 3D LBPs provides performance larger and statistically different from those attained excluding these descriptors ($p < 0.1$). In the case of total ranking, at the same level of confidence, the results of the test show that the use of 3D LBP provides larger and statistically different performance for AdaBoost+PTV and XGBoost+CTV.

Comparison with The-State-Of-The-Art The results were compared against the signature presented by Aerts et al. [8] for OS prediction in NSCLS, as introduced in section 3.1.2.2. Further, to be one of the most cited works in the state-of-the-art, such a radiomics signature was also used by Lambin et al. [7] and Kwan et al. [113].

Table 3.11 reports the performance computed with Aerts et al.'s signature on our dataset: the largest AUC value was obtained considering the CTV with Random Forest as a classifier, and it is equal to 76.92%. Using the same VOI our signature gets an AUC equal to 82.78%, which suggests that the inclusion of the 3D LBP descriptors boosts the discrimination power and leads to larger performance. Even though the best results of Table 3.9 and Table 3.11 are obtained with different classifiers, comparing the corresponding columns of the two tables further validates the previous assertion. Also note that the best combination in Table 3.11 is based on a Random Forest classifier with the CTV, which is also one of the two best configurations in our experiments: this confirms our results that the CTV is the most informative volume of interest.

Table 3.11: Performance with the signature presented by Aerts et al. combined with the classification algorithms used here. The largest performance is highlighted in bold.

	Adaboost			**Decision Tree**			**Random Forest**			**XGBoost**		
	GTV	**CTV**	**PTV**	**GTV**	**CTV**	**PTV**	**GTV**	**CTV**	**PTV**	**GTV**	**CTV**	**PTV**
Accuracy	63.16	67.37	60.00	63.16	66.32	63.16	73.68	**77.89**	73.68	46.32	46.32	46.32
AUC	62.88	67.42	59.94	63.13	66.13	62.88	72.84	**76.92**	72.68	50.00	50.00	50.00
Precision	65.38	70.83	63.27	67.39	68.63	65.38	71.67	**74.19**	70.97	0	0	0
Recall	66.67	66.67	60.78	60.78	68.63	66.67	84.31	**90.20**	86.27	0	0	0

Comparison with a Deep Learning Approach We conducted a comparison with two deep learning approaches to further assess the power of the handcrafted features, as described in section 3.1.2.2.1. Hence, we performed two experiments: the first tests the AlexNet and the ResNet50 as a whole classification approach, whereas the second tests the two-deep networks as feature extractors feeding the same four classifiers used herein-before. All of the results are summarised in Tables 3.12 and 3.13 and, in general, they show that the most performing network is the ResNet50.

In the first experiment, this deep network achieves an AUC equal to 76.49% and an accuracy equal to 71.11%. Furthermore, the performance differences between the ResNet50 and the AlexNet are statistically significant according to the Wilcoxon signed-rank test ($p = 0.0396$).

In the second experiment, when the networks are used as feature extractors, we find that the automatic features they compute combined with the Adaboost classifier

provide accuracy and AUC equal to 73.33% and 73.42%, respectively. Furthermore, using this learner, the performance differences using the features computed by the AlexNet and by ResNet50 are statistically significant with $p < 0.1$ according to the Wilcoxon signed-rank test. We do not find statistical differences in the pairwise comparison between the automatic features while using the other three classifiers.

As it is evident, these values were lower than those achieved by the proposed approach (accuracy = 83.51% and AUC = 82.78%), showing the successful discriminative power of hand-crafted features. We also assess this difference using again the Wilcoxon signed-rank test, comparing our proposal (AdaBoost+CTV in Table 3.9) with the best one provided by the deep learning approach (ResNet50 features in Table 3.12). The results of such two approaches are statistically different with $p < 0.05$. A possible reason for this finding could be the size of our dataset: indeed, this could hinder the power of deep approaches, leading to the superiority of handcrafted features compared to measures automatically learned.

Table 3.12: Performance of the deep features extracted with two state-of-the-art deep neural networks.
The results are computed evaluating the four classification algorithms that were used in the proposed pipeline in 10-fold cross-validation. The largest performance is highlighted in bold.

	ResNet50 features				**AlexNet features**			
	AdaBoost	**DT**	**RF**	**XGB**	**AdaBoost**	**DT**	**RF**	**XGB**
Accuracy	**73.33**	66.67	70.0	71.11	63.33	66.67	63.33	70.0
AUC	**73.42**	66.08	69.5	71.17	63.08	67.67	64.00	70.67
Precision	**72.81**	68.02	68.12	69.94	63.86	70.52	63.89	72.16
Recall	**85.67**	82.17	89.67	89.67	78.17	79.00	89.17	89.17

Table 3.13: Results of the AlexNet and the ResNet50 trained from scratch on our dataset.

	Accuracy	**AUC**	**Precision**	**Recall**
ResNet50	71.11	76.49	69.49	83.67
AlexNet	57.78	64.37	60.38	65.31

3.1.2.2.3 Conclusions

In this experiment, we proposed new radiomics features to predict the Overall Survival in a cohort of patients suffering from locally advanced non-small-cell lung cancer. The LBPs features proved to be determinant, improving the classification performance in comparison compared to other texture measures. The promising result suggested a future direction worthy of investigation.

3.2 Imbalance learning

3.2.1 Definition and possible solutions

An imbalanced dataset is defined as a dataset with an unequal class distribution, where the class with less data is called the minority class and the class with more data is called the majority class. The different distribution of samples in the different classes leads to incorrect results during the ML algorithms application during the learning phase and the subsequent prediction of the outcome. To show the impact of the class distribution on the predictive accuracy, the confusion matrix can be analysed (Figure 3.11) [114].

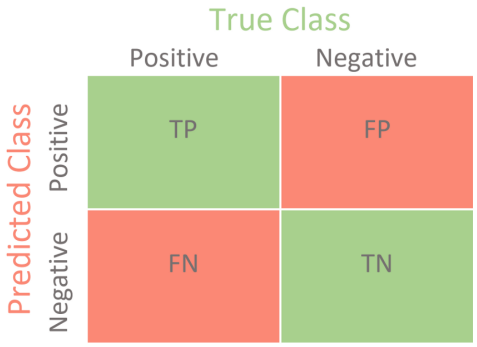


Figure 3.11: Confusion Matrix
The confusion matrix contains an overview of the prediction results of a classification problem. The number of correct and incorrect predictions are represented for each class. TP: true positive, FP: false positive, FN: false negative, TN: true negative

The predictive Accuracy is defined as $Acc = (TP+TN)/(TN+TP+FP+FN)$. For example, consider the classification of pixels in mammogram images as possible cancerous. Typically, 98% of a mammography dataset contains *healthy pixels* while only 2% contains *cancerous pixels*. A simple ML strategy on the majority class would give a 98% of accuracy, which seems to be an excellent result in the first place, but clearly this application requires a high rate of detection for the minority class.

Four main approaches can be used to learn under class skew, namely, an *Over-* and *Under-Sampling* approaches (Figure 3.12), biasing the learning process and multi-experts system.

In the Over-sampling approach, methods creating new "synthetic" samples belonging to the minority class are created to increase the total number of samples belonging to the minority class to restore the balance between the different subgroups. In the Under-sampling method, the majority class samples are deleted to obtain two balanced subgroups. When biasing the learning process a typical ap-

proach applies costs to the model so that such penalities leads the model to pay more attention to the minority class. Finally, multi-expert systems were developed with the aim to overcome the limitations of the available systems. In this system several classifiers are trained on different subsets which included the whole minority class and a subset of the majority class, with the aim to create balanced subsets of data. After that all classifiers were trained, the decisions taken on the test set were combined to obtain the final output [115].

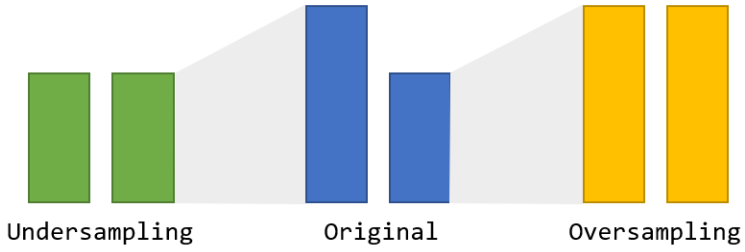


Figure 3.12: Under and Over-sampling: a qualitative example.

3.2.2 Application of imbalance learning on a clinical example

As reported in the previous section, class skew is a problem that affects the performance and results. In this thesis we investigated two alternative methods to learn under class skew: the first uses an Over-sampling apporach (section 3.2.2.1), while the second biases the learning process (section 3.2.2.2). Both methods were applied to the same dataset of acoustic neuromas in MRI images. Acoustic neuroma is a benign, intracranial tumour caused by an overproduction of the Schwann cells surrounding the 8^{th} cranial nerve [116]. The acoustic neuroma is very rare and has an incidence rate of 1.1/100000 cases [117] every year in the US. Initial symptoms include hearing loss and tinnitus, while advanced symptoms are related to increased intracranial pressure.

One of the treatment possibilities is the *stereotactic radiotherapy/ radiosurgery (SRS)*, which can be performed using the CyberKnife device [118], a 6-degrees-of-freedom robotic radiosurgery system that minimises the irradiation to adjacent healthy tissues due to its ability to deliver defined and adapted doses. Currently, after a CyberKnife treatment on an acoustic neuroma, it is necessary to wait up to two years to check if it has been effective or not. The responding rate to the treatment is about 90% and the remaining 10% of the patients need to undergo traditional surgery in case of large tumours. The disadvantage of the radiosurgery is that the tissue becomes fibrotic and with adhesions making a subsequent surgery even more difficult.

On these premises, the study's objective was to predict the response to CyberKnife treatment of patients suffering from acoustic neuroma by analysing MR images acquired before the therapy. This could provide significant advantages in terms of sparing radiation toxicity to patients who will not respond to the treatment. Fortunately, less than 10% of the patients do not respond to the treatment causing an important class imbalancement [119, 120].

DATASET This study includes a cohort of 38 patients presenting an acoustic neuroma treated with CyberKnife at Centro Diagnostico Italiano, Milan, Italy (Table 3.14). Patients have a follow-up up to 10 years and an average age of 61 years. All patients underwent a T1-w 3D MRI with contrast media (ProHance [121]) and, after a few days, a CyberKnife treatment on the monolateral acoustic neuroma. Patients with neuroma on different nerves (e.g. on the 7^{th} facial nerve) were excluded from the study.

Images were acquired on a 1.5T Philips Achieva and a Signa GE 1.5T scanner. Contrast-enhanced T1-weighted images were acquired in the axial plane with both machines. Repetition times were 25 ms and 15.27 ms with the Philips and the GE machine, respectively, while the echo times were 4.5 ms and 6.93 ms. Images used in this study were acquired prior to the CyberKnife treatment.

As SRS aims to reduce the tumour or limit its growth, the physicians ask to predict if the patient will respond to the treatment using quantitative biomarkers extracted from images routinely collected. To this aim, an expert neuroradiologist and an expert radiotherapist compared the pre-treatment MR images and the last available follow-up MR images after the CyberKnife treatment. This permits to divide the population into the following three classes. The first class contains 25 (65.8%) patients who had a volumetric reduction (referred to as *patients with volume reduction*, V.R.); the second class contains 10 patients (26.3%) for whom the size of the lesion had not varied (i.e. *patients with volume stability*, V.S.), whilst the third and last class includes 3 patients (7.9%) with an increase in tumour size (i.e. *patients with volume increase*, V.I.). Note that tumour size was considered stable if the neuroradiologist and the radiotherapist could not find any differences in measuring the two largest diameters of the tumour. The accuracy of this measurement is 0.1mm.

SEGMENTATION In this study a semi-automatic tumour segmentation was carried out using the 3DSlicer image analysis software [122, 123], which is an open source software platform for medical image processing and three-dimensional visualisation. The operator, a trained radiologist, is required to define the segmented region interactively by moving the mouse over the region of interest letting the software automatically adjust the border so that all the pixels have the same or a similar

Table 3.14: Acoustic neuroma dataset.
"V.R." identifies *patients with volume reduction*, "V.S." refers to *patients with volume stability* and "V.I." stands for *patients with volume increase*.Mean age is expressed in years.

Characteristic		**Total**	**Class**		
			V.R.	V.S.	V.I.
Number of Patients		38	25	10	3
Male/Female		18/20	11/14	5/5	2/1
Age (at treatment)	< 50	10	5	4	1
	50-70	14	10	3	1
	70-90	14	10	3	1
Mean age		61.2	63.0	58.3	56.3

grey-level intensity as the seed previously selected. After tumour segmentation, images containing the Volume Of Interest (VOI) were resampled with a linear resampling method to get an isotropic voxel with size equal to $[1 \times 1 \times 1mm]$. This step was performed since texture features can be affected by image resolution [124], a characteristic that varies between the two scanners.

Semantic and Radiomics Feature Computation Before radiomics feature computation, for each patient, age, gender and length of follow-up were collected. Following, we gathered various quantitative biomarkers extracted from the MRI after tumour segmentation.

Shape-based, intensity-based and texture-based features were extracted from the area of interest using the open-software platform IBEX [125]. This choice mitigates a common issue with feature extraction, i.e., the lack of standardised and reproducible tools in order to follow and compare results [62, 126, 127].

Extracted features are divided into the following categories: *Gradient Orientation Histogram*, *Gray level Cooccurrence Matrix*, *Gray Level Intensity*, *Intensity Histogram*, *Intensity Histogram Gaussian Fitting*, *Neighbour Intensity Difference* and *Shape*. Each category is described in section 2.1.2.3

3.2.2.1 On the use of Oversampling technique

Figure 3.13 shows a schematic representation of the proposed radiomics system. The upper and lower panels depict the training and test phase of the approach, respectively. The former first semi-automatically segments the images in the training set. Then it computes the pool of radiomics features (feature computation) and applies a filter-based feature selection method. As the dataset is imbalanced, we first oversample the data and then we train the classification algorithm. The test phase,

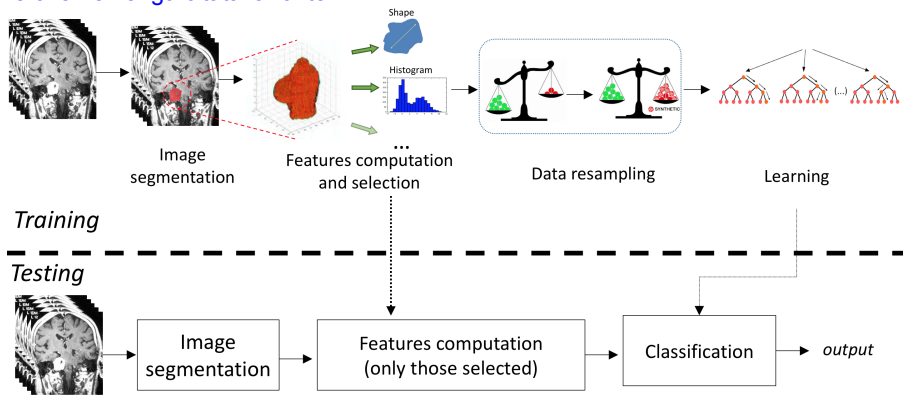


Figure 3.13: Schematic representation of the machine learning approach adopted using Smote as oversampling tecnique.

after image segmentation, gets as input the ROI of the test patient and computes the descriptors selected during the training phase. Next, the trained classifier labels the patient images as patients with volume reduction, patients with volume stability, or patients with volume increase. In the rest of this section, we detail each of such steps. For the sake of presentation, the whole set of computed features is divided into semantic and radiomics descriptors, which are described in section 3.2.2.1.1.

3.2.2.1.1 Semantic and Radiomics Feature Computation

Before radiomics feature computation, for each patient, age, gender and length of follow-up were collected. Following, we gathered various quantitative biomarkers extracted from the MRI after tumour segmentation.

In total, 1135 shape-based, intensity-based and texture-based features were extracted from the area of interest using the open-software platform IBEX [125]. This choice mitigates a common issue with feature extraction, i.e., the lack of standardised and reproducible tools in order to follow and compare results [62, 126, 127].

Extracted features are divided into the following categories: *Gradient Orientation Histogram*, *Gray level Cooccurrence Matrix*, *Gray Level Intensity*, *Intensity Histogram*, *Intensity Histogram Gaussian Fitting*, *Neighbour Intensity Difference* and *Shape*. Each category is described in section 2.1.2.3

3.2.2.1.2 Feature selection

Using the Feature Computation method described in paragraph *Semantic and Radiomics Feature Compuation* of section 3.2.2, we extracted 1135 features. It is well-known that a limited yet salient feature set simplifies both the pattern representation and the classifiers that are built on the selected representation, as it alleviates the curse of dimensionality. Hence, we applied a step of feature selection before data resampling, as also suggested by Zhang et al. [128] where the authors experimentally

found that feature selection before resampling outperforms the opposite schema.

The used feature selection method is based on the following rationale. From a full recovery point of view, patients whose size of tumour increases or is stable in comparison with the first MR scan (see section 3.2.2 paragraph *Dataset*) have poorly responded to the treatment as well. On the contrary, when the lesion size decreases, the patients have responded well to the therapy. Therefore, it is important that a feature can discriminate between well responding from poorly responding patients, i.e. whose tumour size is stable or even has increased after the therapy. Furthermore, it is less important to discriminate between patients with stable tumour size or for whom the lesion has increased.

On these grounds, all features that better represent the data should be very discriminative for reducing volume patients with respect to other classes, whereas they may present similar values for patients with stable or increasing lesion volume.

The method utilised to select the features, which aims at discovering the relationship mentioned above, is based on *Mann-Whitney U test* [129] applied to data using a pairwise approach while testing each feature over the three classes. For the sake of presentation, we labelled the three classes as follows:

- -1 : class of patients whose tumour size has increased;
- 0 : class of patients with the lesion size that has not varied;
- 1 : class of patients whose tumour size has reduced.

Furthermore, we defined F_{ij} as all observations belonging to the *i-th* class measured with the *j-th* feature . To select the features, we performed the aforementioned statistical test three times for each feature, computing the statistical significance of the following comparisons: F_{0j} vs. F_{1j}, F_{-1j} vs. F_{1j} and F_{-1j} vs. F_{0j}. We looked for features able to show that F_{-1j} and F_{0j} were different samples coming from the same population and that, at the same time, F_{-1j} and F_{1j}, as well F_{0j} and F_{1j} belong to separate pools. Therefore, selected features fulfill all the following *p-value* constrains:

- $p < 0.05$, in case of F_{0j} vs. F_{1j}
- $p < 0.05$, in case of F_{-1j} vs. F_{1j}
- $p >= 0.05$, in case of F_{-1j} vs. F_{0j}

We, therefore, looked for features that significantly differ when comparing data coming from classes 1 and 0, or classes 1 and -1. When comparing samples coming from classes 0 and -1 we related these constraints. Graphically, this approach is shown in Figure 3.14 where each feature corresponds to a dot in the 3D space and where the feature category gives the colour of the dots. Selected features are those falling within the red parallelepiped.

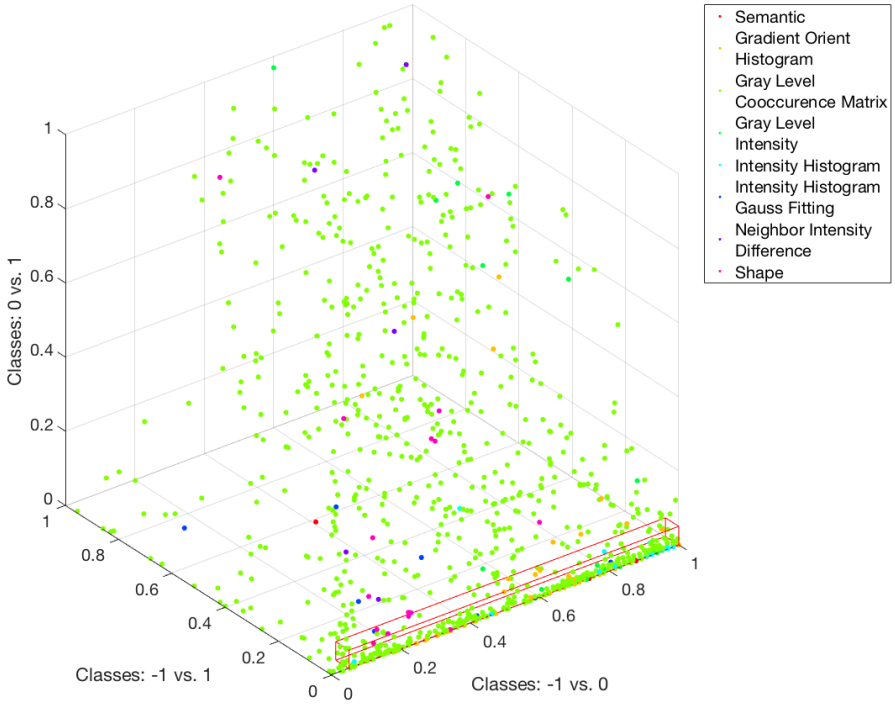


Figure 3.14: Graphical representation of the feature selection approach.
Selected features are those within the red parallelepiped.

