POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Statistical assessment of Radiomics role in prognosis of patients with Intrahepatic Cholangiocarcinoma

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Author: **Noemi Rossi**

Student ID: 939699
Advisor: Prof. Francesca Ieva
Academic Year: 2020-21

Abstract

Intrahepatic cholangiocarcinoma (IHC) is an aggressive tumor whose incidence has increased considerably in recent years. The main treatment is surgery, but an optimal strategy has not yet been characterized and prognostic factors are still debated. Recently, a quantitative analysis approach based on diagnostic images, called Radiomics, has gained attention. This technique is able to extract a large amount of high-dimensional and minable data from diagnostic images, produced by Computed Tomography (CT) scans in our case, capturing additional information to that usually considered by clinicians to formulate prognoses. Thanks to the amount of data collected by Humanitas University, in a multicentre study involving six different hospitals, we have tried to understand the importance and the role of radiomics in predicting pathology data and survival response in patients with IHC. We also focused on the multilevel nature of the data, analysing whether there are differences between the various hospitals. In order to predict, and thus classify, pathology data taking into account its multicentre aspect, we focused on Generalized Linear Mixed Effects Models as modelling techniques. For time-to-event data we employed Cox type regression models with shared frailties, being able to consider the inherent grouping factor in the data. Furthermore, as we were provided with radiomic data covering three phases of CT (Arterial, Portal and Late) we have exploited the multiview aspect of the data by using Multi-view Learning techniques, to understand the importance of each of the three phases. With the analysis carried out within this thesis, we have demonstrated the importance of considering all radiomics information, together with clinical one, in order to have an adequate prognosis in patients with IHC.

Keywords: Intrahepatic Cholangiocarcinoma, Radiomics, Mixed Effects Models, Shared Frailty Models, Multi-view Learning

Sommario

Il colangiocarcimona intraepatico (IHC) è un tumore aggressivo la cui incidenza è cresciuta notevolmente negli ultimi anni. Il trattamento principale è la chirurgia, tuttavia una strategia ottimale non è ancora stata caratterizzata e i fattori prognostici sono tuttora dibattuti. Recentemente, un approccio basato sull'analisi quantitativa di immagini diagnostiche, chiamato Radiomica, ha guadagnato sempre più attenzione. Con questa tecnica si è infatti in grado di estrarre dalle immagini diagnostiche, prodotte dagli scan della Tomografia Computerizzata (TC) nel nostro caso, una grande quantità di *high-dimensional and minable data*, riuscendo a catturare informazioni aggiuntive rispetto a quelle solitamente considerate dai medici per formulare prognosi. Grazie ai dati raccolti da Humanitas University, in uno studio multicentrico comprendente sei diversi ospedali, abbiamo cercato di capire l'importanza e il ruolo che ha radiomica nella predizione dei dati patologici e della risposta di sopravvivenza nei pazienti con IHC. Abbiamo focalizzato l'attenzione anche sull'aspetto multi-livello del dato, analizzando l'eventuale presenza di differenze tra i vari ospedali. Per poter predire, e quindi classificare, i dati patologici tenendo conto dalla natura multicentrica dei dati, ci siamo focalizzati sui modelli lineari generalizzati a effetti misti come tecnica modellistica. Per i dati time-to-event abbiamo impiegato modelli di sopravvivenza di tipo Cox con shared frailties, con i quali siamo riusciti a considerare il fattore di raggruppamento intrinseco nei dati. Inoltre, siccome ci sono stati forniti dati radiomici riguardanti tre fasi della TC (arteriosa, portale e tardiva), abbiamo pensato all'aspetto multiview del dato utilizzando tecniche di Multi-view Learning per poter capire l'importanza di ognuna delle tre fasi. Con le analisi svolte in questa tesi, abbiamo dimostrato l'importanza di considerare tutte le informazioni fornite dalla radiomica, insieme a quelle cliniche, al fine di avere una prognosi adeguata nei pazienti con IHC.

Parole chiave: Colangiocarcinoma Intraepatico, Radiomica, Modelli Cox, Modelli Shared Frailty, Multi-view Learning

Contents

Introduction

In recent years, the incidence of a disease named Intrahepatic Cholangiocarcinoma (IHC) has increased. IHC is an aggressive tumor that affects the liver and its five-years survival rate ranges from 25% to 40%. The prognostic factors associated to IHC are still debated, robust biomakers are lacking and a precision medicine approach with an adequate non-invasive preoperative assessment of tumor biology is still not available. Recently, a new technique called Radiomics, that allows to extract a large amount of quantitative data from diagnostic images, has arisen. With Radiomics we are able to mine information not currently considered by clinicians to formulate prognoses, that can be used in conjunction with traditional clinical ones.
The data analysed within this thesis are collected by Humanitas University and are related to patients that have undergone a liver resection for IHC. These data are embedded in a multicentre study that aims to understand the importance and the role of radiomics in predicting the targeted outcomes. The objective of this work is developed robust models capable of classifying pathology data and predicting survival response in patients with IHC. Building the models, we want to understand if radiomics usage can be decisive in increasing their predictive ability, focusing on the importance of every radiomic data provided. Indeed, at first we aim to investigate whether both tumour radiomics (core) and peritumour area (margin) contribute to improve performances. Afterwards, as we are provided with radiomic data covering three phases of Computed Tomography scan (the technique used for collecting the diagnostic images) we are interested in analyse the importance of each of these three phases. In particular, we want to understand if they carry the same information or if each one gives its own contribution in prediction. Moreover, in performing these tasks, the multicentre nature of the data present in this study needs to be considered, discovering if there are differences among centres.

The rest of this thesis is structured as follows.
In Chapter 1 we describe in detail the problem setting, the dataset used and the clinical questions we aim to answer within this work. After giving this preliminary information, we describe data engineering procedures used to prepare our dataset for the analysis.

In Chapter 2 we deal with the classification of pathology data. In the first part of the chapter, the methodologies used to build the models for classification are described, i.e. Logistic Regression and Mixed Effects Models. With Logistic Regression we select the best models for describing the outcomes, trying several variable selection techniques. With Mixed Effects Model we take into account the multilevel nature of the data in the best models determined, understanding if differences between hospitals are present. At the end, results and conclusion about the classification are provided.
In Chapter 3 we handle Survival Analysis. In the first part of the chapter, the methodologies used to build the model for survival responses are described, i.e. Cox Proportional Hazards (Cox-PH) models and Shared Frailty models. With Cox-PH models we select the best models for describing the outcomes. With Shared Frailty models we take into account the multilevel nature of the data in the best models determined, understanding if differences between hospitals are present. At the end, results and conclusion about the survival analysis are provided.
In Chapter 4 we consider all the information provided by radiomics, adopting a multiview approach to represent the data. In the first part of the chapter, the techniques adopted to perform Multiview Dimensionality Reduction of the data are described, i.e. Multiview Canonical Correlation Analysis and Kernel Multiview Canonical Correlation Analysis. Afterwards, results of classification and survival analysis performed considering the multiview aspect are provided, together with considerations regarding the usefulness of exploiting all radiomic information, possibly adopting a multiview approach.
The work is concluded by a discussion of the results obtained, which provides the answer to the proposed clinical questions.

The analysis present in this thesis are carried out using R [1] and Python [2] programming languages.

1 | Problem Setting and Data Engineering

In this Chapter we give an overview of a disease named Intrahepatic Cholangiocarcinoma (IHC) and describe the associated clinical dataset provided by a collaboration with Humatitas University that is used within this thesis. After introducing the database, we explain the objectives of this thesis and the techniques used to answer the proposed clinical questions. At the end, we describe the data engineering needed to prepare the data for further analysis.

1.1. Multicentre Study for Intrahepatic Cholangiocarcinoma

Intrahepatic Cholangiocarcinoma (IHC) is an aggressive disease that affects the liver arising from the bile duct epithelium, with anatomical position that is difficult to access. It is the second most common primary hepatic tumor, accounting for less than 2% of all malignancies. Its incidence is increasing over last decades and diagnosis is difficult at early stages, due to IHC complicated biology [3–5]. The main treatment is surgery, chemotherapy has a limited effectiveness and an optimal strategy for patients with resectable IHC is not well characterized. Five-years survival rate ranges from 25% to 40% [6–9]. The main factors associated with prognosis are IHC size, number and distribution; tumor differentiation; vascular invasion; lymph node metastases; metabolic tumor volume; R status. However, these prognostic factors are still debated, robust biomarker are lacking and precision medicine approach with an adequate non-invasive preoperative assessment of tumor biology and prognosis is still not available [10–13].

In recent years, an approach of non-invasive image-based tissue analysis, namely Radiomics, has emerged [14, 15]. Radiomics is able to capture additional information not currently considered by clinicians in making prognosis, and more specifically, it recognizes patterns which are related to clinical feature. In this way, relevant data are mined

from any image modalities such as computed tomography (CT) and magnetic resonance imaging (MRI) and these data can be used jointly to traditional clinical information. In literature concerning IHC, it has been demonstrated that models including radiomic features outperform traditional ones, predicting pathology data and patients' outcome with high accuracy [16–18].

In this thesis, we aim to build robust models that are capable to predict pathology data and survival response in patients with IHC, making use of the information provided by patients' clinical history and radiomics. In particular, we want to understand if radiomics can be an added value in the prognosis of IHC, looking for evidence in our data that shows that using radiomics together with clinical information leads to improved predictive performance. We are not only interested in understanding the impact of radiomics in its entirety, but we also want to comprehend the usefulness of an investigation focused not only on the core part of the tumour but also on the marginal peritumoral zone.

1.2. IHC Dataset

The data analysed in this thesis are provided by Department of Hepatobiliary and General Surgery of Humanitas Clinical and Research Center, based in Rozzano. The information provided concerns patients that have undergone a liver resection for IHC confirmed at final pathology, from 2009 to 2019. The patients comply strictly inclusion/exclusion criteria and come from 6 different hospital centres:

- Humanitas Clinical and Research Center (Milano) with 83 patients
- Mauriziano Hospital (Torino) with 73 patients
- Policlinico Rossi (Verona) with 43 patients
- S. Orsola Hospital (Bologna) with 28 patients
- Gemelli Hospital (Roma) with 25 patients
- Ospedale Morgagni-Pierantoni (Forlì) with 9 patients

Therefore, 261 patients are present in the final cohort. Each row of the dataset corresponds to a patient with an associated anonymous code to identify him/her while maintaining privacy. In addition, the hospital which they belong to, described by the variable *CENTRE*, was recorded for each of them. Information collected for each individual concerns clinical and radiomic characteristics.

1.2.1. Clinical Variables

Clinical variables include the features illustrated in Table 1.1 and their types are summarised in Table 1.2.

Table 1.1: Description of the clinical variables in the IHC dataset

Variable name	Description
CENTRE	Hospital of origin of the patient
ID CODE	Anonymous code that identifies the patient
AGE	Patient's age
SEX	Patient's sex
HBV	Presence of the hepatitis B virus in the patient
HCV	Presence of the hepatitis C virus in the patient
CIRRHOSIS	Presence of Cirrhosis in the patient
CA 19-9	Value of the tumor maker Ca 19-9
CA 19-9 $\geq$ 55	Binary variables that identifies when Ca 19-9 is greater or equal than 55
NEOADJUVANT CHEMOTHERAPY	Variable that indicates if the patient has undergone chemotherapy before surgery
FIRST RESECTION	Variable that indicates if the patient has undergone first resection before major surgery treatment
MAJOR HEPATECTOMY	Variable that indicates if the patient has undergone major hepatectomy before major surgery
BILIARY RESECTION	Variable indicating whether a biliary resection was performed during surgery
LYMPHADENECTOMY	Variable indicating whether a lymphadenectomy was performed during surgery
ASSOCIATED RESECTION	Variable indicating whether only part of the liver has been removed during surgery
SEVERE COMPLICATIONS	Variable indicating whether the patient experienced severe complications after surgery
PATTERN	Variable with values in [0,1,2] that describes the number and location of tumor
DIMENSION	Patient's IHC maximum dimension in mm
SINGLE NODULE	Variables that indicates if there is only one tumor
T VII ed	Variables with values in [1a,1b,2,3,4] that indicates the extension of the tumor
R status	Variable with values in [0,1,2] that indicates the presence of tumor residuals

M status	Variable that indicates the presence of metastasis
N status	Variable with values in [0, x ,1] that indicates the presence and extension of regional lymph nodes. 'x' indicates between 0 and 1 but correct value impossible to determine
GRADING	Variable with values in [1,2,3] that indicates the aggressiveness of the tumor
MICROSCOPIC VASCULAR INVASION	Variable that indicates the presence of Microscopic Vascular Invasion
PERINEURAL INFILTRATION	Variable that indicates the presence of Perineural Infiltration
SATELLITE NODULES	Variables that indicates the presence of Satellite Nodules
ADJUVANT CHEMOTHERAPY	Variables that indicates if the patient has undergone chemotherapy after surgery
MORTALITY	Censoring state of death observation
RECURRENCE	Censoring state of recurrence observation
OVERALL SURVIVAL	Days until death/censoring
RELAPSE FREE SURVIVAL	Days until recurrence/censoring

Table 1.2: Types of the clinical variables in the IHC dataset

Variable name	Type	Timing
CENTRE	Categorical	Preoperative
ID CODE	String	Preoperative
AGE	Numerical	Preoperative
SEX	Binary	Preoperative
HBV	Binary	Preoperative
HCV	Binary	Preoperative
CIRRHOSIS	Binary	Preoperative
CA 19-9	Numerical	Preoperative
CA 19-9 $\geq$ 55	Binary	Preoperative
NEOADJUVANT CHEMOTHERAPY	Binary	Preoperative
FIRST RESECTION	Binary	Preoperative
MAJOR HEPATECTOMY	Binary	Preoperative
BILIARY RESECTION	Binary	Postoperative
LYMPHADENECTOMY	Binary	Postoperative
ASSOCIATED RESECTION	Binary	Postoperative
SEVERE COMPLICATIONS	Binary	Postoperative
PATTERN	Ordinal	Preoperative
DIMENSION	Numerical	Preoperative
SINGLE NODULE	Binary	Preoperative
T VII ed	Categorical	Postoperative

R status	Categorical	Postoperative
N status	Categorical	Postoperative
M status	Binary	Postoperative
GRADING	Categorical	Postoperative
MICROSCOPIC VASCULAR INVASION	Binary	Postoperative
PERINEURAL INFILTRATION	Binary	Postoperative
SATELLITE NODULES	Binary	Postoperative
ADJUVANT CHEMOTHERAPY	Binary	Postoperative
MORTALITY	Binary	Postoperative
RECURRENCE	Binary	Postoperative
OVERALL SURVIVAL	Numerical	Postoperative
RELAPSE FREE SURVIVAL	Numerical	Postoperative

Clinical variables correspond to information that is known to clinicians without the use of radiomics. Some of these features, called preoperative, correspond to details that are available prior to the curative surgery, while postoperative covariates are derived from information obtained from histological pathological samples after surgery.

1.2.2. Radiomics

Nowadays, with high-throughput computing it is possible to rapidly extract quantitative features from tomographic images. This conversion from digital medical images into mineable high-dimensional data is called radiomics. This process is motivated by the fact that the medical image contains underlying pathophysiology information that can be revealed via quantitative image analysis. This approach is in contrast to the traditional practice of treating medical images as pictures intended solely for visual interpretation. Quantitative image features based on intensity, shape, size or volume, and texture have the potential to offer information on tumor phenotype and microenvironment, that is distinct from that provided by clinical reports and laboratory test results [19]. Radiomics features can be used in conjunction with other current information related to clinical history for the patient, to model clinical outcomes, for evidence-based clinical decision support. Therefore, radiomics has the potential of improving diagnostic, prognostic, and predictive accuracy.

Once the diagnostic image is available, the so-called Volume of Interest (VOI) is identified. It corresponds to the region of the image where tumor and suspected tumor are present. The most crucial and challenging part of radiomics, i.e. segmentation, takes place in the area selected by VOI. Segmentation defines the segmented volume in which feature data are generated. The critical part of this work is that not all tumour types have the distinct border, and this makes it difficult to make segmentation. Once the contour

has been defined, using voxels (pixels) information, high-dimension quantitative information is extracted producing radiomics features. The procedure employed by radiomics is represented in Figure 1.1.

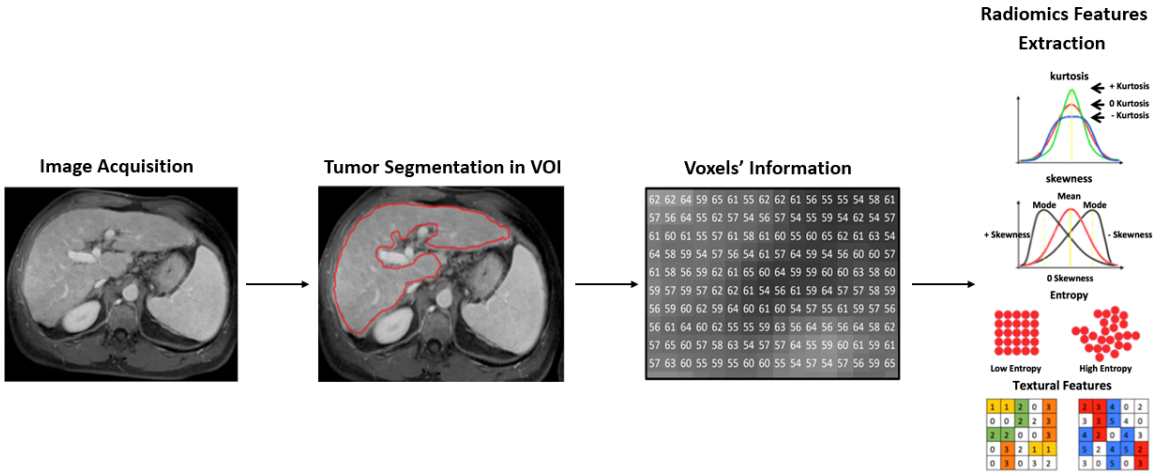


Figure 1.1: Flowchart of the radiomic process

In this study, radiomics is applied to patients' preoperative CT scans. CT is carried out in three different phases:

- **Arterial:** it is the first phase that is acquired 20-30 seconds after contrast agent infusion. It allows to see hypervascularized lesions from branches of the hepatic artery.
- **Portal:** it is the second phase that is acquired 60-80 seconds after contrast agent infusion. It evaluates which lesions are vascularized by branches of the portal vein.
- **Late:** it is the third phase that evaluates how the contrast medium is discarded.

For each of these phases, tumour segmentation was done by generating a VOI that included not only the core part of the IHC, but also the marginal peritumoral zone. In this way, radiomics is extracted from two different zones, producing two different insight of the IHC:

- **Core**: it corresponds to the area where the tumour is identified.
- **Margin**: it corresponds to the surrounding the tumour.

For each of these phases, for both core and margin, 50 radiomic variables are collected, producing 300 covariates in total. The procedure is outlined in Figure 1.2.

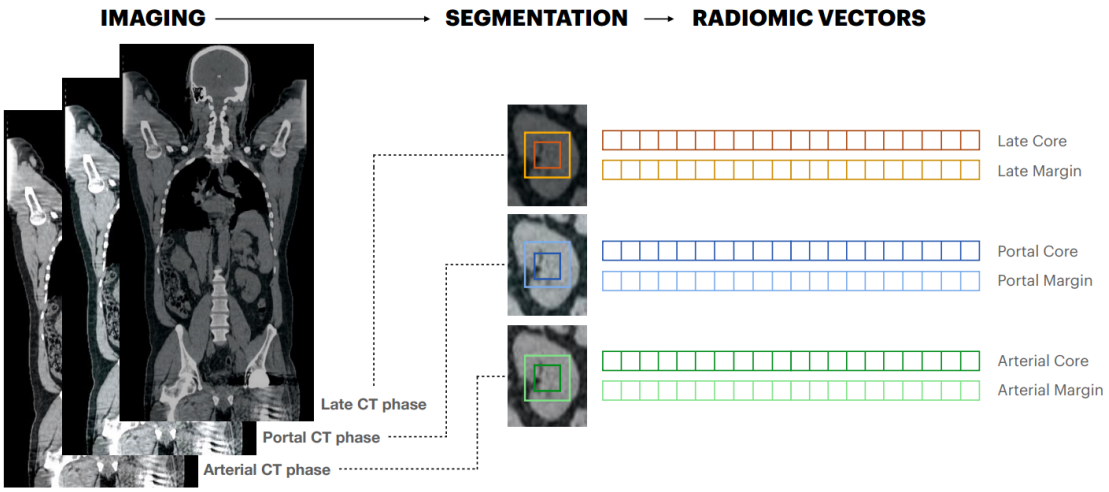


Figure 1.2: Representation procedure of radiomic features extraction from the three phases of CT scan for both core and margin areas

Radiomic variables represents how voxels intensity values are distributed in the target area and are divided into:

- Basic parameters that are intensity related
- First order parameters that describe how the voxel are distributed in an image
- Second order parameters, i.e. matrices (GLCM, GRLM, NGLDM, GLSZM) that are obtained imposing filter grids on the image to extract repetitive and nonrepetitive patterns

The complete set of radiomic variables and the list of filters used for the second order covariates are given in the Appendix A.
The data recorded for the three CT scan phases produces three different representations of the tumor. The main representation is the one obtained analysing the Portal phase: it is used to decide treatment and surgery procedure and make prognosis. It is always taken into account, possibly, in conjunction with the other phases. For this reason, Portal Phase is registered for all patients, while the other two phases present several missing records, whose numbers are detailed in the Table 1.3. Because of this, in the first part of the work only the Portal phase, with core and margin insights, will be initially be considered. Afterwards, all three phases of radiomics will be analysed.

	N° Missing Patients	(%)	N° Available Patients	(%)
PORTAL	1	0.4%	260	99.6%
ARTERIAL	32	12%	229	88%
LATE	46	18%	215	82%
PORTAL + ARTERIAL + LATE	58	22%	203	78%

Table 1.3: Numbers and percentages of missing and available patients in every phase of the CT scan, individually and jointly.

1.2.3. Multiview interpretation of Radiomics

The fact of considering all three phases of the CT scan together, which are nothing else than the description of the same object in three different moments, led us to reflect on what is the most correct way to represent the patient with this data. The easiest way to use all information provided by radiomics is to concatenate all covariates belonging to the three phases producing a very large dataset. However, this concatenation causes overfitting, since we have a small size training sample, and does not take into account the fact that the three radiomic phases basically represent, even if in different ways, the same subject. This led us to think the three radiomic phases as three different *views* that describes the same instance, considering *Multi-view Learning* techniques to model the data. With Multi-view Learning we are able to exploit the redundant nature of the views, since it aims to learn one function to model each view and jointly optimizes all the functions to improve the generalization performance [20]. At the end of this work, Multi-view representation will be considered in Chapter 4 to describe patients considering all available information and multiview nature of the data.

1.3. Outcomes and Endpoints

The outcomes that will be analysed within this thesis are the following clinical covariates:

- **Microscopic Vascular Invasion (MVI)**: it is a binary variable that assess the presence or not of microscopic vascular invasion.
- **Grading**: it is a categorical variable with values {1, 2, 3} that describes the aggressiveness of the tumor. Even if there are three possible values, we are interested

in the two classes classification 1-2 vs 3.

- **Overall Survival (OS)**: OS is a numerical variable that represents the overall survival time of the patient, measured in days. A vector named Mortality defines the related censoring states of the patient at the end of the study (2019).
- **Relapse-Free Survival (RFS)**: RFS is a numerical variable that represents the relapse-free time of the patient, measured in days. A vector named Recurrence defines the related censoring states of the patient at the end of the study (2019).

Given the nature of the outcomes, i.e. binary variables for pathology outcomes and time-to-event censored data for survival outcomes, two different modelling approaches must be used: at first we will focus a *Classification* problem in Chapter 2, then a *Survival Analysis* one in Chapter 3. For both Classifications and Survival we need to find robust models that are capable of predicting the outcomes using clinical and radiomic covariates. Regarding radiomics, we want to understand if its use can be decisive in increasing the predictive ability of the model. In particular, we aim to investigated if both radiomics of tumor (core) and peritumoral area (margin) contribute both in performance improvement. Initially, in Chapters 2 and 3, we only focus on the Portal phase of the CT scan to answer these questions, as this is the main phase on which decisions are made and as we have data for all patients. To understand the impact that radiomics may have on the predictive ability of the model, we test the models with different sets of input covariates: we start considering the clinical covariates alone, which correspond to the information used by clinicians to make prognoses, and add the radiomic covariates corresponding to the tumour and then also those corresponding to the margin area. Once this task is completed, the next step is to understand whether the different phases of the CT scan provide the same information or whether each one adds its own value to the prognostic impact of the model. This aspect is explored in Chapter 4.
In performing these tasks, the multicentre aspect present in this study needs to be considered, modelling the grouping of the hospitals, to understand if there are differences among centres.
Classification problem is addressed in Chapter 2 using Logistic Regression and Mixed Effects Models as modelling techniques. With Logistic Regression we select the best models for describing MVI and Grading, choosing among different sets of input covariates (reported in Section 2.2) and trying several variable selection techniques (listed in Section 2.1.1). Using Logistic Regression the multicentre aspect present in the data is not considered, but it is necessary to use this procedure as a preliminary step in order to perform feature selection identifying which of the covariates are extracted in the best model. Once these covariates are determined, they enter as input into a Mixed Effects Model. With

Mixed Effects Models we are able to perform classification taking care of the multilevel nature of the data, producing the final models. At the end of the Chapter, results are illustrated and conclusions are drawn analysing predictive performances of the models.
Survival Analysis is addressed in Chapter 3 using Cox Proportional Hazards Models (Cox-PH) and Shared Frailty Models as modelling techniques. With Cox-PH model we select the best models for describing OS and RFS, choosing among different sets of input covariates (reported in Section 3.2) and using Stepwise Selection to reduce the number of features. Using Cox-PH models the hierarchy of the data is not considered, but it is necessary to apply this method as a preliminary step to perform variable selection, in order to identify which of covariates are extracted in the best models. Once covariates present in the best models are determined, we take care of the multilevel nature of the data using Shared Frailty Models producing the final models. At the end of the Chapter, results are reported and clinical questions are addressed analysing predictive performances of the models.
Multi-view aspect of the data is exploited in Chapter 4. In this chapter two Multi-view Dimensionality Reduction approaches, Multiview Correlation Analysis and Kernel Multiview Correlation Analysis, are used to reduce the number of the feature in the dataset considering all radiomic phases jointly, accounting properly the multiview nature of the data. Performing Classification and Survival Analysis on Multi-view reduced datasets, we make conclusions about the usefulness of using all the information provided by radiomics.

1.4. Exploratory Analysis and Preprocessing

As previously mentioned, in our study there are 261 patients coming from 6 different centres. Moreover, we have 300 radiomic variables plus clinical ones. The large number of covariates leads to overfitting problems, that are taken into account with features selection and dimensionality reduction in Chapters 2, 3 and 4. The number of patients is not enough large to avoid overfitting and there is a risk that, with a high number of missing values, it could be reduced significantly. We will consider these issues by first dealing with missing values and possible imputation (Section 1.4.1), then we take care of possible presence of outliers (Section 1.4.2) and at the end we analyse the correlation among the radiomic features (Section 1.4.3).

1.4.1. Missing Values and Imputation

The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data [21]. Therefore,

missing values should be handled properly. In both the clinical and radiomic parts of the study missing data are present, due to the difficulty of collecting information.

At first, we consider missing values in radiomic data. These correspond to patients for whom there was not the possibility to collect all data concerning the three CT phases. Hence, in these data, for patients with missing values, all radiomic variables are lacking. Therefore, it is impossible to impute this data. The number of missing values of radiomic variables is summarized in Table 1.5. The amount of missing data is consistent. However, as there is no possibility of doing anything about these values, we focus on the analysis of the missing values in the clinical variables.
The distribution of this latter among the different features is summarized in Figure 1.3 and Table 1.4.

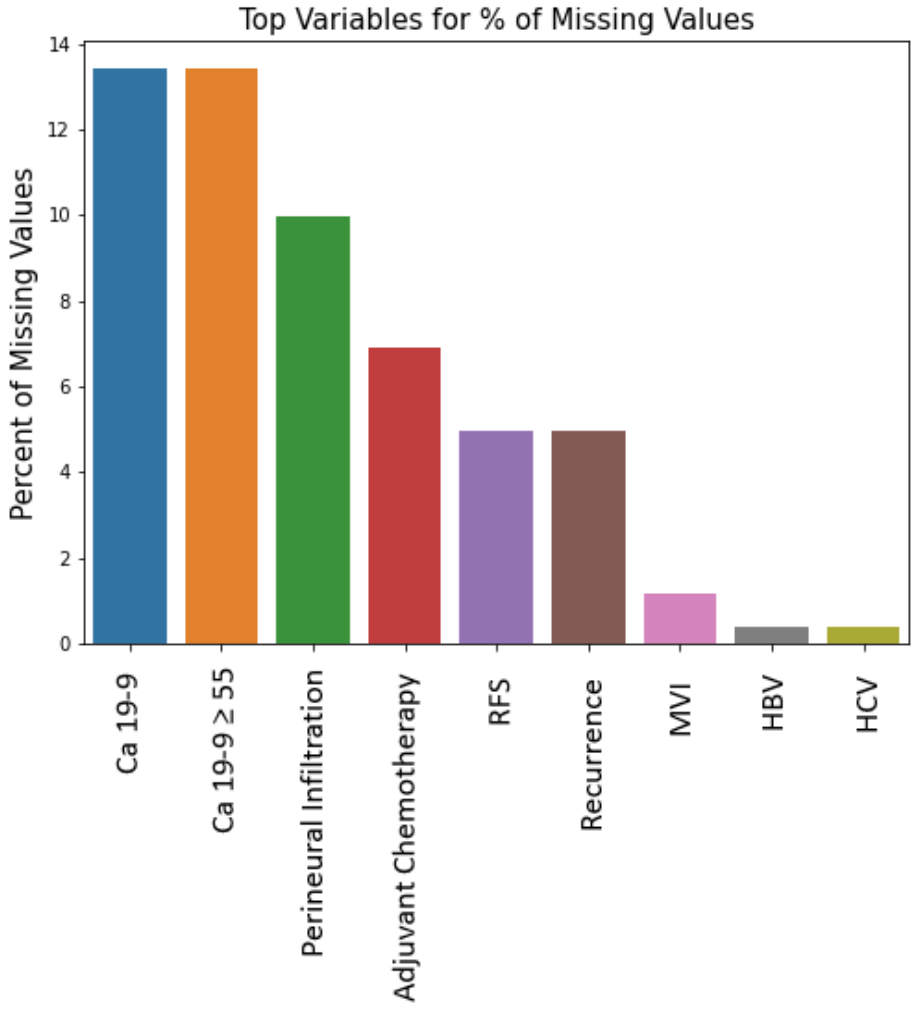


Figure 1.3: Barplot with percentage of Missing Values in Clinical Features of IHC dataset with at least one missing data.

	N° Missing Values	Missing Ratio
Ca 19-9	35	13.41%
PERINEURAL INFILTRATION	26	9.96%
ADJUVANT CHEMOTHERAPY	18	6.90%
RFS (Days)	13	4.98%
RECURRENCE	13	4.98%
MICROSCOPIC VASCULAR INVASION	3	1.15%
HBV	1	0.38%
HCV	1	0.38%

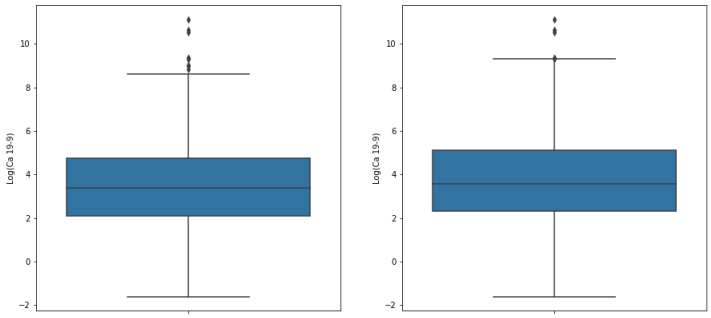
Table 1.4: Numbers and Percentages of Missing Values in Clinical Features in IHC dataset with at least one missing data.

	N° Missing Values	(%)	Remaining Patients	(%)
PORTAL	1	0.4%	260	99.6%
ARTERIAL	32	12%	229	88%
LATE	46	18%	215	82%
PORTAL + ARTERIAL + LATE	58	22%	203	78%

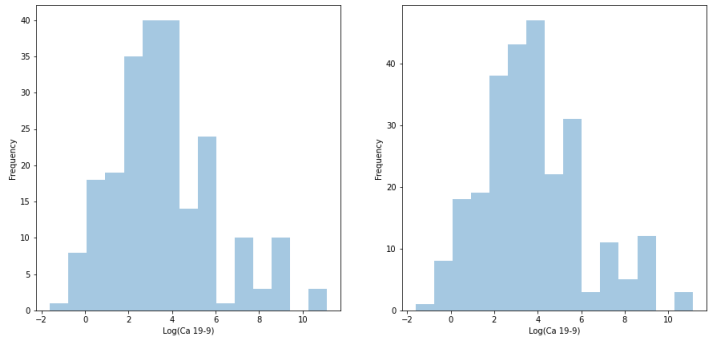
Table 1.5: Number and Percentage of Missing Values in the three phases of radiomics, individually and jointly.

A first strategy to deal with missing values could be to eliminate every patient in which at least one missing value is present in clinical variables, namely perform a listwise deletion. This would eliminate 76 patients, that correspond to 29% of the dataset. The number of missing values is not negligible and the approach of listwise deletion is not feasible. The strategy we used to solve the problem is Multiple Imputation. It is a methodology for the problem of missing data that aims to allow for the uncertainty about the missing data by creating several different plausible imputed data sets and appropriately combining

results obtained from each of them [22]. Multiple imputation has potential to improve the validity of medical research, helping by reducing bias or increasing precision.
In our study only missing values in the input covariates are imputed, while patients with missing data in at least one of the output coviariates (MVI, RFS and RECIDIVA) are deleted. To perform the Multiple Imputation from these data the Python package `miceforest` is used [23]. It fills the missing data through Multiple Imputation by Chained Equation. Using this technique 5 imputed datasets were created using almost all most important covariate and then aggregated into a single one averaging the values. After having produced the final imputed dataset, we took care to check that the distribution of imputed variables did not deviate too much from the original one. This was done by comparing the distribution of the imputed variables through barplots and boxplots.
Concering the numerical variable Ca19-9, results are reported in Table 1.4:



(a) Boxplot of Log(Ca19-9) distribution Original vs Imputed

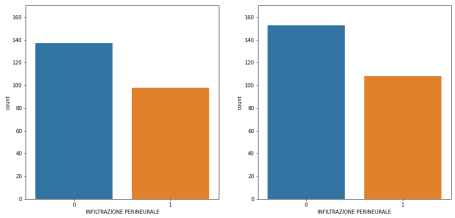


(b) Histogram of Log(Ca19-9) Original vs Imputed

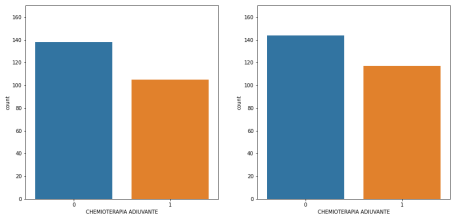
Figure 1.4: Comparison of Ca19-9 distribution in Original vs Imputed Data, using the logarithm function for visualization purposes.

For the categorical imputed variables, to assess the correctness of the imputation proce-

dure, it is aimed that the distribution of the input variables grouped by output covariates is as similar as possible. To verify this, the following barplots are provided in Figures 1.5, 1.6, 1.7:

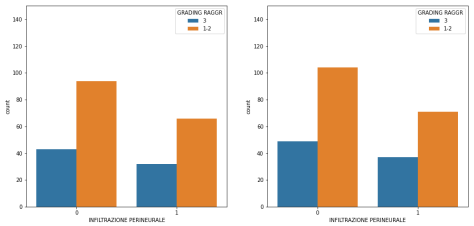


(a) Barplots of Perineural Infiltration Original vs Imputed

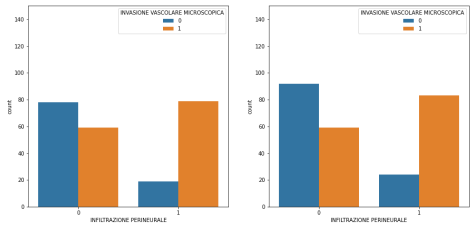


(b) Barplots of Adjuvant Chemotherapy Original vs Imputed

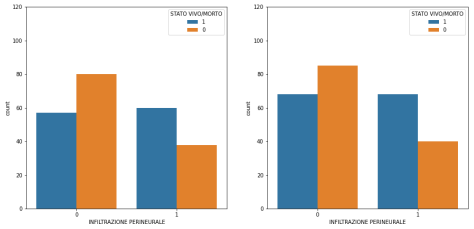
Figure 1.5: Barplots of Perineural Infiltration and Adjuvant Chemotherapy in Original vs Imputed Data



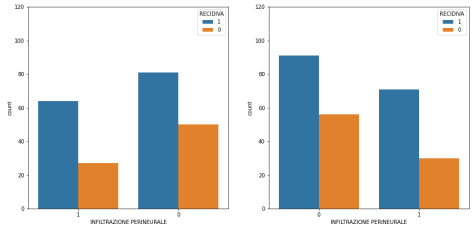
(a) Distribution of Grading 1-2 vs 3 in Perineural Infiltration



(b) Distribution of Microscopic Vascular Invasion in Perineural Infiltration

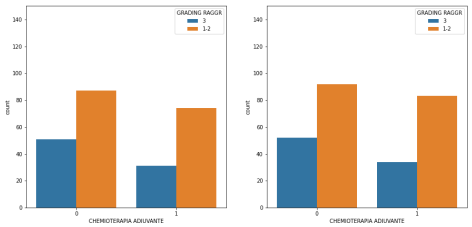


(c) Distribution of Mortality in Perineural Infiltration

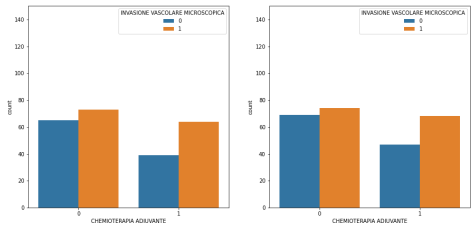


(d) Distribution of Recurrence in Perineural Infiltration

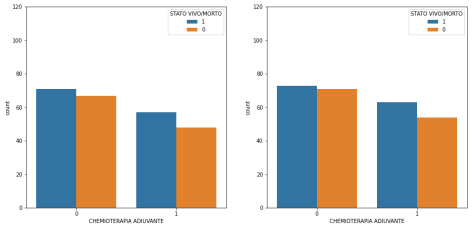
Figure 1.6: Comparison of distribution of Perineural Infiltration grouped by outcome in Original vs Imputed Data



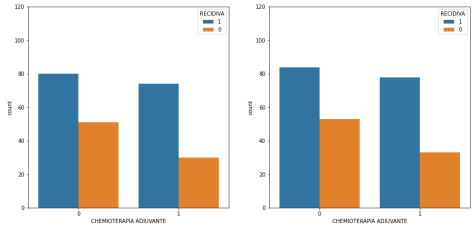
(a) Distribution of Grading 3 vs 1-2 in Adjuvant Chemotherapy



(b) Distribution of Microscopic Vascular Invasion in Adjuvant Chemotherapy



(c) Distribution of Mortality in Adjuvant Chemotherapy



(d) Distribution of Recurrence in Adjuvant Chemotherapy

Figure 1.7: Comparison of distribution of Adjuvant Chemotherapy, thorough barplots grouped by outcome in Original vs Imputed Data

The result is fulfilling, because the distributions of imputed variables do not change significantly after Multiple Imputation. Therefore, the imputed dataset will be used for the following analysis, thus not significantly reducing the number of patients.
Using this technique to deal with missing values, the number of remaining patients according to the different cases of radiomic covariates are:

	Remaining Patients	(%)	Deleted Patients	(%)
PORTAL	244	93%	17	7%
PORTAL + ARTERIAL + LATE	190	73%	71	27%

Table 1.6: Table with final numbers and percentages of patients left and deleted after dealing with Missing Values, in case of using only Portal features or all radiomics covariates jointly with clinical.

In Chapters 2 and 3 only Clinical and Portal(Core+Margin) covariates are used in the analysis, with a total of 244 patients; while in Chapter 4 all information is exploited leading with a number of samples equals to 190.

1.4.2. Outliers Detection

After dealing with the problem of Missing Values, the next step is to analyse outliers. It is important that these anomalies, if present, are found, because they can create problems if attention is not paid [24]. For this reason, once the outliers have been found, decisions about what to do with them must be taken. Most common causes of outliers are data entry errors, measurement errors, experimental errors, data processing errors and so on.
In this context we decided to use a multivariate and non-parametric method to identify outliers, namely DBSCAN [25]. DBSCAN is a density-based clustering algorithm, focused on finding neighbours by density on a multidimensional sphere of a certain radius. DBSCAN is able to identify three different class of points: Core points, Border points and Outliers. Outliers, in this context, are points that lie is no cluster and that are not density reachable nor density connected to any other point. Finding points classified as Outliers by the DBSCAN is the goal. To use the algorithm hyperparameters, must be set. These are the minimum of number of points contained in a sphere to consider the point a Core point and the radius of the sphere. The choice of these parameters can influence the result.
DBSCAN is applied in our case to identify outliers in numerical clinical features: Ca19-9, Dimension of IHC, OS and RFS. To be able to visualize the results in order to have feedback on the goodness of the process, these covariates are considered pairwise: in every pair Outliers are searched in a 2-dimensional space. Figure 1.8 suggests the presence of anomalies in the dataset. Instead of scaling the data, a radius that varies according to the maximum distance of the points is chosen. After several attempts, hyperparameters are decided: 4 as a minPoints and 15% of maximum points distance as radius; Euclidian distance is used.
In this manner, three Outliers are identified in pairs that contain Ca19-9 as a covariate. In Figure 1.9 it can be seen that these points correspond to high values of the biomarker. Moreover, it can be noticed the correctness of the outliers identification, thanks to the accuracy of the algorithm and thanks to the fact that DBSCAN is a very intuitive algorithm with ease of visualizing the result.

