POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Distant supervised learning for cancer subtyping with multiphase imaging data integration

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Author: **Matteo Stefano Savino**

Student ID: 953406
Advisor: Prof. Francesca Ieva
Co-advisors: Dr. Lara Cavinato
Academic Year: 2020-21

Abstract

Radiomics is a field of image analysis that consists in the high throughput extraction of quantitative features from medical images. It has become more and more important thanks to its advantage to non-invasively give access to tumor characterization. Thanks to the fact that the CT scans are composed by three images (phases) and that from an image it is possible to segment more than one region of interest (ROI), it is possible to obtain a multi-view or multiphase dataset. In particular, we have analysed two different types of multi-view dataset of patients affected by Intrahepatic cholangiocarcinoma (ICC). In the first one, also called multiphase, each view was composed by a phase while in the second one two views were composed by two ROIs: the core and the margin of the tumor. Both these regions were extracted from the Portal phase of the CT scans. We have employed these datasets to spot and assess the potential relevance of radiomic data trough different analyses. First of all, we worked in a classic framework for radiomic data performing some supervised analyses. Then, we have employed a distant supervision model, borrowed from the genomic literature, to perform cancer subtyping in Intrahepatic Cholangiocarcinoma patients. In this setting we have studied both the overall survival and the diseases free survival time of the patients affected by ICC. Our purposes were multiple, both methodological and clinical. The methodological ones were assessing whether the different phases and ROIs provide complementary information to create a good stratification for patients affected by ICC. Instead, the clinical purposes regard finding patients that share both a similar prognosis and imaging characterization. Moreover, we wanted to find some risk factors that are linked with the imaging variables and therefore, we were interested in obtaining easily interpretable results. Finally, we also wanted to establish whether the multi-view approach on radiomic data could be generalized also to other settings. Therefore, we have performed an analysis also on patients affected by Colorectal liver metastases to understand if, also in this setting, the addition of radiomic features is useful to improve the performances.

Keywords: Radiomics, CT scans, medical imaging, Intrahepatic cholangiocarcinoma, cancer subtyping, multi-view, multiphase, distant-supervised approach

Abstract in lingua italiana

La radiomica è una branca dell'imaging medico che consiste nell'estrarre variabili quantitative dalle immagini mediche e che sta diventando sempre più importante grazie alla sua capacità di rendere accessibile la caratterizzazione del tumore in modo non invasivo. Grazie al fatto che le TAC sono composte da tre immagini (chiamate fasi) e che da ognuna di esse è possibile segmentare più di una regione di interesse (ROI), è possibile ottenere un dataset multi-view o multifase. In particolare, abbiamo analizzato due tipi di dataset multi-view composti da pazienti affetti da Colangiocarcinoma intraepatico (ICC). Nel primo, chiamato anche multifase, ogni view è composta da una fase mentre nel secondo le due view sono formate da due ROI: il core e il margine del tumore. Entrambe queste regioni sono state segmentate dalla fase portale della TAC. Abbiamo utilizzato questi dataset per individuare e stabilire la potenziale utilità dei dati radiomici attraverso diverse analisi. Come prima cosa abbiamo adottato un framework classico per i dati radiomici facendo delle analisi di tipo supervisionato. In seguito, abbiamo utilizzato un approccio distant-supervised, che era stato proposto nella letteratura genomica, per fare sottotipazione del cancro nel caso di pazienti affetti da ICC. In questo caso abbiamo studiato sia la sopravvivenza che la recidiva dei pazienti. I nostri obbiettivi erano molteplici, sia di tipo metodologico che clinico. Dal punto di vista metodologico eravamo interessati a valutare se le diverse fasi della TAC e le diverse ROI apportano informazioni complementari per creare una buona stratificazione dei pazienti affetti da ICC. Dal punto di vista clinico eravamo interessati a trovare dei gruppi di pazienti che fossero caratterizzati da una simile prognosi e da caratteristiche dell'imaging simili. Inoltre, eravamo interessati ad individuare dei fattori di rischio legati all'imaging tramite dei risultati facili da interpretare. Infine eravamo anche interessati a valutare se questo approccio multi-view potesse essere generalizzato anche ad altri casi. Per questo abbiamo studiato se anche nel caso di pazienti affetti da metastasi epatiche derivanti da un tumore al colon-retto l'utilizzo di variabili radiomiche provenienti da diverse ROI porta a dei risultati migliori.

Parole chiave: Radiomica, TAC, immagini mediche, Colangiocarcinoma intraepatico, sottotipizzazione del cancro, multi-view, multifase, approcci distant-supervised

Contents

Abstract **i**

Abstract in lingua italiana **iii**

Contents **v**

Introduction **1**

1 General context **7**

1.1 Intrahepatic cholangiocarcinoma 7

1.2 Medical Imaging and Radiomics 8

1.3 Cancer subtyping literature 12

1.4 ICC Data 16

1.4.1 Radiomic Features 16

1.4.2 Clinical Variables 21

1.4.3 Data Preprocessing 24

2 Supervised analysis on ICC patients **27**

2.1 Classification Problem 27

2.1.1 Logistic Regression 28

2.1.2 Results 29

2.2 Survival Analysis 31

2.2.1 Cox Model 31

2.2.2 Results 32

2.3 Conclusions 33

3 Cancer subtyping methodological pipeline **35**

3.1 Supervised Survival Graph Clustering model 36

3.1.1 Model definition 38

3.1.2 Survival Analysis based loss term 38

3.1.3 Penalization terms 39

3.1.4 Adaptive graph estimation term 40

3.1.5 Hyperparametrs selection 41

3.1.6 Clustering technique 41

3.2 Core/Margin assessment in Portal CT imaging 41

3.2.1 Construction of the multi-view patient representation 42

3.2.2 Parameters Optimization 42

3.2.3 Spectral Clustering 49

3.2.4 Feature importance 49

3.2.5 Cluster characterization 50

3.3 View assessment in three-phases CT imaging: Overall survival 50

3.3.1 Construction of the multi-view patient representation 51

3.3.2 Parameters Optimization 52

3.3.3 Spectral Clustering 62

3.3.4 Feature importance 63

3.3.5 Cluster characterization 63

3.4 View assessment in three-phases CT imaging: Recurrence 64

3.4.1 Construction of the multi-view patient representation 64

3.4.2 Parameters Optimization 65

3.4.3 Spectral Clustering 75

3.4.4 Feature importance 75

3.4.5 Cluster characterization 76

3.5 General messages 76

4 Results and clinical findings 79

4.1 Core/Margin assessment in Portal CT imaging 80

4.1.1 Clinical Characterization 81

4.1.2 Imaging Characterization 85

4.2 View assessment in three-phases CT imaging: Overall survival 88

4.2.1 Clinical Characterization 89

4.2.2 Imaging Characterization 93

4.3 View assessment in three-phases CT imaging: Recurrence 100

4.3.1 Clinical Characterization 100

4.3.2 Imaging Characterization 105

4.4 General messages 111

5 Discussion 113

5.1 Methodological discussion 113

5.2 Clinical discussion . . . 116
5.2.1 Overall survival . . . 116
5.2.2 Recurrence . . . 119
5.2.3 General messages . . . 120

6 The colorectal liver metastases case study **123**
6.1 Data . . . 124
6.2 Methods . . . 125
6.3 Results . . . 126
6.4 Conclusions . . . 129

7 Conclusions and future developments **131**

Bibliography **133**

List of Figures **141**

List of Tables **147**

Acknowledgements **149**

Introduction

Intrahepatic cholangiocarcinoma (ICC) is an aggressive disease of the family of cholangiocarcinomas [1]. Cholangiocarcinoma (CCA) is a heterogeneous group of neoplasm that can emerge at every point of the biliary tree, from the canals of Hering to the main bile duct. This family of malignancies can be classified in three different groups according to their anatomical location [2]. These groups are: intrahepatic (ICC), perihilar (PCC) and distal CCA (DCC). In Figure 1 it is possible to see the liver divided in the three regions that correspond to the three different class of cholangiocarcinomas. The whole family of cholangiocarcinomas represents the 3% of all the gastrointestinal tumors and it is the second most frequent tumor in the liver [3]. The intrahepatic cholangiocarcinoma represents the 5-15 % of all the cholangiocarcinomas and its incidence is sensibly increasing in the last two decades [4]. For this reason it has become a significant global concern, especially since the increase of incidence occurred in the past decades has implied also an increase in mortality rates [5]. The highest incidence of ICC occurs in Southeast Asia, particularly in Thailand, Laos, Cambodia, and Vietnam [6]. Intrahepatic cholangiocarcinoma is usually diagnosed at an advanced stage, with a poor prognosis and a short survival time. An aspect that leads to a late diagnosis is that this tumor occurs sporadically in patients without recognizable risk factors [7]. The possibility of early detection of ICC would be desirable and have potentially an important impact on the public health in the resource-poor regions where this cancer is most prevalent [6]. Currently, surgical resection is considered a central component of potentially curative treatment for ICC [8]. However, in large part due to a high rate of tumor recurrence, survival remains poor even among resected patients. Indeed, survival rates at 5 years are estimated to be $< 25\%$ for localized disease, $< 10\%$ for regional disease, and $< 5\%$ for distantly metastatic disease [8].

In general, the cancer is a big family of potentially lethal diseases characterized by a big heterogeneity. This is not only between different cancers but also within the same cancer and for this reason can be really important to find sub-types. Once the sub-types have been found it is possible to perform a more accurate prognosis and to study a more personalized treatment. These are the motivations that induced the numerous works that exist in the literature on the cancer subtyping. Cholangiocarcinomas is not different from the

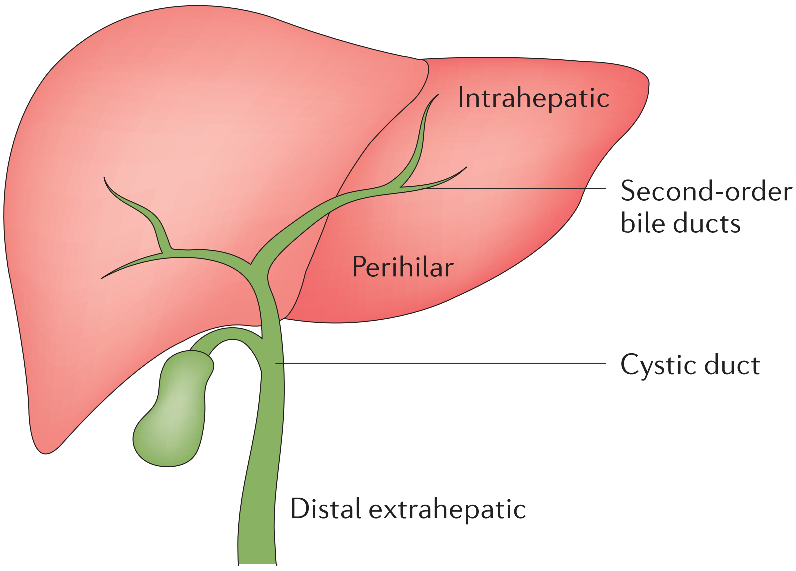


Figure 1: Cholangiocarcinomas are classified according to the anatomical location into intrahepatic (ICC), perihilar (PCC) and distal (DCC) (figure taken from [9]).

others cancers and present a big heterogeneity. Indeed, CCAs comprise a group of cancers with different locations and pronounced inter-tumoural and intra-tumoural heterogeneity [9]. As seen, it is divided in three different groups depending on their anatomical location. This division is important because the three classes have some different characteristic and they are characterized by some difference in prognosis. Unluckily, this separation is not sufficient and also the intrahepatic cholangiocarcinoma is characterized by patients with different prognosis. In particular, ICC shows a inter-tumor heterogeneity, leading to the classification into two main different histological subtypes [10]. In support to the fact that the intrahepatic cholangiocarcinoma has a big heterogeneity it is possible also to notice that it can present three different patterns of growth: mass-forming, periductal infiltrating and intraductal growing [9]. Therefore, due to its increasing incidence and mortality over the past two decades, is now more than ever arising the urgency to further characterize the disease at early stages. In fact, we have seen that this disease is often diagnosed at an advanced stage but the possibility of detecting baseline information can change this. An early stage diagnosis can be crucial in order to design more efficient lines of treatments. Furthermore, since surgical resection is at the moment the most applied curative treatment for patients affected by ICC, is crucial to detect this disease when it is still operable. For these reasons we are interested in applying a cancer subtyping approach to find patients that share common characteristics and in particular a similar prognosis. The majority of the methods proposed to perform cancer subtyping are in the genomic literature. This kind of data requires some invasive analysis in order to be collected. In this thesis we have worked, instead, with radiomic data. Radiomics is a new field that non-invasively provides rich information on diseases by quantitatively analyzing a large number of features extracted from traditional medical images [11]. This technique, which is independent of the subjective visual interpretations of the radiologist, objectively quantifies the heterogeneity of lesions [12][11]. Therefore, radiomics has become a prominent component of medical imaging research and many studies show that can be a very powerful tool for clinical decision-making processes. Indeed, it has been successfully used to improve the diagnosis and risk stratification of several types of cancer, such as gliomas [13], lung cancer [14], breast cancers [15], and rectal cancer [16].

In our setting we work with a multi-view radiomic dataset. With radiomic multi-view dataset we intend a dataset that is composed of radiomic features extracted from different images of the patients, or from different regions of interest of the same image. This can be obtained thanks to the fact that the CT scans produce three different images and from each image it is possible to segment more than one region of interest (ROI).
The main interpretation of a multi-view dataset is the one where the three phases, corresponding to the Arterial, Portal and Late phases of the CT scans, are employed to build three different views. Each phase can be seen as a view since they are different representations of the same patient. What is different from a classic multi-view dataset is that in this case the phases composing the views are of different temporal phases and not different views of the same moment. For this reason this dataset can be also called multiphase. Our interest is in assessing if the three phases of the CT scans are all important or if the Portal one, which is the main one, is sufficient.
The other interpretation of a multi-view dataset relies on different regions of interest of an image composing the views, as it will be explained in Section 3.2.1. In our case, the ROIs correspond to the core and the margin of the tumor, which are, respectively, the tumor region and a peritumoral liver tissue of 5-mm. In this setting we are interested in further investigating the core-margin interface, as suggested in [17]. Indeed, the authors of this work suggest to study a wider area of the tumor and not only the core region since also the tumor-tissue interface can provide important information.

Purposes and outline of our study

Within this context lies this thesis, whose purposes were multiple, both methodological and clinical.
The methodological ones are: assessing whether it can be useful to employ all three CT phases to classify and to create a good stratification for patients affected by ICC. We were

also interested in a more specific analysis on the core-margin interface to assess whether the information provided by these two regions is complementary. Finally we wanted to establish whether the multi-view approach on radiomic data can be generalized also to other settings.
Instead, the clinical purposes regard finding patients that share both a similar prognosis and imaging characterization. Moreover, we wanted to find some risk factors that are linked with the imaging variables and therefore, we were interested in obtaining easily interpretable results.

As first, to establish whether the three phases of the CT scans provide different information, we have performed some classical analysis through supervised methods. As it will be explained in detail in the followings chapters of this work, the supervised approaches have some limitations in the context of radiomic data and they were not sufficient to tackle all our purposes. For this reason after these studies we have employed an algorithm that performs cancer subtyping using multi-view data. Then, once ended the analysis on the ICC patients, we have performed a different analysis on patients affected by Colorectal liver metastases to establish whether the possibility of using a multi-view radiomic dataset can be generalized also to other contexts.

More in detail our work is structured in the following way:

- Chapter 1: General context
  This chapter deepens the content of the current introduction. In particular, some important knowledge useful to have a better comprehension of our work are described together with a presentation of the dataset employed in our analysis. At first, the intrahepatic cholangiocarcinoma with some known risk factors and the typical therapies are presented. Then, medical imaging, its usefulness, and the concept of radiomics are introduced. To end the part of contextualization an overview of the cancer subtyping literature is proposed. Finally the study case data are addressed in detail. A brief explanation of all the, radiomic and clinical, variables available in the dataset is proposed. We also have reported two tables reporting some basic statistics for the clinical variables and all the preprocessing applied before employing the data for our analysis;

- Chapter 2: Supervised analysis on ICC patients
  This chapter regards the supervised analysis that we have performed as first attempt to analyse the information included in the multiphase dataset described in the previous chapter. In particular, the methods used to perform a classification of the survival and to perform a survival analysis are presented together with their

respective results. In the conclusions of the chapter, considerations regarding the results obtained and the main disadvantages of these approaches are pointed out.

- Chapter 3: Cancer subtyping methodological pipeline
  In Chapter 3, a semi-unsupervised approach to analyse the multi-view aspect of our dataset is proposed. This method has the potentiality of overcoming the main problems that the complete supervised models had. In particular, it allowed us to tackle all our purposes, both the both methodological and the clinical ones.
  More in detail, this method is a distant supervised approach called Supervised Survival Graph Clustering (S2GC) model that has been proposed in [18] to perform cancer subtyping. Originally this model has been built to work with genomic dataset but we were interested in following the works done in [19] where for the first time have been employed with the radiomic data.
  After the presentation of the model we, also, report how it has been applied to the data of the patients affected by ICC. In particular, three different applications have been studied in order to tackle all our purposes. The first one have been employed to perform the analysis on the core-margin interface. The second one, that is the main one, has the goal of assess whether the three phases of the CT scans provide different and complementary information for the stratification of the patients affected by ICC when the overall survival time is studied. In last application have been, instead, analysed the disease free survival time, also in this case employing all the three phases. For each application we have reported: how the multi-view representation of the patient have been intend in that setting, the tuning of the parameters of the S2GC model and the techniques that we have employed to present the results in the following chapter;

- Chapter 4: Results and clinical findings
  In this chapter are presented all the results of the three applications of the S2GC model to the ICC data explained in Chapter 3. For these applications we have reported the groups found and their clinical and radiomic characterizations;

- Chapter 5: Discussion
  In this chapter the results of all the analysis performed on the ICC data have been analysed. We have discussed their implications from both a methodological and a clinical point of view. Thanks to this discussion we have given an answer to our research questions;

- Chapter 6: The colorectal liver metastases case study
  In this chapter a different application of our framework is discussed. In particular, we

wanted to study whether what found about the addition of more radiomic variables was still valid also in a different setting. The setting was different since both the disease and the clinical question were different. Indeed, for this application we were studying patients affected by colorectal liver metastases and we would like to study the answer of each lesion to the chemotherapy. In this case we have that each sample of the dataset was a lesion, instead of a patient, as it was for the case of the ICC dataset. For this reason we have studied the answer of each lesion and not of the patient;

- Chapter 7: Conclusions and future developments
  With this chapter we sum up our results and discuss possible future developments.

1 | General context

As previously anticipated, the purpose of this chapter is to deepen the content of the introduction. In particular, at first some important notions to have a better comprehension of our work are explained and then the data employed in our study are presented. Indeed, some additional information on the intrahepatic cholangiocarcinoma (ICC) is reported in Section 1.1. In particular, the known risk factor and the typical therapy for patients affected by this disease were the focus of this section. Then, a small overview of the imaging techniques used in medicine is proposed together with an explanation of the radiomics in Section 1.2. Instead, Section 1.3 concerns the cancer subtyping literature where some algorithm proposed in this research field are reported. The last section of this chapter (section 1.4) regards our data of the patients affected by ICC. In particular, we have reported an explanation of the variables available in the dataset together with some descriptive tables and the preprocessing applied.

1.1. Intrahepatic cholangiocarcinoma

Intrahepatic cholangiocarcinoma (ICC) is an aggressive neoplasm that comes from the epithelium of the intrahepatic bile ducts [20]. Thanks to Figure 1 it is possible to see where the intrahepatic bile ducts are located in the liver. As already anticipated, the incidence of ICC has been increasing in the last decades and it represents the 5-15 % of all the cholangiocarcinomas [4]. Risk factors include chronic hepatitis and cirrhosis, biliary inflammatory diseases and hepatobiliary flukes, although in most cases, no known risk factor is identified [21]. ICC is usually asymptomatic and often diagnosed incidentally on imaging studies [22]. Therefore, it is often diagnosed at an advanced stage and for this reason many patients are not eligible for curative surgical resection [23]. The possible therapies applied to a patient affected by ICC are: surgical resection, chemotherapy and radiation therapy. Surgical resection is the only potentially curative treatment but it is associated with high tumor recurrence rates [24]. The choice of the treatment is determined by the patient's performance status, the local extent of the tumor (including vascular involvement) and the absence or presence of metastatic disease [22]. Curing

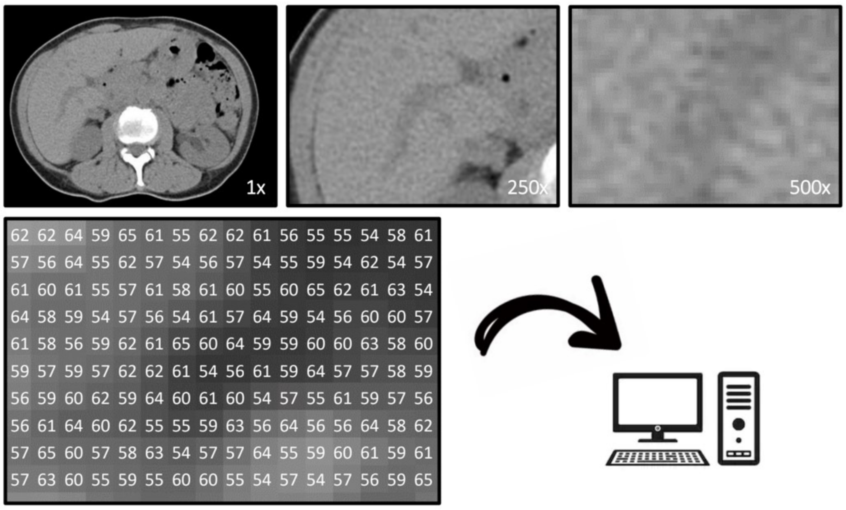
ICC requires a complete surgical resection with histologically negative margins (ie, R0 resection status) [22]. In fact, due to its high aggressiveness, long-term survival is only observed in patients with a complete R0 surgical resection [21]. Indeed, complete R0 resection is the major predictor of disease-free and overall survival after surgical resection for ICC [25]. In fact, some studies have suggested that outcomes in patients who undergo resection with positive margins may be no better than in patients treated non surgically [22]. Other features that are associated with a poor prognosis include factors connected to the extent of the disease, such as lymph node involvement, vascular invasion and distant metastases [25]. Since ICC is usually diagnosed in late stages, R1 resections are frequent even when surgery is performed with curative intent. Adjuvant therapy is not routinely used, although chemotherapy and/or radiotherapy can improve survival following R1 resections [21],[23]. The outcomes from surgical resection have improved in recent years compared with those reported in the past. Surgery with curative intent is associated with 5-year survival rates of up to 22–36% [25]. For patients with ICC, the presence of underlying diabetes mellitus, liver cirrhosis, primary sclerosing cholangitis or a history of smoking, alcohol use, or inflammatory bowel disease do not significantly alter median survival [23].

1.2. Medical Imaging and Radiomics

Medical imaging is one of the major factors that have informed medical science and treatment [26]. The first X-ray was taken by Wilhelm Conrad Roentgen in 1895 and from that moment imaging has become a fundamental part in medicine. Indeed, today medical imaging is an essential component of the entire health-care system, from screening to early diagnosis and treatment selection [27]. By assessing the characteristics of human tissue noninvasively, imaging is often used in clinical practice for oncological diagnosis and treatment guidance [26]. One important characteristic of the imaging is that it can provide a comprehensive view of the entire organ or tumor. This is already an advantage with respect to exams like biopsies or invasive surgeries that extract and analyse what are generally small portions of tissue. In this way they do not allow for a complete characterization of the whole mass, organ, or, in the case of oncological patients, tumour [26]. There are different kind of medical imagines such as: X-ray Computed Tomography (CT), Magnetic Resonance Imaging (MR) and Positron Emission Tomography (PET) [27]. One of the most widely used imaging modality in oncology is the X-ray Computed Tomography (CT), which assesses tissue density [26]. Computed Tomography is one of the most important medical innovations in human history. Images display soft tissue contrasted with anatomic detail, facilitating unprecedented diagnostic accuracy [27]. This is due to the fact that the absorption of the X-ray beam is dependent on the density. The absorption/attenuation coefficient of radiation within a tissue is used during CT reconstruction to produce a gray-scale image. A linear transformation of the linear attenuation coefficient of the X-ray beam produces a Hounsfield scale that displays as gray tones. The Hounsfield unit (HU) is a relative quantitative measurement of radio density used by radiologists in the interpretation of CT images [28]. More dense tissue has positive values while less dense tissue has negative values. For example the bones have 1000 Hu while the air -1000 Hu. The difference in the Hounsfield values of the different soft tissue is not big. For this reason in order to improve the contrast of the captured images, and therefore the accuracy of the diagnosis, contrast media injection is widely used [29]. The patient receives the injection, and 3 series are taken at three different times: the first one, just after the injection, is called the Arterial phase. The second, a few tens of seconds later, the Portal phase. The last one, a few minutes after the injection: the Late phase. In this way three different images of the same patient are available. The use of the contrast injection is used, as already said, to improve the contrast and because it is possible that a lesion indistinguishable from the healthy liver in one phase will be revealed in another phase [29].

Between the aspects that lead the images to be so important in the medical science the noninvasively is a crucial one. The possibility of performing a good diagnosis without the need of an invasive analysis is really important. The imaging has a great potential in this aspect. Further, imaging is already often repeated during treatment in routine practice, on the contrary of genomics, which are still challenging to implement into clinical routine [26]. However, in most cases, visual inspection of CT scans could not be sufficient for proper image interpretation. The definitive diagnosis often requires invasive procedures like biopsy or even surgery, which carry a risk of complications [30]. For these reasons new ways of extracting useful information from the images have been subject of a lot of study in the last years. Techniques that are able to extract information not normally detected by human eye could reduce or even eliminate the necessity of performing the invasive techniques. These techniques can also be very important for the personalized medicine, that is a research field based on efficiently extracting insights from multi-source patient data to shape clinical practice. From all these motivations in recent years has born a new area of research termed Radiomics.

Radiomics is a way to perform medical image analysis studying them voxel by voxel and so extracting microscopic information in a semi-automatic way. In Figure 1.1 is schematically represented this procedure. Starting from an image this is studied analysing the voxels reaching in this way a level of detail that is not possible with a simple enlargement and



62	62	64	59	65	61	55	62	62	61	56	55	55	54	58	61
57	56	64	55	62	57	54	56	57	54	55	59	54	62	54	57
61	60	61	55	57	61	58	61	60	55	60	65	62	61	63	54
64	58	59	54	57	56	54	61	57	64	59	54	56	60	60	57
61	58	56	59	62	61	65	60	64	59	59	60	60	63	58	60
59	57	59	57	62	62	61	54	56	61	59	64	57	57	58	59
56	59	60	62	59	64	60	61	60	54	57	55	61	59	57	56
56	61	64	60	62	55	55	59	63	56	64	56	56	64	58	62
57	65	60	57	58	63	54	57	57	64	55	59	60	61	59	61
57	63	60	55	59	55	60	60	55	54	57	54	57	56	59	65

Figure 1.1: Sketch of the procedure used to extract the radiomic features from the CT scans.

by visual inspection. The voxels are analysed thanks to different statistical procedure in order to extract useful information. Indeed, radiomics consists in high-throughput quantitative features extracted from regions of interest in medical images such as CT scans or MRI or PET. These features, also known as radiomic or texture features, can be many and are agnostic with respect to the clinical application. Texture analysis provides an objective, quantitative assessment of tumor heterogeneity by analyzing the distribution and relationship of voxel gray levels [31]. The radiomic features are not build in order to answer to a specific clinical problem, but to extract all the possible information that is entailed in medical images [11]. This information is transformed into matrix-shaped data, easier to handle and study. The philosophy under the radiomic procedure is to extract as much information as possible and after try to understand which are the variables that has the highest prognostic value. Radiomics try to answer to the need of having a standardized way to extract information from the medical images, allowing for a more deep study of them and not, as was traditionally performed, just a visual interpretation. Indeed it is near to be reader independent while the standard medical imaging relies on the experience of the reader. However, it is known to have some limitations, among all instability with respect to segmentation procedures and complexity in exhaustively shape the imaging representation of the lesions [32]. Another limitation is caused by the acquisition device and conditions. Despite these limitation this technique has become more and more important thanks to its advantage to non-invasively help to characterize the tumor heterogeneity, that is known to be a relevant factor in the tumor prognosis [33]. Furthermore, as already said, imaging is often repeated during treatment in routine practice and this leads to have a big quantity of images. Therefore, it is possible to create a big dataset that can be studied in order to perform a better characterization of the tumor. It is estimated that every year more or less 3.6 billions of X-Ray examinations are done worldwide and these are mostly evaluated once and stored. This enormous amount of images can be a source of a huge quantity of information that can be mined, but with the classical visual inspection this is not possible. Another important aspect is that in this way one can prevent patients with unclear results of a CT/PET/MRI exam from other X-ray or even worst some invasive diagnostic with the associated risks. In fact, extracting useful information that is not visible by visual inspection can help the doctor in his work without the need of more exams. Tumor heterogeneity characterization and consequently the tumor prognosis is not the only area where radiomics can be applied. There are several areas where the use of radiomics can be useful such as for treatment selection, the enabling diagnosis and the decision where to biopsy or resect.

Radiomic features are extracted thanks to different mathematical techniques that are used to analyse the grey-level intensity distribution and spatial organization of the image. These features can be divided in: shape variables, first order variables, second order variables and higher-order variables, with different levels of abstraction and consequently a more difficult interpretation. The shape features describe the shape of the selected ROI and its geometric properties such as volume, maximum surface and sphericity [31]. First order statistics features describe the distribution of individual voxel values neglecting the spatial relationships. These are generally based on histograms methods and represent the entire region of interest with just a number as for example the minimum, the maximum, some quantiles, the kurtosis or the skewness of the historgam of the values. The second order ones, instead, describe the statistical interrelationships between neighbouring voxels. They, indeed, provide a measure of the spatial arrangement of the voxel intensities, and hence of intra-lesion heterogeneity [31]. The higher order features are built thanks to filter grids or mathematical transformations that are applied to the image in order to extract patterns or highlighting details. These include fractal analysis, Minkowski functionals and Laplacian transforms of Gaussian-filtered images. Higher order features are more difficult to interpret, but they have the advantage of evaluating voxels in their local context, taking the relationship with neighboring voxel into account [31].

A very important step in the extraction of radiomic features is the identification of the volume of interest. In Figure 1.2 we can see an example of CT scan where a region of interest has been segmented. In our work we used two different regions of interest. In particular, for each lesion we have used both the features extracted from the core region of the tumor and the ones extracted from the margin. The margin was computed as the

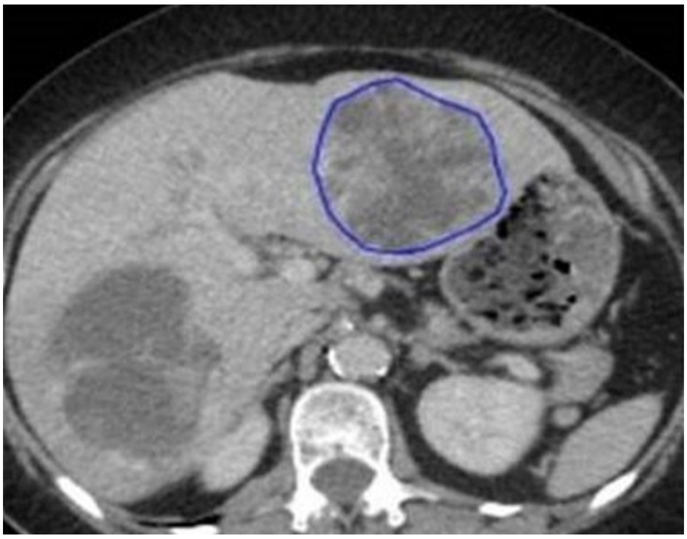


Figure 1.2: Example of a region of interest (ROI) segmented in a CT scan.

5-mm region around the tumor in order to include the peritumoral liver tissue. Recently it has been proposed to study both these two regions in order to capture also the information of the tumor-tissue interface [17]. We are interested in continuing this study and try to understand whether the two regions provide different information.

Radiomics' analysis usually include 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 [32]. An example of dimensionality reduction which is often used is principal component analysis (PCA). For what concerns the supervised analysis, the chosen algorithm depends on the application and can consist for example in random forests [34], logistic regression [35] or Cox proportional hazard regression [36]. In our analysis we have used this two-steps procedure as first attempt to establish whether the different phases add useful information. Then, we have employed a different kind of analysis. Instead of using a feature selection algorithm and a supervised one we have used only a distant supervised procedure. In this way we have obtained clusters of patients that share some imaging characteristic, but that are also grouped accordingly with their prognosis. A big difference is that with this approach we are not directly forcing the radiomic data to predict a label in a supervised way.

1.3. Cancer subtyping literature

Cancer subtyping consists in finding clusters of patients within a cancer type that have shared characteristics and a similar survival time. This is really important since cancer is a large family of lethal diseases characterized by a big heterogeneity both inter and intratumoral. Due to this heterogeneity it is really difficult to develop general and effective treatments against cancer [37]. Therefore, to be able to improve the prognosis and discover specific treatments, it is needed to build a good characterization also of the intra-turmoral heterogeneity. In the last years a research field, called personalized medicine, that tries to design efficient lines of treatments specific for the patient in analysis, has become more and more important. To be able to perform this treatment optimization it is important to be able to detect at baseline some information and for this reason cancer subtyping is currently a trending research topic. Therefore, in literature a lot of different methods that performs cancer subtyping have been proposed. In the following we present the main ones:

1. LRAcluster [37]: LRAcluster is a low-rank approximation based integrative probabilistic model. It is an unsupervised method to find the principal low-dimension subspace of large-scale and high-dimensional multi-omics data for molecular classification. In particular, it can deal with real-type data, binary data and count data. In case of only real-type data the LRScluster solution is the same as the PCA;

2. iCluster [38]: iCluster is one of the first proposal for cancer subtyping and uses a joint latent variable model. It is based on probabilistic principal component analysis, which used generalized linear models to transform continuous, discretized and count variables as a sparse linear regression on a set of latent driving factors. It uses a lasso regularization term to reduce the variance and improve the clustering performance;

3. SNF [39]: SNF consists of two main steps: construction of a sample-similarity network for each data type and integration of those networks into a single similarity network using a nonlinear combination method. An advantage of this integrative procedure is that weak similarities, that correspond to low edges in the graph, disappear, helping to reduce the noise. Once the final similarity network has been estimated the spectral clustering algorithm has been used to identify homogeneous cancer sub-types;

4. CIMLR [40]: CIMLR stays for Cancer Integration via Multikernel Learning. This method learns a measure of similarity between each pair of samples in a multi-omic dataset by combining multiple gaussian kernels per data type, corresponding to different, complementary representations of the data. It enforces a block structure in the resulting similarity matrix, which is then used for dimension reduction and k-means clustering. An important characteristic of this method is that it learns the weight for each data type and not assign equal importance to each of them;

5. Another method, proposed in [41], use multi-omic data to build a patient-to-patient

similarity graph for each data type. Then, it merge those intermediate representation through subspace analysis on a Grassmann manifold;

6. JIVE [42]: JIVE stays for Joint and Individual Variation Explained and is an extension of the Principal Component Analysis. This method, as expressed in the name, try to separate the joint and the individual effects. This exploratory method decomposes a dataset into a sum of three terms: a low-rank approximation capturing joint structure between data types, low-rank approximations capturing structure individual to each data type, and residual noise. The main difference between this approach and the iCluster one, both based on the PCA, is that this distinguish between join and individual effect, while the other no;

7. PINSPlus [43]: PINSPlus is an unsupervised approach that want to find the subtypes without using any prior knowledge. In order to find this sub-types and to make them reliable they study three different scenarios. In particular they estimate how often two patients are grouped together in the following scenarios: (1) when the data are perturbed, (2) when using different data type and (3) when using different clustering techniques. So the groups are formed by patients that are strongly connected in all the different scenarios;

8. MOFA [44]: Multi-Omics Factor Analysis is a method that can be viewed as a generalization of the principal component analysis (PCA) for the multi-omics data. Given several data matrices with measurements of multiple omics data types on the same or on partially overlapping sets of samples, MOFA infers an interpretable low-dimensional data representation in terms of (hidden) factors. These learnt factors capture major sources of variation across data modalities, thus facilitating the identification of continuous molecular gradients or discrete subgroups of samples. In comparison with iCluster is less computational demanding and it deals with missing values while iCluster doesn't;

9. In the last years also some model using deep learning techniques have been proposed. This has been done for the first time in [45] where there is proposed an algorithm that uses the autoencoder framework.

What characterize all these methods is that they are proposed in the genomic literature and are based on the study of genomic and biological data. These kind of data are often collected in multi-omic or multi-view dataset and for this reason different ways to integrate these views have been proposed. These data are, also, often characterized by huge dimensionality therefore many ways to reduce the dimension have been analysed. Another important characteristic of the majority of the methods proposed in the literature is that they are completely unsupervised. This means that they try to subdivide the patients in cluster that have similar median survival time, but they don't use any information on the survival of the patient in order to build that stratification. However, in the last years this has been overcome and some supervised method that exploit also the survival information of the patients has been proposed. In particular, it has been proposed a method called Survival Supervised Graph Clustering (S2GC [18]) that thanks to a distant supervision approach use both the information relative to the genomic data and the survival one. As we will explain in the Section 3.1, we want to borrow this method, that have been proposed in the genomic literature, as done in [19] to perform our studies on the radiomic data.

In order to have a more complete overview of the cancer subtyping, we present also some recent works proposed in the field of image clustering or in general some approaches that work with other kind of data. For example in [46] it is proposed an approach that use unsupervised consensus clustering [47] to discover intrinsic imaging sub-types. Then, the authors have created an association map between imaging and genomic to show that the imaging subtypes are associated with differing molecular pathways and that patients stratified by the imaging sub-types have distinct prognoses. Instead, in [48] it is proposed another consensus cluster analysis [47] this time based on demographic information, principal diagnoses, comorbidities, and laboratory data to study patients affected by hypernatremia. The goal of this study was to characterize hypernatremia patients at hospital admission into clusters and to evaluate the mortality risk among these distinct clusters. In [49] it is instead proposed a review for some unsupervised deep learning models such as autoencoders and their variants used for the study of medical images. An important aspect of the cancer subtyping research field is the difficulty in use effectively the vast amount of multimodal data that is available for cancer patients. To tackle this problem in [50] is proposed a multimodal neural network-based model that employs: clinical data, mRNA expression data, microRNA expression data and histopathology whole slide images to predict the survival of patients for 20 different cancer types. A work more focused on how extracting useful information from the medical images is proposed in [51]. In this work, authors make use of three pre-trained image classification models in order to extract features from histopathological images and then use the extracted features to perform supervised classification using SVM classifier to discriminate between cancer and normal samples. Instead a work that similarly to us want to use both the imaging information and the survival one is proposed in [52]. Indeed in that paper an approach that can be used for survival clustering on medical image data sets along with survival annotations in order to obtain meaningful image clusters

with respect to their survivability is described. This approach is called Medical Image Clustering for Survival risk group identification (MICSurv) that is a DeepConvSurv neural network. The DeepConvSurv neural network is a convolutional neural network (CNN) which differs from traditional CNNs in its final layer that aims for survival risk estimation using a loss function designed specifically for this task.
In general, finding effective ways to perform cancer subtyping is currently a trending research topics for therapy optimization and personalized medicine. In the last years the research in this filed have been also about images and how to extract useful information from this kind of data and our work goes in this direction.

1.4. ICC Data

All the data that have been used in this thesis were made available to us by the Humanitas Clinical and Research Center of Rozzano. The overall number of patients diagnosed with ICC included in our study is 259. These patients come from six different centers: Humanitas Clinical and Research Center of Rozzano, Policlinico Rossi of Verona, Gemelli Hospital of Rome, S. Orsola Hospital of Bologna, Mauriziano Hospital of Turin and Morgagni-Pierantoni Hospital of Forlì. For this reason, the statistical unit was the patient and they could be grouped according to their center of provenience. For these patients, both the overall survival time and the disease free survival time have been collected. These data have been used to perform multiple studies. Depending on the study the number of patients actually available changes. For each patient, radiomic features, extracted from pre-operative CT scans, together with personal and clinical variables and qualitative disease information were collected. The rest of the section deals with radiomic features and clinical ones separately, in subsection 1.4.1 and 1.4.2 respectively. This study was performed according to the Declaration of Helsinki [53]. The local review board approved the study and informed consent was waived given the observational retrospective design of the study.

1.4.1. Radiomic Features

The radiomic features presented in this study have been extracted from the three phases (Arterial, Portal and Late) of pre-operative CT scans. This operation, together with the segmentation of the regions of interest, has been carried out by experienced radiologist using the LIFEx software (www.lifexsoft.org, [54]). For each phase two distinct regions of interest have been selected. Indeed, not only the core of the tumor has been segmented, but also the margin region. The margin represents a peritumoral liver tissue. In particular, a region of 5mm that the software automatically generates around the tumor and that has been manually corrected to obtain a more precise result. From each region of interest 50 features have been extracted. Since there are three different images and for each one two different ROI are available, the radiomic variables were 300. Thanks to the fact that for each patient the radiomic features of more than one image are available, we can refer to the dataset as multi-view, where each view correspond to a phase of the CT scans. The extracted features are the following, stated according to LIFEx documentation [54]:

- Conventional features
  The values of these features are extracted from the original grey level values of the image in analysis
  - *HU_min*: is the minimum value of the Hounsfield values within the volume of interest;
  - *HU_mean*: is the average value of the Hounsfield values within the volume of interest;
  - *HU_std*: is the standard deviation value of the Hounsfield values within the volume of interest;
  - *HU_max*: is the maximum value of the Hounsfield values within the volume of interest;
  - *HU_Q1*: is the value of the first quartile of the Hounsfield distribution within the volume of interest. It corresponds to the value that has 25% of the data below it and 75% above it. It can be also defined as the middle value between the smallest value and the median of the dataset;
  - *HU_Q2*: is the second quartile of the Hounsfield distribution that corresponds to the median of the dataset. It is defined as the point that has 50% of the data points below its value;
  - *HU_Q3*: is the value of the third quartile of the Hounsfield distribution within the volume of interest. It corresponds to the value that has 75% of the data that are below it and 25% above it. It can be also defined as the middle value between the median of the dataset and the highest value;
  - *HU_Skewness*: is the asymmetry of the Hounsfield distribution;
  - *HU_Kurtosis*: reflects the shape of the Hounsfield distribution (peaked or flat) relative to a normal distribution;

  - *HU_ ExcessKurtosis*: compares the kurtosis coefficient with the one of a normal distribution.

- Histogram related features
  These are the variables extracted from a histogram. To build a histogram, it is necessary to determine a bin width ("bin" parameter) and the indices derived from the histogram will depend on this parameter. In LIFEx this parameter is computed starting from the number of grey level (*nbGreyLevel*) with the following formula:

$$bin = \frac{max - min}{nbGreyLevel}.$$

  Where max is the maximum value of intensity in the ROI and min the minimum one.

  - *HISTO_ Entropy_ log10*, *HISTO_ Entropy_ log2* : reflect the randomness of the distribution;
  - *HISTO_ Energy* : reflects the uniformity of the distribution.

- Shape features
  The shape features describe the shape of the selected ROI and its geometric properties.

  - *SHAPE_ Volume(mL)*, *SHAPE_ Volume(vx)*: are the Volume of Interest in terms of mL and voxels;
  - *SHAPE_ Sphericity*: is how spherical a Volume of Interest is. Sphericity is equal to 1 for a perfect sphere;
  - *SHAPE_ Surface*: surface of the region of interest, how much it measures in mm;
  - *SHAPE_ Compacity*: reflects how compact the Volume of Interest is.