3.2.2.1.3 Data resampling and classification

Class imbalance, a.k.a, class skew, refers to the problem that the training set has a disproportion among different classes, as happens in this dataset (section 3.2.2 paragraph *Dataset*). Since traditional learning algorithms are designed to minimize errors over the majority samples, ignoring or paying less attention to instances of the minority classes, this usually results in poor predictive accuracy over the minority ones. For this reason, in the literature, there exist many techniques to combat this issue, including internal approaches that tailors an algorithm to imbalanced data, data selection approaches (also named as resampling approaches), cost-sensitive learning, and ensemble learning [115, 130, 131].

In this work, a priori sample distribution is skewed when pairwise comparing the classes. To cope with this issue, we applied a resampling approach, named as Synthetic minority over-sampling technique (SMOTE) [132], which resizes the training set to make the class distribution more balanced to match the size of the other class(es). Indeed, if we denote as N^+ and N^- the set of samples in the minority and in the majority class, respectively, it generates a set $\overline{N^+}$, with $|\overline{N^+}| \approx |N^-|$, where $\overline{N^+}$ is composed of all samples in N^+ and others generated by the method under consideration.

Table 3.15: Results

Class	*Accuracy*	*Precision*	*Recall*	*AUC*
volume reduction (1)	85.33	1.00	1.00	1.00
no volume variation (0)		0.73	0.88	0.94
volume increase (-1)		0.85	0.68	0.92

After the data resampling step, we proceeded with the learning and classification block using a Random Forest, which has the beneficial property of avoiding overfitting.

3.2.2.1.4 Results

The feature selection results reveal that, starting from the pool of 1135 features, the final signature is composed of 427 measures. Furthermore, of the 7 radiomics feature categories, only the *Neighbor Intensity Difference* was wholly rejected, denoting that a measure based on the grey level difference between adjacent pixel/voxel does not discriminate. On the contrary, texture information is caught by features computed from the *Gray level Cooccurrence Matrix*, and such measures are effective for the task at hand being the most selected ones. Moreover, it is worth noting that the selected features from the semantic category are the *follow-up* measuring the time passed between the MR scans and the patient's *gender*.

Let us focus on the results achieved in the classification stage. First, it is worth noting that the experiments were performed in 10-fold cross-validation. The performances, reported in Table 3.15, are measured not only in terms of accuracy, but we also report precision, recall and ROC AUC for each of the three classes. The results shown were promising: the best performances were achieved in the classification of *volume reduction* class where all the patients were correctly classified. This result was consistent with the feature selection applied: indeed, it focused on discriminating patients who have clinically responded to the therapy with respect to those who have not. Finally, it is worth noting that in case of the class named as *volume increase* (-1) the precision is greater than the recall: this suggests that the system has a more conservative behaviour in predicting results when they deviate from the original volume condition. For the sake of completeness in Table 3.16 we also report the performances achieved without the resampling step given by SMOTE. The results show a considerable drop in accuracy and precision of prediction for all classes, whereas the increased recall of *volume increase* class denotes the tendency of the classifier to prefer this group of patients. This can be expected as such class is the largest one, therefore confirming that the skewness of the dataset biases the results.

For the described approach we used Python- 3.8.3, scikit-learn-0.23.1, pandas-1.0.5, numpy-base-1.18.5, and two NVIDIA GeForce RTX 2080 Ti, each with 11 GB

Table 3.16: Results without resampling

Class	*Accuracy*	*Precision*	*Recall*	*AUC*
volume reduction (1)	63.16	0.67	0.67	0.94
no volume variation (0)		0.20	0.10	0.38
volume increase (-1)		0.70	0.84	0.66

of memory.

3.2.2.1.5 Conclusion

This work proposed a machine-learning based radiomics pipeline that predicts the outcome of patients affected by acoustic neuroma and treated with the CyberKnife therapy. The method works with MR images routinely collected. It extracts the features after tumour volume segmentation. Next, it selects the features using the Mann Whitney U test and generates synthetic samples using the SMOTE technique to prevent unbalancing issues. A Random Forest classifier is adopted to predict if a patient will vary (increase or decrease) the volume of the tumour lesion in the following year.

Although the achieved results are promising, further research is needed for several reasons. First, there is the need to enlarge the number of patients to confirm the robustness of the approach. Second, more data would also permit to investigate the effect of using images belonging to different scans, possibly developing strategies to better homogenize the data. For instance, in our dataset, the two scans have small differences in pixel size (3%). Third, we also plan to enlarge the cohort of tested classifiers.

3.2.2.2 On the use of a cascade of cost-sensitive decision trees

The solution proposed in section 3.2.2.1 presented two main disadvantages; the first was the univariate feature selection that conducts a pair-wise analysis of the dependency between each feature and the target concept. However, its pair-wise nature prevents it from identifying interacting features that, as a group, can be used to predict the target concept, but individually have no detectable dependency relationship with the target concept, being also more prone to noise. The second is that even if SMOTE is a widely used oversampling method, it can introduce a small bias if the number of samples to be increased is too small. In this respect, the second analysis of the dataset extends the previous contribution by introducing a simultaneous feature ranking method, i.e. the Relief-f, and presenting a new classification approach based on a cascade of cost-sensitive decision trees and performing a wider assessment of different methods for imbalance learning.

3.2.2.2.1 Methods

The radiomics method we developed is schematically illustrated in Figure 3.15, where the upper and lower panels depict the training and test phase of the approach, respectively. The former first semi-automatically segments the images in the training set. Next, it computes the pool of radiomics features that, together with some semantic descriptors, are afterwards processed in a feature selection step. The peculiarities of the dataset, as well as our preliminary results [133], suggest us to tackle the learning issue by developing an ad hoc learning approach, which can be framed within cost-sensitive learning methods [130]. Let us recall that cost-sensitive learning is one of the three main approaches that can be used to address class imbalance, a.k.a. class skew.

To overcome the class imbalance, further to cost-sensitive learning, many techniques in the literature can be grouped into data selection approaches (also named as resampling approaches) and ensemble learning [115, 130, 131]. For the sake of presentation, such other methods are briefly discussed in section 3.2.2.2.4, while in the following section 3.2.2.2.3 we will introduce the cost-sensitive approach we developed.

The lower part of Figure 3.15 represents the test phase that, after image segmentation, gets the ROIs of the test patient as input and then computes the features selected during the training phase. Next, the trained model labels the patient images as *patient with volume reduction*, *patient with volume stability* or as *patient with volume increase*.

In the rest of this section, we detail each of the steps depicted in Figure 3.15 and introduced so far.

3.2.2.2.2 Feature selection

Using the Feature Computation method described in paragraph *Semantic and Radiomics Feature Compuation* of section 3.2.2, we extracted 1107 features. It is well-known that a limited yet salient feature set simplifies both the pattern representation and the classifiers that are built on the selected representation, as it alleviates the curse of dimensionality. Furthermore, radiomics image features are sensitive to noise, which could degrade the quality of the features themselves [134]. Hence, to estimate the quality of the available attributes and improve the classification performance, we apply a feature selection step before the learning phase.

Differently than in the previous application where the feature selection was performed finding features that differ using the Mann-Whitney U test, here we used here the Relief-F feature selection method. This methods selects relevant features in supervised multi-class problems, both in noise-free and noisy data [135]. Furthermore, it is a simultaneous feature ranking method that permits to identify interacting

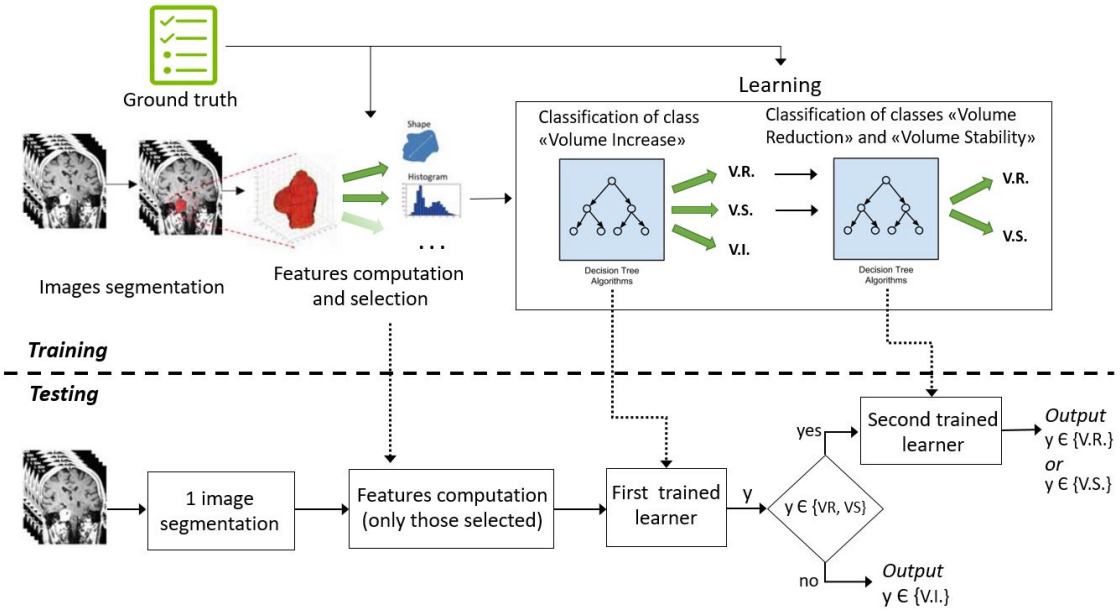


Figure 3.15: Schematic representation of the proposed machine learning approach. During the training phase (top panel) radiomics and semantic features are computed and given to the feature selection phase. Then, a cost-sensitive three-class decision tree is trained using the available samples, classifying them as *patient with volume increase* (V.I.), *patient with volume reduction* (V.R.) or *patient with volume stability* (V.S.). Samples classified in the latter two classes are passed to a further decision tree, which is trained for the binary classification of V.R. and V.S. classes. During the test phase (bottom panel), after feature computation, the trained learners are used in cascade to label the samples.

features that, as a group, can be used to predict the target concept, although individually, they have no detectable dependency relationship with the target concept. Relief-F estimates the quality of attributes according to how well their values distinguish between instances that are near to each other: for each instance R_i it searches the k nearest neighbours from the same class and the k nearest neighbours from each of the other classes. The algorithm assigns a score between -1 and 1 to each of the considered predictors, penalising those that give different values to neighbours of the same class, and rewarding those that give different values to neighbours of different classes, ranking them from the least to the most significant for the classification task. In this work, we consider only those features to which Relief-F assigns a score greater than 0, i.e. the features that make a positive contribution to the discriminating capacity of the system.

3.2.2.2.3 Learning model

In this work, a priori sample distribution is skewed when pairwise comparing the classes. To cope with this issue, we use cost-sensitive learning by directly introducing

and utilising misclassification costs into the learning algorithm, which are reported in Table 3.17.

Table 3.17: Error cost matrix. V.R.: volume reduction, V.S.: volume stability, V.I.: Volume increase

		Predicted Outcome		
		V.R.	**V.S.**	**V.I.**
True Outcome	**V.R.**	0	0.25	0.75
	V.S.	0.25	0	0.75
	V.I.	1.00	1.00	0

Furthermore, the proposed approach consists in a cascade of two classifiers: the rationale stems from the nature of the classification task where two of the three classes, namely V.R. and V.S., correspond to a positive response to the therapy, whereas V.I. corresponds to a therapy without any effect. Therefore, we expect samples belonging to former classes to have a more similar representation in the feature space. This suggested us to introduce a second decision tree in the cascade specialised in distinguishing V.R. and V.S. samples.

From a clinical point of view, the error cost matrix can be explained as follows. The primary aim of the Cyberknife treatment for acoustic neuromas is to achieve maximal tumour control whilst guaranteeing optimal functional outcome and minimal morbidity risk [136]. A good response to treatment is considered based on clinical outcomes and achievement of tumour reduction or stability. Thus, we select the error costs in the light of the following considerations:

- misclassification between classes *patient with volume stability* and *patient with volume reduction*: in case the algorithm classifies as *patient with volume stability* a patient whose tumour volume reduces, or it classifies as *patient with volume reduction* a patient whose tumour volume remains stable, the error cost can be set to a relatively small value, such as 0.25.

- patients whose tumour volume is stable or reduces are classified as *patient with volume increase*: we consider an error cost of 0.75 since these patients will not be treated although they would have a positive result.

- *patients with volume increase* are misclassified as stable or with volume reduction: this is the worst case since an ineffective or potentially harmful treatment would be performed on the patient. As a consequence, we consider an error cost of 1.

After the feature selection phase, relevant predictors to the task are passed to a first learner, a three-class CART decision tree. According to the cost sensitive learning

strategy, we use the aforementioned cost metric during the training phase so that the learner can correctly discriminate the minority class.

We use samples from all classes to train the three-classes decision tree. Then, we also use samples not belonging to the minority class to train a binary CART decision tree where we do not utilise cost sensitive learning since it classifies the two most represented classes.

The testing phase follows the same rationale and it therefore applies a cascade of decision trees to classify unknown samples. When a test sample undergoes the first decision tree two situations can occur: if it is classified as *patient with volume increase* the testing phase ends; otherwise, the sample is consequently passed to the second binary classifier in the cascade, which classifies it as *patient with volume reduction* or as *patient with volume stability*.

3.2.2.2.4 Experimental design

The proposed method was tested on the dataset described in section 3.2.2 paragraph *Dataset* using a leave-one-patient-out scheme. We first selected the features on each training/test set pair and then applied the proposed classification scheme.

Since we are dealing with a skewed classification task, we compared our proposal with several competitors belonging to the two other families of approaches for class imbalance learning, i.e. resampling approaches and ensemble learning. For the sake of presentation, other methods belonging to cost-sensitive learning are not reported since they perform worse than our proposal.

The remainder of this section shortly introduces the tested competitors and the metrics we compute to measure the performances.

Resampling approaches A plethora of resampling approaches exists, which can be divided into under- and oversampling. Our experiments considered the following popular oversampling approaches, whereas the number of samples per class suggested not to consider undersampling methods.

Ensemble learning This approach employs an ensemble of learners, where each composing classifier C_i was trained on a subset of the majority class and a subset of the minority accounting; however, for a large portion of the minority class samples [137, 138, 139].

Then, the decisions taken by all C_i on the test sample were combined to set the final output according to a given rule, such as the majority voting. The rationale lied in observing that an ensemble of classifiers generally produced better results than those obtained by individual composing experts [140, 141]. Furthermore, base classifiers C_i were now trained on sub-problems more balanced than the original one,

which also had the desired properties of containing samples representing different aspects of the original set N.

On these general premises, in this work, we considered the following popular approaches:

- *Balanced Bagging classifier*: Bagging methods build several learners on different randomly selected subsets of data. Therefore this approach balances each subset of data by undersampling the majority class so that the number of selected samples matches the number of samples extracted from the minority class.

- *Forest of randomized trees*: this is a variation of the original random forest method that builds an ensemble of trees induced from balanced down-sampled data. First, each iteration in random forest draws a bootstrap sample from the minority class and randomly draws the same number of cases, with replacement, from the majority class. Second, it induces a CART classification tree from the data to maximum size, without pruning. At each node, instead of searching through all variables for the optimal split, it only searches through a set of randomly selected variables. Third, the previous two steps are repeated and, after training, the final decision is given by majority voting each tree decision [142].

- *XGBoost*: it is a scalable, portable and distributed implementation of Gradient Boosting [143], which has recently gained much popularity as the algorithm used by several teams to win machine learning competition. Its derivation follows the same idea in gradient boosting, where the authors improved the regularised objective.

3.2.2.2.5 Base classifiers

The methods introduced in previous paragraphs were tested considering three base classifiers belonging to different learning paradigms. In detail, we employed a k-Nearest Neighbour (3NN for the oversampling approach and 1NN for the ensemble learning) as a statistical classifier, a Support Vector Machine (SVM) as a kernel machine, and a CART algorithm as a decision tree. Furthermore, the forest of randomized trees and the XGBoost were tested using CART only. We set the classifier parameters to the default values used in the Scikit-learn library [144] for CART, while a Gaussian Radial Basis Function (RBF) and a linear kernel are adopted together with one vs one decomposition for the SVM. This also holds for the decision trees used in our proposal described in section 3.2.2.2.3. Although we acknowledge that their tuning could lead to better results, the No Free Lunch Theorems for Optimization tell us that all configurations perform equally well when averaged over

all possible experiments [145]. As basis for comparison, we therefore preferred to maintain the baseline parameter set, so that the only way a method can outperform another one relies upon its specialization to the structure of the problem under consideration [146]. Furthermore, in a framework where the classifier is not tuned, the winning method tends to correspond to the most robust, which is also a desirable characteristic [147].

3.2.2.2.6 Model assessment

Traditionally, the most frequently used performance metrics are the accuracy (acc) and its counterpart, the error rate. The former is defined as

$$acc = \frac{\sum_{j=1}^{K} n_{jj}}{N} \tag{3.1}$$

where n_{jj} is the number of elements of class j correctly labelled, N is the total number of samples, and K is the number of classes. However, the accuracy fails to reflect the extent of minority class misclassifications since it is based on values from all confusion matrix rows, whose values depend on class distribution.

Hence, it would be more interesting to use a performance measure dissociating the hits (or the errors) that occur in each class. From the confusion matrix we compute the accuracy per class, known also as recall, which is defined as $acc_j = n_{jj}/N_j$, with $j = 1, \ldots, K$, where N_j denotes the number of samples belonging to the class j. Since each value of acc_j is estimated considering only one row of the confusion matrix, it is independent of prior probabilities. From this quantity we derive the geometric mean of accuracies (g) given by

$$g = \left(\prod_{j=1}^{K} acc_j \right)^{\frac{1}{K}} \tag{3.2}$$

It is a non-linear measure and a change in one of its arguments has a different effect depending on its magnitude; for instance, when a classifier misclassifies all samples of the jth class, acc_j and also g are zero.

3.2.2.2.7 Results

The results of the Relief-F feature selection revealed that starting from the pool of 1109 features, the final signature of each iteration of the leave-one-patient-out was averagely composed of 300 ± 8 measures. This suggested that there was a bulk of features that was selected in all the iterations, i.e. it was robust to variations in the training set. Furthermore, we noted that only semantic and texture features (HOG and GLCM) were selected, whilst the others based on grey level difference among neighbour pixels/voxels were completely rejected. In addition, it is worth noting that both semantic features, i.e. the patient's gender and age, are always selected.