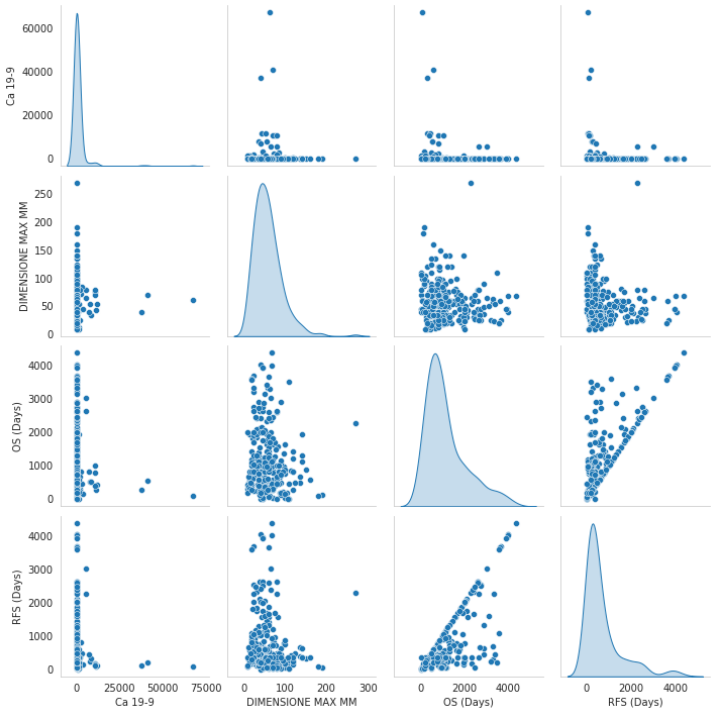


Figure 1.8: Pairwise plot of numerical clinical features of IHC Dataset, i.e. Ca 19-9, Dimension, RFS and OS

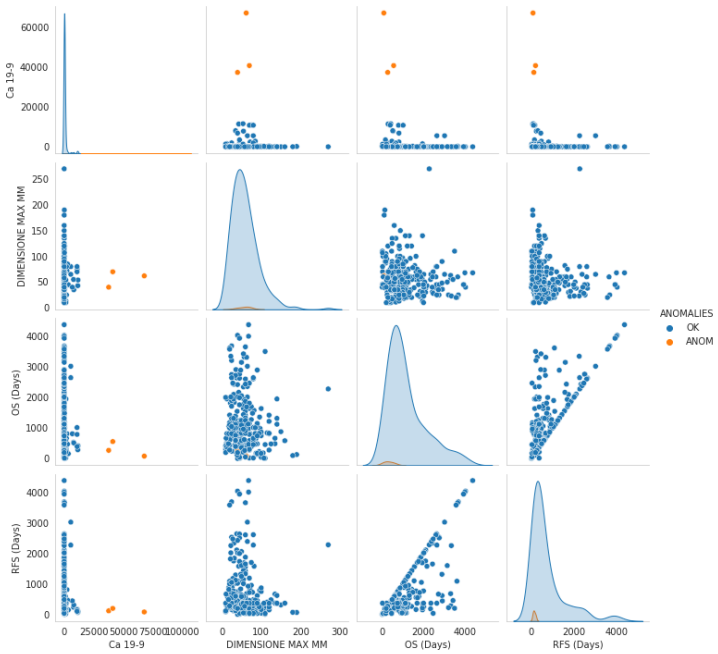


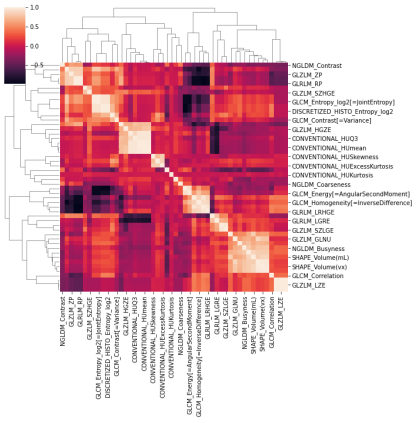
Figure 1.9: Pairwise plot with outliers detected in numerical clinical features of IHC Dataset

Once outliers are identified, decisions about them must be taken. The first thing that comes to mind is that these elevated Ca19-9 values may be due to data entry errors. If this were the reason, patients with such high numbers should be eliminated from the dataset. Because the data were retrospectively collected, there was the opportunity to verify the correctness of the values: the number reported in the dataset are right, no errors have been made in recording the data. For this reason, since the values are plausible and are not mistakes, the outliers are not removed from the dataset. In this way, the numerosity of the patients is preserved.

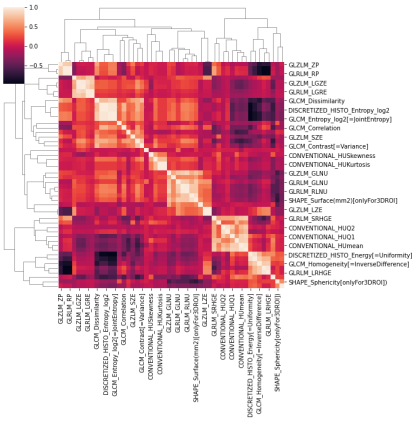
Outliers are not searched in radiomic covariates because the high dimensionality of the data and the small sample size make it impossible, even using multivariate outliers detection techniques such as DBSCAN, to address the problem properly. Outliers detection algorithm performs poorly on dataset with small size and large number of features, which is our case.

1.4.3. Correlation Analysis

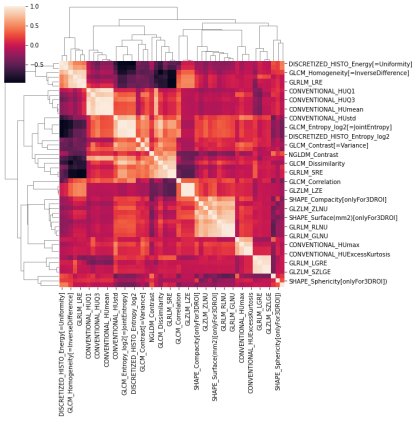
The main issue of our data is the high number of the radiomic features and, for this reason, techniques of features selection and dimensionality reduction must be used. As an initial step, a correlation analysis is used to obtain a first skimming of the covariates. The goal is to eliminate highly correlated features in order to reduce the number of them. For every group of radiomic covariates, namely Portal, Arterial and Late, separately for Core and Margin, a clustermap is used to see if correlation is present among variables. In Figure 1.10 it can be noted that there are groups of highly correlated features in every case and the strategy is to deleted them. A threshold of 0.85 is used for the removal of the variables: covariates that are correlated more than 0.85 are eliminated from the dataset and the further analysis.



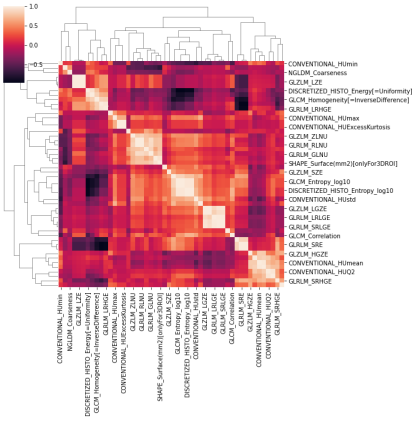
(a) Clustermap of Portal Core covariates



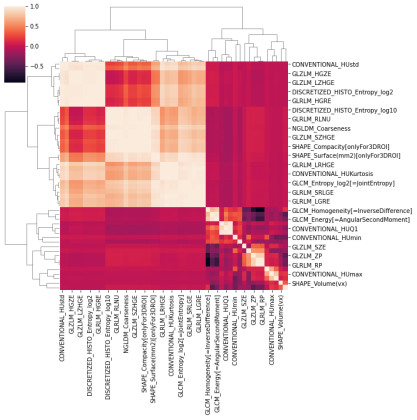
(b) Clustermap of Portal Margin covariates



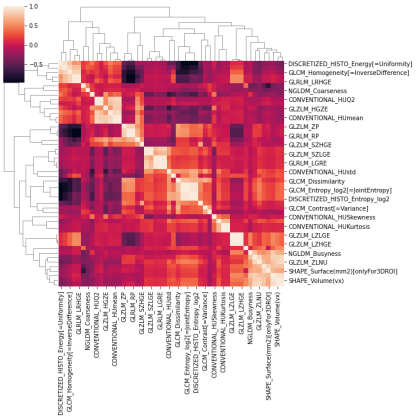
(c) Clustermap of Arterial Core covariates



(d) Clustermap of Arterial Margin covariates



(e) Clustermap of Late Core covariates



(f) Clustermap of Late Margin covariates

Figure 1.10: Correlation Analysis with Clustermaps of each radiomic covariate subgroup in IHC Dataset

Using this approach, many features are deleted and information on changes in their number is summarised in the Table 1.7:

	Remaining Covariates	(%)	Deleted Covariates	(%)
PORTAL CORE	24	48%	26	52%
PORTAL MARGIN	30	60%	20	40%
ARTERIAL CORE	21	42%	29	58%
ARTERIAL MARGIN	25	50%	25	50%
LATE CORE	14	28%	36	72%
LATE MARGIN	24	48%	26	52%

Table 1.7: Final numbers and percentages of remaining and deleted covariates in each radiomics subgroup after removal subsequent to correlation analysis

In this manner the number of features is significantly decreased, reducing the problem of overfitting.

2 | Classification of Pathology Data

In this Chapter we deal with the classification of pathology data. The outcomes taken into account are:

- **Microscopic Vascular Invasion (MVI):** MVI is an established adverse prognostic factor in patients with IHC. It is currently diagnosed on IHC tissue histological examination typically after surgical resection.
- **Grading:** it expresses the differentiation of the tumor, describing how much it deviates from the normal tissue from which it originated. It is an indicator of how quickly a tumor is likely to grow and spread. This variable is categorical over 3 levels, being:
  - Level 1: Well differentiated, i.e. tumours look very similar to surrounding normal cells
  - Level 2: Moderately differentiated. i.e. tumour cells have a clearly abnormal appearance, but still share some characteristics with surrounding normal cells
  - Level 3: Poorly differentiated, i.e. tumours appear very abnormal

  According to clinicians suggestion, we dichotomize the variable in two classes classification coded as 1-2 (0) vs 3 (1).

In order to have an adequate prognosis and treatment, it is crucial to be able to correctly predict values of MVI and Grading. For this reason, it is important to find a robust model for classification.

2.1. Methodologies for Classification

In this Section we describe the methodologies used to classify pathology data. In Section 2.1.1 we explain the importance of using variable selection techniques to select the features

that enter the model, trying to mitigate the problem of overfitting. Then, in Section 2.1.2 we describe Logistic Regression procedure for classification, that is used to find the best model in which to account for the multilevel nature of the data subsequently. In Section 2.1.3 are illustrated Mixed Effects Models, i.e. modelling techniques to consider the grouping factor present in the data.

2.1.1. Feature Selection and Dimensionality Reduction

Since in our problem a huge number of radiomic variables are present, Feature Selection or Dimensionality Reduction are needed. Feature Selection works by keeping only the most relevant variables from the original dataset, deleting the others; while Dimensionality Reduction aims to find a smaller set of new variables, each being a combination of the input features, exploiting the redundancy of input data. [26]. These methods help in understanding data, reducing computation requirement, reducing the effect of curse of dimensionality, namely the difficulty of dealing with high-dimensional data, and improving the predictor performance [27]. These potentialities of variable selection techniques make them suitable for our problem, as we want to obtain a robust and understandable model for classification starting with a large number of input covariates and a small sample size. Together with these, we have also tried to examine Regularization techniques. Regularization involves fitting a model considering all predictors in which the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage has the effect of reducing variance of the model[28]. Therefore, using Regularization allows us to increase the generalisation capacity of the model, making the performance in training more similar to that in test, mitigating overfitting.
Among all these possibilities, the techniques we have used within this work to decrease the number of covariates where radiomic is included are:

- **Forward Selection** [29]: to perform Forward Selection within this thesis, we use the `SFS` function from `mlxtend` Python Package. The best model is selected among models from 3 to 20 covariates using Stratified K-fold Cross-Validation with k=10.
- **Backward Selection** [29]: to perform Backward Selection within this thesis, we use the `SFS` function from `mlxtend` Python Package. The best model is selected among models from 3 to 20 covariates using Stratified K-fold Cross-Validation with k=10.
- **Stepwise Selection** [29]: to perform Stepwise Selection within this thesis, we use the `SFS` function from `mlxtend` Python Package. The best model is selected among models from 3 to 20 covariates using Stratified K-fold Cross-Validation with k=10.

- **Ridge Regression** [28]
- **Lasso Regression** [28, 30]
- **Principal Component Regression** [31, 32]: Principal Components are found separately for tumour (core) and peritumoral zone (margin) by means of the `sklearn` Python Package [33]. The number of components necessary to explain 90% of the variability in the data for both core and margin was retained.

2.1.2. Logistic Regression

To perform two classes classification we decided to use Logistic Regression as a modelling technique, because the ratio between sample size and number of features does not allow for stability of the results employing other machine learning methods such as KNN, CART or Random Forest. Moreover, Logistic Regression enables a higher and more direct explainability of the covariate relevance, that is important to the clinical counterpart. *Logistic Regression* is a particular case of Generalized Linear Model in which the outcome variable Y is binary. Let $Y \sim Be(p(\mathbf{X}))$ be the binary response variable, where $p(\mathbf{X}) = \mathbf{P}(Y = 1|\mathbf{X})$ denotes the probability that Y belongs to the positive class and let $\mathbf{X} = [X_1, X_2, \ldots, X_p]$ be the independent variables. In Logistic Regression model, $p(\mathbf{X})$ is expressed through logistic function, so that:

$$p(\mathbf{X}) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j X_j}} \tag{2.1}$$

To find the values of the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ maximum likelihood estimation is employed. The formula of the likelihood is:

$$L(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)) \tag{2.2}$$

With coefficients found through maximum likelihood maximization, probabilities $\hat{p}(x_i)$ are predicted. Given the values of $\hat{p}(x_i)$ and a threshold $\tilde{p}$, the sample x_i is assigned to the positive class if $\hat{p}(x_i) \geq \tilde{p}$

Logistic Regression is used in Section 2.2.1 to find the best model to classify outcomes considering samples independent and identically distributed. This step is necessary in order to be able to consider the multilevel nature of the data in the model afterwards.

2.1.3. Mixed Effects Models

With Logistic Regression, the multicentre aspect of the study has not been considered and modelled in the analysis: patients were considered independent identically distributed. Due to the fact that individuals are grouped by hospitals the assumption of independence may not be respected, therefore the effect of the centre has to be accounted for appropriately. *Mixed Effects Models* (MEMs) take this aspect into account by providing a flexible and powerful tool for the analysis of grouped data [34]. MEMs incorporate both *fixed effects*, which are parameters associated with an entire population and *random effects* related to the grouping factor, that is the same of all observations of the same group but differs from group to group. Since we are dealing with a classification, a Generalized Linear Mixed Effects Models (GLMM) has to be considered. In general, a GLM is express with the following form:

$$g(\mathbf{E}(Y)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{2.3}$$

where $X_1, \ldots, X_p$ are the fixed covariates, Y is the outcome, g is the link function, β_j with $j \in 1, \ldots, p$ are the unknown parameters that we want to estimate. For classification Y is a Bernoulli random variable, and its mean, which we call p, is the probability that the outcome is one. The link function is the logarithm of the odds, so that the formula of this particular case of GLMM for $i-th$ sample belonging to the $j-th$ group is:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 X_{1ij} + \cdots + \beta_p X_p - b_j \tag{2.4}$$

b_j is the random effect and it is assumed that is normally distributed, namely $b_j \sim \mathcal{N}(0, \sigma_b^2)$. These terms are unobserved and treated as varying randomly among clusters and their estimates provide a measure of the cluster effect.
To recap, MEMs are able to capture centre-to-centre variations, since patients in the same group share the same effect. Therefore, with this modelling technique, the multicentre aspect of this study will be properly modelled in Section 2.2.2.

2.2. Results of Classification

Separately for each outcome, we apply Logistic Regression to clinical and radiomics covariates belonging to core and margin of the Portal phase, applying various methods of variable selection as described in Section 2.1.1. To analyse the importance of the radiomic features in classification, three scenarios of grouped covariates are considered, following a

clinical rationale:

- Clinical
- Clinical + Portal(Core)
- Clinical + Portal(Core+Margin)

Numerical features are standardized before the analysis.
Logistic Regression is used in Section 2.2.1 to identify the best model for each of the above cases, not taking into account multicentre aspect of the data. Then, with covariates identified in each of the best models centre effect is analysed fitting Mixed Effects Models in Sections 2.2.2.

The metrics examined for choosing the best model and make considerations about the importance of radiomics are the following, which notation is defined in Figure 2.1:

Figure 2.1: Schematisation of the confusion matrix, defining the terms with which performance metrics are expressed

		PREDICTED CLASS	
TRUE CLASS		1	0
	1	TP	FN
	0	FP	TN

- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ It is the number of correctly predicted data points out of all the data points.
- $Specificity = \frac{TN}{TN+FP}$ It is the True Negative Rate that estimates the probability to correctly identify the elements of the negative class.
- $Sensitivity = \frac{TP}{TP+FN}$ It is the True Positive Rate that estimates the probability to correctly identify the elements of the positive class.
- $Precision = \frac{TP}{TP+FP}$ It is the percentage of items classified as positive that are actually positive.
- *Precision-Recall AUC*: It is the Area under Precision-Recall curve that plots precision against recall.
- *ROC AUC*: It is the Area under the ROC curve that plots Sensitivity against Specificity.

To identify the covariates retained by the various variable selection techniques, the models have been trained on the entire dataset and the performances tested on the whole dataset are reported. However, the estimates of the performances on the entire dataset are overestimated, because samples that have been used for training are also tested. In order to have a more realistic estimation of the performances, two different techniques of cross-validation are used to test the Logistic Regression model with the best identified features:

- **Method 1 of Cross-Validation:** usually Stratified K-fold Cross-Validation with k=50. Results are reported in terms of mean and standard deviation.
- **Method 2 of Cross-Validation:** the data are split into a training set (80%) and a test set (20%) stratifying the outcome. The validation procedure was repeated 100 times over 100 different samples. The performances for each metric produced on each individual sample was collected in a dataset of 100 sample for later use in Section 2.2. Results are reported in terms of mean and standard deviation.

The best models are selected looking at performances in cross-validation.

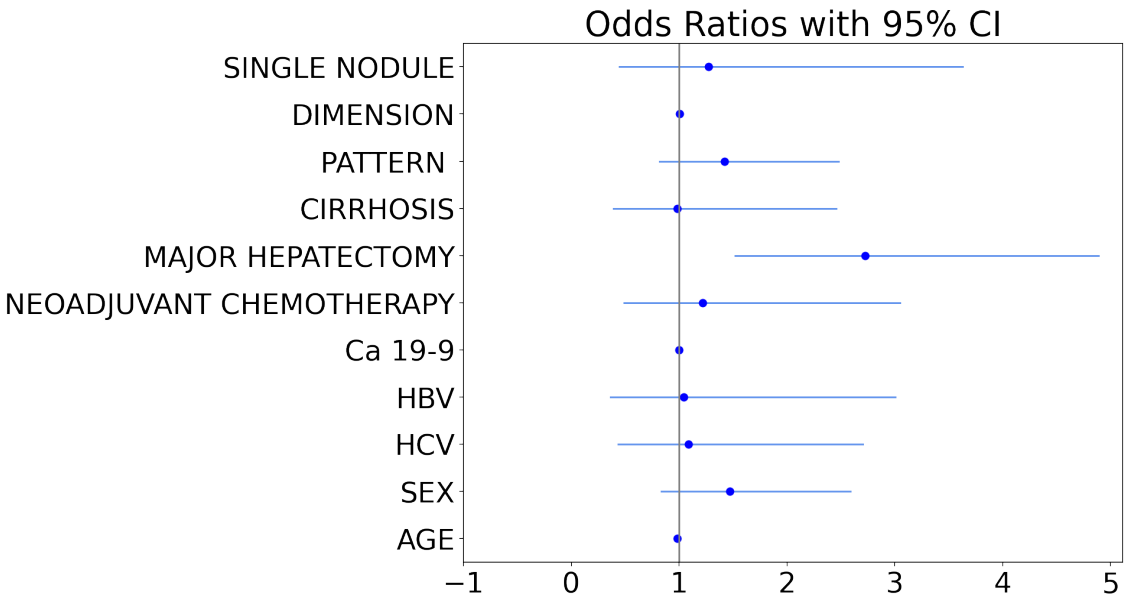
2.2.1. Logistic Regression for identifying the best model

In this Section results of Logistic Regression models are reported. For each different grouping of covariates considered, Logistic Regression is used to classify first MVI and then Grading. For every model, jointly with performances, forest plots of the odds ratios are provided. To perform Logistic Regression `sklearn` [33] Python Package is used. For sake of simplicity, only the best models are reported in this Chapter, while performances of other attempts are summarized in the Appendix C.

Logistic Model for MVI with Clinical Features only

In the case of clinical features only, no variable selection technique was applied as the number of covariates is sufficiently small. Results are displayed in the forest plot reported in Figure 2.2.

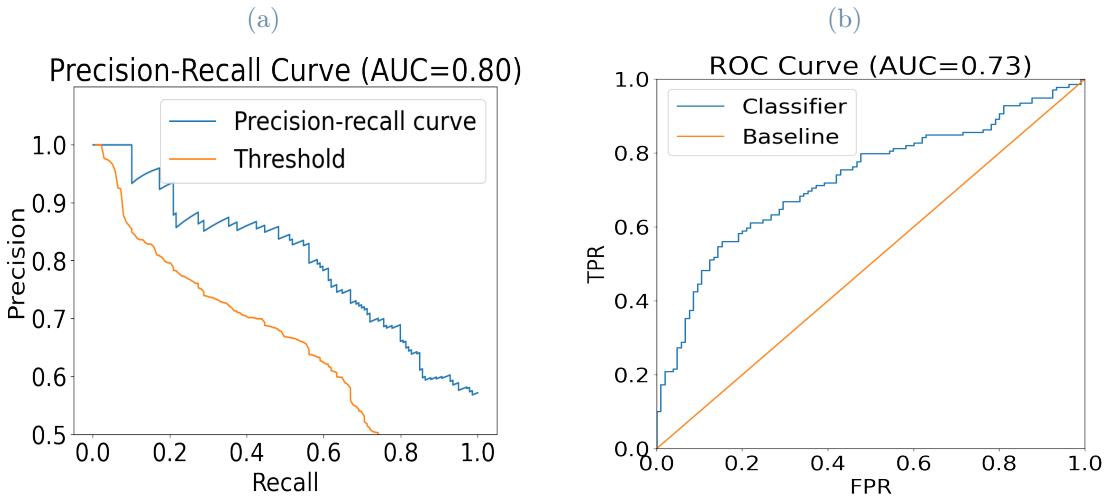
Figure 2.2: Odds ratios with 95% CI obtained applying Logistic Regression for MVI with Clinical features only



MAJOR HEPATECTOMY is strongly significant in the model and from its odds ratio can be deduced that the probability of having MVI increases if a patient has undergone Major Hepatectomy.

Performances are summarized in Figure 2.3 and Table 2.1.

Figure 2.3: Precision-Recall and ROC Curves for MVI LR for Clinical covariates only



Performance are not particularly impressive, but not bad either.

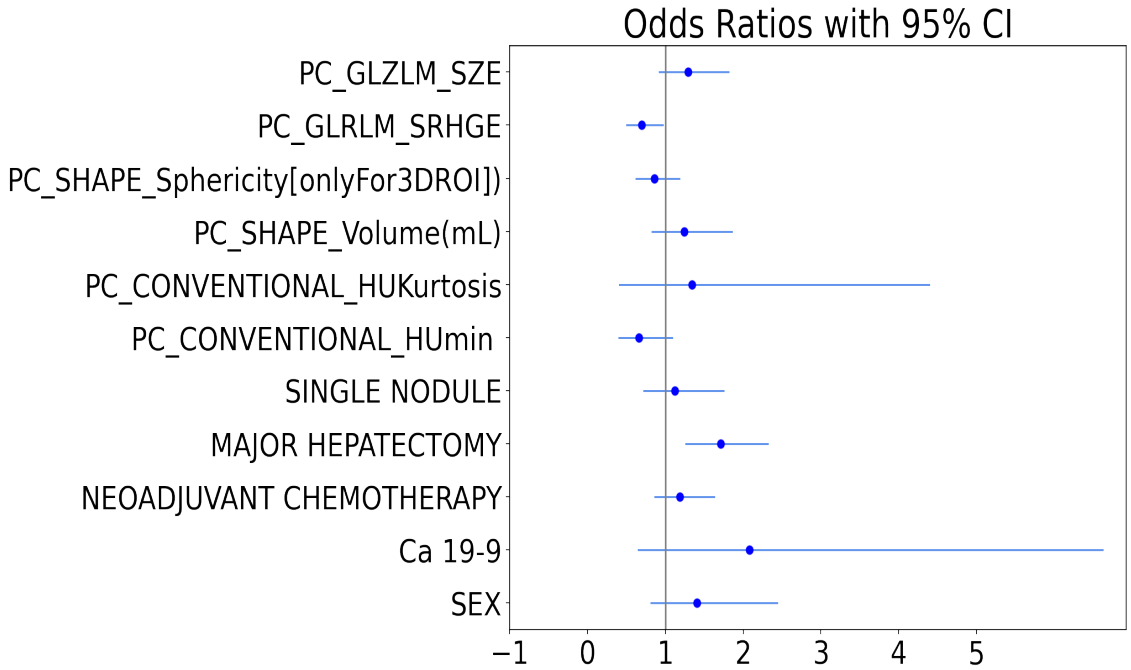
Table 2.1: Performances of MVI LR with Clinical features only

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.67	0.659	0.224	0.652	0.062
SPECIFICITY	0.58	0.557	0.366	0.551	0.104
SENSITIVITY	0.74	0.740	0.285	0.727	0.086
PRECISION	0.70	0.709	0.235	0.686	0.056
PR AUC	0.80	0.848	0.139	0.762	0.058
ROC AUC	0.73	0.715	0.256	0.686	0.068

Logistic Model for MVI with Clinical + Portal(Core) Features

The best model is the one obtained applying Backward Selection. Results are displayed in the forest plot reported in Figure 2.4.

Figure 2.4: Odds ratios with 95% CI obtained applying Logistic Regression for MVI with Clinical+Portal(Core) features



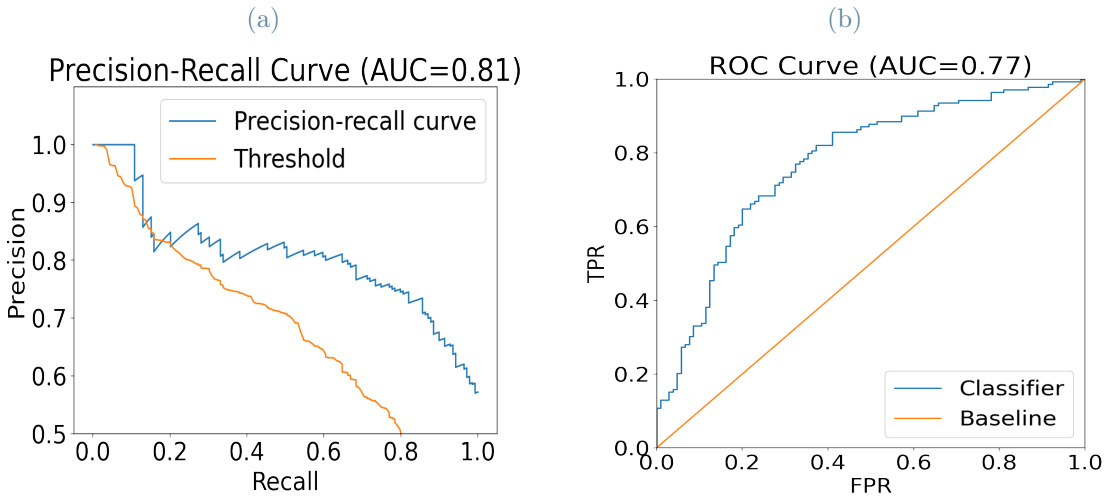
The model selects basic, first order and second order portal core radiomic covariates (for the differentiation we refer the reader to Appendix A) and it is inferred that patients that have undergone MAJOR HEPATECTOMY have an increased risk of developing MVI.

Performances are reported in Table 2.2 and Figure 2.5.

Table 2.2: Performances of MVI LR with Clinical+Portal(Core) features with Backward Selection

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.73	0.724	0.200	0.706	0.062
SPECIFICITY	0.64	0.630	0.355	0.609	0.110
SENSITIVITY	0.80	0.800	0.238	0.779	0.076
PRECISION	0.74	0.773	0.203	0.730	0.060
PR AUC	0.81	0.862	0.141	0.788	0.055
ROC AUC	0.77	0.760	0.235	0.748	0.062

Figure 2.5: Precision-Recall and ROC Curves for MVI LR for Clinical+Portal(Core) covariates with Backward Selection

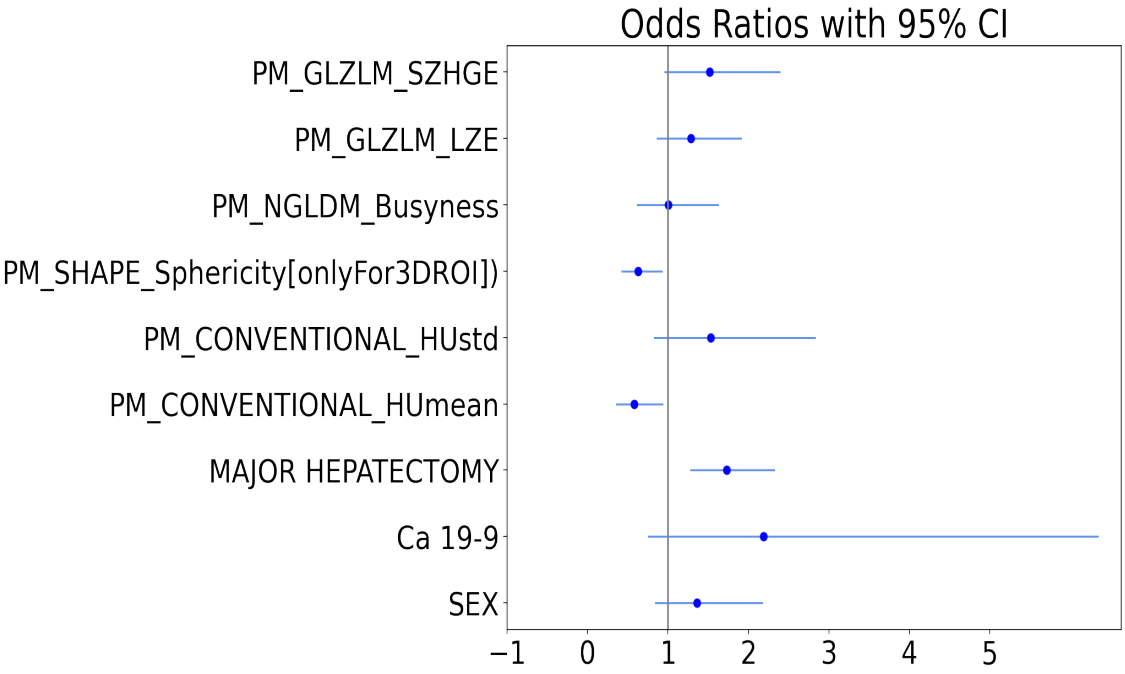


Values of all performance metrics increase both in training and cross-validation with respect to the previous case, in which only clinical features are included in the model. The fact that all values of performance metrics increase by including part of the radiomic covariates suggests the added value that radiomics can bring to the predictive ability of the model.

Logistic Model for MVI with Clinical + Portal(Core+Margin) Features

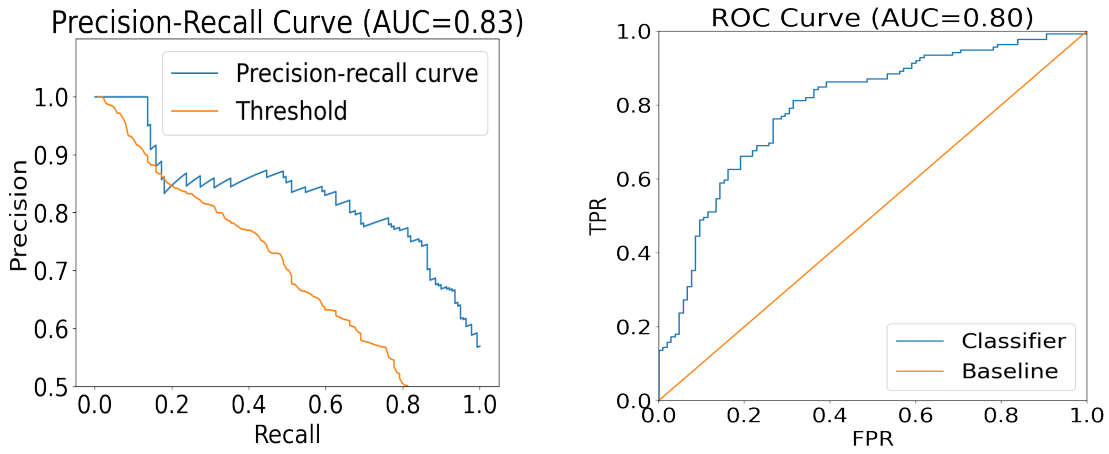
The best model is the one obtained applying Backward Selection. Results are displayed in the forest plot reported in Figure 2.6.

Figure 2.6: Odds ratios with 95% CI obtained applying Logistic Regression for MVI with Clinical+Portal(Core+Margin) features



The only clinical covariate that is significant is MAJOR HEPATECTOMY and all radiomic features that are selected belong to the margin.

Performances are reported in Table 2.3 and Figure 2.7.



(a) (b)

Figure 2.7: Precision-Recall and ROC Curves for MVI LR for Clinical+Portal(Core+Margin) covariates with Backward Selection

Table 2.3: Performances of MVI LR with Clinical+Portal(Core+Margin) features with Backward Selection

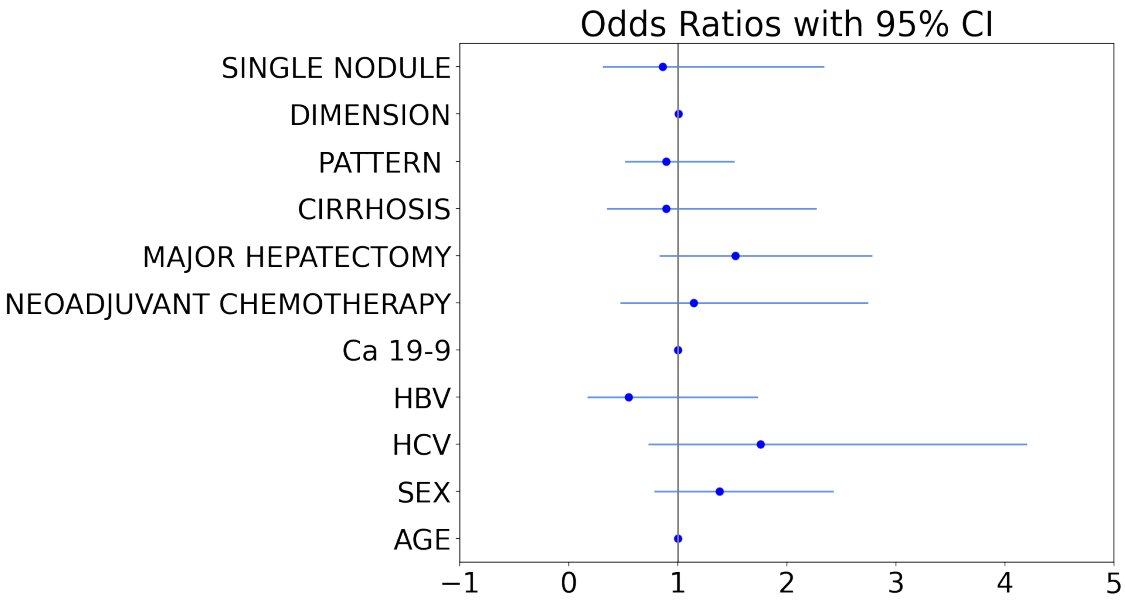
Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.76	0.727	0.187	0.730	0.060
SPECIFICITY	0.69	0.647	0.347	0.646	0.102
SENSITIVITY	0.81	0.790	0.221	0.793	0.078
PRECISION	0.77	0.783	0.199	0.752	0.057
PR AUC	0.83	0.869	0.132	0.820	0.050
ROC AUC	0.80	0.777	0.206	0.777	0.062

With respect to the previous case, in which only radiomics of the core is considered, values of all performance metrics increase. Moreover, the model largely selects variables belonging to the margin. These facts make it clear that it is crucial to include in the model not only the radiomics of the tumour, but also those of the margin area, in order to have better predictions of MVI value.

Logistic Model for Grading with only Clinical Features

In the case of clinical features only, no variable selection technique was applied as the number of covariates is sufficiently small. Results are displayed in the forest plot reported in Figure 2.8.

Figure 2.8: Odds ratios with 95% CI obtained applying Logistic Regression for Grading with Clinical features only



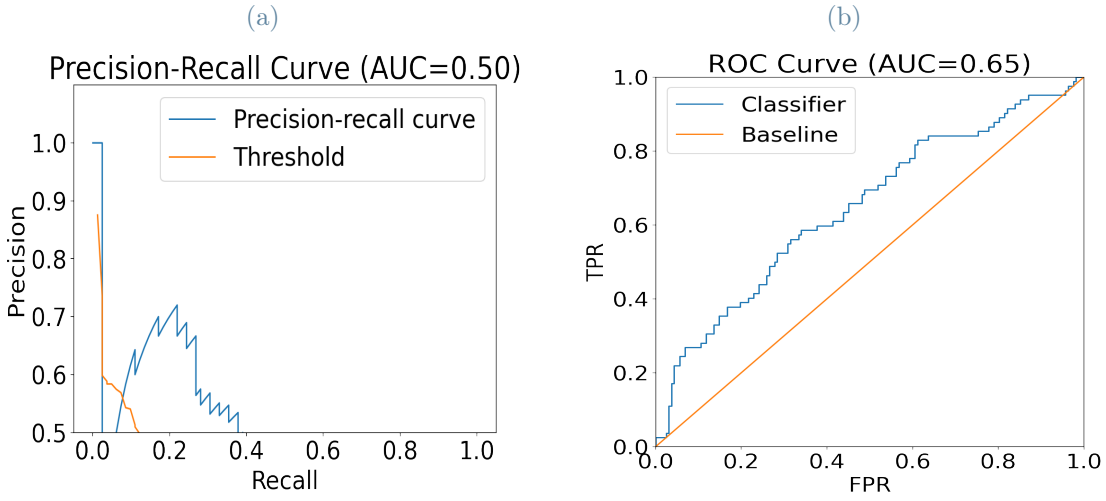
None of the features has a coefficient significantly different from zero.

Performances are summarized in Table 2.4 and Figure 2.9.

Table 2.4: Performances of Grading LR with Clinical features only

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.68	0.650	0.118	0.658	0.037
SPECIFICITY	0.96	0.948	0.112	0.938	0.051
SENSITIVITY	0.11	0.070	0.224	0.081	0.073
PRECISION	0.60	0.090	0.277	0.361	0.318
PR AUC	0.50	0.598	0.250	0.413	0.079
ROC AUC	0.65	0.570	0.295	0.539	0.078

Figure 2.9: Precision-Recall and ROC Curves for Grading LR for Clinical covariates only

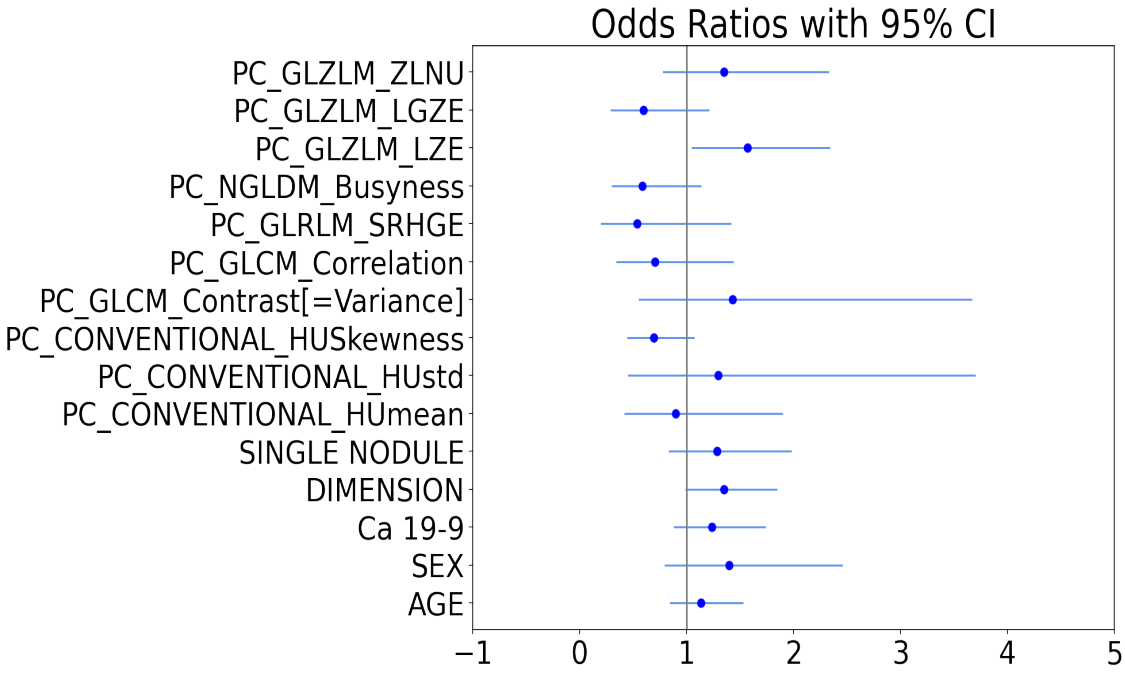


Performance are not satisfactory in particular providing a very high specificity and a very low sensitivity, so that the model has difficulty in recognising positive samples.

Logistic Model for Grading with Clinical + Portal(Core) Features

The best model is the one obtained applying Backward Selection. Results are displayed in the forest plot reported in Figure 2.10.

Figure 2.10: Odds ratios with 95% CI obtained applying Logistic Regression for Grading with Clinical+Portal(Core) features

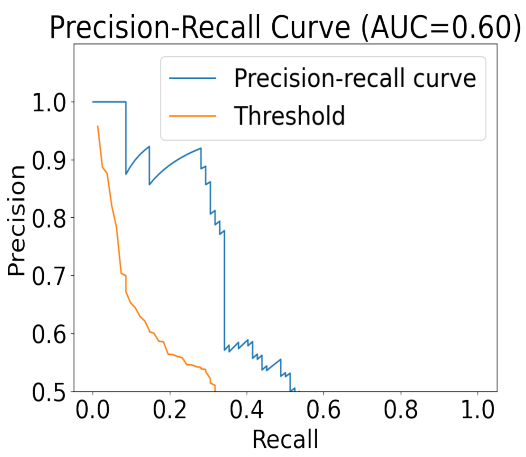


Only PC_GLRLM_SRHGE covariate is significant and it is a radiomic feature of second order.