- Grey Level Co-occurrence matrix-derived features - (GLCM)
  The grey level co-occurrence matrix (GLCM) describes the distribution of co-occurring pixel values at a given offset. It is computed from 13 different directions in 3D with a δ-voxel distance relationship between neighboured voxels. Each direction is associated to with a matrix. Given a certain direction, the element (i, j) of the respective matrix corresponds to the frequency with which there are two voxels, one of intensity i and the other of intensity j, separated by a given distance offset δ. The index value of the GLCM matrix is the average of the index over the 13 directions in space

(X, Y, Z). The offset (δ) is set to 1 by default, which means that only neighbour voxels are used to calculate GLCM.
Seven textural indices are computed from this matrix:

  - *GLCM_Homogeneity*: is the homogeneity of grey-level voxel pairs;
  - *GLCM_Energy*: also called Uniformity or Second Angular Moment, is the uniformity of grey-level voxel pairs;
  - *GLCM_Contrast*: also called Variance or Inertia, is the local variations in the GLCM;
  - *GLCM_Correlation*: is the linear dependency of grey-levels in GLCM;
  - *GLCM_Entropy_log10*, *GLCM_Entropy_log2*: are the randomness of grey-level voxel pairs;
  - *GLCM_Dissimilarity*: is the variation of grey-level voxel pairs.

- Grey Level Run Length matrix-derived features - GLRLM
  The grey-level run length matrix (GLRLM) gives the size of homogeneous runs for each grey level. With gray-level run length it is meant the set of consecutive voxels having the same gray level and where the "length" term indicate the number of voxels. This matrix is computed for the 13 different directions in 3D. The element (i, j) of GLRLM corresponds to the number of homogeneous runs of j voxels with intensity i in an image and is called GLRLM(i, j) thereafter. For each of the 11 texture indices derived from this matrix, the 3D value is the average over the 13 directions in 3D.

  - *GLRLM_SRE*, *GLRLM_LRE*: Short-Run Emphasis and Long-Run Emphasis are the distribution of the short and the long homogeneous runs in the ROI;
  - *GLRLM_LGRE*, *GLRLM_HGRE*: Low Grey-level Run Emphasis and High Grey-level Run Emphasis are the distribution of the low and high grey-level runs;
  - *GLRLM_SRLGE*, *GLRLM_SRHGE*: Short-Run Low Grey-level Emphasis and Short-Run High Grey-level Emphasis are the distribution of the short homogeneous runs with low and high grey-levels;
  - *GLRLM_LRLGE*, *GLRLM_LRHGE*: Long-Run Low Grey-level Emphasis and Long-Run High Grey-level Emphasis are the distribution of the long homogeneous runs with low and high grey-levels;

  - *GLRLM_ GLNU*, *GLRLM_ RLNU*: Grey-Level Non-Uniformity for run and Run Length Non-Uniformity are the nonuniformity of the grey-levels and the length of the homogeneous runs;
  - *GLRLM_ RP*: Run Percentage measures the homogeneity of the homogeneous runs.

- Grey Level Zone Length matrix-derived features - (GLZLM)
  The grey-level zone length matrix (GLZLM) provides information on the size of homogeneous zones for each grey-level in 3 dimensions. It is also named Grey Level Size Zone Matrix (GLSZM). Therefore, it is based on the concept of zone. A gray level zone is defined as a the number of connected voxels that share the same gray level intensity. A voxel is considered connected if the distance is 1 according to the infinity norm (26-connected region in a 3D). Element (i, j) of GLZLM corresponds to the number of homogeneous zones of j voxels with intensity i in an image and is called GLZLM(i, j) thereafter. From this matrix, 11 texture indices are computed:
  - *GLZLM_ SZE*, *GLZLM_ LZE*: Short-Zone Emphasis and Long-Zone Emphasis are the distribution of the short and the long homogeneous zones in the ROI;
  - *GLZLM_ LGZE*, *GLZLM_ HGZE*: Low grey-level Zone Emphasis and High Grey-level Zone Emphasis are the distribution of the low and high grey-level zones respectively;
  - *GLZLM_ SZLGE*, *GLZLM_ SZHGE*: Short-Zone Low Grey-level Emphasis and Short-Zone High Grey-level Emphasis are the distribution of the short homogeneous zones with low and high grey-levels;
  - *GLZLM_ LZLGE*, *GLZLM_ LZHGE*: Long-Zone Low Grey-level Emphasis and Long-Zone High Grey-level Emphasis are the distribution of the long homogeneous zones with low and high grey-levels;
  - *GLZLM_ GLNU*, *GLZLM_ ZLNU*: Grey-Level Non-Uniformity for zone and Zone Length Non-Uniformity are the nonuniformity of the grey-levels and the length of the homogeneous zones;
  - *GLZLM_ ZP*: Zone Percentage measures the homogeneity of the homogeneous zones.

- Neighbour Grey Level Difference matrix-derived features - (NGLDM)
  The neighborhood grey-level difference matrix (NGLDM) corresponds to the difference of grey-levels between one voxel and its 26 neighbours in 3 dimensions. Three

texture indices can be computed from this matrix.

  - *NGLDM_ Coarseness*: is the level of spatial rate of change in intensity;
  - *NGLDM_ Contrast*: is the difference of intensity between neighbouring regions;
  - *NGLDM_ Busyness*: is the spatial frequency of changes in intensity.

1.4.2. Clinical Variables

Together with the radiomic features also some clinical and qualitative disease information were collected. The variables present in this study are listed below, and a summary descriptive for continuous variables is reported in Table 1.1 and in Table 1.2 for the categorical ones:

- *Age*: variable representing the age in years of the patient;
- *CA19-9*: carbohydrate antigen 19-9 is commonly used as tumor marker even if it has a wide variation in sensitivity (50–90%) and specificity (54–98%) [55];
- *Max dimension*: Maximum dimension of the lesion in analysis. In case of multiple lesions the biggest lesion has been considered;
- *Number of lymph nodes removed*: number of lymph nodes removed during the surgery;
- *Number of metastatic lymph nodes*: number of metastatic lymph nodes that have been detected by the pathological analysis;
- *Sex*: variable representing the sex of the patient;
- *HCV*: binary variable (dichotomic) describing whether the patient has or not hepatitis C. HCV equals to one means that the patient is affected by hepatitis C;
- *HBV*: binary variable (dichotomic) describing whether the patient has or not hepatitis B. HBV equals to one means that the patient is affected by hepatitis B;
- *Neoadjuvant chemotherapy*: binary variable indicating whether the patient has gone under chemotherapy before the surgery. If Neoadjuvant chemotherapy is equal to one it means that the patient has gone under this therapy while if is equal to zero not;
- *Major Hepatectomy*: binary variable indicating whether the patient has been subject to surgical excision of three or more hepatic segments following the classification system of Couinaud [56];

- *Severe complications*: binary variable representing if the patient has experienced some severe complications during the hospitalization or in the following 90 days. The definition of severe complications follows the classification of Clavien-Dindo [57]. In this study the label of severe complication has been given to the cases of order three or superior. A complication of the third order is one where a postoperative procedure under local or general anesthesia is needed, a fourth grade implies intensive care while the grade 5 corresponds to the death of the patient;

- *Cirrhosis*: variable representing if the patient is affected by cirrhosis. This disease consists in an hepatic alteration with subversion of the organ anatomy and impairment of its functionality caused by chronic inflammation, more frequently secondary to hepatitis virus infection or alcohol abuse. It is characterized by the combination of necrosis with subsequent nodular regeneration and fibrosis;

- *R status*: variable representing the oncological completeness of the surgery. A value of R status equal to zero corresponds to complete excision of the tumor with the addition of a healthy hepatic parenchyma margin around the tumor, which guarantees complete excision. Instead, a R status equal to one indicates evidence of a resection without a safety margin in the anatomo-pathological analysis, which exposes the patient to the risk of a microscopic residual of the disease left in the part of the liver not removed;

- *Macroscopic vascular invasion*: binary variable describing whether there is macroscopic evidence (visible in the preoperative CT imaging, TAC or MRI, or visible with the naked eye during surgical dissection or at the section of the surgical piece) of a tumor growth within the blood vessels of the liver;

- *Microscopic vascular invasion*: binary variable describing whether there is evidence of anatomo-pathological analysis under the microscope (but not in the imaging nor with the naked eye) of infiltration of the peri-tumor vessels by the tumor;

- *Pattern*: variable representing the index of the tumor burden and of the distribution of the disease. It is defined according to Baheti et al.[58] where: pattern type 1 means single tumor, pattern type 2 means single tumor with satellite nodules or multiple tumor in the same hepatic segment and pattern type 3 means multifocal tumor in multiple liver segments;

- *Single Nodule*: variable representing the presence of a single lesion;

- *Grading*: variable representing a tumor cell differentiation index. This variable can assume three different values. If Grading is equal to one it means that the tumor is well differentiated, if it is equal to two is moderately differentiated and for a value of three the tumor is poorly differentiated.;

- *Perineural infiltration*: binary variable representing if there is evidence of tumor infiltration of peri-tumor nerve structures under anatomo-pathological analysis with the microscope;
- *Adjuvant chemotherapy*: binary variable indicating whether the patient has gone under chemotherapy after the surgery;

As already explained, the aim of this thesis was to spot and assess the potential relevance of radiomic data. In particular, this translates in studying both the core-margin interplay and the contribution of the three phases of the CT scans. We were also interested in find a good stratification for the patients affected by ICC and some risk factors linked with the imaging characterization. For this reason we have preferred to employ only the imaging features to perform cancer subtyping, while the clinical ones have, instead, been used to clinically characterize the clusters founded. In this way we are able to validate the stratification procedure thanks to the clinical variables that are exogenous to the model. Thus, we were also able to compare this clinical characterization with the known clinical risk factor seen in Section 1.1.

Table 1.1: Statistical summary of the numeric clinical variables

Variables	mean ± std.dev.	median (range)
Age	66.22 ± 10.56	67 (25.61 - 86.59)
CA19-9	1145.6 ± 6206.9	30.25 (0.20 - 67456)
Max dimension	56.05 ± 34.69	55 (10 - 270)
Number of lymph nodes removed	5.30 ± 5.66	4 (0 - 30)
Number of metastatic lymph nodes	0.62 ± 1.61	0 (0 - 12)

Table 1.2: Percentage of patients presenting a given clinical characteristic

Variables	number of patients (%)	Variables	number of patients (%)
Sex = Female	135 (52%)	*Adjuvant chemotherapy = TRUE*	105 (40%)
HCV = TRUE	31 (12%)	*HBV = TRUE*	24 (9%)
Neoadjuvant chemotherapy = TRUE	27 (10%)	*Major Hepatectomy = TRUE*	129 (49%)
Severe complications = TRUE	50 (19%)	*Cirrhosis = TRUE*	32 (12%)
Macroscopic vascular invasion = TRUE	75 (29%)	*Microscopic vascular invasion = TRUE*	142 (54%)
R status	79 (30%)	*Pattern = 1*	161 (62%)
Pattern = 2	65 (25%)	*Pattern = 3*	33 (13%)
Single Nodule = TRUE	219 (84%)	*Grading = 1*	30 (12%)
Grading = 2	143 (55%)	*Grading = 3*	86 (33%)
Perineural infiltration = TRUE	98 (38%)		

1.4.3. Data Preprocessing

The data set, as it is described in the previous sections, consists of 259 rows, each row corresponding to a patient affected by ICC, and 328 columns (300 radiomic features, 24 clinical variables and 4 response terms, namely overall survival time and disease free survival time together with the corresponding censoring indicators). The inclusion criteria for this study, that led us to 259 patients, were the followings:

- adult patients ($>$18 y.o.);
- patients undergoing liver resection for ICC confirmed at final pathology;
- resection performed in the period 01/01/2009 – 31/12/2019 (both open and minimally-invasive hepatectomies are allowed);
- preoperative CT imaging available for radiomic analysis;
- at least 1 ICC with diameter $>$10 mm.

While the exclusion criteria were:

- patients undergoing explorative laparotomy/laparoscopy without liver resection;
- mixed HCC-ICC at final pathology;
- preoperative imaging performed $>$90 days before surgery;
- loco-regional treatment of ICC before liver resection, including ablation, chemoembolization, or radio-embolization. Neoadjuvant chemotherapy is not an exclusion criterion.

All the data have been processed with MATLAB [59] and R [60]. After this, before proceeding with any of the studies described in the following chapters, the radiomic data have been normalized. In particular, they have been shifted by their mean and scaled accordingly with their variance.

A peculiar preprocessing of the clinical variables has been applied for the first application of the S2GC model [18], where only the Portal phase of the CT scans has been used. In that case some of the variables have been summarized thanks to the creation of two additional variables. The first one is called *Therapies* and it is a variable that represents which therapy the patient has undergone.
This categorical variable takes the following values:

- *0* if the patient has not gone under any of the three therapies: Neoadjuvant chemotherapy, Major Hepatectomy and Adjuvant chemotherapy;

- *1* if the patient has gone under only the Neoadjuvant chemotherapy;
- *2* if the patient has gone under only the Major Hepatectomy;
- *3* if the patient has gone under only the Adjuvant chemotherapy;
- *4* if the patient has gone under both the Major Hepatectomy and the Adjuvant chemotherapy, but not under the Neoadjuvant chemotherapy;
- *5* if the patient has gone under both the Neoadjuvant chemotherapy and the Adjuvant chemotherapy, but not under the Major Hepatectomy;
- *6* if the patient has gone under both the Neoadjuvant chemotherapy and the Major Hepatectomy, but not under the Adjuvant chemotherapy;
- *7* if the patient has not gone under all the three therapies: Neoadjuvant chemotherapy, Major Hepatectomy and Adjuvant chemotherapy;

The second one is called *Comorbidity* and it is a categorical variable that represents the general health status of the patients and it is defined according to the number of other diseases affecting the patient. The comorbidities takes into account are: *HCV*, *HBV*, *Severe complications*, *Cirrhosis*, *R status*, *Macroscopic vascular invasion*, *Microscopic vascular invasion* and *Perineural infiltration*. This variable is defined as follows:

$$Comorbidity = \begin{cases} 0 & if \sum cmmorbidities = 0 \\ 1 & if \sum comorbidities \geq 1 \ \& \ \sum comorbidities \leq 3 \\ 2 & if \sum comorbidities \geq 4 \end{cases} . \tag{1.1}$$

Instead for the other two applications of the S2GC model [18] we have created another variable that establish if the diseases is metastatic or not. This variable is called *Metastatic disease* defined as follows:

$$Metastatic\ disease = \begin{cases} 0 & if Number\ of\ metastatic\ lymph\ nodes = 0 \\ 1 & if Number\ of\ metastatic\ lymph\ nodes \geq 1 \end{cases} . \tag{1.2}$$

After this important contextualization and the presentation of the data that we have used in our analyses, in the next chapter we are going to present our first study. In particular, we will present two analyses that we have performed on the ICC data with two supervised method, following a quite standard procedure to analyse radiomic data.

2 | Supervised analysis on ICC patients

As anticipated, in this chapter we are going to explain our firsts analyses on the multiphase radiomic dataset of the patients affected by ICC. As described in Chapter 1, the radiomic data are often studied through supervised analysis. Therefore, as first attempt to assess whether the different phases of the CT scans proved different and complementary information, we have applied a typical workflow. We recall that this is composed by a feature selection step followed by a supervised method to analyse an outcome. In this setting we have performed two different analyses. The first one was a classification problem where we were interested in classify the patients affected by ICC accordingly with their survival. This means that we wanted to build a model to predict whether the death of the patient has occurred within the experiment time. The second one, instead, was a survival analysis on the overall survival time of these patients. In these two analyses we have employed both the clinical variables and the imaging ones. For this reason we have kept only the patients for which no missing values were present in both the radiomic and the clinical variables. This leads us to a dataset composed of only 134 patients.

The following part of this chapter is divided in two main part. In the first one the classification problem has been tackled while the second part regards the survival analysis. In the end of the chapter we have highlighted the limitations of this supervised approach and proposed an alternative method to study our multi-view dataset.

2.1. Classification Problem

In this section we describe how we have tackled the classification problem. We recall that this problem consists in building a model to predict whether the death of the patient has occurred within the experiment time. In particular, since we were interested in assessing whether the use of a multiphase dataset can improve the predictive performances, we have built three different models based on an increasing number of radiomic features and compared their performances. For all the three models we have employed the clinical variables available in our dataset, while the radiomic ones have been added one phase at the time. Therefore, for the first model only the variables extracted from the Portal phase together with the clinical ones have been employed. Instead, in the second one also the features relative to the Arterial phase have been used and in the third we have used all the radiomic features available.

The model that we have adopted to solve this classification problem was the logistic regression model. Before proceeding with the description of how we have exploited the radiomic data in Section 2.1.1 the logistic regression model is explained.

2.1.1. Logistic Regression

As anticipated, the model that we have adopted to solve this problem is the logistic regression model. This model is the most important one for categorical response data and it is commonly used for a wide variety of applications.

The logistic regression models are Generalized Linear Models (GLM) with binomial random component, and logit link function [61]. Indeed, a binary response variable Y, as in our case the variable representing if the patients is dead or not, can be modelled as a binomial. Each observation can be treated as a single Bernulli trial. Therefore, the mean $E(Y)$ is equal to $P(Y = 1)$ and we can denote $P(Y = 1)$ by π. We can express the probability mass function as

$$\begin{aligned} f(y; x) &= \pi^y (1-\pi)^{1-y} = (1-\pi)[\pi/(1-\pi)]^y \\ &= (1-\pi) exp\left[y \left(log \frac{\pi}{1-\pi} \right) \right] \end{aligned} \tag{2.1}$$

with y that can assume only 0 and 1. We recall that the natural exponential family has a probability mass function of form

$$f(y; \theta) = a(\theta) b(y) exp[y Q(\theta)]. \tag{2.2}$$

Therefore it is possible to notice that equation 2.1 correspond to a probability mass function inside the natural exponential family. In particular, it is possible to identify θ with π, $a(\pi)$ is equal to $1 - \pi$, $b(y)$ is equal to 1 and $Q(\pi) = log \frac{\pi}{1-\pi}$. Thus, the natural parameter $(Q(\pi))$ is the log odds of response outcome 1, called logit function of π. The link function that transforms the mean of the response variable (π) to the natural parameter is called the canonical link and in this setting it is the logit function.

Introducing the explanatory variable X we can redefine $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$ reflecting its dependence on the values of the explanatory variable.

Thus, the logistic regression model, as defined in [61], is

$$\pi(x) = \frac{exp(\alpha + \beta x)}{1 + exp(\alpha + \beta x)}. \tag{2.3}$$

To see that the link function needed in this case is the logit function we can see that the odds are

$$\frac{\pi(x)}{1 - \pi(x)} = exp(\alpha + \beta x). \tag{2.4}$$

Therefore, the log odds has the linear relationship

$$log\frac{\pi(x)}{1 - \pi(x)} = \alpha + \beta x. \tag{2.5}$$

Thus, we can conclude that the appropriate link function for the logistic regression is the log odds called logit function.

In our setting the binary response variable Y correspond to the variable that establish whether the death of the patient has occurred within the observation time. The explanatory variable X, instead, change depending on the model. As previously said, we have built three different models. In the first one X is composed by the clinical variables and by the radiomic features extracted from the Portal phase of the CT scans. In the second model X is instead composed also by the features relative to the Arterial phase and in the last one also by the variables extracted from the Late phase. The parameter α in equation 2.3 is the intercept parameter, while β is the weights of the explanatory variable X. The sign of β determines whether $\pi(x)$ is increasing or decreasing as x increases and the magnitude of β determines the rate of this increase or decrease. Therefore it establish whether the variable x is a risk or a protective factor.

2.1.2. Results

As described in Chapter 1, the typical first step when analysing radiomic data is a feature selection step. Since this kind of data are for their nature highly correlated first of all we have performed a correlation analysis. Indeed, since to solve the classification problem we have adopted a logistic regression model we need to eliminate the features that are too highly correlated. To do this we have kept only the variables that had a correlation under 0.8 while all the others have been excluded. After this, we have performed the feature selection step thanks to an approach based on the stepwise logistic regression [62]. As result for all the three models both clinical and radiomic features have been selected as relevant. In particular, in the complete model, the one where we have exploited all the three phases of the CT scans, have been selected features coming from all the phases and all the ROIs. Then, as anticipated, to perform the second step of the procedure described in Chapter 1, we have used the logistic regression model described in Section 2.1.1. With this model we have built all the three models that we were interested in comparing. Indeed, we recall that we have built three different models with an increasing number of phases employed and that we were interested in assess whether this addition of phases implies an improvement in the classification performances. To obtain more reliable results we have applied a cross-validation procedure. Therefore, in this section we are going to present the results found with a 10-fold cross-validation procedure. Then, in order to asses whether adding the features of a different phase of the CT scans have improved the performances we have employed the McNemar's test [63].

In Table 2.1 the mean accuracy and the relative standard deviation for the three model are presented. As it is possible to notice the accuracy increases as we add more phases to the model. Unfortunately, the standard deviation is high for all the models. Furthermore the performances are not as good as we would like.

Table 2.1: Results of the three logistic regression model implemented to predict whether the patient is dead during the time of the study

Model	mean(accuracy)	std.
Clinical variables + Portal phase	0.6763736	0.1081812
Clinical variables + Portal phase + Arterial phase	0.7302198	0.1201959
Clinical variables + Portal phase + Arterial phase + Late phase	0.7675824	0.1219304

From these results we have a first evidence that adding more phases can produce a better result. Thanks to some McNemar's tests we have found that with a significance level of $\alpha = 0.05$, the increment between the model with only the Portal phase and the model with also the Arterial one is statistically significant (p-value = 0.023). The same can be said for the comparison between the first and the complete model (p-value = 0.001). Instead, the improvement between the model that employs the Portal and Arterial phases and the complete one is not statistically significant (p-value = 0.227).

As already anticipated this approach has some limitations. For example the radiomic features, as described in Section 1.2, are built in order to extract as much information as possible. Doing this also some variables that are highly correlated have been created and these variables can not be used in this setting. To solve the problem of highly correlated features before preforming the feature selection, we have performed a correlation analysis. The problem in this approach is that discarding some variables it might leads to a loss of useful information. Another limitation is that dealing with radiomic data are always needed a lot of samples, due to the fact that these variables can be many, but this need is more evident for the supervised models and in this setting we have only few patients available. Another important limitation is that, as discussed in [64], [65] and [66], the classification task performed with conventional statistical techniques is often biased and difficult to generalize. Indeed, the authors of these works claim the importance of using an external validation in order to assess the correct performances of a model.

2.2. Survival Analysis

In this section we present the survival analysis that we have performed on the overall survival time of patients affected by ICC. As previously said, in this setting we have employed all the radiomic features, coming from the three phases of the CT scans, together with the clinical variables. For this reason the patients available were 134.

In section 2.2.1 the Cox proportional hazards model [36] is presented together with some notion on the survival analysis useful to better introduce this model. Then in Section 2.2.2 the results obtained with this model are reported.

2.2.1. Cox Model

In survival analysis the object under study is the time to a given event of interest. In this case the time that we are interested in studying is the overall survival time of the patients affected by ICC. This time, for a patient i, can be either an observed time O_i or a censored one C_i. For this reason it is needed a variable that states if the time is observed or censored. So, we can define δ_i as the censoring variable which identifies if the event for the $i-th$ patient has been observed or not. This variable correspond to the one that we were trying to predict in the previous classification problem. Thanks to δ_i it is possible to define the time T_i as follows:

$$T_i = \begin{cases} O_i & if \delta_i = 1 \\ C_i & if \delta_i = 0 \end{cases}. \tag{2.6}$$

The survival function gives the probability that a subject (i) will survive past time t, so that his survival time T_i is greater than t. This function is defined as:

$$S_i(t) = \mathbb{P}(T_i \geq t). \tag{2.7}$$

Instead the hazard function $h_i(t)$ is the instantaneous rate at which events occur, given no previous events:

$$h_i(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T \leq t + \Delta t | T > t)}{\Delta t}. \tag{2.8}$$

We can than introduce the model that we have employed to represent the survival analysis data, that is the Cox proportional hazards model [36]. This model is a regression model commonly used in statistical medical research for investigating the association between the survival time of patients and one or more predictor variables. The purpose of the model is to examine how specified factors influence the rate of a particular event happening (e.g. death in our case) at a particular point in time. This rate is often called hazard ratio. The Cox model is expressed by the hazard function, that can be interpreted as the risk of dying at time t. Calling $h_0(t)$ the baseline hazard function the hazard function is expressed as:

$$h_i(t|X_i) = h_0(t) exp(X_i w) \tag{2.9}$$

where t is time in analysis, w are the weights of the variables X_i relative to patient i. Variables X_i that in our case corresponds to the clinical variables and the radiomic ones extracted from all the three phases of the CT scans.

2.2.2. Results

In this section we are going to present the results obtained for the survival analysis performed, as explained in Section 2.2.1, through the Cox model. To build this model we have employed the radiomic features extracted from all the three phases of the CT scans and the clinical variables. As already described for the classification problem (Section

2.1), also in this setting we have performed a correlation analysis. We have kept the variables for which the correlation was smaller then 0.8. Then, a Principal Component Analysis (PCA) on the radiomic feature has been performed in order to reduce the dimensionality of our problem. We have kept the principal components that explain the 95% of the variance. This leads us with 30 principal component that have been added to the clinical variables to build the model. The performances have been evaluated thanks to the c-index and to obtain a more reliable result we have adopted a 10-fold cross validation approach. For this reason the result are presented as mean value and the relative standard deviation. The mean c-index reached was 0.67 with a standard deviation of 0.09. Unfortunately, the performances of this model were not satisfactory and perhaps even more importantly, the interpretation was not easy. Indeed, one of our goals was to build a easily interpretable model and the Cox model could have been a good choice in this direction. The problem relies on the fact that to preserve as much information as possible after the correlation analysis we have performed a PCA. This approach helped us to exploit the information provided by the radiomic features but it led to a big loss for what concerns the interpretability. Furthermore, since, as for the classification model, we needed to performed a correlation analysis we were not able to exploit all the information included in the radiomic data that are for their nature often highly correlated. Another limitation is that, as previously said, a lot of samples are needed to build a good model and in this setting we don't have enough patient, especially after having excluded the missing values on the clinical variables.

2.3. Conclusions

Summing up, thanks to the classification of the overall survival performed with the logistic regression we have obtained the first evidence that the three phases provide different and useful information. However, performing this classification task and the survival analysis, we have highlighted some important limitations of the supervised analysis in this setting. In both cases there were some problems in exploiting all the information included in the radiomic data that are for their nature highly correlated. In addition to this, the performances of both the two models were not satisfactory. Also problems due to the small number of patients available have been highlighted. For what concerns the survival analysis, the loss of interpretability due to the PCA was not indifferent. For all these reasons, the supervised framework was not the optimal choice for our study. Therefore, we decided to focus on a different approach and, in particular, on a unsupervised one that allowed us to perform a change of perspective. Indeed, with a supervised approach the goal was to predict an output thanks to the radiomic data. Instead, now we are interested in analysing the similarity/dissimilarity between the radiomic features of the patients to cluster them in groups. Groups that are characterized by patients with a similar prognosis and a similar imaging characteristics. This procedure is called cancer subtyping and in particular the method that we have used is called Supervised Survival Graph Clustering model [18]. This algorithm was proposed in the genomic literature and it will be analyzed in detail in the following chapter. As we will discuss later, already the name says that it is not a completely unsupervised approach. Indeed, it is a distant-supervised approach. This means that it performs also another task and use it to better perform the main one. Also the concept of distant supervision will be explained more in detail in the next chapter.

3 | Cancer subtyping methodological pipeline

As anticipated, this chapter concerns the discussion of the Supervised Survival Graph Clustering (S2GC) model [18] that we have used to overcome the limitations seen with the supervised analysis and to achieve all the purposes of our work. We have exploited this chapter also to introduce the three different applications of this model to the ICC data. Indeed, after a complete discussion of the S2GC model where all its components have been analysed we have presented how we have applied this method to our data. In particular, for each of the three application we explain: the goal of that application, the data that we have used, how we have optimized the parameters of this model, the choice of the number of clusters and two techniques important to have a better comprehension of the results presented in the following chapter. We start by anticipating that the first application have the goal of assessing whether the addition of radiomic features can improve the stratification of patients affected by ICC by focusing on the contribution of the core and of the margin region. Indeed, in this case only the Portal phase of the CT scans have been used and the goal was to study the core-margin interface trying to establish if these two regions provide different information. In the second, instead, the complete version of the multiphase dataset has been used. Indeed, in this setting the goal was to assess whether the three phases of the CT scans provide different and complementary information. For these two applications we were interested in provide a stratification of the patients when their overall survival probability was studied. In the last application of the S2GC model to the ICC data the goal was the same of the second one but the medical problem was different. Indeed, in this case the stratification analysed was for the disease free survival time and not for the overall survival time. In all the three applications we were interested in obtaining groups that were homogeneous for both the prognosis and the clinical characterization. As we will see in detail in the following part of the chapter we have not employed the clinical variables to compute the stratification but only to validate this procedure. In this way we were able to give a validation of our procedure using variables not employed in the model.

The structure of the Chapter is the following:

- In section 3.1 the Supervised Survival Graph Clustering model presented in [18] is discussed.

- In section 3.2 an application of the S2GC model [18] to the data on the patients affected by ICC is presented. In particular, radiomic features extracted from the Portal phase of the CT scans are used. The focus of this application was to study the core-margin interface (the tumor-tissue interface) and compare the information provided by the two regions. Furthermore, it has been useful to build a baseline for the stratification of patients affected by ICC where only one phase of the CT scans have been employed and the overall survival time was studied.

- In section 3.3 the main application of this work is described. In this case all three phases of the CT scans (Arterial, Portal and Late) have been exploited. The purposes of this application were: to establish whether these three views provide different and complementary information and to build a complete characterization for the patients affected by ICC, when the overall survival of the patients was studied.

- In section 3.4 another application of the S2GC model to the ICC data is presented, studying the disease free survival time instead of the overall survival;

- In section 3.5 a sum up of the principal results found in the previous sections is reported.

3.1. Supervised Survival Graph Clustering model

In this section the algorithms that we have employed for our main study are presented. In particular, all the techniques needed to perform the cancer subtyping are introduced. To perform this kind of analysis two steps are needed: (1) a patient-to-patient similarity graph needs to be estimated and (2) this graph has to be clustered in a homogeneous population of nodes. In order to estimate the patient-to-patient similarity graph we have employed the Supervised Survival Graph Clustering model [18] (S2GC) that we are going to discuss in detail. First of all, this algorithm has been proposed in [18] to perform cancer subtyping with the genomic data. As we have seen in Section 1.3 the majority of the algorithms that performs this kind of analysis has been proposed in the genomic field. This was true especially till the last years, indeed recently also some works in the imaging clustering field or that more in general exploit also other kind of data have been proposed. Despite this, we were interested in borrow this model from the genomics field to use it with the radiomic data as it was for the first time proposed in [19]. The first reason for which we want to borrow this model from the genomic literature is that, as typically done in recent works proposed in that field, it works for multi-view dataset. This was really important for us since we want to exploit a radiomic patient representation that is multi-view. Another major characteristic of this algorithm to which we was interested in was the fact that it is not a completely unsupervised method. Indeed, as we have seen in Section 1.3, the majority of the models proposed are completely unsupervised, while the S2GC model exploits a distant supervised approach. In fact, as we will see later, it performs two tasks simultaneously: the survival analysis learning model and the adaptive graph estimation, and use the first to compute the second. The concept of distant supervision comes from the Natural Language Processing field and in that setting it is used to do relation extraction and sentiment analysis [67]. This approach consists in training a model for a task different from the final scope, using labels that are not completely pertinent with the problem to be tackled. It thus brings the possibility to solve tasks with non-retrievable labels in a supervised way. In this setting we wanted to cluster patients in groups with different prognosis exploiting their imaging characteristic to predict their survival estimates. Doing this way we were able to combine some of the advantages of the unsupervised setting with some of the ones of the supervised. Indeed, thanks to the fact that we were using an unsupervised technique we were not anymore building a model to predict a label. We were, instead, analysing the similarity/dissimilarity of the imaging characteristic of the patients. Therefore, we were able to distinguish cluster of patients with similar imaging characteristic and then analyse the clinical condition and the risk factor linked with the radiomic features. This was a big change of perspective with respect to the one usually had with the supervised analysis. However, we were also exploiting an advantage of the supervised setting. Indeed, this setting allowed us to inform our clustering with the information about the survival analysis. This was crucial because helped us to obtain groups of patients characterized by patients with an homogeneous prognosis while the different groups have different survival risks.

As previously said, to perform cancer sub-tying two steps are need: (1) estimation of the patient-to-patient similarity graph and (2) this graph has to be clustered. For this reason in the following part of this section as first, we present the S2GC model with a discussion of all its term and the techniques used to select the best hyperparametrs. Then, the clustering technique used to cluster the patient-to-patient similarity graph and how we have selected the best correct number of groups is explained.

3.1.1. Model definition

Here the Supervised Survival Graph Clustering model is defined. As already said this algorithm performs two tasks jointly. Indeed, it learns the survival analysis model and use it together with the imaging data to do the adaptive graph estimation. For this reason, it is composed by different parts, more precisely four, that we are going to analyse in detail. The four components are: 1) the survival analysis part, 2) an L1 regularization term, 3) a co-regularization between the views and 4) the adaptive graph estimation. So the S2GC model is based on the minimization of the following loss function:

$$\begin{aligned}
&\min_{w;S} \sum_{k=1}^{m} \left(- \sum_{i=1}^{n} \delta_i \left(X_i^k w^k - log \sum_{j \in R_i} exp(X_j^k w^k) \right) \right) \\
&+ \lambda \sum_{k \neq j} \|X^k w^k - X^j w^j\|_2^2 + \eta \sum_{k=1}^{m} \|w^k\|_1 \\
&+ \min_S \gamma \sum_{i=1}^{n} \sum_{j=1}^{n} (\|X_i - X_j\|^2 + \|X_i w - X_j w\|^2) S_{i,j} + \mu S_{i,j}^2 \\
&s.t. \sum_{j}^{n} S_{I,j} = 1, S_i \succeq 0; i = 1, 2, \ldots, n.
\end{aligned} \tag{3.1}$$

From the equation 3.1 can be clearly noticed the four terms cited above. The first one consist in the first line of the loss function. The second line represent the two penalization terms and the last part is the one referred to the adaptive graph estimation. So let's analyse each point in detail.

3.1.2. Survival Analysis based loss term

The survival analysis model learning is based on the Cox proportional hazards model [36]. For this reason, we can continue what seen in Section 2.2.1, where some notation and the Cox model have been introduced. In this setting the time to event T_i can be both the Overall survival and the diseases free survival, depending on the application. The X_i, instead, are just the radiomic variables of patient i without the clinical ones. For greater clarity we report here the equation of the hazard function:

$$h_i(t|X_i) = h_0(t) exp(X_i w)$$

where w is the vector of the coefficients relative to the variables X_i. These weights can be estimated by solving the negative partial log-likelihood:

$$-\sum_{i=1}^{n} \delta_i \left(X_i w - log \sum_{j \in R_i} exp(X_j w) \right) \tag{3.2}$$

where R_i is the risk set at time T_i. Namely the set of patients that are at risk at that time, that are the ones with observed times not less than T_i. Then we need to introduce the multi-view, since till now all the discussion have been made for just one view. Indeed, we have employed X_i that was a vector in $\mathbb{R}^p$ and assuming to have n patient $X \in \mathbb{R}^{p \times n}$. Instead now, having $k = 1, ..., m$ views, we have $X^k \in \mathbb{R}^{p^k \times n}$ that is the matrix representing the radiomic data of the $k-th$ view for the n patient. Indeed, p^k represent the number of radiomic features in the $k-th$ view. So, thanks to this, we can define the multi-view Cox model seen in equation 3.1 and that for simplicity we report here:

$$\min_{w} \sum_{k=1}^{m} \left(-\sum_{i=1}^{n} \delta_i \left(X_i^k w^k - log \sum_{j \in R_i} exp(X_j^k w^k) \right) \right). \tag{3.3}$$

This part, together with the penalization one, is the core of our distant supervision approach. Indeed here we compute the estimate of the survival-related risks w that will be exploited, along with the radiomic features, to compute the similarity between patients. Thanks to this weights we are able to give more importance to some variables and less to others also for what concerns the similarity graph. This is really useful in order to obtain groups that have different prognosis and that are homogeneous within the group.

3.1.3. Penalization terms

The second and the third terms of the equation 3.1 are two penalization terms. The first one of the two is a L2 regularization term that works as co-regularization between the imaging views' contributions. In particular, as explained in [18], it performs a shrinkage on the agreement of the prediction between a pair of different views. The parameter λ is the one that controls this regularization. We report here the equation for a more clear discussion:

$$\lambda \sum_{k \neq j} \|X^k w^k - X^j w^j\|_2^2. \tag{3.4}$$

This term has been crucial for our analysis because by analyzing the control parameter λ we were able to study multiple characteristics of the radiomic data. More specifically, the relationship of the core-margin interface with respect to prognostic risks and later the

difference between the three phases of the CT scans for what concerns the estimation of the same risks. These analysis will be discussed in detail in the followings sections.
The second penalization terms, as can be seen from the equation that we have reported here:

$$\eta \sum_{k=1}^{m} \|w^k\|_1 \tag{3.5}$$

is a L1 regularization term for the weights w of the Cox model. This term penalizes the single variable, while at contrary the previous one penalize all the features in the view. For this reason, the sparsity control parameter η addresses the problem of high-dimensional data, in a feature selection fashion. This term has been used also to compute a ranking on the importance of the radiomic variables, as we will see for example in Section 3.2.4.

3.1.4. Adaptive graph estimation term

The last term of the equation 3.1 is the one that has the scope of building the patient-to-patient similarity graph thanks to an adaptive graph estimation algorithm. For greater clarity we report here the equation:

$$\begin{aligned} &\min_S \gamma \sum_{i=1}^{n} \sum_{j=1}^{n} (\|X_i - X_j\|^2 + \|X_i w - X_j w\|^2) S_{i,j} + \mu S_{i,j}^2 \\ &s.t. \sum_{j}^{n} S_{I,j} = 1, S_i \succeq 0; i = 1, 2, \ldots, n. \end{aligned} \tag{3.6}$$

Thanks to the minimization of this equation S, that is the $\mathbb{R}^{n \times n}$ affinity matrix, is estimated. Since it is the patient-to-patient similarity graph, $S_{i,j}$ represents the similarity between patients i and j. In equation 3.6 there are two terms that compute a distance between patient i and patient j. Indeed, the first one compute the squared distance between the radiomic features of the two patients, while the second one does the same for what concerns their survival analysis estimate. In fact, in the second term are present the weights w of the Cox model. This is done because, as explained before, this is a distant supervision approach. Thus, thanks to the Cox model a survival analysis have been performed and here this information is exploited to estimate the patient-to-patient similarity graph S. The other terms in the equation 3.6 are γ that is the learning rate and μ that is a trade-off parameter.

3.1.5. Hyperparametrs selection

The parameters that we have to select are λ (the co-regularization parameter), η (the L1 penalization parameter) and γ (the learning rate). To select the best ones we have built a grid search optimization procedure by maximizing the Harrell's concordance index (c-index) of the estimated survival risks [68]. We have studied the performances of the model as function of the different parameters and selected the ones that return the maximum value of c-index as optimum.

3.1.6. Clustering technique

Once we have computed the patient-to-patient similarity graph we need a clustering algorithm in order to divide that graph in sub-graphs representing different groups with similar characteristics. To isolate these homogeneous groups we have used a spectral clustering algorithm [69]. In order to be able to perform a good clustering, a crucial step is to select the right number of clusters nc. We solved this problem using the eigengap heuristic, which can be applied to graph Laplacians, either normalized or non-normalized [70]. This consists in choosing nc such that all the eigenvalues up to the $nc - th$ one are zeros whereas the $(nc + 1) - th$ one is different from zero or at least the first nc are small while the $(nc + 1) - th$ one is relatively large.

3.2. Core/Margin assessment in Portal CT imaging

This section concerns the first application of the S2GC method [18] to the data of the patients affected by ICC. This analysis has two scopes: the first is to study in detail the core-margin interface; the second is to build a baseline for the stratification of patients affected by ICC when the focus is on the overall survival. For this reason the data that have been employed are all the patients for which the radiomic features of the Portal phase of the CT scans were available. In this case the study comprehends all the 259 patients of the dataset described in Section 1.4. Since the patient representation is build only on the radiomic features, we have not excluded the patients that have some missing values in the clinical variables. The patient representation is explained in detail in the following section. Furthermore, since we are interested in the overall survival, the time T_i used to define R_i in equation 3.1 are based on the overall survival time.

3.2.1. Construction of the multi-view patient representation

Only the radiomic features have been employed to build the patient representation. Instead, the clinical variables, exogenous to the model building, have been used to clinically characterize the subpopulations of patients found thanks to the Portal CT imaging. In this way the stratification procedure has been validated. Furthermore, the S2GC model [18], as explained in Section 3.1, works with multi-view datasets. For these reasons the patient representation, as shown in Figure 3.1, has been built creating two different views. In the first view there are all the radiomic features extracted from the core of the tumor, while in the second view there are all the variables that come from the marginal region. In this way, two views composed by 50 features each have been built. So in this case X_i^k, of the equation 3.1, is composed by the 50 features extracted from the core of the Portal phase of the CT scans of patient i for $k = 1$ while by the ones extracted from the margin region for $k = 2$.

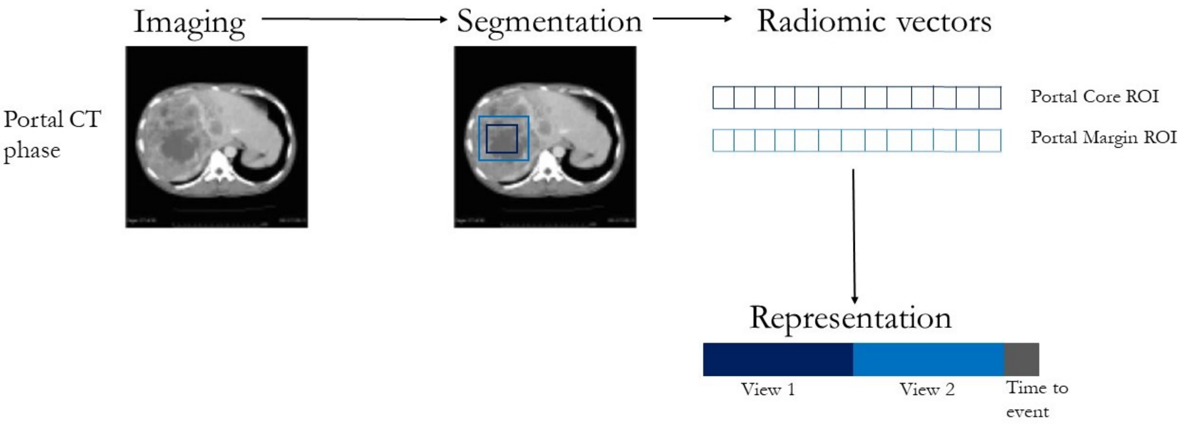


Figure 3.1: Sketch of the patient representation used in the S2GC model when core and margin views of the Portal phase of the CT scans is used.

3.2.2. Parameters Optimization

The parameters that we need to optimize are: γ (the learning rate), λ (the co-regularization parameter) and η (the L1 penalization parameter). In order to select the best parameters a grid search algorithm has been built. The tuple of parameters that returns the highest value of c-index [68] has been selected as the optimal one.