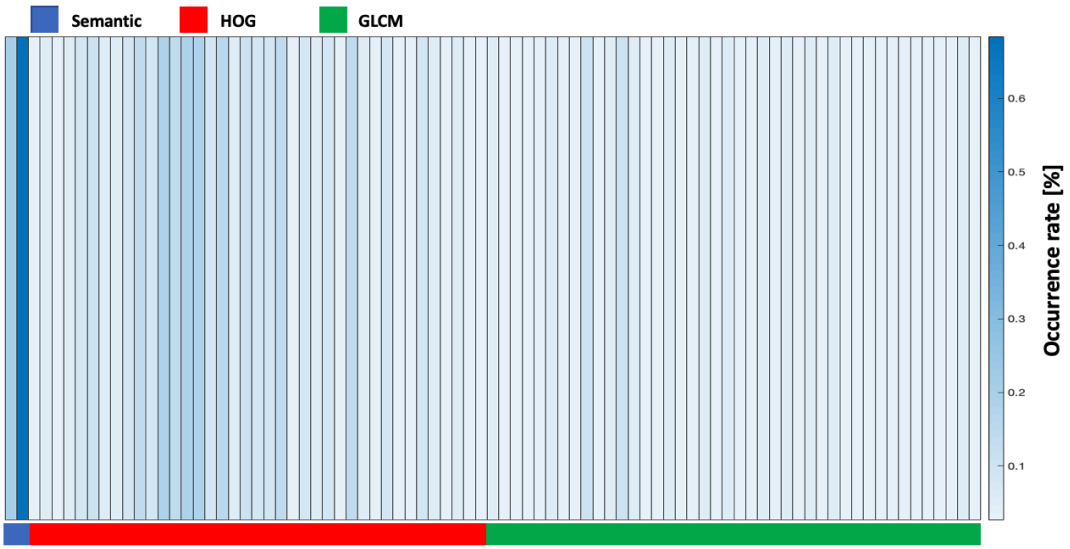


Figure 3.16: Heat-map of the features selected by the decision tree, represented according to the rate of occurrence in the different iterations.
Dark colors represent higher occurrence rates, while light colors represent seldom selected features. The illustrated features belong to the pool of features selected by the Relief-F, that are *Semantic*, *Histogram of Oriented Gradient* (HOG) and *Gray Level Co-occurrence Matrix* (GLCM) features.

After Relief-F, during the training phase the decision trees evolve considering only the most representative features among those already selected by the medium of Relief-F. As an example, Figure 3.16 illustrates the rate of occurrence of the features selected by the three-class decision tree in all the iterations of the leave-one-patient-out scheme. Notably, both semantic features were selected in most iterations, while HOG features had a higher rate of appearance than those from GLCM.

Table 3.18 reported the average results of the performed tests. The performances of the proposed method were illustrated in the first row, while other results referred to methods listed in section 3.2.2.2.4 that we tested on our dataset. In particular, we tested oversampling and ensemble learning approaches, the other two alternatives to cost-sensitive learning available in the literature for skewed tasks. The proposed method achieves interesting results: the accuracy is equal to 0.92 and the g-mean value is 0.85. Concerning precision and recall, the performances achieved for the two majority classes are greater than or equal to 0.91, whereas a value of 0.67 was achieved for the minority class for both the metrics. Compared to the tested competitors, these results are noteworthy: most of them, both from oversampling (SMOTE, Borderline SMOTE1 and Borderline SMOTE2) and from ensemble learning approaches (balanced bagging classifier) have precision and recall values for the minority classes equal to zero. This implies that they classify every sample as belonging to the majority class. We also observe that the other two methods from

ensemble learning (forest of randomized trees and XGBoost) correctly classify the same rate of samples in the most underrepresented class (V.I.), but this happens at the detriment of the other two majority classes. This result can be expected since it is well known that it is a typical side effect that can happen when a classifier tries to improve the recall on the minority class. Furthermore, Table 3.18 show that the SVM classifier in combination with ADASYN oversampling results to be the best competitor; nevertheless, it is considerably outperformed by the proposed model.

Table 3.18: Results of the experiment using a cascade of cost-sentitive decision trees.
As in Figure 3.15, V.R., V.S. and V.I. stand for patient with volume reduction, with volume stability and with volume increase, respectively. In round parenthesis we report also the learning paradigms achieving the best performances, that are the values reported in the columns of the table.

Learning approaches		**Accuracy**	**G-mean**	**Recall**			**Precision**		
				V.R.	**V.S.**	**V.I.**	**V.R.**	**V.S.**	**V.I.**
Cascade of cost-sensitive decision trees (proposed method)		0.92	0.85	0.92	1.00	0.67	0.96	0.91	0.67
Oversampling	SMOTE (RBF SVM)	0.66	0	1.00	0	0	0.66	0	0
	Borderline SMOTE1 (RBF SVM)	0.66	0	1.00	0	0	0.66	0	0
	Borderline SMOTE2 (RBF SVM)	0.66	0	1.00	0	0	0.66	0	0
	ADASYN (LINEAR SVM)	0.50	0.41	0.40	0.50	0.33	0.67	0.29	0.17
Ensemble learning	Balanced Bagging Classifier (RBF SVM)	0.66	0	1.00	0	0	0.66	0	0
	Forest of randomized trees	0.34	0.36	0.36	0.20	0.67	0.69	0.13	0.20
	XGBoost	0.47	0.34	0.60	0.10	0.67	0.60	0.10	0.67

Finally, let us compare the results achieved here with those reported in our previous work [133], which achieved a 0.85 accuracy after SMOTE oversampling. Again, our proposal outperforms such results mostly thanks to the cascade of cost-sensitive trees we designed. We also investigated how our proposal as well as the competitors perform using a feature set determined by the feature selection algorithm presented in our previous work [133], which is based on Mann-Whitney U test. Although not reported here for the sake of brevity, the results show that the feature set determined by the Relief-F used in this manuscript permits to achieve better recognition results. This could be expected since we introduce the Relief-F to overcome the limitations of the feature selector previously used, which runs only univariate and pair-wise analysis and is therefore unable to detect interacting features that as a group can be used to classify the samples.

3.2.2.2.8 Conclusion

This work proposed a machine-learning-based radiomics pipeline that predicts the outcome of patients affected by acoustic neuroma and treated with the CyberKnife therapy. The method worked with MR images routinely collected and extracted the features after tumour volume segmentation. Moreover, it selected the features

using a Relief-F approach. Next, we presented a new classification approach that leverages on a cost-sensitive learning technique. It uses a cascade of two classifiers that, inspired by some biomedical observations on the disease, is composed of a three-class CART decision tree followed by a binary CART tree. The performances of the proposed method were compared to two imbalance learning families of approaches, the resampling approach and the ensemble learning, using different classification algorithms. The outperforming results of the proposed method indeed show the potential of using a cost-sensitive approach to handle the imbalance nature of these radiomics data. Moreover, this work proves again the possibility of identifying a valid radiomics signature for response to treatment prediction in acoustic neuroma.

Despite all these promising findings, there are still some limitations, mainly due to the small sample size used. In fact, future work could be directed towards enlarging the number of patients to confirm the approach's robustness and to test deep learning-based methods. Besides this, more data would permit to investigate the effect of using images belonging to different scans, possibly developing strategies to better homogenise the data.

To summarise, this study has demonstrated that ML associated with radiomics has a great potential in distinguishing patients belonging to the different classes undergoing radiosurgery, prior to the treatment. Knowing in advance the chance of treatment success as well as side effect could greatly complement the clinical knowledge so far used to optimise radiosurgical treatments and eventually change the therapeutic modality to a different one or to a multimodality approach. Lastly, the proposed approach can also be applied to similar pathologies treated by CyberKnife (e.g. brain metastasis [148], lung cancer), generalising the value and applicability of the method.

3.3 Hybrid Approach of ML and DL algorithms

As reported by Ibrahim et al. [149] and in section 2.3.2, usually DL applications for clinical purposes are mainly focused on the VOI/ROI segmentation, the diagnosis or the prognosis of clinical outcome. When combined with ML methods, DL is used for the segmentation step, avoiding the necessity to involve physics for the manual or semi-automatic segmentation. One example of this kind of application was developed by Zhang et al. [150], who uses DL algorithms to replace the manual segmentation of the standard radiomics workflow. In Figure 3.17 the radiomics workflow is represented showing the steps usually done by DL methods when combined with ML algorithms as presented by Zhang et al., compared to our proposal, that used DL methods for the segmentation and feature extraction, while using ML for the feature selection and analysis.

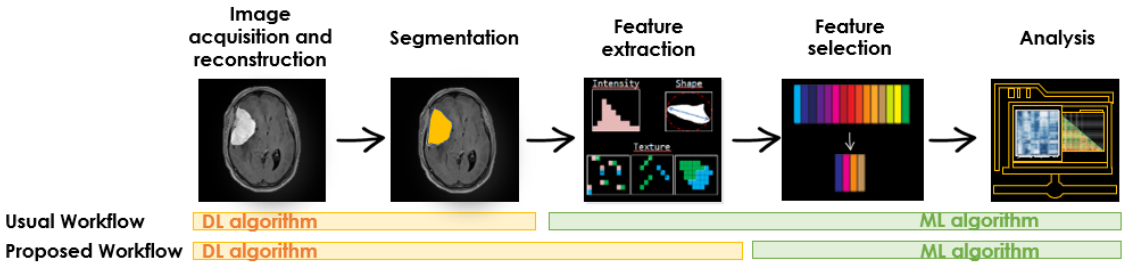


Figure 3.17: Workflow of Hybrid approach integrating DL and ML methods comparing the usual involvement of DL and ML methods with the proposed method.

The pipeline worked as follows: first, we applied a pre-trained deep neural network to segment the lungs; second, a convolutional neural network was trained to extract relevant features from the CXR images; third, we concatenated the deep features with the clinical ones; fourth, we performed a feature selection step as reported in section 3.1.1.2.3; fifth, we trained a supervised classifier to accomplish the binary classification task. In the following, we will illustrate these steps.

As mentioned before, in section 3.1.1.1, the image repository was composed of CXR images collected in multiple hospitals, using different machines with different acquisition parameters. This resulted in a certain degree of variability among the images, where the lungs have also different sizes. To cope with this issue, we adopted the segmentation discussed in section 3.4.1, which boosts the performance of the feature extraction network by locating the lungs.

Each cropped image was then passed to a deep neural network to extract the features, where we performed a transfer learning process as follows. Indeed, preliminary experiments showed that such an approach gave better results than starting the training from scratch.

Furthermore, to alleviate the risk of overfitting and reduced generalisation typical of learning models working with medical images, we pre-trained several state-of-the-art network architectures previously initialised on other repositories. To this end, we used the chest X-ray images dataset presented by Kermany et al. [151]: Kermany et al. [152], which consists of 5,863 CXR images classified as pneumonia or normal by two expert physicians. This would allow the networks to learn modality-specific feature representations [103, 153]. After this step, such models were fine-tuned on our image dataset. In a first stage we tested in 10-fold cross validation these networks: Alexnet [154], VGG-11, VGG-11 BN, VGG-13, VGG-13 BN, VGG-16, VGG-16 BN, VGG-19, VGG-19 BN [155], ResNet-18, ResNet-34, ResNet-50, ResNet-101, ResNet-152 [112], ResNext [156], Wide ResNet-50 v2 [157], SqueezeNet-1.0, SqueezeNet-1.1 [158], DenseNet-121, DenseNet-169, DenseNet-161, DenseNet-201 [159], GoogleNet [160], ShuffleNet v2 [161]and MobileNet v2 [162]. Then, to reduce the computational burden, the top-five networks (i.e. VGG-11, VGG-19, ResNet-18, Wide ResNet-50 v2, and GoogleNet) underwent all the experiments described in section 3.1.1.2.4. In all the cases, we changed the output layer of the CNNs, using two neurons, one for each class. Moreover, image standardization as described in section 3.1.1.2.1 was performed. We also augmented the training data by independently applying the following transformations with a probability equal to 30%: vertical and horizontal shift (-7, +7), y-axis flip, rotation (-175°, +175°) and elastic deformation (σ =7, α = [20,40]). Training parameters were: a batch size of 32 with a cross-entropy loss, a SGD optimiser with learning rate of 0.001 and momentum of 0.9, with max epochs sets equal to 300 and an early stopping criterion fixed at 25 epochs, using the accuracy on the validation set. In this respect, it is worth noting that we also performed a preliminary optimisation of CNN hyperparameters using Bayesian Optimisation [163], and we found that the results did not statistically differ from those achieved using the aforementioned values, according to the Wilcoxon's test (with $p = 0.05$). Furthermore, this finding agrees also with what reported byArcuri and Fraser [164], already summarised at the end of section 3.1.1.2.5.

Once the deep networks were trained, we integrated the automatic features they computed with the clinical information. To this goal, we extracted the last fully connected layer for each network, which was used as a vector of features for each patient; accordingly, on the basis of the network we were using, the number of automatic features varied between 512 and 4096 (i.e. it is 512 for ResNet-18, 1024 for GoogleNet, 2048 for Wide-ResNet-50 v2, and 4096 for VGG-11 and VGG-19) Each of such sets of automatically computed descriptors was combined with the clinical data and, to avoid to overwhelm the latter, the number of features in the former was reduced by a coarse selection stage using the univariate approach already described in section 3.1.1.2.3.

Furthermore, we then applied the same wrapper approach to investigate if the combination of automatic and clinical features had a degree of redundancy. Straightforwardly, to avoid any bias, all the operations described so far were performed respecting the training, validation and test split introduced before, and ensuring that the test was not used in any stage except for the final validation.

Finally, the selected features were used to classify each patient in the two already mentioned classes, i.e. mild and severe, as reported in the last part of the previous subsection, using the same learners already mentioned.

3.3.1 Results

This section reports the results obtained using the hybrid approach described above in staging patients with COVID-19 in severe and mild outcome classes.

Table 3.19 and Table 3.20 presents the performances using the hybrid approach when the experiments were done using a 10-fold cross validation and a LOCO cross validation, repeated 20 times respectively (section 3.1.1.2.4) where in the tables the mean value is represented. Furthermore, to see if there exists a statistically significant difference between the various performance, we ran the Kruskal-Wallis and the Dunn test with Bonferroni correction for multiple comparisons ($p < 0.05$): the results are reported in Figure 3.18. The results showed that each of the best learner reported in Table 3.19 and 3.20 has performances that are statistically different by large part of the other learners.

With reference to the results attained by the hybrid approach on the CXR images only, we found that the best results in terms of accuracy are statistically lower according to the Kruskal-Wallis test than those attained by the clinical descriptors for 10-fold cross-validation ($p < 0.001$) but no differences were found with LOCO cross-validation ($p = 0.24$). Clinical results were reported in Table 3.7. Among the three learners used with the hybrid approach, the best results with 10-fold cross validation are obtained with RF ($p < 0.001$, Kruskal-Wallis and Dunn's test) while no differences were found with LOCO validation. Furthermore, comparing with a full DL approach, the hybrid provides lower performance ($p < 0.05$ and $p = 0.24$ for 10-fold and LOCO cross-validation, respectively), suggesting that a fully connected layer better exploits the automatic features computed by the convolutional layers of the CNNs. As in Figure 3.4, we show in Figure 3.19 the feature importance of the 40 most selected descriptors by the RFECV wrapper during the experiments in 10-fold and LOCO cross-validation using the GoogleNet. The plot shows that the features most frequently detected as discriminative are clinical measures with some neurons of the dense layer that, although few in number, permits to improve the classification accuracy.

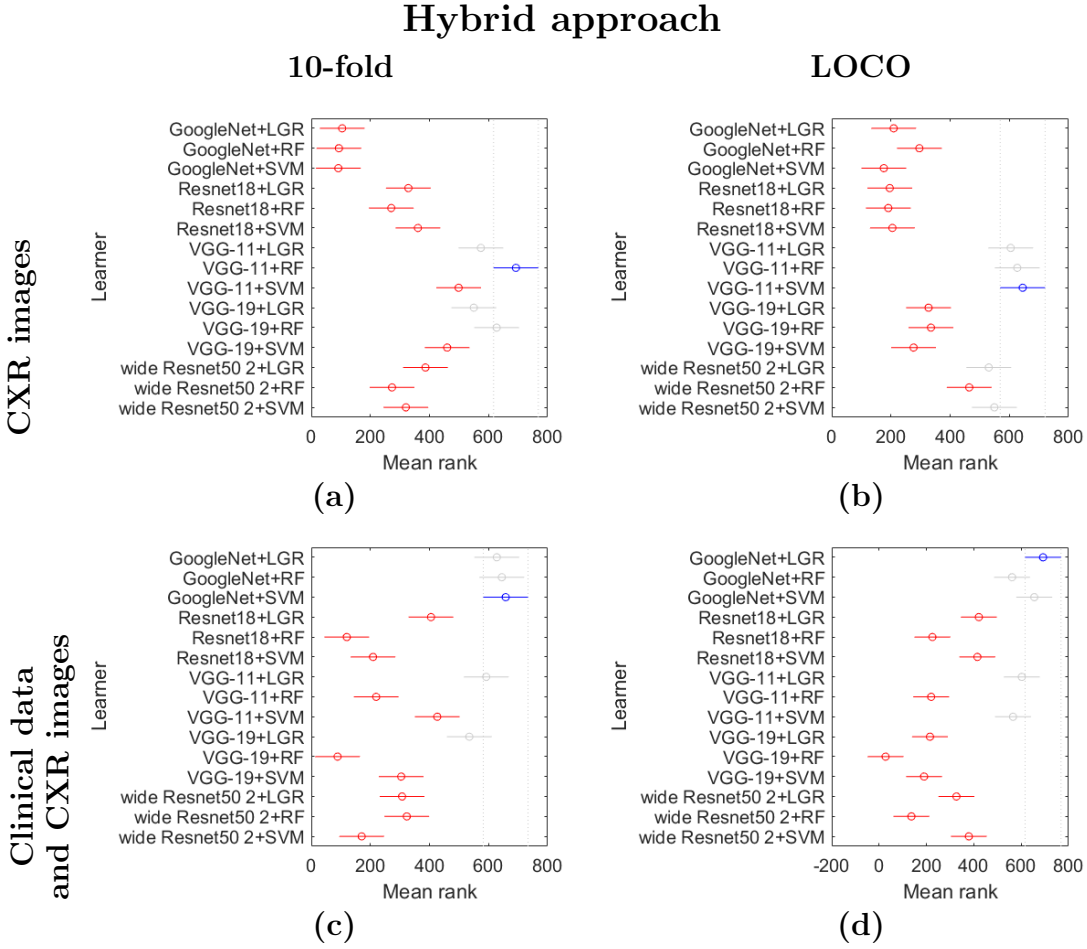


Figure 3.18: Mean Rank results of used learners for the hybrid approach.
In blue the learner which gave the best result reported in Table 3.19 and Table 3.20 The learners in red are those having a mean rank significantly different from the best one.

Table 3.19: Recognition performance of the hybrid approach achieved by all the learners considered, and specified in the last column of the table, when the experiments were executed according to the 10-fold cross-validation (20 repetitions).

Input Data	**Accuracy**	**Sensitivity**	**Specificity**	**Learner**
CXR images	$.672 \pm .048$	$.718 \pm .079$	$.630 \pm .057$	GoogleNet + LGR
	$.672 \pm .052$	$.733 \pm .089$	$.605 \pm .063$	GoogleNet + SVM
	$.677 \pm .038$	$.786 \pm .040$	$.556 \pm .055$	GoogleNet + RF
	$.711 \pm .037$	$.742 \pm .066$	$.671 \pm .073$	Resnet18 + LGR
	$.705 \pm .050$	$.738 \pm .072$	$.668 \pm .078$	Resnet18 + SVM
	$.703 \pm .049$	$.794 \pm .075$	$.600 \pm .088$	Resnet18 + RF
	$.719 \pm .042$	$.730 \pm .090$	$.701 \pm .078$	VGG-11 + LGR
	$.720 \pm .058$	$.729 \pm .104$	$.718 \pm .109$	VGG-11 + SVM
	$.728 \pm .038$	$.769 \pm .072$	$.68 \pm .076$	VGG-11 + RF
	$.721 \pm .047$	$.756 \pm .078$	$.666 \pm .062$	VGG-19 + LGR
	$.713 \pm .043$	$.741 \pm .094$	$.689 \pm .077$	VGG-19 + SVM
	$.719 \pm .049$	$.771 \pm .063$	$.672 \pm .075$	VGG-19 + RF
	$.711 \pm .051$	$.732 \pm .086$	$.686 \pm .068$	Wide Resnet50 2 + LGR
	$.701 \pm .050$	$.726 \pm .077$	$.701 \pm .091$	Wide Resnet50 2 + SVM
	$.708 \pm .046$	$.738 \pm .071$	$.668 \pm .087$	Wide Resnet50 2 + RF
Clinical data and CXR images	$.754 \pm .044$	$.768 \pm .063$	$.755 \pm .039$	GoogleNet + LGR
	$.769 \pm .054$	$.788 \pm .064$	$.747 \pm .059$	GoogleNet + SVM
	$.771 \pm .061$	$.784 \pm .061$	$.753 \pm .107$	GoogleNet + RF
	$.746 \pm .048$	$.753 \pm .043$	$.726 \pm .064$	Resnet18 + LGR
	$.729 \pm .054$	$.72 \pm .069$	$.754 \pm .106$	Resnet18 + SVM
	$.723 \pm .025$	$.774 \pm .063$	$.679 \pm .052$	Resnet18 + RF
	$.761 \pm .064$	$.749 \pm .091$	$.771 \pm .08$	VGG-11 + LGR
	$.742 \pm .04$	$.727 \pm .048$	$.764 \pm .073$	VGG-11 + SVM
	$.755 \pm .021$	$.777 \pm .047$	$.725 \pm .074$	VGG-11 + RF
	$.758 \pm .047$	$.78 \pm .083$	$.74 \pm .065$	VGG-19 + LGR
	$.737 \pm .014$	$.793 \pm .078$	$.691 \pm .05$	VGG-19 + SVM
	$.72 \pm .039$	$.765 \pm .091$	$.686 \pm .048$	VGG-19 + RF
	$.735 \pm .036$	$.752 \pm .048$	$.723 \pm .07$	Wide Resnet50 2 + LGR
	$.737 \pm .027$	$.758 \pm .07$	$.729 \pm .085$	Wide Resnet50 2 + SVM
	$.736 \pm .053$	$.737 \pm .094$	$.735 \pm .058$	Wide Resnet50 2 + RF

Table 3.20: Recognition performance of the hybrid approach achieved by all the learners considered, and specified in the last column of the table, when the experiments were executed according to the LOCO cross-validation.