Performances are reported in Table 2.6 and Figure 2.13

Figure 2.11: Precision-Recall and ROC Curves for Grading LR for Clinical+Portal(Core) covariates with BS

(a)



(b)

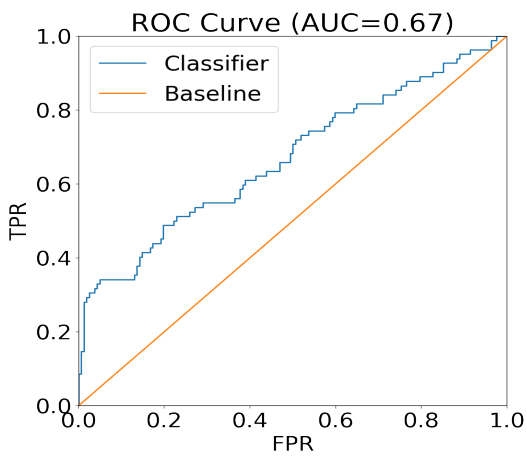


Table 2.5: Performances of Grading LR with Clinical+Portal(Core) features with BS

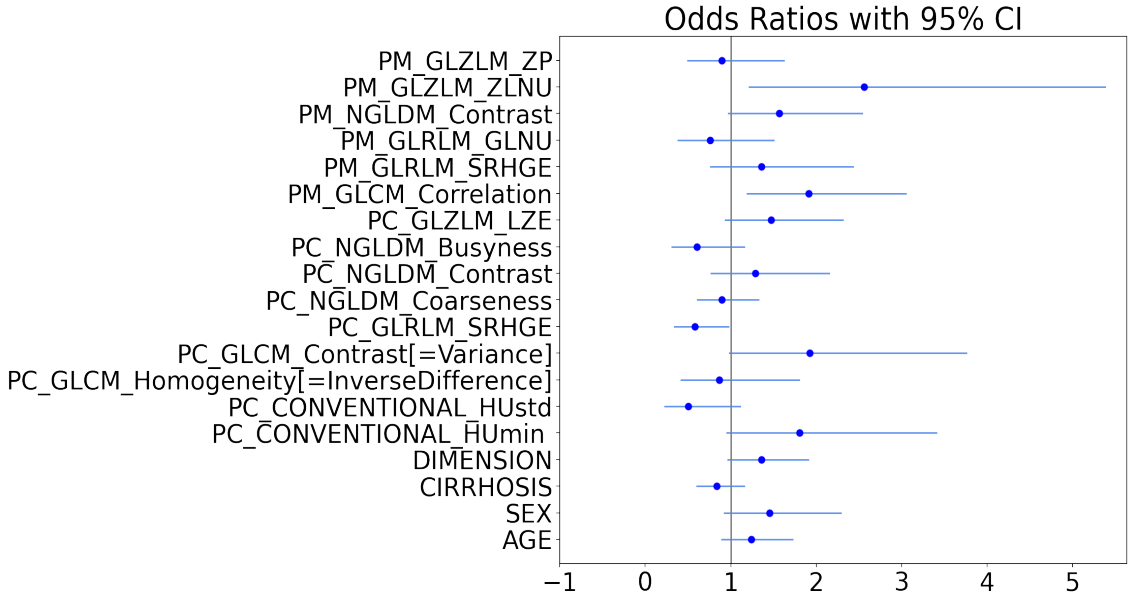
Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.74	0.714	0.200	0.693	0.055
SPECIFICITY	0.96	0.925	0.186	0.905	0.061
SENSITIVITY	0.32	0.320	0.397	0.254	0.108
PRECISION	0.79	0.403	0.476	0.585	0.213
PR AUC	0.60	0.694	0.279	0.489	0.099
ROC AUC	0.67	0.645	0.339	0.583	0.089

Values of all performance metrics increases in training with respect to the previous case, in which only clinical features are included in the model. In cross-validation the only index that does not increase is specificity. However, sensitivity increases by 0.25 in method 1 and 0.17 in method 2. The fact that the predictive ability of the model increases including part of the radiomic covariates in the model suggests the added value of radiomics in improving prediction of Grading.

Logistic Model for Grading with Clinical + Portal(Core+Margin) Features

In this case the best model is the one obtained applying Backward Selection. Results are displayed in the forest plot reported in Figure 2.12.

Figure 2.12: Odds ratios with 95% CI obtained applying Logistic Regression for Grading with Clinical+Portal(Core+Margin) features



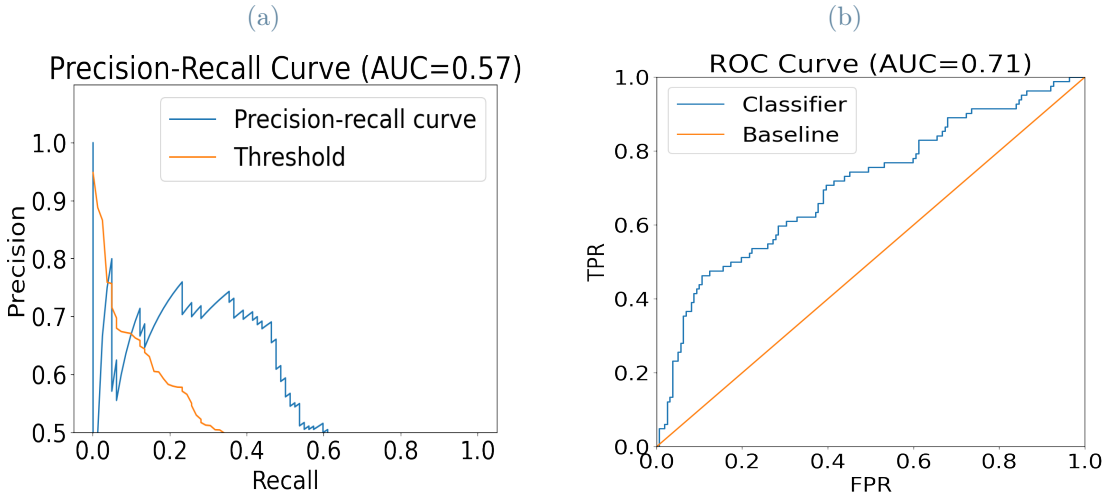
None of the clinical covariates is significant at 5% level. With regard to radiomics, features belonging the margin are significant.

Performances are reported in the Table 2.6 and Figure 2.13

Table 2.6: Performances of Grading LR with Clinical+Portal(Core+Margin) features with Backward Selection

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.73	0.716	0.173	0.716	0.037
SPECIFICITY	0.94	0.925	0.150	0.917	0.044
SENSITIVITY	0.33	0.310	0.386	0.301	0.096
PRECISION	0.73	0.387	0.461	0.650	0.140
PR AUC	0.57	0.685	0.261	0.527	0.088
ROC AUC	0.71	0.663	0.294	0.654	0.074

Figure 2.13: Precision-Recall and ROC Curves for Grading LR for Clinical+Portal(Core+Margin) covariates with Backward Selection

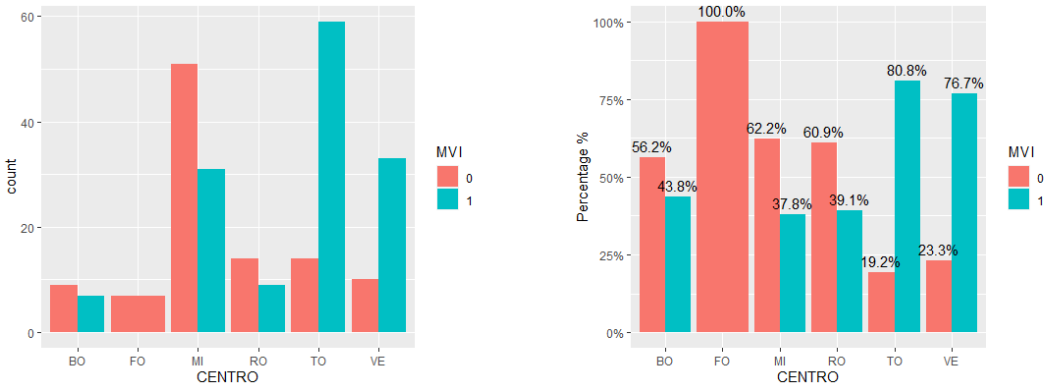


In this case, in which the margin radiomic covariates are considered, values of all performance metrics increases only in cross-validation2. This, together with the fact that margin features included are significant, suggests that, also in the case of grading including radiomic for the peritumoral area is important in order to have more accurate predictions of the model in cross validation.

2.2.2. Mixed Effects Models for accounting multicentre nature of the data

In this Section results of Mixed Effects Models are reported. For all best models identified with Logistic Regression, Mixed Effects Models for first MVI and then Grading are fitted with the covariates selected. With MEMs where the grouped data structure data can be addressed. To understand how strong the centre effect is present in the data the value of the Variance Partition Coefficient [35] is examined. It indicates the amount of variability explained by the grouping factor (the centre here). To fix MEMs the function `glmer` of `lme4` [36] R package is used.

Before examining the results for MVI, to have a qualitative idea on the possible presence of a centre effect, proportions of the event, grouped by the different hospitals, are analysed. From Figure 2.14 it can be seen that distribution of MVI is different among different centres, suggesting that the random effect describing membership of different hospitals is present.



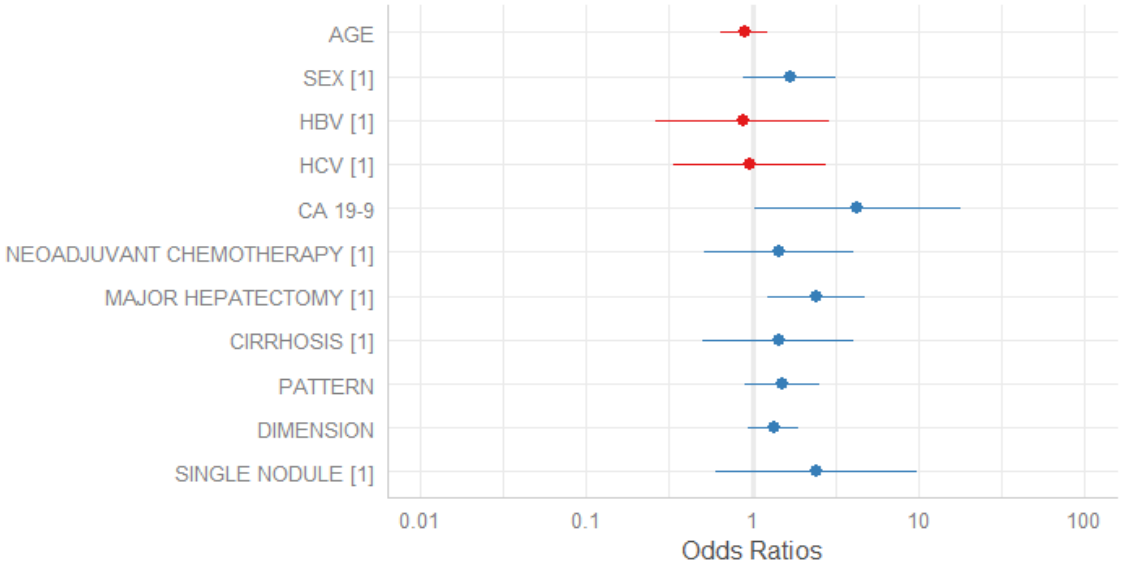
(c) (d)

Figure 2.14: Frequency and Percentages of positive and negative cases of MVI in IHC dataset grouped by variable centre

Mixed Effects Model for MVI with Clinical Features only

These are the results of the MVI MEM with features identified in the best model of Logistic Regression with clinical variables only. The results of the fixed effect are reported in the forest plot in Figure 2.15.

Figure 2.15: Odds ratios with 95% CI obtained applying MEMs for MVI with Clinical features only



The only variables that are significant are MAJOR HEPATECTOMY and CA 19-9. The values of the odds ratio indicate that people that have undergone major hepatectomy and with larger values of Ca19-9 have higher probability to present MVI.

The random effect estimates are illustrated in Figure 2.16.

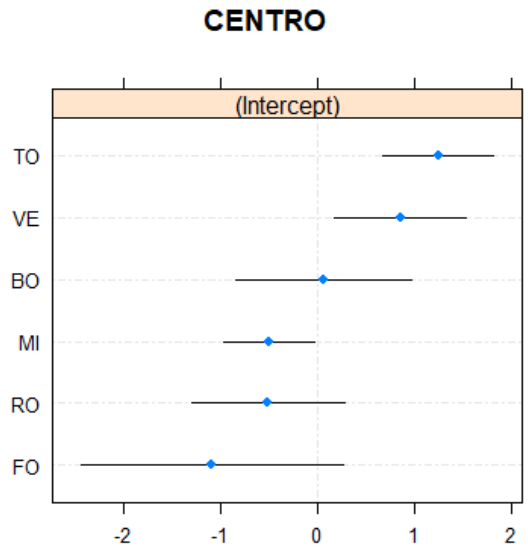


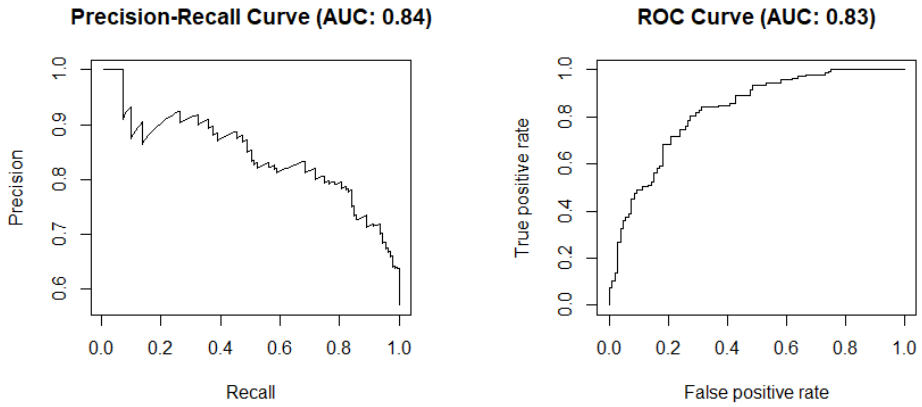
Figure 2.16: Random Effect in MVI MEM with Clinical Feature only

The VPC is 21.24 %. The value is very high, meaning that the centre effect is strongly present.

Performances are summarized in Table 2.7 and Figure 2.17.

Table 2.7: Performances of MVI MEM with Clinical features only

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.77	0.661	0.249	0.701	0.061
SPECIFICITY	0.747	0.645	0.326	0.662	0.082
SENSITIVITY	0.786	0.713	0.266	0.737	0.058
PRECISION	0.82	0.71	0.285	0.748	0.092
PR AUC	0.825	0.818	0.203	0.793	0.063
ROC AUC	0.836	0.762	0.233	0.762	0.063



(a) (b)

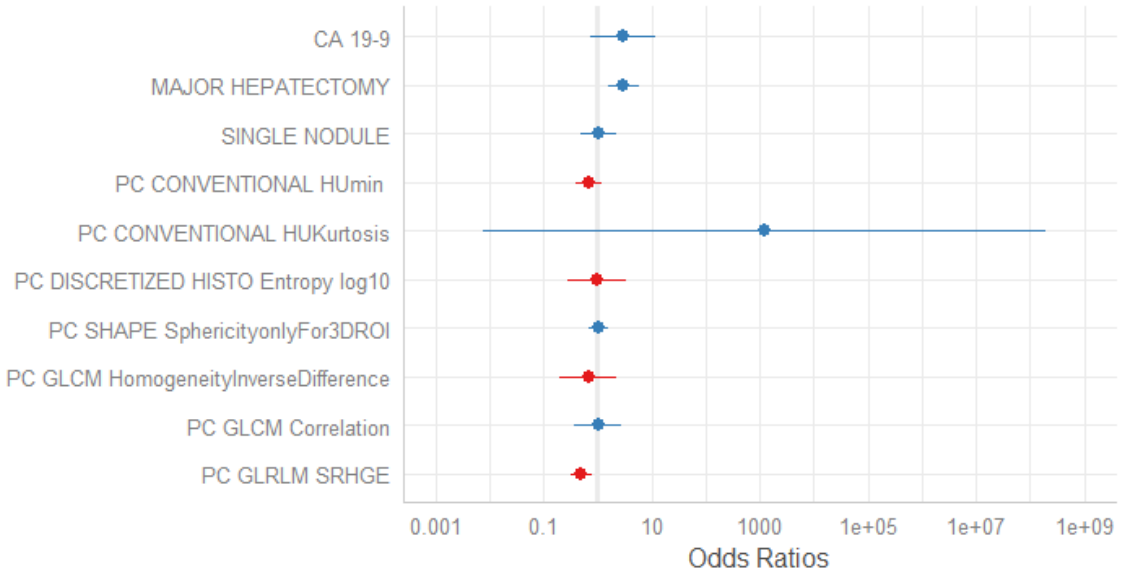
Figure 2.17: Precision-Recall and ROC Curves for MVI MEM for Clinical covariates only

Values of all performance metrics in training and cross-validation 2 improve with respect to the case of Logistic Regression in which the grouping factor is not considered.

Mixed Effects Model for MVI with Clinical+Portal(Core) Features

These are the results of the MVI MEM with features identified in the best model of Logistic Regression with Clinical+Portal(Core) variables. The results of the fixed effect are reported in Figure 2.18.

Figure 2.18: Odds ratios with 95% CI obtained applying MEMs for MVI with Clinical+Portal(Core) features



The only variables that are significant in the model are MAJOR HEPACTECTOMY, which is a clinical covariate and PC_GLRLM_SRHGE, which is a radiomic covariate. Odds ratio of MAJOR HEPATECTOMY indicates a higher risk of present MVI for patients who have undergone major hepatectomy.

The random effect estimas are illustrated in Figure 2.19

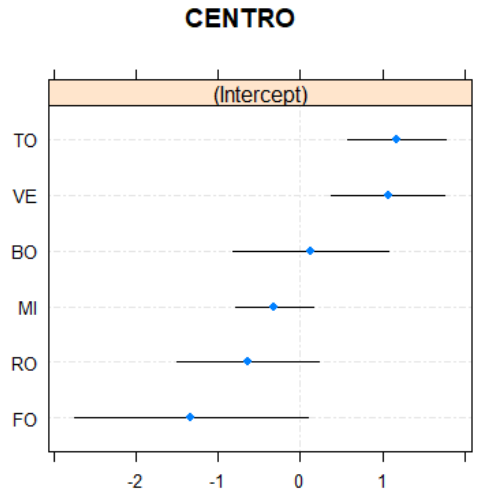


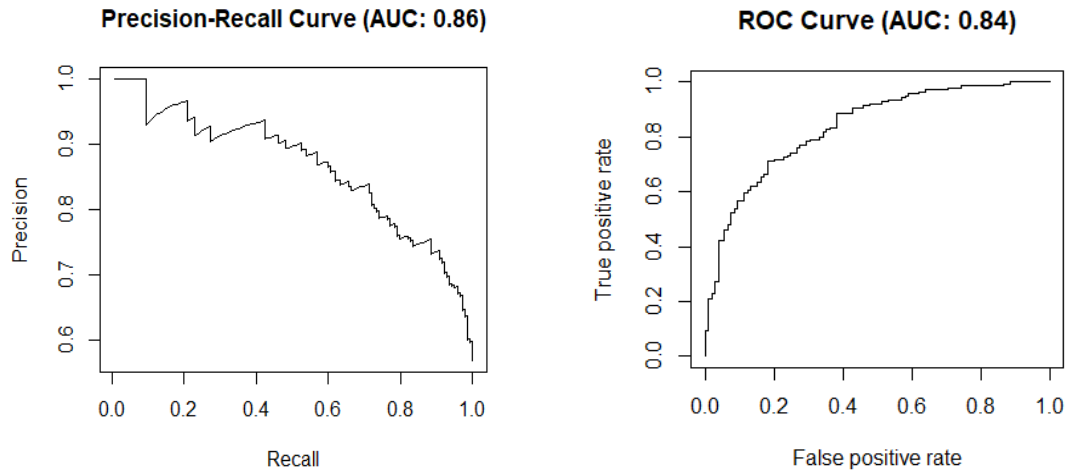
Figure 2.19: Random Effect in MVI MEM with Clinical+Portal(Core) Feature

The VPC is 24.4%. The value is very high, indicating that the effect of the centre is strongly present. In Figure 2.19, it can be seen the effects of the different hospitals: for Torino, Verona there is the statistical evidence of an increasing risk of MVI, given the patients conditions.

Performances are summarized in Table 2.8 and Figure 2.20.

Table 2.8: Performances of MVI MEM with Clinical+Portal(Core) features

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.75	0.651	0.287	0.718	0.056
SPECIFICITY	0.712	0.668	0.293	0.682	0.081
SENSITIVITY	0.779	0.713	0.307	0.75	0.055
PRECISION	0.784	0.677	0.315	0.765	0.077
PR AUC	0.858	0.84	0.203	0.815	0.057
ROC AUC	0.836	0.788	0.225	0.782	0.057



(a) (b)

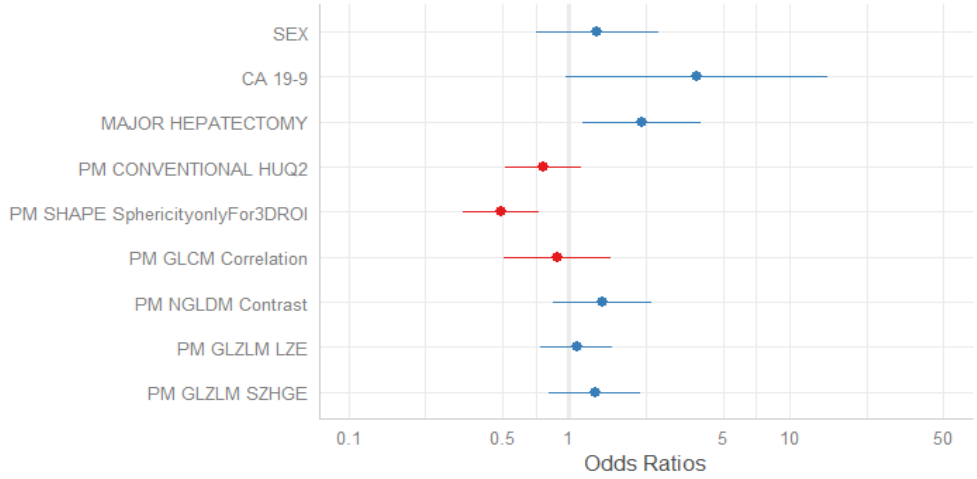
Figure 2.20: Precision-Recall and ROC Curves for MVI MEM for Clinical+Portal(Core) covariates

Adding radiomic covariates of the tumor increases values of all performance metrics in training and cross-validation 2, while in cross-validation 1 only values of specificity, precision-recall AUC and ROC AUC increase. It indicates that considering radiomics information regarding the tumor, increases the predictive ability of the model.

Mixed Effects Model for MVI with Clinical+Portal(Core+Margin) Features

These are the results of the MVI MEM with the features identified in the best model of Logistic Regression with Clinical+Portal(Core+Margin) variables. The results of the fixed effect are reported in Figure 2.21.

Figure 2.21: Odds ratios with 95% CI obtained applying MEMs for MVI with Clinical+Portal(Core+Margin) features



Clinical covariates that are significant in the model are CA 19-9 and MAJOR HEPATECTOMY. Odds ratios values indicate that people that have undergone major hepatectomy and with larger values of Ca19-9 have higher probability present MVI. Among radiomic covariates only one is significant.

The random effect estimates are illustrated in Figure 2.22

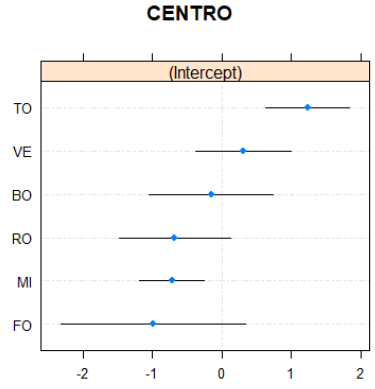


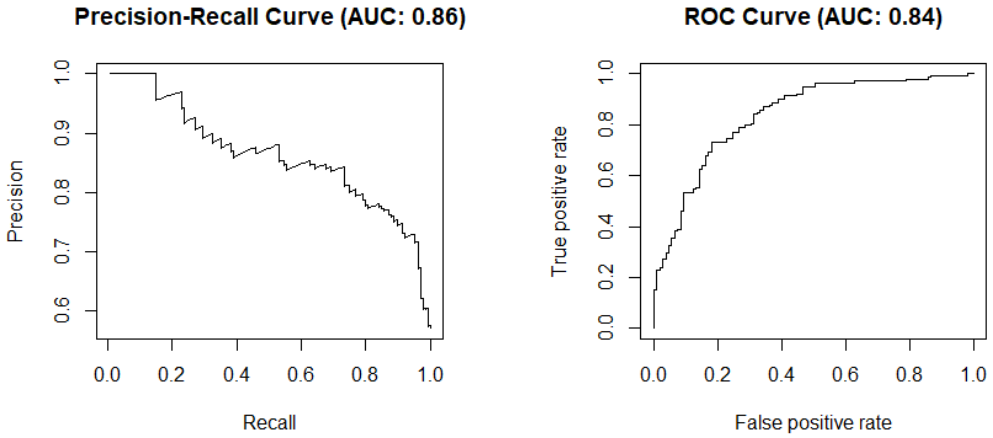
Figure 2.22: Random Effect in MVI MEM with Clinical+Portal(Core+Margin) Feature

The VPC is 19.91%. This testifies that the centre effect is strongly present in the data and it can be seen that there is evidence to say that Torino and Milano hospitals have an effect that is different from zero. Among the various models considering the different sets of covariates, it can be seen that the estimated random effects are consistent with each other.

Performances are summarized in Table 2.9 and Figure 2.23.

Table 2.9: Performances of MVI MEM with Clinical+Portal(Core+Margin) features

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.762	0.674	0.281	0.733	0.058
SPECIFICITY	0.724	0.674	0.261	0.697	0.081
SENSITIVITY	0.791	0.737	0.31	0.768	0.055
PRECISION	0.791	0.673	0.309	0.77	0.086
PR AUC	0.856	0.843	0.193	0.825	0.053
ROC AUC	0.837	0.795	0.219	0.795	0.059

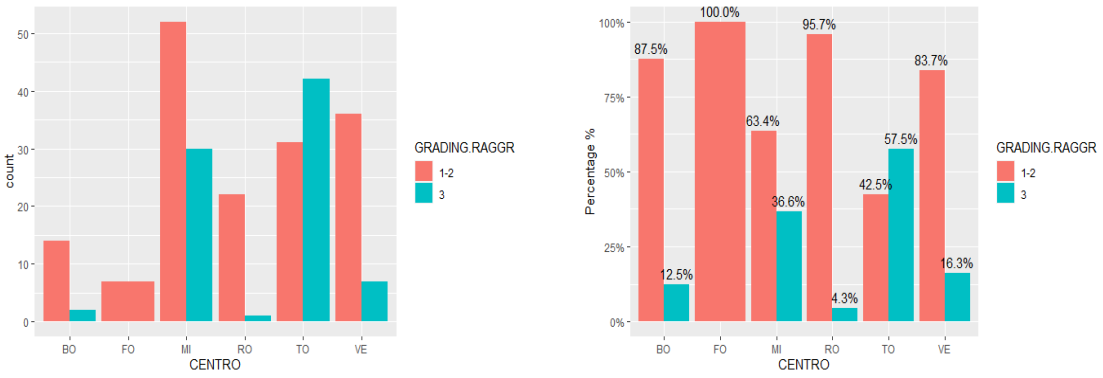


(a) (b)

Figure 2.23: Precision-Recall and ROC Curves for MVI MEM for Clinical+Portal(Core+Margin) covariates

Including margin covariates in the model, all values of performances in cross-validation improve. The conclusion is that radiomics of the peritumoral area is as important as radiomics of the tumour for predicting MVI.

Before examining the results of Grading, to have qualitative feedback on the possible presence of a centre effect, distribution of the event among different hospitals is analysed. From the Figure 2.24 it can be seen that distribution of Grading is different among different centres, suggesting that the random effect describing membership of different hospitals is present.



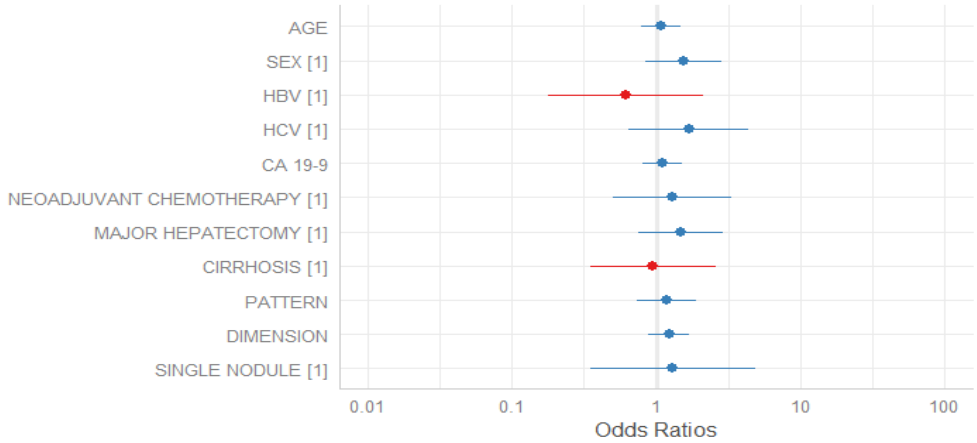
(a) (b)

Figure 2.24: Frequency and Percentages of positive and negative cases of Grading in IHC dataset grouped by variable centre

Mixed Effects Model for Grading with Clinical Features only

These are the results of the Grading MEM with the features identified in the best model of Logistic Regression with clinical variables only. The results of the fixed effect are reported in Figure 2.25.

Figure 2.25: Odds ratios with 95% CI obtained applying MEMs for Grading with Clinical features only



None of the covariates is significant in the model.

The random effect estimates are illustrated in Figure 2.26

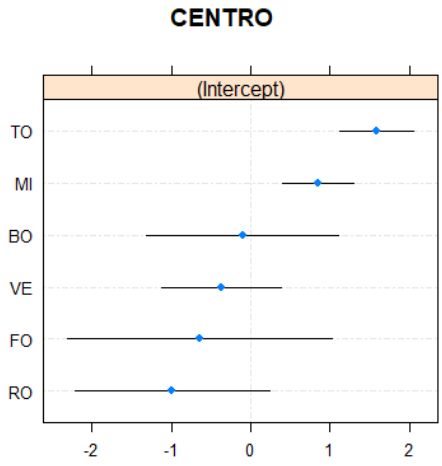
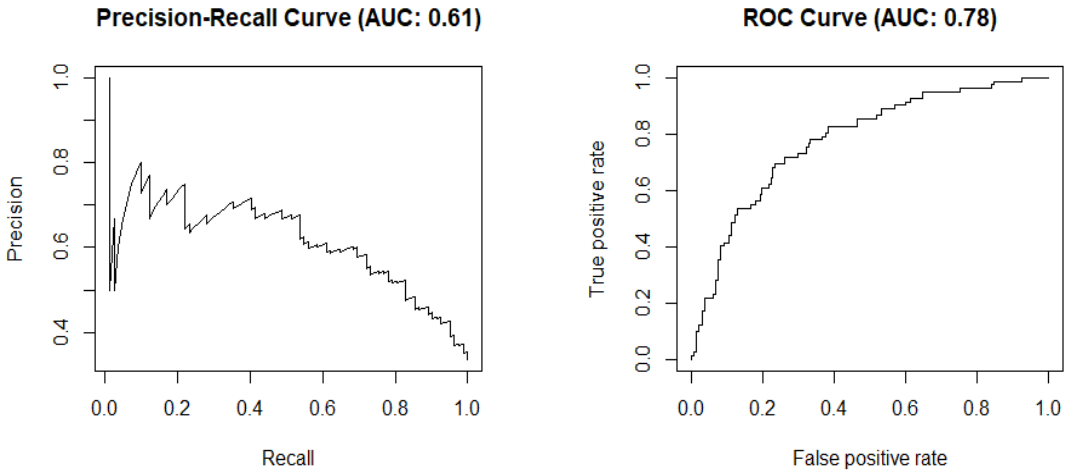


Figure 2.26: Random Effect in Grading MEM with Clinical Feature only

The VPC is 25.97 %. A huge amount of variability of the data is explained by the grouping factor. Therefore, the effect of the centre is present.

Performances are summarized in Table 2.10 and Figure. 2.27



(a) (b)

Figure 2.27: Precision-Recall and ROC Curves for Grading MEM for Clinical covariates

Table 2.10: Performances of Grading MEM with Clinical features only

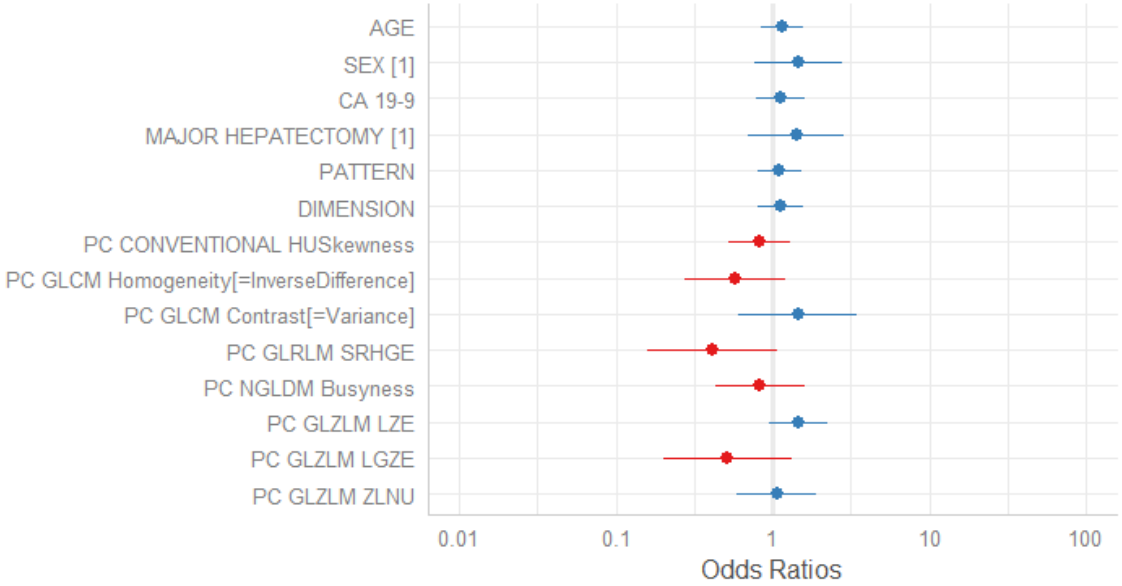
Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.754	0.525	0.362	0.698	0.055
SPECIFICITY	0.78	0.766	0.175	0.745	0.041
SENSITIVITY	0.677	0.423	0.437	0.42	0.433
PRECISION	0.512	0.42	0.568	0.411	0.121
PR AUC	0.614	0.659	0.333	0.514	0.097
ROC AUC	0.78	0.732	0.282	0.709	0.068

With respect to the Logistic Regression case, in which the grouping factor was not considered, the value of the specificity improves by almost 0.5 in training and by almost 0.4 in cross-validation. Therefore, including the grouping factor in the model increases its ability to predict positive samples.

Mixed Effects Model for Grading with Clinical+Portal(Core) Features

These are the results of the Grading MEM with the features identified in the best model of Logistic Regression with Clinical+Portal(Core) variables. The results of the fixed effect are reported in Figure 2.28.

Figure 2.28: Odds ratios with 95% CI obtained applying MEMs for Grading with Clinical+Portal(Core) features



None of the clinical covariates are significant, whereas two radiomics are significant at level 10%.

The random effect estimates are illustrated in Figure 2.29

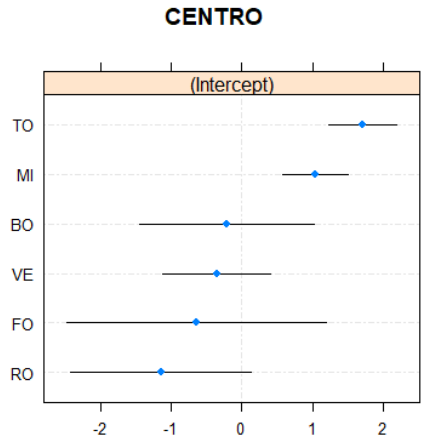
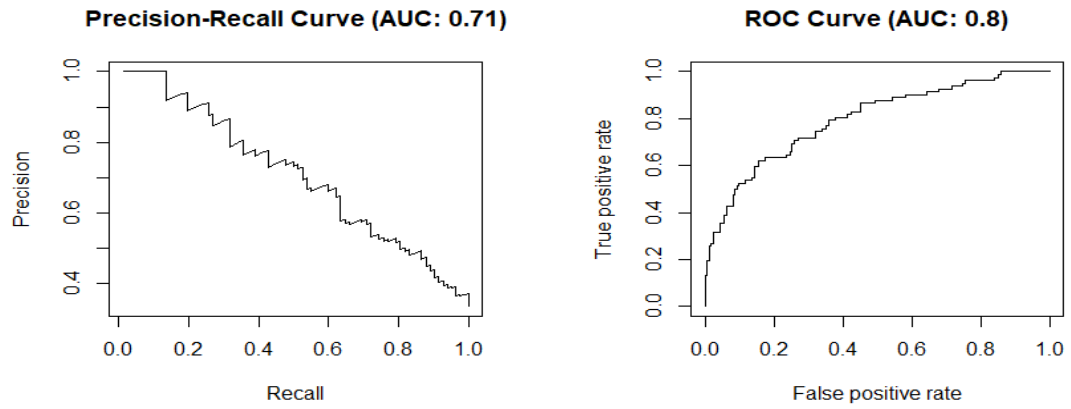


Figure 2.29: Random Effect in Grading MEM with Clinical+Portal(Core) Feature

The VPC is 29.86%. The value is very high, so that the effect of the centre is strongly present. In Figure 2.29 can be seen that for Torino and Milano there is statistical evidence to say that they have random effect different from zero

Performances are summarized in Table 2.11 and Figure 2.30



(a) (b)

Figure 2.30: Precision-Recall and ROC Curves for Grading MEM for Clinical+Portal(Core) covariates

Table 2.11: Performances of Grading MEM with Clinical+Portal(Core) features

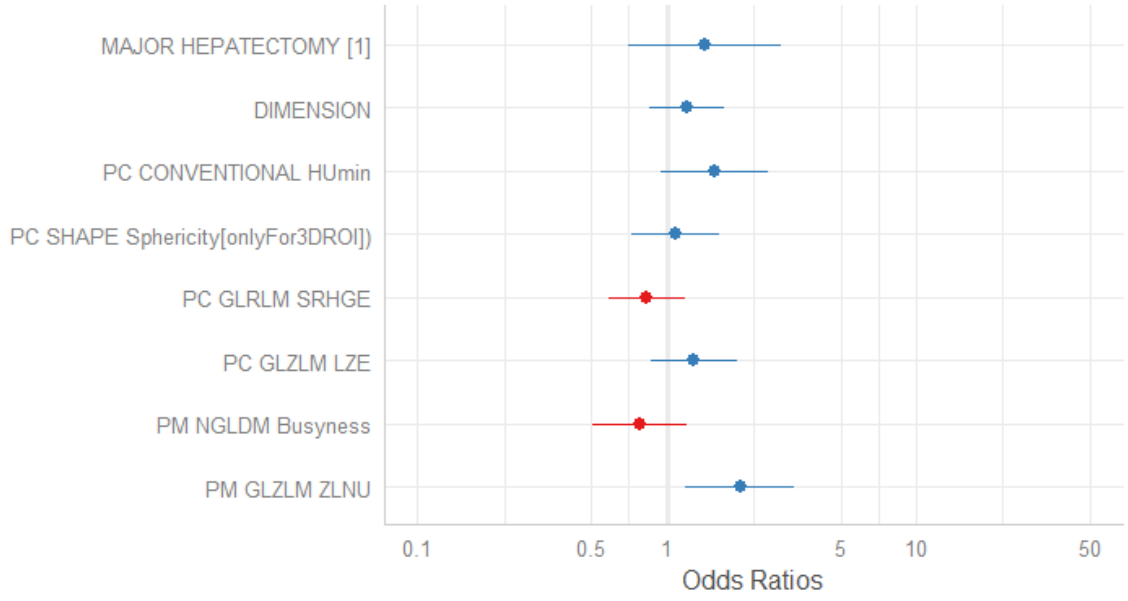
Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.77	0.573	0.362	0.719	0.064
SPECIFICITY	0.788	0.785	0.185	0.765	0.046
SENSITIVITY	0.717	0.503	0.ing	0.598	0.123
PRECISION	0.524	0.5	0.429	0.462	0.125
PR AUC	0.708	0.696	0.345	0.577	0.112
ROC AUC	0.795	0.74	0.322	0.711	0.077

Adding radiomic covariates of the tumor to the model produces an increase in the values of all performances. Therefore, radiomic covariates belonging to the core are important in improving the predictive ability of the model.

Mixed Effects Model for Grading with Clinical+Portal(Core+Margin) Features

These are the results of the Grading MEM with the features identified in the best model of Logistic Regression with Clinical+Portal(Core+Margin) variables. The results of the fixed effect are reported in Figure 2.31.

Figure 2.31: Odds ratios with 95% CI obtained applying MEMs for Grading with Clinical+Portal(Core+Margin) features



One radiomic variable belonging to the margin is significant, but other variables lose significance with respect to the Logistic Regression case because of the presence of the centre.

The random effect estimates are illustrated in Figure 2.32

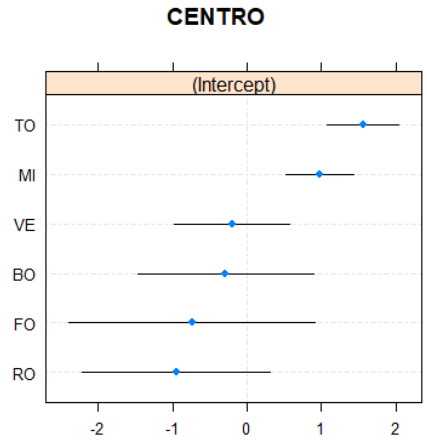


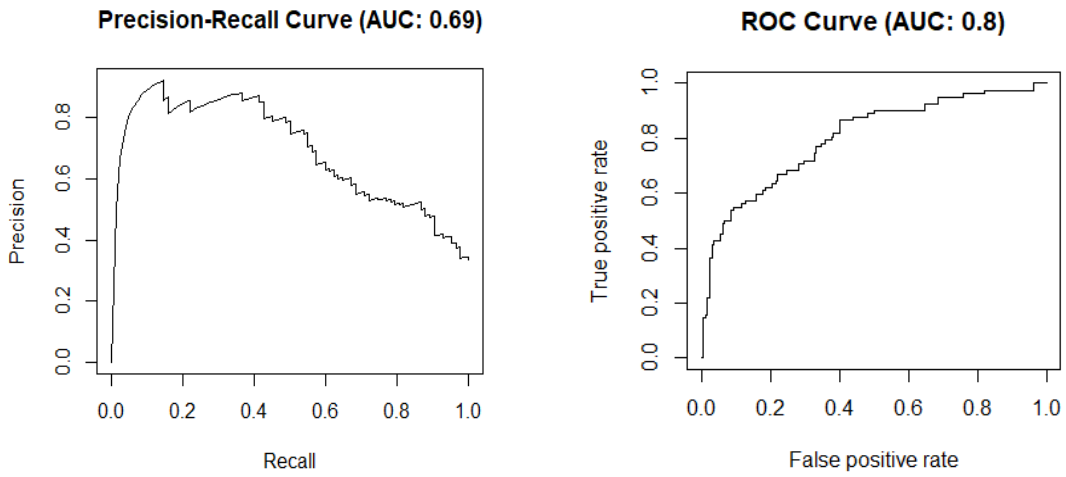
Figure 2.32: Random Effect in Grading MEM with Clinical+Portal(Core+Margin) Feature

The VPC is 26.62%. This testifies that the centre effect is strongly present in the data and it can be seen that there is evidence to say that Torino and Milano hospitals have an effect that is different from zero. Among the various models considering the different sets of covariates, it can be seen that the estimated random effects are consistent with each other.

Performances are summarized in Table 2.12 and Figure 2.33.

Table 2.12: Performances of Grading MEM with Clinical+Portal(Core+Margin) features

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.775	0.554	0.402	0.748	0.057
SPECIFICITY	0.783	0.794	0.173	0.773	0.042
SENSITIVITY	0.745	0.51	0.454	0.679	0.143
PRECISION	0.5	0.48	0.44	0.456	0.121
PR AUC	0.685	0.709	0.327	0.622	0.105
ROC AUC	0.802	0.765	0.276	0.753	0.066



(a) (b)

Figure 2.33: Precision-Recall and ROC Curves for Grading MEM for Clinical+Portal(Core+Margin) covariates

As in the case where the grouping factor was not considered, values of all performance metrics, except precision in cross-validation 2, including margin covariates, improve. This, joint with the fact the only covariate significant in the model belongs to the margin, testifies that radiomics of the peritumoral area is important for predicting more accurately the value of Grading.