Starting from γ, we can look at Figure 3.2. In this figure: on the left is represented the number of iterations needed by the S2GC algorithm to converge for the different values of γ, while on the right is shown how the c-index varies according to γ. The curve that

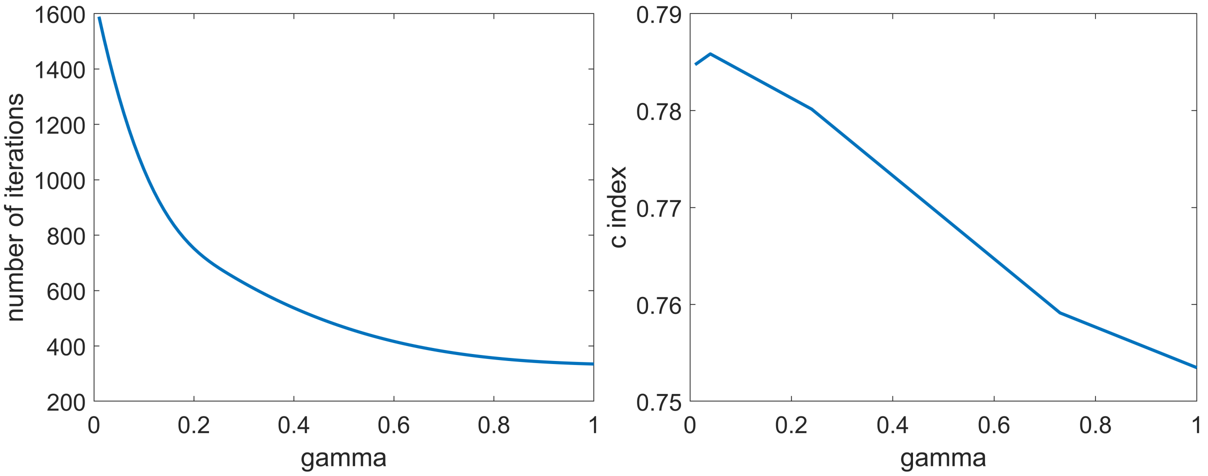


Figure 3.2: Left panel: the number of iterations needed by the S2GC algorithm to converge for the different values of γ. Right panel: c-index variation according to γ. Both in the case where we are studying the overall survival using only the Portal phase of the CT scans.

represents the number of iterations needed to converge as function of γ is monotonically decreasing. Instead the function representing the c-index as function of the learning rate has a maximum in $\gamma = 0.04$ and then it decreases. So, as expected, it is possible to notice that, for a small value of the learning rate, the algorithm needs more iterations in order to converge but the performances are better than having a big value of γ. For this reason the optimal value (γ^*) selected for the learning rate is 0.04.

Looking at Figure 3.3, it is possible to study how the c-index varies as function of both λ and η. In particular, in this figure there are four lines that represents how the c-index varies as function of η for four fixed values of λ (0, 0.2041, 0.4082, 10). First of all, it is clear that the best result is for the blue line that is the one obtained with $\lambda = 0$. Analysing this line we have found that the maximum is for $\eta = 0$. So we can conclude that the best results are achieved in case of no co-regularization between the views, in this case between the core and the margin, and no L1 penalization of the weights w of the Cox model.

A study on the weights w of the Cox model [36] has been performed as well. In Figure 3.4 are shown the weights of the Cox model as function of the L1 regularization parameter η for the the radiomic variables extracted from the core once γ and λ have been fixed to their optimal value γ^* and λ^*. In order to make a more understandable figure the variables have been divided in six different groups. They correspond to the ones presented in section 1.4.1 where the Conventional and the Histogram related features have been united. In Figure 3.5 is represented the same analysis with the same division in groups,

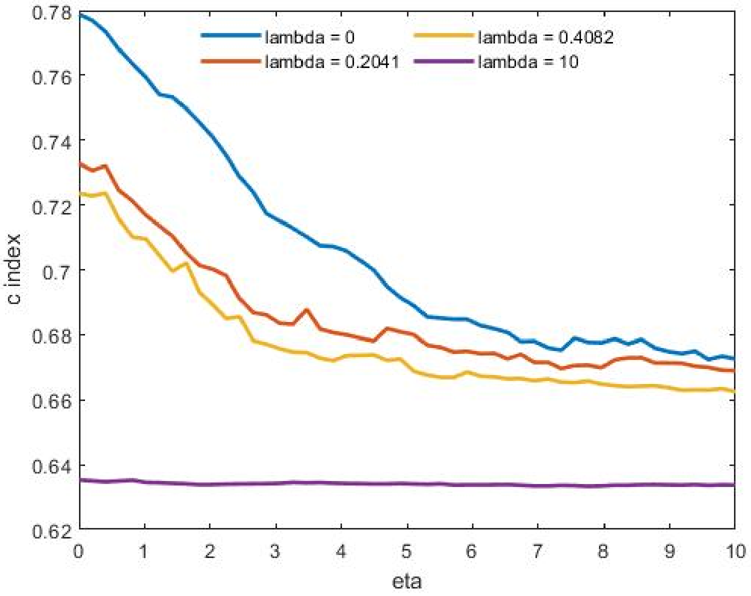
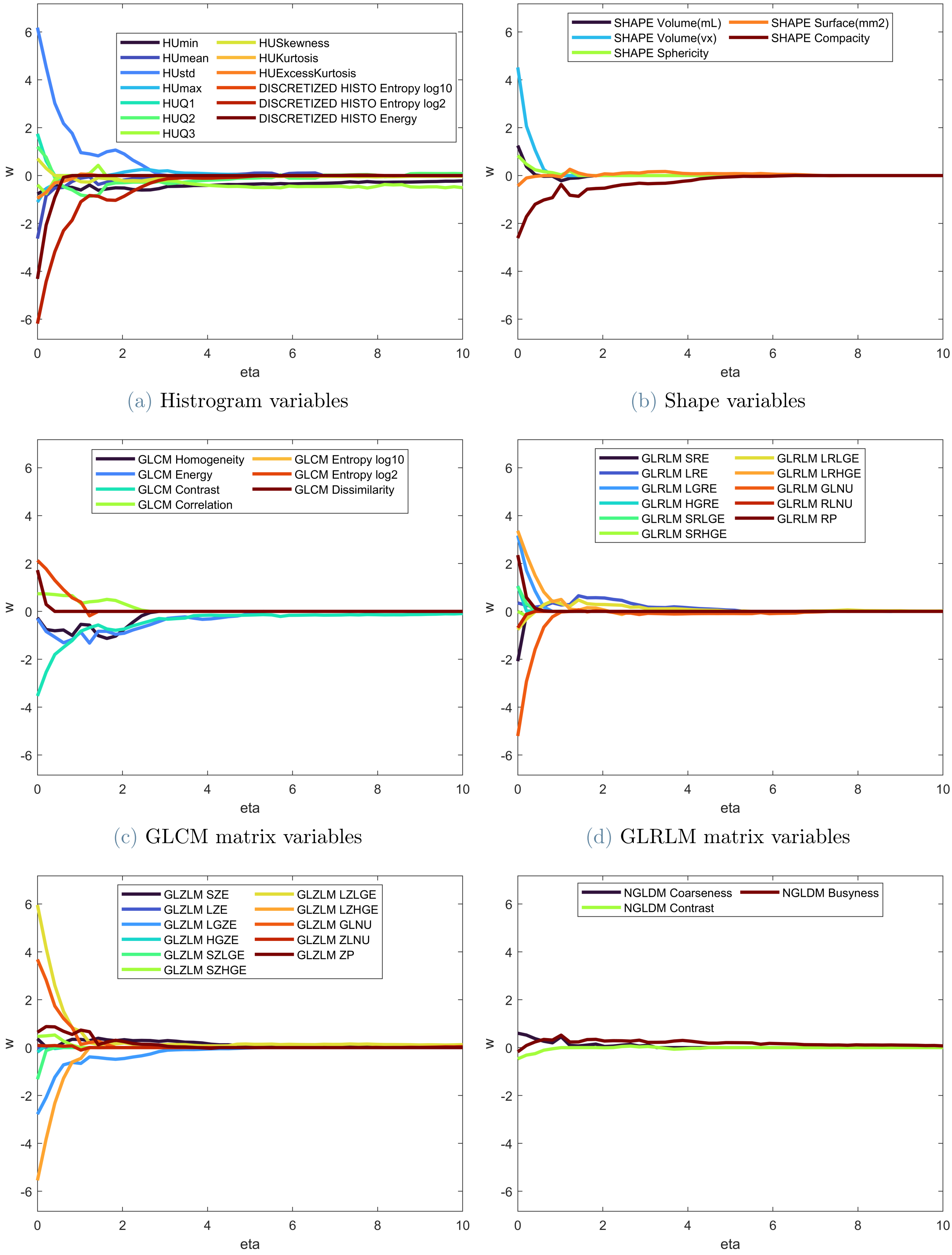


Figure 3.3: C-index variation as function of η for some fixed values of λ in the case where we are studying the overall survival using only the Portal phase of the CT scans.

but for the variables extracted from the margin region of the tumor. Thanks to these two figures we have analysed how the selection of η influences the weights. As expected, a bigger value of the L1 regularization parameter leads to smaller weights.

After this study an analysis on the influence of λ on the weights w has also been performed. In Figure 3.6 we have reported how the weights vary accordingly with the L1 regularization parameter η for the the radiomic variables extracted from the core once γ have been fixed to γ^* while λ is equal to 0.2041. The same analysis for the variables extracted from the margin are reported in Figure 3.7. In both these figures we have followed the same division of the features used in Figure 3.4 and Figure 3.5 in order to make the comparison more easy. Thus, comparing Figures 3.6 and 3.7 with Figure 3.4 and Figure 3.5 respectively, it is possible to notice that an increase of the co-regularization term (λ) also induces the weights to be smaller. This is coherent with the fact that we are penalising the response of the different views, forcing them to be equal even if they are not by reducing the weights towards zero.



(a) Histrogram variables

(b) Shape variables

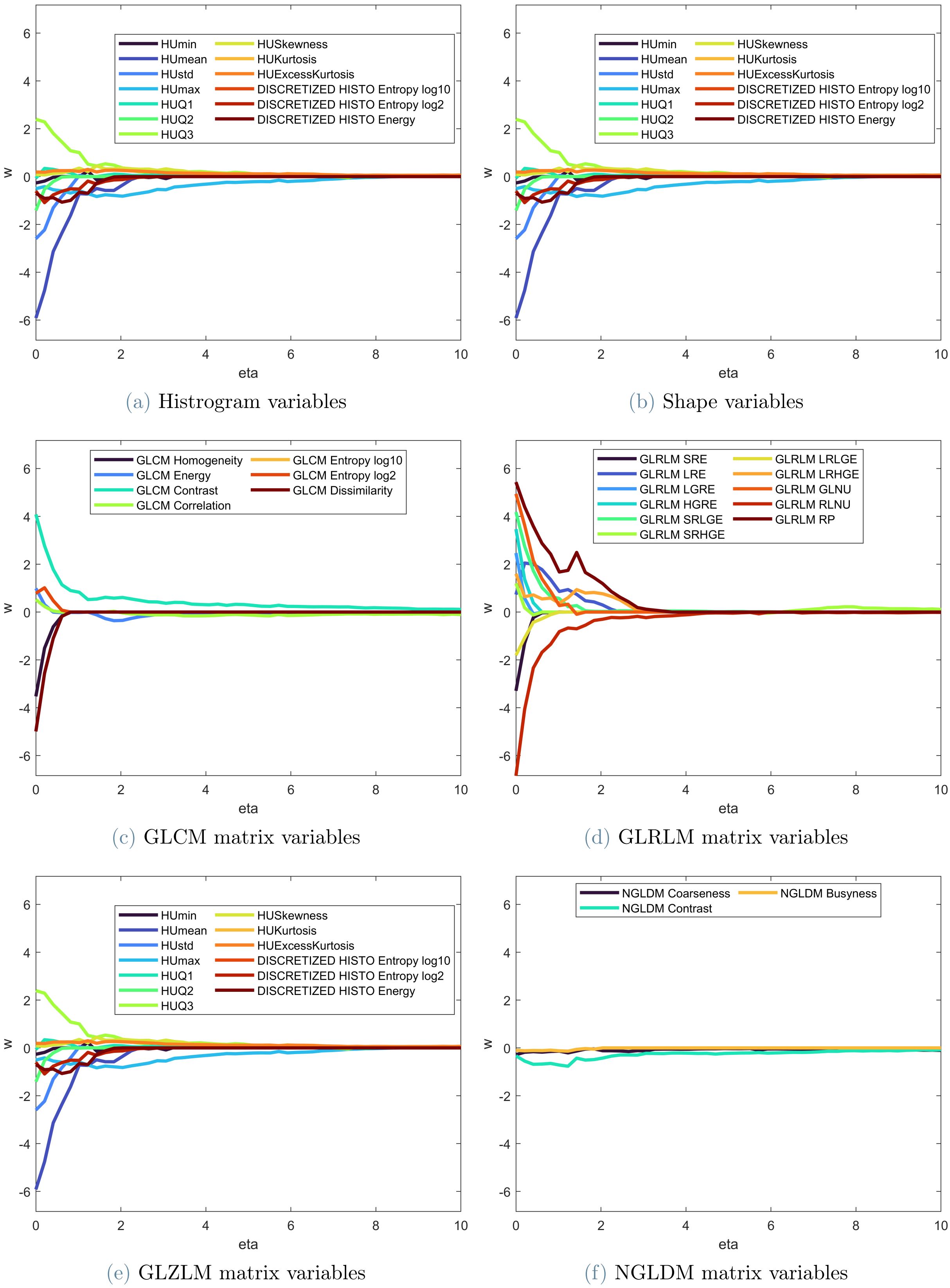
(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.4: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core once γ and λ have been fixed to their optimal value γ^* and λ^*



(a) Histrogram variables

(b) Shape variables

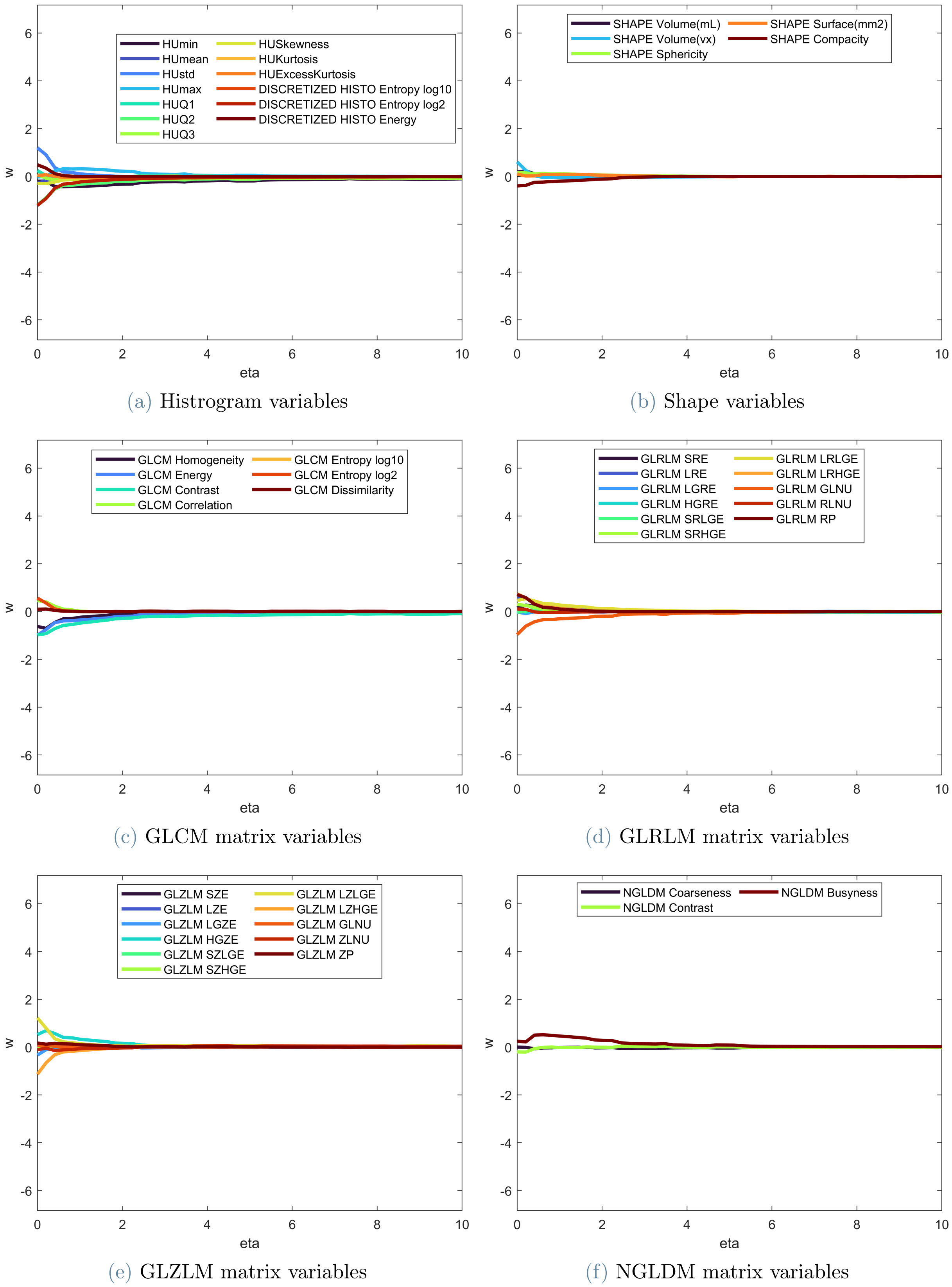
(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.5: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin once γ and λ have been fixed to their optimal value γ^* and λ^*



(a) Histrogram variables

(b) Shape variables

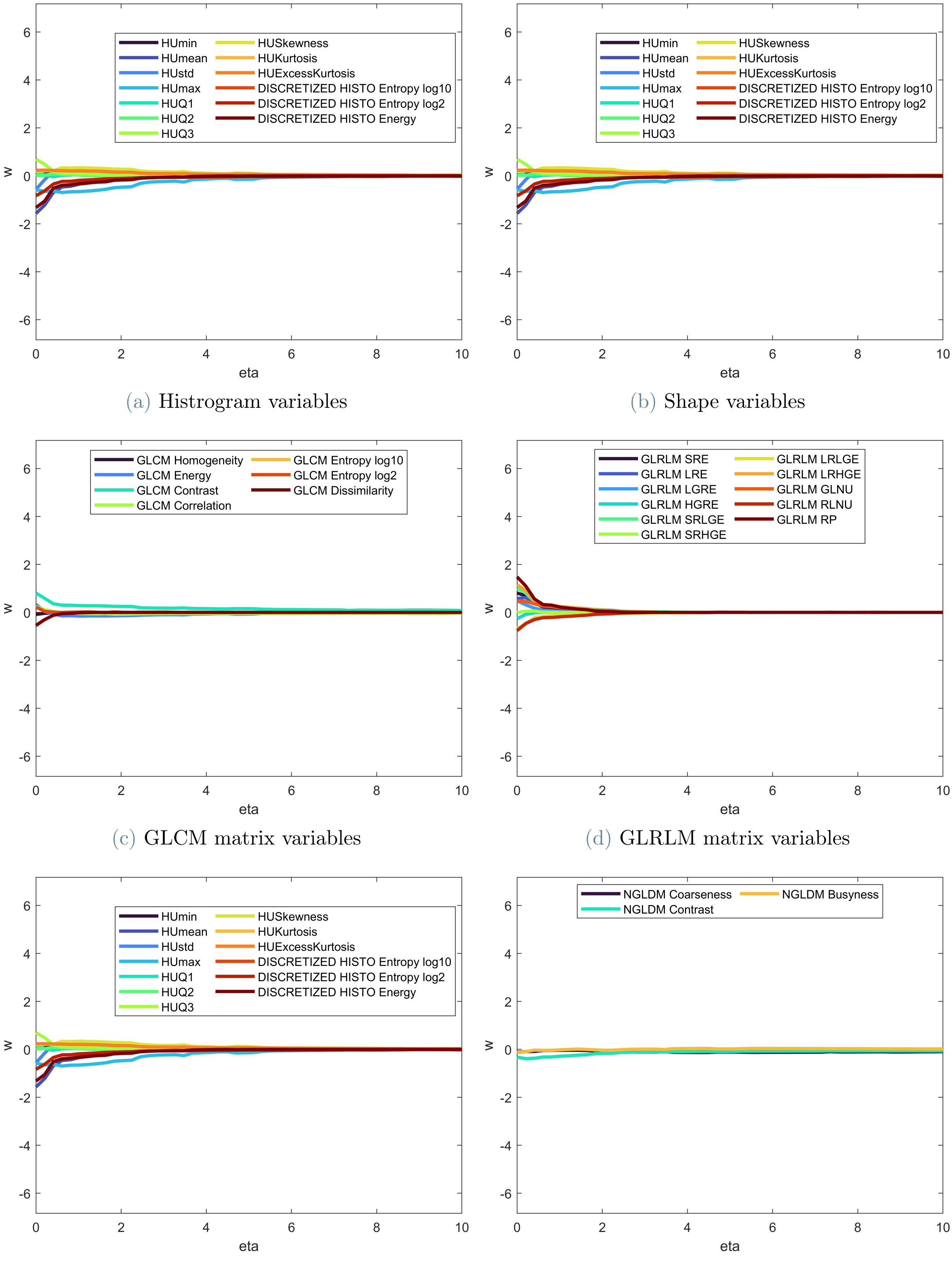
(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.6: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core once γ is fixed as γ^* and $\lambda = 0.2041$



(a) Histrogram variables

(b) Shape variables

(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.7: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin once γ is fixed as γ^* and $\lambda = 0.2041$

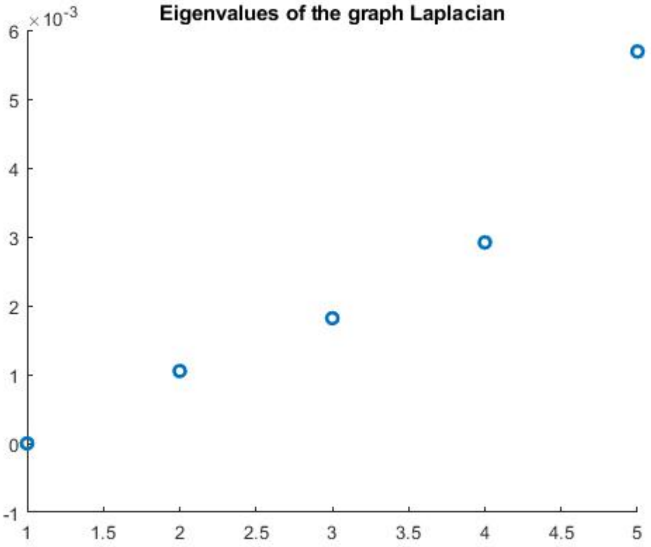


Figure 3.8: Values of the first five eigenvalues of the graph Laplacian

3.2.3. Spectral Clustering

Once the optimal parameters have been selected, the optimal patient-to-patient similarity graph is obtained, as described in Section 3.1.4, and it can be clustered. In order to cluster this graph in homogeneous population of nodes a spectral clustering technique has been applied. The number of clusters (nc) has been chosen following the eigengap heuristic [69]. As said in section 3.1.6, this technique consists in choosing nc such that the first nc eigenvalues are zeros, or at least small, whereas the $(nc + 1) - th$ one is respectively different from zero or relatively large. In Figure 3.8 are shown the values of the first five eigenvalues of the graph Laplacian obtained in this application. Since the first four eigenvalues are relatively small with respect to the fifth one, it is clear that the optimal value for the number of groups is 4. In this case there are no eigenvalue that are actually zero and for this reason the similarity graph is not divided in distinct subgraphs.

3.2.4. Feature importance

In this section we briefly explain how we have selected the most important features. We have created a ranking on the radiomic features using their weights w of the Cox model. In particular, we have studied for which value of the L1 penalization parameter η the different weights go to zero. As already seen in Section 3.2.2, these weights become null for different values of η. Therefore, we have ordered the radiomic features according to the value of η for which the corresponding weight becomes zero. The features for which this value of η is higher are the first ones in the ranking. In this way we have built a ranking on the radiomic features and this ordering has been used to present the result in the following Chapter.

3.2.5. Cluster characterization

In this section are explained the techniques used in order to characterize the four clusters found with the spectral clustering. We have clinically characterized these groups thanks to the clinical variables, exogenous to the model, and to asses if they are significantly different in the four groups we have performed two different tests according to the nature of the variables. In case the variable is numeric, as for *Age*, we have performed a non-parametric ANOVA test. Instead, in case of categorical variables we have used a proportion test. For all the test a p-value lower than 0.05 has been considered as significant and a Bonferroni correction for multiple testing has been used. For the survival assessment the Kaplan-Meier [71] overall survival probability curves have been employed. This method is a non-parametric method used to estimate the survival probability from observed survival times.

3.3. View assessment in three-phases CT imaging: Overall survival

This section concerns the second application of the S2GC model [18] to the data of the patients affected by ICC. All the patients for which we have the radiomic features extracted from all the three phases of the CT scans are employed. This leads us to 203 patients for which we have all the needed information. Since for each image the features have been extracted from both the core and the margin region it is possible to study a complete multi-view dataset where the features have been extracted not just from different regions of interest but also from different images. We have not excluded the patients for which there are some missing values on the clinical variables since these variables have not been employed for building the patient representation. The scope of this application was to establish whether the different views provide additional information in order to build a better stratification of patients affected by ICC. In particular, we analysed if the contribution of the three phases of the CT scans provide the same information or a different one.

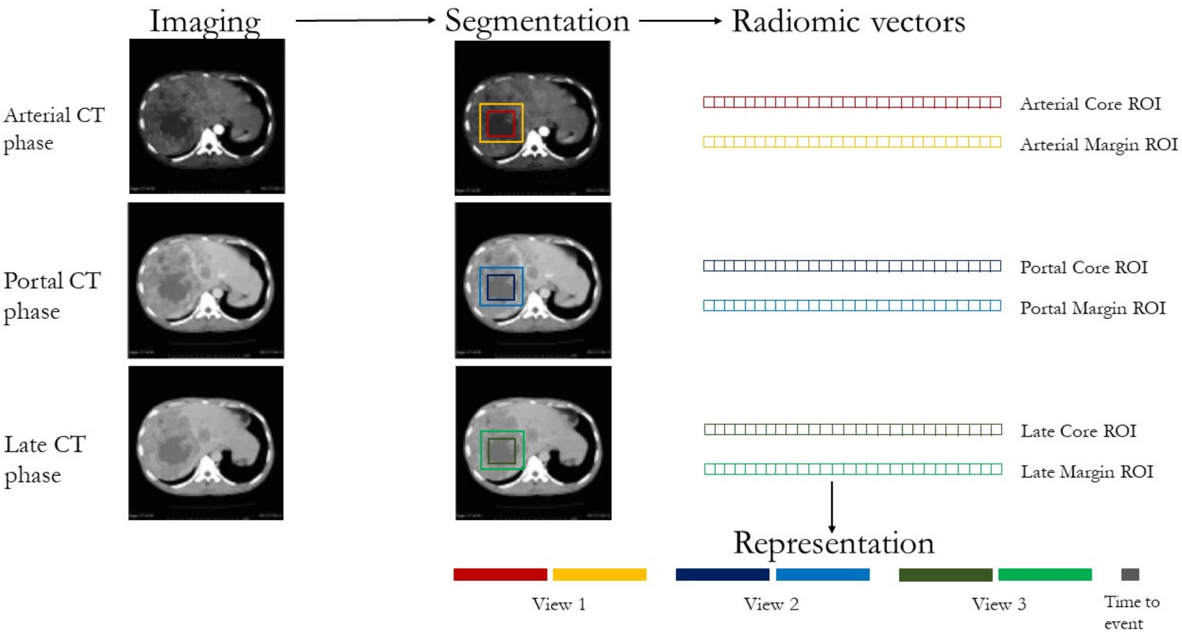


Figure 3.9: Sketch of the patient representation used in the S2GC model when all the three phases of the CT scans are used.

3.3.1. Construction of the multi-view patient representation

In this setting all the 300 radiomic features present in the dataset have been used. In Figure 3.9 is shown how these feature have been employed to build the patient representation. Starting from the the left of the figure, all the three images of the CT scans are used for this application and for each of them two regions of interest (core and margin of the tumor) have been segmented. From each of these regions, six in total, 50 radiomic features have been extracted for a total of 300 variables. These variables have been divided in three different views, where each view corresponds to the 100 features extracted from the two ROI of each phase. For this reason we are going to call the different views with the names of the different phases of the CT imaging. Recovering the notation of the S2GC model we have that for patient i, the vector X_i^k, where k is the $k-th$ view, is composed by 100 radiomic variables, 50 from the core and 50 from the margin. k in this case can be equal to 1, 2 or 3 which correspond respectively to the Portal, Arterial and Late phase of the CT scans. Also in this application only the radiomic features have been used to build the patient representation. Instead, all the clinical ones, as it will be shown in Chapter 4, have been used for the characterization of the clusters found thanks to the imaging features.

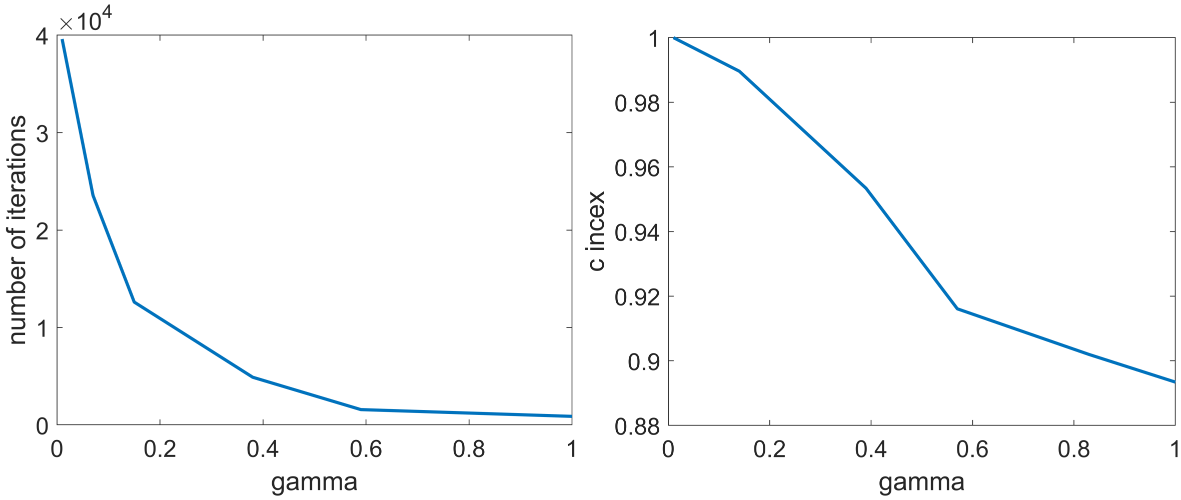


Figure 3.10: Left panel: the number of iterations needed by the S2GC algorithm to converge for the different values of γ. Right panel: the c-index variation according to γ. Both in the case where we are studying the overall survival using all the three phases of the CT scans.

3.3.2. Parameters Optimization

The procedure applied in order to find the optimal parameters is the same one described in Section 3.2.2. This consists in a grid search algorithm where we want to find the tuple of parameters that maximize the c-index [68].

In Figure 3.10 on the left is represented the number of iterations needed to the S2GC model to converge as function of γ, while on the right there is the variation of the c-index according to γ. From this figure it is possible to observe that the behaviour of both the number of iterations and the c-index as function of γ are similar to the one presented in the previous case. However the number of iterations needed in order to converge is now much bigger. Indeed, if before around 1500 iterations were needed to converge, now the iterations are between 30 and 40 thousands. This is due to the fact that before we were using just 2 views composed by 50 features each, for a total of 100 variables, while now the total number of features is 300. In this setting the value of γ that corresponds to the highest value of c-index is 0.01 and this value is selected as γ^*.

For what concerns λ and η we can refer to Figure 3.11. In this figure is represented how the c-index varies as function of η for three different fixed values of λ. The first two curves are decreasing as η increases while the last one (the yellow one) is almost constant. Looking at this figure it is clear that the line providing the best results is the blue one, that corresponds to $\lambda = 0$ and the maximum value in this line is achieved for $\eta = 0$.

Therefore, the optimal values are $\lambda^* = 0$ and $\eta^* = 0$ that correspond respectively to no co-regularization between the views and to no L1 penalization of the weights w of the Cox model. This means that the best result is achieved when all the information provided by the different features and the different views is exploited. Thus the information provided by the different views is not the same and all the features are in some way important. By penalizing some of the features there is a loss in performance. The fact that the views provide different information means that it is important to not force the response given by the different views to be the same, imposing a penalization thought λ.

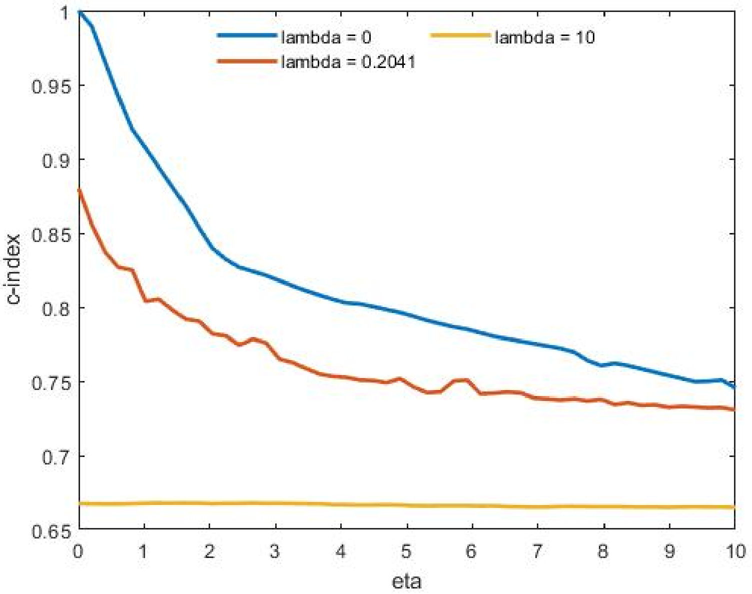
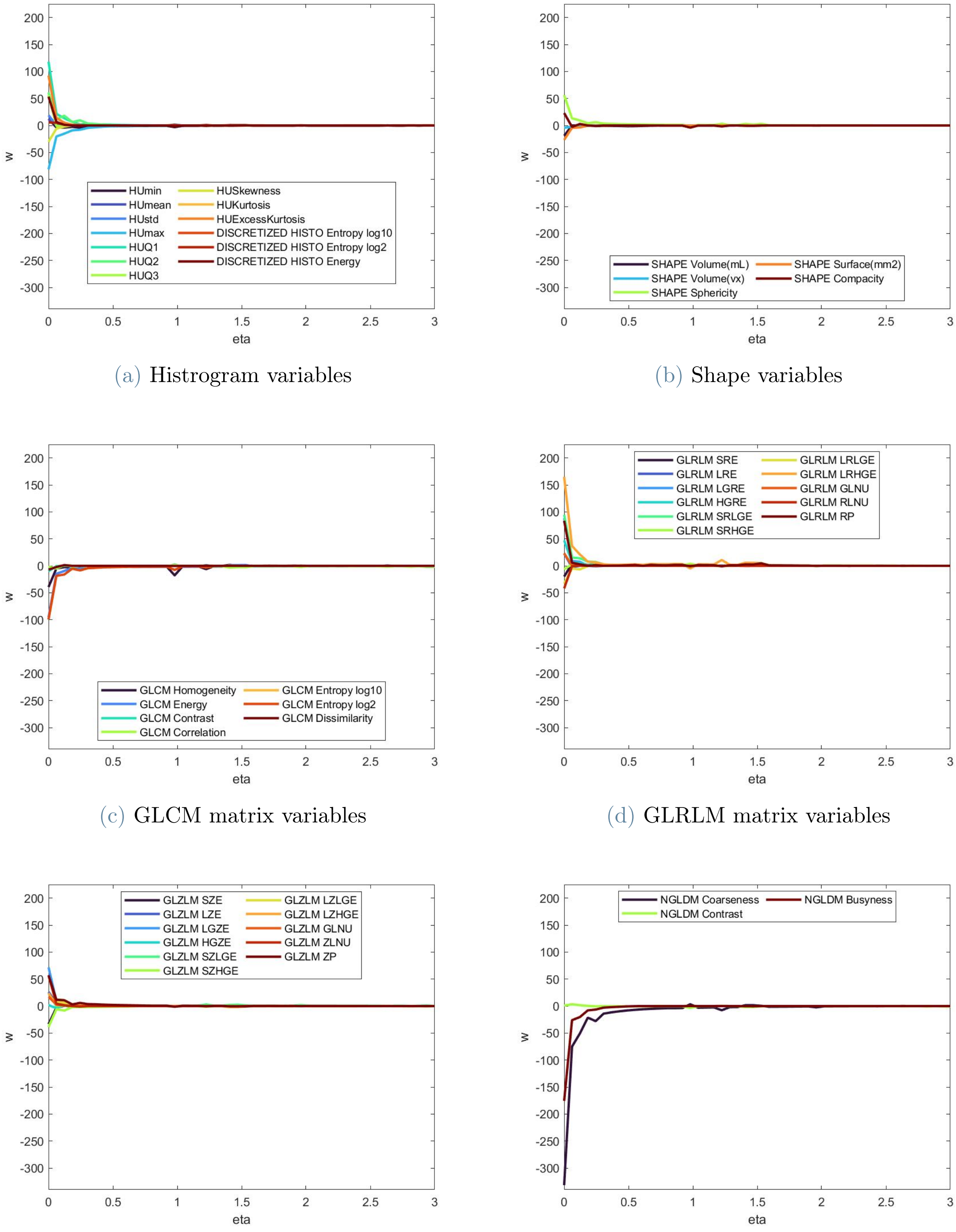


Figure 3.11: C-index variation as function of η for some fixed values of λ in the case where we are studying the overall survival using all the three phases of the CT scans.

Before proceeding with the choice of the number of clusters, some additional analysis on the impact of these parameters have been executed. The first one, as in the previous application, has the scope of studying how the weights w vary according to the different choices of η. In Figure 3.12 is shown how the weights of the radiomic features extracted from the core of the tumor of the Portal phase vary accordingly with the L1 penalization parameter η. In order to make this figure clearer, we have divided these features in six different groups. These groups are: a) Histogram derived variables, b) Shape derived variables, c) GLCM matrix derived variables, d) GLRLM matrix derived variables, e) GLZLM matrix derived variables and f) NGLDM matrix derived variables. In Figure 3.13 the same plots of Figure 3.12 are reported this time for all the features extracted from the margin region of the tumor, again for the Portal phase. What described for the

Portal phase has been applied also for the other two phases. In Figure 3.14 and Figure 3.15 are reported the plots of the weights relative to the Arterial phase respectively of the core and of the margin. For what concerns the Late phase we have reported the same analysis in Figures 3.16 and 3.17. Looking to these six figures it is possible to notice that, as expected, all the weights go to zero as η increases. So, at the end, all the weights become null, even though not all for the same value of η. Thanks to the fact that not all the weights go to zero with the same speed, we have built a ranking of the radiomic features. This procedure will be explained more in detail in the Section 3.3.4. Another important thing that can be noticed from these figures is that the weights for η^* are different in the diverse phases and also in the different ROI. The difference can be both in magnitude or in sign. For example there is a difference of magnitude between some of the weights relative to the shape variables in the core, generally small, and their counterparts in the margin region (bigger) of the Arterial phase. An example of difference in sign can be seen for the *NGLDM_ Coarseness* that is negative for the core and positive in the margin of the Portal phase.