Input Data	Accuracy	Sensitivity	Specificity	Learner
CXR images	.657 ± .043	.706 ± .117	.562 ± .105	GoogleNet + LGR
	.653 ± .037	.72 ± .096	.572 ± .123	GoogleNet + SVM
	.662 ± .04	.762 ± .099	.533 ± .126	GoogleNet + RF
	.646 ± .053	.744 ± .183	.571 ± .231	Resnet18 + LGR
	.643 ± .05	.73 ± .176	.584 ± .221	Resnet18 + SVM
	.66 ± .042	.779 ± .143	.527 ± .222	Resnet18 + RF
	.692 ± .049	.747 ± .16	.639 ± .187	VGG-11 + LGR
	.694 ± .053	.806 ± .161	.549 ± .213	VGG-11 + SVM
	.682 ± .056	.759 ± .157	.619 ± .21	VGG-11 + RF
	.658 ± .082	.711 ± .158	.625 ± .246	VGG-19 + LGR
	.645 ± .071	.757 ± .141	.605 ± .252	VGG-19 + SVM
	.654 ± .059	.739 ± .146	.625 ± .222	VGG-19 + RF
	.68 ± .031	.705 ± .129	.665 ± .138	wide Resnet50 2 + LGR
	.676 ± .032	.731 ± .107	.602 ± .139	wide Resnet50 2 + SVM
	.668 ± .066	.795 ± .131	.552 ± .256	wide Resnet50 2 + RF
Clinical data and CXR images	.736 ± .061	.769 ± .189	.685 ± .155	GoogleNet + LGR
	.731 ± .036	.726 ± .152	.76 ± .135	GoogleNet + SVM
	.721 ± .031	.775 ± .138	.676 ± .16	GoogleNet + RF
	.713 ± .05	.74 ± .162	.682 ± .19	Resnet18 + LGR
	.705 ± .042	.777 ± .15	.645 ± .214	Resnet18 + SVM
	.679 ± .022	.742 ± .175	.647 ± .211	Resnet18 + RF
	.731 ± .04	.698 ± .186	.733 ± .158	VGG-11 + LGR
	.733 ± .033	.719 ± .163	.733 ± .149	VGG-11 + SVM
	.693 ± .079	.753 ± .137	.641 ± .218	VGG-11 + RF
	.689 ± .035	.732 ± .144	.67 ± .206	VGG-19 + LGR
	.682 ± .027	.718 ± .155	.647 ± .136	VGG-19 + SVM
	.648 ± .054	.775 ± .186	.522 ± .225	VGG-19 + RF
	.691 ± .041	.716 ± .148	.706 ± .149	wide Resnet50 2 + LGR
	.707 ± .031	.762 ± .141	.658 ± .102	wide Resnet50 2 + SVM
	.675 ± .044	.753 ± .154	.648 ± .127	wide Resnet50 2 + RF

To deepen the results, Figure 3.20 shows how much the selected neurons contribute to the network predictions. To this goal, we first depict the regions of input that are important for outputs provided by the CNN (panels a and c in the figure) by applying the Gradient-weighted Class Activation Mapping (Grad-CAM) approach [165]. In a nutshell, Grad-CAM uses the class-specific gradient information flowing into the final convolutional layer to produce a coarse localisation map of the relevant regions in the image. Next, we ran the same algorithm using only the 40 neurons in the dense layer that were mostly selected by the RFECV wrapper (panels b and d in the same figure). The visually inspection of the figure shows

that the regions activated by the 40 neurons cover most of the areas activated by the whole dense layer, confirming that the wrapper correctly identifies the neurons carrying most of the information. Finally, as in Figure 3.5, Figure 3.21 shows that using all the features automatically learnt does not help the learner improving the accuracy, whilst a limited and small number of descriptors is beneficial.

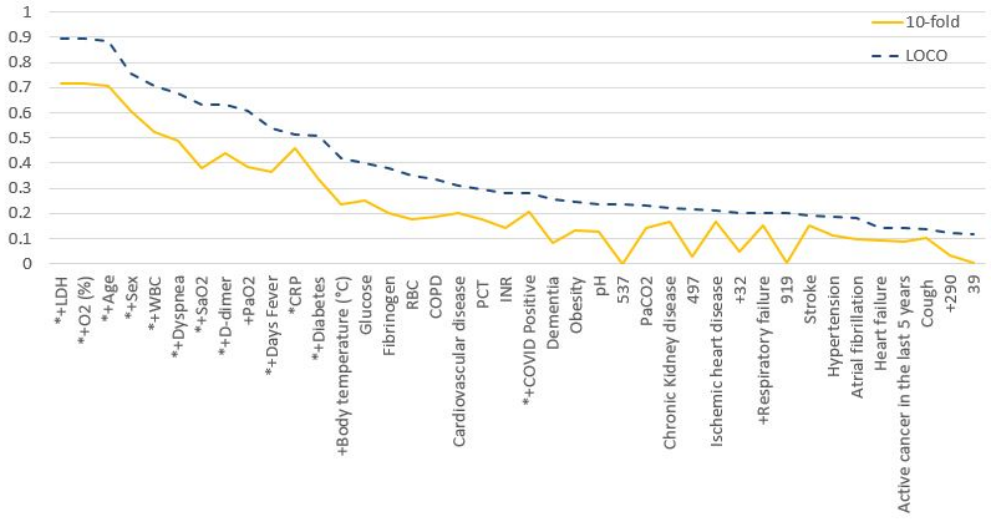
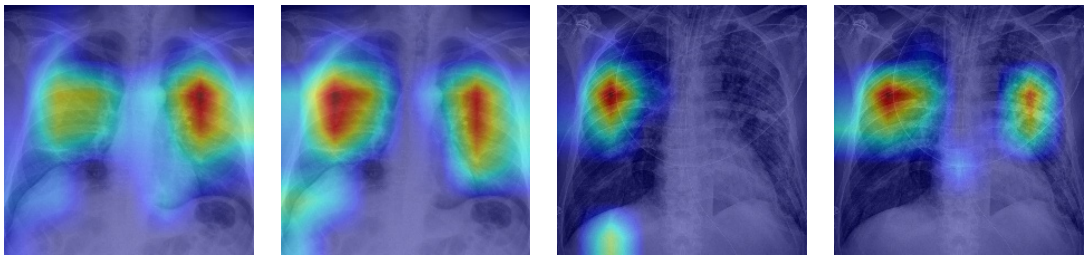


Figure 3.19: Importance of clinical automatically learnt features
It is measured as the rate each descriptor was selected by the RFECV wrapper during the 10-fold and LOCO cross-validation experiments considering all the the three classifiers employed. The y axis scale is normalised to one. Moreover, we add a "*" or a "+" before each feature name if it is included in the feature set used to get the best hybrid results reported in the last section of Table 3.19 and 3.20, respectively.



(a) Mild class, all neurons

(b) Mild class, 40 most selected neurons

(c) Severe class, all neurons

(d) Severe class, 40 most selected neurons

Figure 3.20: Two examples of the activation maps provided by the Grad-CAM approach, using all the neurons in the dense layer of the CNN dense layer or all the 40 neurons selected by the RFECV wrapper.

We now delve into the effect of pre-training the CNNs on another CXR dataset that would help the models learning modality-specific feature representations, as already mentioned in section 3.3. In the case of the hybrid approach, we did not find

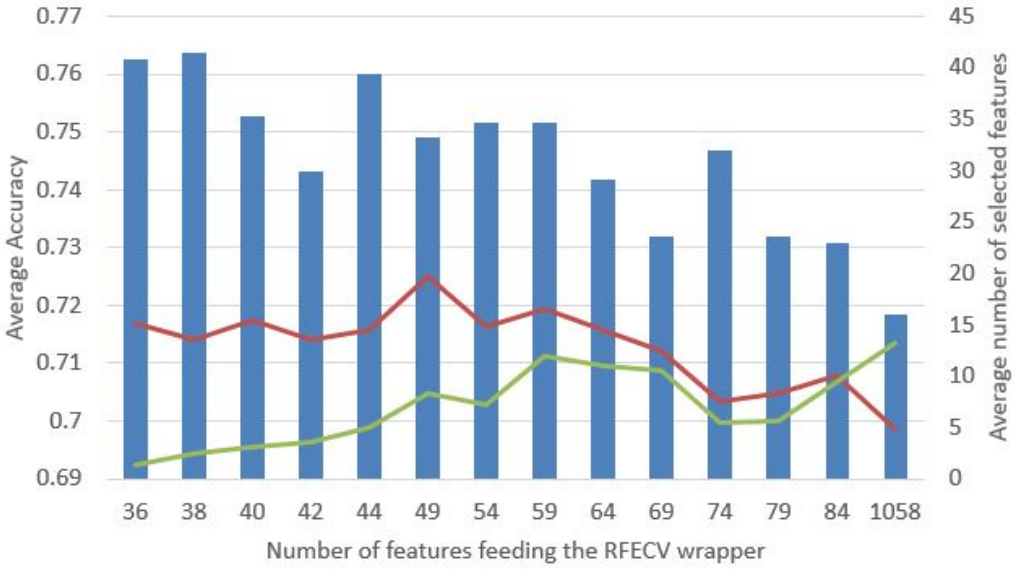


Figure 3.21: Variation of the average classification accuracy (blue bars) with the number of features feeding the RFECV wrapper.
The red and green curves show the number of clinical and texture features selected by the RFECV wrapper, respectively. The experiments plotted here refer to the best results shown in Table 3.19 integrating clinical and imaging features for the hybrid approach

any significant difference in the performance achieved with and without this step: for instance, when using pre-training and the same configuration providing the best result using CXR images in 10-fold cross-validation (ninth line in Table 3.19) we get an accuracy equal to 0.712 ± 0.047. Similarly, in the case of using clinical data and CXR images (second line in Table 3.19) we get an accuracy equal to 0.768 ± 0.036 when we pre-trained the CNN.

For the hybrid approach we used Python 3.7, PyTorch-1.8.1, scikit-learn-0.23.1, and an NVIDIA TESLA V100 with 16 GB of memory for the Deep Learning part, while we used Python- 3.8.3, scikit-learn-0.23.1, pandas-1.0.5, numpy-base-1.18.5, and two NVIDIA GeForce RTX 2080 Ti, each with 11 GB of memory for the Machine Learning part.

To deepen the use of semantic data as model input we also take into consideration the CXR radiological severity score proposed by [1] and further investigated by [2]. To this end, an expert radiologist with more than 10 years of experience assigned such lung damage burden score to a cohort of 240 images randomly selected from the dataset. Then, this score was added to the clinical feature set as an additional image-derived feature. An SVM with the REFCV feature selection, which is the best performing architecture on clinical data as shown in Table 3.7, was used to classify the samples in 10-fold cross-validation with 20 repetitions according to the following three different experiments. First, to have a performance baseline, we ran

the experiment on this subset of images again, using the clinical features only (first row of Table 3.21); second, we tested what happens using such a score only (second row of Table 3.21); third, we ran another experiment using a feature set given by the clinical descriptors plus the radiological score (last row of the same table). It is worth noting that in this last experiment the severity score is included in the set of selected features in 184 out of the 200 runs. The results show that the CXR severity score provides lower performance than the use of clinical descriptors only, regardless if it is used alone or in conjunction with such descriptors. Furthermore, the accuracies are not statistically different according to the Kruskal-Wallis test ($p = 0.545$). A Dunn's test with Bonferroni correction confirmed the result. This suggests that the use of a human-based score assessing lung damage burden is not beneficial.

Table 3.21: Recognition performance attained on a cohort of 240 images from the whole dataset when the human-based CXR radiological score proposed by [1] and [2] is added to the clinical data.

Input data	**Accuracy**	**Sensitivity**	**Specificity**
Clinical data	$.728 \pm .018$	$.701 \pm .039$	$.758 \pm .032$
Only radiological score	$.718 \pm .021$	$.682 \pm .032$	$.750 \pm .023$
Clinical data + radiological score	$.719 \pm .021$	$.720 \pm .030$	$.724 \pm .018$

Let us now discuss how performances vary when Anterior Posterior (AP) and Posterior-Anterior (PA) projections are used. To this end, we measure the performance for each of the best learners reported in Table 3.19 and 3.20 distinguishing between the accuracies achieved on AP and PA images belonging to the different cross-validation instances of the test sets or to the different centres involved. The results, show that in most of the cases the accuracies obtained on the AP images are larger than those achieved on the PA images; nevertheless, the AP scores are not much larger than the average results shown in Table 3.19 and 3.20. Although AP images were mainly used for acquisitions of bedridden patients using portable machines producing a poorer quality image when compared with a PA chest radiograph performed in a dedicated radiography facility [166], we deem that the larger accuracies they provide are due to their larger prior probability than the PA projections in the available repository (Table 3.1). This should support the proposed approach because the use of AI has revealed the possibility to predict the prognosis of the patients even in spite of the limitations of AP CXR scans, e.g. their more difficult interpretation and the sub-optimal imaging resulting from patient's positioning that may reduced inspiratory effort [166].

Finally, let us now focus on the population characteristics, where we found interesting reports on the age and gender distribution. Women were both less and older, suggesting that they become less ill and suffer from more serious conditions

at an older age than men; also the women mortality was lower, as 72% were male confirming the male mortality reported in China (73%) by [167]. The male-related susceptibility and the higher male-mortality rate was also reported by [168], who analyses the data of 59,254 patients from 11 different countries. The second main finding was that 87% of patients had at least one comorbidity (Figure 3.22), suggesting that, in most cases, the conditions leading to hospitalisation occur in patients with coexisting disorders. The most common disease (in 45 % of cases) was hypertension, confirming the results reported by [169], who meta-analysed the data of 1,576 infected patients from seven studies and reported a hypertension prevalence of 21%.

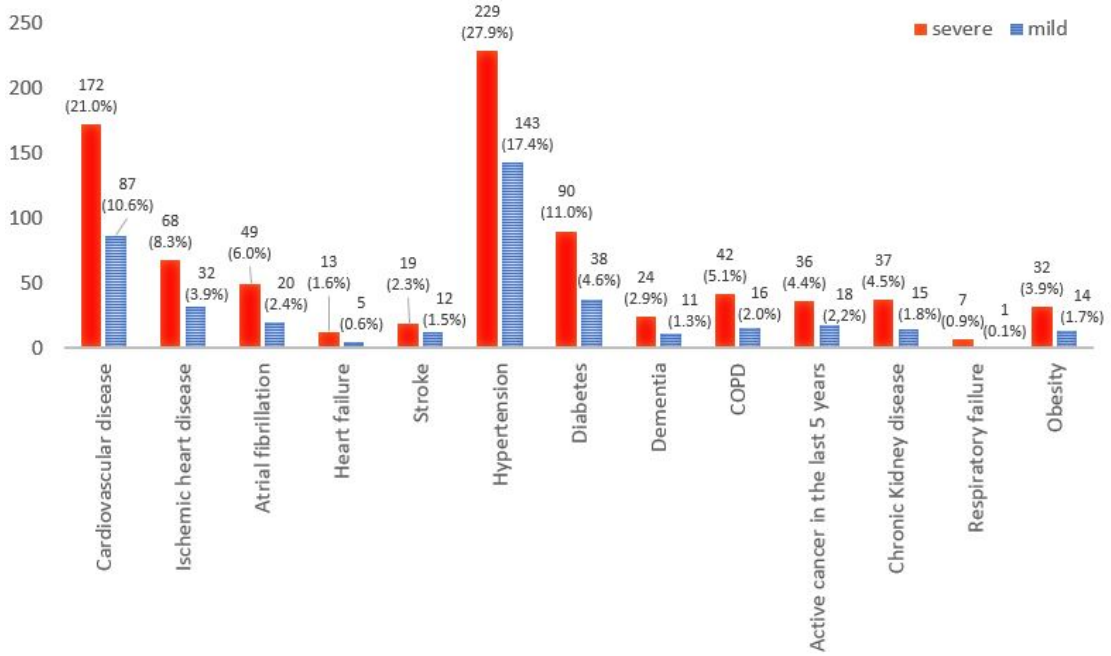


Figure 3.22: Comorbidity distributions between groups. For all data, value and percentage referred to the total population was indicated.

3.4 Segmentation analysis

During this PhD I studied two cases where the segmentation can effect the results. In the first case I analysed the results obtained by the hybrid approach previously introduced, using two different segmentation methods. In the second case I analysed the value of tree different segmentations in NSCLC and how they affect the results.

3.4.1 Segmentation application in AIforCovid

As described in the literature the importance of the segmentation in ML applications is highly demonstrated since different segmentations leads to different extracted features and therefore to different results. Usually, in DL applications, the segmentation step is not necessary since the whole image is provided to the DL algorithm, but in examples such as the project presented in section 3.3 where images have a certain degree of variability due to the fact that the dataset was acquired using different acquisition parameter and machines, a segmentation step may be necessary. Therefore, we investigated if the use of lung regions automatically segmented differently impacts the final performance with respect to the use of lung masks manually delineated.

Segmentation using U-Net The network adopted for the automatic segmentation was a U-Net, which is a convolutional neural network architecture for fast and precise lung segmentation. This system was already trained on non-COVID-19 lung CXR datasets[2], namely the Montgomery County CXR set (MC) [171] and the Japanese Society of Radiological Technology (JSRT) repository [172], using an Adam optimiser and with a binary cross-entropy loss function. Furthermore, during training, a random augmentation phase composed of rotation ($\pm 10°$), horizontal and vertical shift (± 25 pixels), and zoom ($0 - 0.2$) was applied. Furthermore, the batch size was set to 8, and the number of epochs was equal to 100. The MC dataset contains 7,470 CXR collected by the National Library of Medicine within the Open-i service, whereas the JSRT repository is composed of 247 CXR with and without a lung nodule. The U-net requires input images represented as 3-channel 256×256 matrices and, hence, greyscale images were copied to all the channels and then resized. Furthermore, we normalised the pixel intensities as detailed in section 3.1.1.2.1. After these transformations, each image was passed through the convolutional network and pixels were classified as foreground (i.e. the lung) or background. Unlike the handcrafted approach described in section 3.1.1, where the U-Net was used to pre-segment the lungs whose borders were manually refined, in the hybrid approach described in section 3.3, we adopted a fully automated approach since the segmen-

[2]The network is available as detailed in the reference denoted as Imlab-UIIP [170].

tation mask given by the network was used to extract the rectangular bounding box containing the ROI. In the hybrid approach, there was no need for manual intervention since the performance at the level of ROI bounding box segmentation was satisfactory compared with human annotation. Indeed, the Jaccard index and the Dice score were now equal to 0.929 and 0.960, respectively[3]. Next, each ROI was resized to a square so that the longest side of the ROI was mapped to the square side, and the other ROI side was resized accordingly.

To check if the network worked well, all the images in the repository were segmented by two expert radiologists working in parallel using a consensus strategy (Figure 3.23), permitting us to assess the U-Net segmentation performance. We found that the network provides a Jaccard index and a Dice score equal to 0.896 and 0.942, respectively. We deem that such performance was satisfactory since it is only needed to recover the bounding box, as in the hybrid approach, while it would not be sufficient for the exact lung delineation needed by the following handcrafted approach.