2.2.3. Summary for Classification

In order to have a more powerful way of demonstrating the added value of radiomics and margin than simply looking at the increase in cross-validation performance, that is nonetheless present, the performances on the test set of the various samples of the cross-validation 2 are exploited. We have collected the performances on each single test set sample, producing a new dataset with 100 rows that correspond to a sample and a column for each performance matrix, 6 in total. This dataset is produced for every MEM that is fitted and it is used to perform *Permutation Tests* on the mean. Permutation tests [37] are nonparametric test procedures to test the null hypothesis that two different groups come from the same distribution. This type of tool is very useful because it does not require any assumption about sampling distribution. Permutation tests are used by us to test, for each performance metric, whether there is statistical evidence to say that, by including radiomics and margin, the average of a given metric is higher than in the case where less information is considered.

Formally, we will do the following one-sided tests for the means:

1. H_0 : Clinical $\geq$ Clinical + Core vs H_1 : Clinical $<$ Clinical + Core
2. H_0 : Clinical $\geq$ Clinical + Core + Margin vs H_1 : Clinical $<$ Clinical + Core + Margin
3. H_0 : Clinical + Core$\geq$ Clinical + Core + Margin vs H_1 : Clinical + Core $<$ Clinical + Core + Margin

With tests 1 and 2 we want to prove that there is evidence to say that adding radiomic covariates to the model makes the mean of the performances greater, while with test 3 we want to understand if adding the margin the mean of the performances improves.
The test statistic used to perform the tests on population X_1 and X_2 is $T = mean(X1) - mean(X2)$ with null hypothesis $H_0 : mean(X_1) > mean(X_2)$ that we aim to reject. Result of the test is summarized with the p-value, and are carried out for MVI and Grading for every performance metric.

The result for MVI are reported in Table 2.13

Table 2.13: P-values of Permutation Tests applied to values of performances obtained in Cross-validation 2 method while classifying MVI with MEMs

	Cliniche vs Core	Cliniche vs Core + Margin	Core vs Core + Margin
ACCURACY	0.0221	0.0001	0.0301
SPECIFICITY	0.0408	0.0009	0.0935
SENSITIVITY	0.0544	0.0001	0.011
PRECISION	0.0734	0.0473	0.361
PR AUC	0.0043	$<$ 0.0001	0.1126
ROC AUC	0.0066	$<$ 0.0001	0.0549

The result for Grading are reported in Table 2.14

The results of these tests further reinforce what has already been said. In conclusion, we can say that radiomics bring added value to the predictive performances of the models they are inserted. In particular, these data also highlight the additional information that the part of radiomics associated with the area surrounding the tumour brings. It should also be remembered that in the analysis of pathology data, both for MVI and Grading, there is a strong centre effect. Therefore, the variable recording the hospital of origin of the patients is crucial in explaining part of the variability of the data. It is very important that this aspect is taken into account by clinicians: the centre effect in our case

Table 2.14: P-values of Permutation Tests applied to values of performances obtained in Cross-validation 2 method while classifying Grading with MEMs

	Cliniche vs Core	Cliniche vs Core + Margin	Core vs Core + Margin
ACCURACY	0.0068	< 0.0001	0.0004
SPECIFICITY	0.0006	< 0.0001	0.1111
SENSITIVITY	0.0125	< 0.0001	< 0.0001
PRECISION	0.0005	< 0.0001	0.6422
PR AUC	$<$0.0001	< 0.0001	0.0013
ROC AUC	0.4055	< 0.0001	< 0.0001

could be caused by the different case mix between hospitals, but it could also indicate inhomogeneities between hospitals in the implemented protocols for CT scans and for the analysis of histological samples.

3 | Survival Analysis

In this Chapter we focus on the study of time-to-event data involving patient survival. The outcome analysed are:

- **Overall Survival (OS):** it is the time from either the date of diagnosis or the start of treatment up to patient's death or the end of the study [38].
- **Relapse-Free Survival (RFS):** it is the time from either the date of diagnosis or the start of treatment up to patient's recurrence or the end of the study without any sign or symptoms of cancer [39].

In order to study this type of data Survival Analysis is used. Survival Analysis techniques are used in order to identify the best model to describe the survival response in the case of OS and RFS.

3.1. Methodologies for Survival Analysis

In this Section we describe the methodologies used to analyse the time-to-event data. In Section 3.1.1 a short Introduction of Survival Analysis is given. Section 3.1.2 introduces Log-Rank Test. In Section 3.1.3 we introduce Cox Proportional-Hazard model to describe survival response of the patient with independence assumption. In Section 3.1.4 Shared Frailty model for considering the multicentre nature of the data is illustrated.

3.1.1. Introduction to Survival Analysis

Survival Analysis [40] is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs. It differs from other techniques due to the presence of censoring.

Censoring

In this study, censored data are represented by patients that survived (for OS) or that did not present tumour recurrence (for RFS) until the follow-up period. Censoring occurs

because we do not know the survival time exactly, the only thing we are aware of is that survival time is longer than censoring time. In this case the observations are indicated as *right-censored*.
Formally, for each patient i let T_i^* be the random variable denoting the true event time and let C_i be the censored time. The survival time observed is:

$$T_i = min(T_i^*, C_i) \tag{3.1}$$

Another piece of information known to us is whether or not the observed data corresponds to the censored time. For this quantity is defined the indicator random variable:

$$\delta_i = I(T_i^* \geq C_i) \tag{3.2}$$

Hence, the observation related to the *time-to-event* data for a patient i is the pair (T_i, δ_i)

Survival and Hazard Function

To model the survival time, denoted by the random variable T, two equivalent characterization are used: *survival function* and *hazard function*.

Definition 3.1.1. The **survival function** of T at time t, denoted by $S(t)$, is the probability that an individual survives longer that t:

$$S(t) = \mathbf{P}(T > t) = 1 - \mathbf{P}(T \leq t) = 1 - F(t) \tag{3.3}$$

where $F(t)$ is the cumulative density function of T.

The survival function $S(t)$ is an estimate of the percentage of individuals in a cohort who are still event free at time t. Therefore, the property of this function are:

- $S(0) = 1$: at the beginning of the study all individuals are alive
- $S(t)$ is non-increasing
- $S(t)$ may never reach zero if all the subjects do not experience the event by the end of the study

The graph of $S(t)$ is the *survival curve*. To estimate the survival function $S(t)$, the *Kaplan-Meier estimator* is used. It is a non-parametric statistic that is defined as the probability of surviving in a given length of time while considering in small intervals. Under proper assumptions, Kaplan-Meier estimator is computed through the maximization

of the likelihood estimation of the hazard function.

Definition 3.1.2. Given $j \in 1, \ldots, J$ as the failure event index, $0 < t_1^* < \cdots < t_J^* < \infty$ as the observed ordered times of deaths, n_j as the number of individuals alive just before t_j^*, d_j as the number of observed events at t_j^* and p_j as the conditional probability of surviving time t_j^*, the **Kaplan-Meier estimator** of the survival function $S(t)$ is:

$$\widehat{S}(t) = \prod_{j:t^*} p_j = \prod_{j:t^*} \left(1 - \frac{d_j}{n_j}\right) \tag{3.4}$$

With the *Greenwood's formula* for the **estimated variance**, which is:

$$\widehat{Var}\left(\hat{S}(t)\right) = \left[\hat{S}(t)\right]^2 \sum_{j:t^*} \frac{d_j}{n_j(n_j - d_j)} \tag{3.5}$$

the 95% confidence interval for the Kaplain-Meier survival estimator is expressed as:

$$CI_{0.95}\big(S(t)\big) = \left[\hat{S}(t) \pm z_{0.975}\hat{se}(t)\right] \tag{3.6}$$

This procedure produces the Kaplan-Meier curve which is a step function with jumps at the observed death times.

The other characterization used to model the survival time T is the hazard function.

Definition 3.1.3. The **hazard function** of T, denoted by $h(t)$, is the instantaneous risk of failure at time t, conditional on survival to that time:

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbf{P}(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{3.7}$$

Hazard function can be expressed in terms of survival function $S(t)$ and probability density function $f(t)$ of T as:

$$h(t) = \frac{f(t)}{S(t)} \tag{3.8}$$

It is a measure of the proneness to failure as a function of the age of the individual [41].

Another measure that is important in Survival Analysis is the cumulative hazard function, that can be interpreted as the cumulative force of mortality.

Definition 3.1.4. The **cumulative hazard function** of T at time t, denoted by $H(t)$, is:

$$H(t) = \int_0^t h(u)du = -\ln[S(t)] \tag{3.9}$$

3.1.2. Log-Rank Test

The *Log-Rank test* is the most commonly-used non-parametric statistical test for comparing the survival distribution of two or more groups. With this test, we try to disprove the null hypothesis that all survival curves of the various groups are equal. Formally:

$$H_0 : S_1(\cdot) = \cdots = S_K(\cdot) \qquad vs \qquad H_1 : \text{Survival curves are not identical}$$

To perform the test, the test statistic must be found. Its calculation is based on a contingency table of group by status at each observed survival time, as shown in Table 3.1 [42]. In this Table n_{kj} is the number at risk in group k at observed survival time t_j^*, d_{kj} is the number of observed death in group k, n_j is the total number at risk and d_j is the total number of deaths.

Event/Group	K	...	1	0	Total
Die	d_{Kj}	...	d_{1j}	d_{0j}	d_j
Not Die	$n_{Kj} - d_{Kj}$	...	$n_{1j} - d_{1j}$	$n_{0j} - d_{0j}$	$n_j - d_j$
At Risk	n_{Kj}	...	n_{1j}	n_{0j}	n_j

Table 3.1: Table Used for Log-Rank Test in K groups at Observed Survival time t_j^*

With these quantities just introduced, we can define the number of expected events in group k at time t_j^* as:

$$e_{kj} = \frac{d_j}{n_j} n_{kj} \tag{3.10}$$

We can now state the formula and the distribution of the approximated **Log-Rank test statistic**:

$$\chi^2 = \sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} \sim \chi^2_{K-1} \quad \text{with} \quad O_K = \sum_{j=1}^{J} d_{kj} \quad \text{and} \quad E_K = \sum_{j=1}^{J} e_{kj} \tag{3.11}$$

H_0 is rejected at statistical level α if $\chi^2 > \chi^2_{K-1,\alpha}$.

In our analysis clinical covariates are many, particularly categorical ones. We use Log-Rank Test in Section 3.2.1 to have a reduction in the number of categorical clinical covariates, eliminating those for which there is no evidence to say that the survival curves are different.

3.1.3. Cox Proportional Hazard Model

Since in our case we are provided with censored observation, we cannot apply a standard regression method; but survival methods are able to handle censored data which are not considered in traditional models. The *Cox Proportional Hazard* (Cox-PH) [43] is the mostly used mathematical model for doing regression with time-to-event data. It also allows exploring the relationship between the survival of an individual and several explanatory variables. In this setting, individuals are considered independent and identically distributed.
The Cox-PH model is written in terms of the hazard function. Let $\mathbf{x}_i$ be the vector of the predictor variables, let $h_0(t)$ be a non-negative function of time called *baseline hazard* and let β be the vector of the coefficients that need to be estimated, the Hazard function in Cox-PH model is assumed to have the following formula:

$$h_i(t|\mathbf{x}_i) = h_0(t)exp(\mathbf{x}_i^T\beta) \tag{3.12}$$

The model is semiparametric because of the presence of the unspecified function $h_0(t)$.
The assumption on which the Cox-PH model is based is that the ratio of the hazard function of two patients with fixed covariates is constant over time. This quantity is called **Hazard Ratio**:

$$HR = \frac{h_i(t|\mathbf{x}_i)}{h_k(t|\mathbf{x}_k)} = exp(\mathbf{x_i} - \mathbf{x}_k)^T\beta \tag{3.13}$$

With Hazard Ratio we can assess the effect of a change in a predictor variable, since we are able to quantify the change of the Hazard Ratio as one covariate increases by one unit:

$$HR_k = \frac{h(t|x_1, \ldots, x_k, \ldots, x_p)}{h(t|x_1, \ldots, x_k + 1, \ldots, x_p)} = e^{\beta_k} \tag{3.14}$$

Depending on the value of the HR_k, the interpretation of the effect of the predictor

variable is different:

- $HR_k = 1$ means that the k-th covariate has no effect
- $HR_k < 1$ results into a reduction in the hazard, so that the k-th covariate is a good prognostic factor
- $HR_k > 1$ results into an increase in the hazard, so that the k-th covariate is a bad prognostic factor

The Cox-PH model parameters β are derived by maximizing the likelihood function. In the case of Cox-PH model we talk about *partial* likelihood, because it considers probabilities only for those subjects who fail and does not explicitly consider probabilities for subjects who are censored.
The Cox partial likelihood is express as the product of several likelihoods, one for each failure time. Let J be the total number of deaths, let $0 < t_1^* < \cdots < t_J^*$ be the ordered observed deaths times and let $R(t_j^*)$ be the risk set just before t_j^*, we define:

$$L_j = \frac{\exp(\mathbf{x}_j^T \beta)}{\sum_{k \in R(t_j^*)} \exp(\mathbf{x}_k^T \beta)} \tag{3.15}$$

as the conditional probability that the individual j-th dies at t_j^* given that one individual from the risk set on $R(t_j^*)$ dies at t_j^*. The Cox partial likelihood $\mathcal{L}(\beta)$ is the product of these latter conditional probabilities L_j:

$$\mathcal{L}(\beta) = \prod_{j=1}^{J} L_j = \prod_{j=1}^{J} \frac{\exp(\mathbf{x}_j^T \beta)}{\sum_{k \in R(t_j^*)} \exp(\mathbf{x}_k^T \beta)} \tag{3.16}$$

The parameters are found through log-likelihood maximization, namely:

$$\hat{\beta} = \underset{\beta \in \mathrm{R}^p}{\arg\max} \ln(\mathcal{L}(\beta)) \tag{3.17}$$

In the context of Cox-PH model, in which several covariates are present, it is interesting to study the effect of the explanatory variable via **adjusted survival curves**. These are curves obtained through the Cox model that adjust for the explanatory variables. Adjusted survival function formulation is the following:

$$S_i(t|\mathbf{x}_i) = [S_0(t)]^{\exp(\mathbf{x}_i^T \beta)} \quad \text{with} \quad S_0(t) = \exp\left\{ - \int_0^t h_0(u) du \right\} \tag{3.18}$$

where $S_0(t)$ is the **baseline survival curve**.

Its estimate is given by the formula:

$$\widehat{S}_i(t|\mathbf{x}_i) = \prod_{j:t_j^*<t} \left(1 - \frac{1}{\sum_{k(t_j^*)} \exp(\mathbf{x}_k^T \beta)}\right) \tag{3.19}$$

Cox-PH models are used in Section 3.2.2 to find the model best that describes survival response of the patients, considering them independent and identically distributed patients. This is a preliminary step, which is necessary in order to be able to consider the multicentre nature of the data in the model afterwards.

3.1.4. Shared Frailty Model

With Cox-PH models we are not able to focus on the multicentre nature of our data. Cox-PH models consider patients independent and identically distributed, but the fact that the patients in the study came from different hospitals might not guarantee the assumption of statistical independence. To overcome this aspect and to explore the centre effect we introduce Shared Frailty Models.
The concept of *frailty* provides a convenient way of introducing unobserved heterogeneity and association [44]. Without this modelling technique the population is implicitly assumed homogeneous, meaning that the individuals share the same risk of death. The aim is therefore to investigate whether the centre effect is a determinant of heterogeneity. Frailty model is a random effect model for time-to-event data, where the random effect (frailty) has a multiplicative effect on the baseline Hazard. Frailty represents an unobservable random effect shared by subjects with similar (unmeasured) risks in the analysis of mortality rates [45]. To model the frailty, a proportional hazard structure that is conditional on the random effect is assumed: the hazard function depends on the time-independent random variable Z. It enters the function in a multiplicative way so that:

$$h(t|Z) = Zh_0(t) \tag{3.20}$$

Z is a nonnegative random mixture variable, that varies across groups. Frailty is a measure of relative risk: the greater the frailty, the greater is the susceptibility to the cause of death. The variability of Z determines the degree of the heterogeneity among the groups [46]. Introducing the presence of predictors variable into the model, as in the Cox-PH model, the formula is:

$$h(t|\mathbf{x}_i, Z) = Zh_0(t)exp(\mathbf{x}_i^T \beta) \tag{3.21}$$

In shared frailty model [47] individuals in the same group share the same risk, so that the frailty is associated with the group. The value of the frailty term is common to all individuals in the cluster and all failure times in a cluster are conditionally independent given the frailties and event times from different clusters are considered to be independent. Let n be the number of clusters and let j cluster has n_j observation associated with the same unobserved frailty Z_j and let $\mathbf{x}_{ji}$ the vector that contains the information about the i-th observation in the j-th cluster, the hazard function of the survival times in cluster j conditional on the frailty Z_j is:

$$h(t|\mathbf{x}_{ji}, Zj) = Z_j h_0(t)\exp(\mathbf{x}_{ji}^T\beta) \tag{3.22}$$

the frailties Z_j are assumed to be independent and identically distributed with density function $f(z)$. From equation 3.22 the joint conditional multivariate survival function for the individuals in the j-th cluster can be derived. It holds that:

$$S(t_{j1}, \ldots, t_{jn_j}|\mathbf{x}_j, Z_j) = S(t_{j1}|\mathbf{x}_j 1, Z_j)\ldots S(t_{jn_j}|\mathbf{x}_{jn_j}, Z_j) = \exp\big(-Z_j\sum_{i=1}^{n_j} H_0(tji)e^{\mathbf{x}_{ji}^T\beta}\big) \tag{3.23}$$

where $H_0(t) = \int_0^{tji} h_0(u)du$. With this step we can derive the unconditional joint survival function. Averaging 3.23 with respect to Z_j the marginal survival function can be obtained:

$$S(t_{j1}, \ldots, t_{jn_j}|\mathbf{x}_j) = \mathbf{E}\big[S(t_{j1}, \ldots, t_{jn_j}|\mathbf{x}_j, Z_j)\big] = \mathbf{L}\big(\sum_{i=1}^{n_j} H_0(tji)e^{\mathbf{x}_{ji}^T\beta}\big) \tag{3.24}$$

where L is the Laplace transformation of the frailty variable. Because of the assumption of independence between cluster, we have that:

$$S(t_{11}, \ldots, t_{nn_j}|\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{j=1}^{n}\mathbf{L}\big(\sum_{i=1}^{n_j} H_0(tji)e^{\mathbf{x}_{ji}^T\beta}\big) \tag{3.25}$$

The univariate unconditional survival functions can be expressed with Laplace transformation:

$$S(t_{ji}|\mathbf{X}_{ji}) = \mathbf{E}\big[S(t_{ji}|\mathbf{X}_{ji}, Z_j)\big] = \mathbf{L}\big(H_0(tji)e^{\mathbf{x}_{ji}^T\beta}\big) \tag{3.26}$$

In this work we assume that the frailty follows a gamma distribution [47]. Despite the fact that there are no biological reasons for preferring the gamma distribution over the

others, it is nevertheless better for mathematical and computational aspects. It is usually used because of its simplicity of the derivatives of the Laplace transform [48].
Assuming for the frailty a gamma distribution with mean 1 and variance σ^2 the survival function is:

$$S(t_{j1}, \ldots, t_{jn_j}|\mathbf{x}_j) = \big(1 + \sigma^2 \sum_{i=1}^{n_j} H_0(tji) e^{\mathbf{x}_{ji}^T \beta}\big)^{-\frac{1}{\sigma^2}} \tag{3.27}$$

Since frailties are unobserved, Expectation Maximization (EM) algorithm is used to estimate the model. It is a combination of Expectation step (E-step), in which the expectation of the full likelihood is found given the current estimates of the parameters, and Maximization step (M-step), in which the parameters are updated maximizing the expected value. The full likelihood of which the expected value is calculated is:

$$\mathcal{L}_{full} = \prod_{j=1}^{n} \prod_{i=1}^{nj} Z_j^{\delta_{ji}} h_o(tji)^{\delta_{ji}} \exp(\delta_{ji} \mathbf{x}_{ji}^T \beta) \exp\big(- Z_j \sum_{i=1}^{n_j} H_0(tji) e^{\mathbf{x}_{ji}^T \beta}\big) f(Z_j) \tag{3.28}$$

The full likelihood is the product of the conditional and the density of the frailty. The information that we assume of the frailty is that it has distribution $f(z)$ with the unknown parameter θ.
Therefore, the steps of the EM algorithm are:

Algorithm 3.1 EM Algorithm for estimation of frailty model

1: Provide initial values of β, h_0 and θ
2: In the E-step plug values of β,h_0 and θ into the $\mathcal{L}_{full}$ and calculate the conditional expectation of Z_j
3: In the M-step plug the expectation in the partial likelihood $\mathcal{L}(\beta)$ and update the parameters β and h_0, and plug into $\mathcal{L}_{full}$ to update the estimate of θ
4: Repeat E-step and M-step until convergence

Summarizing, with heterogeneity some unexpected results can be explained and centre-to-centre variations can be described by the frailty [44]. In particular, in Shared Frailty models individuals in the same group share the same risk. Therefore, we use Shared Frailty models in Sections 3.2.3 to take into account the grouping present in our multicentre study.

3.2. Results of Survival Analysis

In the case of this study, in the first part, Survival Analysis is applied only to radiomic covariate belonging to Portal phase, jointly with clinical ones. Apart from the primary

objective of finding the best model capable of describing the survival response and identifying important features, in these analyses we want to understand the radiomics role, and thus the importance of tumor (Core) and peritumoral area (Margin) in predicting outcome. In addition, in order to understand whether radiomics can provide adequate non-invasive preoperative assessment, we decided to consider preoperative and postoperative clinical covariates separately, following a clinical rationale (for the differentiation we refer the reader to Section 1.2.1 and Appendix A). To answer these questions, in order to study the prognostic impact of the radiomic features on the responses, six scenarios of grouped covariates are analysed:

- Clinical Preoperative
- Clinical Preoperative + Portal(Core)
- Clinical Preoperative + Portal(Core+Margin)
- Clinical Postoperative
- Clinical Postoperative + Portal(Core)
- Clinical Postoperative + Portal(Core+Margin)

As a first step, to reduce the number of input variables in the model, we use Log-Rank Test and the results of the skimming can be seen in Section 3.2.1. Afterwards, for each of the above cases, in the Section 3.2.2, not taking into account the multicentre nature of the data (summarized in the variable *centre*), Cox Proportional Hazard models are fitted, in order to find the best model. With the covariates given by the latter, a Shared Frailty model is used in Section 3.2.3 to study the grouping effect.

3.2.1. Log-Rank Test for Variables Skimming

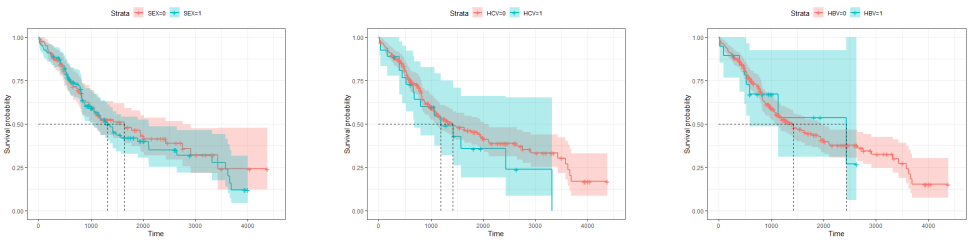
In this Section, since in our analysis clinical covariates are many, particularly categorical ones, we proceed by reducing the number of the latter, using Log-Rank Test. It is carried out for each of categorical covariates: with this test we can see for which one we have evidence to state that the Kaplan-Meier curves are different, depending on the value assumed by the variable. Features for which we cannot say that there is a difference in the curves will be removed in subsequent Survival Analysis, in order to have a first skim of the variables. The p-values of the tests are summarized in Table 3.2 and plots of the curves for all categorical variables are provided in Figures 3.1 and 3.2. Only variables with p-value less than 0.10 are considered in subsequent analysis, reducing the number of clinical postoperative covariates by approximately 40% for both OS and RFS.

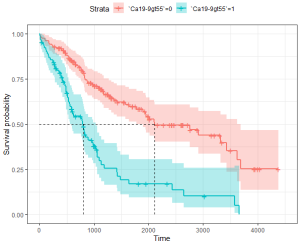
Table 3.2: P-values of Log Rank Tests performed for OS and RFS for each clinical categorical feature in IHC dataset

Feature	P-value (OS)	P-value (RFS)
SEX	0.7	0.8
HCV	0.3	0.9
HBV	1	0.3
Ca19-9 $\geq$ 55	3e-10	1e-05
NEOADJUVANT CHEMOTHERAPY	1	0.08
FIRST RESECTION	0.7	0.04
MAJOR HEPATECTOMY	1e-03	0.4
BILIARY RESECTION	1e-04	0.8
LYMPHADENECTOMY	0.06	0.2
ASSOCIATED RESECTION	0.2	1e-03
SEVERE COMPLICATIONS	2e-07	0.03
CIRRHOSIS	0.6	1
PATTERN	1e-05	3e-06
SINGLE NODULE	1e-03	1e-06
T VIII ed	2e-04	1e-03
N	2e-06	2e-05
M	1e-08	5e-06
GRADING(1-2 vs 3)	0.02	0.3
R	8e-05	1e-03
MICROSCOPIC VASCULAR INVASION	1e-03	4e-03
PERINEURAL INFILTRATION	2e-04	7e-03
SATELLITE NODULES	3e-06	1e-04
ADJUVANT CHEMOTHERAPY	0.2	0.04

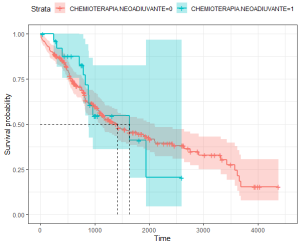
Figure 3.1: OS Kaplain-Meier Curves Estimates with 95% CI for clinical categorical feature in IHC dataset

(a) SEX (b) HCV (c) HBV

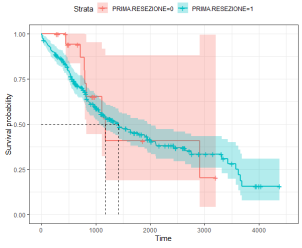




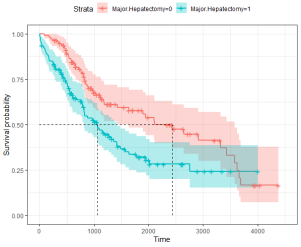
(d) Ca19-9≥55



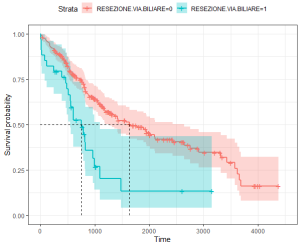
(e) NEOADJUVANT CHEMOTHERAPY



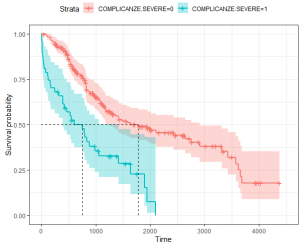
(f) FIRST RESECTION



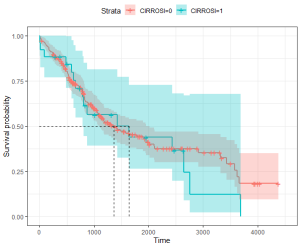
(g) MAJOR HEPATECTOMY



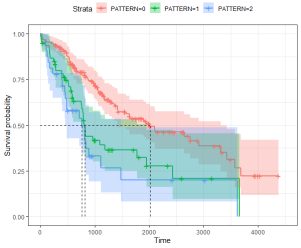
(h) BILIARY RESECTION



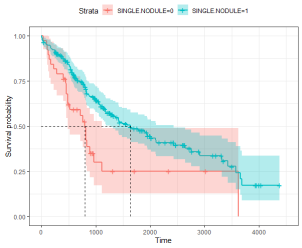
(i) SEVERE COMPLICATIONS



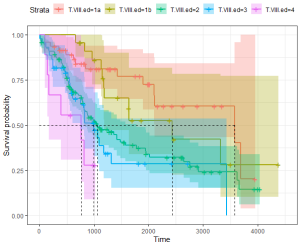
(j) CIRRHOSIS



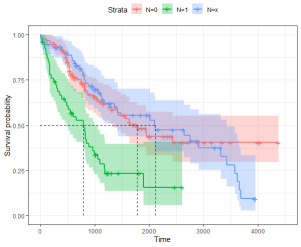
(k) PATTERN



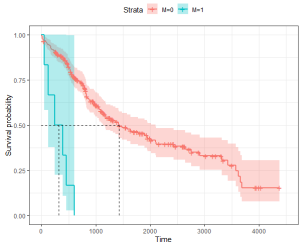
(l) SINGLE NODULE



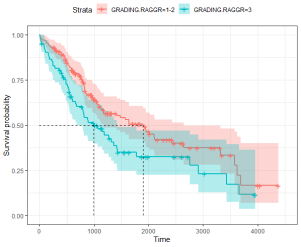
(m) T VIII ed



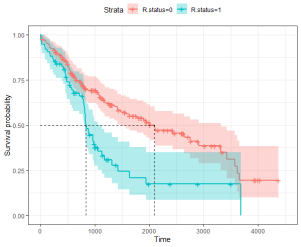
(n) N



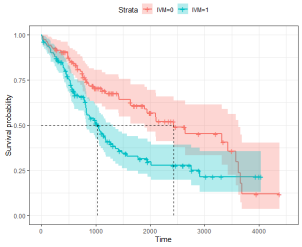
(o) M



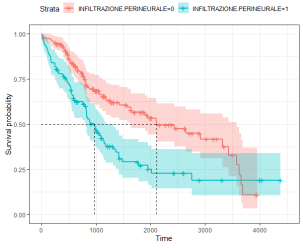
(p) GRADING (1-2 vs 3)



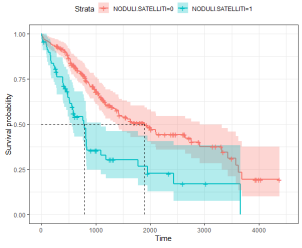
(q) R



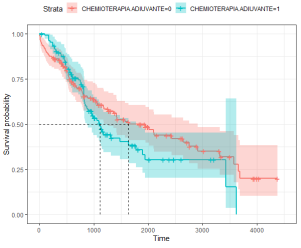
(r) MICROSCOPIC VASCULAR INVASION



(s) PERINEURAL INFILTRATION



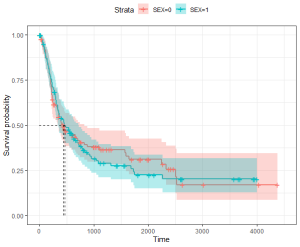
(t) SATELLITE NODULES



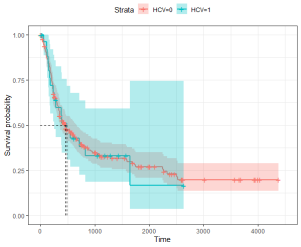
(u) ADJUVANT CHEMOTHERAPY

Figure 3.2: RFS Kaplain-Meier Curves Estimates with 95% CI for clinical categorical feature in IHC dataset

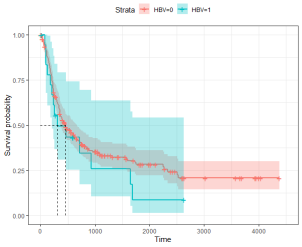
(a) SEX



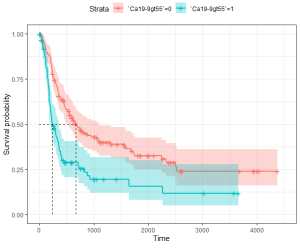
(b) HCV



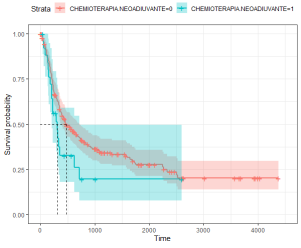
(c) HBV



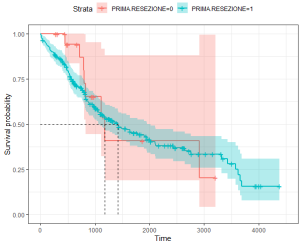
(d) Ca19-9≥55



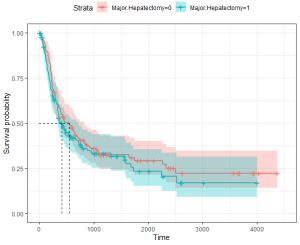
(e) NEOADJUVANT CHEMOTHERAPY



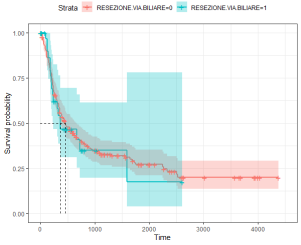
(f) FIRST RESECTION



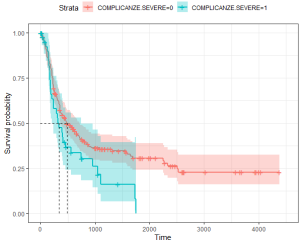
(g) MAJOR HEPATECTOMY

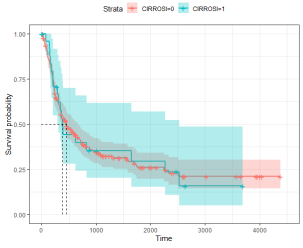


(h) BILIARY RESECTION

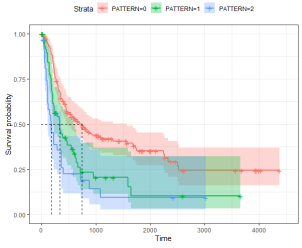


(i) SEVERE COMPLICATIONS

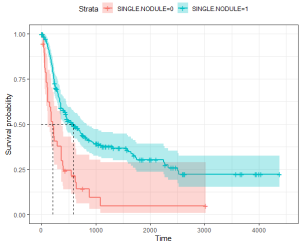




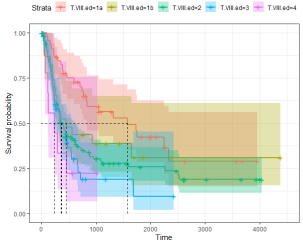
(j) CIRRHOSIS



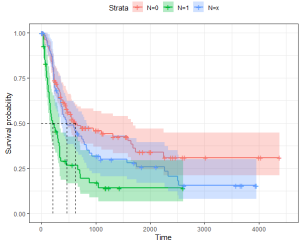
(k) PATTERN



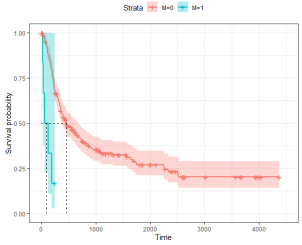
(l) SINGLE NODULE



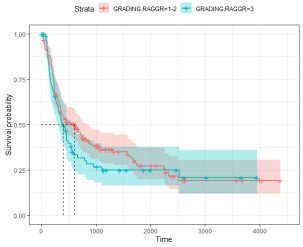
(m) T VIII ed



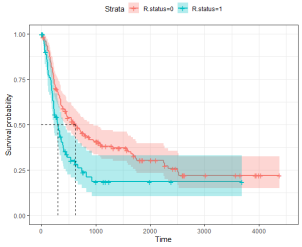
(n) N



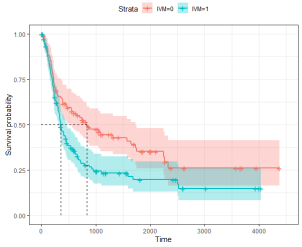
(o) M



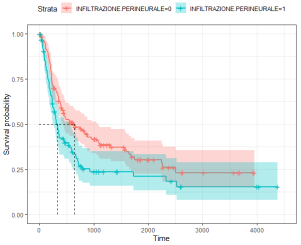
(p) GRADING (1-2 vs 3)



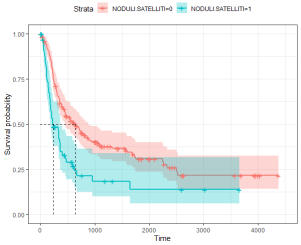
(q) R



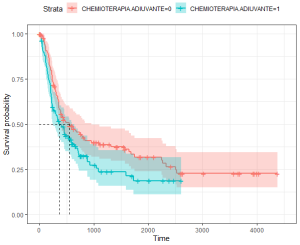
(r) MICROSCOPIC VASCULAR INVASION



(s) PERINEURAL INFILTRATION



(t) SATELLITE NODULES



(u) ADJUVANT CHEMOTHERAPY

3.2.2. Cox-PH Models for identifying the best models

In this Section, results of Cox-PH models are reported. For each different grouping of covariates considered, Cox-PH models are used to estimate survival curves first for OS and then for RFS, combined with the Stepwise Algorithm as a feature selection technique. For each model, statistics about coefficients are provided, jointly with Hazard Ratio and Estimate of the Baseline Survival Curve.

The Concordance index (C-Index) [49] is used as evaluation metric, since it validates the predictive ability of a survival model. The model selected is the one with the highest C-index.
For sake of simplicity, only the best models are reported in this Chapter, while other attempts are summarized in the Appendix D.

Overall Survival Cox-PH Best Model

The best model for representing OS response is the one in which all clinical and portal covariates (tumor and peritumor area) enters as input, namely **Postoperative + Portal (Core + Margin)** case. The model selects clinical, core and margin covariates in different percentages:

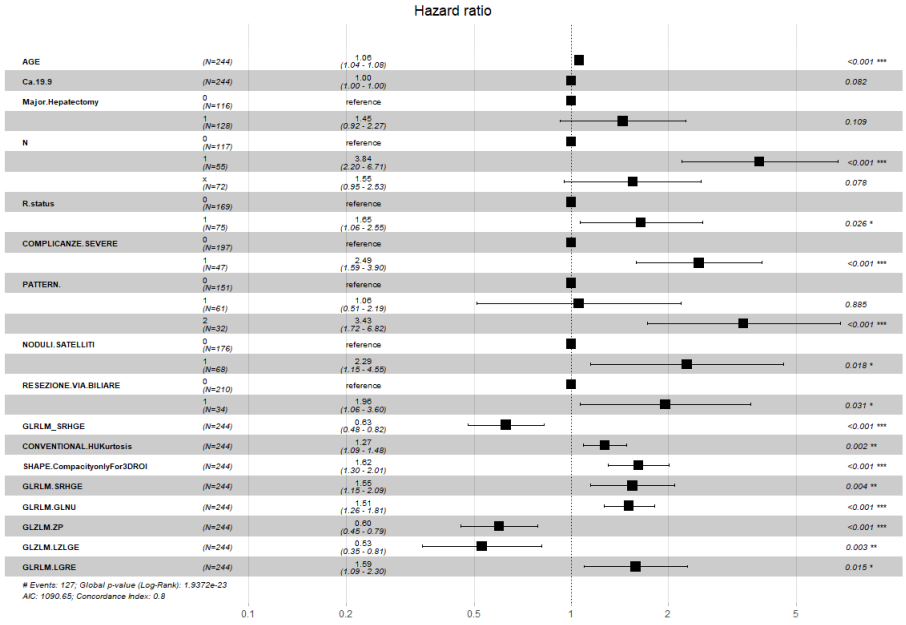
- **CLINICAL:** 9 features selected among all clinical covariates (53%)
- **PORTAL CORE:** 1 features selected in core (4%)
- **PORTAL MARGIN:** 7 features selected in margin (23%)

The coefficients of the model are summarized in Table 3.3:

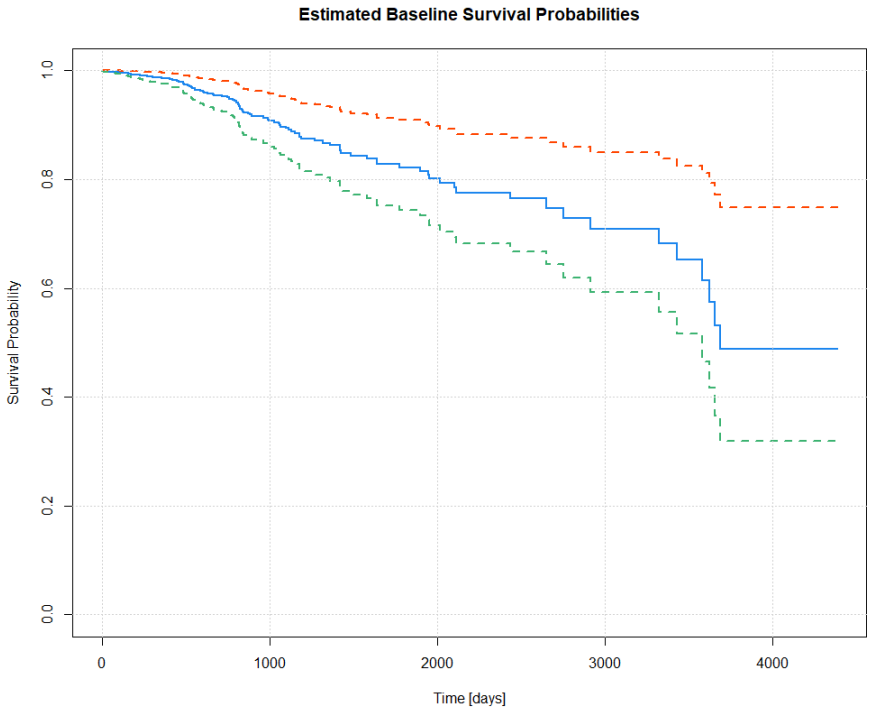
	coef	exp(coef)	se(coef)	z	Pr(>|z|)
AGE	5.885e-02	1.061e+00	9.928e-03	5.928	3.06e-09
CA 19-9	2.132e-05	1.000e+00	1.225e-05	1.741	0.081759
MAJOR HEPATECTOMY=1	3.684e-01	1.445e+00	2.296e-01	1.604	0.108668
N=1	1.346e+00	3.841e+00	2.848e-01	4.725	2.30e-06
N=x	4.400e-01	1.553e+00	2.493e-01	1.765	0.077506
R=1	4.987e-01	1.647e+00	2.235e-01	2.231	0.025659
SEVERE COMPLICATIONS=1	9.122e-01	2.490e+00	2.290e-01	3.984	6.79e-05
PATTERN=1	5.398e-02	1.055e+00	3.732e-01	0.145	0.885000
PATTERN=2	1.232e+00	3.427e+00	3.510e-01	3.509	0.000450
SATELLITE NODULES=1	8.276e-01	2.288e+00	3.511e-01	2.357	0.018420
BILIARY RESECTION=1	6.725e-01	1.959e+00	3.110e-01	2.162	0.030597
CORE GLRLM SRHGE	-4.669e-01	6.269e-01	1.385e-01	-3.370	0.000751
MAR CONV HUKurtosis	2.387e-01	1.270e+00	7.853e-02	3.040	0.002369
MAR SHAPE Compacity	4.794e-01	1.615e+00	1.111e-01	4.317	1.58e-05
MAR GLRLM SRHGE	4.385e-01	1.550e+00	1.526e-01	2.874	0.004055
MAR GLRLM GLNU	4.134e-01	1.512e+00	9.199e-02	4.494	6.98e-06
MAR GLZLM ZP	-5.138e-01	5.982e-01	1.410e-01	-3.643	0.000269
MAR GLZLM LZLGE	-6.370e-01	5.289e-01	2.172e-01	-2.933	0.003353
MAR GLRLM LGRE	4.613e-01	1.586e+00	1.892e-01	2.438	0.014776

Table 3.3: Coefficient Summary of Cox-PH model for OS with Postoperative+Portal(Core+Margin) covariates

The Hazard ratios with corresponding 95% CI and Estimated Baseline Survival Curve are reported in Figure 3.3:



(a) Hazard Ratio



(b) Estimated Baseline Survival Curves with 95% CI

Figure 3.3: Hazard ratios with corresponding 95% CI and Estimated Baseline Survival Curve for Cox-PH model with OS Postoperative+Portal(Core+Margin) covariates.