(a) Histrogram variables

(b) Shape variables

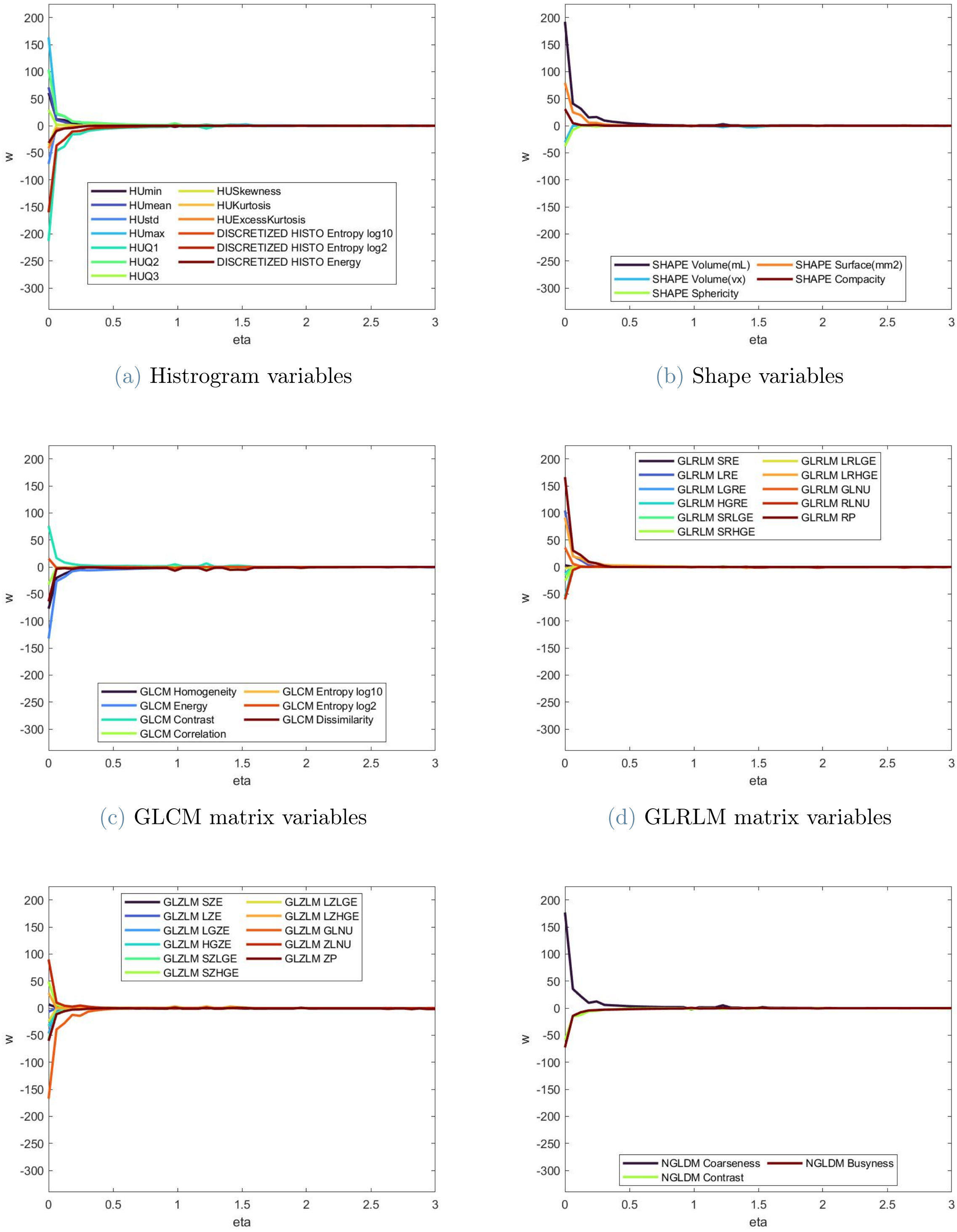
(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.12: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core of the Portal phase once γ and λ have been fixed to their optimal value γ^* and λ^*



(a) Histrogram variables

(b) Shape variables

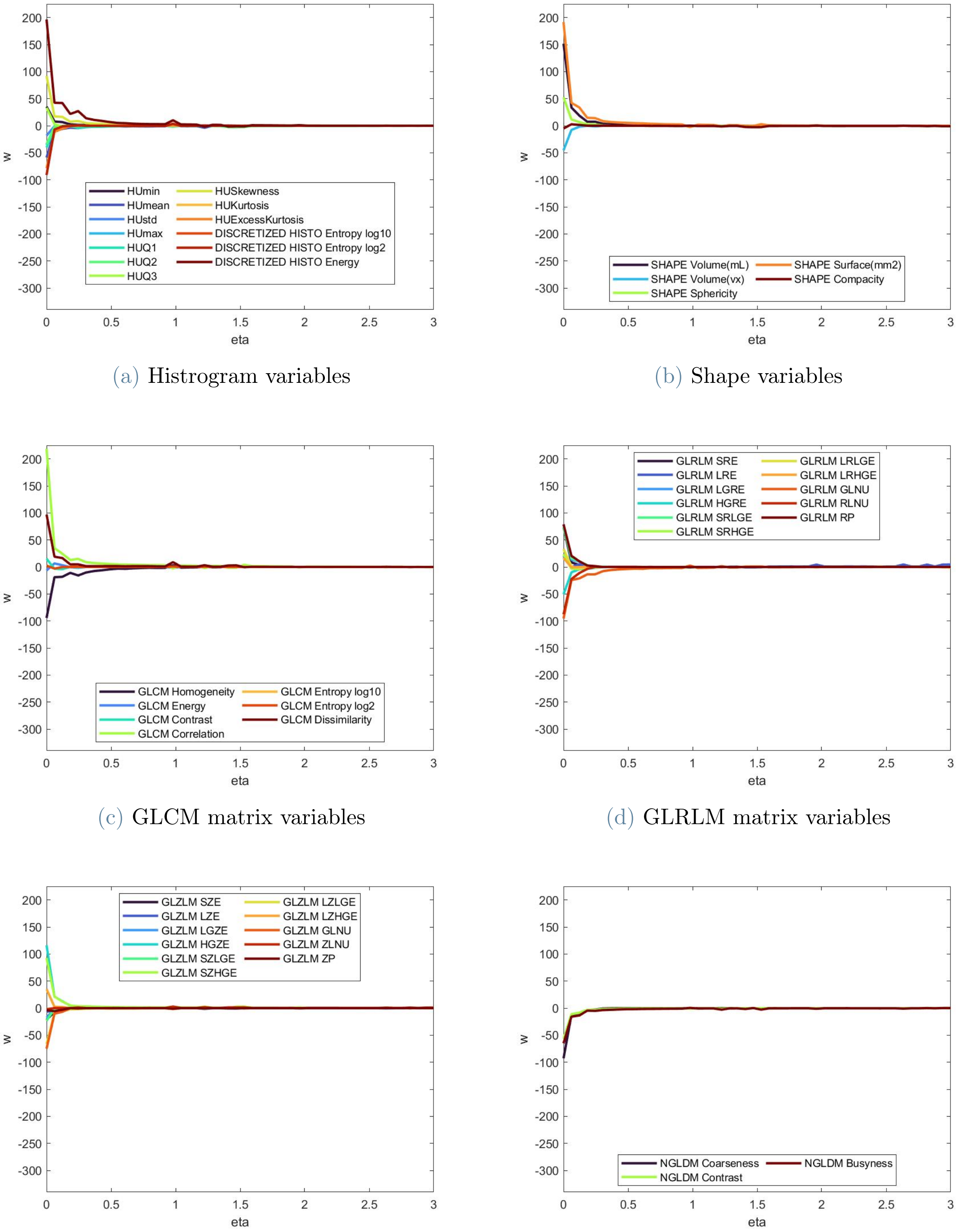
(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.13: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin of the Portal phase once γ and λ have been fixed to their optimal value γ^* and λ^*



(a) Histrogram variables

(b) Shape variables

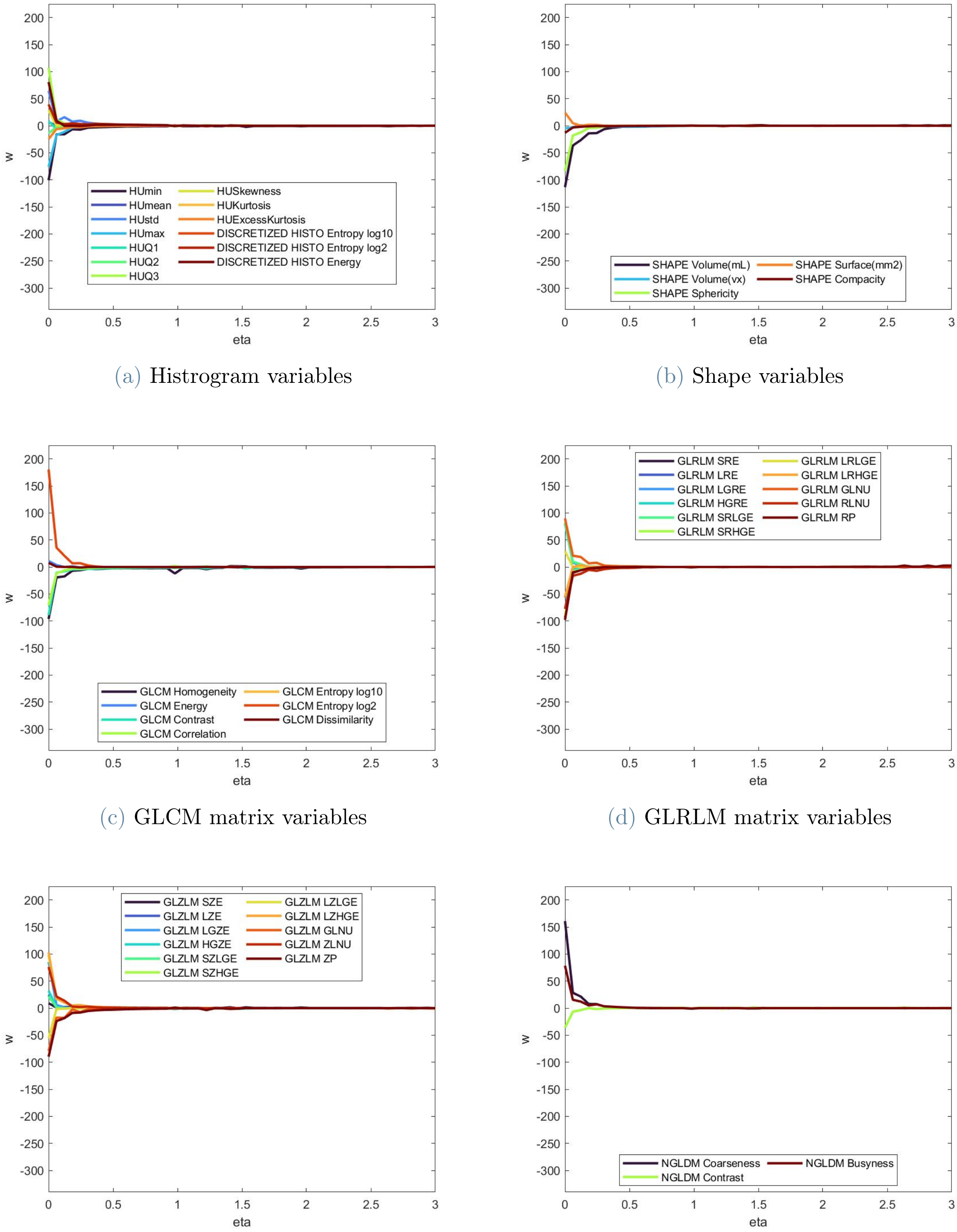
(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.14: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core of the Arterial phase once γ and λ have been fixed to their optimal value γ^* and λ^*



(a) Histrogram variables

(b) Shape variables

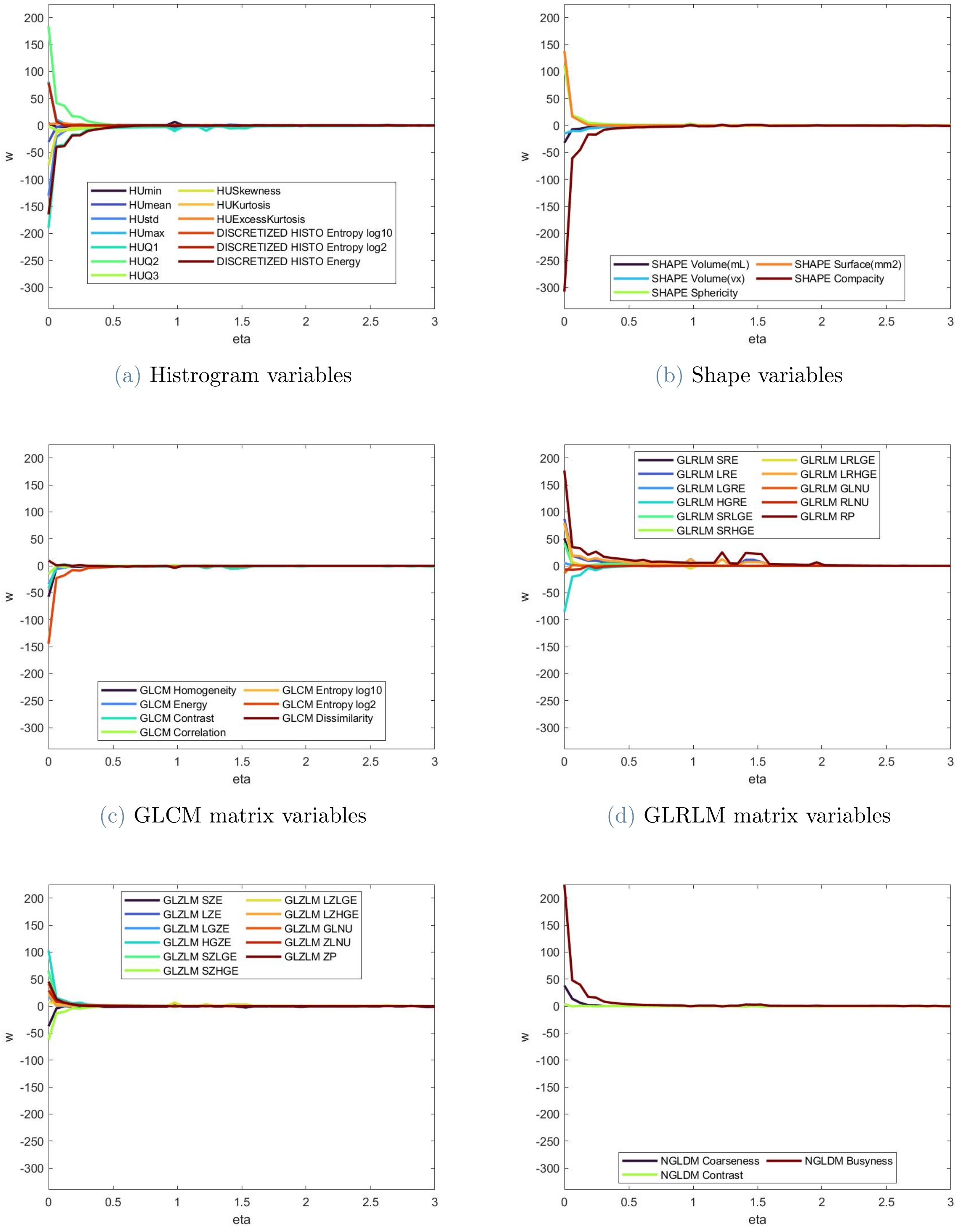
(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.15: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin of the Arterial phase once γ and λ have been fixed to their optimal value γ^* and λ^*



(a) Histrogram variables

(b) Shape variables

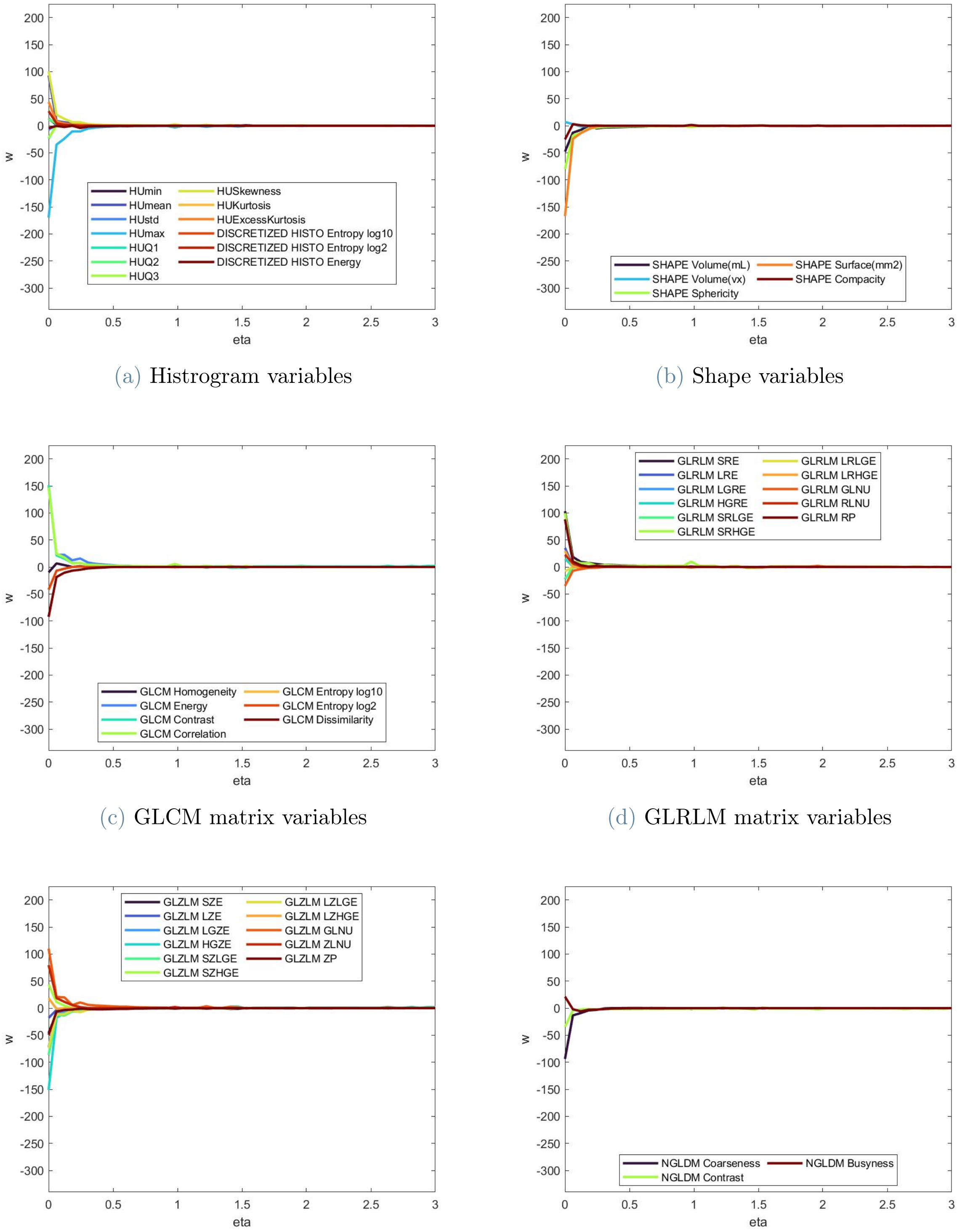
(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.16: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core of the Late phase once γ and λ have been fixed to their optimal value γ^* and λ^*



(a) Histrogram variables

(b) Shape variables

(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.17: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin of the Late phase once γ and λ have been fixed to their optimal value γ^* and λ^*

The second study that has been performed was on the patient-to-patient similarity graph and in particular on how it varies according to the penalization terms η and λ. To do this we have reported in Figure 3.18 the patient-to-patient similarity graph obtained for different values of λ, while γ and η are fixed to their optimal values. In particular, starting from the left and going to the right we have a decrease in the co-regularization term. Observing these figure, we can see that the model with the smallest penalization, the one on the right, is the one that has the richest information. In fact, for high values of λ all patients are connected to each other obtaining a random graph, while, decreasing the value of the penalization, the patients (nodes) start to rearrange in an ordered way and detach in groups. This procedure ends with $\lambda = 0$, that corresponds to no penalization between the views, for which 5 different subgraphs are individuated. Every subgraph is intended as a group of patients with similar prognosis and radiomic description. In fact, we are interested in finding groups of patients that share similar properties. In particular, we want to find groups composed by patients with similar prognosis and distinguish groups that have the same survival curves but different radiomic characterization. For this reason we have employed the S2GC model that uses both the survival analysis and the radiomic features in order to build the patient-to-patient similarity graph. A similar analysis has been carried out also for η, when instead it is λ that is fixed to its optimal value, and the results are shown in Figure 3.19. Also in this case going from left to right the penalization parameter is decreased to zero and the respective patient-to-patient similarity graph is reported. In this case the value of η does not need to be 0 since the detachment is evident also for other values. This is due to the fact that with η we are penalizing the single feature and losing only the information provided by the less important ones. Instead, with λ all the features of the view are penalized and for this reason the information that is lost is much more. For this reason smaller values of λ are needed in order to see the detachment between the subgraphs compared to the values needed for η. Therefore looking at Figure 3.19 it is possible to notice that we could have chosen any value of η small enough, although the choice fell on $\eta = 0$ according to the performance optimization previously performed.

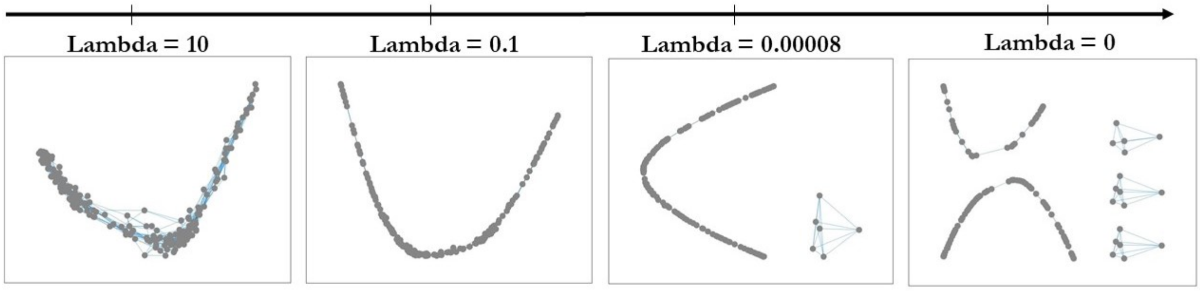


Figure 3.18: Patient-to-patient similarity graph variation according with λ while γ and η are fixed to their optimal values

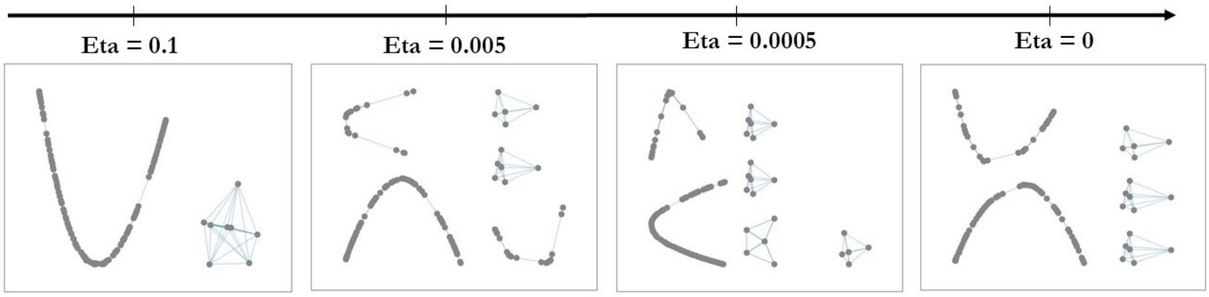


Figure 3.19: Patient-to-patient similarity graph variation according with η while γ and λ are fixed to their optimal values

3.3.3. Spectral Clustering

After having selected the optimal parameters needed by the S2GC model [18], the procedure adopted is the same of the case with only the Portal phase. In Figure 3.20 are reported the first nine eigenvalues of the graph Laplacian. The first five are equal to zero, while the other four are small but different from zero and increasing. For this reason the eigengap heuristic [69] suggests 5 as optimal number of clusters. Since in this case the first five eigenvalues are zero, and not just small, the graph, as can be seen in Figures 3.18 and 3.19 on the right, is divided in five distinct subgraphs.

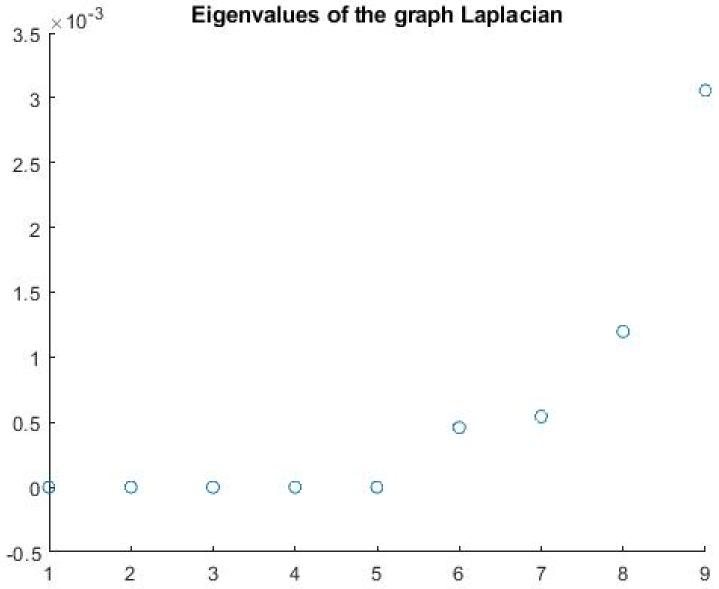


Figure 3.20: Values of the first nine eigenvalues of the graph Laplacian

3.3.4. Feature importance

In this section we explain how the selection of the most important features has been performed. As anticipated in section3.3.2 we have used the analysis done on the weights of the Cox model to build a ranking of the radiomic features. In particular, we have studied the speed with which the different weights go to zero when η is increased. We have assumed that the most important variables are the ones that are more persistent and less prone to go to zero when the L1 regularization parameter is increased. So we have built a ranking on the radiomic features where the ones considered as most relevant were the ones that correspond to the weights that go to zero for the higher values of η. This ranking has been used in order to present the result in the following Chapter.

3.3.5. Cluster characterization

Once the patient-to-patient similarity graph has been estimated and divided in homogeneous clusters, we were interested in describing these groups both for the survival analysis and for their clinical characterization. For what concerns the survival assessment, the Kaplan-Meier [71] overall survival probability curves have been employed. This is a non-parametric method used to estimate the survival probability from observed survival times. Instead, for the clinical characterization some tests have been performed according to the nature of the clinical variables, exogenous to the model, in order to establish which are significantly different in the five groups. In case the variable is numeric, as for *Age* we have performed a non-parametric ANOVA test. Instead, in case of categorical variables we have used a proportion test. For all the tests a p-value lower than 0.05 has been considered as significant and a Bonferroni correction for multiple testing has been used. Since for this application we are also interested in studying more thoroughly the radiomic features, we have performed some tests also on these variables in order to find which ones are significantly different in the five groups. In particular, also in this case we have performed a non-parametric ANOVA test.

3.4. View assessment in three-phases CT imaging: Recurrence

The structure of this section is similar to the previous ones, especially to the last one, where all the phases of the CT scans have been used. The main difference between this application and the other two is that now the T_i that has been used is not any longer the overall survival time, but the disease free survival time. The inclusion/exclusion criteria are almost the same. The only difference is that, while there were no missing values on the Overall survival, for what concerns the Recurrence, there are 10 patients for which we don't have this datum. For this reason the patients included in this analysis are only 193. Also in this setting the complete multi-view version of the dataset has been employed, giving in this way the possibility to further investigate the contribution of the different phases of the CT scans. This allowed us also to compare this model with the one presented in the previous application for the Overall survival, both in terms of performance and of clinical messages.

3.4.1. Construction of the multi-view patient representation

The patient representation is the same as the one described in section 3.3.1 where all the three phases of the CT scans are used to build a multi-view representation. We recall that the multi-view representation was built thanks to the use of all the three images of the CT scans where each one of them forms a view. For this reason each view takes its name from the respective phase of the CT imaging and it is composed by 100 variables, 50 from the core and 50 from the margin. The only difference between this case and the one described in Figure 3.9, is in the time used to perform the survival analysis, since this time our interest is in the disease free survival time.

3.4.2. Parameters Optimization

To chose the best parameters a grid search approach has been implemented and the parameters that return the highest c-index are selected as optimal.
First of all, we can start by analysing γ. The behaviour of both the number of iterations needed to converge and the c-index are reported in Figure 3.21. In fact, on the left there is a plot of the number of iterations needed to the model to converge for the different values of γ. Instead on the right we report the graph of the c-index, again, as function of the learning rate. As in the previous cases, both the number of iterations and the c-index decrease when γ increases. We select $\gamma^* = 0.01$ since it is the value of γ that corresponds to the highest value of c-index, as it was in the previous application. It is also possible to notice that the number of iterations needed to converge is similar to the previous case and this is coherent with the fact that the number of variables used in these two cases is the same.

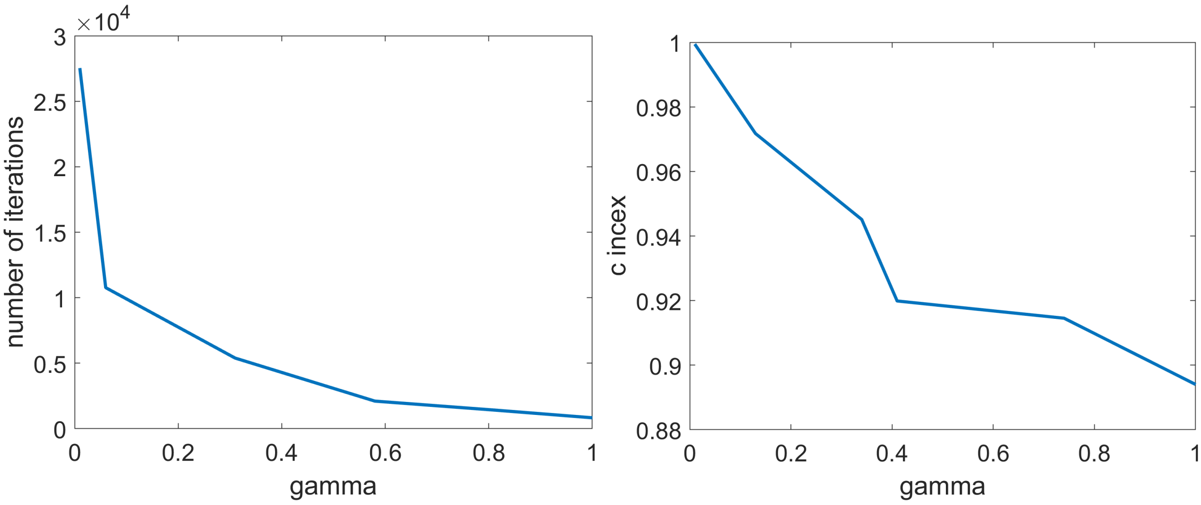


Figure 3.21: Left panel: the number of iterations needed by the S2GC algorithm to converge for the different values of γ. Right panel: the c-index variation according to γ. Both in the case where we are studying the disease free survival time using all the three phases of the CT scans.

For what concerns λ and η we have reported the results of the grid search in Figure 3.22. In this figure there are represented three different curves: blue, orange and yellow, that correspond to three different values of λ, respectively to 0, 0.2041 and 10. Each curve shows how the c-index varies as function of η. Therefore looking at the three different curves it is clear that the one that achieves the best result is the blue one. Indeed, for all the values of η this curve is higher than the other two.
For what concerns η this line is not sufficient and we need to look at Figure 3.23 too,

where a zoom of this curve is reported. Thanks to the fact that in this figure the results are reported only for some very small values of η, it is possible to analyse better which is the optimal choice of this parameter. In this case the choice falls on $\eta = 0.005$ that is the value for which we have the maximum value of c-index. Therefore in this case the optimal values are $\lambda^* = 0$ and $\eta^* = 0.005$ that correspond to a zero co-regularization between the views and a small L1 penalization of the weights w of the Cox model. This is consistent with what we have already found in the previous case. We can deduce that also in this case the three views provide different information and forcing them to produce the same risk prediction is an error that causes a big loss of information. A very small penalization of the L1 norm is needed, which means that the best results are achieved when the less informative variables are a little constrained.

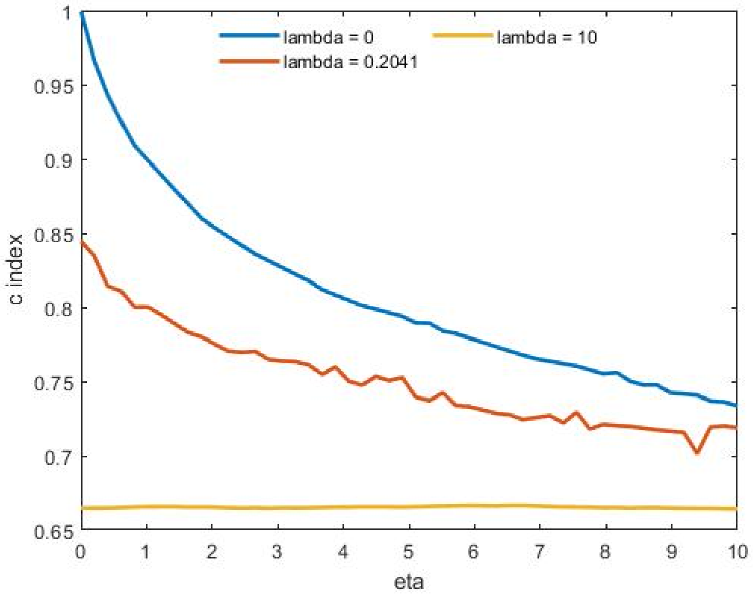


Figure 3.22: C-index variation as function of η for some fixed values of λ in the case where we are studying the disease free survival using all the three phases of the CT scans.

As for Section 3.3.2, also here some additional analysis on the implications of the choice of the parameters have been made. The first one has the scope of studying how the weights w of the Cox model [36] vary according to the different choices of η. Instead the second is performed to study the patient-to-patient similarity graph as function of λ and η. To complete the first study we have reported the weights as function of η in some figures and to make them clearer we have divided the features, to which the weights refer, in six different groups. So in Figure 3.24 six different plots are reported, each one for one of the six different groups of radiomic features extracted from the core of the tumor in the Portal phase. In Figure 3.25 the same is done for the variables extracted from the

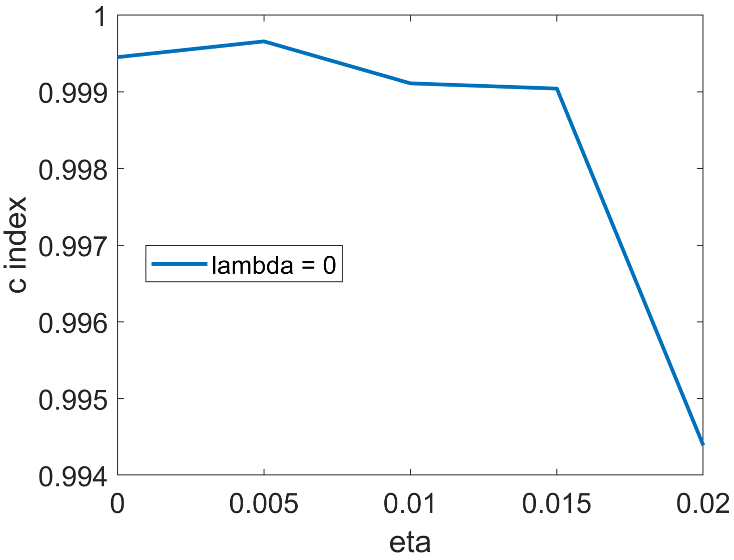
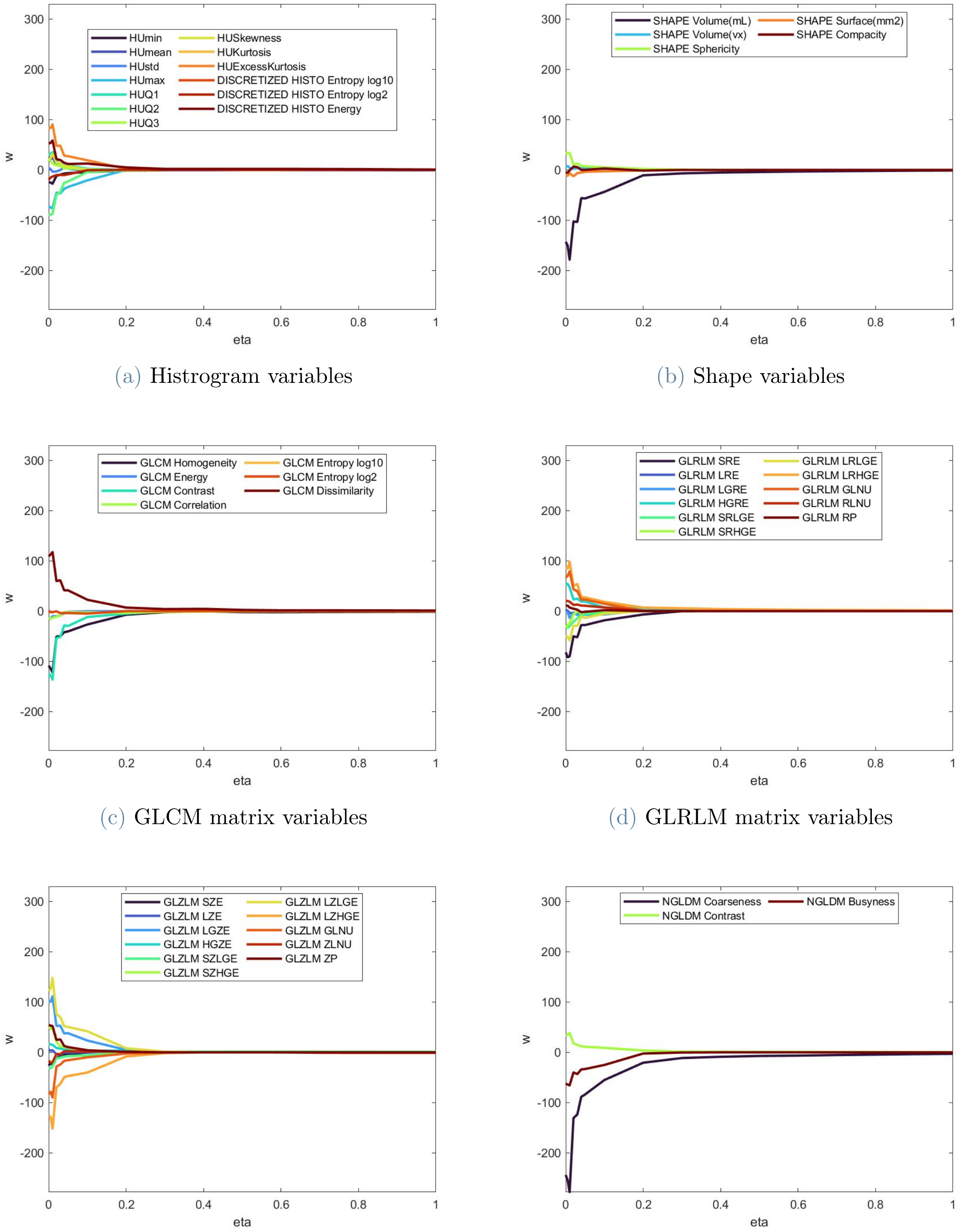


Figure 3.23: Zoom of the blue line, that corresponds to $\lambda = 0$, for small value of η. It is useful to study the optimal value of this parameter

margin. Comparing these two plots we can see that there are some weights that have really different values for the core and for the margin. For example looking at plot b) and f), that correspond respectively to the shape and to the NGLDM matrix derived variables, it is possible to notice that there are weights of some features that are negative for the core and positive for the margin or the opposite. While focusing on the histogram derived variables, it is possible to see that there are some weights that are much more bigger in the margin than in the core region. This will be seen more in detail in the following chapter were we will present a more clear comparison between some of the most important weights. In Figure 3.26 and Figure 3.27 the same plots of the weights are reported respectively relative to the core and to the margin of the Arterial phase. The weights of the Late view relative to the core can be seen in Figure 3.28, while in Figure 3.29 the ones extracted from the margin are shown. Thanks to the fact that the scale of all these figures is the same, it is possible to make an easy comparison between the weights of the different phases. Therefore one can notice that the weights relative to the same radiomic features can be really different in the three views. Furthermore, as expected, it is clear that all the weights go to zero when the L1 regularization parameter (η) increases, but with different speeds. In fact, this can be seen in the plot b) of Figure 3.24 where it is clear that the weights relative to the shape variables go to zero for different values of η. Indeed the weight relative to *SHAPE_ Volume(ml)* goes clearly to zero more slowly than the others. A similar behaviour can be noticed also in the others figures for the different views and ROI.



(a) Histrogram variables

(b) Shape variables

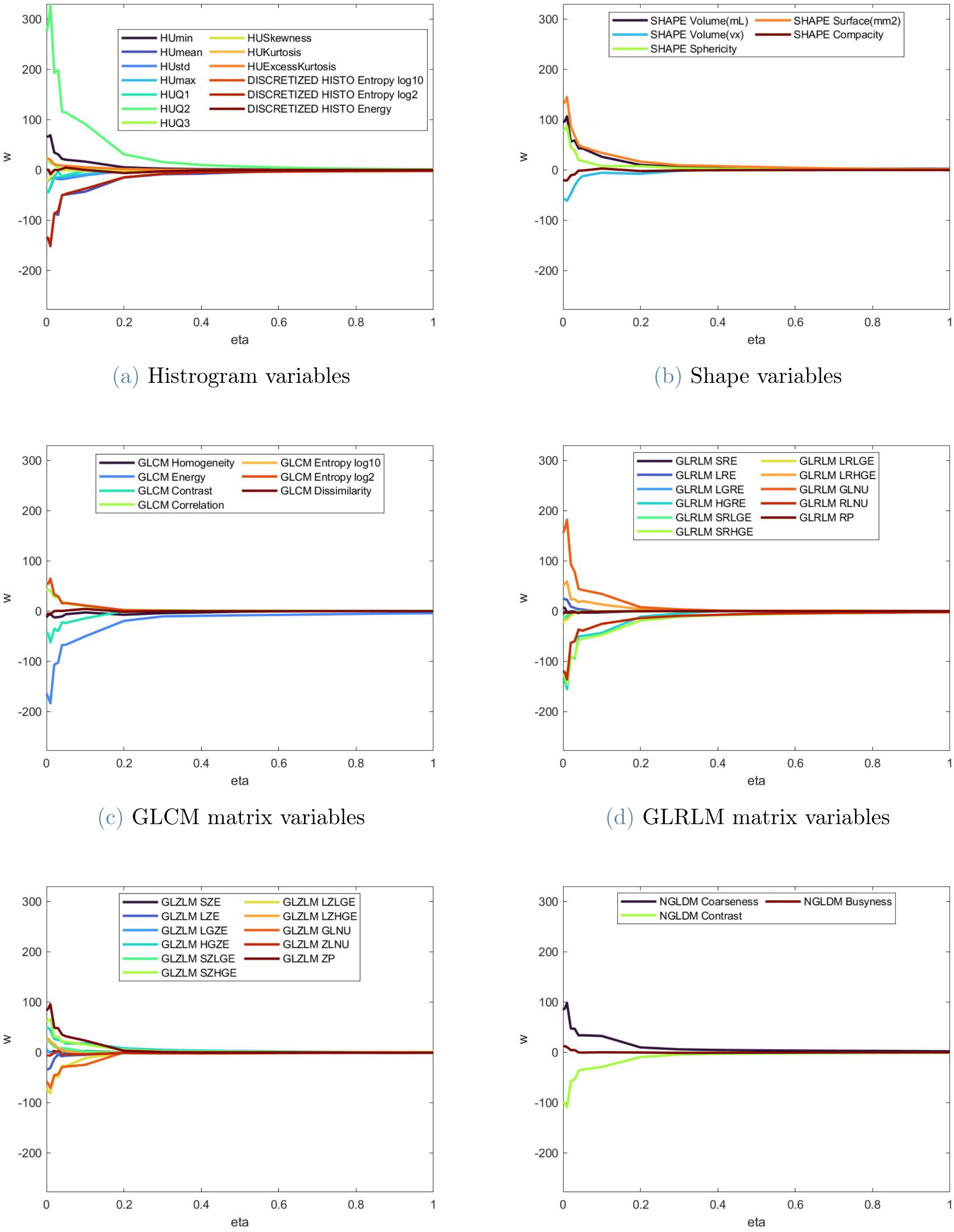
(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.24: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core of the Portal phase once γ and λ have been fixed to their optimal value γ^* and λ^*



(a) Histrogram variables

(b) Shape variables

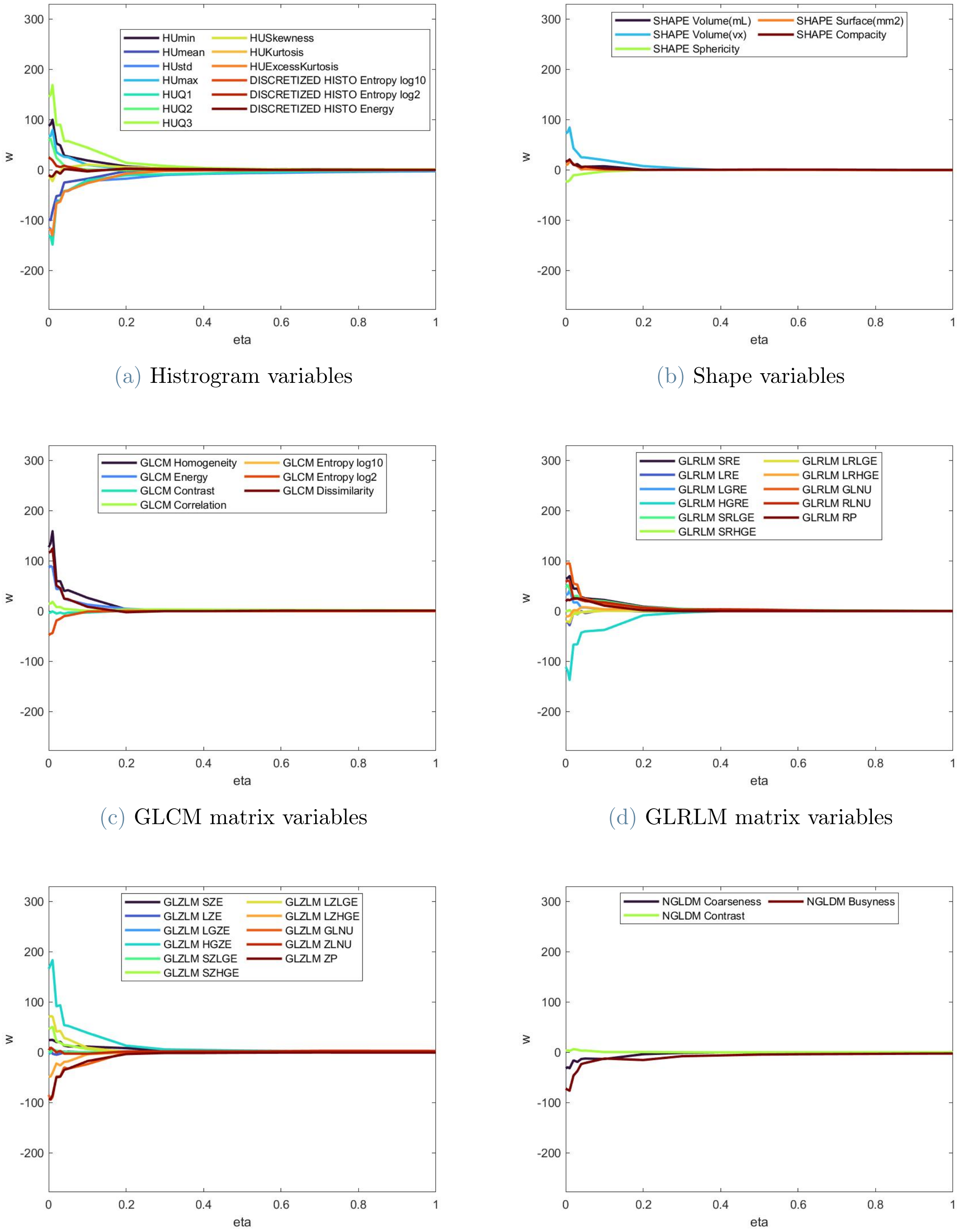
(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.25: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin of the Portal phase once γ and λ have been fixed to their optimal value γ^* and λ^*