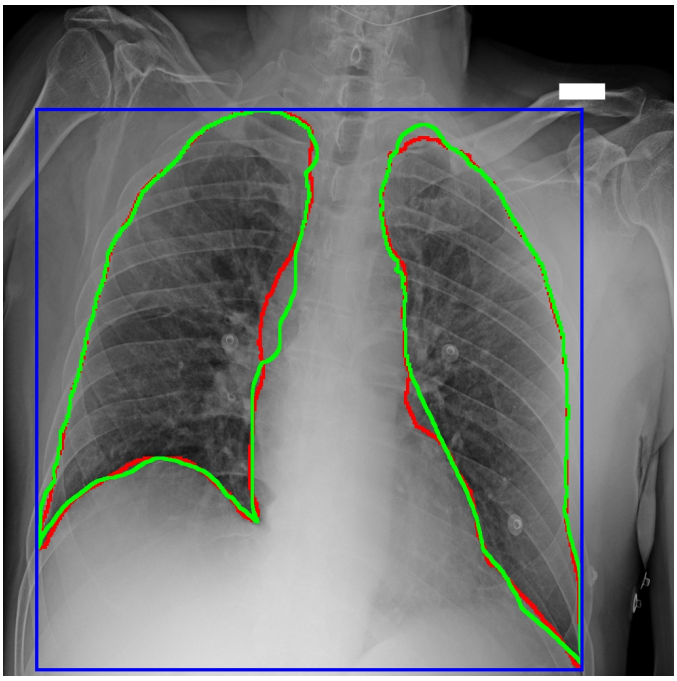


Figure 3.23: Example of the lung segmentation results.
Green line: manual segmentation, red line: segmentation returned by the U-Net, blue line: bounding box from U-Net segmentation.

To compare the segmentation influence on the results, we use the best model combination of the hybrid approach shown in Table 3.19 and Table 3.20 where, however, the CNN is applied on lung regions manually segmented. The results are reported in Table 3.22 showing that the performance attained are almost the

[3]In only one case the segmentation network did not segment the lungs; in this case, the entire original image was used.

same or even slightly worse than those achieved using automatically segmented lung regions. Furthermore, we also find that such differences are not statistically significant according to Wilcoxon's test with $p = 0.05$, except for one case where the results correspond to a classifier with lower performance than those reported in the paper and obtained using the automatic segmentation.

Table 3.22: Performance of the hybrid approach when we use the manual segmentation mask attained using the same setup of learners shown in Table 3.19 and Table 3.20. The last column reports the p-values computed according to Wilcoxon's test.

Validation approach	Input Data	Accuracy	Sensibility	Specificity	p-value
10-fold	CXR images	$.717 \pm .062$	$.759 \pm .077$	$.671 \pm 0.074$	1.160e-03
	Clinical data and CXR images	$.764 \pm .056$	$.764 \pm .073$	$.763 \pm .067$	3.983e-01
LOCO	CXR images	$.686 \pm .093$	$.778 \pm .152$	$.616 \pm .289$	8.182e-01
	Clinical data and CXR images	$.743 \pm .048$	$.712 \pm .123$	$.810 \pm .150$	6.991e-01

3.4.2 Application on NSCLC of different Segmentations

As reported in section 2.3.1, in the literature the problem of the segmentation was widely studied, and authors tried to find approaches providing good performances while introducing a small feature- and result-variability. The literature explored so far shows a clear predominance of studies that focus on radiomics descriptors extracted from the segmentation obtained from the area that can be seen in imaging and no different kind of segmentations were developed.

To address the segmentation problem and the possibility of using another segmentation, we investigated the use of different Volumes-Of-Interest (VOIs) in NSCLC. We exploited quantitative biomarkers computed not only from the Gross Tumour Volume (GTV), which is what can be seen, palpated, or imaged, but also from other two larger VOIs commonly used in radiation treatment clinical practice, which are Clinical Target Volume (CTV) and Planning Target Volume (PTV). The first take into account the subclinical disease spread and the last allows for uncertainties in planning or treatment delivery. The idea, inspired by the radiotherapy practice adopted when planning the treatments, is that the area surrounding the tumour may contain important information that can be quantitatively exploited by a computer-based approach, although it cannot be discerned by the visual interpretation. In fact, this area around the tumour could catch also how the cancer infiltrates the nearby healthy tissues and, hence, it could potentially help predicting the disease behaviour and patient survival.

The analysis of the different segmentation contribution was done in parallel with the development of LBPs features and the comparison with the standard used fea-

tures for radiomics applications. The formulation and results of the LBPs are reported in section 3.1.2.2.

3.4.2.1 Material and Methods

DATASET As described in section 3.1.2.2.1, the dataset consists of 97 patients with NSCLC stage III that were treated with definite concurrent chemoradiotherapy. For each patient, the simulation CT images were acquired before the treatment using a Siemens Somatom Emotion, with 140 Kv, 80 mAs, and 3 mm for slice thickness. Subsequently, all images were preprocessed using a lung filter (kernel B70) and a mediastinum filter (kernel B31). The patients were then clinically followed, for a median follow-up of 18.55 months after that it was possible to divide patients into two classes: 53 dead (class 0) and 44 alive (class 1). The patients had a mean Overall Survival (OS) time of 28.7 ± 26.4 months (min 4.8 months, max 142.6 months). For the study we considered three different VOIs for each patient, the Gross Tumour Volume (GTV), the Clinical Target Volume (CTV), and the Planning Target Volume (PTV) segmented by radiation oncologists. For GTV, the expert radiation oncologists included all the macroscopic disease defined at CT or PET-CT and nodes PET positive and/or larger than 1 cm in the short axis at CT imaging. Starting from GTV, the other volumes can be segmented: CTV former contains GTV plus a margin for sub-clinical disease spread, which therefore cannot be fully imaged, and it represents the volume that must be adequately treated to achieve cure [173]. PTV is a geometric concept that is designed to ensure that the radiotherapy dose is actually delivered to CTV. It is determined expanding CTV with a safety margin, which is necessary to manage internal motion and set-up reproducibility [174].

It is worth noting that the ability to segment tumour mass and the surrounding tissue involved with the tumour to a certain extent is an important professional skill for a medical speciality that is radiation oncology. Several methods are proposed in the literature to reduce interpersonal variability [175, 176]. This is why, in our work, the same radiation oncologist segmented all the images in our repository, and then another lung cancer expert radiation oncologist proofread the segmentations. Furthermore, to mitigate the inter-reader variability, the segmentations were performed according to the international guidelines [177].

Figure 3.24 shows an example of these VOIs: the GTV is drawn in red, and it is the smallest segmentation that precisely delineates the visible tumour; CTV is represented in yellow and it includes areas, where, in the image, there is no evidence of tumour, but where sub-clinical disease may be. PTV, in blu, is the largest segmentation.

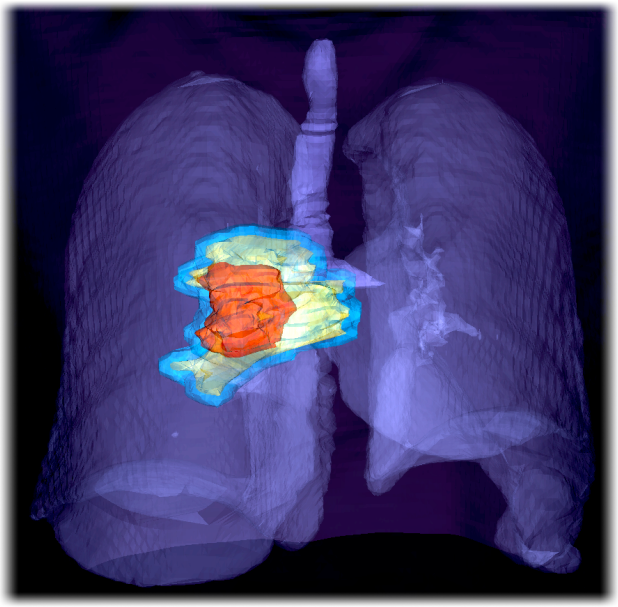


Figure 3.24: Example of different segmentations for a Non-Small Cell Lung Cancer (NSCLC)
The segmentations corresponding to Gross Tumour Volume (GTV) (red), Clinical Target Volume (CTV) (yellow), and Planning Target Volume (PTV) (blue) were highlighted

FEATURES EXTRACTION AND SELECTION As described in paragraph *Feature Computation* of section 3.1.2.2.1, three different types of features were extracted, the first-order statistical features, the 3D Grey level co-occurrence features and the LBP features.

After the feature extraction, a selection step was done to reduce the dimensionality of the problem, lowering the curse of dimensionality and the risk of overfitting. The feature selection pipeline described in Figure 3.10 and introduced in paragraph *Feature selection and Classification* of section 3.1.2.2.1 consists of a wrapper-based approach, which searches and evaluates the best subset of features maximising the performance of a given classification algorithm. To avoid any bias, features are normalised before each step, using a standard scaler, as represented in Figure 3.10.

After the RFE feature selection, which performs a feature ranking according to the feature importance, two alternative paths were considered and referred to as *single ranking* and *total ranking*. On the one hand, in the first path, the ranking procedure was performed independently for each single experiment, i.e., considering a single classifier and a single type of VOI. Straightforwardly, each of the 12 experiments listed in Table 3.23 uses a specific features rank and, in turn, this permits us to optimise the feature set for each learning algorithm and for each VOI.

On the other hand, in the second path, the wrapper first computes the rank of the features for the GTV, CTV, and PTV; then, such rankings computed using the

Table 3.23: Overview of all possible combinations of volume of interests (VOIs) and learning algorithms.

	Classifier			
VOI	***AdaBoost***	***Decision Tree***	***Random Forest***	***XGBoost***
GTV	AdaBoost-GTV	DT-GTV	RF-GTV	XGB-GTV
CTV	AdaBoost-CTV	DT-CTV	RF-CTV	XGB-CTV
PTV	AdaBoost-PTV	DT-PTV	RF-PTV	XGB-PTV

same classifier on the three volumes were summed up to gain a global ranking for every learning algorithm. This implies that we finally have four different feature sets, one for each classifier used in the wrapper, regardless of the VOI used. This permitted designing experiments that are comparable within the same classifier.

3.4.2.2 Results

In Table 3.24, the upper panel shows the performance of the twelve combinations of the classifiers and the VOIs when the single ranking method is used, whereas the lower panel reports the performance for the total ranking case. For the single ranking, the accuracy and AUC values range from 60.82% to 83.51% and from 61.06% to 82.78%, respectively. For the total ranking, the accuracy ranges from 44.33% to 77.32%, and the AUC ranges from 44.43% to 75.96%. The subsets of features selected in each combination, range from two to nine descriptors for the single ranking and from two to 20 descriptors for the total ranking. Note that, in the total ranking case, we obtain a worse range of accuracy, AUC values and larger subsets of selected feature in comparison to the single ranking. This could be expected, since the single ranking approach provides a best feature subset optimised for both the classification algorithm and the considered VOI.

Table 3.24: Performance in all possible combinations
The upper panel refers to the single ranking strategy for feature selection, whereas the lower panel shows the results considering the total ranking option.

Single	**AdaBoost**			**Decision Tree**			**Random Forest**			**XGBoost**		
Ranking	**GTV**	**CTV**	**PTV**	**GTV**	**CTV**	**PTV**	**GTV**	**CTV**	**PTV**	**GTV**	**CTV**	**PTV**
Accuracy	70.10	**83.51**	71.13	70.10	60.82	63.92	73.20	78.35	77.32	68.04	71.13	76.29
AUC	70.13	**82.78**	70.11	70.13	61.06	64.09	71.81	77.29	76.54	66.70	70.11	75.02
Precision	69.81	**90.57**	81.13	69.81	58.49	62.26	86.79	88.68	84.91	67.19	70.49	73.44
Recall	74.00	**81.36**	70.49	74.00	65.96	68.75	70.77	75.81	76.27	81.13	81.13	88.68
Total	**AdaBoost**			**Decision Tree**			**Random Forest**			**XGBoost**		
Ranking	**GTV**	**CTV**	**PTV**	**GTV**	**CTV**	**PTV**	**GTV**	**CTV**	**PTV**	**GTV**	**CTV**	**PTV**
Accuracy	64.95	72.16	68.04	44.33	58.76	65.98	75.26	**77.32**	74.23	64.95	72.16	74.23
AUC	64.84	71.83	67.47	44.43	59.18	65.59	73.89	**75.96**	73.33	62.52	70.09	72.94
Precision	66.04	75.47	73.58	43.40	54.72	69.81	88.68	**90.57**	83.02	62.67	68.06	71.88
Recall	68.63	74.07	69.64	48.94	64.44	68.52	72.31	**73.85**	73.33	88.68	92.45	86.79

A more detailed analysis of Table 3.24 suggests us other three observations. The first is that the performance on GTV in the single ranking is stable across the four classification algorithms. The same does not happen for the total ranking case. We deem this indicates that a system built on this VOI is particularly stable when a volume-specific optimisation occurs for the best feature set selection. Secondly, the overall largest performance was obtained by combining the classifier Adaboost with the CTV in the single ranking procedure. Compared to this result, the best result for the total ranking procedure was obtained combining the Random Forest Classifier with, again, the CTV. Such best results for single and total rankings are highlighted in bold in Table 3.24. As a third observation, we notice that, although both the best experiments exploit the CTV, the largest performance are given by the single ranking approach, where the best feature set was customised on the specific VOI analysed.

We also run the ANOVA test among the VOIs both for the results attained by single and total ranking, whatever the classification paradigm used. When we get a significant result, i.e., at least one of the groups tested differs from the other groups, we proceed with Least Significant Difference (LSD) that detects the statistical differences among a group of results. In the case of results provided by the single ranking, we obtain a p-value equal to 0.0334, suggesting that the performance achieved using the different VOIs are not the same. The LSD tests show that the results of CTV+Adaboost are significantly different ($p < 0.1$) from the others in all cases, except for CTV+Random Forest, PTV+Random Forest, and PTV+XGBoost. In the case of the results provided by the total ranking, the p-value is $1.586 \cdot 10^{-5}$, suggesting again that there are differences between the VOIs. The method with the best AUC, i.e., CTV+Random Forest, shows performance statistically different from all the others except for CTV+AdaBoost, PTV+AdaBoost, GTV+Random Forest, PTV+Random Forest, CTV+XGBoost, and PTV+XGBoost.

Let us now turn the attention to the radiomics signature underlying the best case, i.e., CTV+Adaboost in single ranking: as shown in the first column of Table 3.10 it is composed of eight features belonging to the LBP set, and one belonging to the GLCM set. In detail, such features are: "uniform 3D LBP kurtosis", which estimates the shape of the distribution of all patterns discharging the noisy patterns, "3D LBP energy", which represents the homogeneity of the patterns, "rotation invariant 3D LBP absolute maximum", which is the absolute maximum value among the patterns, regardless of their rotation, "3D LBP energy around the absolute maximum", which estimates the homogeneity of patterns around the absolute maximum value, "rotation invariant 3D LBP energy", which is the homogeneity of the patterns without considering their rotation, "uniform 3D LBP energy around the relative maximum", which estimates the homogeneity of pattern around a relative maximum value discharging the noisy patterns, "uniform 3D LBP entropy", which

is a measure of the variety of patterns discharging the noisy ones, and "uniform 3D LBP skewness", which measures the asymmetry of distribution of all patterns discharging noisy patterns, and "inverse GLCM in direction (−1,−1,0)", which measures the inverse of grey level attenuation in the ROI over the direction (−1,−1,0).

3.4.2.3 Conclusions

In this work, we propose a radiomics approach to predict the Overall Survival in a cohort of 97 patients suffering from locally advanced non-small cell lung cancer. The aim was to analyse three different VOIs commonly used in clinical practice and segmented on CT images routinely collected. Among the different combinations of classification algorithms and VOIs, the best accuracy of the proposed approach is equal to 83.51%, achieved using the Clinical Target Volume, the Adaboost learner, and nine features selected from the pool. The fact that the VOI providing the best results is the CTV, confirms our initial hypothesis that there is information in the surroundings of the visible tumour and that this is important to predict patients' OS.

Chapter 4

Conclusions and future perspectives

Medical Imaging is one of the main methodologies used in medical practice to assess treatments. Its advantage consist in the non-invasive assessment of patients' conditions and the possibility to diagnose pathologies and guide and monitor treatment. Hardware and software in the field of Medical Imaging have significantly improved thanks to the enhancement of imaging contrast agents, the standardisation of acquisition protocols, and the innovation in imaging analysis techniques. These improvements led to the introduction of radiomics, a method that extracts a vast number of features from medical images. As detailed in section 2.2, since the first article on radiomics was published, every year scientists focus their research on different applications of radiomics. A literature review reveals that subsequent research papers generally followed the workflow introduced by the pioneers, using the same techniques, without introducing substantial innovations to the workflow.

This thesis set out to develop alternatives to the radiomic workflow, focusing on four main aspects: the features used in radiomics, the imbalanced dataset problem, the combination of Deep Learning and Machine Learning algorithms and the segmentation analysis. This thesis demonstrates that improvements of the radiomics workflow are possible. First, the introduction of new features discussed in section 3.1 evidences that parametric maps and Local Binary Patterns can significantly improve the classification performance, compared to the standard texture measures commonly used in radiomics. These results indicate that even if medical societies are standardising radiomics features (Image Biomarker Standardisation Initiative [178]), the list of features currently used in radiomics is not exhaustive. Second, it emerges that class imbalance is a problem especially for medical applications, since usually the datasets have much more data on benign than malignant conditions. In the radiomics literature, very few authors acknowledge this issue and try to tackle it. The first method developed in this thesis to address the imbalance dataset problem showes that using an over-sampling method considerably enhances the accuracy and

precision of the predictions for all classes. The second method shed light on the potential of using a cost-sensitive approach to address the imbalanced nature of the radiomics data. Both results evidence the importance of addressing this problem, in order to improve the stability and performance of the models. Third, the thesis proposed a new combination of Deep Learning and Machine Learning algorithms, which demonstrated that joint learning could be another direction of investigation. This combination could allow to extract correlated information across clinical and imaging data to enforce the network. The final section of the thesis 3.4 was built on the hypothesis that there is information in the surroundings of a visible tumour that could be used to predict patients' overall survival. This initial idea was confirmed by comparing the results of different segmentations, which showed that the best results were obtained with the Clinical Target Volume rather than the Gross Tumour Volume. This result marks the potential of targeting not only the visible tumour as an area to be segmented but also the "invisible" tumour cells around the main tumour.

This thesis is part of a larger research project designed to improve the radiomics workflow that has been carried out at the *Campus Bio-medico di Roma*. Three main lines of research are in the pipeline. First, an assessment of the Local Binary Patterns, aimed at developing a way to extract Local Binary Patterns within a Deep Learning model. Second, addressing the imbalanced dataset problem with i) a siamese model that contains two identical subnetworks used to find the similarity of two inputs and ii) a generative model to produce new data similar to the data belonging to the minority class. Third, the exploration of the combinations of Machine Learning and Deep Learning models, in order to find the optimal balance between them.

Bibliography

[1] H. Y. F. Wong, H. Y. S. Lam *et al.*, "Frequency and distribution of chest radiographic findings in patients positive for COVID-19," *Radiology*, vol. 296, no. 2, pp. E72–E78, 2020.

[2] M. A. Orsi, G. Oliva *et al.*, "Feasibility, reproducibility, and clinical validity of a quantitative chest x-ray assessment for COVID-19," *The American journal of tropical medicine and hygiene*, vol. 103, no. 2, pp. 822–827, 2020.

[3] L. Hood and S. H. Friend, "Predictive, personalized, preventive, participatory (P4) cancer medicine," *Nature reviews Clinical oncology*, vol. 8, no. 3, pp. 184–187, 2011.

[4] F. S. Collins, M. Morgan, and A. Patrinos, "The human genome project: lessons from large-scale biology," *Science*, vol. 300, no. 5617, pp. 286–290, 2003.

[5] M. Baumann, T. Hölscher, and A. C. Begg, "Towards genetic prediction of radiation responses: ESTRO's GENEPI project," *Radiotherapy and Oncology*, vol. 69, no. 2, pp. 121–125, 2003.

[6] G. A. Evans, "Designer science and the "omic" revolution," *Nature Biotechnology*, vol. 18, no. 2, pp. 127–127, 2000.