The C-index of the model is 0.797. The value is sufficiently high: the model provides good predictive performances.
From Table 3.3 it can be seen that there is statistical evidence to say that almost all coefficients are significant. From Hazard Ratio in Figure 3.3b, regarding clinical variables, it can be deduced that the presence of metastases (described by variable M), nodules (Pattern and satellite nodules) and residual of the tumour (R status), complications and biliary resection increase the risk of death.

Relapse Survival Cox-PH Best Model

The best model for representing RFS response is the one in which all clinical and portal covariates (tumor and peritumor area) enters as input, namely **Postoperative + Portal(Core + Margin)** case. The coefficients of the model are summarized in Table 3.4:

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
CA 19-9	4.271e-05	1.000e+00	1.252e-05	3.411	0.000648
PATTERN=1	4.631e-01	1.589e+00	2.107e-01	2.198	0.027929
PATTERN=2	9.924e-01	2.698e+00	2.359e-01	4.207	2.58e-05
N=1	7.119e-01	2.038e+00	2.378e-01	2.994	0.002751
N=x	3.303e-01	1.391e+00	2.139e-01	1.544	0.122655
M=1	1.149e+00	3.155e+00	5.155e-01	2.229	0.025800
R=1	3.019e-01	1.352e+00	1.981e-01	1.524	0.127545
CORE SHAPE Sphericity	2.673e-01	1.306e+00	1.115e-01	2.398	0.016495
CORE GLRLM LGRE	2.875e-01	1.333e+00	9.104e-02	3.158	0.001586
CORE GLZLM LZE	3.253e-01	1.385e+00	1.016e-01	3.201	0.001367
MAR SHAPE Sphericity	-1.910e-01	8.261e-01	1.316e-01	-1.452	0.146600
MAR GLZLM SZHGE	3.197e-01	1.377e+00	8.354e-02	3.827	0.000130
MAR GLZLM LZLGE	-6.300e-01	5.326e-01	2.162e-01	-2.914	0.003573
MAR GLZLM ZP	-3.211e-01	7.253e-01	1.129e-01	-2.845	0.004439
MAR GLRLM GLNU	2.856e-01	1.331e+00	9.326e-02	3.062	0.002197
MAR GLCM Correlation	-2.162e-01	8.055e-01	9.720e-02	-2.225	0.026104

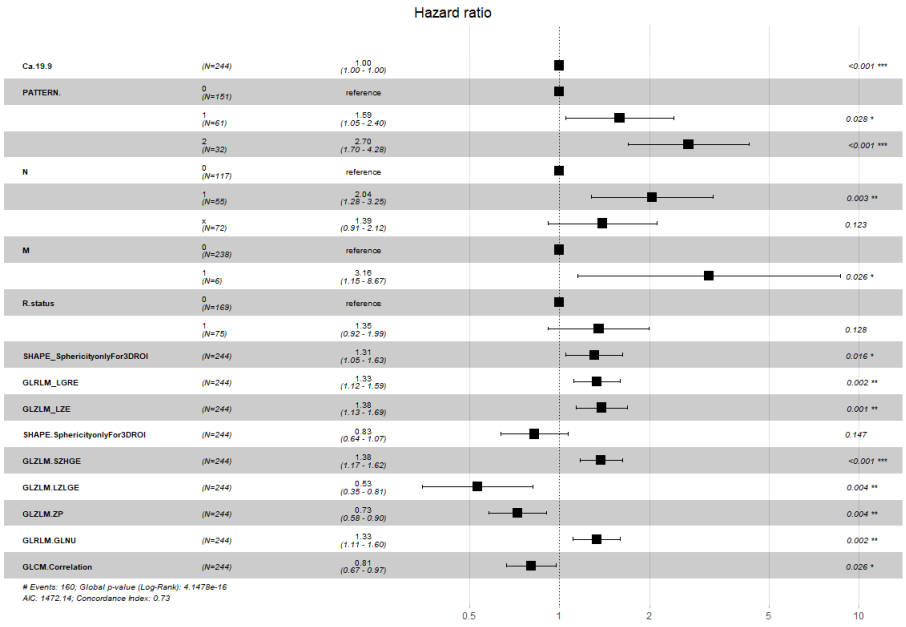
Table 3.4: Coefficient Summary of Cox-PH model for RFS with Postoperative+Portal(Core+Margin) covariates

The model selects clinical, core and margin covariates in different percentages:

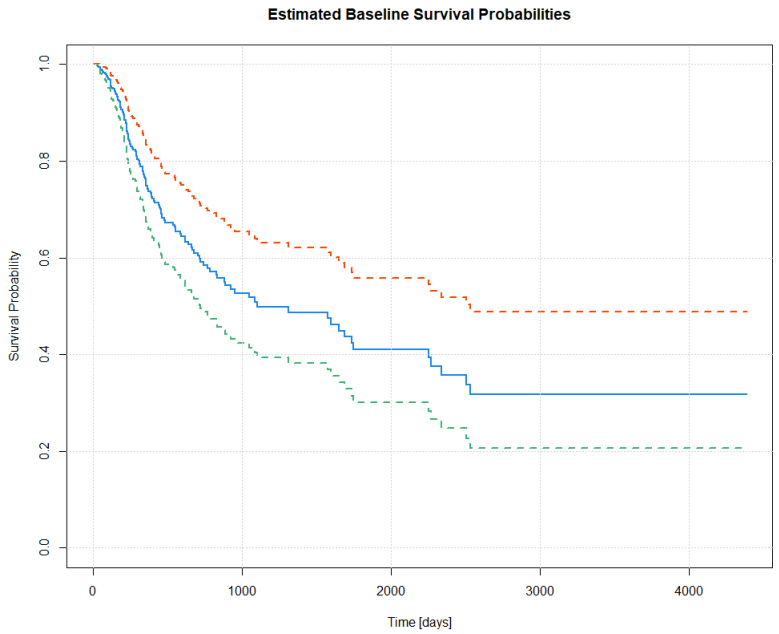
- **CLINICAL:** 5 features selected among all clinical covariates (29%)
- **PORTAL CORE:** 3 features selected in core (12%)

- **PORTAL MARGIN:** 6 features selected in margin (20%)

The Hazard ratios with corresponding 95% CI and Estimated Baseline Survival Curve are reported in Figure 3.4:



(a) Hazard Ratio



(b) Estimated Baseline Survival Curves with 95% CI

Figure 3.4: Hazard ratios with corresponding 95% CI and Estimated Baseline Survival Curve for Cox-PH model with RFS Postoperative+Portal(Core+Margin) covariates.

The C-index of the model is 0.733. The value is fulfilling: the model provides good predictive performances.
From Table 3.4 it can be seen that there is statistical evidence to say that almost all coefficients are significant. From Hazard Ratio in Figure 3.4b, regarding clinical variables, it can be deduced that the presence of metastases (described by variable M and N), nodules (Pattern) and residual of the tumour (R status) increase the risk of relapse.

3.2.3. Shared Frailty Models for considering the grouping factor

In this Section results of Shared Frailty models are reported first for OS, then for RFS. For both best models identified with Cox-PH, a Shared Frailty Model for OS and RFS is fitted with the covariates selected. With Shared Frailty the multicentre nature of the data is considered. In addition, the Commenges-Andersen test is carried out to assess the presence of heterogeneity [50]. For these analyses the R package `frailtyEM` is used [51].

Shared Frailty Model for OS

As a first step, in order to have a qualitative feedback on the possible presence of a centre effect, a boxplot of the OS, grouped by the different hospitals, is analysed in Figure 3.5.

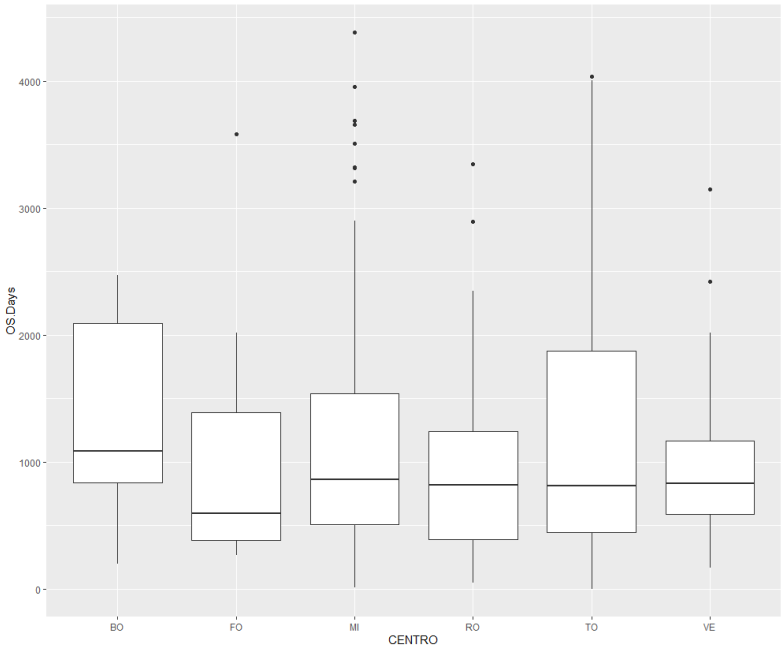


Figure 3.5: Boxplot of OS grouped by Centre

From Figure 3.5, it can be seen that the distributions of OS among the various centres

do not differ significantly between them. To check for heterogeneity the Shared Frailty model is used. In Tables 3.5, 3.6 and 3.7 most important information about the fitted model are reported.

Table 3.5: Coefficient Summary of Shared Frailty model for OS with covariates identified in the best Cox-PH model

	coef	exp(coef)	se(coef)	adj se	z	Pr(>|z|)
AGE	6.13e-02	1.06	1.00e-02	1.01e-02	6.08	< 0.001
CA 19-9	1.93e-05	1.00	1.25e-05	1.25e-05	1.54	0.12
HEPATECTOMY=1	3.89e-01	1.47	2.45e-01	2.45e-01	1.58	0.11
N=1	1.57e+00	4.81	2.91e-01	3.00e-01	5.24	< 0.001
N=x	3.33e-01	1.39	2.76e-01	2.76e-01	1.21	0.23
R status=1	5.57e-01	1.75	2.42e-01	2.46e-01	2.27	0.02
COMPLICANZE=1	8.96e-01	2.45	2.31e-01	2.31e-01	3.88	< 0.001
PATTERN=1	4.07e-01	1.50	3.91e-01	4.08e-01	0.998	0.32
PATTERN=2	1.58e+00	4.86	3.70e-01	3.89e-01	4.07	< 0.001
SATELLITE NODULES1	6.18e-01	1.85	3.66e-01	3.74e-01	1.65	0.10
BILIARY RESECTION=1	1.06e+00	2.90	3.42e-01	3.58e-01	2.98	< 0.001
CORE GLRLM SRHGE	-3.80e-01	0.684	1.40e-01	1.41e-01	-2.70	0.01
MAR CONV HUKurtosis	2.27e-01	1.25	8.26e-02	8.31e-02	2.73	0.01
MAR SHAPE Compacity	1.94e-01	1.21	1.65e-01	1.93e-01	1.00	0.32
MAR GLRLM SRHGE	3.42e-01	1.41	1.61e-01	1.61e-01	2.12	0.03
MAR GLRLM GLNU	5.11e-01	1.67e	9.53e-02	9.91e-02	5.16	< 0.001
MAR GLZLM ZP	-4.42e-01	0.643	1.38e-01	1.39e-01	-3.18	< 0.001
MAR GLZLM LZLGE	-4.57e-01	0.633	2.15e-01	2.22e-01	-2.06	0.04
MAR GLRLM LGRE	3.07e-01	1.36	1.88e-01	1.93e-01	1.59	0.11

Table 3.6: Frailty Summary of Shared Frailty model for OS with covariates identified in the best Cox-PH model

	estimate	lower 95%	upper 95%
Var[Z]	0.350	0.000	1.620
Kendall's tau	0.149	0.000	0.448
Median concordance	0.145	0.000	0.455
E[logZ]	-0.185	-0.997	0.000
Var[logZ]	0.418	0.000	3.473
theta	2.861	0.617	Inf

Table 3.7: Fit Summary of Shared Frailty model for OS with covariates identified in the best Cox-PH model

Commenges-Andersen test p-val	0.654
no-frailty Log-likelihood	-526.406
Log-likelihood	-525.258
LRT p-val	0.0648

From Table 3.7, through the p-value of the Commongen-Andersen test of heterogeneity, which value is 0.654, it can be deduced that there is no statistical evidence to say that the effect of the centre is significant. Consistently with this, comparing the coefficient in Table 3.3 and Table 3.5, it can be noticed that they are very similar.
Since there is no obvious difference between the COX-PH model and the Shared Frailty Model, and the centre effect is not present, we decide to keep the most parsimonious model, namely the Cox-PH model.

Shared Frailty Model for RFS

Also in this case, in order to have a qualitative view of the possible presence of heterogeneity, a boxplot of the RFS, grouped by the different centres, is observed.

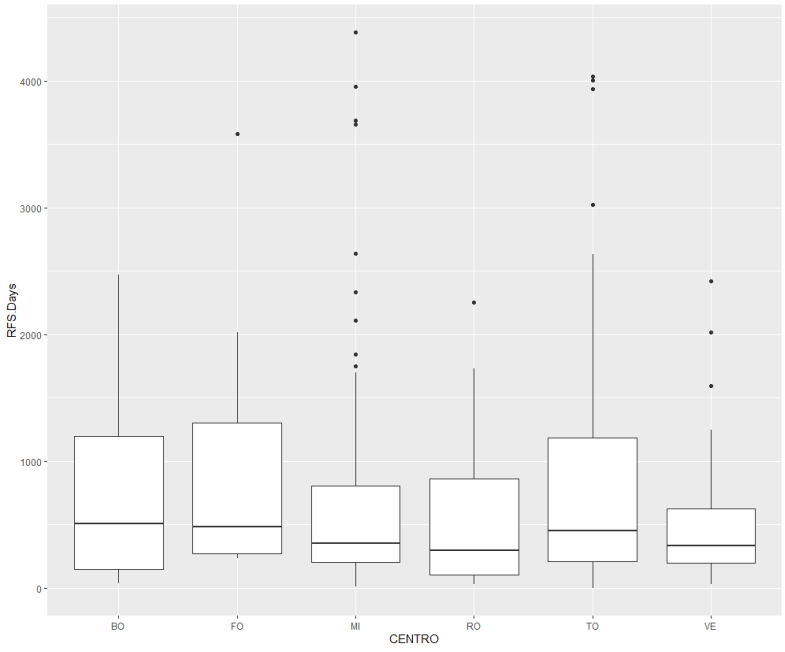


Figure 3.6: Boxplot of RFS grouped by Center

From Figure 3.6 it can be seen that the distributions of RFS among the various hospital do not differ significantly between them. To check for the centre effect the Shared Frailty model is employed. In the Tables In Tables 3.8, 3.9 and 3.10 below most important information about the fitted model are summarized.

Table 3.8: Coefficient Summary of Shared Frailty model for RFS with covariates identified in the best Cox-PH model

	coef	exp(coef)	se(coef)	adj se	z	Pr(>|z|)
CA 19-9	4.28e-05	1.00	1.25e-05	3.41	< 0.001	
PATTERN=1	4.62e-01	1.59	2.11e-01	2.19	0.03	
PATTERN=2	9.91e-01	2.69	2.36e-01	4.20	< 0.001	
N=1	7.11e-01	2.04	2.38e-01	2.99	< 0.001	
N=x	3.30e-01	1.39	2.14e-01	1.54	0.12	
M=1	1.15e+00	3.14	5.15e-01	2.22	0.03	
R status=1	3.01e-01	1.35	1.98e-01	1.52	0.13	
PC SHAPE Sphericity	2.67e-01	1.31	1.11e-01	2.39	0.02	
PC GLRLM LGRE	2.87e-01	1.33	9.10e-02	3.16	< 0.001	
PC GLZLM LZE	3.25e-01	1.38	1.02e-01	3.20	< 0.001	
PM SHAPE Sphericity	-1.91e-01	0.826e	1.32e-01	-1.45	0.15	
PM GLZLM SZHGE	3.19e-01	1.38e	8.35e-02	3.82	< 0.001	
PM GLZLM LZLGE	-6.28e-01	0.533	2.16e-01	-2.91	< 0.001	
PM GLZLM ZP	-3.21e-01	0.726	1.13e-01	-2.84	< 0.001	
PM GLRLM GLNU	2.84e-01	1.33	9.33e-02	3.05	< 0.001	
PM GLCMCorrelation	-2.16e-01	0.806	9.73e-02	-2.22 0.03		

Table 3.9: Frailty Summary of Shared Frailty model for RFS with covariates identified in the best Cox-PH model

	estimate	lower 95%	upper 95%
Var[Z]	0.0	0.000	0.233
Kendall's tau	0.0	0.000	0.104
Median concordance	0.0	0.000	0.102
E[logZ]	0.0	-0.121	0.000
Var[logZ]	0.0	0.000	0.263
theta	25816.3	4.285	Inf

Table 3.10: Fit Summary of Shared Frailty model for RFS with covariates identified in the best Cox-PH model

Commenges-Andersen test p-val	0.865
no-frailty Log-likelihood	-720.252
Log-likelihood	-720.252
LRT p-val	>0.5

From Table 3.10, through the p-value of the Commongen-Andersen test of heterogeneity, which value is 0.865, it can be deduced that there is no statistical evidence to say that the effect of the centre is significant. Coherently, comparing the coefficient in Table 3.8 and Table 3.4, it can be observed that they are basically the same.
Since there is no obvious difference between the COX-PH model and the Shared Frailty Model, and the centre effect is not present, it is decided to keep the most parsimonious model, the COX Proportional Hazard.

3.2.4. Summary of Survival Analysis

First of all, in Appendix D, it can be observed that the variables selected in the different models for OS and RFS are consistent with each other. This reinforces the correctness of the feature selection performed by the Stepwise Algorithm. In addition, the best model is the one with CLINICAL POSTOPERATIVE+PORTAL(CORE+MARGIN) covariates in both cases.
In order to have a unified view of the results, Table 3.11 was produced with the C-index values for each model.

Table 3.11: Summary of Cox-PH models results with C-Index

	OS	**RFS**
CLINICAL PREOP	0.682	0.66
PREOP+CORE	0.713	0.668
PREOP+CORE+MARGIN	0.752	0.71
CLINICAL POSTOP	0.755	0.677
POSTOP+CORE	0.766	0.716
POSTOP + CORE +MARGIN	0.797	0.733

From Table 3.11 different conclusions can be drawn. As a first step, it can be seen that, by adding radiomic covariates to clinical ones, the C-index value increases in both cases. Therefore, radiomics gives added predictive value in Survival Analysis.
Secondly, it can be noticed that, by adding MARGIN covariates to clinical and CORE, the C-index rises. Hence, the features belonging to MARGIN are important for predictive

purposes, as they give added value compared to CORE variables.
Lastly, C-Indexes in PREOP+CORE+MARGIN and POSTOP+CORE cases are comparable. This not only testifies to the additional value of the MARGIN, as already observed in the previous case; but shows how preoperative information integrated with radiomics can achieve similar performance to the postoperative case. Thus, an adequate non-invasive preoperative assessment is possible taking into consideration both tumour and margin radiomics information.
Therefore, radiomics not only of the tumour but also of the peritumoral area, is very informative about the outcome and it can increase the prognostic impact. At the end, a final model with a sufficiently high C-Index is obtained for both OS and RFS, in which the covariates predicting the outcome are identified.
Subsequently, it is observed that the centre effect is not evident: the variability of outcomes between centre is not high and the estimated one is not significant. This could be because an outcome that is long-term is analysed and it is probably more related to the characteristics of the disease than to the protocol of the multicentre study. Since the estimates between the models with and without frailty are similar, for reasons of simplicity and parsimony the Cox-PH models are used to describe OS and RFS.

4 | Multiview Dimensionality Reduction

In Chapters 2 and 3 we have considered radiomic data belonging to core and margin of the Portal phase only within the presented models. Doing this way had a twofold motivation:

1. Portal phase is the reference phase.
2. Our aim was to assess the presence of the added value of including the set of margin information, using only one CT phase as a benchmark.

In this Chapter we consider, for both core and margin, all three radiomic phases of CT scan, i.e. Arterial, Portal and Late, for both pathology data classification and survival analysis. The aim is to assess whether each of the phases can enrich the prediction, understanding whether the phases carry out the same information or whether each is decisive for modelling the outcomes.
In fact, Portal, Arterial and Late phases of the CT scan can be considered as a multiple view representations of the tumor and its surrounding area, as explained in Section 1.2.3, so that *Multiview Learning* techniques are natural candidate to properly account for these kinds of data. Since considering all radiomics involves a very large number of features, in this Chapter we consider Multiview Dimensionality Reduction techniques to decrease the number of input covariates in the models, considering the multiview aspect of the data. The two techniques that we analyse within this thesis are Multiview Canonical Correlation Analysis (MCCA) and Kernel Multiview Canonical Correlation Analysis (KMCCA), respectively described in Sections 4.1 and 4.2.

4.1. Multiview Canonical Correlation Analysis

Multiview Canonical Correlation Analysis (MCCA) is the extension of Canonical Correlation Analysis (CCA), that allows the simultaneous consideration of more than two sets of random variables [52]. MCCA is an unsupervised method that searches for a lower-dimensional common subspace to represent multiview data [20], finding a set of directions

(one per view) which maximize the average correlation, computed for each pair of views [53].

Let $\mathbf{X} \in \mathbb{R}^P$ be a random vector, where $\mathbf{X}$ is composed of m subset of random variables referred to a view, so that $\mathbf{X} = [\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(m)}]$, and $\mathbf{X}^{(j)'} = [X_1^{(j)}, X_2^{(j)}, \ldots, X_{p_j}^{(j)}]$ indicates the j-th set of variable. Each set of variables has got p_j covariates, with $p_1 \leq \cdots \leq p_m$ and $P = \sum_{j=0}^{m} p_j$.

Given $w = [w^{(1)}, \ldots, w^{(m)}] \in \mathbb{R}^P$, we derive m variables $\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_m$, where Z_j identified as:

$$\mathbf{Z}_j := \sum_{i=1}^{pj} \mathbf{X}_i^{(j)} w_i^{(j)} = \mathbf{X}^{(j)T} \cdot w^{(j)} \tag{4.1}$$

To the problem formulation, we need to define the correlation coefficient between $\mathbf{Z}_i$ and $\mathbf{Z}_j$. It can be expressed as:

$$\rho(\mathbf{Z}_i, \mathbf{Z}_j) = \frac{w^{(i)T} Cov(\mathbf{X}^{(i)}, \mathbf{X}^{(j)}) w^{(j)}}{\sqrt{w^{(i)T} Cov(\mathbf{X}^{(i)}, \mathbf{X}^{(i)}) w^{(i)}} \sqrt{w^{(j)T} Cov(\mathbf{X}^{(j)}, \mathbf{X}^{(j)}) w^{(j)}}} \tag{4.2}$$

The initial formulation of the MCCA problem can be stated as finding the set of vectors $w^{(i)}$ which maximize the sum of Correlations (SUMCOR), namely:

$$\sum_{i=1}^{m} \sum_{j=i+1}^{m} \rho(\mathbf{Z}_i, \mathbf{Z}_j) \tag{4.3}$$

If we expand the SUMCOR expression, we get the following maximization problem:

$$\max_{w \in \mathbb{R}^P} \sum_{i=1}^{m} \sum_{j=i+1}^{m} \frac{w^{(i)T} Cov(\mathbf{X}^{(i)}, \mathbf{X}^{(j)}) w^{(j)}}{\sqrt{w^{(i)T} Cov(\mathbf{X}^{(i)}, \mathbf{X}^{(i)}) w^{(i)}} \sqrt{w^{(j)T} Cov(\mathbf{X}^{(j)}, \mathbf{X}^{(j)}) w^{(j)}}} \tag{4.4}$$

From this formulation of the problem, it can be seen that the solution is invariant to block scaling, so that only the direction of the solutions matters. Therefore, if $(w^{(1)}, \ldots, w^{(m)})$ is a solution then $(\alpha_i w^{(1)}, \ldots, \alpha_m w^{(m)})$ is also a solution for $\alpha_i > 0$. For this reason, the constraints $w^{(i)T} Cov(\mathbf{X}^{(i)}, \mathbf{X}^{(i)}) w^{(i)} = 1$ is imposed to the problem. This yield to the following equivalent constrained problem:

$$\begin{aligned} &\underset{w \in \mathbb{R}^P}{\text{maximize}} \sum_{i=1}^{m} \sum_{j=i+1}^{m} w^{(i)T} Cov(\mathbf{X}^{(i)}, \mathbf{X}^{(j)}) w^{(j)} \\ &\text{subject to } w^{(i)T} Cov(\mathbf{X}^{(i)}, \mathbf{X}^{(i)}) w^{(i)} = 1 \quad \forall i = 1, \ldots, m. \end{aligned} \tag{4.5}$$

The optimal solution is not affected even if the objective is manipulated by multiplying

it by 2 and adding a constant m.
Joint with the fact that the equalities $w^{(i)T}Cov(\mathbf{X}^{(i)}, \mathbf{X}^{j)})w^{(j)} = w^{(j)T}Cov(\mathbf{X}^{(j)}, \mathbf{X}^{(i)})w^{(i)}$ and $w^{(i)T}Cov(\mathbf{X}^{(i)}, \mathbf{X}^{(i)})w^{(i)} = 1$ hold, we obtain:

$$\begin{aligned} &\underset{w\in\mathbb{R}^P}{\text{maximize}} \sum_{i=1}^{m}\sum_{j=1}^{m} w^{(i)T}Cov(\mathbf{X}^{(i)}, \mathbf{X}^{(j)})w^{(j)} \\ &\text{subject to } w^{(i)T}Cov(\mathbf{X}^{(i)}, \mathbf{X}^{(i)})w^{(i)} = 1 \quad \forall i = 1, \ldots, m. \end{aligned} \tag{4.6}$$

In this way we get the final formulation of the problem in which the objective function is transformed into a quadratic form, so that the problem can be solved [53].

We will use MCCA in Section 4.3 to perform Multiview Dimensionality Reduction on both core and margin of all three radiomic phases, in order to be able to reduce the dimensionality of the radiomics covariates considering the multiview aspect of the data. To implement the method, the `mvlearn` Python Package [54] has been used.

4.2. Kernel Multiview Canonical Correlation Analysis

Kernel Multiview Canonical Correlation Analysis (KMCCA) is the extension of MCCA to use kernels. In fact, the traditional MCCA aims to find useful projections of covariates, computing a weighted sum, but may not extract useful descriptor of the data because of its linearity. KMCCA allows to first project the data onto a higher dimensional feature space before performing MCCA in the new feature space:

$$\Phi : \mathbf{x} = (x_1, \ldots, x_m) \mapsto \Phi(\mathbf{x}) = (z_1 \ldots, z_N), (m << N) \tag{4.7}$$

Kernel function is defined as scalar product between the feature vectors of two data samples:

$$k(x, x') = \Phi(x)^T\Phi(x') \tag{4.8}$$

Kernels can be interpreted as a similarity measure between x and x' that computes inner products in the higher dimensional feature space, with a method known as Kernel trick. With Kernel trick it is not required to explicitly compute the feature mapping, because it uses the similarities between each pair of samples.

A commonly used kernel is the Gaussian Kernel that is defined as:

$$k(x, x') = \exp\left(- \frac{\|x - x'\|^2}{2\sigma^2} \right) \tag{4.9}$$

In Equation 4.6 we focused only on manipulating the covariance matrices, but to be able to apply the Kernel Trick the expression of their estimation based on finite samples is required.
Let $\mathbf{X} \in \mathbb{R}^{nxp}$ represent a sample of n observation of $\mathbf{X}$. The empirical covariance of $\mathbf{X}$, computed using regularization techniques, is:

$$\overline{Cov(\mathbf{X})} = (1 - \mathcal{K})\frac{1}{n-1}\mathbf{X}\mathbf{X}^T +_{\mathcal{K}} \mathbf{I}_P \tag{4.10}$$

where $\mathcal{K} \in [0, 1]$. The higher the value of $\mathcal{K}$, the better is the numerical stability, but less optimal the solutions, since the problem is different from the originally posed. Substituting the expression of covariance in Equation 4.5, the problem becomes:

$$\begin{aligned} &\max_{w \in \mathbb{R}^P} \frac{1}{n-1} \sum_{i=1}^{m} \sum_{j=i+1}^{m} w^{(i)T} \mathbf{X}^{(i)} \mathbf{X}^{(j)T} w^{(j)} \\ &\text{subject to } w^{(i)T} \left(\frac{1 - \mathcal{K}}{n-1} \mathbf{X}^{(i)} \mathbf{X}^{(i)T} +_{\mathcal{K}} \mathbf{I}_P \right) w^{(i)} = 1 \quad \forall i = 1, \ldots, m. \end{aligned} \tag{4.11}$$

Let $y^{(i)} \in \mathbb{R}^n$ be a block of the dual variable $y \in \mathbb{R}^{m \cdot n}$, so that we can express each component $w^{(i)}$ in terms of columns of $\mathbf{X}^{(i)}$. We obtain:

$$w^{(i)} = X^{(i)} y^{(i)} \tag{4.12}$$

Letting $\mathbf{K}^{(i)} = \mathbf{X}^{(i)T} \mathbf{X}^{(i)} \in \mathbb{R}^{nxn}$ be the Gram matrix, we express the problem in terms of the dual variables:

$$\begin{aligned} &\max_{y \in \mathbb{R}^{m \cdot n}} \frac{1}{n-1} \sum_{i=1}^{m} \sum_{j=i+1}^{m} y^{(i)T} \mathbf{K}^{(i)} \mathbf{K}^{(j)T} y^{(j)} \\ &\text{subject to } y^{(i)T} \left(\frac{1 - \mathcal{K}}{n-1} \mathbf{K}^{(i)} \mathbf{K}^{(i)T} +_{\mathcal{K}} \mathbf{K}^{(i)} \right) y^{(i)} = 1 \quad \forall i = 1, \ldots, m. \end{aligned} \tag{4.13}$$

Typically, $\mathbf{K}(i)$ matrices are ill conditioned or even singular. This problem is addressed

by introducing the following quantity:

$$\widetilde{\mathbf{K}^{(i)}} := \left(\sqrt{\frac{1-\mathcal{K}}{n-1}}\mathbf{K}^{(i)} + \frac{\mathcal{K}}{2}\sqrt{\frac{n-1}{1-\mathcal{K}}}\mathbf{I}_n\right) \tag{4.14}$$

that makes possible the approximation:

$$\overline{Cov(\mathbf{X}^{(i)})}_{\mathcal{K}} = \frac{1-\mathcal{K}}{n-1}\mathbf{K}^{(i)}\mathbf{K}^{(i)T} +_{\mathcal{K}} \mathbf{K}^{(i)} \approx \widetilde{\mathbf{K}^{(i)}}\widetilde{\mathbf{K}^{(i)T}} \tag{4.15}$$

With this approximation, that is invertible and in a factorized form the optimization problem becomes:

$$\begin{aligned} &\max_{y\in\mathbb{R}^{m\cdot n}} \frac{1}{n-1}\sum_{i=1}^{m}\sum_{j=i+1}^{m} y^{(i)T}\mathbf{K}^{(i)}\mathbf{K}^{(j)T}y^{(j)} \\ &\text{subject to } y^{(i)T}\widetilde{\mathbf{K}^{(i)}}\widetilde{\mathbf{K}^{(i)}}^{T}y^{(i)} = 1 \quad \forall i = 1,\dots,m. \end{aligned} \tag{4.16}$$

Expressing the problem in terms of Gram matrices makes us able to use Kernel methods. In this way we get the final formulation in which the objective function is transformed into a quadratic form and solving this problem we obtain a one-dimensional representation for each view. However, one component is not sufficient to capture all the information present in the data, so that higher dimensional subspaces are needed. After computing the first set of canonical vectors we proceed to computing the next set of components. The next set should be almost as highly correlated as the first one, but essentially different from the first one. To obtain the desired number of components we impose additional constraints for every view.

Formally, letting $Y = [y_1,\dots,y_k] \in \mathbb{R}^{m\cdot n\times k}$ represent the k sets of canonical vectors where for each view $Y^{(\ell)T}\overline{K_\ell^2}Y^{(\ell)} = I_k \ \forall \ell = 1,\dots,m$, the formulation of the problem for k sets of canonical components is:

$$\begin{aligned} &\max_{y\in\mathbb{R}^{m\cdot n}} \frac{1}{n-1}\sum_{i=1}^{m}\sum_{j=i+1}^{m} y^{(i)T}\mathbf{K}^{(i)}\mathbf{K}^{(j)T}y^{(j)} \\ &\text{subject to } y^{(i)T}\widetilde{\mathbf{K}^{(i)}}\widetilde{\mathbf{K}^{(i)}}^{T}y^{(i)} = 1 \quad \forall i = 1,\dots,m. \\ &\qquad\qquad Y^{(i)T}\widetilde{\mathbf{K}^{(i)}}\widetilde{\mathbf{K}^{(i)}}^{T}y^{(i)} = 1 \quad \forall i = 1,\dots,m. \end{aligned} \tag{4.17}$$

Solving the problem in 4.17, we are able to compute the desired number of canonical components.

We will use KMCCA in Section 4.3 to perform Multiview Dimensionality Reduction on

both core and margin of all three radiomic phases, trying to understand if using Kernels we are able to find a more suitable mapping to extract information. To implement the method, the `mvlearn` Python Package [54] has been used.

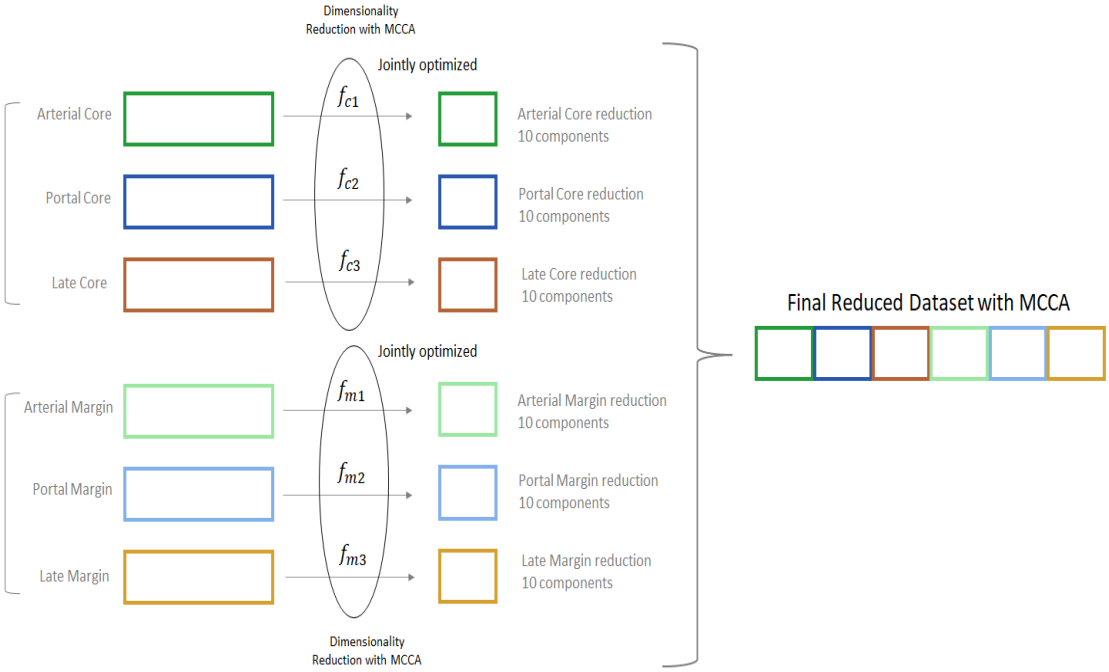
4.3. Results of Multiview Learning

In this Section we illustrate the results of analysing clinical and radiomic data concerning all three phases of CT scan using Multiview techniques. In order to study the effect of considering all radiomics information and multiview modelling on the prediction performances, four settings of different covariates are considered:

- **BASELINE** - Only Portal Covariates: this case is the baseline adopted to see if it is worth considering the three radiomic phases, instead of just the Portal one.
- **ALL** - All Radiomic covariates (Portal + Arterial + Late) simply concatenated: in this case the Multiview aspect of the data is not considered, as the dimensionality reduction is not optimized considering all the views jointly. This setting is adopted in order to be able to understand the difference made by using a Multiview approach to model the problem or a simpler one.
- **MCCA**: in this case the results of MCCA method applied on both core and margin of all three radiomic phases are analysed. In performing MCCA, we consider core and margin separately, because they are two different regions that provide different insights. Hence, the process is repeated twice, once for the core and once for the margin. In each of the two areas, there are three views related to the three phases of CT scan. We reduce the dimensionality of each view considering 10 components per view, taking into account all phases simultaneously. At the end, the results of the reduction for core and margin are concatenated. In this way, for radiomics, a dataset of 60 features, 10 for each of the 3 views for both core and margin, is produced. The process is schematised in Figure 4.1.
- **KMCCA**: in this case the results of KMCCA applied on both core and margin of all three radiomic phases are analysed. The procedure by which the dataset has been reduced is the same as the one employed for MCCA. The difference in this case is that a Gaussian Kernel used. The expression of the Kernel is:

$$k(x, x') = \exp(-\gamma \|x - x'\|) \text{with} \gamma = \frac{1}{\text{n° of features}}$$

Figure 4.1: Schematisation of MCCA process performed separately on both core and margin of all the three radiomic phases



For each of these cases, always jointly considering clinical information, the results and performances of both Classification and Survival Analysis are analysed. In Section 4.3.1, results of Classification are reported. Classification is performed employing Logistic Regression and Mixed Effects Models, jointly with Backward Selection as a feature selection technique. In Section 4.3.2 results of Survival Analysis are illustrated. Survival Analysis is carried out using Cox-PH model with Stepwise Algorithm of variable selection. Since in Chapter 3 we have deduced that the grouping factor present in the data is not relevant to the analysis of the survival response, Shared Frailty models have not been applied. Considering all radiomics, the number of patients, due to missing values, drops to a total of 190 individuals, so that all this following analysis are carried out with 190 samples.

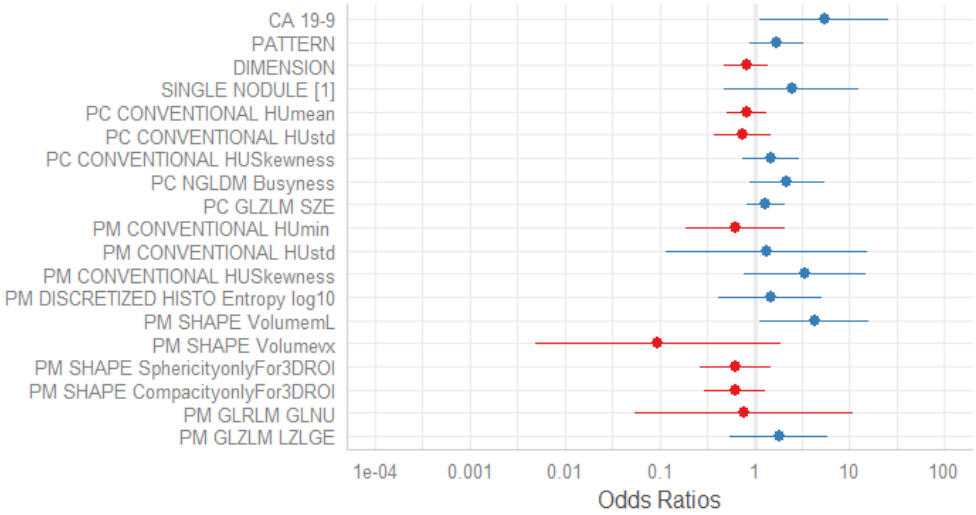
4.3.1. Classification with Multiview Dimensionality Reduction

In this section the result for MVI and Grading outcomes are illustrated. For sake of simplicity, only the MEMs results are reported, illustrating odds ratios with 95% CI and performances.

Mixed Effects Model for MVI with Only Portal covariates - BASELINE case

These are the results of the MEM for MVI, with the features identified in Logistic Regression with Backward Selection on Portal phase covariates only. The results of the fixed effect are reported in Figure 4.2.

Figure 4.2: Odds ratios with 95% CI obtained applying MEMs for MVI with Portal phase features only



The only feature that is significant in the model is CA 19-9. Its odds ratio values indicates that people with larger value of CA 19-9 have higher risk to present MVI.

Performances are summarized in Table 4.1.

Table 4.1: Performances of MVI MEM with Portal Features only

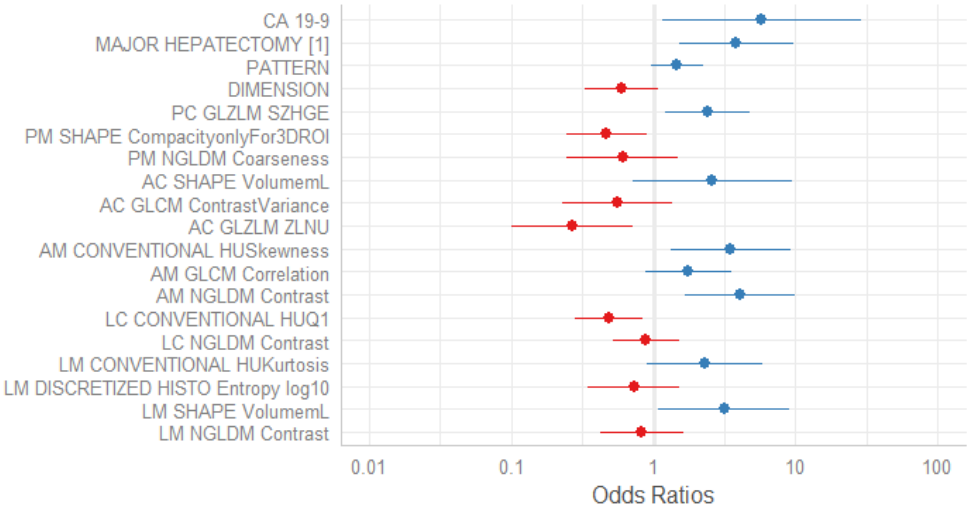
Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.784	0.726	0.27	0.663	0.068
SPECIFICITY	0.785	0.647	0.341	0.652	0.11
SENSITIVITY	0.784	0.767	0.339	0.684	0.094
PRECISION	0.792	0.695	0.347	0.639	0.122
PR AUC	0.867	0.786	0.283	0.744	0.08
ROC AUC	0.863	0.726	0.315	0.729	0.071

These performances are considered as baseline to understand if including all radiomic features could lead to an improvement in predictive ability of the model.

Mixed Effects Model for MVI with all radiomics concatenated - ALL case

These are the results of the MEM for MVI, with the features identified in Logistic Regression with Backward Selection applied on all radiomics features concatenated. The results of the fixed effect are reported in Figure 4.3.

Figure 4.3: Odds ratios with 95% CI obtained applying MEMs for MVI with all radiomic features concatenated



Regarding Clinical features, CA 19-9 and MAJOR HEPATECTOMY are significant. It can be deduced that patients with larger values of CA 19-9, that have undergone major hepatectomy have higher risk to present MVI. Regarding Radiomics, the model selects variable both of core and margin, that belong to every of the three different phases. This suggests that it is important to consider all the three phases of CT scan, because they contain some feature that are informative in predicting the presence of MVI.

Performances are summarized in Table 4.2. All values of performances increase both in training and validation, with respect to the case in which only portal phase covariates are included. Therefore, we can guess that including all radiomics, instead of only Portal phase, improves the prognostic impact of the model, predicting the presence of MVI more accurately.