(a) Histrogram variables

(b) Shape variables

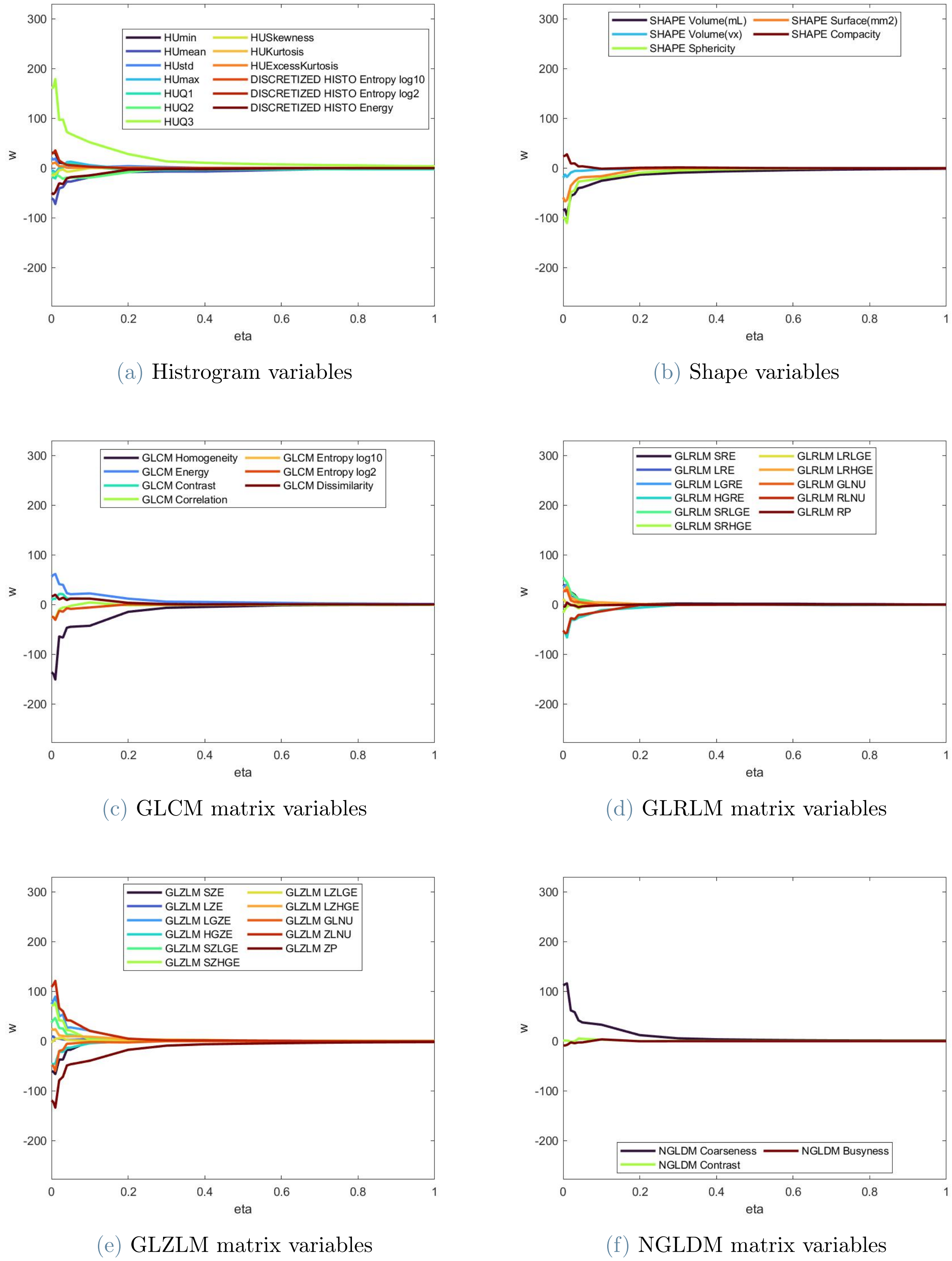
(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.26: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core of the Arterial phase once γ and λ have been fixed to their optimal value γ^* and λ^*



(a) Histrogram variables

(b) Shape variables

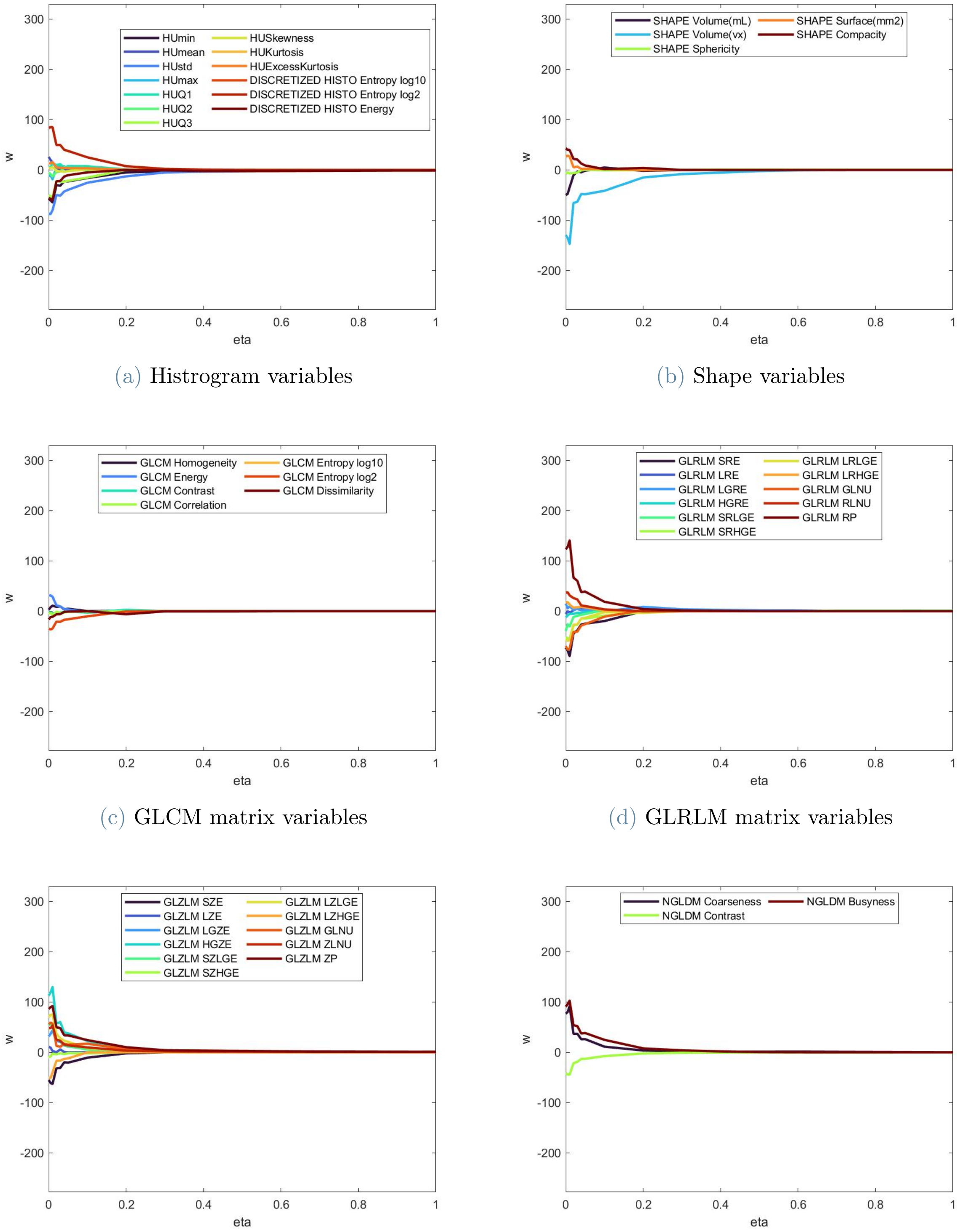
(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.27: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin of the Arterial phase once γ and λ have been fixed to their optimal value γ^* and λ^*



(a) Histrogram variables

(b) Shape variables

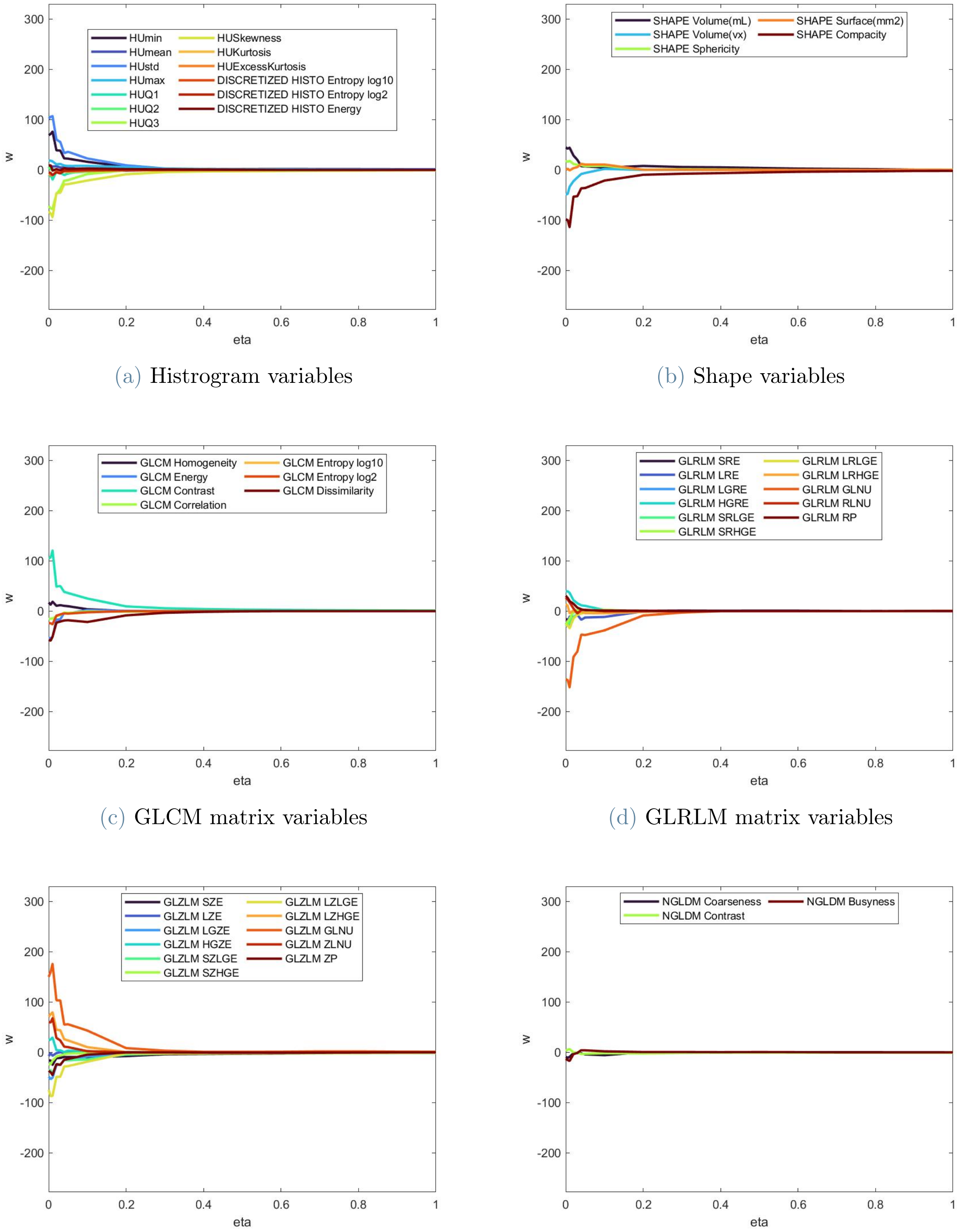
(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.28: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core of the Late phase once γ and λ have been fixed to their optimal value γ^* and λ^*



(a) Histrogram variables

(b) Shape variables

(c) GLCM matrix variables

(d) GLRLM matrix variables

(e) GLZLM matrix variables

(f) NGLDM matrix variables

Figure 3.29: Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin of the Late phase once γ and λ have been fixed to their optimal value γ^* and λ^*

The second study is instead performed to analyse how the choices of η and λ impact on the patient-to-patient similarity graph. For this reason we have reported the graphs obtained for the different values of λ, while η is fixed to η^*, in Figure 3.30. In particular, going from the left to the right, decreasing values of λ with the corresponding graph are reported. From this figure it is clear that only the model with $\lambda = 0$ has enough information in order to distinguish the groups and divide them into different subgraphs. In fact, also for a small penalization all the patients are connected to each other in a random graph. For what concerns η a similar plot has been reported in Figure 3.31. The structure of the figure is the same as the one shown for λ, but here only three values of η are described and it is λ that is fixed to its optimal value. In this case it is clear the it is not necessary to have $\eta = 0$ in order to have some disjoint subgraphs. Indeed, already for $\eta = 0.1$ there are three disjoint subgraphs and for the optimal value (0.005) four different groups can be seen.

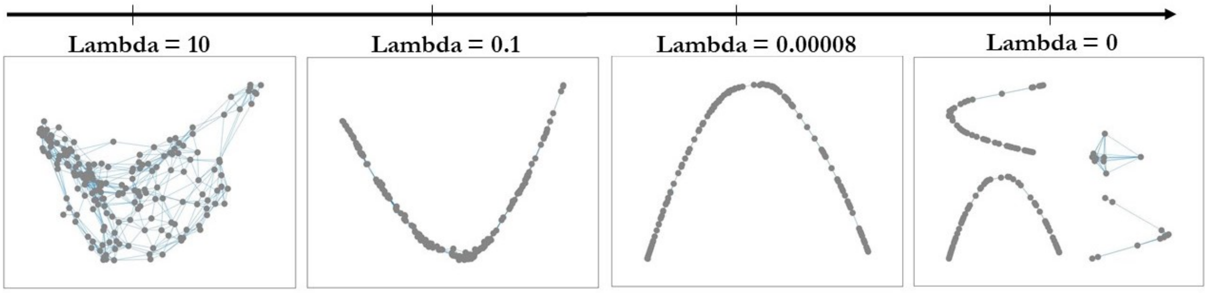


Figure 3.30: Patient-to-patient similarity graph variation according with λ

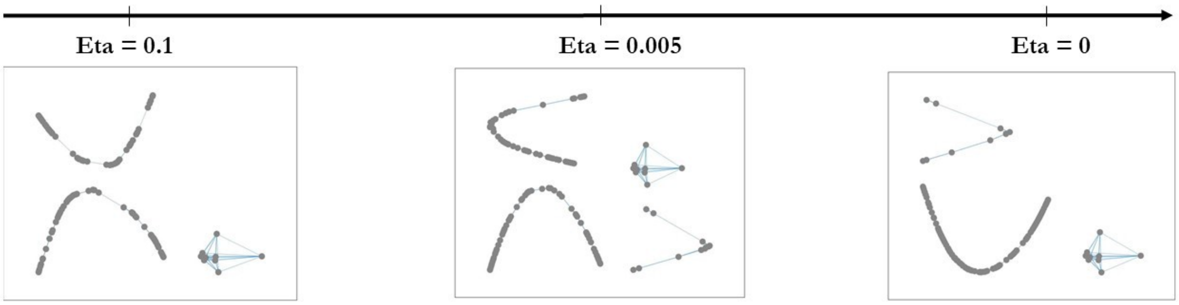


Figure 3.31: Patient-to-patient similarity graph variation according with η

3.4.3. Spectral Clustering

Once the patient-to-patient similarity graph is computed with the optimal parameters, it has to be divided in homogeneous clusters of patients. To do this an eigengap heuristic has been used. This technique states that the optimal number of clusters to chose is equal to the number of null eigenvalues of the graph Laplacian. In case there are no eigenvalues that are null, the number of clusters (nc) to choose is the one for which the first nc eigenvalues are small and the $(nc+1)-th$ one is relatively large. The eigenvalues of the graph Laplacian found in this application are reported in Figure 3.32. Looking at this figure it is possible to notice that there are four eigenvalues that are equal to zero and so the optimal number of clusters for this case is four. Having four eigenvalues equal to zero means that the graph is divided into four distinct subgraphs. This is in line with what already seen in Figure 3.30 and Figure 3.31, where, in correspondence of the optimal parameters, four separated subgraphs can be noticed.

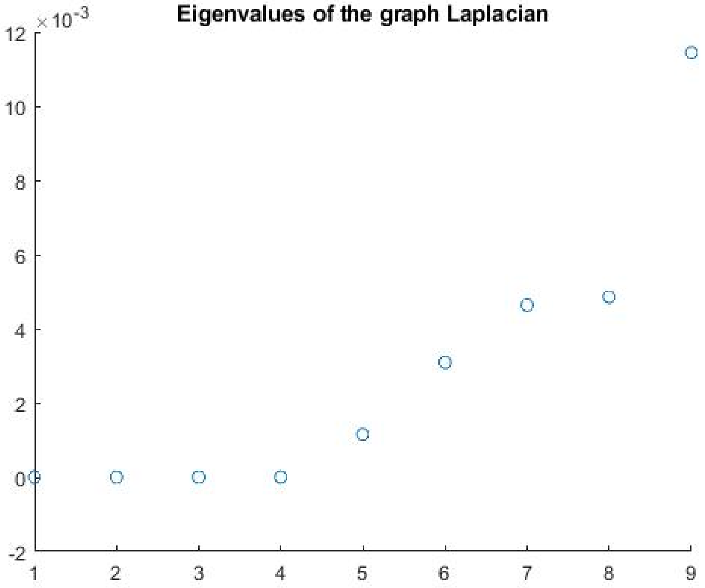


Figure 3.32: Values of the first nine eigenvalues of the graph Laplacian

3.4.4. Feature importance

In this section we explain how the most important features have been selected. This method is based on the analysis performed on the weights w of the Cox model. In particular, on the speed with which the different weights go to zero when the L1 penalization parameter is increased. Indeed, as discussed in section 3.4.2, not all the weights go to zero with the same speed. Thanks to this we have built a ranking on the radiomic features, assuming that the ones with the weights that go to zero for the higher values of η are the more relevant. This ranking has been used in order to present the results in the following Chapter.

3.4.5. Cluster characterization

Once the patient-to-patient similarity graph has been estimated and divided into homogeneous clusters, we are interested in finding a good characterization of these groups. First of all, we are interested in the survival assessment, which has been done using the Kaplan-Meier [71] disease free survival probability curves. This method is a non-parametric method used to estimate the survival probability from observed survival times. Moreover we are interested in studying the clinical variables, that are exogenous to the model. To do this we have used different statistical tests according to the nature of the variable. For a numeric variable a non-parametric ANOVA test has been performed, while for a categorical variable a proportion test has been employed. This last kind of test is used to establish if the proportion of patients having a certain clinical characteristic is the same or not in the different groups. For all the tests a p-value lower than 0.05 has been considered as significant and a Bonferroni correction for multiple testing has been used. Since for this application we are interested also in studying more thoroughly the radiomic features and finding which ones are significantly different, in the four groups we have performed some tests also on these variables. In particular, also in this case we have performed a non-parametric ANOVA test.

3.5. General messages

In this chapter we have discussed in detail the S2GC model [18] that we have employed to perform the cancer subtyping. Together with this model we have also presented the three applications of this model the data of the patients affected by ICC. From these applications we have learned that in all the three cases the optimal value of the parameter that controls the co-regularization between the views (λ) is 0. This is a first evidence that in every case the different views provide different and complementary information. Therefore, depending on the application, we have a sign of the importance of the two ROIs (core and margin) or of the three phases (Arterial, Portal and Late). We have also noticed that, as expected, all the weights w of the Cox model goes to zero when the L1 penalization parameter η is increased but, that there are some weights that go to zero later than others. For the application where only the radiomic features extracted from the core have been employed we have seen that the optimal number of cluster is 5 even if the S2GC model is not able to divide the patient-to-patient similarity graph in 5 disjoint subgraphs. For the other two applications, instead, we have noticed that the patient-to-patient similarity graph correspond to 4 and to 5 detached subgraphs. For this reason the optimal number of groups were 4 and 5 for the case of the overall survival and the one of the disease free survival respectively. We have also presented how we have tested the clinical variables, exogenous to the model, and how we have built a ranking on the importance of the radiomic features. Both useful to have a better comprehension of the results presented in the next chapter. Indeed, in Chapter 4 we will present the groups found with the S2GC algorithm with their characterization both from a clinical and an imaging point of view. Then, in Chapter 5, we will discuss both what found in this chapter with the hyperparametrs tuning and the groups characterization described in the next one.

4 | Results and clinical findings

The object of this Chapter is the exploration of all the results of our analyses performed with the S2GC model [18] explained in Chapter 3. In particular, we present the groups found in the three different settings presented in the previous chapter. For each application we will analyse the prognosis of the different groups, using the Kaplan-Meier survival probability curves estimate and then both their clinical and imaging characterization. The study on the prognosis has been done to control that the goal of having groups with different prognosis has been reached. The clinical characterization has the role of validate our stratification procedure. In particular, we are interested in verify that the groups found correspond also to patients with a different clinical characterization. In the following chapter we will also compare the clincial characterization of the groups that we have found thanks to the imaging with the well known risk factor described in Chapter 1. The imaging characterization, instead, has been analysed to find the risks linked with the imaging features and to analyse in detail the role of each view.

This Chapter is organized as follows:

- In Section 4.1 the results of the first application of the S2GC model to the data of the patient affected by ICC are shown. We recall that in this application we have employed only the radiomic features extracted from the Portal phase of the CT scans. In this way we have created two views, the first one composed by the featrues extracted from the core region and the second one composed by the ones relative to the margin. This was done to study the core-margin interface and to build a baseline for the stratification of these patients. Indeed, having the two regions in two different views we were able to analyse their contribution thanks to the study of the regularization parameter described in Section 3.1.3;
- In Section 4.2 the results of the main application of the S2GC model are presented. In this case all the three phases of the CT scans were used to build the patient representation. In particular, we recall that each phase corresponds to a view. The focus was to analyse the contribution of the different views both for the survival analysis and for the performances of the patient stratification. In particular, we were

interested in assess whether these three views provide complementary information to build the stratification of the patients affected by ICC. Another goal of this application was to establish which are the risk factors linked with the imaging;

- In Section 4.3 the results of the last application of the S2GC model to the patient affected by ICC are shown. In this case we were studying the disease free survival time. The goal of this application was to study again the contribution of all the phases of the CT scans, but in a different setting. Indeed, the two times, overall survival and disease free survival time, have a different medical meaning and they are both interesting to study. Therefore, we were not only interested in assess if the three views provide different and complementary information also for this problem, but also to establish whether the clinical characterization of the groups is the same to the one found in the previous application. We were also interested in comparing the risk factors linked with the imaging in the two settings;
- In Section 4.4 a summary of the principal results found in the previous sections is reported.

4.1. Core/Margin assessment in Portal CT imaging

Before presenting the results, we recall that this is the first application of the S2GC [18] model to the data of the patient affected by ICC. In this case the overall survival time has been studied, using only the radiomic features that comes from the Portal phase of the CT scans. The patient included in this study were all the 259 patients available in the dataset. As discussed in section 3.2, the number of cluster nc in this setting is 4, obtaining in this way four different risk classes. In Figure 4.1 Kaplan-Meier curves estimating clusters' survival probability curves for such groups are displayed. The two main groups are the central ones, Group 1 (red line) and Group 2 (blue line), with respectively 138 and 99 patients. Group 3 (yellow line) and group 4 (grey line), however containing only few patients (15 and 7), were associated to better and worse prognosis, with not-achieved and 42 days median survival time respectively. The yellow group is almost exclusively composed by censored patients, while the grey one, conversely, is composed only by non-censored patients. The two centrals groups, instead, have respectively a median survival time of 3324 days and of 779 days.

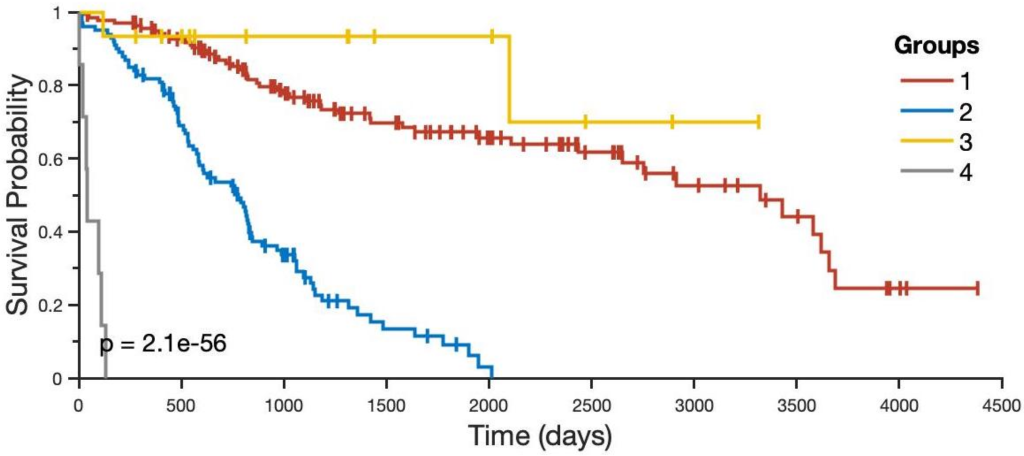


Figure 4.1: Kaplan-Meier curves estimating clusters' survival probability

4.1.1. Clinical Characterization

Beside life expectancy, some test on the clinical variables have been performed in order to establish if the four groups were different in terms of qualitative tumor assessment. For all these tests a p-value lower than 0.05 was considered significant and Bonferroni correction for multiple testing has been used. In Table 4.1 we have reported the results of these tests. In particular, we have reported the variables tested with their correspondent p-value. The variables that have been found statistically different in the four groups have been highlighted in blue.

Table 4.1: P-values of the tests performed on the mean or on the proportion of the clinical variables in the different groups

Variables	P-value	Variables	P-value
Age	0.9720	*CA19-9*	0.0600
Max dimension	0.0050	*Number of lymph nodes removed*	0.1250
Sex	0.4874	*Pattern = 1*	0.0130
Pattern = 2	0.0216	*Pattern = 3*	0.0030
Single Nodule	0.0089	*Grading = 1*	0.9509
Grading = 2	0.3123	*Grading = 3*	0.4279
Therapies = 0	0.0627	*Therapies = 1*	0.0000
Therapies = 2	0.0025	*Therapies = 3*	0.5997
Therapies = 4	0.2955	*Therapies = 5*	0.4950
Therapies = 6	0.7456	*Therapies = 7*	0.8898
Comorbidity = 0	0.0535	*Comorbidity = 1*	0.7458
Comorbidity = 2	0.0166		

The first variable that has been found to be significant is the *Maximum dimension*. This feature corresponds to the maximum dimension of the lesion. In Figure 4.2 the box-plot for this variable is displayed. From this figure it is possible to notice that Group 3 is the one with the smaller median maximum dimension and the smaller variance. Group 4 is, instead, the one characterized by the higher median maximum dimension and the bigger variance. This was expected since Group 3 is the one with the best prognosis while Group 4 has the worst prognosis.

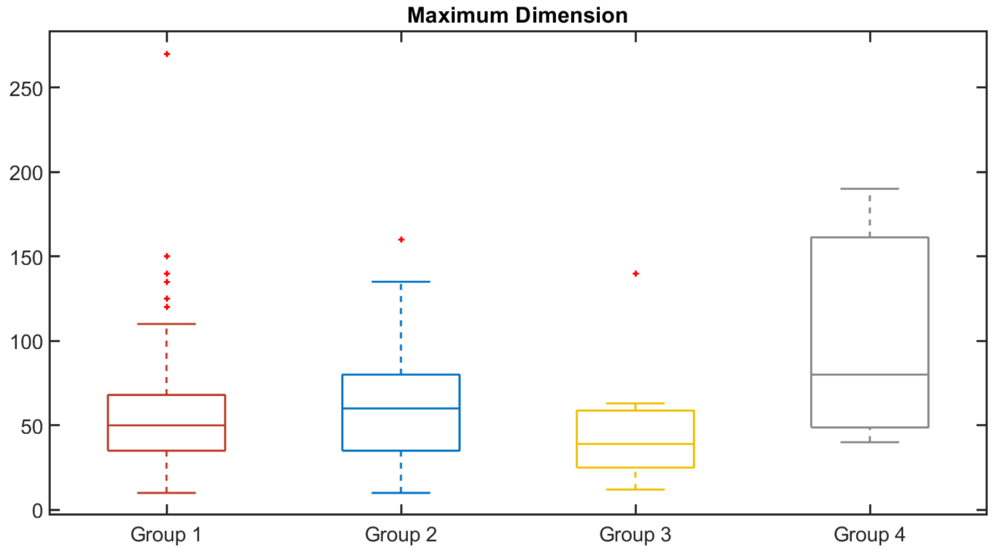


Figure 4.2: Boxplot of the maximum dimension for the four groups

Another variable that resulted significant is the *Pattern*. In particular, *Pattern* = 1 that was significantly more present in the better-prognosis groups and *Pattern* = 3, that instead was associated to bad-prognosis groups. This can be seen in Figure 4.3 where the percentage of patients characterized by the different values of *Pattern* in the four groups is reported. Indeed, from this figure it is clear that Group 4 has the higher percentage of *Pattern* equal to 3, while Group 1 and Group 3, that are the two groups with the best prognosis, have the higher percentage of *Pattern* 1. Instead, Group 2 has an intermediate situation. Indeed, it presents a smaller number of Pattern 1 with respect to Group 1 and Group 3 and more Pattern 2.

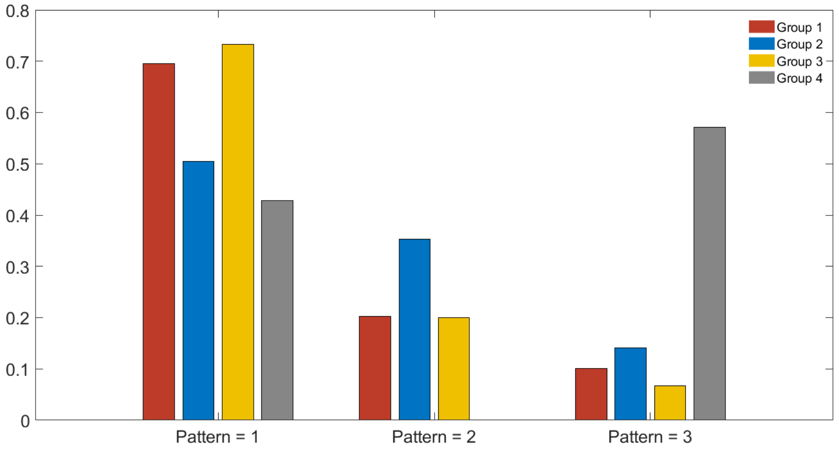


Figure 4.3: Percentage of patients presenting the different values of Pattern for the four groups

It is possible to notice from Table 4.1 that also the *Single nodule* is statistically different in the four groups. From Figure 4.4 can be seen that the groups with an higher median survival time are characterized by more patients having a single nodule while the groups containing the patients that are more at risk have a higher number of nodules. In Figure 4.4 is indeed reported the proportion of patient having a single nodule in the different groups. For each group there are two bars. The one on the left correspond to the percentage of patient characterized by a number of nodules bigger than one, while the bar on the right represent the percentage of patients with a single nodule. As expected, among the left bars the one of Group 4 is the highest. All the others groups are characterized by an high percentage of patients with a single nodule and a small one with a bigger number.

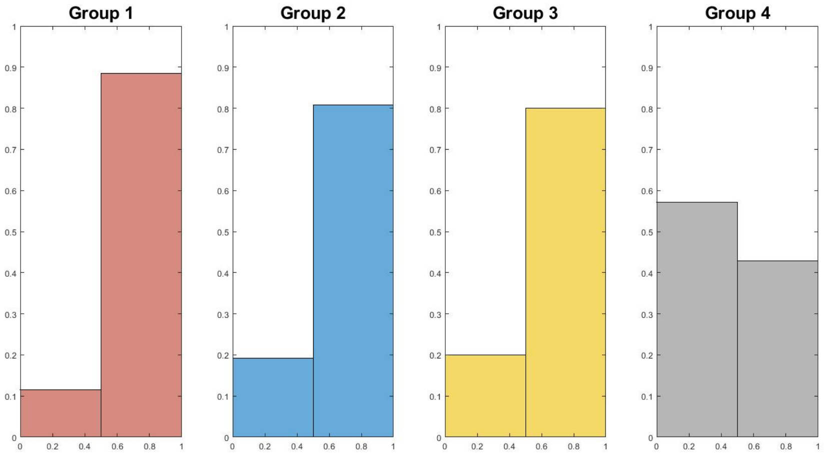


Figure 4.4: Percentage of patients presenting a single nodule for the different groups

Also the number of comorbities have been found to be statistically significant (see Table 4.1). In particular, an high value characterize the groups with patient at higher risks as it is shown in Figure 4.5. In this figure it is displayed a bar chart representing for each group the proportion of patients characterized by a value of comorbity equal to 2. We recall that, as described in Section 1.4.3, a value of 2 for this variable is the highest value and it correspond to a severe general situation of the patient. Indeed, it means that the patient has at least four comorbities. Therefore, the fact that the Group 4 is the one with the biggest proportion of high value of comorbity was expected.

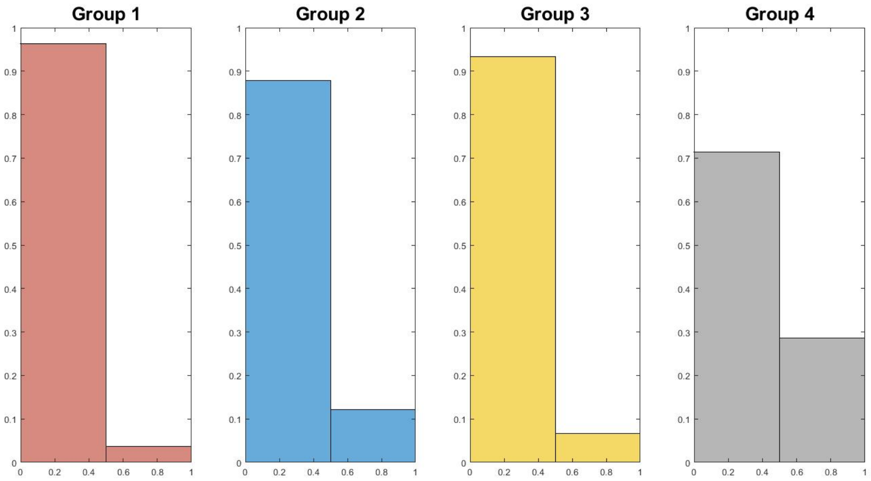


Figure 4.5: Percentage of patients presenting a high value of comorbities for the different groups

Additionally, the majority of patients who underwent Neoadjuvant Chemotherapy and Minor Hepatectomy were found in the better-prognosis groups while patients mainly undergoing Major Hepatectomy without perioperative chemotherapy in the worse-prognosis groups. This shows that patients undergoing minor surgery without perioperative chemotherapy and those undergoing major surgery with chemotherapy have intermediate prognosis. In Figure 4.6 the proportions of patients who underwent the different therapies for the four groups are shown.

Thanks to these studies we can say that the four groups corresponds not only to a different median survival time, but also to a different clinical characterization.

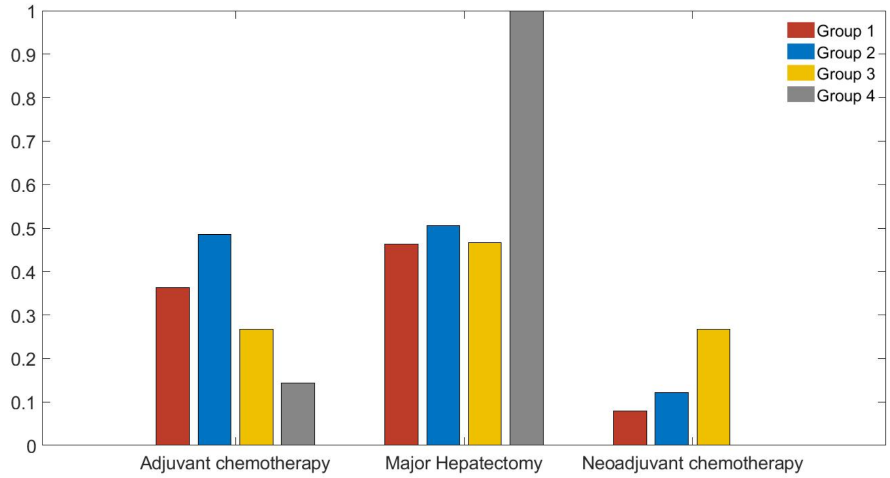


Figure 4.6: Percentage of patients who underwent the different therapies for the four groups

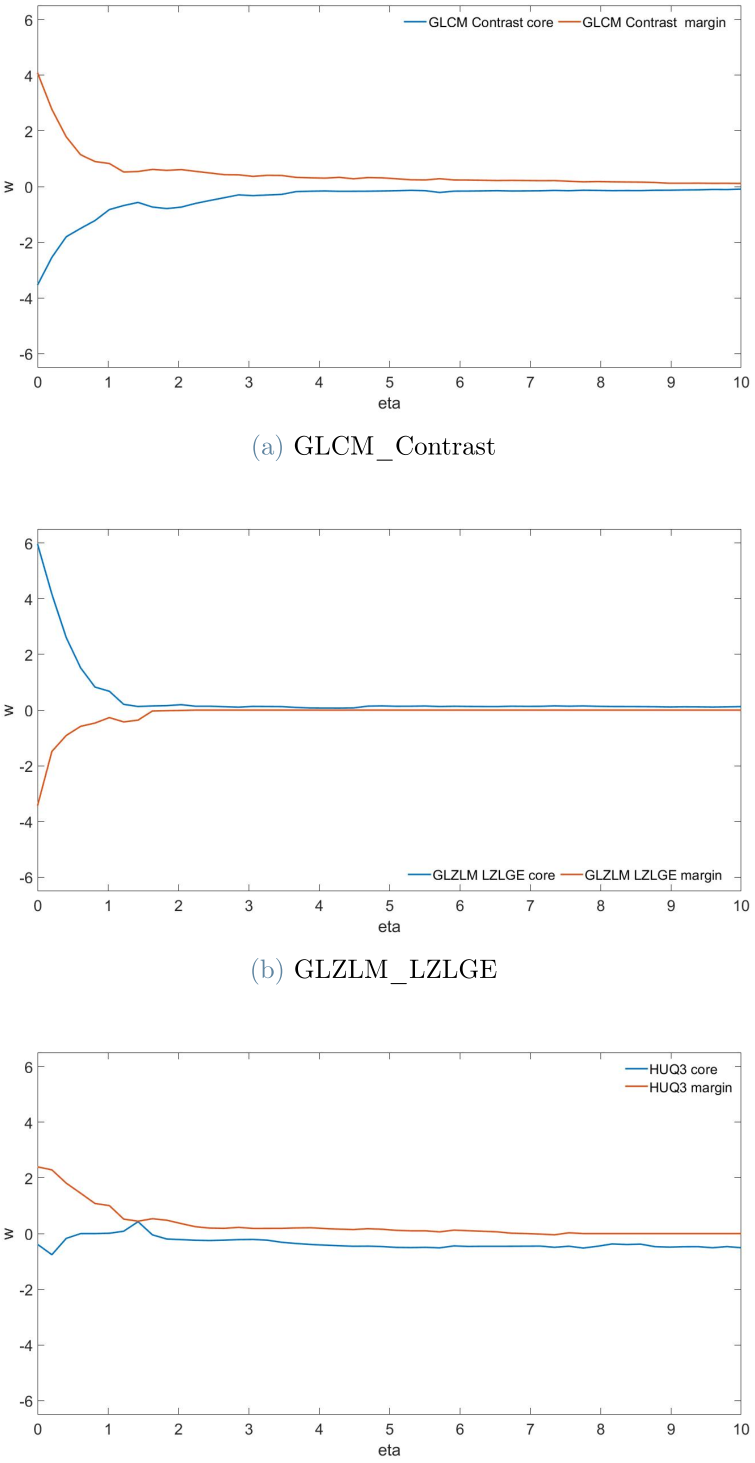
4.1.2. Imaging Characterization

Since one of the goals of this application was to analyse the core-margin interface, a study on the different contribution of these region is reported. As described in section 3.2.4 a ranking on the radiomic features has been computed. We recall that the features considered as the most important are the ones for which the correspondent weight w goes to zero more slowly when the L1 penalization parameter (η) is increased. In this analysis we have used this ranking. For the ten most relevant variables, the weights w of the Cox model, each one with its own counterpart in the other view, are reported in Table 4.2. Interestingly for the majority of these features the weights associated with the core have a sign that is the opposite to the one of the weights associated to the margin region. The weights reported in Table 4.2 are the ones obtained when the parameters have been fixed to their optimal value γ^*, λ^* and η^*. For three of these weights we have, also, reported their behaviour for some value of η in Figure 4.7, while the other two parameters were still fixed to their optimal values. Thanks to this figure it is possible to compare these weights more completely. Indeed, it can be noticed that for all the three variables represented the weights goes to zero when η is increased. This is coherent with the study done in section 3.2.2. It is, also, interesting to notice that in all the cases for the optimal value of η, that is zero, the weights have different sign. In case of *HUQ3* the weight of the core is negative, but smaller in absolute value with respect to the one of the margin that, instead, is bigger and positive. For the *GLCM_ Contrast*, again, the weight of the core is negative while the one of the margin is positive. The difference with respect to the

Table 4.2: Weights of the ten most relevant radiomic features for the application to only the Portal phase of the CT scans

Variable	Core Risk	Margin Risk
HUQ3	-0.3862	2.3952
HUmin	-0.7687	-0.2666
GLZLM_ ZP	0.6417	-0.9747
GLCM_ Contrast	-3.5331	4.0841
GLZLM_ LZLGE	5.9590	-3.4327
NGLDM_ Contrast	-0.4712	-0.3298
HUExcessKurtosis	-0.9529	0.1930
NGLDM_ Coarseness	0.6010	-0.2838
NGLDM_ Busyness	-0.1775	-0.1023

case of *HUQ3* is that for this variable the weight of the core is bigger in absolute value. For the last feature presented in this figure (*GLZLM_ LZLGE*) is, instead, the opposite. Indeed, as can be noticed in Figure 4.7, the blue line, the one representing the core, is in the positive part and the red one in the negative part. This means that the contribution of the two ROIs are different. In fact, it may happen that a variable is a risk factor for the core and a protective one in the margin or the opposite. So we can deduce that is really important to exploit the information of both the two regions in order to achieve a better stratification of the patients affected by Intrahepatic cholangiocarcinoma.



(a) GLCM_Contrast

(b) GLZLM_LZLGE

(c) HUQ3

Figure 4.7: Comparison between the weights of the core and the margin for three of the most relevant variables.

4.2. View assessment in three-phases CT imaging: Overall survival

In this section we present the results obtained for the main application of the S2GC model to the patient affected by ICC. We recall that in this case we were studying the overall survival time using all the three phases of the CT scans to build the patient representation. The patients included in this study were 203, the ones for which all the radiomic features were available. In this setting the number of clusters nc is 5 since, as presented in section 3.3, there are five null eigenvalues of the graph Laplacians, corresponding to five separate subgraphs. For what concerns the survival probability we report Figure 4.8 where are shown the Kaplan-Meier overall survival probability curves for the five groups.