[7] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. Van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker *et al.*, "Radiomics: extracting more information from medical images using advanced feature analysis," *European journal of cancer*, vol. 48, no. 4, pp. 441–446, 2012.

[8] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld *et al.*, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature communications*, vol. 5, no. 1, pp. 1–9, 2014.

[9] K.-C. Song, Y.-H. Yan, W.-H. Chen, and X. Zhang, "Research and Perspective on Local Binary Pattern," *Acta Automatica Sinica*, vol. 39,

no. 6, pp. 730–744, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1874102913600518

[10] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher *et al.*, "Radiomics: the process and the challenges," *Magnetic resonance imaging*, vol. 30, no. 9, pp. 1234–1248, 2012.

[11] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification," *IEEE Trans Syst Man Cybern*, vol. SMC-3, pp. 610–621, 01 1973.

[12] M. Sollini, L. Antunovic, A. Chiti, and M. Kirienko, "Towards clinical application of image mining: a systematic review on artificial intelligence and radiomics," *European journal of nuclear medicine and molecular imaging*, vol. 46, no. 13, pp. 2656–2672, 2019.

[13] C. Parmar, E. Rios Velazquez, R. Leijenaar, M. Jermoumi, S. Carvalho, R. H. Mak, S. Mitra, B. U. Shankar, R. Kikinis, B. Haibe-Kains *et al.*, "Robust radiomics feature quantification using semiautomatic volumetric segmentation," *PloS one*, vol. 9, no. 7, p. e102107, 2014.

[14] W. L. Bi, A. Hosny, M. B. Schabath, M. L. Giger, N. J. Birkbak, A. Mehrtash, T. Allison, O. Arnaout, C. Abbosh, I. F. Dunn *et al.*, "Artificial intelligence in cancer imaging: clinical challenges and applications," *CA: a cancer journal for clinicians*, vol. 69, no. 2, pp. 127–157, 2019.

[15] M. Avanzo, J. Stancanello, and I. El Naqa, "Beyond imaging: the promise of radiomics," *Physica Medica*, vol. 38, pp. 122–139, 2017.

[16] S. Rizzo, F. Botta, S. Raimondi, D. Origgi, C. Fanciullo, A. G. Morganti, and M. Bellomi, "Radiomics: the facts and the challenges of image analysis," *European radiology experimental*, vol. 2, no. 1, pp. 1–8, 2018.

[17] J. E. van Timmeren, R. T. Leijenaar, W. van Elmpt, J. Wang, Z. Zhang, A. Dekker, and P. Lambin, "Test–retest data for radiomics feature stability analysis: generalizable or study-specific?" *Tomography*, vol. 2, no. 4, pp. 361–365, 2016.

[18] R. S. of North America, "Quantitative Imaging Biomarkers Alliance," https://www.rsna.org/research/quantitative-imaging-biomarkers-alliance, accessed: 2021-11-29.

[19] I. El Naqa, "The role of quantitative PET in predicting cancer treatment outcomes," *Clinical and translational imaging*, vol. 2, no. 4, pp. 305–320, 2014.

[20] E. C. Ehman, G. B. Johnson, J. E. Villanueva-Meyer, S. Cha, A. P. Leynes, P. E. Z. Larson, and T. A. Hope, "PET/MRI: where might it replace PET/CT?" *Journal of Magnetic Resonance Imaging*, vol. 46, no. 5, pp. 1247–1262, 2017.

[21] J. Antunes, S. Viswanath, M. Rusu, L. Valls, C. Hoimes, N. Avril, and A. Madabhushi, "Radiomics analysis on FLT-PET/MRI for characterization of early treatment response in renal cell carcinoma: a proof-of-concept study," *Translational oncology*, vol. 9, no. 2, pp. 155–162, 2016.

[22] B.-B. Chen, "PET-MRI of the Pancreas and Kidneys," *Current Radiology Reports*, vol. 5, no. 8, pp. 1–9, 2017.

[23] N. England and N. Improvement, "Diagnostic imaging dataset statistical release," *London: Department of Health*, vol. 421, 2016.

[24] L. Duron, J. Savatovsky, L. Fournier, and A. Lecler, "Can we use radiomics in ultrasound imaging? impact of preprocessing on feature repeatability," *Diagnostic and Interventional Imaging*, 2021.

[25] K. Rajendran, M. Petersilka, A. Henning, E. R. Shanblatt, B. Schmidt, T. G. Flohr, A. Ferrero, F. Baffour, F. E. Diehn, L. Yu *et al.*, "First clinical photon-counting detector ct system: Technical evaluation," *Radiology*, p. 212579, 2021.

[26] J. R. Rajagopal, J. Hoye, M. Robins, E. C. Jones, and E. Samei, "Accuracy and variability of radiomics in photon-counting ct: texture features and lung lesion morphology," in *Medical Imaging 2019: Physics of Medical Imaging*, vol. 10948. SPIE, 2019, pp. 1102–1110.

[27] P. Lambin, R. T. Leijenaar, T. M. Deist, J. Peerlings, E. E. De Jong, J. Van Timmeren, S. Sanduleanu, R. T. Larue, A. J. Even, A. Jochems *et al.*, "Radiomics: the bridge between medical imaging and personalized medicine," *Nature reviews Clinical oncology*, vol. 14, no. 12, pp. 749–762, 2017.

[28] Y. Balagurunathan, Y. Gu, H. Wang, V. Kumar, O. Grove, S. Hawkins, J. Kim, D. B. Goldgof, L. O. Hall, R. A. Gatenby *et al.*, "Reproducibility and prognosis of quantitative features extracted from ct images," *Translational oncology*, vol. 7, no. 1, pp. 72–87, 2014.

[29] J. Zhao, Z. Meng, L. Wei, C. Sun, Q. Zou, and R. Su, "Supervised brain tumor segmentation based on gradient and context-sensitive features," *Frontiers in neuroscience*, vol. 13, p. 144, 2019.

[30] L. Bischof and R. Adams, "Seeded region growing," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, pp. 641–647, 1994.

[31] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[33] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020.

[34] R. Anantharaman, M. Velazquez, and Y. Lee, "Utilizing mask R-CNN for detection and segmentation of oral diseases," in *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE, 2018, pp. 2197–2204.

[35] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in *Deep learning in medical image analysis and multimodal learning for clinical decision support*. Springer, 2018, pp. 3–11.

[36] M. M. Galloway, "Texture analysis using gray level run lengths," *Computer graphics and image processing*, vol. 4, no. 2, pp. 172–179, 1975.

[37] F. Tixier, C. C. Le Rest, M. Hatt, N. Albarghach, O. Pradier, J.-P. Metges, L. Corcos, and D. Visvikis, "Intratumor heterogeneity characterized by textural features on baseline 18F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer," *Journal of Nuclear Medicine*, vol. 52, no. 3, pp. 369–378, 2011.

[38] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[39] S. University, "CS231n: Convolutional Neural Networks for Visual Recognition," https://cs231n.github.io/convolutional-networks/, accessed: 2021-11-29.

[40] G. I. Webb, *Overfitting*. Boston, MA: Springer US, 2010, pp. 744–744. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_623

[41] S. Wang, J. Tang, and H. Liu, "Feature Selection." 2017.

[42] V. Kumar and S. Minz, "Feature selection: a literature review," *SmartCR*, vol. 4, no. 3, pp. 211–229, 2014.

[43] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[44] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*. Ieee, 2015, pp. 1200–1205.

[45] J. Miao and L. Niu, "A survey on feature selection," *Procedia Computer Science*, vol. 91, pp. 919–926, 2016.

[46] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and H. J. Aerts, "Machine learning methods for quantitative radiomic biomarkers," *Scientific reports*, vol. 5, no. 1, pp. 1–11, 2015.

[47] M. Robnik-Šikonja and I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF," *Machine learning*, vol. 53, no. 1, pp. 23–69, 2003.

[48] M. Kuhn and K. Johnson, "Applied predictive modelling springer," *New York Heidelberg Dordrecht London*, 2013.

[49] I. Guyon *et al.*, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[50] H. Belyadi and A. Haghighat, "Chapter 5 - Supervised learning," in *Machine Learning Guide for Oil and Gas Using Python*, H. Belyadi and A. Haghighat, Eds. Gulf Professional Publishing, 2021, pp. 169–295. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128219294000044

[51] E. Konukoglu and B. Glocker, "Chapter 19 - Random forests in medical image computing," in *Handbook of Medical Image Computing and Computer Assisted Intervention*, ser. The Elsevier and MICCAI Society Book Series, S. K. Zhou, D. Rueckert, and G. Fichtinger, Eds. Academic Press, 2020, pp. 457–480. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128161760000247

[52] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.* " O'Reilly Media, Inc.", 2019.

[53] A. Subasi, "Chapter 7 - Clustering examples," in *Practical Machine Learning for Data Analysis Using Python*, A. Subasi, Ed. Academic Press, 2020, pp. 465–511. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128213797000072

[54] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, no. 2, pp. 373–440, 2020.

[55] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[56] A. Vial, D. Stirling, M. Field, M. Ros, C. Ritz, M. Carolan, L. Holloway, and A. A. Miller, "The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review," *Transl Cancer Res*, vol. 7, no. 3, pp. 803–816, 2018.

[57] V. S. Parekh and M. A. Jacobs, "Deep learning and radiomics in precision medicine," *Expert review of precision medicine and drug development*, vol. 4, no. 2, pp. 59–72, 2019.

[58] F. Chollet, *Deep Learning with Python*. Manning, Nov. 2017.

[59] S. Saha, "A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way," https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53, accessed: 2021-12-13.

[60] Y. Wang, Y. Li, Y. Song, and X. Rong, "The influence of the activation function in a convolution neural network model of facial expression recognition," *Applied Sciences*, vol. 10, no. 5, p. 1897, 2020.

[61] I. Castiglioni, L. Rundo, M. Codari, G. Di Leo, C. Salvatore, M. Interlenghi, F. Gallivanone, A. Cozzi, N. C. D'Amico, and F. Sardanelli, "AI applications to medical images: From machine learning to deep learning," *Physica Medica*, vol. 83, pp. 9–24, 2021.

[62] S. S. Yip and H. J. Aerts, "Applications and limitations of radiomics," *Physics in Medicine & Biology*, vol. 61, no. 13, p. R150, 2016.

[63] J. W. Lee and S. M. Lee, "Radiomics in oncological PET/CT: clinical applications," *Nuclear medicine and molecular imaging*, vol. 52, no. 3, pp. 170–189, 2018.

[64] A. Varrone, S. Asenbaum, T. Vander Borght, J. Booij, F. Nobili, K. Någren, J. Darcourt, Ö. L. Kapucu, K. Tatsch, P. Bartenstein *et al.*, "EANM procedure

guidelines for PET brain imaging using [18 F] FDG, version 2," *European journal of nuclear medicine and molecular imaging*, vol. 36, no. 12, pp. 2103–2110, 2009.

[65] J.-h. Jung and B.-C. Ahn, "Current radiopharmaceuticals for positron emission tomography of brain tumors," *Brain tumor research and treatment*, vol. 6, no. 2, pp. 47–53, 2018.

[66] M. Zhou, J. Scott, B. Chaudhury, L. Hall, D. Goldgof, K. W. Yeom, M. Iv, Y. Ou, J. Kalpathy-Cramer, S. Napel *et al.*, "Radiomics in brain tumor: image assessment, quantitative feature descriptors, and machine-learning approaches," *American Journal of Neuroradiology*, vol. 39, no. 2, pp. 208–216, 2018.

[67] A. Chaddad, M. J. Kucharczyk, P. Daniel, S. Sabri, B. J. Jean-Claude, T. Niazi, and B. Abdulkarim, "Radiomics in glioblastoma: current status and challenges facing clinical implementation," *Frontiers in oncology*, vol. 9, p. 374, 2019.

[68] E. Lotan, R. Jain, N. Razavian, G. M. Fatterpekar, and Y. W. Lui, "State of the art: machine learning applications in glioma imaging," *American Journal of Roentgenology*, vol. 212, no. 1, pp. 26–37, 2019.

[69] N. Soni, S. Priya, and G. Bathla, "Texture analysis in cerebral gliomas: a review of the literature," *American Journal of Neuroradiology*, vol. 40, no. 6, pp. 928–934, 2019.

[70] S. Ather, T. Kadir, and F. Gleeson, "Artificial intelligence and radiomics in pulmonary nodule management: current status and future applications," *Clinical radiology*, vol. 75, no. 1, pp. 13–19, 2020.

[71] M. Avanzo, J. Stancanello, G. Pirrone, and G. Sartor, "Radiomics and deep learning in lung cancer," *Strahlentherapie und Onkologie*, pp. 1–9, 2020.

[72] B. Chen, R. Zhang, Y. Gan, L. Yang, and W. Li, "Development and clinical application of radiomics in lung cancer," *Radiation Oncology*, vol. 12, no. 1, pp. 1–8, 2017.

[73] J. Constanzo, L. Wei, H.-H. Tseng, and I. El Naqa, "Radiomics in precision medicine for lung cancer," *Translational lung cancer research*, vol. 6, no. 6, p. 635, 2017.

[74] M. Rabbani, J. Kanevsky, K. Kafi, F. Chandelier, and F. J. Giles, "Role of artificial intelligence in the care of patients with nonsmall cell lung cancer," *European journal of clinical investigation*, vol. 48, no. 4, p. e12901, 2018.

[75] R. Thawani, M. McLane, N. Beig, S. Ghose, P. Prasanna, V. Velcheti, and A. Madabhushi, "Radiomics and radiogenomics in lung cancer: a review for the clinician," *Lung cancer*, vol. 115, pp. 34–41, 2018.

[76] M. Scrivener, E. E. de Jong, J. E. van Timmeren, T. Pieters, B. Ghaye, and X. Geets, "Radiomics applied to lung cancer: a review," *Transl Cancer Res*, vol. 5, no. 4, pp. 398–409, 2016.

[77] P. S. van Rossum, C. Xu, D. V. Fried, L. Goense, L. E. Court, and S. H. Lin, "The emerging field of radiomics in esophageal cancer: current evidence and future potential," *Translational cancer research*, vol. 5, no. 4, p. 410, 2016.

[78] B.-R. Sah, K. Owczarczyk, M. Siddique, G. J. Cook, and V. Goh, "Radiomics in esophageal and gastric cancer," *Abdominal radiology*, vol. 44, no. 6, pp. 2048–2058, 2019.

[79] W. K. Jeong, N. Jamshidi, E. R. Felker, S. S. Raman, and D. S. Lu, "Radiomics and radiogenomics of primary liver cancers," *Clinical and molecular hepatology*, vol. 25, no. 1, p. 21, 2019.

[80] T. Wakabayashi, F. Ouhmich, C. Gonzalez-Cabrera, E. Felli, A. Saviano, V. Agnus, P. Savadjiev, T. F. Baumert, P. Pessaux, J. Marescaux *et al.*, "Radiomics in hepatocellular carcinoma: a quantitative review," *Hepatology international*, vol. 13, no. 5, pp. 546–559, 2019.

[81] F. Fiz, L. Viganò, N. Gennaro, G. Costa, L. La Bella, A. Boichuk, L. Cavinato, M. Sollini, L. S. Politi, A. Chiti *et al.*, "Radiomics of liver metastases: A systematic review," *Cancers*, vol. 12, no. 10, p. 2881, 2020.

[82] H. J. Park, B. Park, and S. S. Lee, "Radiomics and deep learning: hepatic applications," *Korean journal of radiology*, vol. 21, no. 4, pp. 387–401, 2020.

[83] A. D. de Leon, P. Kapur, and I. Pedrosa, "Radiomics in kidney cancer: Mr imaging," *Magnetic Resonance Imaging Clinics*, vol. 27, no. 1, pp. 1–13, 2019.

[84] N. Dinapoli, C. Casà, B. Barbaro, G. V. Chiloiro, A. Damiani, M. Di Matteo, A. Farchione, M. A. Gambacorta, R. Gatta, V. Lanzotti *et al.*, "Radiomics for rectal cancer," *Transl Cancer Res*, vol. 5, no. 4, pp. 424–431, 2016.

[85] F. Valdora, N. Houssami, F. Rossi, M. Calabrese, and A. S. Tagliafico, "Rapid review: radiomics and breast cancer," *Breast cancer research and treatment*, vol. 169, no. 2, pp. 217–229, 2018.

[86] P. Crivelli, R. E. Ledda, N. Parascandolo, A. Fara, D. Soro, and M. Conti, "A new challenge for radiologists: radiomics in breast cancer," *BioMed research international*, vol. 2018, 2018.

[87] B. Reig, L. Heacock, K. J. Geras, and L. Moy, "Machine learning in breast MRI," *Journal of Magnetic Resonance Imaging*, vol. 52, no. 4, pp. 998–1018, 2020.

[88] R. D. Chitalia and D. Kontos, "Role of texture analysis in breast MRI as a cancer biomarker: A review," *Journal of Magnetic Resonance Imaging*, vol. 49, no. 4, pp. 927–938, 2019.

[89] R. Cuocolo, M. B. Cipullo, A. Stanzione, L. Ugga, V. Romeo, L. Radice, A. Brunetti, and M. Imbriaco, "Machine learning applications in prostate cancer magnetic resonance imaging," *European radiology experimental*, vol. 3, no. 1, pp. 1–8, 2019.

[90] Y. Sun, H. M. Reynolds, B. Parameswaran, D. Wraith, M. E. Finnegan, S. Williams, and A. Haworth, "Multiparametric MRI and radiomics in prostate cancer: a review," *Australasian physical & engineering sciences in medicine*, vol. 42, no. 1, pp. 3–25, 2019.

[91] R. Stoyanova, M. Takhar, Y. Tschudi, J. C. Ford, G. Solórzano, N. Erho, Y. Balagurunathan, S. Punnen, E. Davicioni, R. J. Gillies *et al.*, "Prostate cancer radiomics and the promise of radiogenomics," *Translational cancer research*, vol. 5, no. 4, p. 432, 2016.

[92] A. Stanzione, M. Gambardella, R. Cuocolo, A. Ponsiglione, V. Romeo, and M. Imbriaco, "Prostate MRI radiomics: A systematic review and radiomic quality score assessment," *European journal of radiology*, vol. 129, p. 109095, 2020.

[93] R. T. Larue, G. Defraene, D. De Ruysscher, P. Lambin, and W. Van Elmpt, "Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures," *The British journal of radiology*, vol. 90, no. 1070, p. 20160665, 2017.

[94] G. J. Cook, M. Siddique, B. P. Taylor, C. Yip, S. Chicklore, and V. Goh, "Radiomics in PET: principles and applications," *Clinical and Translational Imaging*, vol. 2, no. 3, pp. 269–276, 2014.

[95] H. J. Aerts, "The potential of radiomic-based phenotyping in precision medicine: a review," *JAMA oncology*, vol. 2, no. 12, pp. 1636–1642, 2016.

[96] B. Q. Huynh, H. Li, and M. L. Giger, "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks," *Journal of Medical Imaging*, vol. 3, no. 3, p. 034501, 2016.

[97] L. Boldrini, J.-E. Bibault, C. Masciocchi, Y. Shen, and M.-I. Bittner, "Deep learning: a review for the radiation oncologist," *Frontiers in oncology*, vol. 9, p. 977, 2019.

[98] K. Suzuki, "Overview of deep learning in medical imaging," *Radiological physics and technology*, vol. 10, no. 3, pp. 257–273, 2017.

[99] V. S. Parekh and M. A. Jacobs, "Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric mri," *NPJ breast cancer*, vol. 3, no. 1, pp. 1–9, 2017.

[100] R. L. Gonzalez and A. Alberich Bayarri, "Imaging, AI and radiomics to understand and fight coronavirus Covid-19," https://quibim.com/imaging-ai-and-radiomics-to-understand-and-fight-coronavirus-covid-19/, accessed: 2021-11-15.

[101] X. Yang, Y. Yu, J. Xu, H. Shu, H. Liu, Y. Wu, L. Zhang, Z. Yu, M. Fang, T. Yu *et al.*, "Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study," *The Lancet Respiratory Medicine*, 2020.

[102] M. H. Hesamian, W. Jia, X. He, and P. Kennedy, "Deep learning techniques for medical image segmentation: achievements and challenges," *Journal of digital imaging*, vol. 32, no. 4, pp. 582–596, 2019.