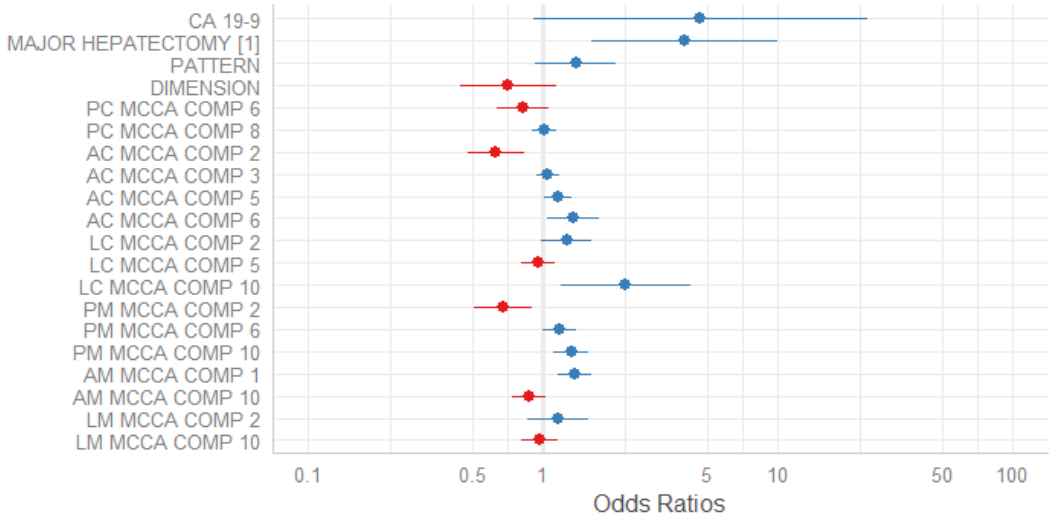
Table 4.2: Performances of MVI MEM with all radiomic features concatenated

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.805	0.762	0.242	0.738	0.056
SPECIFICITY	0.8	0.679	0.315	0.726	0.089
SENSITIVITY	0.811	0.806	0.317	0.75	0.091
PRECISION	0.802	0.741	0.322	0.727	0.096
PR AUC	0.889	0.816	0.292	0.805	0.077
ROC AUC	0.88	0.76	0.327	0.795	0.063

Mixed Effects Model for MVI with MCCA - MCCA case

These are the results of the MEM for MVI, with the features identified in Logistic Regression with Backward selection applied on the result of MCCA dimensionality reduction. The results of the fixed effect are reported in Figure 4.4.

Figure 4.4: Odds ratios with 95% CI obtained applying MEMs for MVI on MCCA dimensionality reduction result



The variables selected by the model are components belonging to all views of core and margin. This indicates that the information conveyed by radiomics is not the same for each view, but all of them bring added value to the prediction.

Performances are summarized in Table 4.3.

Table 4.3: Performances of MVI MEM applied on MCCA dimensionality reduction result

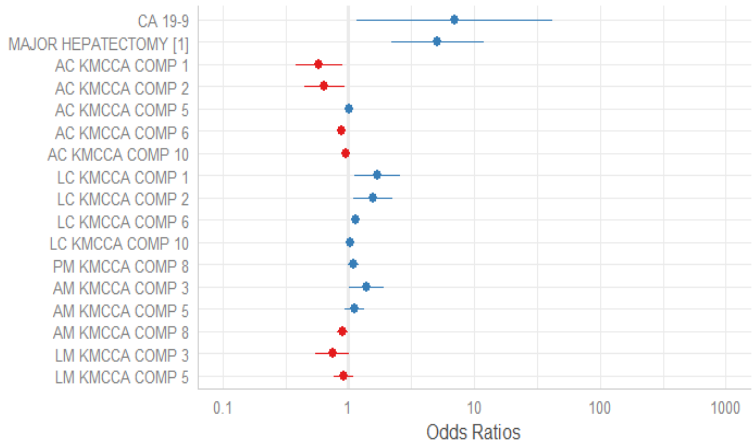
Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.789	0.738	0.259	0.709	0.071
SPECIFICITY	0.77	0.668	0.335	0.696	0.106
SENSITIVITY	0.811	0.759	0.327	0.73	0.095
PRECISION	0.76	0.75	0.325	0.686	0.118
PR AUC	0.884	0.795	0.298	0.775	0.082
ROC AUC	0.882	0.756	0.32	0.765	0.072

All performance metrics value increases in validation, except sensitivity in cross-validation 1, with respect to the case in which only the portal covariates are included. From this fact we can conclude that it is important to consider all radiomics information in the prediction of MVI value, in order to increase the predictive ability of the model. As regards comparison with the performance of radiomics concatenated, in this case, performances in cross-validation 2 are worse for every index.

Mixed Effects Model for MVI with KMCCA - KMCCA case

These are the results of the MEM for MVI, with the features identified in Logistic Regression with Backward selection applied on the result of KMCCA dimensionality reduction. The results of the fixed effect are reported in Figure 4.5.

Figure 4.5: Odds ratios with 95% CI obtained applying MEMs for MVI on KMCCA dimensionality reduction result



The variables selected by the model are components belonging to all views of core and margin. This indicates that the information conveyed by radiomics is not the same for each view.

Performances are summarized Table 4.4.

Table 4.4: Performances of MVI MEM applied on KMCCA dimensionality reduction result

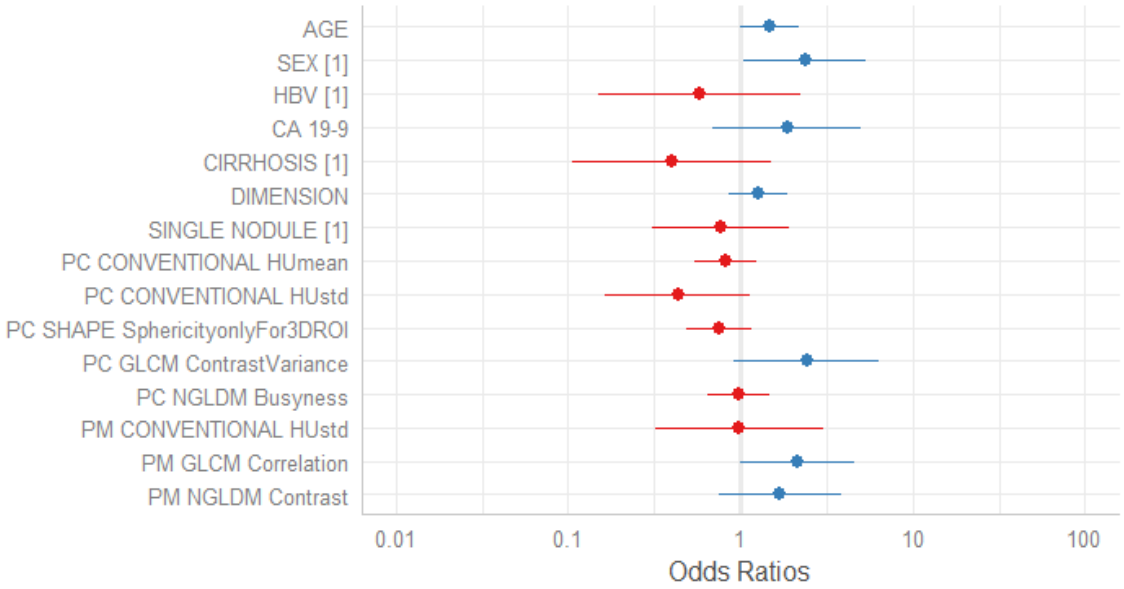
Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.821	0.718	0.217	0.73	0.065
SPECIFICITY	0.8	0.672	0.322	0.722	0.096
SENSITIVITY	0.844	0.752	0.298	0.741	0.094
PRECISION	0.792	0.744	0.291	0.724	0.096
PR AUC	0.888	0.816	0.268	0.794	0.089
ROC AUC	0.881	0.786	0.271	0.78	0.067

All values of performances in cross-validation 2 increase with respect to the case in which only portal phase features are considered. Moreover, they are comparable to the case of all radiomics concatenated. This testifies the importance of considering all CT scan phases in MVI prediction, since all of them contain useful information.

Mixed Effects Model for Grading with only Portal covariates - BASELINE case

These are the results of the MEM for Grading, with the features identified in Logistic Regression with Backward Selection on Portal phase covariates only. The results of the fixed effect are reported in Figure 4.6. The only coviariate that is significant in the model is AGE. Its odd ratio indicates that older people have higher risk to present a IHC of grading equal to 3.

Figure 4.6: Odds ratios with 95% CI obtained applying MEMs for Grading with Portal phase features only



Performances are summarized in Table 4.5.

Table 4.5: Performances of Grading MEM with Portal phase covariates only

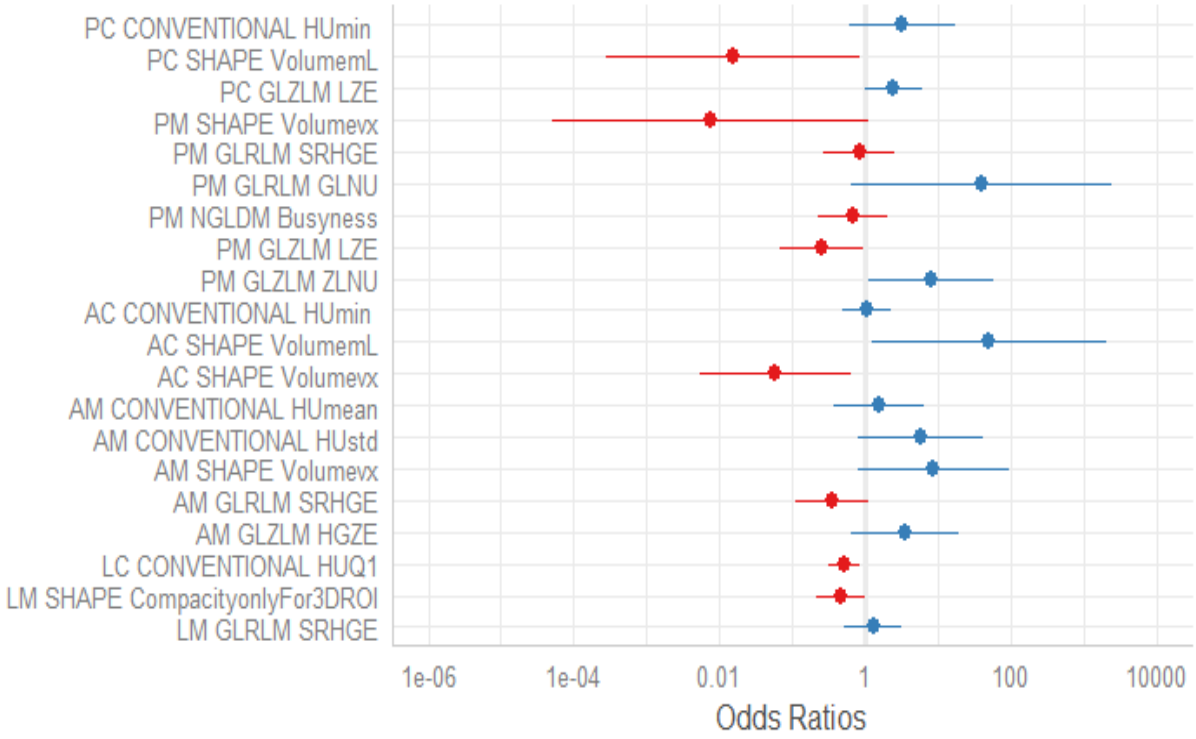
Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.789	0.667	0.224	0.722	0.062
SPECIFICITY	0.791	0.69	0.174	0.766	0.072
SENSITIVITY	0.781	0.611	0.467	0.561	0.175
PRECISION	0.431	0.463	0.393	0.384	0.136
PR AUC	0.684	0.619	0.369	0.502	0.147
ROC AUC	0.802	0.648	0.364	0.671	0.096

These performances are considered as baseline to understand if including all radiomic features could lead to an improvement in predictive ability of the model.

Mixed Effects Model for Grading with all radiomics concatenated - ALL case

These are the results of the MEM for Grading, with the features identified in Logistic Regression with Backward Selection applied on all radiomics features concatenated. The results of the fixed effect are reported in Figure 4.7.

Figure 4.7: Odds ratios with 95% CI obtained applying MEMs for Grading with all radiomic features concatenated



The model selects variables both of core and margin, that belong to every of the three different phases. This suggests that it is important to consider that all the three phases of CT scan, because they contain some feature that are informative in predicting the presence of MVI.

Performances are summarized in Table 4.6.

Table 4.6: Performances of Grading MEM with all Radiomics concatenated

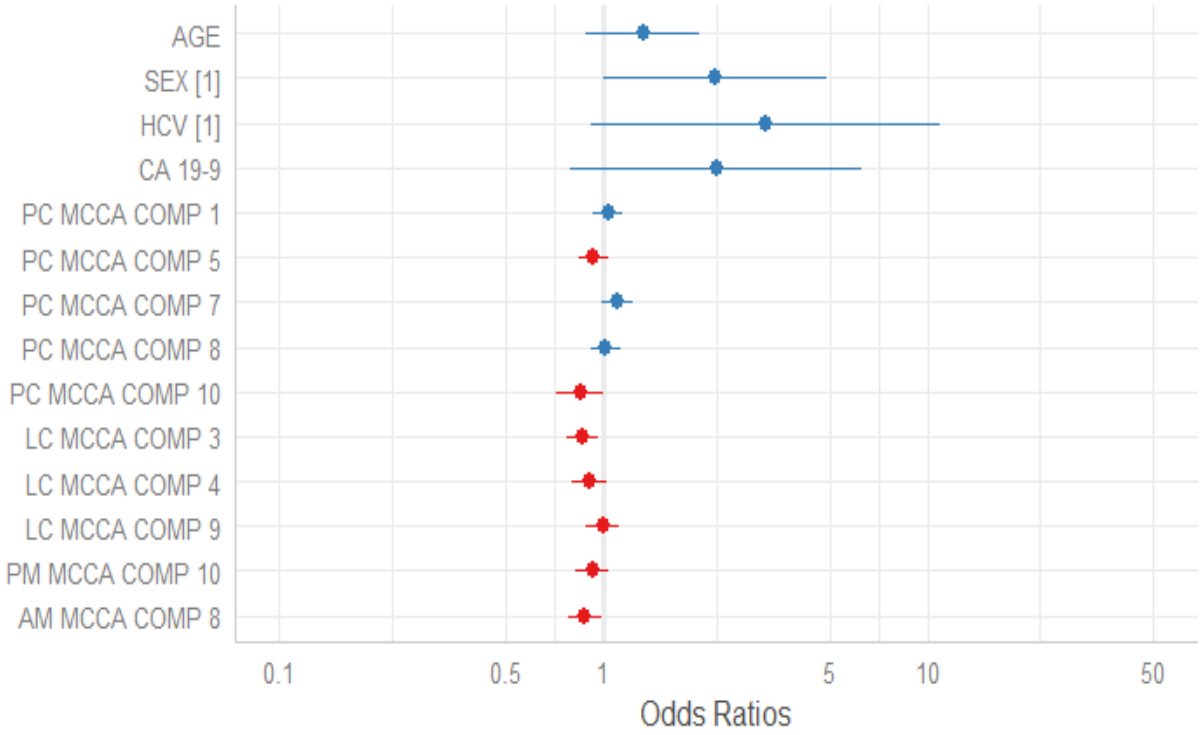
Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.816	0.747	0.196	0.716	0.07
SPECIFICITY	0.817	0.734	0.201	0.758	0.072
SENSITIVITY	0.811	0.722	0.383	0.568	0.205
PRECISION	0.517	0.653	0.39	0.355	0.144
PR AUC	0.754	0.703	0.363	0.537	0.143
ROC AUC	0.823	0.656	0.399	0.656	0.104

Performances in cross-validation 2 are very similar to the one in which only portal phase covariates are considered.

Mixed Effects Model for Grading with MCCA - MCCA case

These are the results of the MEM for Grading, with the features identified in Logistic Regression with Backward selection applied on the result of MCCA dimensionality reduction. The results of the fixed effect are reported in Figure 4.8.

Figure 4.8: Odds ratios with 95% CI obtained applying MEMs for Grading on MCCA dimensionality reduction result



Regarding clinical covariates, the only one that is significant is SEX. The odd ratio associated to SEX indicates that males (coded as SEX=1) have higher risk to present Grading equal to 3. Regarding radiomics, variables selected by the model are components belonging to all views of core and margin. This indicates that the information conveyed by radiomics is not the same for each view, but all of them bring added value to the prediction.

Performances are summarized in Table 4.7.

Table 4.7: Performances of Grading MEM on MCCA dimensionality reduction result

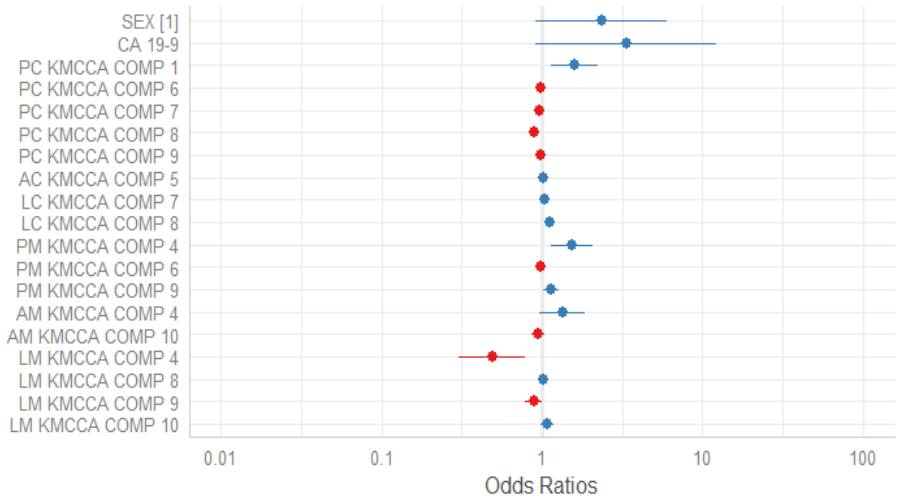
Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.816	0.777	0.221	0.732	0.063
SPECIFICITY	0.817	0.74	0.217	0.774	0.067
SENSITIVITY	0.811	0.727	0.382	0.595	0.192
PRECISION	0.517	0.7	0.382	0.403	0.143
PR AUC	0.738	0.658	0.368	0.57	0.146
ROC AUC	0.816	0.679	0.362	0.692	0.093

All performance metrics values increase in training and cross-validation, with respect to the case in which only the portal covariates are included. From this fact we can conclude that it is important to consider all radiomics information in the prediction of Grading value, in order to increase the predictive ability of the model. As regards comparison with the performance of radiomics concatenated, in this case performances in training are very similar, while in cross-validation 2 they increase. This indicates that the Multiview Dimensionality Reduction is able to mitigate the overfitting.

Mixed Effects Model for Grading with KMCCA - KMCCA case

These are the results of the MEM for Grading, with the features identified in Logistic Regression with Backward selection applied on the result of KMCCA dimensionality reduction. The results of the fixed effect are reported in Figure 4.9.

Figure 4.9: Odds ratios with 95% CI obtained applying MEMs for Grading on KMCCA dimensionality reduction result



The variables selected by the model are components belonging to all views of core and margin. This indicates that the information conveyed by radiomics is not the same for each view.

Performances are summarized Table 4.8

Table 4.8: Performances of Grading MEM with KMCCA

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.863	0.775	0.24	0.769	0.058
SPECIFICITY	0.868	0.772	0.223	0.824	0.066
SENSITIVITY	0.848	0.694	0.411	0.629	0.147
PRECISION	0.672	0.717	0.411	0.579	0.136
PR AUC	0.86	0.778	0.333	0.627	0.147
ROC AUC	0.903	0.817	0.296	0.781	0.077

All values of performances, except sensitivity in training and cross-validation 1, increase. Therefore, we can deduce that, with KMCCA dimensionality reduction, considering all views simultaneously, we are able to find a lower dimensional subspace that it is able to capture additional information about Grading with respect to the one contained in all radiomics simply concatenated.

As it has been done in Section 2.2.3, Permutation tests are used by us to test, for each performance metric, whether there is statistical evidence to say that by including all radiomics information the average of a given metric is higher than in the case where only Portal phase is considered. Formally, we do the following one-sided tests for the means:

1. H_0 : Portal $>$ (=) KMCCA vs H_1 : Portal $\leq$ ($\neq$) KMCCA
2. H_0 : Portal $>$ (=) MCCA vs H_1 : Portal $\leq$ ($\neq$) MCCA
3. H_0 : Portal $>$ (=) All Radiomics vs H_1 : Portal $\leq$ ($\neq$) All Radiomics
4. H_0 : All Radiomics $>$ (=) KMCCA vs H_1 : All Radiomics $\leq$ ($\neq$) KMCCA
5. H_0 : All Radiomics $>$ (=) MCCA vs H_1 : All Radiomics $\leq$ ($\neq$) MCCA

With tests 1, 2 and 3 we want to prove if there evidence to say that considering all radiomics information in the model makes the mean of the performances greater, while with test 4 and 5 we want to understand if employing Multiview Learning techniques can be more powerful for prediction and modelling purposes.
The test statistics used to perform the tests on population X_1 and X_2 are $T = mean(X1) - mean(X2)$ when null hypothesis is $H_0 : mean(X_1) > mean(X_2)$ and $T = |mean(X1) - mean(X2)|$ when null hypothesis is $H_0 : mean(X_1) = mean(X_2)$. Result of the tests are summarized with the p-value, and are carried out for MVI and Grading for every performance metric.
The result for MVI are reported in Figure 4.10

Figure 4.10: Permutation Tests results for MVI MEMs considering multiview aspect of radiomics

ACCURACY	
H0	P-value
Portal>KMCCA	<0.0001
Portal>MCCA	<0.0001
Portal>Radiomica	<0.0001
Radiomica=KMCCA	0.51
Radiomica>MCCA	0.9988

PRECISION	
H0	P-value
Portal>KMCCA	<0.0001
Portal>MCCA	0.0031
Portal>Radiomica	<0.0001
Radiomica=KMCCA	0.805
Radiomica>MCCA	0.9958

SPECIFICITY	
H0	P-value
Portal>KMCCA	<0.0001
Portal>MCCA	0.0032
Portal>Radiomica	<0.0001
Radiomica=KMCCA	0.757
Radiomica>MCCA	0.9839

PR AUC	
H0	P-value
Portal>KMCCA	<0.0001
Portal>MCCA	0.004
Portal>Radiomica	<0.0001
Radiomica=KMCCA	0.337
Radiomica>MCCA	0.995

SENSITIVITY	
H0	P-value
Portal>KMCCA	<0.0001
Portal>MCCA	0.0003
Portal>Radiomica	<0.0001
Radiomica=KMCCA	0.515
Radiomica=MCCA	0.145

ROC AUC	
H0	P-value
Portal>KMCCA	<0.0001
Portal>MCCA	0.0003
Portal>Radiomica	<0.0001
Radiomica=KMCCA	0.09
Radiomica>MCCA	0.9986

The result for Grading are reported in Figure 4.11

Figure 4.11: Permutation Tests results for Grading MEMs considering multiview aspect of radiomics

ACCURACY	
H0	P-value
Portal>KMCCA	<0.0001
Portal=MCCA	0.228
Portal=Radiomica	0.492
Radiomica>KMCCA	<0.0001
Radiomica>MCCA	0.0374

PRECISION	
H0	P-value
Portal>KMCCA	<0.0001
Portal=MCCA	0.325
Portal>Radiomica	0.155
Radiomica>KMCCA	<0.0001
Radiomica>MCCA	0.0089

SPECIFICITY	
H0	P-value
Portal>KMCCA	<0.0001
Portal=MCCA	0.453
Portal=Radiomica	0.418
Radiomica>KMCCA	<0.0001
Radiomica>MCCA	0.0576

PR AUC	
H0	P-value
Portal>KMCCA	<0.0001
Portal>MCCA	0.0008
Portal>Radiomica	0.0491
Radiomica>KMCCA	<0.0001
Radiomica>MCCA	0.0562

SENSITIVITY	
H0	P-value
Portal>KMCCA	0.0012
Portal>MCCA	0.0952
Portal=Radiomica	0.781
Radiomica>KMCCA	0.0089
Radiomica=MCCA	0.349

ROC AUC	
H0	P-value
Portal>KMCCA	<0.0001
Portal>MCCA	0.0575
Portal=Radiomica	0.282
Radiomica>KMCCA	<0.0001
Radiomica>MCCA	0.0058

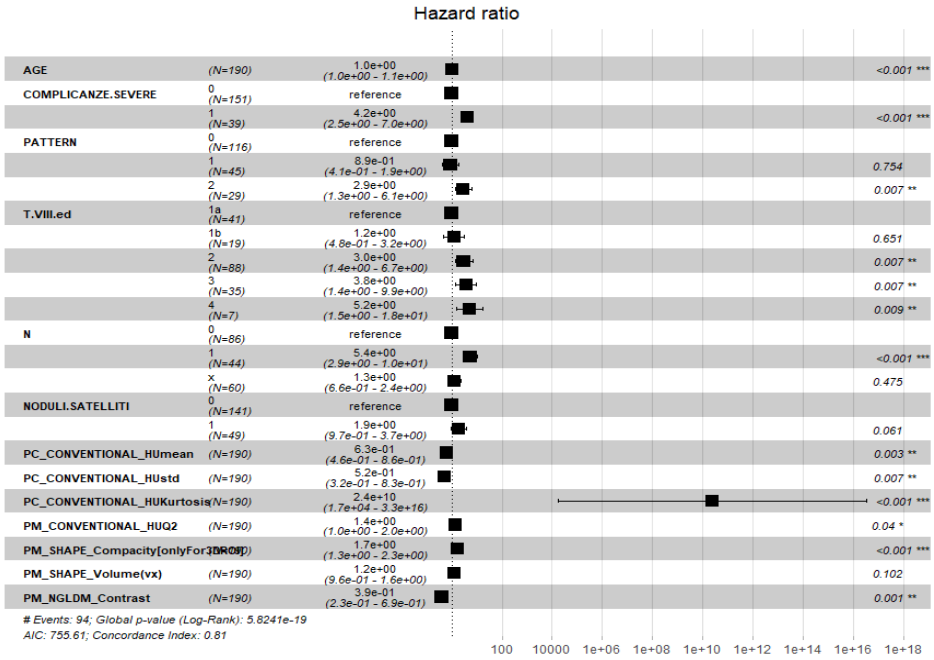
4.3.2. Survival Analysis with Multiview Dimensionality Reduction

In this Section the results of Survival for both OS and RFS are illustrated. For each model, Hazard ratios with 95% CI and C-Index value are provided.

Cox-PH Model for OS with Portal phase features only - BASELINE case

These are the results of the Cox-PH model for OS with Portal phase covariates only. The Hazard ratios are reported in Figure 4.12.

Figure 4.12: Hazard ratios with 95% CI obtained applying Cox-PH model for OS with Portal phase features only

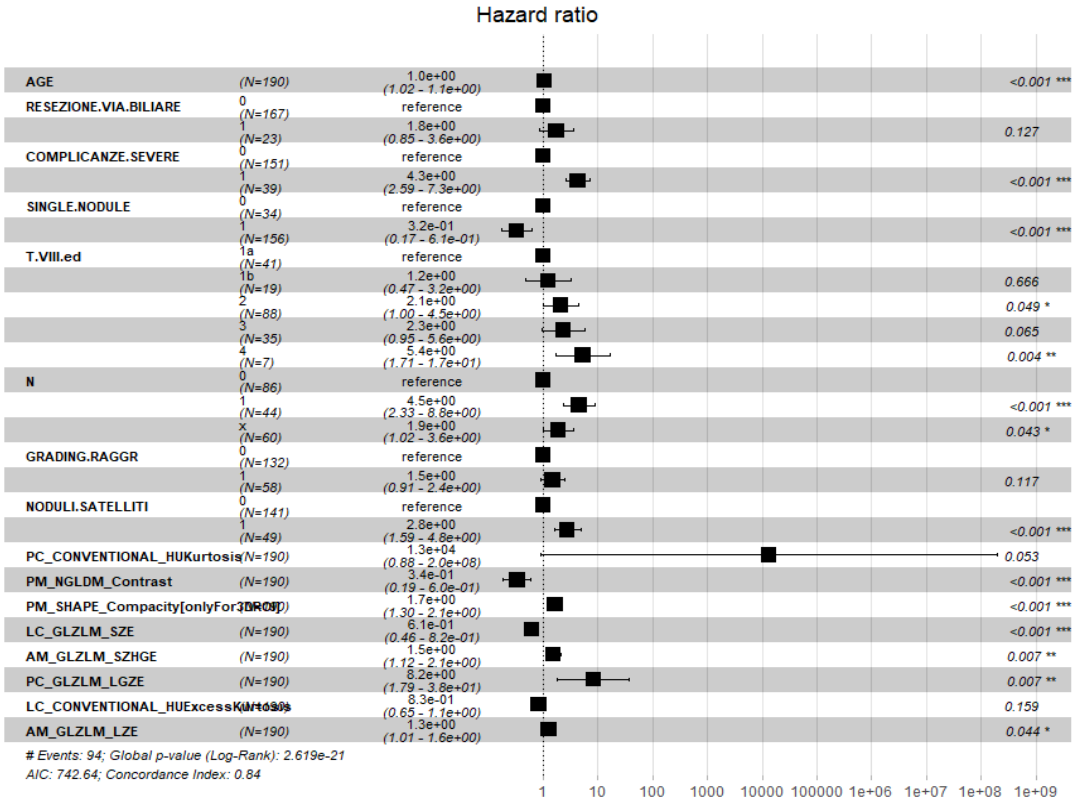


The C-index of the model is 0.807. The value is sufficiently high: the model provides good predictive performances. From Figure 4.12, it can be seen that there is statistical evidence to say that most of the coefficients are significant. Regarding clinical variables, it can be deduced that older age, complications after surgery, nodules (Pattern) and extended tumor (T VII ed and N), increase the risk of death.

Cox-PH Model for OS with all radiomics concatenated - ALL case

These are the results of the Cox-PH model for OS with all radiomics features concatenated. The Hazard ratios are reported in Figure 4.13. The C-index of the model is 0.838. The value is higher than the one provided by the model that considers only Portal phase. This indicates that considering all radiomics we have better predictive performances.
From Figure 4.13, it can be seen that radiomics features that are selected by the model belong to core and margin of all three phases of the CT scan. This reinforces the fact that all radiomic information must be exploited to predict OS more accurately. Regarding clinical variables, it can be deduced that older age, complications after surgery, multiple nodules (SINGLE NODULE and SATELLITES NODULES) and extended tumor (T VII ed and N), increase the risk of death.

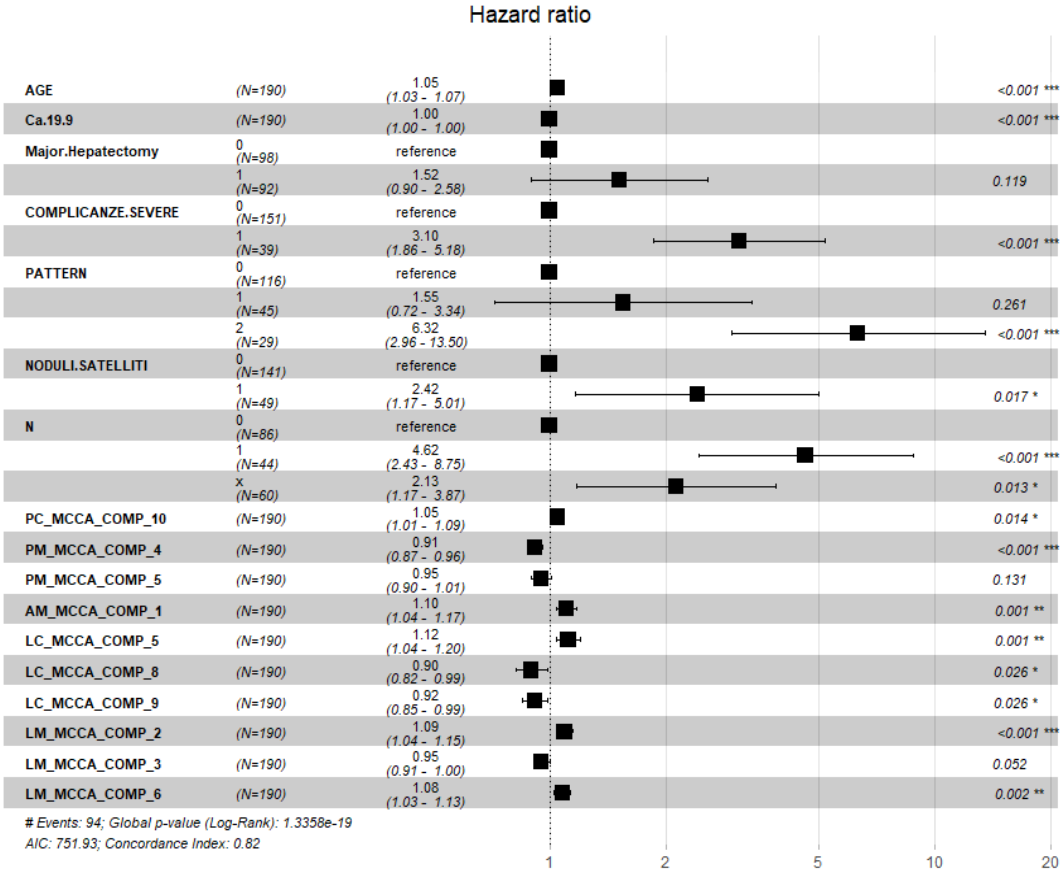
Figure 4.13: Hazard ratios with 95% CI obtained applying Cox-PH model for OS with all radiomics features concatenated



Cox-PH Model for OS with MCCA - MCCA case

These are the results of the Cox-PH model for OS with MCCA dimensionality reduction result. The Hazard ratios are reported in Figure 4.14. The C-index of the model is 0.821. The value is higher than the one provided by the model that considers only Portal phase. This indicates that, considering all radiomics, we have better predictive performances. However, the value of C-Index is less than that with all radiomics concatenated. From Figure 4.14, it can be seen that radiomics features that are selected by the model belong to components extracted from margin of all three phases of the CT scan and from core of Portal and Late. This reinforces the fact that all radiomic information must be exploited to predict OS more accurately. Regarding clinical variables, it can be deduced that older people who have undergone complications after surgery, with multiple nodules (PATTERN and SATELLITES NODULES) and extended tumor(N), have a higher risk of death.

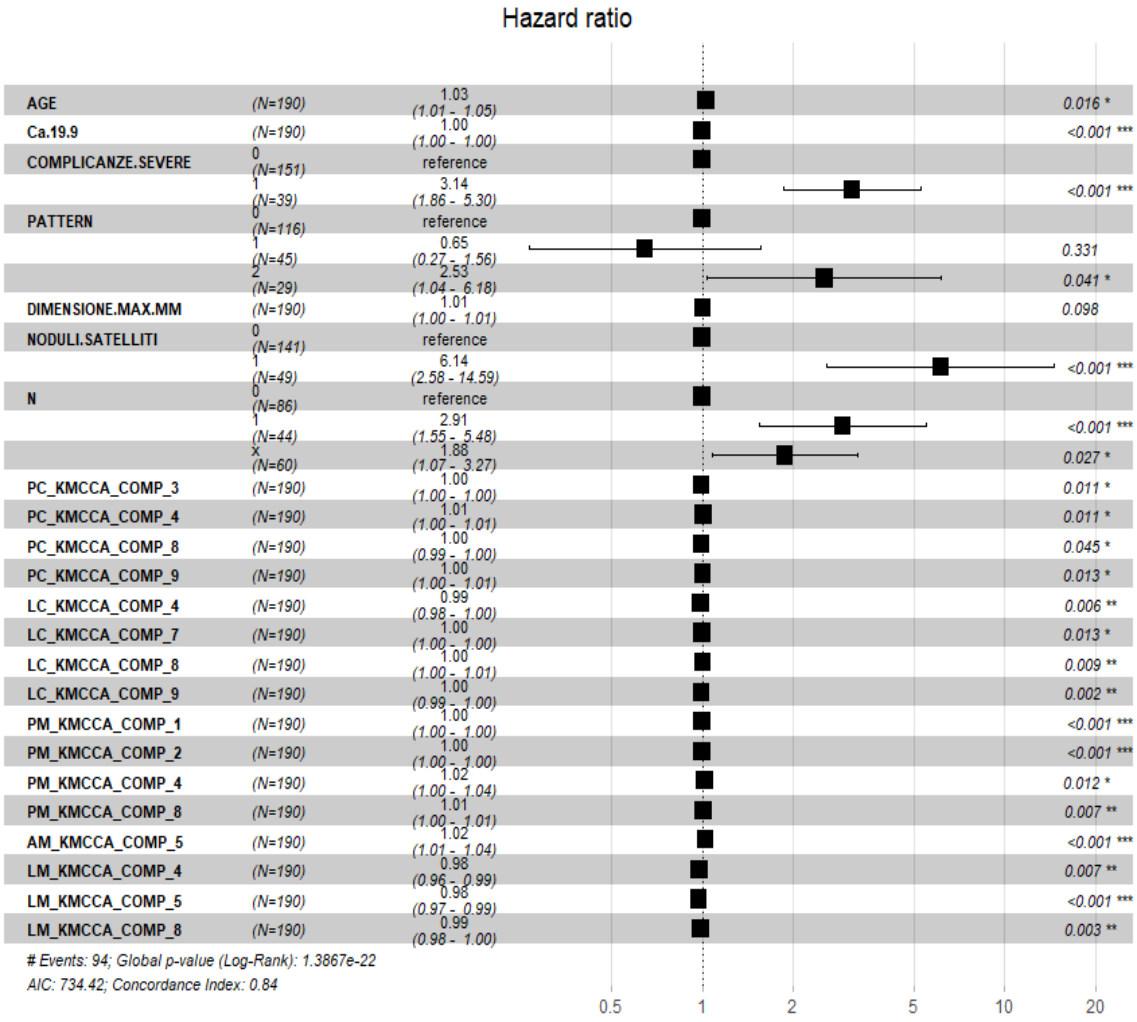
Figure 4.14: Hazard ratios with 95% CI obtained applying Cox-PH model for OS with MCCA dimensionality reduction result



Cox-PH Model for OS with KMCCA - KMCCA case

These are the results of the Cox-PH model for OS with KMCCA dimensionality reduction result. The Hazard ratios are reported in Figure 4.15. The C-index of the model is 0.838. The value is higher than the one provided by the model that considers only Portal phase. This indicates that, considering all radiomics, we have better predictive performances. Moreover, the value of C-Index is equal to that with all radiomics concatenated.
From Figure 4.14, it can be seen that radiomics features that are selected by the model belong to components extracted from margin of all three phases of the CT scan and from core of Portal and Late. This reinforces the fact that all radiomic information must be exploited to predict OS more accurately. Regarding clinical variables, it can be deduced that older people that have undergone complications after surgery with multiple nodules (PATTERN and SATELLITES NODULES) and extened tumor (N), have a higher risk of death.

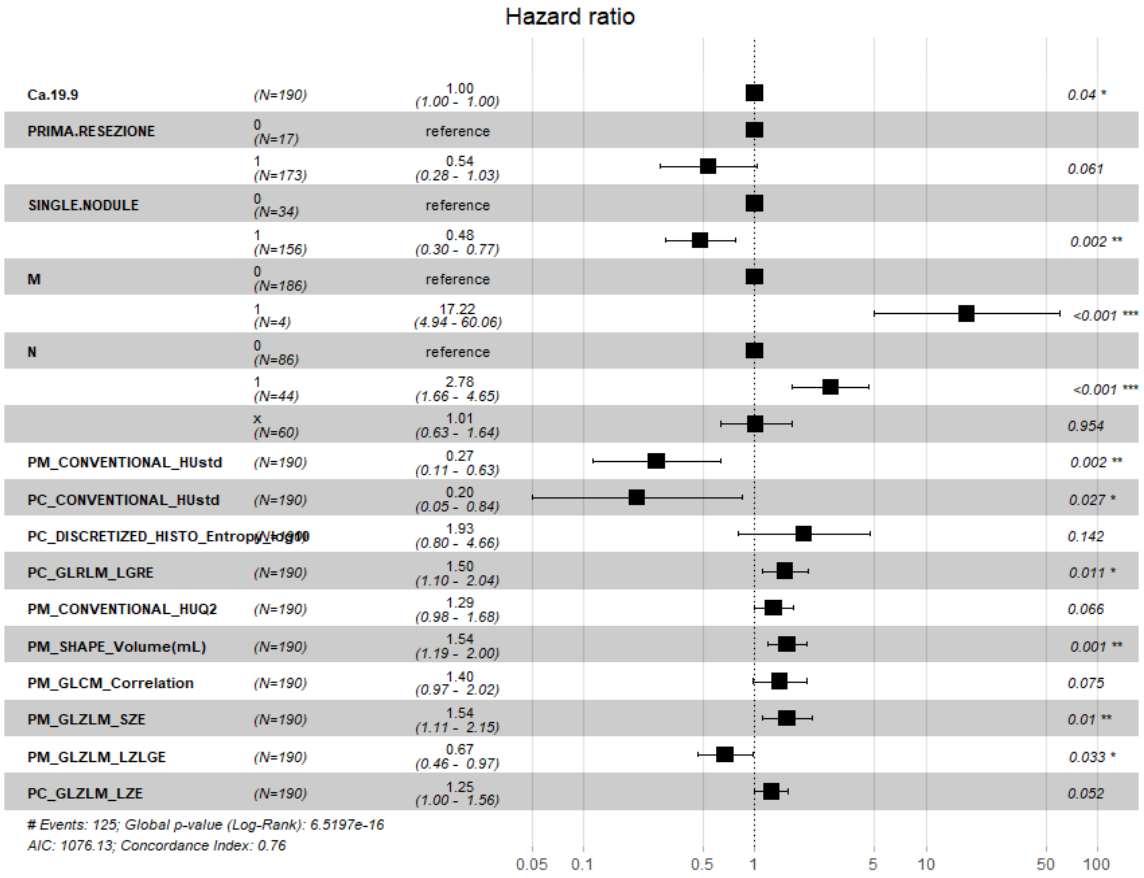
Figure 4.15: Hazard ratios with 95% CI obtained applying Cox-PH model for OS with KMCCA dimensionality reduction result



Cox-PH Model for RFS with Portal phase features only - BASELINE case

These are the results of the Cox-PH model for RFS with Portal phase covariates only. The Hazard ratios are reported in Figure 4.16. The C-index of the model is 0.763. The value is sufficiently high: the model provides good predictive performances.
From Figure 4.16 it can be seen that there is statistical evidence to say that most of the coefficients are significant. Regarding clinical variables, it can be deduced that people with nodules (SINGLE NODULE), metastasis (M) and extended tumor (N), have an increased the risk of recurrence.