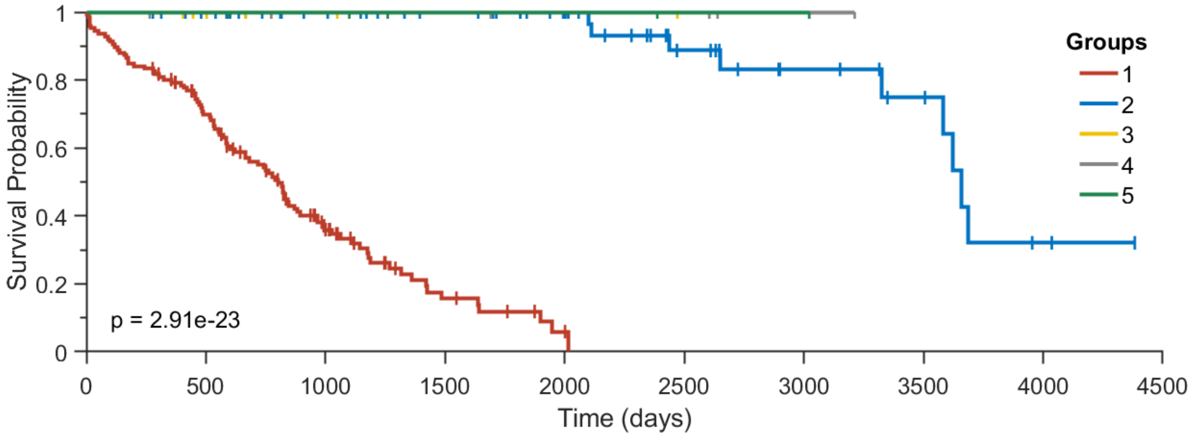


Figure 4.8: Kaplan-Meier curves estimating clusters' survival probability for the Overall Survival when using all the three phases of the CT scans

It is worth to notice that there are three small groups: the yellow one, the grey one and the green one, that have the same survival curve and for which the median survival time is not achieved since they are composed only by censored patient. The other two groups are bigger and clearly characterized by two different survival curves, one with a bad prognosis and the other with a good prognosis. Group 1 (red line) with 126 patients is the biggest one and have a median survival time of 801 days while the Group 2 (blue line), composed by 60 patients, have a median survival time of 3657 days.

4.2.1. Clinical Characterization

Once studied the results of the survival analysis, we were interested in understand if the five groups corresponds to a different clinical characterization or not. For this reason some tests on the clinical variables have been preformed. We recall that these features have been excluded from the patient representation used in the model and employed only to perform this characterization. In Table 4.3 are reported some statistics for the numeric clinical variables and the p-value of the correspondent test. In particular, for the continuous ones we report the mean ± std.dev. while for the discrete ones the median and between the parenthesis the range (minimum value - maximum value). Instead, in Table 4.4 the statistic for the categorical clinical variables and, again, the p-values of the correspondent test are reported. For these variables the number and the percentage of patients presenting that characteristic for the five groups is shown. The variables that have resulted statistically different are the followings: *Severe complications*, *R status*, *Microscopic vascular invasion*, *Grading* and *Metastatic disease*.

Table 4.3: Group characterization according to the exogenous clinical variables (numeric or discrete) with the p-values of the tests performed on the mean of these variables in the different groups.

Variables (mean ± std.dev.) or median (min -max))	Group 1	Group 2	Group 3	Group 4	Group 5	P-value
Numerosity	126	60	6	6	5	
Age	66.60±10.79	65.71 ± 10.30	65.91±12.45	65.36±8.05	64.08±11.85	0.9670
CA19-9	1666.5±7769.5	227.98±854.95	28.36±38.71	17.80±18.89	1869.4±3204.6	0.2910
Max dimension	63.25±33.33	54.10±37.72	40±46.86	45.50±16.86	51.80±13.03	0.1700
Number of lymph nodes removed	4 (0-26)	4 (0-20)	3.5 (0-9)	8 (0-9)	1.5 (0-19)	0.9920
Number of metastatic lymph nodes	0 (0-12)	0 (0-5)	0 (0-0)	0 (0-0)	0 (0-0)	0.0790

For these tests all the five different groups have been taken in count. Then, since the three groups that do not achieve the the median survival time are very small and with the same survival curve, only the two major groups have been compared. The results of these tests have been reported in Table 4.9. From this table it is possible to notice that the variables that have been resulted statistically significant are the same of the previous tests.

Table 4.4: Group characterization according to the exogenous clinical variables (dichotomous or categorical) with the p-values of the tests performed on the proportion of these variables in the different groups.

Variables (% in the group)	Group 1	Group 2	Group 3	Group 4	Group 5	P-value
Metastatic disease	42 (35%)	5 (10%)	0 (0%)	0 (0%)	0 (0%)	0.0011
Sex	64 (51%)	28 (47%)	4 (67%)	5 (83%)	4 (80%)	0.2677
HCV	18 (14%)	6 (10%)	1 (17%)	1 (17%)	0 (0%)	0.8098
HBV	14 (10%)	10 (17%)	0 (0%)	0 (0%)	0 (0%)	0.3601
Neoadjuvant chemotherapy	15 (12%)	4 (7%)	1 (17%)	1 (17%)	0 (0%)	0.6808
Major Hepatectomy	65 (52%)	23 (38%)	1 (17%)	3 (50%)	1 (20%)	0.1659
Severe complications	38 (30%)	3 (5%)	0 (0%)	0 (0%)	0 (0%)	0.0003
Cirrhosis	14 (11%)	8 (13%)	2 (33%)	1 (17%)	0 (0%)	0.4795
R status	52 (41%)	13 (22%)	0 (0%)	1 (17%)	1 (20%)	0.0220
Macroscopic vascular invasion	38 (30%)	11 (18%)	2(33%)	0 (0%)	1 (20%)	0.2512
Microscopic vascular invasion	70 (56%)	22 (37%)	3 (50%)	1 (17%)	1 (20%)	0.0375
Pattern = 1	68 (54%)	43 (72%)	4 (67%)	4 (67%)	4 (80%)	0.1730
Pattern = 2	35 (28%)	12 (20%)	1 (17%)	2 (33%)	0 (0%)	0.4856
Pattern = 3	23 (18%)	5 (8%)	1 (17%)	0 (0%)	1 (20%)	0.3603
Single Nodule	98 (78%)	52 (87%)	5 (83%)	6 (100%)	4 (80%)	0.4682
Grading = 1	18 (14%)	8 (13%)	0 (0%)	0 (0%)	0 (0%)	0.5990
Grading = 2	60 (48%)	44 (73%)	3 (50%)	5 (83%)	4 (80%)	0.0074
Grading = 3	48 (38%)	8 (13%)	3 (50%)	1 (17%)	1 (20%)	0.0081
Perineural infiltration	52 (47%)	17 (33%)	0 (0%)	3 (60%)	1 (20%)	0.0653
Adjuvant chemotherapy	56 (47%)	20 (38%)	2 (33%)	2 (33%)	1 (20%)	0.5717

Table 4.5: P-values of the tests performed on the mean or on the proportion of the clinical variables in the different groups

Variables	P-value	Variables	P-value
Age	0.6040	*CA19-9*	0.2880
Max dimension	0.1110	*Number of lymph nodes removed*	0.2470
Number of metastatic lymph nodes	0.3740	*Metastatic disease*	0.0011
Sex	0.7118	*HCV*	0.5485
HBV	0.2512	*Neoadjuvant chemotherapy*	0.3988
Major Hepatectomy	0.1247	*Severe complications*	0.0002
Cirrhosis	0.8447	*R status*	0.0140
Macroscopic vascular invasion	0.1252	*Microscopic vascular invasion*	0.0235
Pattern = 1	0.0323	*Pattern = 2*	0.3368
Pattern = 3	0.1213	*Single Nodule*	0.2165
Grading = 1	1.0000	*Grading = 2*	0.0017
Grading = 3	0.0011	*Perineural infiltration*	0.1358
Adjuvant chemotherapy	0.3577		

For the variables which were found to be significant we have displayed some plots to better analyse the differences between the five groups. Figure 4.9 shows the percentages of patients characterized by the presence of *Microscopic vascular invasion*, *R status* and *Severe complications* as bar plot for the five groups. Focusing only on the red bar and on the blue one we can analyse the two major groups. In this way can be clearly seen that the patients of Group 1 (the red one), that are the ones with the worst prognosis, are characterized by an higher percentage of *severe complications*, *R status* and *Microscopic vascular invasion* that are well-known clinical risk factors. Group 2 (blue), that is the group with a better prognosis with respect to the red one, instead, is characterized by a less severe situation for what concerns these three variables.

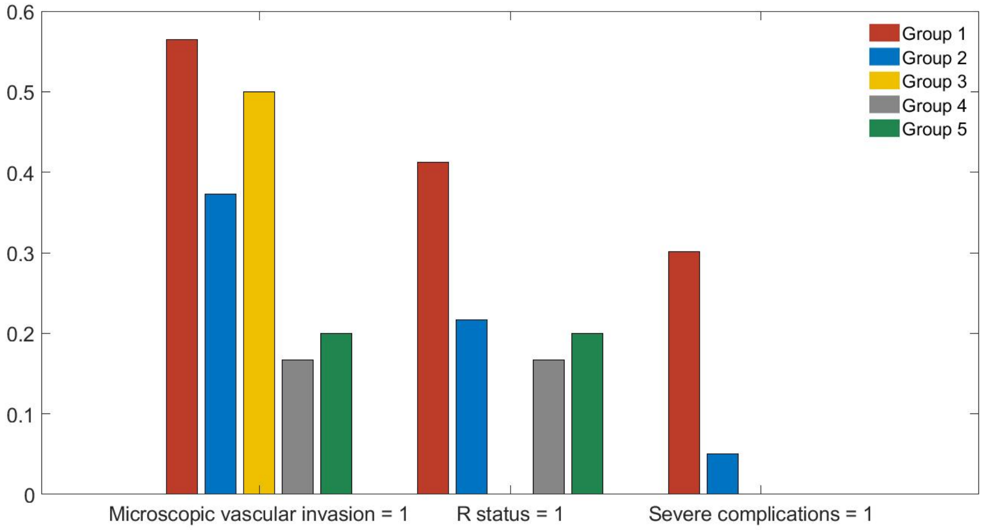


Figure 4.9: Percentage of patient presenting: severe complications, R status and Microscopic vascular invasion for all the five groups

In Figure 4.10 we report the percentage of patients of the five groups characterized by the different values of *Grading*. Focusing, again, on the first two groups, it is possible to notice that Group 2 has an higher percentage of Grading 2 with respect to Group 1 while for Grading 3 is the opposite. This is coherent with the fact that patients in the red group have a worst prognosis, indeed Grading equal to 3 implies a more severe situation. These two groups have, instead, almost the same percentage of Grading 1.

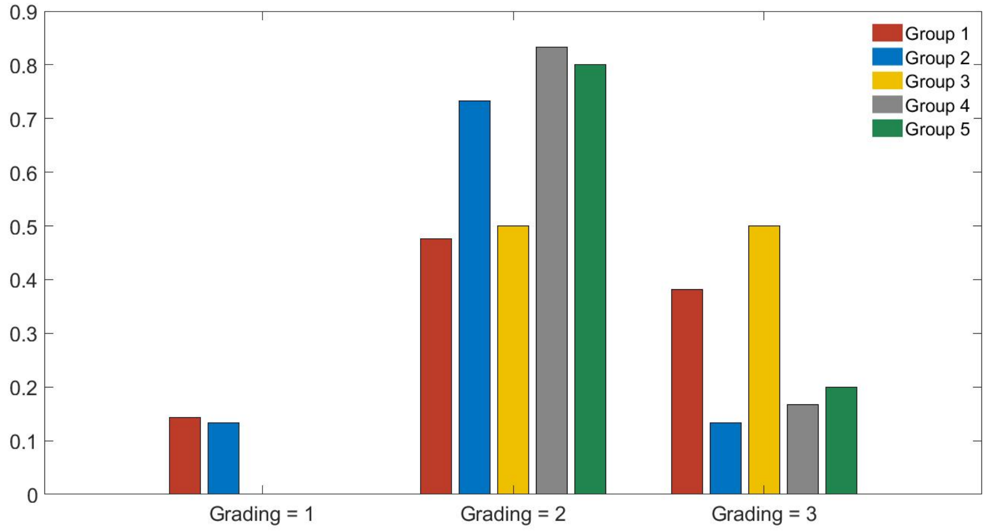


Figure 4.10: Percentage of patient presenting the different values of Grading for all the five groups

Another important aspect to notice is that Group 1 is the one that have the higher percentage of patients that are affected by a metastatic disease, and that all the groups that have the best prognosis are composed by patients with a non metastatic disease (Figure 4.11).

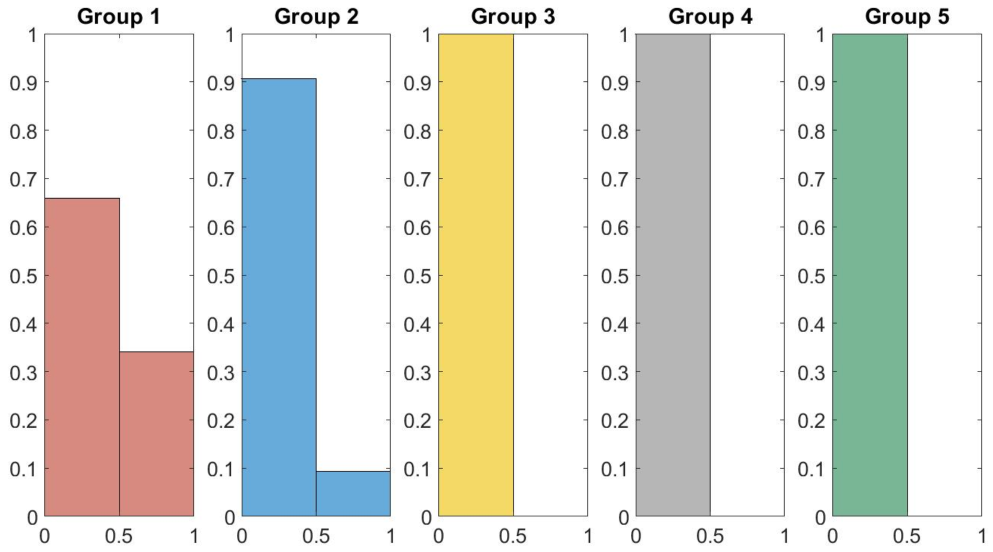


Figure 4.11: Percentage of patient affected by a Metastatic Disease for all the groups

Focusing now on the three small groups (the yellow, the grey and the green ones) we have that the characterization is less definite, especially for Group 4 and Group 5 that are very similar. From Figure 4.9 we can see that these two groups contain patients with no severe complications, mild vascular invasion and small percentage of R1 status. In Figure 4.10 it is, instead, possible to notice that such patients mostly present tumors with Grading 2. Group 3 is instead a group that presents a more severe condition but also do not experience death within the observation time, in this case probably due to the limited time of the study. This group is characterized by patients with no severe complications and no R1 status, but high vascular invasion (Figure 4.9).

4.2.2. Imaging Characterization

After the clinical characterization of the five groups we want to study how the choices of η and λ impact on the patient-to-patient similarity graph. In particular, we were interested in analyse the subgraphs corresponding to the different groups. To do this in Figure 4.12 and Figure 4.13 are reported the same study that have been explained in section 3.3. The difference is that, since in this study we were interested in analyse the different groups, the points of the graphs, that represents the patients, are colored according with the colours of their group. Indeed, in Figure 4.12, looking from the left to the right, four values of λ in decreasing order are reported. Each value of λ is matched with the corresponding similarity graph, obtained when the other two parameters have been fixed to their optimal values. In this way it is possible to notice that only for very small values of λ some patients start to detach. The first group to separate from the other patients is the one of the yellow group. The same plot for four values of η is proposed in Figure 4.13. From this plot we can notice that the first subgraphs to separate are the three small groups (the yellow, the grey and the green ones), while only at the end also the two major groups (the blue and the red ones) separate. Since the three small groups are separated but they have the same survival time, what makes the difference between a patient of one of these groups and one of another group are the radiomic features. This can be interesting because the fact that they are the first to separate seems to confirm this difference in the radiomic variables.

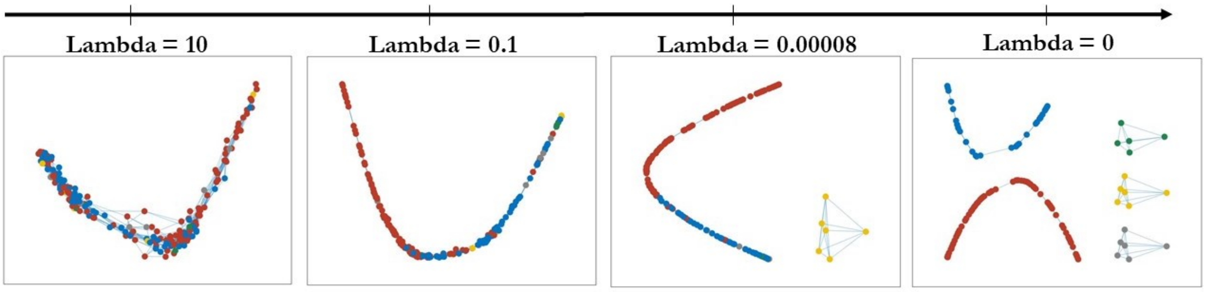


Figure 4.12: Patient-to-patient similarity graph variation according with λ where the colours represent the groups found with the optimal parameters.

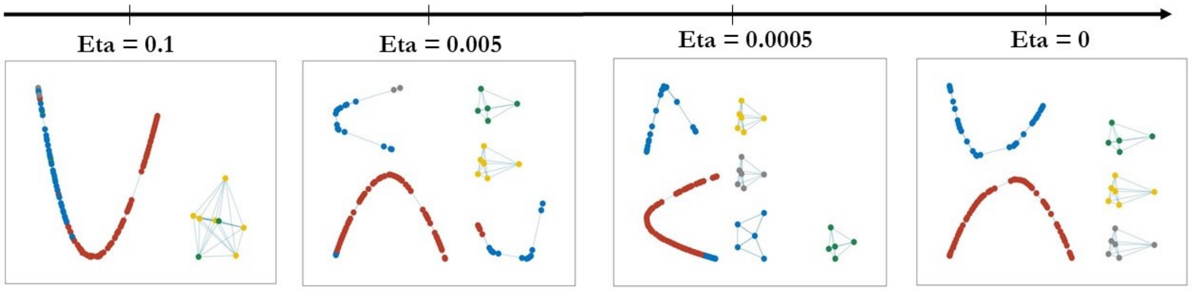
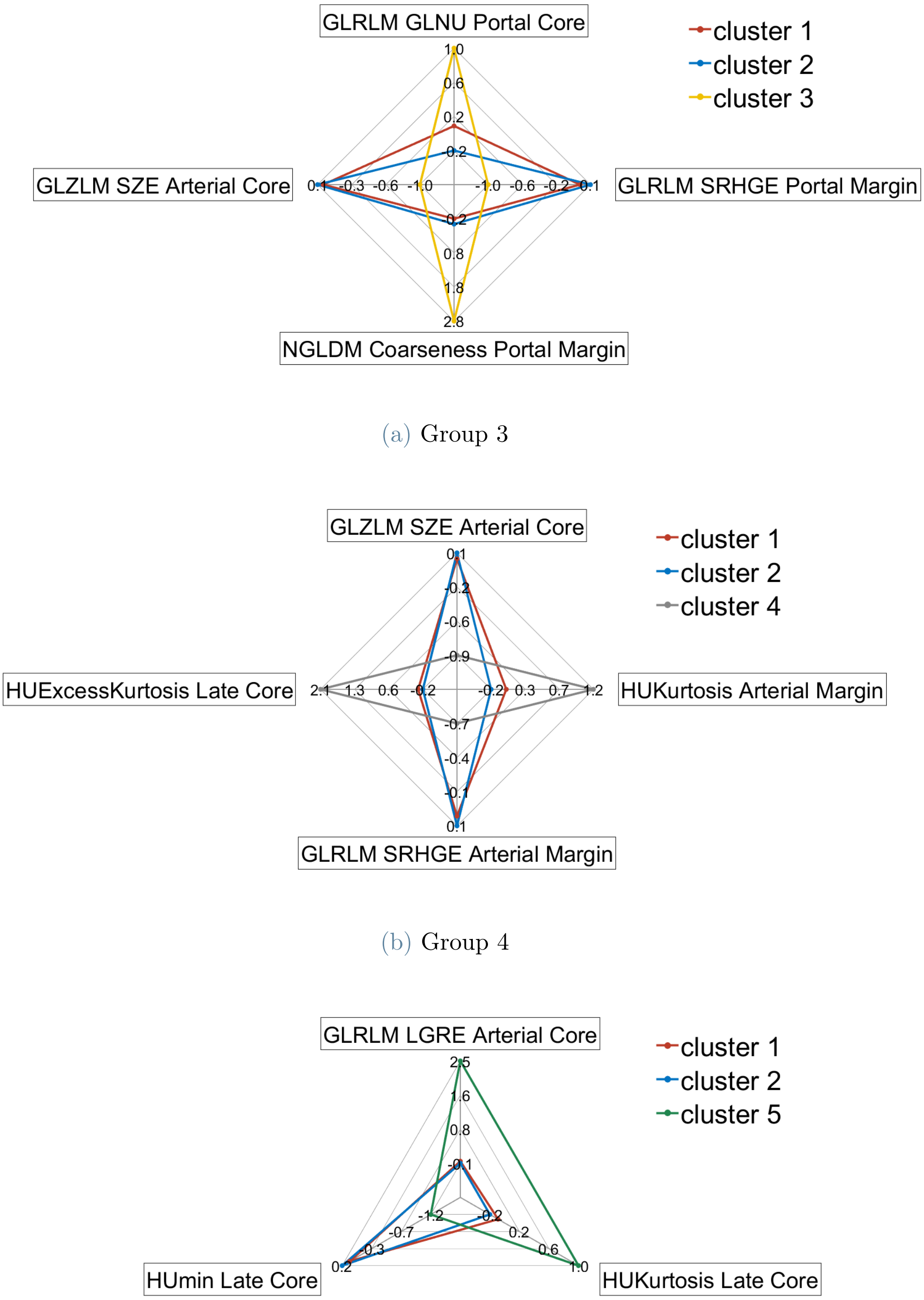


Figure 4.13: Patient-to-patient similarity graph variation according with η where the colours represent the groups found with the optimal parameters.

Once performed this analysis on the penalization parameters, we want to study the imaging features in these groups. Since the model has individuated two major groups and three very small ones, we were interested in studying the difference between the first two and each one of the other three. For this reason a comparative study has been performed. In Figure 4.14 we present a study where each of the small groups has been compared to the two major groups. Therefore, the figure is divided in three parts: (a) corresponds to the study related to Group 3, (b) corresponds to the study related to Group 4 and (c) to the one related to Group 5. To perform these studies, we have selected, respectively, four radiomic variables for Group 3 and Group 4 and three radiomic variables for Group 5. To select the variables for each group, we have performed a correlation analysis between the variables that resulted significantly different between the group in analysis and the two major groups. For each of the three groups, the features that resulted as the less correlated have been chosen. In Figure 4.14 we report the mean values of the different groups for the selected variables. Looking at the figure (a), it is clear that, for all the four variables, the mean values of the first two groups (in red and blue) are very similar, while the ones of Group 3 (in yellow) have an different behaviour. The same pattern can be noticed in figure (b), analysing the values of Group 4 (in grey) and in figure (c), analysing the ones of Group 5 (in green). These results highlight that the prognosis is not the only factor separating the three small groups from the other two. Indeed, as desired, the radiomic variables are also relevant for this distinction.

Since these three groups have been separated by the algorithm, but have the same survival time curve, an analysis on the difference between them has been performed. In Figure 4.15 two different plots are reported. Four radiomic variables have been selected among the ones that have resulted significantly different in these three groups. The ones selected have been chosen in order to minimize the correlation between them. In figure (a) the mean values of these variables is reported for each group. In figure (b), instead, all the patients of these three groups are reported. From both these two figures it is possible to see that, as expected, the three groups are different for some radiomic values. Looking to figure (a) it is possible to see that all the three groups are different. Instead, thanks to figure (b) it is clear that the most different group is the yellow one, while the other two are more similar for the majority of the patients. The similarity for what concerns the radiomic variables is consistent with the result presented in Section 4.2.1. Indeed, in that section we have seen that Group 4 and Group 5 have also a similar clinical characterization, while Group 3 present more differences. The result found on the radiomic variables is also coherent with what seen in Figure 4.12 and Figure 4.13 where the graphs relative to different values of respectively λ and η are shown. Indeed, in Figure 4.12 it is possible to see that the yellow group is the only one that separate also for a value of λ different from 0. This means that Group 3 is the group composed by the patients with the biggest difference for what concerns the radiomic variables. In Figure 4.13 is shown that this group, together with some patients of the green one, is again the first one to detach from all the other patients. This support the thesis that says that the yellow group is the one with the most different radiomic values among the three small groups.

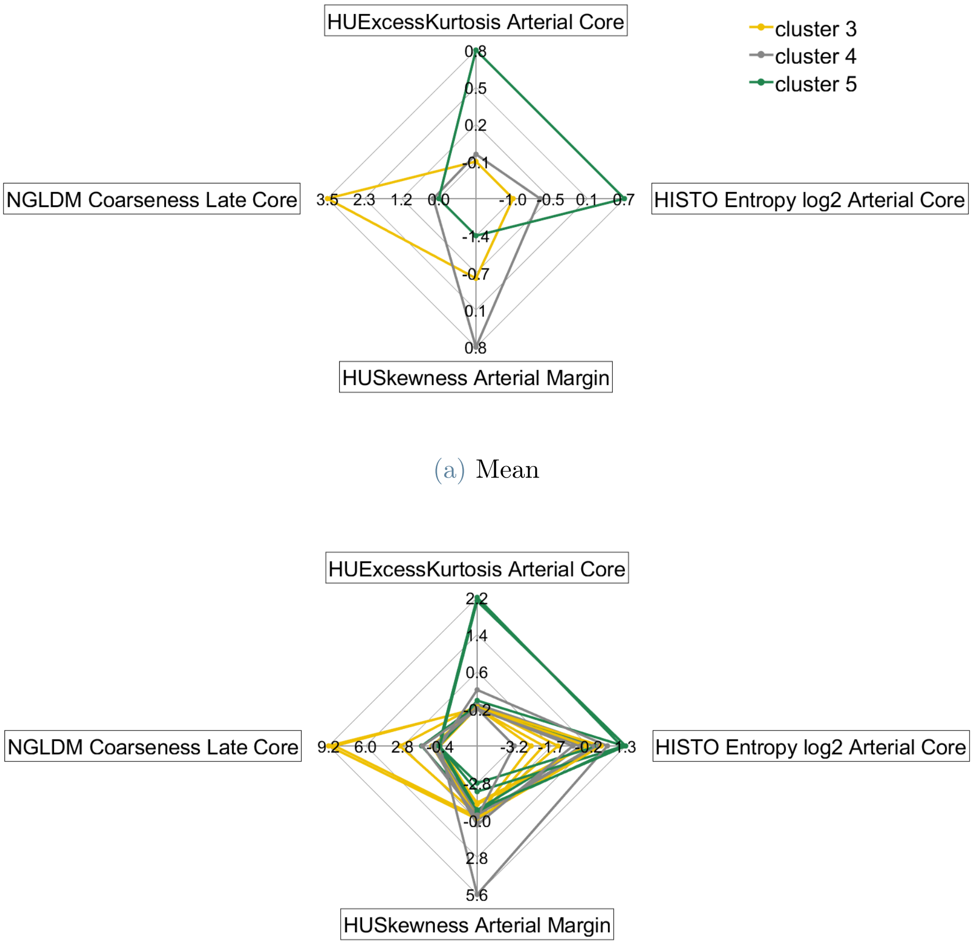


(a) Group 3

(b) Group 4

(c) Group 5

Figure 4.14: Comparisons between the mean values of some radiomic features for Group 1, Group 2 and the smaller groups.



(a) Mean

(b) All patients

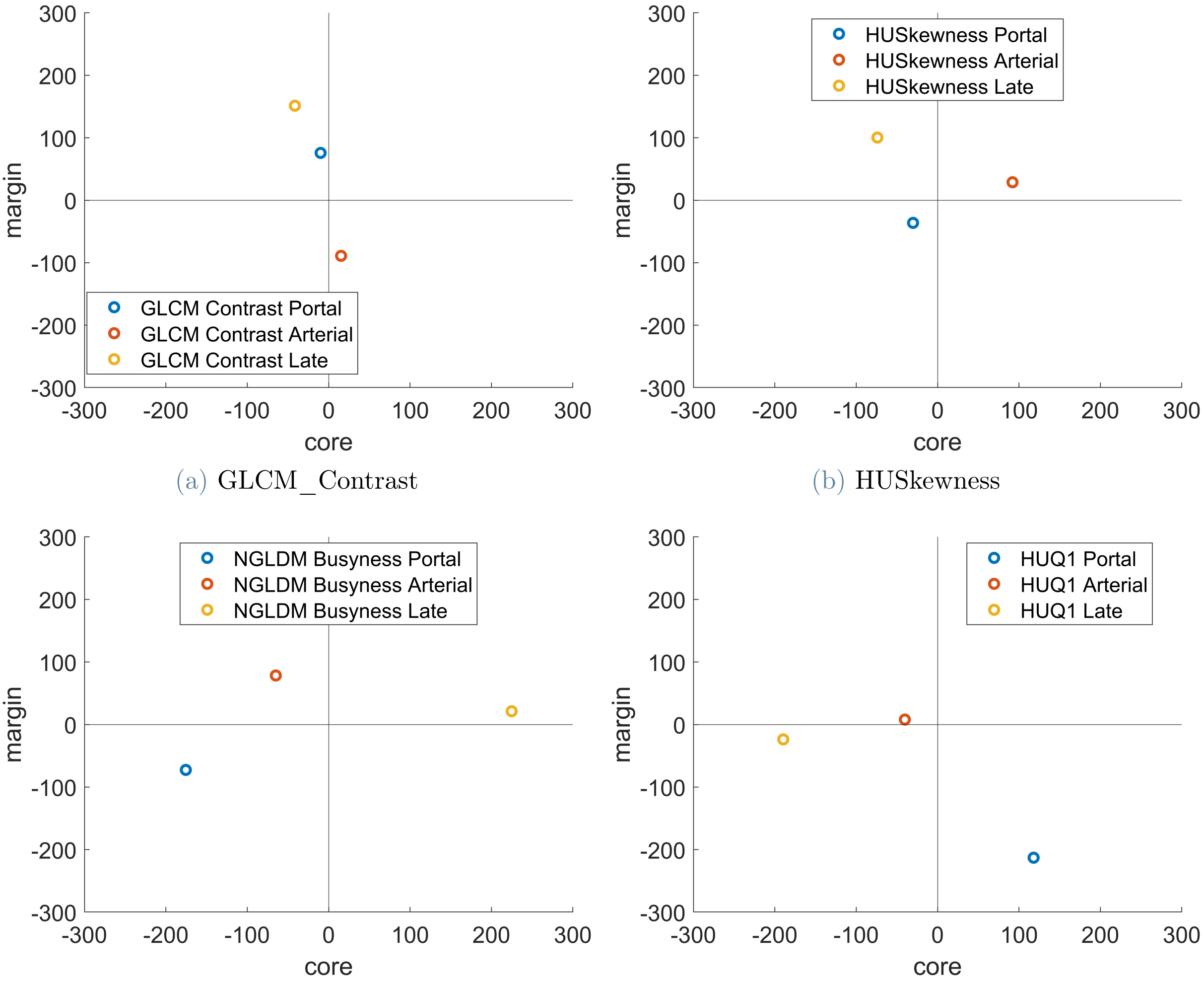
Figure 4.15: (a): comparison between the mean values of five radiomic features for group 3, group 4 and group 5. (b): comparison between the values of five radiomic features for all the patients of group 3, group 4 and group 5.

After having characterized the groups with the clinical variables and having studied the three small groups, we have performed an analysis on the weights w of the Cox model. As already explained in section 3.3.4, a ranking on the radiomic variables has been built accordingly with their weights' speed in going to 0 while η increases. In Table 4.6 the weights of the nine most relevant features are reported together with their counterpart in the other region of interest. The weights reported are the ones obtained with γ^*, λ^* and η^*. As it is possible to notice from this table, also in this case the majority of the features have a weight with opposite sign in the two different ROIs (core and margin). For each of the three phases we can spot cases where this happens. Therefore, we can establish that, as we have already done in the case where only the Portal phase have been employed, the two regions provide different and complementary information. For this reason, also when all the phases of the CT scans are used to build the patient representation is important to include both the core and the margin region. Another important aspect to notice is that all the three phases are represented in the nine most important variables and this is a support to the fact that they provide different information and that are all important.

Table 4.6: Weights of the ten most relevant radiomic features for the application to all the phases of the CT scans for the Overall Survival

Phase	Variable	Core Risk	Margin Risk
Arterial	*GLZLM_ ZLNU*	-1.9867	75.9652
Portal	*HUmin*	12.2855	60.6932
	HUKurtosis	92.5141	-41.9207
Late	*HUQ1*	-189.6116	-23.7804
	GLCM_ Contrast	-41.5134	151.2986
	NGLDM_ Busyness	224.8132	21.2330
	GLZLM_ LZLGE	7.5971	-73.4562
	HUSkewness	-73.9927	100.357
	NGLDM_ Contrast	3.6208	-34.4342

In Figure 4.16 the weights of four of these relevant variables are reported together with the weights of the same variables but relative to a different phase of the CT scans. Two of these features are the biggest for what concerns the absolute value of the weight associated to the core and the other two are the same but with respect to the margin. Each point in this figure represents the ratio between the core risk coefficient and the margin one of a given variable and a given phase. In case the point is in the second or in the fourth quadrant it means that the risk associated to the core and to the margin region have opposite sign, while in case they are in the first or in the third the signs of the two risks are coherent. For example, it is possible to see that for *GLCM_ Contrast* all the points are in the second or in the fourth quadrant and for this reason their risks have opposite sings. In particular, for Portal and Late phases the risk associated with the margin is positive and the one associated with the core is negative, while for the Arterial phase is the opposite. In case of *HUSkewness* all the points are in a different quadrant. The ones of Portal and Late are respectively in the third and in the first quadrant. This means that for the first one both the weights are negative while for the second both are positive. The point of the Arterial phase is, instead, in the second quadrant, meaning that the weight of the core is negative while the one linked with the margin is positive. In general what is possible to notice from Figure 4.16 is that the three phases of the CT scans does not behave in the same way. Indeed, it is possible to notice that for all the four variables reported in this figure the points representing the different views are in different quadrants. In particular, for *HUSkewness*, *NGLDM_ Busyness* and *HUQ1* all the points are in a different quadrant, while for *GLCM_ Contrast* two points are in the same quadrant, but the third is in a different one. This supports what already said: the three views provide important and different information, useful to build a good stratification of patients affected by ICC. It confirm also that analyse the core-margin interface is crucial in patients affected by this disease.



(a) GLCM_Contrast

(b) HUSkewness

(c) NGLDM_Busyness

(d) HUQ1

Figure 4.16: Comparison between the weights of the core and the margin of the different phases of the CT scans.

4.3. View assessment in three-phases CT imaging: Recurrence

In this section the results of the last application of the S2GC model [18] to the ICC data are presented. We recall that in this case it has been used the Disease Free survival time instead of the Overall Survival time. As in the previous case, all the phases of the CT scans have been used. In this setting the number of cluster nc, selected thanks to the eigengap heuristic explained in section 3.4.5, is 4. In Figure 4.17 the Kaplan-Meier disease free survival probability curves for the four groups are shown.

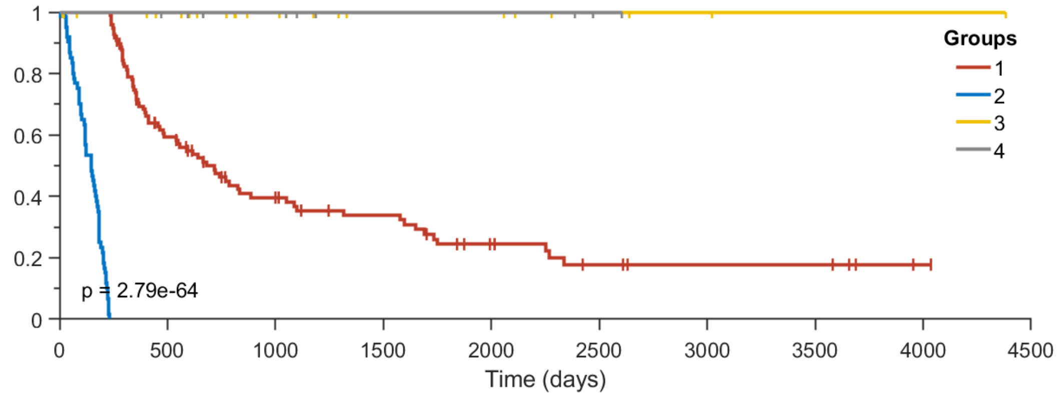


Figure 4.17: Kaplan-Meier curves estimating clusters' survival probability for the Disease Free Survival when using all the three phases of the CT scans

From this figure it is clear that there are two groups with the same disease free survival probability curve and other two with a completely different one. The two groups characterized by the same survival probability are Group 3 (yellow line) and Group 4 (grey line). The other two groups are also the bigger ones where: Group 1 (red line) is composed by 97 patients and Group 2 (blue line) by 63. Between the two smaller group, Group 3 is bigger than Group 4 and they are composed respectively of 24 and 9 patients. These two groups have a median disease free survival time that is not achieved while for Group 1 is 769 and for Group 2 161.

4.3.1. Clinical Characterization

In this section we want to give a clinical characterization to the four groups. To do this we have performed tests on the clinical variables in order to find which are the variables statistically different. The results of these tests, together with a summary statistic for the four groups, are reported in Table 4.7 and Table 4.8. In the first table are reported the tests that refers to the numeric variables. For these variables we have reported the mean ± std.dev. in case they are continuous and the median with the range (minimum value - maximum value) for the discrete ones. In table 4.8 are, instead, reported the results of the categorical variables. For these variables the percentage of patients presenting that clinical characteristic is shown. The variables resulted as statistically significant were: *Number of lymph nodes removed*, *Major Hepatectomy* and *Adjuvant chemotherapy*.

Table 4.7: Group characterization according to the exogenous clinical variables (numeric or discrete) with the p-values of the tests performed on the mean of these variables in the different groups

Variables (mean ± std.dev.) or median (min -max))	Group 1	Group 2	Group 3	Group 4	P-value
Numerosity	97	63	24	9	
Age	67.11 ± 10.99	64.34 ±10.58	68.85 ±10.32	65.25±11.03	0.2430
CA19-9	480.51±1888.9	1446.2±5969.6	3749.1±14656	70.53±123.64	0.1240
Max dimension	56.48±27.99	62.27±37.38	59.46±50.35	70.89±36.34	0.5440
Number of lymph nodes removed	3 (0-21)	4 (0-26)	9 (0-30)	5 (0-9)	0.0100
Number of metastatic lymph nodes	0 (0-8)	0 (0-12)	0 (0-5)	0 (0-1)	0.3530

For the *Number of lymph nodes removed* we can see, from Table 4.7, that it is bigger for the groups with a better prognosis and lower for the other two. This can be noticed also in Figure 4.18 where we have reported a box plot for the number of lymph nodes removed. Indeed, in this figure it can be clearly seen that the Group 3 is the one with the highest number of lymph nodes removed.

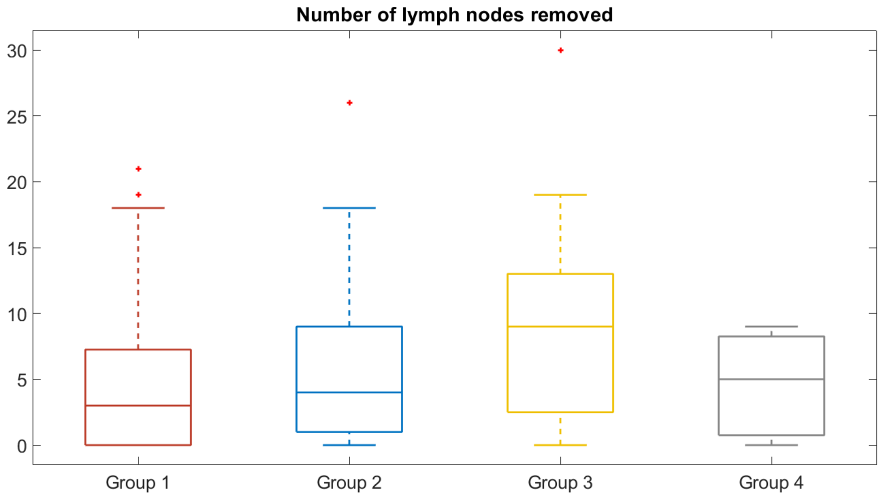


Figure 4.18: Boxplots of the Number of lymph nodes removed for the four groups

Table 4.8: Group characterization according to the exogenous clinical variables (dichotomous or categorical) with the p-values of the tests performed on the proportion of these variables in the different groups

Variables (% in the group)	Group 1	Group 2	Group 3	Group 4	P-value
Metastatic disease	18 (20%)	21 (35%)	5 (21%)	2 (23%)	0.1567
Sex	46 (47%)	34 (54%)	12 (50%)	5 (56%)	0.8588
HCV	14 (14%)	7 (11%)	3 (13%)	2 (22%)	0.8168
HBV	12 (12%)	6 (10%)	2 (8%)	0 (0%)	0.6613
Neoadjuvant chemotherapy	13 (13%)	6 (10%)	0 (0%)	1 (11%)	0.2847
Major Hepatectomy	40 (41%)	33 (52%)	19 (79%)	1 (11%)	0.0008
Severe complications	17 (18%)	17 (27%)	7 (29%)	0 (0%)	0.1465
Cirrhosis	13 (13%)	4 (6%)	2 (8%)	3 (33%)	0.0901
R status	33 (34%)	25 (40%)	8 (33%)	0 (0%)	0.1369
Macroscopic vascular invasion	25 (26%)	14 (22%)	11 (46%)	2 (22%)	0.1554
Microscopic vascular invasion	47 (49%)	32 (52%)	12 (50%)	4 (44%)	0.9599
Pattern = 1	65 (67%)	28 (44%)	16 (67%)	7 (78%)	0.0187
Pattern = 2	21 (22%)	19 (30%)	5 (21%)	2 (22%)	0.6322
Pattern = 3	11 (11%)	16 (25%)	3 (13%)	0 (0%)	0.0505
Single Nodule	79 (81%)	47 (75%)	20 (83%)	9 (100%)	0.2931
Grading = 1	10 (10%)	8 (13%)	5 (21%)	1 (11%)	0.5774
Grading = 2	54 (55%)	35 (56%)	13 (54%)	7 (78%)	0.6234
Grading = 3	33 (34%)	20 (32%)	6 (25%)	1 (11%)	0.4763
Perineural infiltration	32 (39%)	22 (40%)	13 (62%)	3 (33%)	0.2526
Adjuvant chemotherapy	32 (35%)	35 (61%)	7 (32%)	4 (50%)	0.0100

In Figure 4.19, instead, a bar plot for the two statistically significant therapies is shown. For what concerns the *Major Hepatectomy* it is possible to notice that the majority of the patients of Group 3 has undergone through this therapy. This is coherent with the fact that Group 3 is also the group with the highest number of lymph nodes removed. Group 4, the other group with a good prognosis, has, instead, a very small percentage of patients that has undergone through the major hepatectomy. Regarding Group 1 and Group 2 this percentage is around fifty percent. For *Adjuvant chemotherapy* we have a different situation. Indeed, Group 4 has an high percentage of patients that has undergone this therapy. The other group that has a high percentage is Group 2. However, probably due to the fact that the majority of patients has undergone a major hepatectomy, it is a bad prognosis group.