[103] S. Rajaraman, J. Siegelman, P. O. Alderson, L. S. Folio, L. R. Folio, and S. K. Antani, "Iteratively Pruned Deep Learning Ensembles for COVID-19 Detection in Chest X-Rays," *IEEE Access*, vol. 8, pp. 115 041–115 050, 2020.

[104] B. Ross, "Mutual Information between Discrete and Continuous Data Sets," *PloS one*, vol. 9, p. e87357, 02 2014.

[105] A. Arcuri and G. Fraser, "Parameter tuning or default values? An empirical investigation in search-based software engineering," *Empirical Software Engineering*, vol. 18, no. 3, pp. 594–623, 2013.

[106] W. D. Penny, K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols, *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

[107] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[108] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[109] T. Chen *et al.*, "Xgboost: extreme gradient boosting," *R package version 0.4-2*, pp. 1–4, 2015.

[110] V. Griethuysen *et al.*, "Computational radiomics system to decode the radiographic phenotype," *Cancer research*, vol. 77, no. 21, pp. e104–e107, 2017.

[111] A. Krizhevsky *et al.*, "Imagenet classification with deep convolutional neural networks," pp. 1097–1105, 2012.

[112] K. He *et al.*, "Deep residual learning for image recognition," pp. 770–778, 2016.

[113] J. Y. Y. Kwan *et al.*, "Radiomic biomarkers to refine risk models for distant metastasis in HPV-related oropharyngeal carcinoma," *International Journal of Radiation Oncology* Biology* Physics*, vol. 102, no. 4, pp. 1107–1116, 2018.

[114] N. V. Chawla, "Data mining for imbalanced datasets: An overview," *Data mining and knowledge discovery handbook*, pp. 875–886, 2009.

[115] P. Soda, "A multi-objective optimisation approach for class imbalance learning," *Pattern Recognition*, vol. 44, no. 8, pp. 1801–1810, 2011.

[116] "Vestibular Schwannoma (Acoustic Neuroma) and Neurofibromatosis," https://www.nidcd.nih.gov/health/vestibular-schwannoma-acoustic-neuroma-and-neurofibromatosis, Aug. 2015.

[117] T. J. Gal, J. Shinn, and B. Huang, "Current Epidemiology and Management Trends in Acoustic Neuroma," *Otolaryngology–Head and Neck Surgery: Official Journal of American Academy of Otolaryngology-Head and Neck Surgery*, vol. 142, no. 5, pp. 677–681, May 2010.

[118] "CyberKnife System from Accuray."

[119] C. J. Przybylowski, J. F. Baranoski, G. M. Paisan, K. M. Chapple, A. J. Meeusen, S. Sorensen, K. K. Almefty, and R. W. Porter, "CyberKnife radiosurgery for acoustic neuromas: Tumor control and clinical outcomes," *Journal of Clinical Neuroscience*, vol. 63, pp. 72–76, 2019.

[120] D. Rueß, L. Pöhlmann, A. Hellerbach, C. Hamisch, M. Hoevels, H. Treuer, S. Grau, K. Jablonska, M. Kocher, and M. I. Ruge, "Acoustic neuroma treated with stereotactic radiosurgery: follow-up of 335 patients," *World neurosurgery*, vol. 116, pp. e194–e202, 2018.

[121] "ProHance® (Gadoteridol) injection, 279.3 mg/mL — Bracco Imaging," https://imaging.bracco.com/us-en/products/magnetic-resonance-imaging/prohance, Mar. 2018.

[122] "3D Slicer," https://www.slicer.org/, Mar. 2018.

[123] A. Fedorov *et al.*, "3D Slicer as an Image Computing Platform for the Quantitative Imaging Network," *Magnetic resonance imaging*, vol. 30, no. 9, pp. 1323–1341, Nov. 2012.

[124] R. T. Leijenaar, G. Nalbantov, S. Carvalho, W. J. Van Elmpt, E. G. Troost, R. Boellaard, H. J. Aerts, R. J. Gillies, and P. Lambin, "The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis," *Scientific Reports*, vol. 5, p. 11075, 2015.

[125] L. Zhang *et al.*, "IBEX: An open infrastructure software platform to facilitate collaborative work in radiomics," *Medical Physics*, vol. 42, no. 3, pp. 1341–1353, Mar. 2015.

[126] F. Orlhac, M. Soussan, J.-A. Maisonobe, C. A. Garcia, B. Vanderlinden, and I. Buvat, "Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis," *J Nucl Med*, vol. 55, no. 3, pp. 414–22, 2014.

[127] F. Tixier *et al.*, "Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET," *Journal of Nuclear Medicine*, vol. 53, no. 5, p. 693, 2012.

[128] C. Zhang, J. Bi, and P. Soda, "Feature selection and resampling in class imbalance learning: Which comes first? An empirical study in the biological domain," in *Bioinformatics and Biomedicine (BIBM), 2017 IEEE Int. Conf. on.* IEEE, 2017, pp. 933–938.

[129] P. E. McKnight and J. Najab, "Mann-Whitney U Test," *The Corsini encyclopedia of psychology*, pp. 1–1, 2010.

[130] J. Bi and C. Zhang, "An Empirical Comparison on State-of-the-art Multi-class Imbalance Learning Algorithms and A New Diversified Ensemble Learning Scheme," *Knowledge-Based Systems*, 2018.

[131] G. Iannello *et al.*, "On the use of classification reliability for improving performance of the one-per-class decomposition method," *Data & Knowledge Engineering*, vol. 68, no. 12, pp. 1398–1410, 2009.

[132] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, no. 3, pp. 321–357, 2002.

[133] N. C. D'Amico, R. Sicilia, E. Cordelli, G. Valbusa, E. Grossi, I. B. Zanetti, G. Beltramo, D. Fazzini, G. Scotti, G. Iannello *et al.*, "Radiomics for Predicting CyberKnife response in acoustic neuroma: a pilot study," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 847–852.

[134] J. A. Oliver, M. Budzevich, D. Hunt, E. G. Moros, K. Latifi, T. J. Dilling, V. Feygelman, and G. Zhang, "Sensitivity of image features to noise in conventional and respiratory-gated pet/ct images of lung cancer: uncorrelated noise effects," *Technology in cancer research & treatment*, vol. 16, no. 5, pp. 595–608, 2017.

[135] I. Kononenko, "Estimating attributes: analysis and extensions of RELIEF," in *European conference on machine learning*. Springer, 1994, pp. 171–182.

[136] H. P. Wilson, P. M. Price, K. Ashkan, A. Edwards, M. M. Green, T. Cross, R. P. Beaney, R. Davies, A. Sibtain, N. P. Plowman *et al.*, "CyberKnife Radiosurgery of Skull-base Tumors: A UK Center Experience," *Cureus*, vol. 10, no. 3, 2018.

[137] X. Y. Liu, J. Wu, and Z. H. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 39, no. 2, pp. 539–550, 2009.

[138] S. Kotsiantis and P. Pintelas, "Mixture of expert agents for handling imbalanced data sets," *Annals of Mathematics, Computing and Teleinformatics*, vol. 1, no. 1, pp. 46–55, 2003.

[139] P. Soda, "An experimental comparison of MES aggregation rules in case of imbalanced datasets," in *Computer-Based Medical Systems, 2009. 22nd IEEE Int. Symp. on*, 2009, pp. 1–6.

[140] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[141] P. Soda and G. Iannello, "A multi-expert system to classify fluorescent intensity in antinuclear autoantibodies testing," in *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*. IEEE, 2006, pp. 219–224.

[142] C. Chen, A. Liaw, L. Breiman *et al.*, "Using random forest to learn imbalanced data," *University of California, Berkeley*, vol. 110, pp. 1–12, 2004.

[143] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[144] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[145] D. H. Wolpert and W. G. Macready, "No Free Lunch Theorems for Optimization," *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67–82, 1997.

[146] Y.-C. Ho and D. L. Pepyne, "Simple explanation of the no-free-lunch theorem and its implications," *Journal of optimization theory and applications*, vol. 115, no. 3, pp. 549–570, 2002.

[147] A. Fernández, V. López, M. Galar, M. J. Del Jesus, and F. Herrera, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches," *Knowledge-based systems*, vol. 42, pp. 97–110, 2013.

[148] Z. Zhang, J. Yang, A. Ho, W. Jiang, J. Logan, X. Wang, P. D. Brown, S. L. McGovern, N. Guha-Thakurta, S. D. Ferguson *et al.*, "A predictive model for distinguishing radiation necrosis from tumour progression after gamma knife radiosurgery based on radiomic features from mr images," *European radiology*, vol. 28, no. 6, pp. 2255–2263, 2018.

[149] A. Ibrahim, S. Primakov, M. Beuque, H. Woodruff, I. Halilaj, G. Wu, T. Refaee, R. Granzier, Y. Widaatalla, R. Hustinx *et al.*, "Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework," *Methods*, vol. 188, pp. 20–29, 2021.

[150] X. Zhang, D. Wang, J. Shao, S. Tian, W. Tan, Y. Ma, Q. Xu, X. Ma, D. Li, J. Chai *et al.*, "A deep learning integrated radiomics model for identification of coronavirus disease 2019 using computed tomography," *Scientific reports*, vol. 11, no. 1, pp. 1–12, 2021.

[151] D. Kermany, K. Zhang, M. Goldbaum *et al.*, "Labeled optical coherence tomography (oct) and chest x-ray images for classification," *Mendeley data*, vol. 2, no. 2, 2018.

[152] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[153] S. Rajaraman, S. Sornapudi, P. O. Alderson, L. R. Folio, and S. K. Antani, "Analyzing inter-reader variability affecting deep ensemble learning for COVID-19 detection in chest radiographs," *PloS one*, vol. 15, no. 11, p. e0242301, 2020.

[154] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *arXiv preprint arXiv:1404.5997*, 2014.

[155] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[156] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[157] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[158] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and $<$ 0.5 MB model size," *arXiv preprint arXiv:1602.07360*, 2016.

[159] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[160] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[161] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient CNN architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[162] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[163] J. Mockus, *Bayesian approach to global optimization: theory and applications*. Springer Science & Business Media, 2012, vol. 37.

[164] A. Arcuri and G. Fraser, "Parameter tuning or default values? an empirical investigation in search-based software engineering," *Empirical Software Engineering*, vol. 18, no. 3, pp. 594–623, 2013.

[165] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[166] J. Cleverley, J. Piper, and M. M. Jones, "The role of chest radiography in confirming covid-19 pneumonia," *bmj*, vol. 370, 2020.

[167] T. Chen, D. Wu, H. Chen, W. Yan, D. Yang, G. Chen, K. Ma, D. Xu, H. Yu, H. Wang *et al.*, "Clinical characteristics of 113 deceased patients with coronavirus disease 2019: retrospective study," *Bmj*, vol. 368, 2020.

[168] I. J. Borges do Nascimento, N. Cacic, H. M. Abdulazeem, T. C. von Groote, U. Jayarajah, I. Weerasekara, M. A. Esfahani, V. T. Civile, A. Marusic, A. Jeroncic *et al.*, "Novel coronavirus infection (COVID-19) in humans: a scoping review and meta-analysis," *Journal of clinical medicine*, vol. 9, no. 4, p. 941, 2020.

[169] J. Yang, Y. Zheng, X. Gou, K. Pu, Z. Chen, Q. Guo, R. Ji, H. Wang, Y. Wang, and Y. Zhou, "Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis," *International Journal of Infectious Diseases*, vol. 94, pp. 91–95, 2020.

[170] Imlab-UIIP, "Lung Segmentation (2D)," https://github.com/imlab-uiip/lung-segmentation-2dR, online; accessed 19 October 2020.

[171] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public Chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, no. 6, p. 475, 2014.

[172] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for Chestradiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, 2000.

[173] N. G. Burnet *et al.*, "Defining the tumour and target volumes for radiotherapy," *Cancer Imaging*, vol. 4, no. 2, p. 153, 2004.

[174] S. Ramella *et al.*, "Local control and toxicity of adaptive radiotherapy using weekly CT imaging: results from the LARTIA trial in stage III NSCLC," *Journal of Thoracic Oncology*, vol. 12, no. 7, pp. 1122–1130, 2017.

[175] T. Schimek-Jasch, E. G. Troost, G. Rücker, V. Prokic, M. Avlar, V. Duncker-Rohr, M. Mix, C. Doll, A.-L. Grosu, and U. Nestle, "A teaching intervention in a contouring dummy run improved target volume delineation in locally advanced non-small cell lung cancer," *Strahlentherapie und Onkologie*, vol. 191, no. 6, pp. 525–533, 2015.

[176] P. Giraud, S. Elles, S. Helfre, Y. De Rycke, V. Servois, M.-F. Carette, C. Alzieu, P.-Y. Bondiau, B. Dubray, E. Touboul *et al.*, "Conformal radiotherapy for lung cancer: different delineation of the gross tumor volume (GTV) by radiologists and radiation oncologists," *Radiotherapy and oncology*, vol. 62, no. 1, pp. 27–36, 2002.

[177] U. Nestle, D. De Ruysscher, U. Ricardi, X. Geets, J. Belderbos, C. Pöttgen, R. Dziadiuszko, S. Peeters, Y. Lievens, C. Hurkmans *et al.*, "ESTRO ACROP guidelines for target volume definition in the treatment of locally advanced non-small cell lung cancer," *Radiotherapy and oncology*, vol. 127, no. 1, pp. 1–5, 2018.

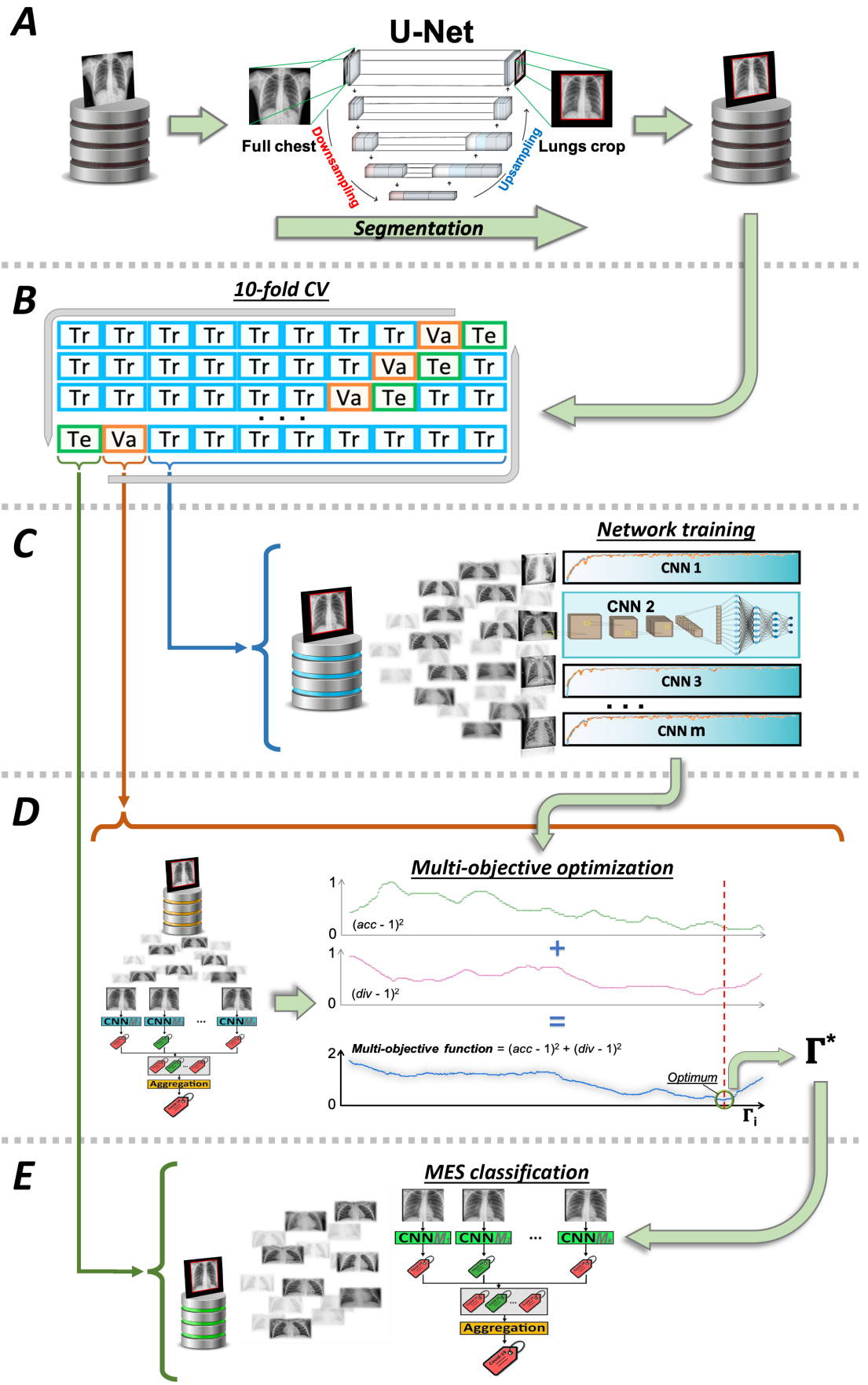
[178] A. Zwanenburg, S. Leger, M. Vallières, and S. Löck, "Image biomarker standardisation initiative," *arXiv preprint arXiv:1612.07003*, 2016.

Appendix A

Publications during PhD

(1) Guarrasi, V., D'Amico, N.C., Sicilia, R., Cordelli, E., Soda, P., Pareto optimization of deep networks for COVID-19 diagnosis from chest X-rays, Pattern Recognition, 2022, 121, 108242

Abstract:
The year 2020 was characterized by the COVID-19 pandemic that has caused, by the end of March 2021, more than 2.5 million deaths worldwide. Since the beginning, besides the laboratory test, used as the gold standard, many applications have been applying deep learning algorithms to chest X-ray images to recognize COVID-19 infected patients. In this context, we found out that convolutional neural networks perform well on a single dataset but struggle to generalize to other data sources. To overcome this limitation, we propose a late fusion approach where we combine the outputs of several state-of-the-art CNNs, introducing a novel method that allows us to construct an optimum ensemble determining which and how many base learners should be aggregated. This choice is driven by a two-objective function that maximizes, on a validation set, the accuracy and the diversity of the ensemble itself. A wide set of experiments on several publicly available datasets, accounting for more than 92,000 images, shows that the proposed approach provides average recognition rates up to 93.54% when tested on external datasets.

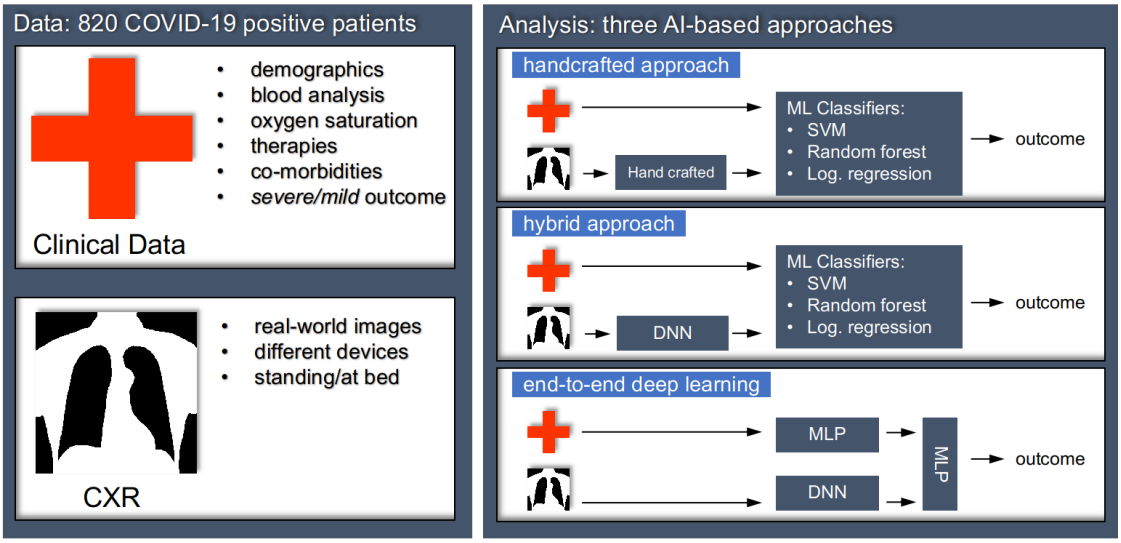


Appendix Figure A1: Pipeline followed for the classification tasks

(2) Soda, P., D'Amico, N.C., Tessadori, J., et.al., AIforCOVID: Predicting the clinical outcomes in patients with COVID-19 applying AI to chest-X-rays. An Italian multicentre study, Medical Image Analysis, 2021, 74, 102216

Abstract:

Recent epidemiological data report that worldwide more than 53 million people have been infected by SARS-CoV-2, resulting in 1.3 million deaths. The disease has been spreading very rapidly and few months after the identification of the first infected, shortage of hospital resources quickly became a problem. In this work we investigate whether artificial intelligence working with chest X-ray (CXR) scans and clinical data can be used as a possible tool for the early identification of patients at risk of severe outcome, like intensive care or death. Indeed, further to induce lower radiation dose than computed tomography (CT), CXR is a simpler and faster radiological technique, being also more widespread. In this respect, we present three approaches that use features extracted from CXR images, either handcrafted or automatically learnt by convolutional neuronal networks, which are then integrated with the clinical data. As a further contribution, this work introduces a repository that collects data from 820 patients enrolled in six Italian hospitals in spring 2020 during the first COVID-19 emergency. The dataset includes CXR images, several clinical attributes and clinical outcomes. Exhaustive evaluation shows promising performance both in 10-fold and leave-one-centre-out cross-validation, suggesting that clinical data and images have the potential to provide useful information for the management of patients and hospital resources.