Figure 4.16: Hazard ratios with 95% CI obtained applying Cox-PH model for RFS with Portal phase features only

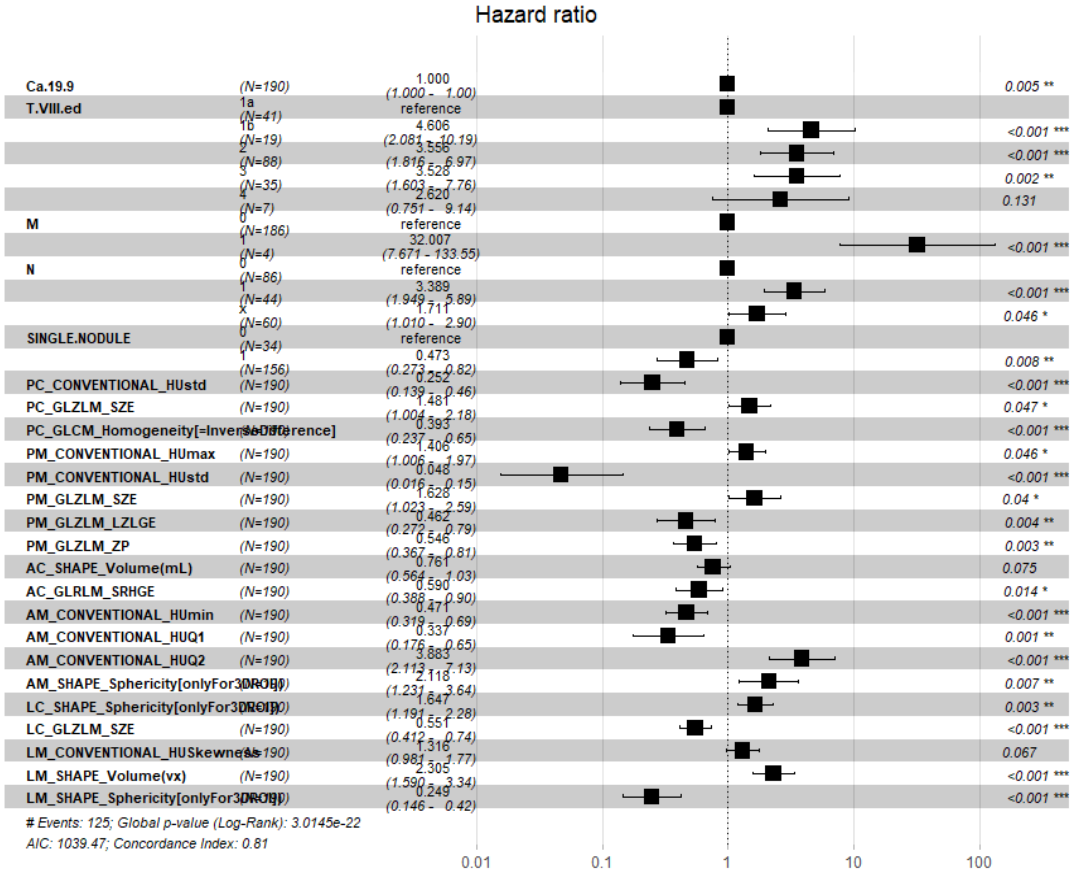


Cox-PH Model for RFS with all radiomics concatenated - ALL case

These are the results of the Cox-PH model for RFS with all radiomics features concatenated. The Hazard ratios are reported in Figure 4.17. The C-index of the model is 0.811. The value is higher than the one provided by the model that considers only Portal phase. This indicates that, considering all radiomics, we have better predictive performances.

From Figure 4.17, it can be seen that radiomics features that are selected by the model belong to core and margin of all three phases of the CT scan. This reinforces the fact that all radiomic information must be exploited to predict OS more accurately. Regarding clinical variables, it can be deduced that people with multiple nodules (SINGLE NODULE), metastasis (M) and extended tumor (T VII ed and N) have higher the risk of recurrence.

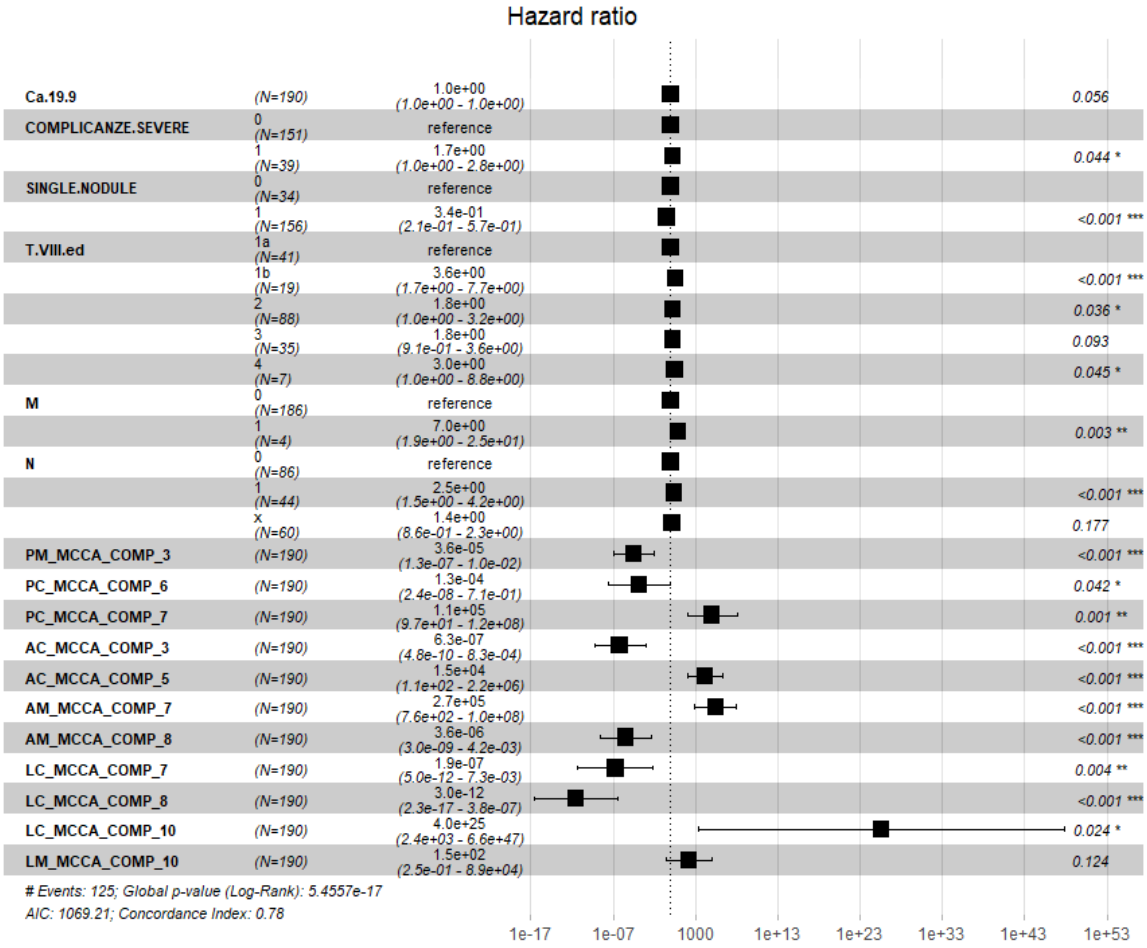
Figure 4.17: Hazard ratios with 95% CI obtained applying Cox-PH model for RFS with all radiomics features concatenated



Cox-PH Model for RFS with MCCA - MCCA case

These are the results of the Cox-PH model for RFS with MCCA dimensionality reduction result. The Hazard ratios are reported in Figure 4.18. The C-index of the model is 0.779. The value is higher than the one provided by the model that considers only Portal phase. This indicates that considering all radiomics we have better predictive performances. However, the value of C-Index is less than that with all radiomics concatenated. From Figure 4.18 it can be seen that radiomics features that are selected by the model belong to components extracted from core and margin of all three phases of the CT scan. This reinforces the fact that all radiomic information must be exploited to predict OS. Regarding clinical variables, it can be deduced that people with multiple nodules (SINGLE NODULE), metastasis (M) and extended tumor (T VII ed and N) have higher the risk of recurrence.

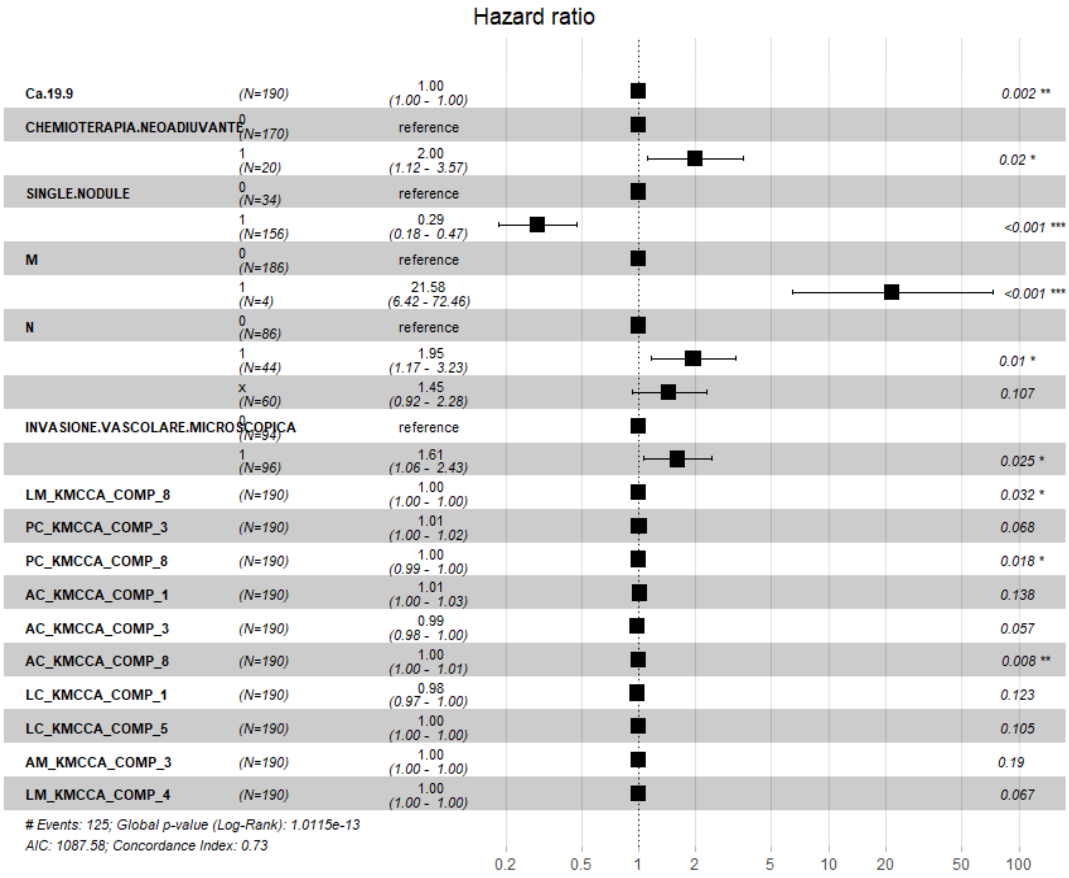
Figure 4.18: Hazard ratios with 95% CI obtained applying Cox-PH model for RFS with MCCA dimensionality reduction result



Cox-PH Model for RFS with KMCCA - KMCCA case

These are the results of the Cox-PH model for RFS with KMCCA dimensionality reduction result. The Hazard ratios are reported in Figure 4.19. The C-index of the model is 0.729. The value is less than the one provided by the model that considers only Portal phase. This indicates that the Multiview Dimensionality Reduction with Gaussian Kernel is not effective in finding a lower dimensional subspace capable to synthesise the RFS. Regarding clinical variables, it can be deduced that people with multiple nodules (SINGLE NODULE), metastasis (M), extended tumor (N) and that present MVI have higher the risk of recurrence. Moreover, people that have undergone chemotherapy before the surgery have lower risk of recurrence.

Figure 4.19: Hazard ratios with 95% CI obtained applying Cox-PH model for RFS with KMCCA dimensionality reduction result



4.3.3. Summary of Multiview Dimensionality Reduction

In this Section we summarize the results about Classification and Survival Analysis, applied considering Multiview Dimensionality Reduction techniques. Regarding Classification, it can be concluded that considering all information provided by the three different phases of radiomics is important in order to improve the predictive ability of the model. Depending on how all radiomic information is considered, whether or not the Multiview aspect of the data is taken into account, there is a greater or lesser increase in the predictive performance of the model. In the case of MVI, using multiview techniques does not exceed the performance of the model in which all radiomics is simply concatenated. In the case of Grading, accounting for the multiview nature of the data using KMCCA, performance improve consistently compared to all other cases.

Regarding Survival Analysis, all values of C-Index are summarized in Table 4.9.

Table 4.9: C-Indexes of Cox-PH models fitted to analyse the benefit of using Multiview Dimensionality Reduction techniques for radiomics

	OS	**RFS**
PORTAL PHASE ONLY	0.807	0.763
ALL RADIOMICS CONCATENATED	0.838	0.811
KMCCA	0.838	0.729
MCCA	0.821	0.779

From Table 4.9, it can be concluded that information supplied by the three phases of radiomics are different from each other, as every phase offers and added value to the model, increasing predictive performances. For survival analysis, considering the multiview nature of the data does not increase the predictive ability of the model, which is at best equalled for OS with KMCCA, compared to the case where all radiomics is considered concatenated. However, it must be remembered that in this case we are not doing validation and the number of starting variables in the case of all radiomics concatenated is very high, unlike the other cases, so that performances may be slightly overestimated.
After all these considerations, we can state that MCCA and KMCCA are valuable methods to perform dimensionality reduction considering the multiview nature of data. However, it must be remembered that using this type of method to decrease the size of the dataset leads to a loss of interpretability of the result. That is because with MCCA and KMCCA we do not deal with original covariates but with a transformation of them.

5 | Conclusions

Within this work, we developed robust models capable to classify pathology data and predict survival response in patients with IHC, using a multicenter trial provided by Humanitas University. In the proposed models we tried to assess the role of radiomics, together with clinical variables, in order to provide clinicians with some relevant and actionable insights. For what concerns Classification, we first employed Generalized Linear Models. Then, we used their random effect version, in order to properly account for hierarchy of the data in the best models obtained. The same approach was followed for time-to-event data: initially we used Cox type regression models to find the best set of covariates among several at disposal. Afterwards, we employed Shared Frailty models with features selected in the best model. At the end of the work, in order to better consider the multiview aspect of the radiomic information available, we used Multiview Canonical Correlation Analysis and Kernel Multiview Canonical Correlation Analysis as dimensionality reduction techniques. With these methods, we were able to decrease the number of covariates to be given as input to the model with respect to optimising the process by considering all views simultaneously. The results of the dimensionality reduction were used in Classification and Survival Analysis to understand the advantages of using all radiomic information with a multiview approach.

We highlighted the importance of considering the information provided by radiomics, in conjunction with clinical data, to have an adequate prognosis in patients with IHC. Using radiomics, predictive performances of classification and survival models improve up to a ROC AUC of 0.795 for MVI and 0.753 for Grading, and up to a C-Index of 0.797 for OS and 0.733 for RFS, so that we are capable of predicting more accurately quantities that are relevant to know in order to find the proper treatment. Moreover, we discovered that both the radiomics of the core tumor zone and surrounding peritumoral area are relevant for the analysis, as together contribute in prediction of the outcomes. With regard to the three different phases of the CT scan, we showed the importance of considering them together, as they are not redundant descriptions of the same subject, but each provides added value to the analysis. How to take this information into account, whether with a multiview approach or not, depends on the modelled outcome, as there is no technique

that always outperforms the others. Furthermore, we pointed out that when modelling pathology data, it is necessary to consider the grouping factor present in the data, since in MEMs the centre-related random effect is strongly present. On the other hand, for Survival Analysis, hospital grouping may not be taken into account as it is not significant in the model, probably because long-term outcomes are examined.

A possible future development of this work could be to explore other Multiview dimensionality reduction/ feature selection techniques. The ideal would be to develop a variable selection technique able to reduce the number of covariates in each view, optimising the procedure by considering all the views jointly. This procedure should produce as output the actual features, not a transformation of them. In this way it would be possible to maintain the interpretability of the results, even when using Multiview techniques.

Bibliography

[1] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

[2] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[3] Sumera Rizvi, Shahid A Khan, Christopher L Hallemeier, Robin K Kelley, and Gregory J Gores. Cholangiocarcinoma—evolving concepts and therapeutic strategies. *Nature reviews Clinical oncology*, 15(2):95–111, 2018.

[4] Supriya K Saha, Andrew X Zhu, Charles S Fuchs, and Gabriel A Brooks. Forty-year trends in cholangiocarcinoma incidence in the us: intrahepatic disease on the rise. *The oncologist*, 21(5):594–599, 2016.

[5] Jesus M Banales, Vincenzo Cardinale, Guido Carpino, Marco Marzioni, Jesper B Andersen, Pietro Invernizzi, Guro E Lind, Trine Folseraas, Stuart J Forbes, Laura Fouassier, et al. Expert consensus document: Cholangiocarcinoma: current knowledge and future perspectives consensus statement from the european network for the study of cholangiocarcinoma (ens-cca). *Nature reviews. Gastroenterology & hepatology*, 13(5):261–280, 2016.

[6] Vincenzo Mazzaferro, Andre Gorgen, Sasan Roayaie, Michele Droz dit Busset, and Gonzalo Sapisochin. Liver resection and transplantation for intrahepatic cholangiocarcinoma. *Journal of hepatology*, 72(2):364–377, 2020.

[7] Guido Torzilli, Luca Viganò, Andrea Fontana, Fabio Procopio, Alfonso Terrone, Matteo M Cimino, Matteo Donadon, and Daniele Del Fabbro. Oncological outcome of r1 vascular margin for mass-forming cholangiocarcinoma. a single center observational cohort analysis. *HPB*, 22(4):570–577, 2020.

[8] Simone Conci, Luca Viganò, Giorgio Ercolani, Esteban Gonzalez, Andrea Ruzzenente, Giulia Isa, Claudia Salaris, Andrea Fontana, Fabio Bagante, Corrado Pedrazzani, et al. Outcomes of vascular resection associated with curative intent

hepatectomy for intrahepatic cholangiocarcinoma. *European Journal of Surgical Oncology*, 46(9):1727–1733, 2020.

[9] Jonathan J Hue, Flavio G Rocha, John B Ammori, Jeffrey M Hardacre, Luke D Rothermel, Kenneth D Chavin, Jordan M Winter, and Lee M Ocuin. A comparison of surgical resection and liver transplantation in the treatment of intrahepatic cholangiocarcinoma in the era of modern chemotherapy: An analysis of the national cancer database. *Journal of surgical oncology*, 123(4):949–956, 2021.

[10] Fabio Bagante, Gaya Spolverato, Matthew Weiss, Sorin Alexandrescu, Hugo P Marques, Luca Aldrighetti, Shishir K Maithel, Carlo Pulitano, Todd W Bauer, Feng Shen, et al. Defining long-term survivors following resection of intrahepatic cholangiocarcinoma. *Journal of Gastrointestinal Surgery*, 21(11):1888–1897, 2017.

[11] Alexandre Doussot, Mithat Gonen, Jimme K Wiggers, Bas Groot-Koerkamp, Ronald P DeMatteo, David Fuks, Peter J Allen, Olivier Farges, T Peter Kingham, Jean Marc Regimbeau, et al. Recurrence patterns and disease-free survival after resection of intrahepatic cholangiocarcinoma: preoperative and postoperative prognostic models. *Journal of the American College of Surgeons*, 223(3):493–505, 2016.

[12] Michael N Mavros, Konstantinos P Economopoulos, Vangelis G Alexiou, and Timothy M Pawlik. Treatment and prognosis for patients with intrahepatic cholangiocarcinoma: systematic review and meta-analysis. *JAMA surgery*, 149(6):565–574, 2014.

[13] Masayo Tsukamoto, Yo-ichi Yamashita, Katsunori Imai, Naoki Umezaki, Takanobu Yamao, Hirohisa Okabe, Shigeki Nakagawa, Daisuke Hashimoto, Akira Chikamoto, Takatoshi Ishiko, et al. Predictors of cure of intrahepatic cholangiocarcinoma after hepatic resection. *Anticancer research*, 37(12):6971–6975, 2017.

[14] Philippe Lambin, Ralph TH Leijenaar, Timo M Deist, Jurgen Peerlings, Evelyn EC De Jong, Janita Van Timmeren, Sebastian Sanduleanu, Ruben THM Larue, Aniek JG Even, Arthur Jochems, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology*, 14(12):749–762, 2017.

[15] Jiangdian Song, Yanjie Yin, Hairui Wang, Zhihui Chang, Zhaoyu Liu, and Lei Cui. A review of original articles published in the emerging field of radiomics. *European journal of radiology*, 127:108991, 2020.

[16] Gu-Wei Ji, Fei-Peng Zhu, Yu-Dong Zhang, Xi-Sheng Liu, Fei-Yun Wu, Ke Wang, Yong-Xiang Xia, Yao-Dong Zhang, Wang-Jie Jiang, Xiang-Cheng Li, et al. A ra-

diomics approach to predict lymph node metastasis and clinical outcome of intrahepatic cholangiocarcinoma. *European radiology*, 29(7):3725–3735, 2019.

[17] Hyo Jung Park, Bumwoo Park, Seo Young Park, Sang Hyun Choi, Hyungjin Rhee, Ji Hoon Park, Eun-Suk Cho, Suk-Keu Yeom, Sumi Park, Mi-Suk Park, et al. Preoperative prediction of postsurgical outcomes in mass-forming intrahepatic cholangiocarcinoma based on clinical, radiologic, and radiomics features. *European radiology*, 31(11):8638–8648, 2021.

[18] Jiahui Zhang, Xiaoli Wang, Lixia Zhang, Linpeng Yao, Xing Xue, Siying Zhang, Xin Li, Yuanjun Chen, Peipei Pang, Dongdong Sun, et al. Radiomics predict postoperative survival of patients with primary liver cancer with different pathological types. *Annals of Translational Medicine*, 8(13), 2020.

[19] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2016.

[20] Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

[21] John W Graham. Missing data analysis: Making it work in the real world. *Annual review of psychology*, 60:549–576, 2009.

[22] Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338, 2009.

[23] Samuel Wilson. miceforest Documentation. URL `https://pypi.org/project/miceforest/`.

[24] Malcolm Gladwell. *Outliers: The story of success*. Little, Brown, 2008.

[25] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[26] Carlos Oscar Sánchez Sorzano, Javier Vargas, and A Pascual Montano. A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*, 2014.

[27] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[28] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[29] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[30] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[31] I. T. Jolliffe (auth.). *Principal Component Analysis*. Breakthroughs in Statistics. Springer New York, 1986.

[32] Ian T Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(3):300–303, 1982.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[34] José Pinheiro and Douglas Bates. *Mixed-effects models in S and S-PLUS*. Springer science & business media, 2006.

[35] Harvey Goldstein, William Browne, and Jon Rasbash. Partitioning variation in multilevel models. *Understanding statistics: statistical issues in psychology, education, and the social sciences*, 1(4):223–231, 2002.

[36] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

[37] Edwin JG Pitman. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937.

[38] NCI Dictionary of Cancer Terms - OS Definition, . URL `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/overall-survival`.

[39] NCI Dictionary of Cancer Terms - RFS Definition, . URL `https://www.cancer.gov/publications/dictionaries/cancer-terms/def/relapse-free-survival`.

[40] Mitchel Klein and David G. Kleinbaum. *Survival Analysis: A Self-Learning Text*. Springer-Verlag New York, 2012.

[41] Elisa T. Lee and John Wang. *Statistical Methods for Survival Data Analysis*. Wiley, 2003.

[42] David W. Hosmer and Stanley Lemeshow. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley, 2008.

[43] Cox R David et al. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, 34(2):187–220, 1972.

[44] Wienke and Andreas. *Frailty models in survival analysis*. CRC press, 2010.

[45] Vaupel, James W, Manton, Kenneth G, Stallard, and Eric. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454, 1979.

[46] David D Hanagal. *Modeling survival data using frailty models*. Springer, 2011.

[47] Clayton and David G. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151, 1978.

[48] David Oakes. A model for association in bivariate survival data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(3):414–422, 1982.

[49] Michael J Pencina and Ralph B D'Agostino. Overall c as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in medicine*, 23(13):2109–2123, 2004.

[50] Per Kragh Andersen, John P Klein, and Mei-Jie Zhang. Testing for centre effects in multi-centre survival studies: a monte carlo comparison of fixed and random effects tests. *Statistics in medicine*, 18(12):1489–1500, 1999.

[51] Theodor Adrian Balan and Hein Putter. frailtyEM: An R package for estimating semiparametric shared frailty models. *Journal of Statistical Software*, 90(7):1–29, 2019. doi: 10.18637/jss.v090.i07.

[52] Jon R Kettenring. Canonical analysis of several sets of variables. *Biometrika*, 58(3): 433–451, 1971.

[53] Jan Rupnik and John Shawe-Taylor. Multi-view canonical correlation analysis. In *Conference on data mining and data warehouses (SiKDD 2010)*, pages 1–4, 2010.

[54] Ronan Perry, Gavin Mischler, Richard Guo, Theodore Lee, Alexander Chang, Arman Koul, Cameron Franz, Hugo Richard, Iain Carmichael, Pierre Ablin, Alexandre

Gramfort, and Joshua T. Vogelstein. mvlearn: Multiview machine learning in python. *Journal of Machine Learning Research*, 22(109):1–7, 2021.

[55] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

[56] Mary M Galloway. Texture analysis using gray level run lengths. *Computer graphics and image processing*, 4(2):172–179, 1975.

[57] Moses Amadasun and Robert King. Textural features corresponding to textural properties. *IEEE Transactions on systems, man, and Cybernetics*, 19(5):1264–1274, 1989.

[58] G Thibault, B Fertil, C Navarro, S Pereira, P Cau, N Levy, J Sequeira, and J Mari. Texture indexes and gray level size zone matrix. *Application to Cell Nuclei Classification. PRIP*, pages 140–145, 2009.

A | Appendix A

In this Appendix we report detailed information about radiomics covariates.

Table A.1: Types of radiomic variables present in IHC dataset

Radiomic Variable	Type
CONVENTIONAL HUmin	Basic
CONVENTIONAL HUmean	Basic
CONVENTIONAL HUstd	Basic
CONVENTIONAL HUmax	Basic
CONVENTIONAL HUQ1	Basic
CONVENTIONAL HUQ2	Basic
CONVENTIONAL HUQ3	Basic
CONVENTIONAL HUSkewness	Basic
CONVENTIONAL HUKurtosis	Basic
CONVENTIONAL HUExcessKurtosis	Basic
DISCRETIZED HISTO Entropy log10	First Order
DISCRETIZED HISTO Entropy log2	First Order
DISCRETIZED HISTO Energy Uniformity	First Order
SHAPE Volume mL	First Order
SHAPE Volume vx	First Order
SHAPE Sphericity onlyFor3DROI	First Order
SHAPE Surface mm2 onlyFor3DROI	First Order
SHAPE Compacity onlyFor3DROI	First Order
GLCM Homogeneity InverseDifference	Second Order
GLCM Energy AngularSecondMoment	Second Order
GLCM Contrast Variance	Second Order
GLCM Correlation	Second Order
GLCM Entropy log10	Second Order
GLCM Entropy log2 JointEntropy	Second Order
GLCM Dissimilarity	Second Order
GLRLM SRE	Second Order
GLRLM LRE	Second Order
GLRLM LGRE	Second Order
GLRLM HGRE	Second Order

Radiomic Variable	Type
GLRLM SRLGE	Second Order
GLRLM SRHGE	Second Order
GLRLM LRLGE	Second Order
GLRLM LRHGE	Second Order
GLRLM GLNU	Second Order
GLRLM RLNU	Second Order
GLRLM RP	Second Order
NGLDM Coarseness	Second Order
NGLDM Contrast	Second Order
NGLDM Busyness	Second Order
GLZLM SZE	Second Order
GLZLM LZE	Second Order
GLZLM LGZE	Second Order
GLZLM HGZE	Second Order
GLZLM SZLGE	Second Order
GLZLM SZHGE	Second Order
GLZLM LZLGE	Second Order
GLZLM LZHGE	Second Order
GLZLM GLNU	Second Order
GLZLM ZLNU	Second Order
GLZLM ZP	Second Order

Second order parameters are found imposing filter grids on the image. The filters are the following:

- GLCM: It stands for Gray Levels Co-occurrence Matrix. The matrix describes the second-order joint probability function of an image region constrained by the mask and is defined as $P(i, j|\delta, \theta)$. The $(i, j)^{th}$ element of this matrix represents the number of times the combination of levels i and j occur in two pixels in the image, that are separated by a distance of δ pixels along angle θ [55].
- GLRLM: It stands for Gray Level Run Length Zone Matrix. The matrix quantifies gray level runs, which are defined as the length in number of pixels, of consecutive pixels that have the same gray level value. In a gray level run length matrix $P(i, j|\theta)$, the $(i, j)^{th}$ element describes the number of runs with gray level i and length j occur in the image (VOI) along angle θ [56].
- NGLDM: It stands for Neighboring Gray Level Difference Matrix. The matrix quantifies the difference between a gray value and the average gray value of its neighbours within distance δ [57].

- GLZLM: It stands for Gray Level Zone Length Matrix. The matrix quantifies gray level zones in an image. A gray level zone is defined as the number of connected voxels that share the same gray level intensity. In matrix $P(i, j)$ the $(i, j)^{th}$ element equals the number of zones with gray level i and size j appear in image [58].

B | Appendix B

In this Appendix the main parts of the code used to develop this work are reported.

B.1. Classification Code

The following code contains the function used to compute the performances of Logistic Regression models without cross-validation.

```
def performace_LR(model, X_train, y_train, X_test, y_test):
  model.fit(X_train, y_train);
  yp=model.predict(X_test)
  yprob = model.predict_proba(X_test)
  #Compute performances
  accuracy=model.score(X_test, y_test)
  cm = confusion_matrix(y_test, yp)
  PrintConfusionMatrix(y_test, yp)
  Specificity = cm[0,0]/(cm[0,0]+cm[0,1])
  Sensitivity = cm[1,1]/(cm[1,0]+cm[1,1])
  precision=precision_score(y_test,yp, pos_label='1')
  recall=recall_score(y_test,yp,pos_label='1')
  pr_auc= average_precision_score(y_test, yprob[:,1],pos_label='1')
  roc_auc = roc_auc_score(y_true=y_test, y_score = yprob[:,1])
  print("Accuracy %3.2f"  % accuracy)
  print("Specificity %3.2f" % Specificity)
  print("Sensitivity    %3.2f" % Sensitivity)
  print("Precision    %3.2f" % precision)
  print("Precision-Recall AUC  %3.2f" %pr_auc)
  print("ROC AUC  %3.2f" %roc_auc)
  precision, recall, thresholds = precision_recall_curve(y_true=y_test,
   probas_pred=yprob[:,1],pos_label='1')
  auc = average_precision_score(y_test, yprob[:,1],pos_label='1')
  #Plot Precision-Recall Curves
  plt.figure(1, figsize=(8, 6));
  font = {'family':'sans', 'size':24};
  plt.rc('font', **font);
  plt.plot(recall, precision, label="Precision-recall curve");
```

```
  plt.xlabel('Recall');
  plt.ylabel('Precision');
  plt.ylim([0.5,1.1])
  plt.yticks(np.arange(0.5,1.01,.1))
  plt.title('Precision-Recall Curve (AUC=%3.2f)'%auc);
  plt.plot(recall[:-1],thresholds, label="Threshold");
  plt.legend()
  plt.show()
  fpr, tpr, thresholds = roc_curve(y_true=y_test, y_score = yprob[:,1],
   pos_label='1')
  #Plot ROC Curve
  plt.figure(1, figsize=(8, 8));
  font = {'family':'sans', 'size':24};
  plt.rc('font', **font);
  plt.xlabel('FPR');
  plt.ylabel('TPR');
  plt.plot(fpr,tpr,label='Classifier')
  #plt.plot(fpr,thresholds,label='Thresholds')
  plt.plot([0.0,1.0],[0.0,1.0],label='Baseline')
  plt.yticks(np.arange(0.0,1.01,.2))
  plt.title('ROC Curve (AUC=%3.2f)'%roc_auc)
  plt.ylim([0.0,1.0])
  plt.xlim([0.0,1.0])
  plt.legend()
  plt.show();

  return accuracy, Specificity, Sensitivity, precision, recall, pr_auc,
   roc_auc
```

Logistic Regression model is identified with `sklearn` function
`LogisticRegression(solver='liblinear', random_state=0, max_iter=500)`

The following code contains the function used to compute performances with Cross-validation Method 1.

```
def cross_validation1(model, X, y, n_split):
  cv = StratifiedKFold(n_splits=n_split, random_state=0, shuffle=True)
  kfold = cv.split(X, y)

  accu=[]
  spec=[]
  sens=[]
  prec=[]
  rec=[]
```

```
  pr_auc=[]
  roc_auc=[]

  for k, (train, test) in enumerate(kfold):
    result=performace_LR(model, X.iloc[train, :], y.iloc[train], X.iloc[
   test, :], y.iloc[test])
    accu.append(result[0])
    spec.append(result[1])
    sens.append(result[2])
    prec.append(result[3])
    rec.append(result[4])
    pr_auc.append(result[5])
    roc_auc.append(result[6])

  print('\n Cross-Validation 1 Accuracy: %.3f +/- %.3f' %(np.mean(accu),
    np.std(accu)))
  print('\n Cross-Validation 1 Specificity: %.3f +/- %.3f' %(np.mean(
   spec), np.std(spec)))
  print('\n Cross-Validation 1 Sensitivity: %.3f +/- %.3f' %(np.mean(
   sens), np.std(sens)))
  print('\n Cross-Validation 1 Precision: %.3f +/- %.3f' %(np.mean(prec)
   , np.std(prec)))
  print('\n Cross-Validation 1 ecall: %.3f +/- %.3f' %(np.mean(rec), np.
   std(rec)))
  print('\n Cross-Validation 1 PR_AUC: %.3f +/- %.3f' %(np.mean(pr_auc),
    np.std(pr_auc)))
  print('\n Cross-Validation 1 ROC_AUC: %.3f +/- %.3f' %(np.mean(roc_auc
   ), np.std(roc_auc)))
```

The following code contains the function used to compute performances with Cross-validation Method 2.

```
def cross_validation2(model, X, y, n_test, percentage):
  accu=[]
  spec=[]
  sens=[]
  prec=[]
  rec=[]
  pr_auc=[]
  roc_auc=[]

  for i in range(n_test):
     X_train, X_test, y_train, y_test = train_test_split(X, y, test_size
   =percentage, random_state=i, stratify=y)
     result=performace_LR(model, X_train, y_train, X_test, y_test)
```

```
13      accu.append(result[0])
14      spec.append(result[1])
15      sens.append(result[2])
16      prec.append(result[3])
17      rec.append(result[4])
18      pr_auc.append(result[5])
19      roc_auc.append(result[6])
20
21    print('\n Cross-Validation 2 Accuracy: %.3f +/- %.3f' %(np.mean(accu),
        np.std(accu)))
22    print('\n Cross-Validation 2 Specificity: %.3f +/- %.3f' %(np.mean(
       spec), np.std(spec)))
23    print('\n Cross-Validation 2 Sensitivity: %.3f +/- %.3f' %(np.mean(
       sens), np.std(sens)))
24    print('\n Cross-Validation 2 Precision: %.3f +/- %.3f' %(np.mean(prec)
       , np.std(prec)))
25    print('\n Cross-Validation 2 Recall: %.3f +/- %.3f' %(np.mean(rec), np
       .std(rec)))
26    print('\n Cross-Validation 2 PR_AUC: %.3f +/- %.3f' %(np.mean(pr_auc),
        np.std(pr_auc)))
27    print('\n Cross-Validation 2 ROC_AUC: %.3f +/- %.3f' %(np.mean(roc_auc
       ), np.std(roc_auc)))
28
29    return accu, spec, sens, prec, rec, pr_auc, roc_auc
```

The following code concerns the fitting of the Mixed Effects Model in the case for IVM with Clinical+Portal(Core+Margin). For Grading and other cases of covariates, the code is basically the same.

```
1  #Setting the working directory
2  setwd("C:/Users/N/Desktop/Polimi/Magistrale/Tesi/Materiale/Datasets")
3
4  #Prepare the dataset
5  X <- read_excel("IVM.xlsx")
6  X$CENTRO<-as.factor(X$CENTRO)
7  X$`MAJOR HEPATECTOMY`<-as.numeric(X$`MAJOR HEPATECTOMY`)
8  X$SEX<-as.numeric(X$SEX)#lo considero numerico perch ne voglio solo uno
       se no sono correlati e da problemi
9  X$INVASIONE.VASCOLARE.MICROSCOPICA<-as.factor(X$INVASIONE.VASCOLARE.
       MICROSCOPICA)
10
11 #Fit the model
12 ivm_mem = glmer(INVASIONE.VASCOLARE.MICROSCOPICA ~ 0+SEX+`CA 19-9`+`
       MAJOR HEPATECTOMY`+PM_CONVENTIONAL.HUQ2+
13                    PM_SHAPE.SphericityonlyFor3DROI+PM_GLCM.Correlation+PM
```

```
   _NGLDM.Contrast+
                  PM_GLZLM.LZE+PM_GLZLM.SZHGE+(1|CENTRO),
                data=X,
                family = binomial,
                control=glmerControl(optimizer="bobyqa",optCtrl=list(
   maxfun=2e5)))

plot_model(ivm_mem)
summary(ivm_mem)

#Compute VPC
print(vc <- VarCorr(ivm_mem), comp = c("Variance", "Std.Dev."))
sigma2_eps <- as.numeric(get_variance_residual(ivm_mem))
sigma2_eps
sigma2_b <- as.numeric(get_variance_random(ivm_mem))
sigma2_b
VPC <- sigma2_b/(sigma2_b+sigma2_eps)

#Plot the random effects
library(lattice)
rand_intercept = ranef(ivm_mem, condVar=TRUE)
lattice::dotplot(rand_intercept,strip=T, lty= 4)

#Performance on all the dataset
pred <- predict(ivm_mem,type='response')
class_pred <- ifelse(pred>0.5,1,0)
tb <- table(true=X$INVASIONE.VASCOLARE.MICROSCOPICA, assigned=class_pred
   )
accuracy <- (tb[1]+tb[4])/sum(tb)
spec<-tb[1]/(tb[1]+tb[2])
sens<-tb[4]/(tb[3]+tb[4])
prec<-tb[4]/(tb[4]+tb[2])

#Plot ROC Curve
prob <- predict(ivm_mem, type="response")
pred <- prediction(prob, X$INVASIONE.VASCOLARE.MICROSCOPICA)
roc_auc <- performance(pred, measure = "auc")
roc_auc <- roc_auc@y.values[[1]]
roc_auc
prob <- predict(ivm_mem, type="response")
pred <- prediction(prob, X$INVASIONE.VASCOLARE.MICROSCOPICA)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")
plot(perf, main= paste0("ROC Curve (AUC: ", round(roc_auc, 2), ")"))

```

```
#Plot Precision-Recall Curve
perf_pr <- performance(pred,"prec","rec")
x = perf_pr@x.values[[1]]
y = perf_pr@y.values[[1]]
idx = 2:length(x)
testdf=data.frame(recall = (x[idx] - x[idx-1]), precision = (y[idx] + y[
    idx-1]))
testdf = subset(testdf, !is.na(testdf$precision))
pr_auc = sum(testdf$recall * testdf$precision)/2
plot(perf_pr, main= paste0("Precision-Recall Curve (AUC: ", round(pr_auc
    , 2), ")"))

print(paste0("   Accuracy:    ", round(accuracy, 3), "   "))
print(paste0("   Specficity:  ", round(spec, 3), "   "))
print(paste0("   Sensitivity:  ", round(sens, 3), "   "))
print(paste0("   Precision:  ", round(prec, 3), "   "))
print(paste0("   ROC AUC:    ", round(roc_auc, 3), "   "))
print(paste0("   PR AUC:    ", round(pr_auc, 3), "   "))
```

B.2. Survival Analysis Code

The following code is about using Log Rank Tests for skimming the clinical categorical variables before Survival Analysis

```
library(readxl)
library(car)
library(caret)
library(survival)
library(survminer)

set.seed(1)

main <- read_excel("main_imputed.xlsx")
main<-main[-261,]
main[which(main[,21]=='X'),21]<-'x'
main <- na.omit(main)

main$OS<-as.numeric(main$OS)
main$RFS<-as.numeric(main$RFS)
main$RECIDIVA<-as.numeric(main$RECIDIVA)
main$STATO.VM<-as.numeric(main$STATO.VM)

survdiff(Surv(main$OS, main$STATO.VM) ~ main$SEX, data = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$HCV, data = main)
```

```
survdiff(Surv(main$OS, main$STATO.VM) ~ main$HBV, data = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$`Ca19-9gt55`, data = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$CHEMIOTERAPIA.NEOADIUVANTE,
    data = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$PRIMA.RESEZIONE, data =
   main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$Major.Hepatectomy, data =
   main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$RESEZIONE.VIA.BILIARE, data
    = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$LINFOADENECTOMIA, data =
   main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$ASSOCIATED.RESECTION, data
   = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$COMPLICANZE.SEVERE, data =
   main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$CIRROSI, data = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$PATTERN, data = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$SINGLE.NODULE, data = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$T.VIII.ed, data = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$N, data = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$M, data = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$GRADING.RAGGR, data = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$R.status, data = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$IVM, data = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$INFILTRAZIONE.PERINEURALE,
   data = main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$NODULI.SATELLITI, data =
   main)
survdiff(Surv(main$OS, main$STATO.VM) ~ main$CHEMIOTERAPIA.ADIUVANTE,
   data = main)

fit1<-survfit(Surv(main$OS, main$STATO.VM) ~ main$SEX, data = main)
fit2<-survfit(Surv(main$OS, main$STATO.VM) ~ main$HCV, data = main)
fit3<-survfit(Surv(main$OS, main$STATO.VM) ~ main$HBV, data = main)
fit4<-survfit(Surv(main$OS, main$STATO.VM) ~ main$`Ca19-9gt55`, data =
   main)
fit5<-survfit(Surv(main$OS, main$STATO.VM) ~ main$CHEMIOTERAPIA.
   NEOADIUVANTE, data = main)
fit6<-survfit(Surv(main$OS, main$STATO.VM) ~ main$PRIMA.RESEZIONE, data
   = main)
fit7<-survfit(Surv(main$OS, main$STATO.VM) ~ main$Major.Hepatectomy,
   data = main)
```

```
fit8<-survfit(Surv(main$OS, main$STATO.VM) ~ main$RESEZIONE.VIA.BILIARE, data = main)
fit9<-survfit(Surv(main$OS, main$STATO.VM) ~ main$LINFOADENECTOMIA, data = main)
fit10<-survfit(Surv(main$OS, main$STATO.VM) ~ main$ASSOCIATED.RESECTION, data = main)
fit11<-survfit(Surv(main$OS, main$STATO.VM) ~ main$COMPLICANZE.SEVERE, data = main)
fit12<-survfit(Surv(main$OS, main$STATO.VM) ~ main$CIRROSI, data = main)
fit13<-survfit(Surv(main$OS, main$STATO.VM) ~ main$PATTERN, data = main)
fit14<-survfit(Surv(main$OS, main$STATO.VM) ~ main$SINGLE.NODULE, data = main)
fit15<-survfit(Surv(main$OS, main$STATO.VM) ~ main$T.VIII.ed, data = main)
fit16<-survfit(Surv(main$OS, main$STATO.VM) ~ main$N, data = main)
fit17<-survfit(Surv(main$OS, main$STATO.VM) ~ main$M, data = main)
fit18<-survfit(Surv(main$OS, main$STATO.VM) ~ main$GRADING.RAGGR, data = main)
fit19<-survfit(Surv(main$OS, main$STATO.VM) ~ main$R.status, data = main)
fit20<-survfit(Surv(main$OS, main$STATO.VM) ~ main$IVM, data = main)
fit21<-survfit(Surv(main$OS, main$STATO.VM) ~ main$INFILTRAZIONE.PERINEURALE, data = main)
fit22<-survfit(Surv(main$OS, main$STATO.VM) ~ main$NODULI.SATELLITI, data = main)
fit23<-survfit(Surv(main$OS, main$STATO.VM) ~ main$CHEMIOTERAPIA.ADIUVANTE, data = main)

splots <- list()
splots[[1]]<-ggsurvplot(fit1)
splots[[2]]<-ggsurvplot(fit2)
splots[[3]]<-ggsurvplot(fit3)
splots[[4]]<-ggsurvplot(fit4)
splots[[5]]<-ggsurvplot(fit5)
splots[[6]]<-ggsurvplot(fit6)
splots[[7]]<-ggsurvplot(fit7)
splots[[8]]<-ggsurvplot(fit8)
splots[[9]]<-ggsurvplot(fit9)
splots[[10]]<-ggsurvplot(fit10)
splots[[11]]<-ggsurvplot(fit11)
splots[[12]]<-ggsurvplot(fit12)
splots[[13]]<-ggsurvplot(fit13)
splots[[14]]<-ggsurvplot(fit14)
splots[[15]]<-ggsurvplot(fit15)
```

```
splots[[16]]<-ggsurvplot(fit16)
splots[[17]]<-ggsurvplot(fit17)
splots[[18]]<-ggsurvplot(fit18)
splots[[19]]<-ggsurvplot(fit19)
splots[[20]]<-ggsurvplot(fit20)
splots[[21]]<-ggsurvplot(fit21)
splots[[22]]<-ggsurvplot(fit22)
splots[[23]]<-ggsurvplot(fit23)

arrange_ggsurvplots(splots[1:6], print = TRUE, ncol = 3, nrow = 2)
arrange_ggsurvplots(splots[7:12], print = TRUE, ncol = 3, nrow = 2)
arrange_ggsurvplots(splots[13:18], print = TRUE, ncol = 3, nrow = 2)
arrange_ggsurvplots(splots[19:23], print = TRUE, ncol = 3, nrow = 2)
```