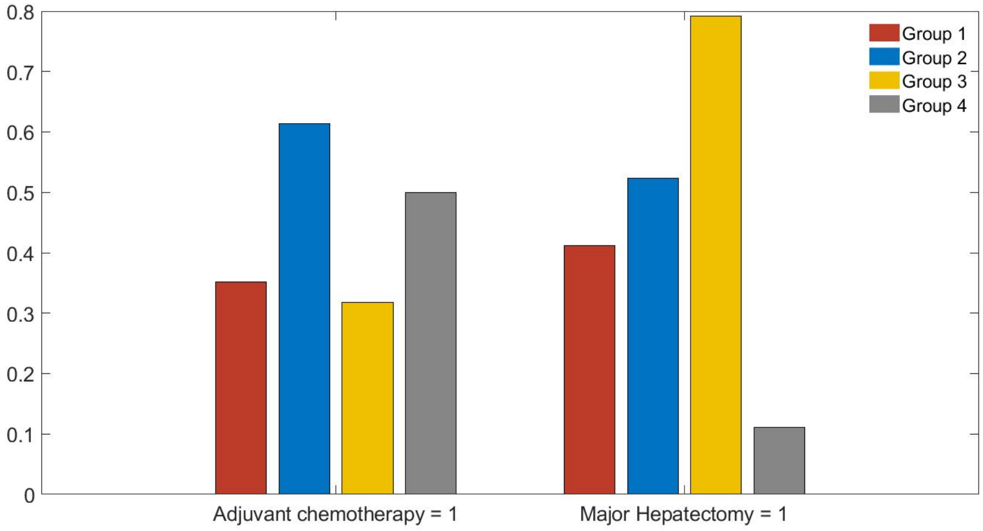


Figure 4.19: Percentage of patient that have undergone: Adjuvant Chemotherapy, Major Hepatectomy for all the groups

Since the last group is composed only by few patients we have performed these tests on the clinical variables also considering only the first three groups. We report the results of these tests in Table 4.9, where we have highlighted the ones that resulted significant. It can be clearly seen that the variables that resulted significant are the same of the case where all the groups have been considered with the only addition of the *Pattern*.

Table 4.9: P-values of the tests performed on the mean or on the proportion of the clinical variables of the three major groups

Variables	P-value	Variables	P-value
Age	0.1270	*CA19-9*	0.0780
Max dimension	0.5990	*Number of lymph nodes removed*	0.0030
Number of metastatic lymph nodes	0.2580	*Metastatic disease*	0.0754
Sex	0.7209	*HCV*	0.8449
HBV	0.7893	*Neoadjuvant chemotherapy*	0.1497
Major Hepatectomy	0.0035	*Severe complications*	0.2556
Cirrhosis	0.3379	*R status*	0.7373
Macroscopic vascular invasion	0.0781	*Microscopic vascular invasion*	0.9122
Pattern = 1	0.0130	*Pattern = 2*	0.4289
Pattern = 3	0.0544	*Single Nodule*	0.5071
Grading = 1	0.3769	*Grading = 2*	0.9910
Grading = 3	0.6966	*Perineural infiltration*	0.1508
Adjuvant chemotherapy	0.0037		

Indeed, in this case, as can be seen in Table 4.9, also *Pattern* becomes statistically different. In particular, Pattern equal to one with a p-value of 0.0130. Therefore, in Figure 4.20 we have shown a bar plot reporting the percentage of patients presenting the different values of *Pattern* for the three major groups. From this figure we can notice that the majority of patients of Group 1 and Group 3 has Pattern 1 while the majority of patients of Group 2 has Pattern 2 and 3. This is coherent with the fact that Group 2 is the one with the worst prognosis.

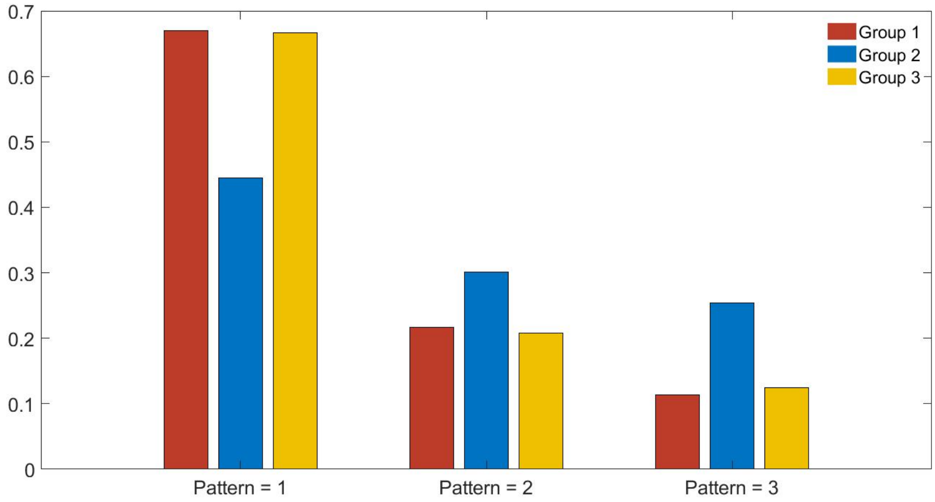


Figure 4.20: Percentage of patient presenting the different values of Pattern for the three most relevant groups

One can notice that in this setting the clinical characterization based on the statistically significant variables is less defined with respect to the one done in the previous case. In fact, in the other one more well known risk factor were significantly different in the groups. However, it is still interesting to notice that in this case the variables that have been found as significant are linked more to the treatment than to the general condition of the patient. This of course is only partially true since generally the treatment is chosen as consequence of the condition of the patients.

Trying to give a more complete clinical characterization of the four groups is possible to say that:

- **Group 1** (red) contains patients with a small number of removed lymph nodes, and few severe complications. They present medium incidence of vascular invasion and major hepatectomy. Few patients are characterized by grading equal to one, while some of them have grading 3. The majority of this patients present Pattern 1 or 2;

- **Group 2** (blue) contains patients with an high incidence of severe complications and R status while a limited number of removed lymph nodes. Major hepatectomy is often administered, along with adjuvant chemotherapy. The patients of this group are characterized by an higher percentage of Pattern 2 and 3 with respect to the other groups;
- **Group 3** (yellow) contains patients with high number of removed lymph nodes, severe complications (but low number of cirrhosis cases) and predominant vascular invasion. Although the majority of patients does not exhibit a good condition, these patients benefit from a heavy application of major hepatectomy and a limited use of adjutant chemotherapy. This might have prevented them from recur within the observation time;
- **Group 4** (grey) contains patients with no severe complication and no R status equal to one but high number of cirrhosis cases, limited incidence of vascular invasion. The majority of the patients have tumors of Pattern 1 and none has Pattern 3. For what concerns the grading the majority has grading equal to 2.

4.3.2. Imaging Characterization

Once performed the clinical characterization of the four groups, we wanted to analyse the imaging features. In particular, we wanted to study the impact of the penalization parameters (η and λ) on the graphs representing the groups of patients. To do this we have reported two figures (Figure 4.21 and Figure 4.22) similar to the ones proposed in Section 3.4. Therefore, in Figure 4.21 are represented the patient-to-patient similarity graphs for four different values of λ. Starting from the left and going to the right these values are displayed in decreasing order. We recall that the graphs shown have been obtained with the other two parameters fixed to their optimal values. In Figure 4.22, instead, the same analysis is represented but with the role of λ and η inverted. Indeed the graphs are displayed for three values of η having γ and λ fixed to respectively to γ^* and λ^*. The difference between these figure and the ones already presented in section 3.4 is that, in this case all the patients, represented by points, are labeled with the colour of the optimal groups. Thanks to this study it is possible to see which are the groups that separate first and in general the behaviour of the different groups. From Figure 4.21 it is possible to notice that as λ decreases the four clusters start to group together but they are still all in a single graph until λ reaches 0. As it is shown in Figure 4.22, for η, the L1 regularization parameter, is different. Indeed, also for η different from 0 there are some groups that are already separated in a subgraphs. In particular, the first to separate is the grey one, that is the smaller one, and what is important to notice is that the two major groups that are separated for η^* are instead again together for $\eta = 0$.

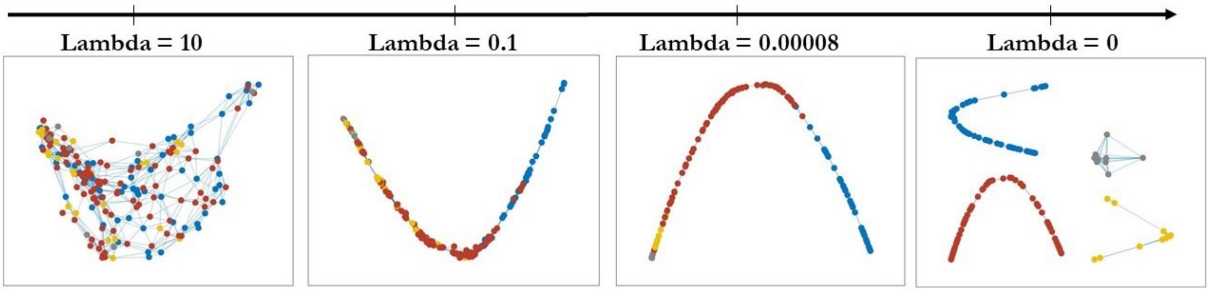


Figure 4.21: Patient-to-patient similarity graph variation according with λ where the colours represent the groups found with the optimal parameters.

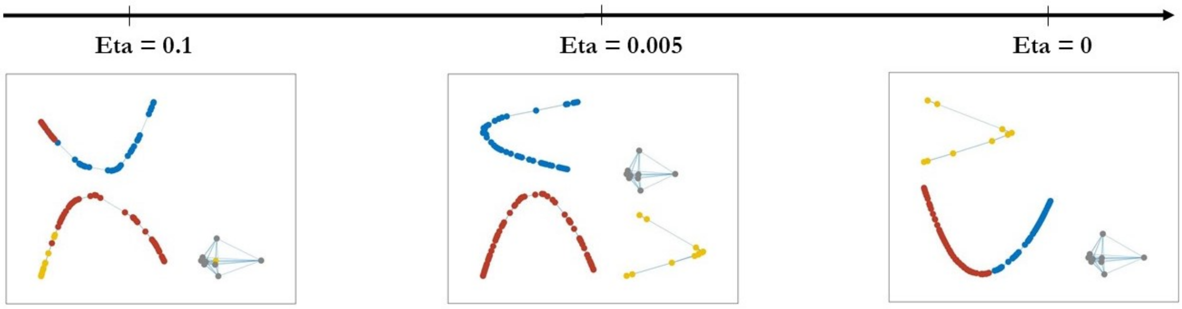
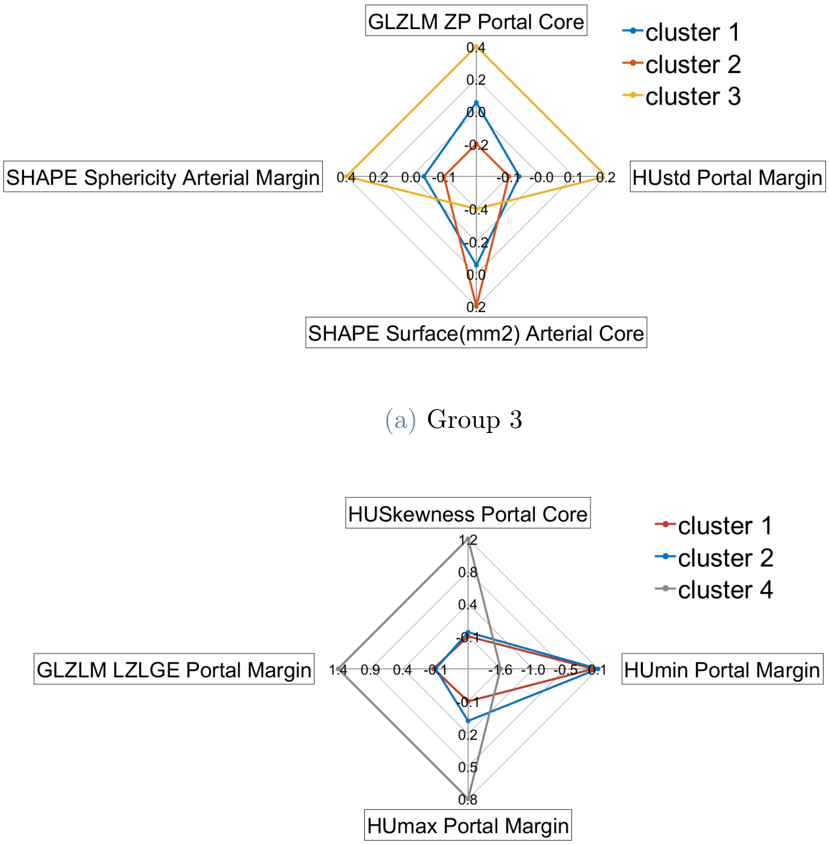


Figure 4.22: Patient-to-patient similarity graph variation according with η where the colours represent the groups found with the optimal parameters.

After this study on the penalization parameters, we were interested in further investigating Group 3 and Group 4 that had been separated by the algorithm, even though they have the same disease survival time. Therefore, some additional studies have been performed to analyse why they are two separated groups. A first study was to test the difference between each one of these two groups and the other two. This study is reported in Figure 4.23. In particular, the figure is divided in two parts: (a) corresponds to the study related to Group 3 and (b) to the one related to Group 4. We selected four radiomic variables for both studies, which were chosen, for each study, through a correlation analysis between the variables that resulted significantly different between the group in analysis and the two major groups. For both groups, the features that resulted as the less correlated have been chosen. In Figure 4.23 we report the mean values of the different groups for the selected variables. Looking at figure (a), it is clear that the behaviour of the mean values of Group 1 is close to the one of Group 2 while relatively different to the one of Group 3. A similar analysis can be performed by referring to figure (b) and substituting Group 3 with Group 4. Therefore Group 3 and Group 4 have been separated by the algorithm from the other two groups, not just because of the great difference in prognosis but also due to the difference regarding the radiomic variables.

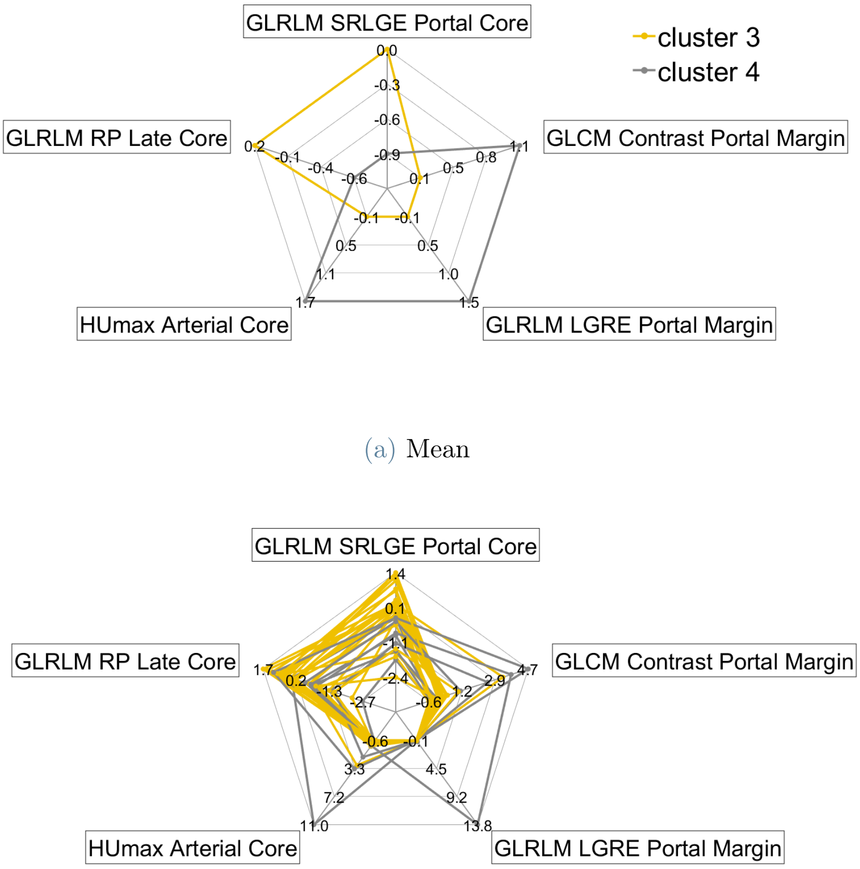


(a) Group 3

(b) Group 4

Figure 4.23: (a): comparison between the mean values of four radiomic features for group 1, group 2 and group 3. (b): comparison between the mean values of four radiomic features for group 1, group 2 and group 4.

The second study that we have performed to better analyse Group 3 and Group 4 was done to test the difference between these two groups. As previously said, this has been done in order to better understand why they are classified in two separate groups although they have the same median disease free survival time. To do this, five radiomic features have been selected, also in this case, through a correlation analysis between the variables that resulted significantly different between the two groups. The five features that resulted as the less correlated have been chosen. This study is divided in two parts and consequently also the Figure 4.24 where we have reported it, is divided in two. In figure (a) there are shown two fictitious patients, one for each group, whose variables correspond to the average value for those variables for all the patients of that group. In figure (b), instead, directly all the patients of these two groups are reported. From both this two figures it is possible to see that the values of these variables for the patients of the two groups are relatively different. This is clear in both figure but especially from figure (a) where having just two lines the difference appears more evident. From figure (b) can be also noticed that the yellow group have value that are similar for all the patients while the grey one is more heterogeneous. Indeed, the yellow lines seems all more compact while the grey ones more spread.



(a) Mean

(b) All patients

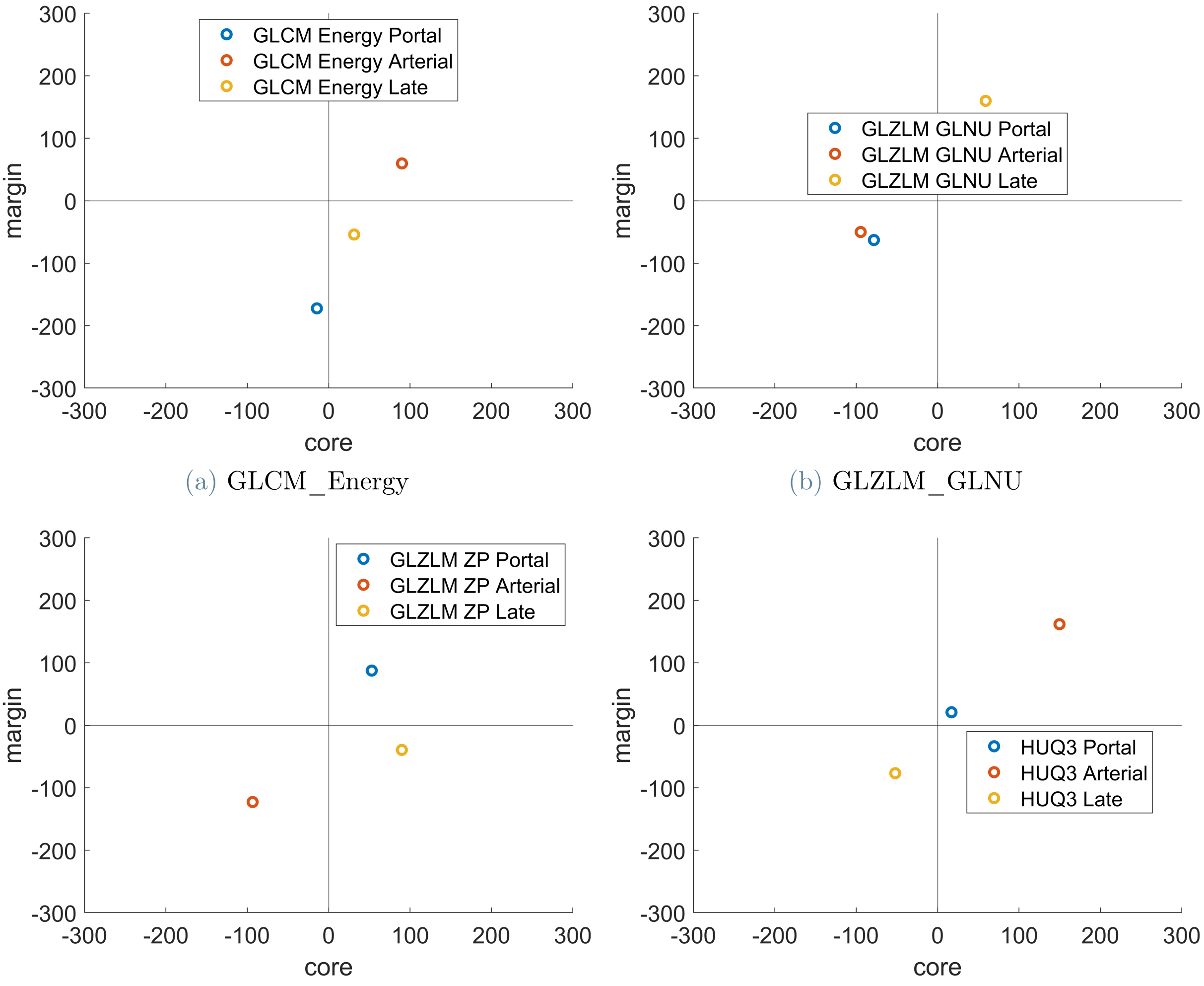
Figure 4.24: (a): comparison between the mean values of five radiomic features for Group 3 and Group 4. (b): comparison between the values of five radiomic features for all the patients of Group 3 and Group 4.

After these studies on Group 3 and Group 4 we present the results of the analysis done on the weights w of the Cox model. Also for this application a ranking on the radiomic features have been built thanks to the convergence to zero of the relative weight when η is increased. Given this ranking we have selected the ten most relevant features and reported them in Table 4.10. This table contains the already cited most relevant features and their counterpart in the other ROI. As already seen in the previous applications also here there are some variables where the sign of the risks linked to the core and the one to the margin have opposite sign. This confirms that the core and the margin regions provide different information and both are important to better analyse the patients affected by ICC. In this case it is, also, interesting to notice that all the three phase of the CT scans are almost equally represented in the 10 most relevant features. The fact that all the phases are represented in the 10 most relevant features can be a support to the theory that says that all the features are important to build a good stratification of patients affected by ICC.

Table 4.10: Weights of the ten most relevant radiomic features for the application to all the phases of the CT scans for the Diseases Free Survival

Phase	Variable	Core Risk	Margin Risk
Arterial	*GLZLM_ ZLNU*	9.4212	113.6525
	HUQ3	149.7209	161.7939
	GLZLM_ ZP	-93.4201	-122.7985
Portal	*HUstd*	1.6775	-41.8953
	GLZLM_ SZLGE	-31.7793	21.9675
	SHAPE_ Compacity	-5.7946	-21.0117
	GLCM_ Energy	-14.4569	-172.2320
Late	*HISTO_ Energy*	-57.5512	9.4263
	GLZLM_ LZLGE	74.5591	-87.2208
	GLZLM_ GLNU	59.0396	159.8932

The weights of four of the features reported in Table 4.10 have been reported in Figure 4.25 together with the weights of the same variables but relative to a different phase of the CT scans. In particular, it is possible to see that for each variable three points have been displayed. Each one of these three points represents the ratio between the core risk coefficient and the margin one for that variable and a given phase. The four features that have been selected are: two the ones with the biggest absolute weight for the core region and the other two the same for the margin. As it is possible to notice from Figure 4.25, the different phases have a behaviour that is not the same. Indeed, for example looking to *GLZLM_ ZP* (Figure 4.25 (c)) it is possible to notice that all the points are in a different quadrant. This means that having an high value for this variable can be a risk factor or a protector one depending on the phase and on the region from which this value come from. In particular, for the Portal phase both weights are positive. Therefore, the *GLZLM_ ZP* is a risk factor for both the core and the margin while for the Arterial phase is exactly the opposite. Instead, for the Late phase the *GLZLM_ ZP* is a risk factor for the core and a protective one for the margin. For what concerns the *GLCM_ Energy* (Figure 4.25 (a)) the discussion is the same only inverting the role of the Portal phase and of the Arterial one. Indeed, in this case is the point of the Arterial phase that relies on the first quadrant while the one of the Portal is in the third. Focusing now on the *GLZLM_ GLNU* (Figure 4.25 (b)) we can see that in this case both the points of the Arterial and of the Portal phase are in the third quadrant. This means that for both the views both the core and the margin are protective factors. The Late phase of the *GLZLM_ GLNU* is instead in the first quadrant meaning that it is a risk factor for both the two ROIs. For *HUQ3* (Figure 4.25 (d)) the situation is symmetric to the one described for *GLZLM_ GLNU*. Indeed, in this case we have that the point of the Portal phase and the one of the Arterial are in the first quadrant while the one of the Late phase in the third. It is also possible to notice that, while for *GLZLM_ GLNU* the weights of the Arterial and Portal phase are similar in magnitude, for *HUQ3* the ones of the Arterial phase are bigger than the ones of the Portal phase. All these results support the thesis for which the three phases provide different and complementary information that is useful to build a good stratification of patients affected by ICC.



(a) GLCM_Energy

(b) GLZLM_GLNU

(c) GLZLM_ZP

(d) HUQ3

Figure 4.25: Comparison between the weights of the core and the margin of the different phases of the CT scans.

4.4. General messages

In this chapter we have presented all the results found performing cancer subtyping with the S2GC model [18] on our multi-view dataset of patients affected by ICC. In this section we sum up the main messages learned from this chapter. These results, together with what found in the hyperparametrs tuning performed in Chapter 3, will be discussed in detail in Chapter 5. In the first scenario, we have analysed in detail the core-margin interface learning that both these two regions are necessary to build a good stratification of the patients affected by ICC. In the second, we have studied the overall survival time of the patients affected by ICC employing all the phases of the CT scans. Therefore, we have been able to analyse the information provided by these three views. We have found evidence that the three views provide different and complementary information to build the patient stratification. In this way we have, also, been able to build a more complete characterization of the patients affected by ICC from both a clinical and an imaging perspective. We have found that the five groups detected by the algorithm correspond to patients with a different clinical and imaging characterization. In particular for the clinical one we have found results that are coherent with the literature for this disease described in Section 1.1. For what concerns the imaging, we had proof of its importance in building the groups of patients thanks to the analysis of the groups characterized by the same prognosis. In the end, with the third application, we have studied the disease free survival time. Also for this study we have employed all the views present in our dataset. This allowed us to further investigate the role of the three phases for patients affected by ICC when the clinical problem in analysis was different. We have found, also in this case, evidence that the three views provide complementary information. From the clinical characterization we have found that the four groups detected from the S2GC algorithm are different also under this point of view. We have also discovered that the clinical variables that characterize the groups of patients in case of the overall survival were different from the ones of the disease free survival. This can be understandable since these two were two different medical problems. Also in this setting we had evidence that the radiomic variables were important to determine the groups thanks to the study performed on the patients of the two groups characterized by the same prognosis. We recall that, also in both the second and third applications, we have found confirm of what said thanks to the first application about the importance of the core and the margin region. For the second and third application of the S2GC model we have also presented an analysis on the regularization parameter λ and η. In particular, we have studied how the patient-to-patient similarity graph varies accordingly with these two parameters. We recall that λ is the co-regularization term that penalize the views' contributions. The role of this parameter will be central in the discussion of Chapter 5 where it will be analysed to understand if the different regions/phases, depending on the application, provide different information. The parameter η is instead the one that penalizes the L1 norm of the weights of the Cox model and it has been fundamental to select the most important radiomic features. All these results will be discussed in detail in Chapter 5 from both a methodological and a clinical point of view.

5 | Discussion

This chapter concerns the discussion of all the results found in the previous chapters. We have divided this analysis in two parts: a methodological part and a clinical one, following the purposes of this work. Indeed, in the first part we have studied all the result from a methodological point of view. Therefore, we have analysed whether it is useful to study the core-margin interface and whether it is important to employ all the phases of the CT scans. Then, since in our purposes there was also to give importance to interpretability and we wanted to find some risk factor linked to the imaging, we have proposed a discussion with a clinical point of view.

5.1. Methodological discussion

For this discussion we have to start with the results of the supervised classification models presented in Chapter 2. These models, even if with all the limitations that a supervised analysis has in case of radiomic data, give us a first evidence that all the phases of the CT scans are important. Indeed, we recall that the prediction accuracy was very low for all the models, but it was increasing while adding more views. This means that this model was not the right one, but that the idea of employ all the three phases is meaningful. Some of the limitations of the supervised analysis for radiomic data have been highlighted also by the survival analysis done with the Cox model. Indeed, looking at the result presented in section 2.2.2 we can see that the performance was not really good also in this case. Another problem that we have found with this model is linked with the interpretabiliy. The Cox model is a regression model that can be of easy interpretation. The problem is that in order to use this model exploiting as much information as possible from the radiomic data we have performed a principal component analysis and this reduced a lot the possibility of easily interpret the results.

Therefore, we have changed the type of analysis performing cancer subtyping using a distant supervision approach. In this way we were not anymore trying to force the radiomic variables to predict a supervised label, but we were looking for patients that shares similar imaging characteristic. We were also interested in having groups composed by patients with similar prognosis and this algorithm, thanks to the fact that is not completely unsupervised, answers also to this problem. Indeed, as deeply discussed in Section 3.1 , it also estimate the survival risk and use it to compute the similarity graph. We now analyse in detail all the result found with the S2GC model [18]. The first application, the one described in section 3.2, had the goal of study the core-margin interface. We recall that in this situation the patient representation was built using two views. The first view was composed by the radiomic features extracted from the core and the second one by the features of the margin. To analyse whether the two views provide different information we have to start looking at section 3.2.2 where the parameters optimization was described. In that section we can see that the optimal value of the co-regularisation term that regularize the imaging views' contributions (λ) is zero. This means that the best results are achieved without any penalization. Therefore, we can say that the two regions provide different and complementary information. To confirm this we can also look at the results presented in section 4.1. Indeed, looking at the plot of the Kaplan-Meier curves (Figure 4.1), we can see that the different cluster have different prognosis and so the stratification under this point of view is meaningful. It doesn't make sense only from a prognosis point of view. In fact, looking at the clinical characterization it is possible to notice that also for these variables the four groups present different characteristics. This means that even if this is just a baseline model the information included in the two region of the Portal phase is important for a correct stratification of patient affected by ICC. The fact that the information provided by the two regions is different can be seen also thanks to Table 4.2 and to Figure 4.7. Indeed, in the table are presented the weights of some of the most important variables computed with the optimal parameters, while in the figure there is represented the behaviour of three of these weights when η is increased. From these, especially from the table, it is possible to notice that the weights are often with different sign in the two regions. This means that the same variable can be a protective factor in the core and a risk one in the margin or the other way around.

Then, we can study the results of the second application, the one where for the first time all the three phases of the CT scans have been used to build the patient representation. This time the goal was to assess whether it is useful to employ the radiomic features of all the phases. As done for the previous application we have to start from the analysis done on λ in section 3.3.2. Also in this case the best results have been found for λ equal to zero. Therefore, we can say that the information provided by the different views is different. Since this time each view correspond to a single phase we can say that all the phases are important in order to find a good stratification of the patient affected by ICC. As it is possible to see from Figure 4.12 just a little increment in the penalization controlled by λ leads to a big loss in the stratification power. Indeed, only for λ equal to zero the groups are really detached. From Table 4.6 we can notice that all the three phases are represented in the nine most important features and this is a support to the fact that they provide different and relevant information. From this table we can also notice that the weights of the core and of the margin have often opposite singe. This is coherent with what analysed in the previous application. Indeed, also this seems to support the thesis that the two regions provide different information. We can now focus on Figure 4.16 where are represented the ratio of the risks linked to the core and to the margin for the different phases for some of the variables presented in Table 4.6. From this figure we can see that the a single variable can have a different role in the different phases, both for what concerns the core and the margin. Indeed, for example *HUSkewness* has both the weights of the core and of the margin positive for the Portal phase and both negative for the Arterial one. Another example can be *GLCM_ Contrast* where for the Arterial phase the weight of the core is positive and the one of the margin is negative while for the other two phases is exactly the opposite. The difference can be also in magnitude. This can be seen for example for the *HUSkewness* where the Portal phase has both weights smaller, in absolute value, than the ones of the the Late phase. It is also important to notice that the performances of this complete model are better than the ones obtained when only the features extracted from the Portal phase were exploited. Indeed, we have obtained a better c-index (see Figure 3.3 and Figure 3.11) and a good clinical characterization. In fact, as we have seen from Section 4.2.1, and as we will discuss later, we can notice that also in this case the clinical variables resulted as significantly different are meaningful. Thanks to all this we can say that the information provided by the three phase is different and all important.

In order to have another proof of this, we have done the same study also for the disease free survival time. Looking at Section 3.4.2 it is possible to notice that also in this case the optimal value of λ is zero. The optimal result is for a very small value of η, small but different from zero. This means that it is correct to slightly penalize the single variable but it is not possible to force the different views to provide the same information since λ is equal to zero. This is a support to the fact that all the phases, that coincide with the views, are relevant and provide different information. Looking at Figure 4.21 the conclusion are even clearer than the ones of the previous application. Indeed, in this case only for λ equal to zero the groups of patients are detached. For all the other values, all the patients are connected all together in a random graph. This is a strong confirm that λ equal to zero is the correct choice for this parameter and consequently that our thesis is correct. For what concerns Table 4.10 and Figure 4.25 the same considerations done for the previous application can be done also in this case.

An important aspect to notice is that in both this two applications the stratification of patient is good. Indeed, we can see from Figure 4.8 and Figure 4.17 that the different groups have different prognosis. Furthermore, both clinical characterization are meaningful and this implies that the stratification obtained thanks to the imaging is really good. Looking at Figure 3.23 and Figure 3.11 we can notice that the c-index obtained with the optimal parameters is high in both cases. We can so conclude that when studying patients affected by ICC it is important to include both the core and the margin region of the lesion but also that is important to use all the three phases of the CT scans to obtain a more complete representation of the patients.

The fact that all the three phases are important for a correct stratification of patients affected by ICC is coherent with what seen in Chapter 1. Indeed, in that chapter we have seen that the contrast is really important in CT scans and that it is possible that a lesion indistinguishable from the healthy liver in one phase will be revealed in another one [29]. So, more in general, it is possible that some characteristics of the region in analysis are highlighted by a phase and other by another one. This is confirmed also by the fact that for all the three applications between the most important variables there is none that is the same in the different views.

5.2. Clinical discussion

In this section we are going to discuss the result from a clinical point of view. In particular we will focus on the last two applications of the S2GC model. This choice has been made since we want to analyse in detail the results found for the stratification for both the overall survival and the disease free survival time. We have not included the first application because the role of that setting was to give us the possibility of studying more in detail the core-margin interface and to build a baseline. Therefore, since we have found that all the three phases were important to build a good stratification of the patients affected by ICC we prefer to directly analyse that result. We are then going to discuss separately the results obtained for the overall survival and the disease free survival time trying then to extrapolate some common characteristics.

5.2.1. Overall survival

Starting from the results of the overall survival time we have divided the discussion in two. First, we are going to analyse the prognosis and the clinical characterization of the five groups. Then, the focus will be on the most important radiomic features and we will discuss the risks linked with them.

First of all, looking at Figure 4.8 it is clearly possible to notice that there are some differences in the prognosis of the different groups. This means that one of our goal is achieved. Indeed, we were interested in obtain groups composed by patients that shares a similar prognosis while the different groups have well distinct survival curves. Said this, we can start to analyse the clinical characterization presented in Section 4.2.1. As first variable we can analyse the *Ca19-9*. This variable has a cutoff of 55. This means that a value under 55 is considered as normal while an higher value is sign of a worst situation. From Table 4.3 it is possible to notice that the first two groups have an high value of this marker and that the three groups with the best prognosis have different values. In particular, Group 3 and Group 4 have small values with the mean under the cutoff, while Group 5 has a big one. This can be a sign that the patient in Group 3 and 4 are healthier while the ones of the Group 5 are in a more severe situation. Unfortunately, the standard deviation is big especially in all the groups with an high mean value meaning that not all the patients have the same situation. Despite this, their mean values seem coherent with the prognosis of the different groups. Looking at Table 4.4 we can see that for what concerns the *R status* there is a big difference both between the first two groups and between the last three. Group 1, the one with the worst prognosis, is the one with the highest percentage of R1 status. This is coherent with the literature described in Chapter 1. Indeed, R0 surgical resection is the major factor that leads to a long-term survival [21]. Group 2, instead, have a smaller percentage of R1 status and consistently a better prognosis. For what concerns the three small groups we have that Group 4 and 5 have a similar percentage of R1 status, while Group 3 has no patient without the safety margin. Another factor that we have seen in the literature that have an influence on the prognosis is the presence of metastatic disease. Again from Table 4.4, it is possible to notice that Group 1 has the highest percentage of patients presenting a metastatic disease while all the groups that have a median survival time not achieved have no patients with a this characteristic. Looking to Group 3 we can notice that in general is characterized by a good situation a part for the high vascular invasion and some Grading 3. Vascular invasion that can be another factor that guides the overall survival probability after the surgery. In this case it seems to be not relevant for the patients of this group, probably also due to the limited time of the study. We recall that our dataset is composed only by patients that have undergone liver resection (major or minor hepatectomy). Looking at *R status* and at *Severe complications* we can say that for the patients of Group 3 the surgery has been successful and also this can be the reason why they do not experience death within the experiment time. Focusing now on the first group, that is the one with the worst prognosis it is possible to see that, as expected, is composed by the patients with the worst situation. Indeed, they are characterized by a severe general condition, by an high vascular invasion and by high values of *Grading*. In fact, Group 1 present a big number of patients with Grading 3. A value of three for this variable means that the tumor is growing fast and in order to continue to proliferate it needs a higher number of blood vessels. This leads to two important consequences. The first one is that generally the tumor grows faster than the blood vessels and, as consequence, there is often more necrosis that makes the surgical resection more difficult. The second one is that this vessels are more fragile with respect to the healthy one and this leads more often to bleeding during the surgical resection. For these reasons, the patients of Group 1 are characterized by a more difficult disease to operate. This is in line with the percentage of *Severe complications* that is higher in Group 1 and null in all the groups with the best prognosis.

For what concerns the radiomic variables we can analyse some of the most important variables that were reported in Table 4.6. In particular, for two reasons we will focus on the margin. The first one is that, as we have seen thanks to the results previously explained, this region is really important and provide useful information. The second one is that from a clinical point of view this region can include the front of progress of the disease that can be really interesting to analyse.
We can look at the *HUKurtosis* that reflects the shape of the Hounsfield distribution relative to a normal distribution. A positive value for this variable means that the distribution is more peaked than the normal one and this implies that the Hounsfield values are more homogeneous. Since the weight for the margin is negative it means that in case of homogeneous values in the margin the risk for the overall survival is smaller. This is dual with what can be deduced from the weight of the *GLCM_ Contrast*. Indeed, the *GLCM_ Contrast* variable represents the local variations in the GLCM matrix. Therefore, it is a measure of the heterogeneity of that matrix. For this reason, the fact that it is associated with a positive weight is coherent with the fact that to a measure of the homogeneity correspond a negative weight. Another interesting variable to analyse is the *HUSkewness* that is a measure of the asymmetry of the Hounsfield distribution. A positive value of *HUSkewness* means that the mean of this distribution is bigger than the median which in turn is bigger than the mode. Therefore, it means that in this distribution there are a lot of small Hounsfield values that usually correspond to a not healthy tissue. This is coherent with the weight of the margin that is positive. A similar analysis can be done also for *HUQ1*. Indeed, it is possible to notice that for this variable both the weights are negative, therefore, implying that having an high first quartile of the Hounsfield distribution is a protective factor. This is meaningful from a clinical point of view because the healthy tissue are usually characterized by higher Hounsfield values than the tumoral ones.

It is really difficult and it doesn't make to much sense to give a precise explanation of every single variable, but it is possible and useful to extract a trend. What can be clearly seen is that in general a big heterogeneity in the margin leads to an higher risk while an high homogeneity to a smaller one. This can be explained by the fact that the margin is the interface between the tumor and the healthy tissues. Therefore, a big heterogeneity can mean that the disease is penetrated in the healthy tissues and as consequence it is stronger and more dangerous.

5.2.2. Recurrence

The discussion for the disease free survival time is divided in the same way as the one already done for the overall survival. First, we are going to analyse the clinical characterization of the four groups described in Section 4.3.1 . Then, we will analyse the weights of the most relevant radiomic variables in order to understand which can be some risk factors for the disease free survival linked with the imaging.

First of all, looking at Figure 4.17 it is possible to notice that the groups have different survival curves. Therefore, since we were interested in find groups of patients characterized by a similar prognosis while the different groups have well distinct survival curves, the first goal is achieved. Analysing now the clinical characterization we can see that there are less variables resulted with a difference statistically significant in the four groups with respect to the overall survival case. Nevertheless, there are some interesting aspect to notice. First of all, it is possible to see from Table 4.8 that the values of *R status* are still coherent with the literature. Indeed, Group 4, the one with the best prognosis, does not have patients with R1 status, while Group 2, the one with the worst prognosis, has the higher percentage. We can also notice that Group 2 has an high number of *Severe complications* and a relevant percentage of Pattern 3 and Grading 3. These characteristics together with the high percentage of vascular invasion show that the patients of this group were in a severe situation. We can see that, instead, Group 1 has a better situation. Indeed, it has lower cases of severe complications and an higher percentage of Pattern 1. For what concerns Group 3 we can see that it has still a relevant incidence of severe complications, but it has less Grading 3 and more Grading 2 with respect to the first two groups. Group 4 is clearly the one with the less severe patients. Indeed, it doesn't have case of Pattern 3 and also the cases of Grading 3 were extremely less than in the other groups. Furthermore, it doesn't have patients with severe complications that strengthens the hypothesis that the disease of this patients was less severe and more easy to operate. We can also see that the general situation of patients of Group 3 was not extremely better than the one of the patients of Group 1. The difference in the survival for these two groups can be also due to the fact that for the patients of Group 3 it was possible to remove more lymph nodes with respect to the ones of Group 1. The fact that Group 4 has the less severe disease is also confirmed by the fact that only a small percentage of these patients has undergone major hepatectomy while for Group 3 the majority of the patients has undergone under this therapy. This last aspect is coherent with the fact that the median number of lymph nodes removed in Group 3 is higher than the ones in the others groups.

Then, we can analyse the radiomic variables to try to understand which are the clinical messages hidden in these variables. In doing this we have take in count the Table 4.6 where the most important variables were reported. In particular, we have focused more on the margin region since, as already explained, is a region rich of information.
First of all, we can look at *HUQ3* that in the margin has a big positive weight. This means that an high value for this variable implies a bigger risk for the disease free survival. Since we are in the Arterial phase, higher Hounsfiled values can be due to an high vascularization. Therefore, this can make sense from a clinical point of view because an high vascularization in the margin can be indication of a tumor that is strong and that it is growing fast. An analogous comment can be done also for the core. Another variable interesting to analyse is the *GLCM_Energy* in the Portal phase which is a measure of the uniformity of the grey-level voxel pairs. Its weight is big in absolute value, but it is negative and so it is a protective factor. A big value for this variable means that the margin is uniform. From this we can deduce that when the margin is uniform the survival risk is smaller. This has a clinical meaning. Indeed, in case of an uniform margin it means that the disease is less strong since it is not penetrated in the that region. This can be confirmed also from what seen for *GLZLM_GLNU* in the Late phase. The *GLZLM_GLNU* variable is a measure of the nonuniformity of the gray-levels and it has a big positive weight, meaning that in case of a big nonuniformity in the margin the risk is bigger. This is coherent with what seen before for *GLCM_Energy*. Furthermore, we can say that having an heterogeneity in the margin increases the risk while an uniformity reduce it. This can be confirmed also by *GLZLM_ZP* that is a measure of the homogeneity of the homogeneous zones and coherently with what already said it has a negative weight.

5.2.3. General messages

We can extrapolate some interesting aspect that are in common or different in the two stratification of the patients accordingly respectively with the overall survival time and the disease free survival time. First, we can see that in both we have found a confirm of what seen in literature about the R status. Indeed, in both cases the groups with the highest percentage of patient with an R1 status are the ones with the worst prognosis.