Appendix Figure A2: Overview of the three developed AI-based-methods

(3) Guarrasi, V., D'Amico, N.C., Sicilia, R., Cordelli, E., Soda, P., A multi-expert system to detect COVID-19 Cases in X-ray images, Proceedings - IEEE Symposium on Computer-Based Medical Systems, 2021, 2021-June, pp. 395–400, 9474724

Abstract:

The year 2020 was marked by the worldwide COVID-19 pandemic, which caused over 2.5 million deaths by the end of February 2021. Different methods have been established since the beginning to identify infected patients and restrict the spread of the virus. In addition to laboratory analysis, used as the gold standard, several applications have been developed to apply deep learning algorithms to chest X-ray (CXR) images to diagnose patients affected by COVID-19. The literature shows that convolutional neural networks (CNNs) perform well on a single image dataset, but fail to generalize to other sources of data. To overcome this limitation, we present a late fusion approach in which multiple CNNs collaborate to diagnose the CXR scan of a patient, improving the generalizability. Experiments on three datasets publicly available show that the ensemble of CNNs outperforms stand-alone networks, achieving promising performance not only in cross-validation, but also when external validation is used, with an average accuracy of 95.18%.

Appendix Table A1: Accuracy of Experiments

(a) #1 (AIforCOVID ∪ *under*-RSNA)

	10 fold CV			*External Validation*	
	Global	non-COVID-19	COVID-19	RSNA[c]	COVIDX+
DenseNet121	97.27	98.007	96.39	**97.00**	49.43
DenseNet161	96.72	96.00	97.59	95.71	62.07
DenseNet169	95.08	96.00	93.98	95.22	60.45
GoogleNet	96.17	99.00	92.77	95.14	47.16
MobileNetV2	95.63	95.00	96.39	96.03	52.35
ResNet101	96.17	97.00	95.18	95.22	**63.37**
ResNet152	96.17	96.00	96.39	96.03	59.00
ResNet18	95.08	96.00	93.98	96.03	54.78
ResNet34	95.63	96.00	95.18	95.47	48.14
ResNet50	97.27	99.00	95.18	96.43	58.67
ResNeXt50(32x4d)	**98.36**	98.00	98.80	94.57	63.05
SqueezeNet1(1)	89.62	91.00	87.95	94.25	42.63
VGG11	93.99	95.00	92.77	94.73	45.38
WideResNet50(2)	96.17	98.00	93.98	96.03	61.26
MES (k=3)	98.90	99.00	98.78	98.51	**83.18**
MES (k=5)	98.90	99.00	98.78	**98.59**	79.74
MES (k=7)	**99.45**	100.00	98.78	98.26	77.02
MES (k=9)	**99.45**	100.00	98.78	98.02	73.78

(b) #2 (COVIDX+ ∪ *under*-COVIDX−)

	10 fold CV			*External Validation*	
	Global	non-COVID-19	COVID-19	RSNA[c]	AIforCOVID
DenseNet121	91.85	92.68	90.16	93.60	83.55
DenseNet161	92.93	93.50	91.80	**94.10**	81.51
DenseNet169	93.48	94.31	91.80	92.20	**84.75**
GoogleNet	92.93	93.50	91.80	92.40	75.03
MobileNetV2	92.93	95.12	88.52	93.70	77.19
ResNet101	90.76	94.31	83.61	94.20	74.91
ResNet152	90.22	93.50	83.61	94.00	79.35
ResNet18	92.93	95.12	88.52	93.60	72.51
ResNet34	**94.02**	95.94	90.16	93.60	78.15
ResNet50	92.93	95.12	88.52	93.80	79.23
ResNeXt50(32x4d)	91.85	93.50	88.52	93.80	80.43
SqueezeNet1(1)	92.93	92.68	93.44	91.50	78.15
VGG11	92.39	91.87	93.44	92.80	71.31
WideResNet50(2)	91.85	92.68	90.16	92.40	80.43
MES (k=3)	97.73	94.83	99.15	**99.40**	**91.60**
MES (k=5)	**98.86**	98.28	99.15	99.20	87.35
MES (k=7)	98.31	99.15	96.61	**99.40**	88.01
MES (k=9)	94.02	95.12	91.80	99.10	86.53

(4) Alì, M., D'Amico, N.C., Interlenghi, M., et.al., A decision support system based on BI-RADS and radiomic classifiers to reduce false positive breast calcifications at digital breast tomosynthesis: A preliminary study, Applied Sciences (Switzerland), 2021, 11(6), 2503

Abstract:

Digital breast tomosynthesis (DBT) studies were introduced as a successful help for the detection of calcification, which can be a primary sign of cancer. Expert radiologists are able to detect suspicious calcifications in DBT, but a high number of calcifications with non-malignant diagnosis at biopsy have been reported (false positives, FP). In this study, a radiomic approach was developed and applied on DBT images with the aim to reduce the number of benign calcifications addressed to biopsy and to give the radiologists a helpful decision support system during their diagnostic activity. This allows personalizing patient management on the basis of personalized risk. For this purpose, 49 patients showing microcalcifications on DBT images were retrospectively included, classified by BI-RADS (Breast Imaging-Reporting and Data System) and analyzed. After segmentation of microcalcifications from DBT images, radiomic features were extracted. Features were then selected with respect to their stability within different segmentations and their repeatability in test–retest studies. Stable radiomic features were used to train, validate and test (nested 10-fold cross-validation) a preliminary machine learning radiomic classifier that, combined with BI-RADS classification, allowed a reduction in FP of a factor of 2 and an improvement in positive predictive value of 50%.

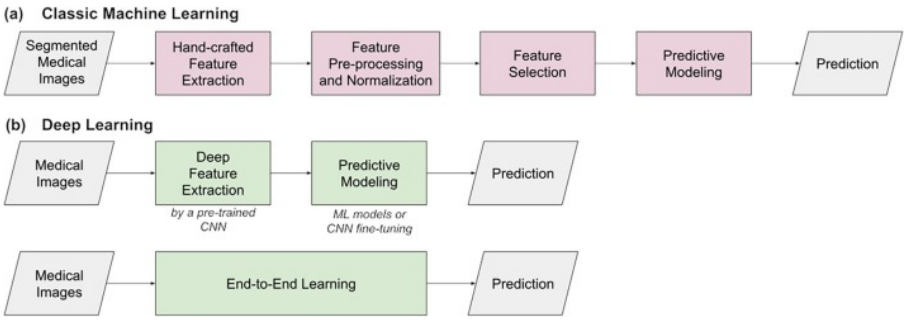
Sensitivity	Specificity	PPV	NPV	Accuracy	AUC
0.78 [0.52–0.94]	0.85 [0.68–0.95]	0.74 [0.49–0.91]	0.88 [0.71–0.96]	0.82*[0.69–0.92]	0.80**[0.67–0.90]

Appendix Table A2: Performances obtained in testing the best ensemble of machine learning radiomic classifiers for benign calcifications versus malignant calcification. Performances are reported as sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and Area Under the Curve (AUC), (95% Confidence Interval), * = p-value < 0.05, ** = p-value > 0.005.

(5) Castiglioni, I., Rundo, L., Codari, M., ...D'Amico, N.C., Sardanelli, F., AI applications to medical images: From machine learning to deep learning, Physica Medica, 2021, 83, pp. 9–24

Abstract:

Purpose: Artificial intelligence (AI) models are playing an increasing role in biomedical research and healthcare services. This review focuses on challenges points to be clarified about how to develop AI applications as clinical decision support systems in the real-world context. Methods: A narrative review has been performed including a critical assessment of articles published between 1989 and 2021 that guided challenging sections. Results: We first illustrate the architectural characteristics of machine learning (ML)/radiomics and deep learning (DL) approaches. For ML/radiomics, the phases of feature selection and of training, validation, and testing are described. DL models are presented as multi-layered artificial/convolutional neural networks, allowing us to directly process images. The data curation section includes technical steps such as image labelling, image annotation (with segmentation as a crucial step in radiomics), data harmonization (enabling compensation for differences in imaging protocols that typically generate noise in non-AI imaging studies) and federated learning. Thereafter, we dedicate specific sections to: sample size calculation, considering multiple testing in AI approaches; procedures for data augmentation to work with limited and unbalanced datasets; and the interpretability of AI models (the so-called black box issue). Pros and cons for choosing ML versus DL to implement AI applications to medical imaging are finally presented in a synoptic way. Conclusions: Biomedicine and healthcare systems are one of the most important fields for AI applications and medical imaging is probably the most suitable and promising domain. Clarification of specific challenging points facilitates the development of such systems and their translation to clinical practice.

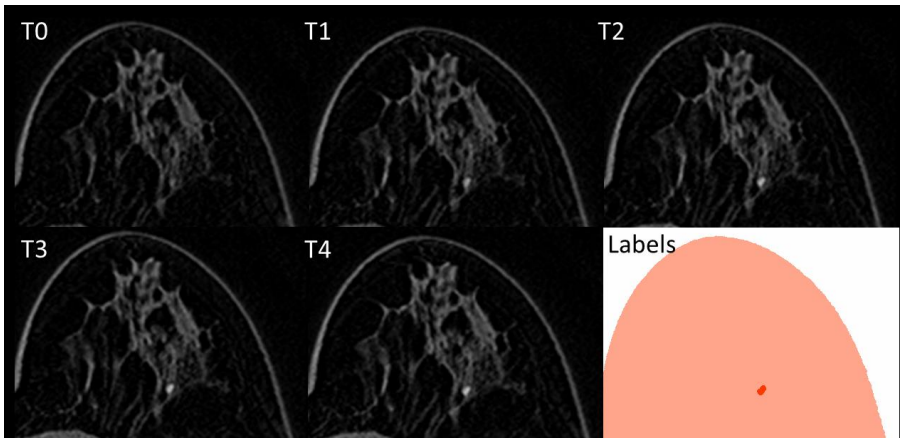


Appendix Figure A3: Typical architecture and workflow of artificial intelligence systems for predictive modelling: a) classic machine learning, with the various processing steps involving hand-crafted features such as in radiomics; b) deep learning considering either deep medical image feature extraction or end-to-end learning.

(6) D'Amico, N.C., Grossi, E., Valbusa, G., et.al., A machine learning approach for differentiating malignant from benign enhancing foci on breast MRI, European Radiology Experimental, 2020, 4(1), 5

Abstract:

Background: Differentiate malignant from benign enhancing foci on breast magnetic resonance imaging (MRI) through radiomic signature. Methods: Forty-five enhancing foci in 45 patients were included in this retrospective study, with needle biopsy or imaging follow-up serving as a reference standard. There were 12 malignant and 33 benign lesions. Eight benign lesions confirmed by over 5-year negative follow-up and 15 malignant histopathologically confirmed lesions were added to the dataset to provide reference cases to the machine learning analysis. All MRI examinations were performed with a 1.5-T scanner. One three-dimensional T1-weighted unenhanced sequence was acquired, followed by four dynamic sequences after intravenous injection of 0.1 mmol/kg of gadobenate dimeglumine. Enhancing foci were segmented by an expert breast radiologist, over 200 radiomic features were extracted, and an evolutionary machine learning method ("training with input selection and testing") was applied. For each classifier, sensitivity, specificity and accuracy were calculated as point estimates and 95% confidence intervals (CIs). Results: A k-nearest neighbour classifier based on 35 selected features was identified as the best performing machine learning approach. Considering both the 45 enhancing foci and the 23 additional cases, this classifier showed a sensitivity of 27/27 (100%, 95% CI 87–100%), a specificity of 37/41 (90%, 95% CI 77–97%), and an accuracy of 64/68 (94%, 95% CI 86–98%). Conclusion: This preliminary study showed the feasibility of a radiomic approach for the characterisation of enhancing foci on breast MRI.

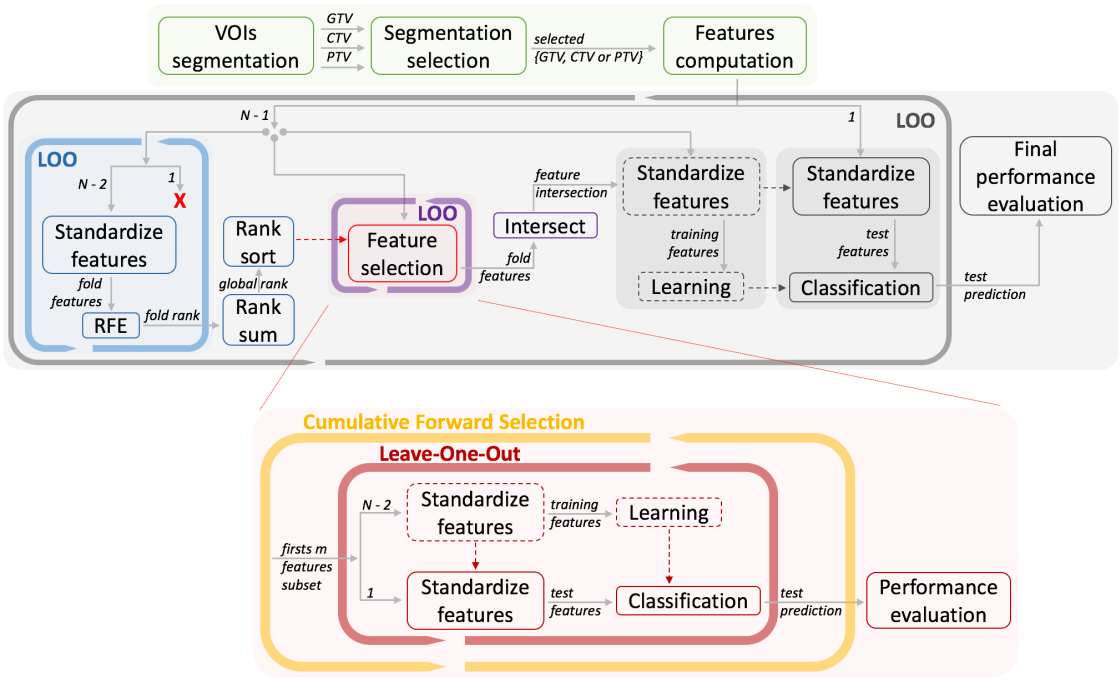


Appendix Figure A4: Breast magnetic resonance imaging showing in T0 the first (unenhanced) image and from T1 to T4 the contrast-enhanced images, where the wash-in and wash-out phenomena give information about the malignant or benign nature of the lesion. In the last image ("labels"), the segmented focus is coloured in red while normal breast tissues are coloured in pink

(7) D'Amico, N.C., Sicilia, R., Cordelli, E., et.al., Radiomics-based prediction of overall survival in lung cancer using different volumes-of-interest, Applied Sciences, 2020, 10(18), 6425

Abstract:

Lung cancer accounts for the largest amount of deaths worldwide with respect to the other oncological pathologies. To guarantee the most effective cure to patients for such aggressive tumours, radiomics is increasing as a novel and promising research field that aims at extracting knowledge from data in terms of quantitative measures that are computed from diagnostic images, with prognostic and predictive ends. This knowledge could be used to optimize current treatments and to maximize their efficacy. To this end, we hereby study the use of such quantitative biomarkers computed from CT images of patients affected by Non-Small Cell Lung Cancer to predict Overall Survival. The main contributions of this work are two: first, we consider different volumes of interest for the same patient to find out whether the volume surrounding the visible lesions can provide useful information; second, we introduce 3D Local Binary Patterns, which are texture measures scarcely explored in radiomics. As further validation, we show that the proposed signature outperforms not only the features automatically computed by a deep learning-based approach, but also another signature at the state-of-the-art using other handcrafted features.

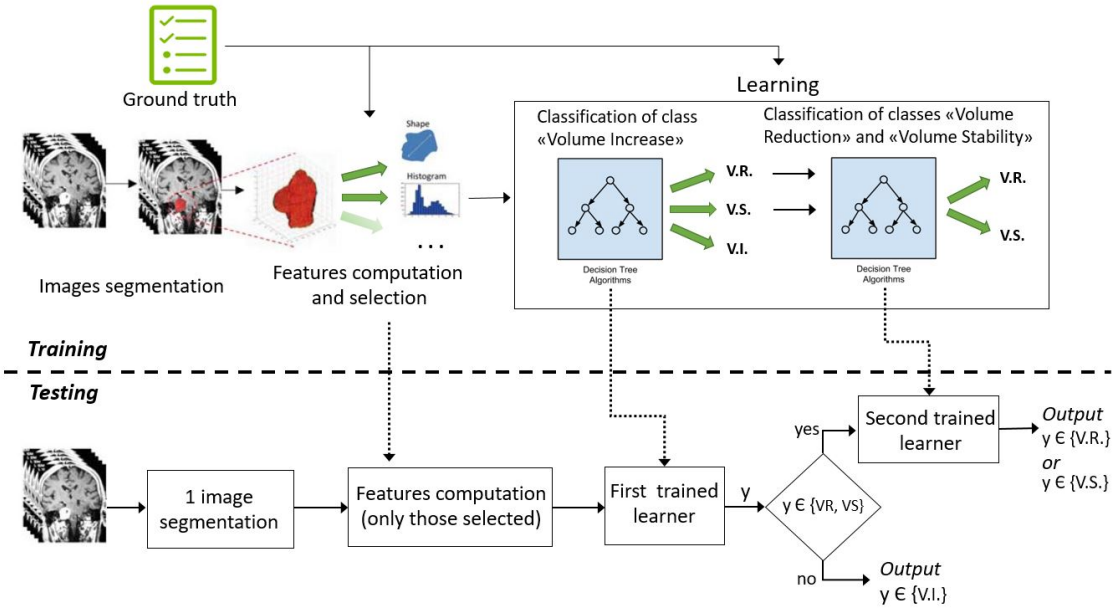


Appendix Figure A5: Overview of the proposed pipeline

(8) D'Amico, N.C., Merone, M., Sicilia, R., et.al., Tackling imbalance radiomics in acoustic neuroma, International Journal of Data Mining and Bioinformatics, 2019, 22(4), pp. 365–388

Abstract:

Acoustic neuroma is a primary intracranial tumour of the myelin-forming cells of the 8th cranial nerve. Although it is a slow growing benign tumour, symptoms in the advanced phase can be serious. Hence, controlling tumour growth is essential and stereotactic radiosurgery, which can be performed with the CyberKnife robotic device, has proven effective for managing this disease. However, this approach may have side effects and a follow-up is necessary to assess its efficacy. To optimise the administration of this treatment, in this work we present a machine learning-based radiomics approach that first computes quantitative biomarkers from MR images routinely collected before the CyberKnife treatment and then predicts the treatment response. To tackle the challenge of class imbalance observed in the available dataset we present a cascade of cost-sensitive decision trees. We also experimentally compare the proposed approach with several approaches suited for learning under class skew. The results achieved demonstrate that radiomics has a great potential in predicting patients response to radiosurgery prior to the treatment that, in turns, can reflect into great advantages in therapy planning, sparing radiation toxicity and surgery when unnecessary.



Appendix Figure A6: Overview of the proposed pipeline