The following code concerns the fitting of the Cox-PH model in the case of Postoperative+Portal(Core +Margin) covariates for OS. For the other cases of covariates and for RFS the code is basically the same.

```
library(readxl)
library(car)
library(caret)
library(survival)
library(survminer)

set.seed(1)

X <- read_excel("X_no_c_scale.xlsx")
X[which(X[,21]=='X'),21]<-'x'
X[which(X[,33]=='1-2'),33]<-'0'
X[which(X[,33]=='3'),33]<-'1'
X$PATTERN.<-as.factor(X$PATTERN.)
X$OS.Days<-as.numeric(X$OS.Days)
X$STATO.VIVO.MORTO<-as.numeric(X$STATO.VIVO.MORTO)

#Elimino
X$CENTRO<-NULL
X$`Codice.PAZ`<-NULL
X$SEX<-NULL
X$HCV<-NULL
X$HBV<-NULL
X$CHEMIOTERAPIA.NEOADIUVANTE<-NULL
X$PRIMA.RESEZIONE<-NULL
X$ASSOCIATED.RESECTION<-NULL
X$CIRROSI<-NULL
X$GRADING<-NULL
```

```
X$CHEMIOTERAPIA.ADIUVANTE<-NULL
X$RECIDIVA<-NULL
X$RFS.Days<-NULL
X$Ca19.9gt55<-NULL

surv_obj <- Surv(X$OS.Days, X$STATO.VIVO.MORTO)

X$STATO.VIVO.MORTO<-NULL
X$OS.Days<-NULL

m_null <- coxph( surv_obj ~ 1, data = X)

mod.cox <- coxph( surv_obj ~ .,
                  data = X,
                  control = coxph.control(iter.max = 100))
summary(mod.cox)

step(m_null, trace = F, scope = list(lower=formula(m_null), upper=
   formula(mod.cox)),
     direction = 'both', data = X)

cox.reduced <- coxph( surv_obj ~ AGE + Ca.19.9 + Major.Hepatectomy + N +
    R.status+ COMPLICANZE.SEVERE  + PATTERN.  + NODULI.SATELLITI +
   RESEZIONE.VIA.BILIARE + GLRLM_SRHGE+  CONVENTIONAL.HUKurtosis+SHAPE.
   CompacityonlyFor3DROI + GLRLM.SRHGE + GLRLM.GLNU + GLZLM.ZP +  GLZLM.
   LZLGE + GLRLM.LGRE,
                      data = X,
                      control = coxph.control(iter.max = 100))

summary(cox.reduced)

#Hazard Ratio and CI plot
x11()
ggforest(cox.reduced, data=X)

prediction = predict(cox.reduced, X)
score = survConcordance(surv_obj ~ prediction, data = X)$concordance
print(paste('Concordance index',score))

# Plot martingale residuals
x11()
ggcoxdiagnostics(cox.reduced, type = "martingale")

x11()
```

```
ggcoxdiagnostics(cox.reduced, type = "deviance")

x11()
ggcoxdiagnostics(cox.reduced, type = "schoenfeld")

print('Proportional Hazard assumption')
test.ph <- cox.zph(cox.reduced)
test.ph

print('schoenfeld residuals')

par(mfrow=c(2,3))
for(i in 1:17){
  plot(test.ph[i])
  abline(h=0, col='red')
}

#### Estimated Baseline Survival Curves ####
####-------------------------------------####
fit<-survfit(cox.reduced, data=X)

x11()
plot(fit, conf.int=TRUE,
     col=c('dodgerblue2','mediumseagreen','orangered'), lwd=2, lty=1,
     xlab='Time [days]', ylab='Survival Probability',
     main='Estimated Baseline Survival Probabilities')
grid()
```

The following code concerns the fitting of the Shared Frailty model in the case for OS. For the case of RFS the code is basically the same.

```
library(readxl)
library(car)
library(caret)
library(survival)
library(survminer)
library(frailtypack)
library(frailtyEM)
library(frailtySurv)

set.seed(1)

X <- read_excel("X_no_c_scale.xlsx")
X[which(X[,21]=='X'),21]<-'x'
X[which(X[,33]=='1-2'),33]<-'0'
```

```
X[which(X[,33]=='3'),33]<-'1'
X$PATTERN.<-as.factor(X$PATTERN.)
X$N<-as.factor(X$N)
X$OS.Days<-as.numeric(X$OS.Days)
X$STATO.VIVO.MORTO<-as.numeric(X$STATO.VIVO.MORTO)

#Elimino
X$`Codice.PAZ`<-NULL
X$SEX<-NULL
X$HCV<-NULL
X$HBV<-NULL
X$CHEMIOTERAPIA.NEOADIUVANTE<-NULL
X$PRIMA.RESEZIONE<-NULL
X$ASSOCIATED.RESECTION<-NULL
X$CIRROSI<-NULL
X$GRADING<-NULL
X$CHEMIOTERAPIA.ADIUVANTE<-NULL
X$RECIDIVA<-NULL
X$RFS.Days<-NULL
X$Ca19.9gt55<-NULL

# library
library(ggplot2)

# grouped boxplot
ggplot(X, aes(x=CENTRO, y=OS.Days)) +
  geom_boxplot()

surv_obj <- Surv(X$OS.Days, X$STATO.VIVO.MORTO)

mod.frailty <- emfrail(surv_obj ~ AGE + Ca.19.9 + Major.Hepatectomy + N
   + R.status+ COMPLICANZE.SEVERE  + PATTERN.  + NODULI.SATELLITI +
   RESEZIONE.VIA.BILIARE + GLRLM_SRHGE + CONVENTIONAL.HUKurtosis+SHAPE.
   CompacityonlyFor3DROI + GLRLM.SRHGE + GLRLM.GLNU + GLZLM.ZP +  GLZLM.
   LZLGE + GLRLM.LGRE + cluster(CENTRO), data = X)

summary(mod.frailty)
```

B.3. Multiview Dimensionality Reduction Code

The following code is about the Multiview Dimensionality reduction of radiomic covariates performed with MCCA.

```
!pip install mvlearn
```

```
!pip install scipy --upgrade

from IPython.core.interactiveshell import InteractiveShell
InteractiveShell.ast_node_interactivity = "all"

# Run this cell only if you are using Colab with Drive
from google.colab import drive
drive.mount('/content/drive')

import pandas as pd
import numpy as np
from numpy import where
import matplotlib
import copy
import random
import matplotlib.pyplot as plt
from scipy import stats
from sklearn.preprocessing import StandardScaler

# %config InlineBackend.figure_format = 'retina' #set 'png' here when
    working on notebook
# %matplotlib inline

data_types={"SEX": np.str,"HCV": np.str,"HBV": np.str,"Ca19-9  55  ": np.
    str,"CHEMIOTERAPIA NEOADIUVANTE": np.str,"PRIMA RESEZIONE": np.str,"
    Major Hepatectomy": np.str,"RESEZIONE VIA BILIARE": np.str,"
    LINFOADENECTOMIA": np.str,"ASSOCIATED RESECTION": np.str,"COMPLICANZE
     SEVERE": np.str,"CIRROSI": np.str,"PATTERN": np.str,"SINGLE NODULE":
     np.str,"T VIII ed": np.str,"N": np.str,"M": np.str,"GRADING": np.str
    ,"R status": np.str,"INVASIONE VASCOLARE MICROSCOPICA": np.str,"
    INFILTRAZIONE PERINEURALE": np.str,"NODULI SATELLITI": np.str,"
    CHEMIOTERAPIA ADIUVANTE": np.str,"STATO VIVO/MORTO": np.str,"RECIDIVA
    ": np.str,"GRADING RAGGR": np.str}

#Import the dataset
main= pd.read_excel ('/content/drive/My Drive/TESI/DATASETS/main_imputed
    .xlsx')
portal_core= pd.read_excel ('/content/drive/My Drive/TESI/DATASETS/
    NO_CORR/portal_core_nocorr.xlsx')
portal_margin= pd.read_excel ('/content/drive/My Drive/TESI/DATASETS/
    NO_CORR/portal_margin_nocorr.xlsx')
arterial_core= pd.read_excel ('/content/drive/My Drive/TESI/DATASETS/
    NO_CORR/arterial_core_nocorr.xlsx')
```

```
arterial_margin= pd.read_excel ('/content/drive/My Drive/TESI/DATASETS/NO_CORR/arterial_margin_nocorr.xlsx')
late_core= pd.read_excel ('/content/drive/My Drive/TESI/DATASETS/NO_CORR/late_core_nocorr.xlsx')
late_margin= pd.read_excel ('/content/drive/My Drive/TESI/DATASETS/NO_CORR/late_margin_nocorr.xlsx')

#Join the dataset
portal=pd.concat([portal_core, portal_margin],axis=1)
arterial=pd.concat([arterial_core, arterial_margin],axis=1)
late=pd.concat([late_core, late_margin],axis=1)
temp=pd.concat([portal, arterial],axis=1)
radiomics=pd.concat([temp, late],axis=1)
X=pd.concat([main, radiomics], axis=1)

#Delete missing ravlues
X.dropna(inplace=True)

#Standardize the data
numeric_feats = X.dtypes[X.dtypes != "object" ].index
scaler = StandardScaler()
X[numeric_feats]= scaler.fit_transform(X[numeric_feats])

#Go back to single radiomics data
main_new=X.iloc[:,0:33]
portal_core_new=X.iloc[:,33:57]
portal_margin_new=X.iloc[:,57:87]
arterial_core_new=X.iloc[:,87:108]
arterial_margin_new=X.iloc[:,108:133]
late_core_new=X.iloc[:,133:147]
late_margin_new=X.iloc[:,147:172]

#Reduce core and margin separately with MCCA dimensionality reduction
from mvlearn.embed import MCCA
mcca_core = MCCA(n_components=10, regs='lw')
mcca_core.fit([portal_core_new, arterial_core_new, late_core_new])
Xs_core = mcca_core.transform([portal_core_new, arterial_core_new, late_core_new])
mcca_margin = MCCA(n_components=10, regs='lw')
mcca_margin.fit([portal_margin_new, arterial_margin_new, late_margin_new])
Xs_margin = mcca_margin.transform([portal_margin_new, arterial_margin_new, late_margin_new])

```

```
#Using invariant propertu of MCCA solution, modify the scale of the components, in order to have values with the same magnitude of clinical standardized variables
Xs_pc=pd.DataFrame(100*Xs_core[0])
Xs_ac=pd.DataFrame(100*Xs_core[1])
Xs_lc=pd.DataFrame(100*Xs_core[2])
Xs_pm=pd.DataFrame(100*Xs_margin[0])
Xs_am=pd.DataFrame(100*Xs_margin[1])
Xs_lm=pd.DataFrame(100*Xs_margin[2])
X_core=pd.concat([Xs_pc, Xs_ac, Xs_lc],axis=1)
X_margin=pd.concat([Xs_pm, Xs_am, Xs_lm],axis=1)

radiomics_new=pd.concat([X_core, X_margin],axis=1)
main_new.reset_index(drop=True, inplace=True)
radiomics_new.reset_index(drop=True, inplace=True)
X_new=pd.concat([main_new, radiomics_new], axis=1)

X_new.to_excel('/content/drive/My Drive/TESI/DATASETS/MULTIVIEW/Dataset_MCCA.xlsx', index= False)
```

The following code is the part to change in the above, from line 61 to 77, to perform KMCCA, instead of MCCA.

```
#Reduce core and margin separately with MCCA dimensionality reduction
from mvlearn.embed import KMCCA
X_core=[portal_core_new, arterial_core_new, late_core_new]
X_margin=[portal_margin_new, arterial_margin_new, late_margin_new]
mcca_core = KMCCA(n_components=10, kernel='rbf')
mcca_core.fit([portal_core_new, arterial_core_new, late_core_new])
Xs_core = mcca_core.transform([portal_core_new, arterial_core_new, late_core_new])
mcca_margin = KMCCA(n_components=10, kernel='rbf')
mcca_margin.fit([portal_margin_new, arterial_margin_new, late_margin_new])
Xs_margin = mcca_margin.transform([portal_margin_new, arterial_margin_new, late_margin_new])Xs_pc=10000*+Xs_core[0]

#Using invariant propertu of MCCA solution, modify the scale of the components, in order to have values with the same magnitude of clinical standardized variables
Xs_pc=pd.DataFrame(10000*Xs_core[0])
Xs_ac=pd.DataFrame(10000*Xs_core[1])
Xs_lc=pd.DataFrame(10000*Xs_core[2])
Xs_pm=pd.DataFrame(10000*Xs_margin[0])
Xs_am=pd.DataFrame(10000*Xs_margin[1])
```

```
Xs_lm=pd.DataFrame(10000*Xs_margin[2])
X_core=pd.concat([Xs_pc, Xs_ac, Xs_lc],axis=1)
X_margin=pd.concat([Xs_pm, Xs_am, Xs_lm],axis=1)
```

C | Appendix C

In this Appendix the performances of the Logistic Regression models with the various variable selection techniques are reported.

Table C.1: Performances of MVI LR with Clinical+Portal(Core) features

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.74	0.652	0.214	0.652	0.063
SPECIFICITY	0.68	0.610	0.387	0.589	0.099
SENSITIVITY	0.79	0.690	0.273	0.700	0.090
PRECISION	0.76	0.726	0.267	0.696	0.057
PR AUC	0.85	0.843	0.162	0.745	0.058
ROC AUC	0.81	0.713	0.284	0.691	0.067

Table C.2: Performances of MVI LR with Clinical+Portal(Core) features with Stepwise/ Forward Selection

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.72	0.721	0.209	0.708	0.060
SPECIFICITY	0.66	0.647	0.341	0.628	0.117
SENSITIVITY	0.77	0.783	0.250	0.769	0.081
PRECISION	0.75	0.776	0.201	0.738	0.061
PR AUC	0.81	0.873	0.136	0.798	0.051
ROC AUC	0.77	0.772	0.233	0.751	0.059

Table C.3: Performances of MVI LR with Clinical+Portal(Core) features with Ridge Regression

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.74	0.652	0.214	0.652	0.063
SPECIFICITY	0.68	0.610	0.387	0.589	0.099
SENSITIVITY	0.79	0.690	0.273	0.700	0.090
PRECISION	0.76	0.726	0.267	0.696	0.057
PR AUC	0.85	0.843	0.162	0.745	0.058
ROC AUC	0.81	0.713	0.284	0.691	0.067

Table C.4: Performances of MVI LR with Clinical+Portal(Core) features with Lasso Regression

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.72	0.670	0.213	0.671	0.060
SPECIFICITY	0.65	0.630	0.393	0.610	0.104
SENSITIVITY	0.77	0.707	0.262	0.718	0.089
PRECISION	0.74	0.753	0.251	0.713	0.057
PR AUC	0.84	0.867	0.150	0.780	0.055
ROC AUC	0.80	0.757	0.264	0.721	0.065

Table C.5: Performances of MVI LR with Clinical+Portal(Core) features with Principal Components Regression

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.73	0.687	0.227	0.677	0.064
SPECIFICITY	0.67	0.647	0.381	0.615	0.107
SENSITIVITY	0.78	0.647	0.381	0.723	0.091
PRECISION	0.76	0.767	0.252	0.718	0.062
PR AUC	0.84	0.862	0.156	0.782	0.056
ROC AUC	0.79	0.752	0.261	0.723	0.064

Table C.6: Performances of MVI LR with Clinical+Portal(Core+Margin) features

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.74	0.589	0.224	0.619	0.062
SPECIFICITY	0.69	0.543	0.407	0.549	0.110
SENSITIVITY	0.78	0.627	0.290	0.672	0.083
PRECISION	0.77	0.681	0.292	0.668	0.057
PR AUC	0.88	0.815	0.177	0.736	0.058
ROC AUC	0.84	0.673	0.297	0.662	0.067

Table C.7: Performances of MVI LR with Clinical+Portal(Core+Margin) features with Forward/Stepwise Selection

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.73	0.736	0.199	0.717	0.056
SPECIFICITY	0.65	0.647	0.347	0.629	0.098
SENSITIVITY	0.80	0.810	0.245	0.783	0.078
PRECISION	0.75	0.788	0.199	0.740	0.053
PR AUC	0.83	0.877	0.126	0.813	0.052
ROC AUC	0.78	0.787	0.210	0.755	0.060

Table C.8: Performances of MVI LR with Clinical+Portal(Core+Margin) features with Ridge Regression

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.74	0.641	0.216	0.654	0.063
SPECIFICITY	0.69	0.610	0.387	0.592	0.112
SENSITIVITY	0.78	0.667	0.287	0.701	0.085
PRECISION	0.77	0.726	0.277	0.699	0.061
PR AUC	0.88	0.851	0.145	0.784	0.053
ROC AUC	0.84	0.737	0.247	0.716	0.063

Table C.9: Performances of MVI LR with Clinical+Portal(Core+Margin) features with Lasso Regression

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.76	0.641	0.216	0.654	0.063
SPECIFICITY	0.67	0.610	0.387	0.592	0.112
SENSITIVITY	0.83	0.667	0.287	0.701	0.085
PRECISION	0.77	0.726	0.277	0.699	0.061
PR AUC	0.87	0.851	0.145	0.784	0.053
ROC AUC	0.83	0.737	0.247	0.716	0.063

Table C.10: Performances of MVI LR with Clinical+Portal(Core+Margin) features with Principal Components Regression

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.75	0.643	0.196	0.657	0.064
SPECIFICITY	0.70	0.627	0.355	0.592	0.103
SENSITIVITY	0.78	0.660	0.273	0.706	0.091
PRECISION	0.78	0.722	0.261	0.700	0.058
PR AUC	0.86	0.854	0.137	0.773	0.056
ROC AUC	0.82	0.740	0.235	0.705	0.071

Table C.11: Performances of Grading LR with Clinical+Portal(Core) features

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.72	0.637	0.193	0.635	0.057
SPECIFICITY	0.92	0.847	0.197	0.832	0.076
SENSITIVITY	0.33	0.230	0.349	0.229	0.102
PRECISION	0.68	0.287	0.424	0.412	0.165
PR AUC	0.62	0.614	0.266	0.415	0.086
ROC AUC	0.71	0.580	0.323	0.521	0.086

Table C.12: Performances of Grading LR with Clinical+Portal(Core) features with Forward Selection

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.71	0.690	0.188	0.684	0.043
SPECIFICITY	0.96	0.927	0.170	0.921	0.052
SENSITIVITY	0.23	0.240	0.377	0.197	0.081
PRECISION	0.76	0.297	0.446	0.572	0.218
PR AUC	0.55	0.646	0.273	0.467	0.092
ROC AUC	0.66	0.603	0.305	0.554	0.093

Table C.13: Performances of Grading LR with Clinical+Portal(Core) features with Stepwise Selection

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.73	0.710	0.165	0.692	0.046
SPECIFICITY	0.97	0.950	0.126	0.936	0.054
SENSITIVITY	0.24	0.250	0.364	0.189	0.086
PRECISION	0.80	0.330	0.454	0.611	0.242
PR AUC	0.54	0.638	0.269	0.461	0.099
ROC AUC	0.63	0.595	0.314	0.548	0.096

Table C.14: Performances of Grading LR with Clinical+Portal(Core) features with Ridge Regression

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.72	0.637	0.193	0.635	0.057
SPECIFICITY	0.92	0.847	0.197	0.832	0.076
SENSITIVITY	0.33	0.230	0.349	0.229	0.102
PRECISION	0.68	0.287	0.424	0.412	0.165
PR AUC	0.62	0.614	0.266	0.415	0.086
ROC AUC	0.71	0.580	0.323	0.521	0.086

Table C.15: Performances of Grading LR with Clinical+Portal(Core) features with Lasso Regression

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.74	0.645	0.178	0.646	0.060
SPECIFICITY	0.96	0.867	0.194	0.844	0.075
SENSITIVITY	0.32	0.220	0.334	0.239	0.100
PRECISION	0.79	0.287	0.424	0.441	0.178
PR AUC	0.61	0.618	0.265	0.436	0.091
ROC AUC	0.70	0.587	0.318	0.546	0.087

Table C.16: Performances of Grading LR with Clinical+Portal(Core) features with Principal Components Regression

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.68	0.640	0.177	0.637	0.048
SPECIFICITY	0.93	0.887	0.182	0.869	0.068
SENSITIVITY	0.20	0.170	0.326	0.158	0.076
PRECISION	0.59	0.210	0.388	0.396	0.201
PR AUC	0.53	0.587	0.247	0.392	0.082
ROC AUC	0.67	0.543	0.298	0.525	0.083

Table C.17: Performances of Grading LR with Clinical+Portal(Core+Margin) features

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.79	0.628	0.202	0.641	0.056
SPECIFICITY	0.93	0.793	0.252	0.789	0.076
SENSITIVITY	0.51	0.320	0.384	0.334	0.113
PRECISION	0.78	0.317	0.393	0.443	0.119
PR AUC	0.75	0.641	0.277	0.459	0.083
ROC AUC	0.83	0.598	0.328	0.587	0.077

Table C.18: Performances of Grading LR with Clinical+Portal(Core+Margin) features with Forward Selection

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.74	0.724	0.190	0.706	0.048
SPECIFICITY	0.92	0.900	0.175	0.880	0.061
SENSITIVITY	0.39	0.360	0.400	0.348	0.107
PRECISION	0.71	0.430	0.458	0.598	0.138
PR AUC	0.63	0.668	0.283	0.531	0.092
ROC AUC	0.73	0.642	0.328	0.649	0.076

Table C.19: Performances of Grading LR with Clinical+Portal(Core+Margin) features with Stepwise Selection

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.74	0.715	0.182	0.710	0.044
SPECIFICITY	0.93	0.900	0.175	0.890	0.054
SENSITIVITY	0.73	0.330	0.395	0.338	0.098
PRECISION	0.71	0.393	0.452	0.610	0.138
PR AUC	0.59	0.683	0.274	0.531	0.086
ROC AUC	0.73	0.662	0.308	0.673	0.069

Table C.20: Performances of Grading LR with Clinical+Portal(Core+Margin) features with Ridge Regression

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.79	0.628	0.202	0.641	0.056
SPECIFICITY	0.93	0.793	0.252	0.789	0.076
SENSITIVITY	0.51	0.320	0.384	0.334	0.113
PRECISION	0.78	0.317	0.393	0.443	0.119
PR AUC	0.75	0.641	0.277	0.459	0.083
ROC AUC	0.83	0.598	0.328	0.587	0.077

Table C.21: Performances of Grading LR with Clinical+Portal(Core+Margin) features with Lasso Regression

Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.76	0.658	0.197	0.659	0.054
SPECIFICITY	0.93	0.825	0.242	0.811	0.070
SENSITIVITY	0.41	0.340	0.380	0.347	0.111
PRECISION	0.76	0.367	0.419	0.478	0.119
PR AUC	0.72	0.684	0.267	0.497	0.088
ROC AUC	0.82	0.678	0.274	0.633	0.075

Table C.22: Performances of Grading LR with Clinical+Portal(Core+Margin) features with Principal Components Regression

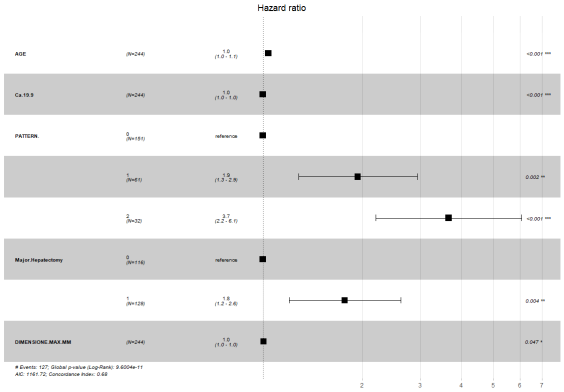
Metrics	Entire Dataset	Cross-Validation 1		Cross-Validation 2	
		Mean	Std	Mean	std
ACCURACY	0.75	0.657	0.183	0.647	0.054
SPECIFICITY	0.91	0.835	0.195	0.805	0.073
SENSITIVITY	0.43	0.310	0.386	0.321	0.105
PRECISION	0.70	0.327	0.407	0.456	0.122
PR AUC	0.66	0.650	0.283	0.458	0.073
ROC AUC	0.77	0.620	0.314	0.599	0.072

D | Appendix D

In this Appendix the Hazard Ratios with 95% CI and Estimates of the Baseline Survival Curves of the various Cox-PH models for OS and RFS are reported.

Figure D.1: HR with 95% CI and estimated baseline survival curves of Cox-PH model for OS with Clinical Preoperative features

(a) Hazard Ratio



(b) Estimated Baseline Survival Curves

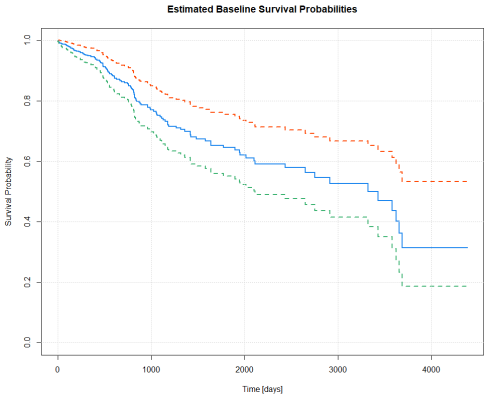
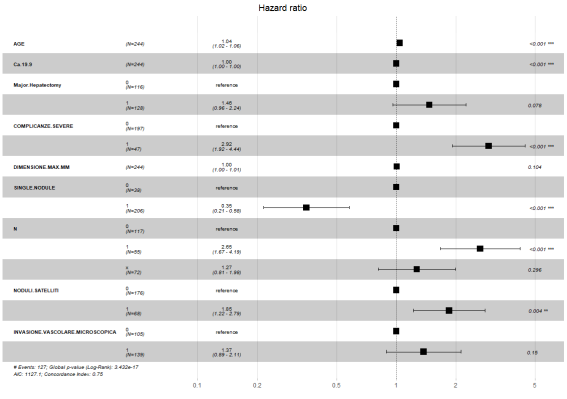


Figure D.2: HR with 95% CI and estimated baseline survival curves of Cox-PH model for OS with Clinical Postoperative features

(a) Hazard Ratio



(b) Estimated Baseline Survival Curves

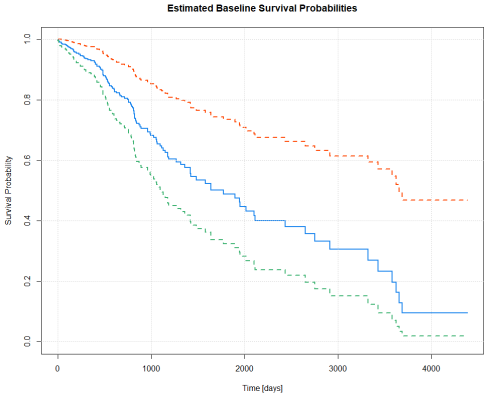
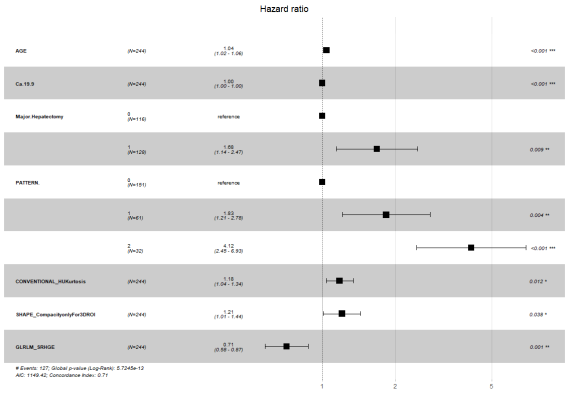


Figure D.3: HR with 95% CI and estimated baseline survival curves of Cox-PH model for OS with Clinical Preoperative + Portal(Core) features

(a) Hazard Ratio



(b) Estimated Baseline Survival Curves

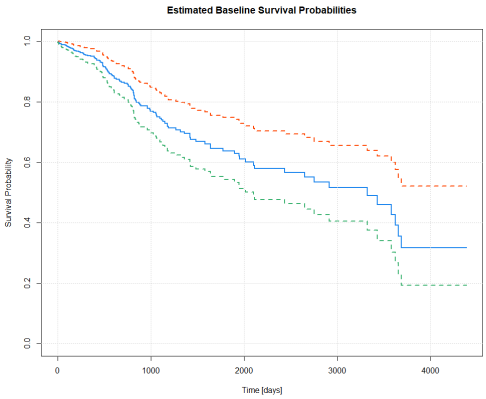
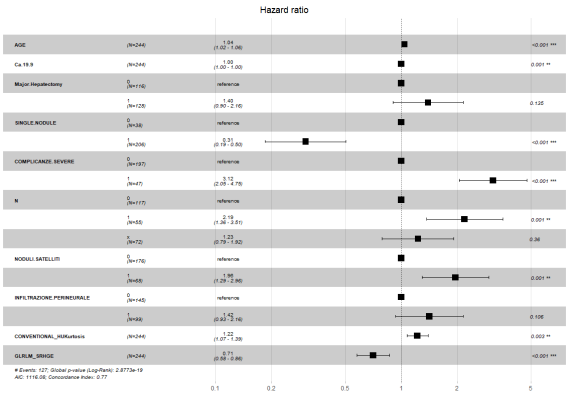


Figure D.4: HR with 95% CI and estimated baseline survival curves of Cox-PH model for OS with Clinical Postoperative + Portal(Core) features

(a) Hazard Ratio



(b) Estimated Baseline Survival Curves

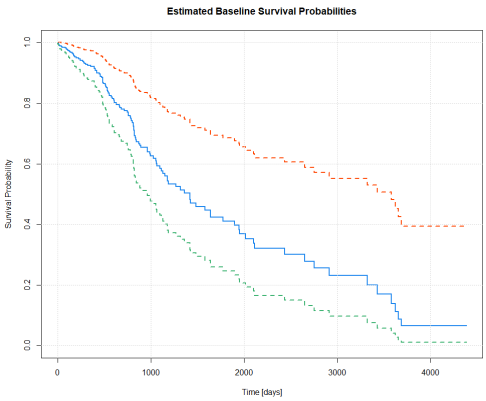
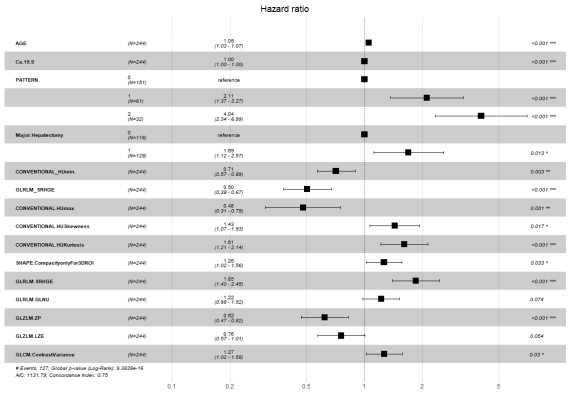


Figure D.5: HR with 95% CI and estimated baseline survival curves of Cox-PH model for OS with Clinical Preoperative + Portal(Core+Margin) features

(a) Hazard Ratio



(b) Estimated Baseline Survival Curves

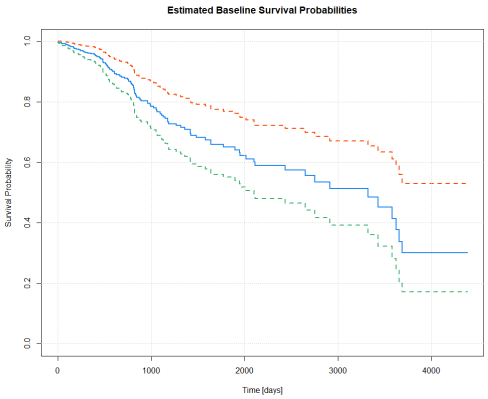
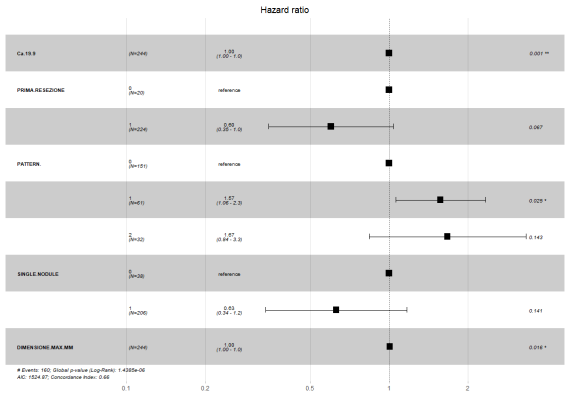


Figure D.6: HR with 95% CI and estimated baseline survival curves of Cox-PH model for RFS with Clinical Preoperative features

(a) Hazard Ratio



(b) Estimated Baseline Survival Curves

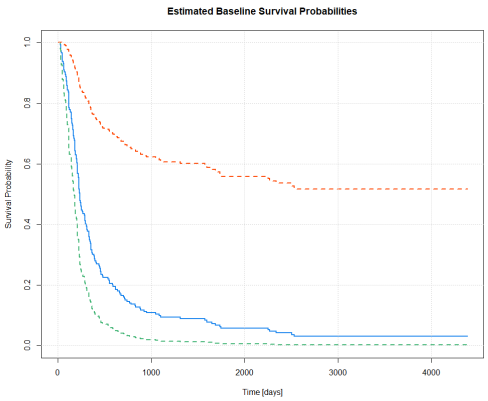
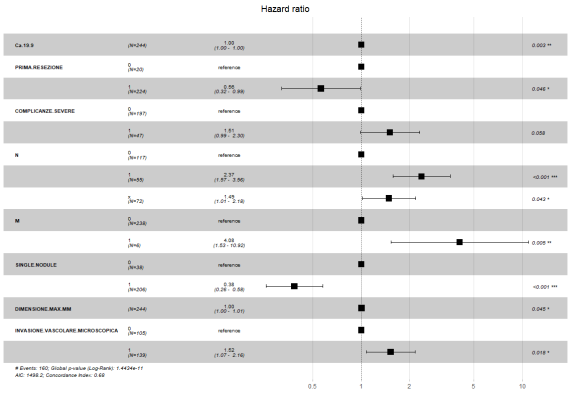


Figure D.7: HR with 95% CI and estimated baseline survival curves of Cox-PH model for RFS with Clinical Postoperative features

(a) Hazard Ratio



(b) Estimated Baseline Survival Curves

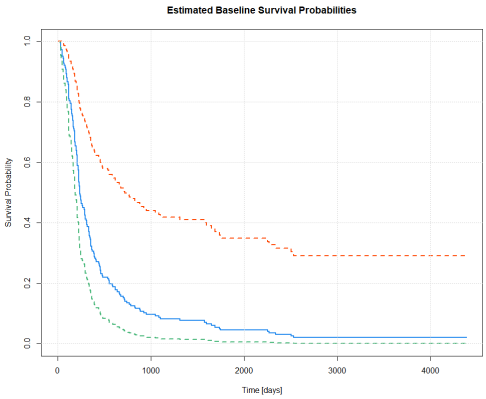
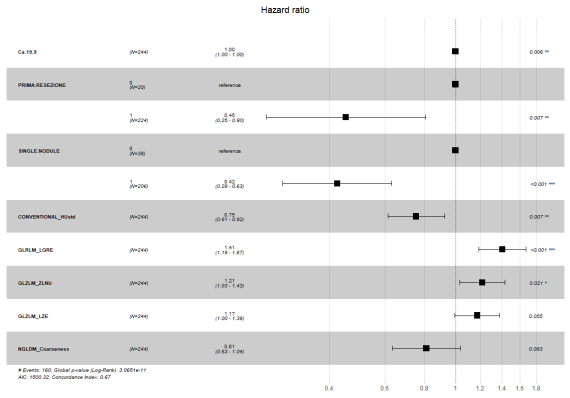


Figure D.8: HR with 95% CI and estimated baseline survival curves of Cox-PH model for RFS with Clinical Preoperative +Portal(Core) features

(a) Hazard Ratio



(b) Estimated Baseline Survival Curves

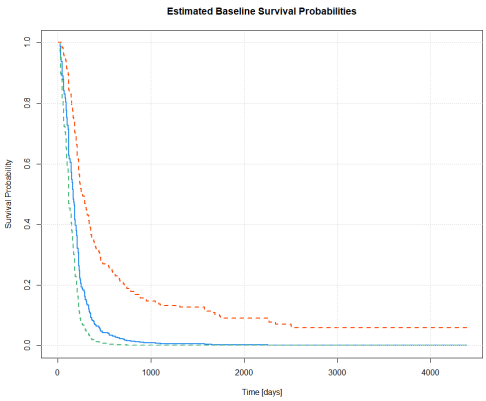
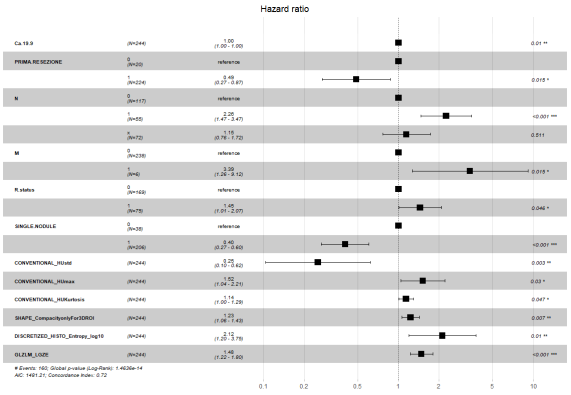


Figure D.9: HR with 95% CI and estimated baseline survival curves of Cox-PH model for RFS with Clinical Postoperative +Portal(Core) features

(a) Hazard Ratio



(b) Estimated Baseline Survival Curves

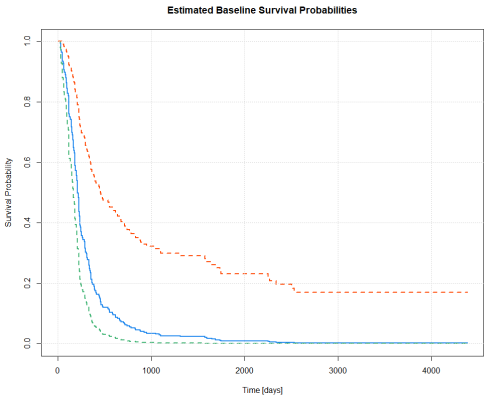
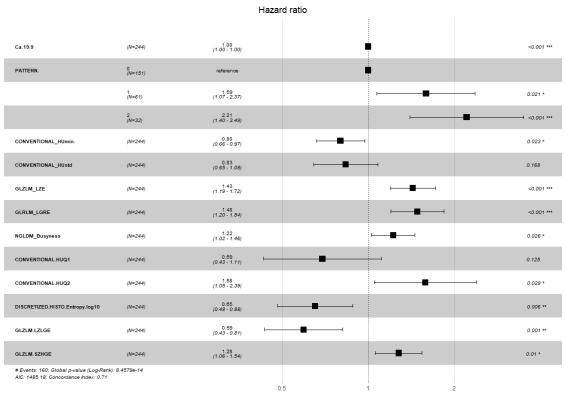
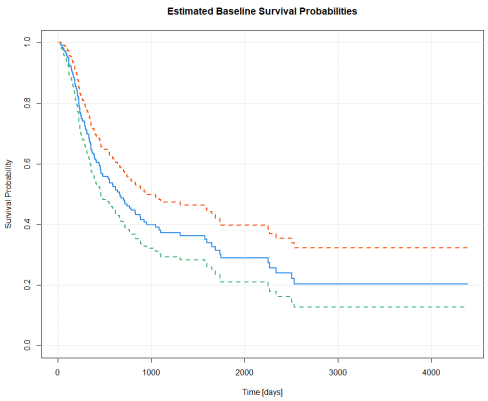


Figure D.10: HR with 95% CI and estimated baseline survival curves of Cox-PH model for RFS with Clinical Preoperative +Portal(Core+Margin) features

(a) Hazard Ratio



(b) Estimated Baseline Survival Curves



List of Figures

List of Tables

Ringraziamenti

Finalmente questo lungo, intenso ed emozionante viaggio giunge al termine e desidero ringraziare tutte le persone che mi sono state accanto e che mi hanno aiutato in questo percorso.

Innanzitutto, ringrazio la Professoressa Francesca Ieva e il Dottor Luca Viganò per avermi dato la possibilità di cimentarmi in questo stimolante lavoro, che mi ha permesso di imparare moltissimo e di accrescere le mie competenze. Vi ringrazio per il supporto che mi avete sempre dimostrato, che mi ha permesso di appassionarmi sempre di più a questo lavoro, cercando di dare sempre il meglio. Vi ringrazio soprattutto per la disponibilità e per il tempo che mi avere dedicato, rispondendo sempre con cortesia a tutti i miei dubbi e le mie domande. E' stato un piacere lavorare con voi.
Un grazie va anche al Dottor Francesco Fiz per la sua appassionante lezione sulla Radiomica.

Ringrazio di cuore i miei genitori, che mi hanno sostenuto in questi anni di studio. Grazie per aver sempre creduto in me, per non avermi mai fatto mancare nulla e per il bene che ogni giorno mi dimostrate. Ogni traguardo che ho raggiunto è in parte merito vostro; con i vostri preziosi insegnamenti e lezioni di vita avete contribuito a rendermi la persona che sono oggi. Vi sono grata per tutto.

Inoltre ringrazio tutti i parenti, nonne, zii e cugini per avermi sempre mostrato affetto e interesse durante questo percorso.

Ringrazio anche tutte le persone che in questi cinque anni ho incontrato qui al Poli che, in diversi modi, hanno fatto parte di questa esperienza. Dai conoscenti agli amici, dai compagni di progetto ai professori, ognuno è riuscito a lasciarmi qualcosa. Ringrazio particolarmente Tommi, persona alla quale mi sono legata di più in questi anni di università. Grazie per aver reso più divertenti e meno noiose le giornate in università con la tua compagnia.

Infine, un grazie speciale va a Francesco, compagno di vita, che è sempre stato presente in questo percorso, nei momenti belli, ma soprattutto in quelli più brutti, nei quali mi ha dato la forza per non mollare e andare avanti. Grazie per esserci sempre stato, condividendo con me ogni istante. Mi hai sopportato e sostenuto ogni volta che avevo bisogno di aiuto o di conforto e hai gioito con me per ogni singolo traguardo raggiunto. Sono felice di averti al mio fianco.

Grazie a tutti.
Noemi