Another factor that in literature is know to influence the prognosis is the presence of metastatic disease and it is more relevant in the case of the overall survival than in the one of the disease free survival. Indeed, in the last case the four groups, even if they are characterized by different prognosis, do not present relevant difference for what concerns this variables. The same can be seen for what concerns the vascular invasion that we have found in literature it may influence the survival time.
From an imaging point of view the main knowledge that we can learn from both the application is that in the margin a big heterogeneity increases the risks. Indeed, in both cases, we have found that some variables that measures the heterogeneity have positive weights. This means that they are risk factors for both the overall survival and the disease free survival. On contrary we have found that the weights of some of the variables measuring the homogeneity in the margin have negative weights meaning that they are protective factor. As already discussed this can have a clinical meaning. Indeed, in case the margin is more heterogeneous it means that the disease is penetrated also in that region. While in case this region is more homogeneous it means that it is still predominantly composed by healthy tissues.

Before concluding our work with a summary of all the results we have found, in the next chapter we propose a study on a different disease to understand if what we have found can be transferred also in different contexts. In particular, we were interested in assess whether also in the case of Colorectal liver metastases it is useful to employ the information of both the core and the margin region of the tumor.

6 | The colorectal liver metastases case study

This chapter concerns our attempt to transfer the methodological approach found in the previous analysis in a different setting. We want to assess whether the addition of radiomic variables can be useful in case of patients affected by colorectal liver metastases (CLM). In particular, we are interested in the study of the core-margin interface of the Portal phase of the CT scans.

Colorectal cancer (CRC) is one of the most common and deadly types of tumors in the world, indeed it accounts for 10% of all the annually diagnosed cancers and cancer-related deaths worldwide [72]. Colorectal liver metastases (CLM) affect about half of patients with colorectal cancer and dictate patients' prognosis. Prediction of prognosis is of paramount importance for patients allocation to the most adequate treatment, but available parameters do not adequately fulfil this role. The main and best treatment for patients affected by hepatic metastases arising from CRC concerns a surgical liver resection, the aim of which is to remove all the metastatic disease. However, only a restricted number of patients with CLM are candidates for surgical resection at the time of the diagnosis [73]. It is increasingly recognized that in patients with initially inoperable liver metastases, chemotherapy can be effectively provided to downstage the disease allowing a potentially curative resection [74],[72].

The purpose of this study was to assess whether the addition of radiomic features can improve a model that uses only some clinical variables to predict the tumor regression grade (TRG). The TRG is a system to evaluate the amount of residual tumor in patients who underwent preoperative therapy [75]. First, we wanted to establish if there were improvements in the prediction of the TRG while exploiting the radiomic features extracted from the core, together with the clinical variables. Then, we wanted to add also the variables relative to the margin and compare the three models: the initial one based only on the clinical variables, the one where we have added also the radiomic variables extracted from the core and the one where all the radiomic variables have been employed.

6.1. Data

The data for this study have been collected by the Humanitas Clinical and Research Center of Rozzano. This dataset is composed by 319 rows and 109 columns. Each row correspond to a lesion and not as in the previous study to a patient. In this case, the response variable is known for each metastasis and for each patient more than one lesion might be available. Unfortunately, the variation between the lesions of a patient is really low and for this reason can not be used a model that groups these lesions accordingly with their patient. Therefore, we have worked only at lesion level without any grouping. The inclusion criteria to collect these data were: patients undergoing liver resection for CLM with available preoperative CT or PET-CT performed within 60 days before surgery, age>18 years, ability to give the informed consent for study participation and no malignancies other than colorectal cancer and CLM in the previous 5 years with at least one CLM $>$ 10 mm. Patients undergoing explorative laparotomy (i.e., the resection was not performed after explorative laparotomy) and patients with incomplete resection have been excluded. This study was performed according to the Declaration of Helsinki [53]. The local review board approved the study and informed consent was waived given the observational retrospective design of the study. The 109 columns are, instead, composed by radiomic varaibles, clinical variables and the response (TRG). The first ones are 96: 48 extracted from the core region of the tumor and 48 extracted from the margin. Also in this case the margin is defined as a 5mm region that the software automatically generates around the tumor and that has been manually corrected to obtain a more precise result. These variables are the same to the one presented in Section 1.4.1, except for *SHAPE_ Volume(vx)* and *SHAPE_ Surface* that were not available. For what concerns the clinical variables there were available the following 12 features:

- *Interval CT Surgery days*: variable representing the number of days between the moment in which the CT scan has been taken and the day of the surgery;
- *Age*: variable representing the age in years of the patient;
- *Sex*: variable representing the sex of the patient;
- *Number metastasis*: variable representing the number of metastasis that have been detected by the pathological analysis;
- *Diameter*: variable representing the diameter of the metastasis;

- *CEA*: CEA stays for carcinoembryonic antigen and it is used as tumor marker in colorectal cancer;
- *Oxaliplatino*: binary variable indicating whether the patient has been subject to a chemotherapy based on the chemotherapy drug Oxaliplatino;
- *Itinotecan*: binary variable indicating whether the patient has been subject to a chemotherapy based on the chemotherapy drug Itinotecan;
- *ANTI-VEGF*: binary variable indicating whether the patient has been subject to a chemotherapy based on the chemotherapy drug ANTI-VEGF;
- *ANTI-EGFR*: binary variable indicating whether the patient has been subject to a chemotherapy based on the chemotherapy drug ANTI-EGFR;
- *Lines of chemotherapy*: categorical variable representing the number of lines under which the patient has undergone. In case the patient has undergone under only one line this variable has the value of one, while in case the patient has undergone under more than one line the value is two;
- *Cycles of chemotherapy*: numeric variable representing the number of cycles of chemotherapy under which the patient has undergone.

Before proceeding with any model all the data have been imported in R [60] and the radiomic data have been normalized. In particular, they have been shifted by their mean and scaled accordingly with their variance.

The response is a categorical variable that can assume three different values. A value of TRG equal to 1 means that the lesion is completely responsive to the chemotherapy. If TRG is, instead, equal to 3 it means that the lesion is partially responsive and if it is equal to 5 it means that is completely not responsive.

6.2. Methods

The models used in this setting to predict the TRG were two: a multinomial model [61] and a random forest [76]. Since the TRG is a categorical variables but not a binomial one it can not be used a logistic model. Therefore we have used instead a multinomial one that is an extension of the logistic model to the case where the response variable is still categorical but not anymore binary. In particular, we have built three different models. In the first one only the clinical variables have been employed. In the second model we have added also the variables extracted from the core region of the lesion, while in the last one we have considered all the radiomic features, always together with the clinical ones. To select the optimal set of variables we have applied a feature selection approach based on the stepwise logistic regression [62]. Then, to asses whether adding the radiomic features of a the different ROIs have improved the performances we have employed the McNemar's test [63].

As said, the second kind of model that we have used in this setting is the Random forest (RF). This method, developed by Breiman (2001), is an ensemble classification scheme that utilizes a majority vote to predict classes based on the partition of data from multiple decision trees [76]. A random forest is composed by a big number of trees and a tree is composed by sequential splitting of the samples accordingly with the variables. Therefore, for each tree the variables that are used for the first splits are the most important ones. For this reason in a random forest model the most important variables are the ones with the smallest mean depth. With RF we have built just one model where directly all the radiomic features together with the clinical one have been employed. To select the most important features we have performed a feature selection procedure thanks to the R package VSURF [77].

6.3. Results

In this section we present the result found thanks to the analysis previously described. We recall that our interest was in assessing whether the radiomic features can add useful information to the clinical one to correctly predict the TRG. Since we were interested in a comparison between the models that uses only the clinical variables or also the radiomic ones we have built three different multinomial models. The first one exploit only the clinical variables and it is used as baseline. The second one use both the clinical and the radiomic features extracted from the core region. The third, instead, is built on the clinical variables and all the radiomic features, both the one related to the core and the ones of the margin. Before fitting these models a correlation analysis have been performed. Only the variables that were correlated less than 0.8 have been kept. Then, as anticipated, we have applied a feature selection approach based on the stepwise logistic regression [62] to find the optimal set of features. In order to evaluate the results in a more reliable way a 10-fold cross-validation has been applied. So all the result presented in Table 6.1 consist in the mean value of the accuracy found in the 10 fold of the cross-validation. In particular, in this table we have reported the values of the mean accuracy and the standard deviation for all the three model previously described.

Table 6.1: Mean accuracy and standard deviation of the three models fitted to predict the TRG

Models	mean accuracy	standard deviation
only clinical variables	0.561	0.089
clinical variables + Core	0.595	0.085
clinical variables + Core + Margin	0.630	0.136

From Table 6.1 it is possible to notice that all the performances are very low and the standard deviation are not so small. Despite this, what is interesting to notice is that the performances are increasing while the radiomic variables have been added to the model. In particular, it is possible to see that the complete model, the one that employs all the radiomic variables, achieves the best result. In order to asses whether this difference in performances is statistically significant we have employed the McNemar's test [63]. We have performed three different tests, one for each couple of models. In the following we report the results of these tests:

- model only clinical variables vs model clinical variables + core
  p-value = 0.0291;
- model only clinical variables vs model clinical variables + core + margin
  p-value = 0.0007;
- model clinical variables + core vs model clinical variables + core + margin
  p-value = 0.0153.

It is possible to notice that all the tests have resulted statistically significant with a level on 0.05. This means that the radiomic features are important to predict the TRG. In particular, it is useful to analyse both the core and the margin region of the lesions and not only the core one.
Unfortunately, the performances are very low for all the three models. For this reason we have also built a random forest model exploiting all the radiomic features. In this model we have directly employed all the radiomic features since we have seen that the addition of these is important. Thanks to this model we have improved the performances. To estimate the accuracy have been always applied a 10-fold cross-validation procedure. The mean accuracy achieved with this model is 0.84 with a standard deviation of 0.072.

In Figure 6.1 we report an importance plot for this model. In this figure are reported the 10 variables with the smallest mean depth. We recall that in a random forest model the most important variables are the ones with the smallest mean depth. Therefore looking at

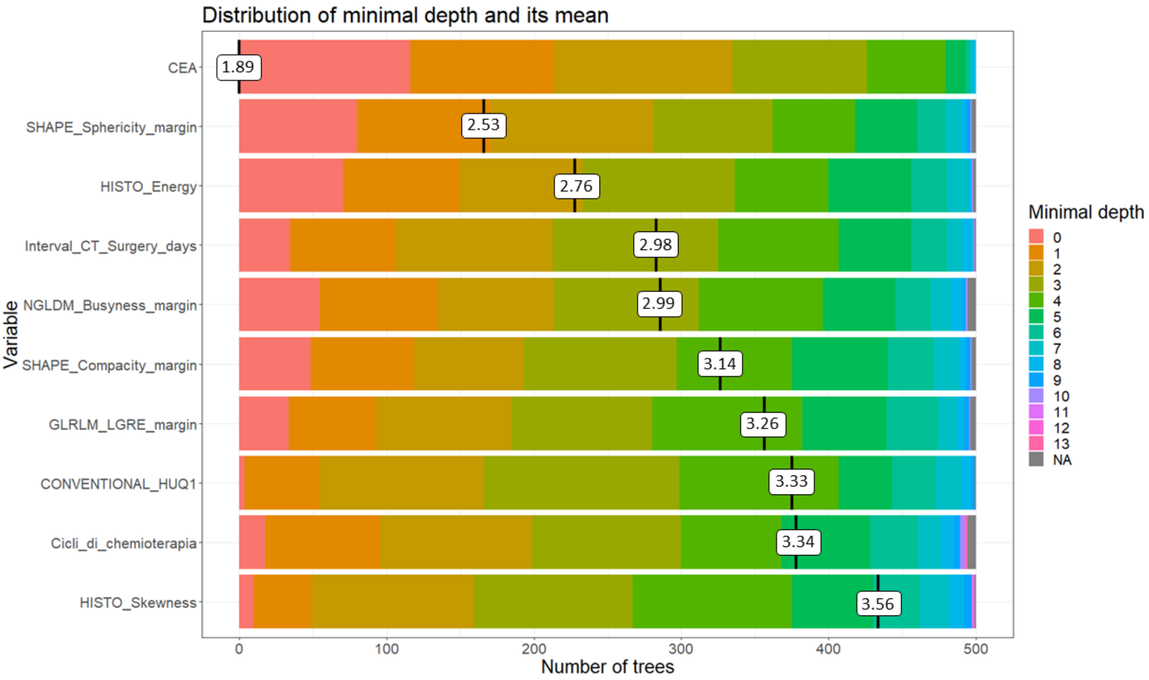


Figure 6.1: Importance plot of the variables used in the random forest model.

Figure 6.1, it is possible to notice that for each variable multiple information is reported. First of all, the features are ordered from the most important one, that is the one on the top, to the fifteenth one that is the last one in the bottom. As said, this ranking has been built on the mean depth and this value is reported in the white boxes for each variable. We can also notice that for each variables there are some bar of different colours and lengths. Each colour stays for a depth at which the variable has been used to split the samples in the trees. The legend can be seen on the right of the figure. The length of the bar, instead, represent the number of trees in which this variable has been used at that depth. From this figure we can notice some interesting aspect. The first is that *CEA* is clearly the most important variable. Indeed, it is both the one with the smallest mean depth and the one used more times as first split. This make sense from a clinical point of view since it is a well known tumor marker for colorectal cancer. The second interesting aspect to notice is that, as already seen with the multinomial model, both the core and the margin region are important. Indeed, between the 10 most important features we can find clinical variables but also radiomic variables extracted from both the core and the margin.

6.4. Conclusions

We can see that also in this application the addition of the radiomic features have improved the performances. In particular, it is interesting to notice that also for the CLM is really important to study the core-margin interface. Indeed, an important improvement in the performance has been obtained exploiting also the features extracted from the margin. Both the multinomial model, thanks also to the McNemar's tests, and the random forest model have confirmed this. Indeed in the random forest model between the first seven most important variables four are variables that have been extracted from the margin region of the lesion.

After having established that the addition of radiomic variables from different ROIs can be useful also in the case of Colorectal liver metastases, in the next chapter we will conclude our work and propose some future developments.

7 | Conclusions and future developments

In this work we have analysed two different types of radiomic multi-view dataset. In the first one each view was composed by the features extracted from a ROI, the core or the margin of the lesion, of the Portal phase of the CT scans. In the second, instead, each view was directly composed by one phase of the CT scans, including the features extracted from both the ROIs. The methodological goals of the analyses that we have performed were to assess whether the addition of more radiomic features, ROIs or phases, could be useful. In particular, we were interested in classifying whether the death of a patient has occurred within the experiment time and in creating a good stratification of the patients affected by ICC. To tackle all these goals, we have performed multiple analyses. First, we have worked in a classic framework for the radiomic data. Indeed, we have employed two supervised methods: a logistic regression model for classification and a Cox model for survival analysis. Thanks to these studies we had the first evidence that the different phases provide complementary information. Then, employing the S2GC model, we have performed three different analyses. Thanks to the first one we had evidence to claim that the core and the margin provide different information and that it is crucial to include both to obtain an accurate stratification of patients affected by ICC when the overall survival is studied. With the second, instead, we have resumed the analysis on the different phases and we have obtained evidence that the different phases provide complementary information. Indeed, if the goal is to create a good stratification for the overall survival of patients affected by ICC all the three phases of the CT scans are needed. Then, with the last application, we have proposed an analogous study in the case where the medical question was the disease free survival time. Also for this stratification we have found that it is important to include all the phases of the CT scans. Thanks to the three applications of the S2GC model to the ICC data we were also able to tackle our clinical purposes. Indeed, we were able to perform cancer subtyping stratifying patients into clinically relevant subpopulations. We have found that some of the clinical variables that characterize the different groups are the ones known in literature to guide the prognosis of patients after the surgery. This was more evident in the case of the overall survival. Moreover, as was one of our goals, we have found some risk factors linked to the imaging. In particular, we have found some evidence that shows that a big heterogeneity in the margin is a risk factor for the survival while homogeneity is a protective one.

From all these findings, we can claim that the three phases provide complementary information that have proven their importance to achieve a good performance in both cancer subtyping and survival analysis. Pertinently, a new frontier of texture analysis is currently rising, that is the delta-texture analysis (DTA) [78]. In fact, evaluating the difference between two regions of interest (spatial DTA) or the same region of interest in separate clinical time instants (temporal DTA) has been shown to be more robust in oncological predictive tasks.

Finally, we have also found that the addition of more radiomic variables, in particular of the margin region, is important also in other medical contexts such as the case of patients affected by colorectal liver metastases.

A possible future development is the analysis of the grouping of the patients affected by ICC. Indeed, as said in Section 1.4, the patients come from six different centers so they can be grouped accordingly with their provenience group. For this reason, it could be interesting to add an additional term to the loss function of the S2GC model to take into account of the center from which each patient comes from.

The code is available at https://github.com/MatteoSavino/RadiomicS2GC.

Bibliography

[1] S. Rizvi and G. J. Gores, “Pathogenesis, diagnosis, and management of cholangiocarcinoma,” *Gastroenterology*, vol. 145, no. 6, pp. 1215–1229, 2013.

[2] A. Nakeeb, H. A. Pitt, T. A. Sohn, J. Coleman, R. A. Abrams, S. Piantadosi, R. H. Hruban, K. D. Lillemoe, C. J. Yeo, and J. L. Cameron, “Cholangiocarcinoma. a spectrum of intrahepatic, perihilar, and distal tumors.,” *Annals of surgery*, vol. 224, no. 4, p. 463, 1996.

[3] R. Bellantone, M. Montorsi, and G. De Toma, *Chirurgia generale. Metodologia, patologia, clinica chirurgica. Con CD-ROM.* Minerva Medica, 2009.

[4] R. Dionigi, P. Cabitza, G. Carcano, P. Castelli, P. Castelnuovo, G. Dionigi, D. Locatelli, G. Parigi, P. Rigatti, and A. Stella, *Chirurgia: Sesta Edizione.* Edra, 2016.

[5] A. F. Peery, S. D. Crockett, C. C. Murphy, J. L. Lund, E. S. Dellon, J. L. Williams, E. T. Jensen, N. J. Shaheen, A. S. Barritt, S. R. Lieber, *et al.*, “Burden and cost of gastrointestinal, liver, and pancreatic diseases in the united states: update 2018,” *Gastroenterology*, vol. 156, no. 1, pp. 254–272, 2019.

[6] J. Peng, Y. Feng, G. Rinaldi, P. Yonglitthipagon, S. E. Easley, T. Laha, C. Pairojkul, V. Bhudhisawasdi, B. Sripa, P. J. Brindley, *et al.*, “The mirnaome of opisthorchis viverrini induced intrahepatic cholangiocarcinoma,” *Genomics Data*, vol. 2, pp. 274–279, 2014.

[7] A. E. Sirica, G. J. Gores, J. D. Groopman, F. M. Selaru, M. Strazzabosco, X. Wei Wang, and A. X. Zhu, “Intrahepatic cholangiocarcinoma: continuing challenges and translational advances,” *Hepatology*, vol. 69, no. 4, pp. 1803–1815, 2019.

[8] M. C. Mason, N. N. Massarweh, C.-W. D. Tzeng, Y.-J. Chiang, Y. S. Chun, T. A. Aloia, M. Javle, J.-N. Vauthey, and H. S. Tran Cao, “Time to rethink upfront surgery for resectable intrahepatic cholangiocarcinoma? implications from the neoadjuvant experience,” *Annals of surgical oncology*, vol. 28, no. 11, pp. 6725–6735, 2021.

[9] J. M. Banales, V. Cardinale, G. Carpino, M. Marzioni, J. B. Andersen, P. Invernizzi,

G. E. Lind, T. Folseraas, S. J. Forbes, L. Fouassier, *et al.*, "Cholangiocarcinoma: current knowledge and future perspectives consensus statement from the european network for the study of cholangiocarcinoma (ens-cca)," *Nature Reviews Gastroenterology & Hepatology*, vol. 13, no. 5, pp. 261–280, 2016.

[10] M. Komuta, O. Govaere, V. Vandecaveye, J. Akiba, W. Van Steenbergen, C. Verslype, W. Laleman, J. Pirenne, R. Aerts, H. Yano, *et al.*, "Histological diversity in cholangiocellular carcinoma reflects the different cholangiocyte phenotypes," *Hepatology*, vol. 55, no. 6, pp. 1876–1888, 2012.

[11] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, 2016.

[12] C. Yang, M. Huang, S. Li, J. Chen, Y. Yang, N. Qin, D. Huang, and J. Shu, "Radiomics model of magnetic resonance imaging for predicting pathological grading and lymph node metastases of extrahepatic cholangiocarcinoma," *Cancer letters*, vol. 470, pp. 1–7, 2020.

[13] C.-F. Lu, F.-T. Hsu, K. L.-C. Hsieh, Y.-C. J. Kao, S.-J. Cheng, J. B.-K. Hsu, P.-H. Tsai, R.-J. Chen, C.-C. Huang, Y. Yen, *et al.*, "Machine learning–based radiomics for molecular subtyping of gliomas," *Clinical Cancer Research*, vol. 24, no. 18, pp. 4429–4436, 2018.

[14] E. R. Velazquez, C. Parmar, Y. Liu, T. P. Coroller, G. Cruz, O. Stringfield, Z. Ye, M. Makrigiorgos, F. Fennessy, R. H. Mak, *et al.*, "Somatic mutations drive distinct imaging phenotypes in lung cancer," *Cancer research*, vol. 77, no. 14, pp. 3922–3930, 2017.

[15] H. Li, Y. Zhu, E. S. Burnside, E. Huang, K. Drukker, K. A. Hoadley, C. Fan, S. D. Conzen, M. Zuley, J. M. Net, *et al.*, "Quantitative mri radiomics in the prediction of molecular classifications of breast cancer subtypes in the tcga/tcia data set," *NPJ breast cancer*, vol. 2, no. 1, pp. 1–10, 2016.

[16] J.-E. Bibault, P. Giraud, M. Housset, C. Durdux, J. Taieb, A. Berger, R. Coriat, S. Chaussade, B. Dousset, B. Nordlinger, *et al.*, "Deep learning and radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer," *Scientific reports*, vol. 8, no. 1, pp. 1–8, 2018.

[17] F. Fiz, G. Costa, N. Gennaro, L. la Bella, A. Boichuk, M. Sollini, L. S. Politi, L. Balzarini, G. Torzilli, A. Chiti, *et al.*, "Contrast administration impacts ct-based radiomics of colorectal liver metastases and non-tumoral liver parenchyma revealing

the "radiological" tumour microenvironment," *Diagnostics*, vol. 11, no. 7, p. 1162, 2021.

[18] C. Liu, C. Wenming, S. Wu, W. Shen, D. Jiang, Z. Yu, and H. San Wong, "Supervised graph clustering for cancer subtyping based on survival analysis and integration of multi-omic tumor data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.

[19] L. Cavinato, N. Gozzi, M. Sollini, C. Carlo-Stella, A. Chiti, and F. Ieva, "Recurrence-specific supervised graph clustering for subtyping hodgkin lymphoma radiomic phenotypes," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2155–2158, IEEE, 2021.

[20] S. Conci, A. Ruzzenente, L. Viganò, G. Ercolani, A. Fontana, F. Bagante, F. Bertuzzo, A. Dore, A. D. Pinna, G. Torzilli, *et al.*, "Patterns of distribution of hepatic nodules (single, satellites or multifocal) in intrahepatic cholangiocarcinoma: prognostic impact after surgery," *Annals of Surgical Oncology*, vol. 25, no. 12, pp. 3719–3727, 2018.

[21] R. El-Diwany, T. M. Pawlik, and A. Ejaz, "Intrahepatic cholangiocarcinoma," *Surgical Oncology Clinics*, vol. 28, no. 4, pp. 587–599, 2019.

[22] N. F. Esnaola, J. E. Meyer, A. Karachristos, J. L. Maranki, E. R. Camp, and C. S. Denlinger, "Evaluation and management of intrahepatic and extrahepatic cholangiocarcinoma," *Cancer*, vol. 122, no. 9, pp. 1349–1369, 2016.

[23] D. Waseem and T. Patel, "Intrahepatic, perihilar and distal cholangiocarcinoma: management and outcomes," *Annals of hepatology*, vol. 16, no. 1, pp. 133–139, 2017.

[24] A. J. Lee and Y. S. Chun, "Intrahepatic cholangiocarcinoma: the ajcc/uicc 8th edition updates," *Chin Clin Oncol*, vol. 7, no. 5, p. 52, 2018.

[25] T. Patel, "Cholangiocarcinoma," *Nature clinical practice Gastroenterology & hepatology*, vol. 3, no. 1, pp. 33–42, 2006.

[26] 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.

[27] E. Bercovich and M. C. Javitt, "Medical imaging: from roentgen to the digital revolution, and beyond," *Rambam Maimonides medical journal*, vol. 9, no. 4, 2018.

[28] T. D. DenOtter and J. Schubert, "Hounsfield unit," 2019.

[29] A. Quatrehomme, I. Millet, D. Hoa, G. Subsol, and W. Puech, "Assessing the classification of liver focal lesions by using multi-phase computer tomography scans," in *MICCAI international workshop on medical content-based retrieval for clinical decision support*, pp. 80–91, Springer, 2012.

[30] D. Duda, M. Kretowski, and J. Bézy-Wendling, "Texture characterization for hepatic tumor recognition in multiphase ct," *Biocybernetics and Biomedical Engineering*, vol. 26, no. 4, p. 15, 2006.

[31] M. G. Lubner, A. D. Smith, K. Sandrasegaran, D. V. Sahani, and P. J. Pickhardt, "Ct texture analysis: definitions, applications, biologic correlates, and challenges," *Radiographics*, vol. 37, no. 5, pp. 1483–1503, 2017.

[32] 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.

[33] E. Scalco and G. Rizzo, "Texture analysis of medical images for radiotherapy applications," *The British journal of radiology*, vol. 90, no. 1070, p. 20160642, 2017.

[34] 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.

[35] S. Rizzo, F. Botta, S. Raimondi, D. Origgi, V. Buscarino, A. Colarieti, F. Tomao, G. Aletti, V. Zanagnolo, M. Del Grande, *et al.*, "Radiomics of high-grade serous ovarian cancer: association between quantitative ct features, residual tumour and disease progression within 12 months," *European radiology*, vol. 28, no. 11, pp. 4849–4859, 2018.

[36] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.

[37] D. Wu, D. Wang, M. Q. Zhang, and J. Gu, "Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification," *BMC genomics*, vol. 16, no. 1, p. 1022, 2015.

[38] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, 2009.

[39] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg, "Similarity network fusion for aggregating data types on a genomic scale," *Nature methods*, vol. 11, no. 3, pp. 333–337, 2014.

[40] D. Ramazzotti, A. Lal, B. Wang, S. Batzoglou, and A. Sidow, "Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival," *Nature communications*, vol. 9, no. 1, pp. 1–14, 2018.

[41] H. Ding, M. Sharpnack, C. Wang, K. Huang, and R. Machiraju, "Integrative cancer patient stratification via subspace merging," *Bioinformatics*, vol. 35, no. 10, pp. 1653–1659, 2019.

[42] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel, "Joint and individual variation explained (jive) for integrated analysis of multiple data types," *The annals of applied statistics*, vol. 7, no. 1, p. 523, 2013.

[43] H. Nguyen, S. Shrestha, S. Draghici, and T. Nguyen, "Pinsplus: a tool for tumor subtype discovery in integrated genomic data," *Bioinformatics*, vol. 35, no. 16, pp. 2843–2846, 2019.

[44] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle, "Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets," *Molecular systems biology*, vol. 14, no. 6, p. e8124, 2018.

[45] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire, "Deep learning–based multi-omics integration robustly predicts survival in liver cancer," *Clinical Cancer Research*, vol. 24, no. 6, pp. 1248–1259, 2018.

[46] J. Wu, Y. Cui, X. Sun, G. Cao, B. Li, D. M. Ikeda, A. W. Kurian, and R. Li, "Unsupervised clustering of quantitative image phenotypes reveals breast cancer subtypes with distinct prognoses and molecular pathways," *Clinical Cancer Research*, vol. 23, no. 13, pp. 3334–3342, 2017.

[47] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine learning*, vol. 52, no. 1, pp. 91–118, 2003.

[48] C. Thongprayoon, M. A. Mao, M. T. Keddis, A. G. Kattah, G. Y. Chong, P. Pattharanitima, V. Nissaisorakarn, A. K. Garg, S. B. Erickson, J. J. Dillon, *et al.*, "Hypernatremia subgroups among hospitalized patients by machine learning consensus clustering with different patient survival," *Journal of Nephrology*, pp. 1–9, 2021.

[49] K. Raza and N. K. Singh, "A tour of unsupervised deep learning for medical image analysis," *Current Medical Imaging*, vol. 17, no. 9, pp. 1059–1077, 2021.

[50] A. Cheerla and O. Gevaert, "Deep learning with multimodal representation for pancancer prognosis prediction," *Bioinformatics*, vol. 35, no. 14, pp. i446–i454, 2019.

[51] L. Lu and B. J. Daigle Jr, "Prognostic analysis of histopathological images using pre-trained convolutional neural networks: application to hepatocellular carcinoma," *PeerJ*, vol. 8, p. e8668, 2020.

[52] G. Marinos, C. Symvoulidis, and D. Kyriazis, "Micsurv: Medical image clustering for survival risk group identification," in *2021 4th International Conference on Bio-Engineering for Smart Technologies (BioSMART)*, pp. 1–4, IEEE, 2021.

[53] W. M. Association *et al.*, "World medical association declaration of helsinki: ethical principles for medical research involving human subjects," *Jama*, vol. 310, no. 20, pp. 2191–2194, 2013.

[54] C. Nioche, F. Orlhac, S. Boughdad, S. Reuzé, J. Goya-Outi, C. Robert, C. Pellot-Barakat, M. Soussan, F. Frouin, and I. Buvat, "Lifex: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity," *Cancer research*, vol. 78, no. 16, pp. 4786–4789, 2018.

[55] M. Jaklitsch and H. Petrowsky, "The power to predict with biomarkers: carbohydrate antigen 19-9 (ca 19-9) and carcinoembryonic antigen (cea) serum markers in intrahepatic cholangiocarcinoma," *Translational Gastroenterology and Hepatology*, vol. 4, 2019.

[56] H. Bismuth and L. Chiche, "Surgery of hepatic tumors," *Progress in liver diseases*, vol. 11, pp. 269–285, 1993.

[57] P. A. Clavien, J. Barkun, M. L. De Oliveira, J. N. Vauthey, D. Dindo, R. D. Schulick, E. De Santibañes, J. Pekolj, K. Slankamenac, C. Bassi, *et al.*, "The clavien-dindo classification of surgical complications: five-year experience," *Annals of surgery*, vol. 250, no. 2, pp. 187–196, 2009.

[58] A. D. Baheti, S. H. Tirumani, A. B. Shinagare, M. H. Rosenthal, J. L. Hornick, N. H. Ramaiya, and B. M. Wolpin, "Correlation of ct patterns of primary intrahepatic cholangiocarcinoma at the time of presentation with the metastatic spread and clinical outcomes: retrospective study of 92 patients," *Abdominal imaging*, vol. 39, no. 6, pp. 1193–1201, 2014.

[59] MATLAB, *9.11.0.1809720 (R2021b) Update 1*. Natick, Massachusetts: The MathWorks Inc., 2021.

[60] R. C. Team *et al.*, "R: A language and environment for statistical computing," 2013.

[61] A. Agresti, *Categorical data analysis*. John Wiley & Sons, 2003.

[62] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*. New York: Springer, fourth ed., 2002. ISBN 0-387-95457-0.

[63] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[64] J. A. Damen, L. Hooft, E. Schuit, T. P. Debray, G. S. Collins, I. Tzoulaki, C. M. Lassale, G. C. Siontis, V. Chiocchia, C. Roberts, *et al.*, "Prediction models for cardiovascular disease risk in the general population: systematic review," *bmj*, vol. 353, 2016.

[65] G. S. Collins, J. A. de Groot, S. Dutton, O. Omar, M. Shanyinde, A. Tajar, M. Voysey, R. Wharton, L.-M. Yu, K. G. Moons, *et al.*, "External validation of multivariable prediction models: a systematic review of methodological conduct and reporting," *BMC medical research methodology*, vol. 14, no. 1, pp. 1–11, 2014.

[66] W. Bouwmeester, N. P. Zuithoff, S. Mallett, M. I. Geerlings, Y. Vergouwe, E. W. Steyerberg, D. G. Altman, and K. G. Moons, "Reporting and methods in clinical prediction research: a systematic review," *PLoS medicine*, vol. 9, no. 5, p. e1001221, 2012.

[67] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N project report, Stanford*, vol. 1, no. 12, p. 2009, 2009.

[68] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," *Jama*, vol. 247, no. 18, pp. 2543–2546, 1982.

[69] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, pp. 849–856, 2002.

[70] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[71] J. T. Rich, J. G. Neely, R. C. Paniello, C. C. Voelker, B. Nussenbaum, and E. W. Wang, "A practical guide to understanding kaplan-meier curves," *Otolaryngology—Head and Neck Surgery*, vol. 143, no. 3, pp. 331–336, 2010.

[72] J. Martin, A. Petrillo, E. C. Smyth, N. Shaida, S. Khwaja, H. Cheow, A. Duckworth, P. Heister, R. Praseedom, A. Jah, *et al.*, "Colorectal liver metastases: Current management and future perspectives," *World Journal of Clinical Oncology*, vol. 11, no. 10, p. 761, 2020.

[73] T. O'Rourke, F. Welsh, P. Tekkis, N. Lyle, A. Mustajab, T. John, D. Peppercorn, and M. Rees, "Accuracy of liver-specific magnetic resonance imaging as a predictor of chemotherapy-associated hepatic cellular injury prior to liver resection," *European Journal of Surgical Oncology (EJSO)*, vol. 35, no. 10, pp. 1085–1091, 2009.

[74] A. Gangi and S. C. Lu, "Chemotherapy-associated liver injury in colorectal cancer," *Therapeutic Advances in Gastroenterology*, vol. 13, p. 1756284820924194, 2020.

[75] Y. Tong, Y. Zhu, Y. Zhao, Z. Shan, J. Zhang, and D. Liu, "Tumor regression grade predicts survival in locally advanced gastric adenocarcinoma patients with lymph node metastasis," *Gastroenterology research and practice*, vol. 2020, 2020.

[76] C. Chen, A. Liaw, L. Breiman, *et al.*, "Using random forest to learn imbalanced data," *University of California, Berkeley*, vol. 110, no. 1-12, p. 24, 2004.

[77] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot, "Vsurf: an r package for variable selection using random forests," *The R Journal*, vol. 7, no. 2, pp. 19–33, 2015.

[78] V. Nardone, A. Reginelli, C. Guida, M. P. Belfiore, M. Biondi, M. Mormile, F. B. Buonamici, E. Di Giorgio, M. Spadafora, P. Tini, *et al.*, "Delta-radiomics increases multicentre reproducibility: a phantom study," *Medical Oncology*, vol. 37, no. 5, pp. 1–7, 2020.

List of Figures

1 Cholangiocarcinomas are classified according to the anatomical location into intrahepatic (ICC), perihilar (PCC) and distal (DCC) (figure taken from [9]). 2

1.1 Sketch of the procedure used to extract the radiomic features from the CT scans. 10

1.2 Example of a region of interest (ROI) segmented in a CT scan. 12

3.1 Sketch of the patient representation used in the S2GC model when core and margin views of the Portal phase of the CT scans is used. 42

3.2 Left panel: the number of iterations needed by the S2GC algorithm to converge for the different values of γ. Right panel: c-index variation according to γ. Both in the case where we are studying the overall survival using only the Portal phase of the CT scans. 43

3.3 C-index variation as function of η for some fixed values of λ in the case where we are studying the overall survival using only the Portal phase of the CT scans. 44

3.4 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core once γ and λ have been fixed to their optimal value γ^* and λ^* 45

3.5 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin once γ and λ have been fixed to their optimal value γ^* and λ^* 46

3.6 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core once γ is fixed as γ^* and $\lambda = 0.2041$ 47

3.7 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin once γ is fixed as γ^* and $\lambda = 0.2041$ 48

3.8 Values of the first five eigenvalues of the graph Laplacian 49

3.9 Sketch of the patient representation used in the S2GC model when all the three phases of the CT scans are used. 51

3.10 Left panel: the number of iterations needed by the S2GC algorithm to converge for the different values of γ. Right panel: the c-index variation according to γ. Both in the case where we are studying the overall survival using all the three phases of the CT scans. 52

3.11 C-index variation as function of η for some fixed values of λ in the case where we are studying the overall survival using all the three phases of the CT scans. 53

3.12 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core of the Portal phase once γ and λ have been fixed to their optimal value γ^* and λ^* 55

3.13 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin of the Portal phase once γ and λ have been fixed to their optimal value γ^* and λ^* 56

3.14 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core of the Arterial phase once γ and λ have been fixed to their optimal value γ^* and λ^* 57

3.15 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin of the Arterial phase once γ and λ have been fixed to their optimal value γ^* and λ^* 58

3.16 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core of the Late phase once γ and λ have been fixed to their optimal value γ^* and λ^* 59

3.17 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin of the Late phase once γ and λ have been fixed to their optimal value γ^* and λ^* 60

3.18 Patient-to-patient similarity graph variation according with λ while γ and η are fixed to their optimal values 62

3.19 Patient-to-patient similarity graph variation according with η while γ and λ are fixed to their optimal values 62

3.20 Values of the first nine eigenvalues of the graph Laplacian 63

3.21 Left panel: the number of iterations needed by the S2GC algorithm to converge for the different values of γ. Right panel: the c-index variation according to γ. Both in the case where we are studying the disease free survival time using all the three phases of the CT scans. 65

3.22 C-index variation as function of η for some fixed values of λ in the case where we are studying the disease free survival using all the three phases of the CT scans. 66
3.23 Zoom of the blue line, that corresponds to $\lambda = 0$, for small value of η. It is useful to study the optimal value of this parameter 67
3.24 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core of the Portal phase once γ and λ have been fixed to their optimal value γ^* and λ^* 68
3.25 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin of the Portal phase once γ and λ have been fixed to their optimal value γ^* and λ^* 69
3.26 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core of the Arterial phase once γ and λ have been fixed to their optimal value γ^* and λ^* 70
3.27 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin of the Arterial phase once γ and λ have been fixed to their optimal value γ^* and λ^* 71
3.28 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the core of the Late phase once γ and λ have been fixed to their optimal value γ^* and λ^* 72
3.29 Weights of the Cox model w as function of the L1 regularization parameter η for the variables extracted from the margin of the Late phase once γ and λ have been fixed to their optimal value γ^* and λ^* 73
3.30 Patient-to-patient similarity graph variation according with λ 74
3.31 Patient-to-patient similarity graph variation according with η 74
3.32 Values of the first nine eigenvalues of the graph Laplacian 75

4.1 Kaplan-Meier curves estimating clusters' survival probability 81
4.2 Boxplot of the maximum dimension for the four groups 82
4.3 Percentage of patients presenting the different values of Pattern for the four groups 83
4.4 Percentage of patients presenting a single nodule for the different groups . 83
4.5 Percentage of patients presenting a high value of comorbities for the different groups 84
4.6 Percentage of patients who underwent the different therapies for the four groups 85

4.7 Comparison between the weights of the core and the margin for three of the most relevant variables. 87
4.8 Kaplan-Meier curves estimating clusters' survival probability for the Overall Survival when using all the three phases of the CT scans 88
4.9 Percentage of patient presenting: severe complications, R status and Microscopic vascular invasion for all the five groups 91
4.10 Percentage of patient presenting the different values of Grading for all the five groups 92
4.11 Percentage of patient affected by a Metastatic Disease for all the groups . . 92
4.12 Patient-to-patient similarity graph variation according with λ where the colours represent the groups found with the optimal parameters. 94
4.13 Patient-to-patient similarity graph variation according with η where the colours represent the groups found with the optimal parameters. 94
4.14 Comparisons between the mean values of some radiomic features for Group 1, Group 2 and the smaller groups. 96
4.15 (a): comparison between the mean values of five radiomic features for group 3, group 4 and group 5. (b): comparison between the values of five radiomic features for all the patients of group 3, group 4 and group 5. 97
4.16 Comparison between the weights of the core and the margin of the different phases of the CT scans. 99
4.17 Kaplan-Meier curves estimating clusters' survival probability for the Disease Free Survival when using all the three phases of the CT scans 100
4.18 Boxplots of the Number of lymph nodes removed for the four groups . . . 101
4.19 Percentage of patient that have undergone: Adjuvant Chemotherapy, Major Hepatectomy for all the groups 103
4.20 Percentage of patient presenting the different values of Pattern for the three most relevant groups 104
4.21 Patient-to-patient similarity graph variation according with λ where the colours represent the groups found with the optimal parameters. 106
4.22 Patient-to-patient similarity graph variation according with η where the colours represent the groups found with the optimal parameters. 106
4.23 (a): comparison between the mean values of four radiomic features for group 1, group 2 and group 3. (b): comparison between the mean values of four radiomic features for group 1, group 2 and group 4. 107
4.24 (a): comparison between the mean values of five radiomic features for Group 3 and Group 4. (b): comparison between the values of five radiomic features for all the patients of Group 3 and Group 4. 108

4.25 Comparison between the weights of the core and the margin of the different phases of the CT scans. 111

6.1 Importance plot of the variables used in the random forest model. 128

List of Tables

1.1 Statistical summary of the numeric clinical variables 23
1.2 Percentage of patients presenting a given clinical characteristic 23

2.1 Results of the three logistic regression model implemented to predict whether the patient is dead during the time of the study 30

4.1 P-values of the tests performed on the mean or on the proportion of the clinical variables in the different groups 81
4.2 Weights of the ten most relevant radiomic features for the application to only the Portal phase of the CT scans 86
4.3 Group characterization according to the exogenous clinical variables (numeric or discrete) with the p-values of the tests performed on the mean of these variables in the different groups. 89
4.4 Group characterization according to the exogenous clinical variables (dichotomous or categorical) with the p-values of the tests performed on the proportion of these variables in the different groups. 90
4.5 P-values of the tests performed on the mean or on the proportion of the clinical variables in the different groups 90
4.6 Weights of the ten most relevant radiomic features for the application to all the phases of the CT scans for the Overall Survival 98
4.7 Group characterization according to the exogenous clinical variables (numeric or discrete) with the p-values of the tests performed on the mean of these variables in the different groups 101
4.8 Group characterization according to the exogenous clinical variables (dichotomous or categorical) with the p-values of the tests performed on the proportion of these variables in the different groups 102
4.9 P-values of the tests performed on the mean or on the proportion of the clinical variables of the three major groups 103
4.10 Weights of the ten most relevant radiomic features for the application to all the phases of the CT scans for the Diseases Free Survival 109

6.1 Mean accuracy and standard deviation of the three models fitted to predict the TRG 127

Acknowledgements

First of all, I would like to thank my supervisor Prof. Francesca Ieva and my co-supervisor Dr. Lara Cavinato who helped me carrying out this project with their experience.

I would like to thank Dr. Luca Viganò and all his group for their support in this project and all their help on the clinical aspects of this work.

I would like to thank the surgical and radiological units of the participating centers (Policlinico Rossi, Verona; Mauriziano Hospital, Torino; Gemelli Hospital, Roma; S. Orsola Hospital, Bologna; Morgagni-Pierantoni Hospital, Forlì) that provided us the data on the patients affected by Intrahepatic cholangiocarcinoma.

Then, I would like to express my gratitude to my family and in particular to my father Andrea, my mother Paola and my sister Sara for their unconditional support.

I would like to thank my great friends Costanza, Daniele, Federica and Giorgio, without you and all the experiences shared together in these years I wouldn't be who I am now.

I would like to thank David and Salvatore, two of my oldest friends, for being always present when needed.

A special thanks goes to Sharon who stood by me during my master degree supporting me in all the difficult moments and celebrating for my successes.