Aus der Radiologischen Klinik der Universität Heidelberg

(Geschäftsführender Direktor: Prof. Dr.med. Dr.rer.nat. Jürgen Debus)

Abteilung für RadioOnkologie und Strahlentherapie

(Ärztlicher Direktor: Prof. Dr.med. Dr.rer.nat. Jürgen Debus)

In Zusammenarbeit mit der Klinischen Kooperationseinheit Translationale Radioonkologie am Deutsches Krebsforschungszentrum (DKFZ)
(Leiter: Prof. Dr.med. Dr.rer.nat. Amir Abdollahi)

Development of Machine Learning-based Spatially and Temporally Resolved 4D Radiomics in Radiation Oncology

Inauguraldissertation

zur Erlangung des Doctor scientiarum humanarum

an der

Medizinischen Fakultät Heidelberg

der

Ruprecht-Karls-Universität

vorgelegt von

Francesco Sforazzini

aus

Italien

November, 2021

Dekan:	Herr Prof. Dr.med. Hans-Georg Kräusslich
Doktorvater:	Herr Prof. Dr.med. Dr.rer.nat. Jürgen Debus

Contents

List of Abbreviations **v**

List of Figures **viii**

List of Tables **ix**

1 Introduction **1**
1.1 Radiomics in medicine 1
1.2 Typical radiomics workflow 2
1.2.1 Data collection 2
1.2.2 Data curation and pre-processing 3
1.2.3 Region of interest segmentation 4
1.2.4 Features extraction and selection 4
1.2.5 Model building and validation 5
1.3 Current challanges in radiomics analyses 5
1.3.1 Data curation 5
1.3.2 Lack of standardization and reproducibility 7
1.4 Translation: immune monitoring using radiomics 9
1.4.1 Link between PD-L1 and tumor response 9
1.4.2 Current pitfalls in PD-L1 expression quantification strategy 10
1.4.3 Assess PD-L1 expression using radiomics 10

2 Aims **12**

3 Material & Methods **14**
3.1 Cohorts 14
3.1.1 Primary GBM 14

3.1.2 Recurrent high grade glioma (rHGG) . . . 15
3.1.3 Pre-clinical mouse lung CT . . . 15
3.1.4 Non-small-cell lung cancer . . . 16
3.1.5 Anal adenocarcinoma . . . 16
3.1.6 Head and neck squamous cell carcinoma . . . 16
3.1.7 TCGA cohorts . . . 16
3.2 Data sorting and curation . . . 17
3.2.1 Folder sorting . . . 17
3.2.2 BP-Class: Body part classification . . . 18
3.2.3 MR-Class: MR image sequence classification . . . 21
3.2.4 DICOM metadata-based radiotherapy sorting and final session labelling . . . 21
3.3 Data pre-processing and radiomics features extraction . . . 22
3.3.1 RADIANTS workflows . . . 23
3.3.2 Gross tumor volume segmentation in high grade gliomas 28
3.3.3 Deep learning-based automatic lung segmentation for healthy and fibrotic lungs in mice from multi-resolution CT . . . 30
3.3.4 Metrics used for quantitative segmentation evaluation . 33
3.4 PD-L1 R: a MR based surrogate for PD-L1 expression in Glioblastoma multiforme . . . 34
3.4.1 Data collection and pre-processing . . . 34
3.4.2 Features selection and PD-L1 R model training . . . 35
3.4.3 PD-L1 R model evaluation . . . 36
3.5 Tumor immune evolution determined by dynamic PD-L1 R in Glioblastoma patients treated with post-surgical radio-chemotherapy . . . 37
3.5.1 Dataset description and pre-processing . . . 37
3.5.2 PD-L1 R "High" longitudinal evolution as prognostic biomarker . . . 37
3.5.3 PDL1R-Dyn model training and evaluation . . . 38
3.5.4 PDL1R-Dyn correlation with overall survival . . . 38

4 Results **39**
4.1 PyCURT: Python-based tool for radiotherapy imaging data curation . . . 41
4.1.1 CT body part classification (BP-Class CT) . . . 41
4.1.2 MR body part classification (BP-class MR) . . . 49

4.1.3 MR image contrast classification (MR-Class) 51
4.1.4 Radiotherapy data curation and sessions labelling . . . 51
4.2 RADIANTS: Radiomics analysis framework for radiotherapy studies 57
4.2.1 Gross tumor volume segmentation in high grade gliomas, from pre-surgical and follow up imaging time points . . 59
4.2.2 Deep learning-based automatic lung segmentation for healthy and fibrotic lungs in mice from multi-resolution CT 65
4.3 PD-L1 R: a MR based surrogate for PD-L1 expression in Glioblastoma multiforme 77
4.3.1 PD-L1 R results 77
4.3.2 Association between PD-L1 R and overall survival . . . 80
4.3.3 Association between PD-L1 R and additional molecular covariates 81
4.4 Tumor immune evolution determined by dynamic PD-L1 R in Glioblastoma patients treated with post-surgical radiochemotherapy 86
4.4.1 PD-L1 R-Dyn model results 87
4.5 Cohorts description 91
4.5.1 Primary glioblastoma 91
4.5.2 Recurrent high-grade glioma 92
4.5.3 Pre-clinical mouse lung CT 92
4.5.4 Non-small-cell lung cancer 93
4.5.5 Anal adenocarcinoma 94
4.5.6 Head and Neck Squamous Cell Carcinoma 95

5 Discussion **96**

6 Summary **108**

7 Zusammenfassung **110**

8 References List **112**

9 Personal Publications **130**

Curriculum Vitae **132**

Eidesstattliche Versicherung **133**

List of Abbreviations

Abd-Pel: Abdominal pelvic

ADC: Apparent Diffusion Coefficients

AUC: Area Under the Curve

BP: Body Part

CI: Confidence Interval

CNN: Convolutional Neural Network

CRC: ColoRectal Cancer

CT: Computer Tomography

DD: Dose Distribution

DICOM: Digital Imaging and Communications in Medicine

DSC: Dice Score Coefficient

FI: Fibrotic Index

FLAIR: Fluid Attenuated Inversion Recovery

FU: Follow Up

GBM: Glioblastoma

GTV: Gross Tumor Volume

HD95: Hausdorff Distance 95^{th} percentile

HNC: Head and Neck

HNSCC: Head and Neck Squamous Cell Carcinoma

HU: Hounsfield Unit

IQR: Interquartile Range

LASSO: Least Absolute Shrinkage and Selection Operator

MGMT: Methylation status of the O(6)-methylguanine-DNA methyltransferase

MR: Magnetic Resonance

NSCLC: Non-Small-Cell Lung Cancer

PD-1: Programmed Cell Death Protein 1

PD-L1: Programmed Cell Death Protein Ligand 1

PET: Positron Emission Tomography

RCHT: Radio-chemotherapy

rHGG: Recurrent High-Grade Glioma

RNA: RiboNucleic Acid

ROC: Receiver Operating Characteristic

ROI: Region Of Interest

RT: Radiotherapy

RTK: Receptor Tyrosine Kinase

SGD: Stochastic Gradient Descent

SOP: Service-Object Pair

SS: Structure Set

SVZ: Sub-Ventricular Zone

TCGA:The Cancer Genome Atlas

TCIA: The Cancer Imaging Archive

TME: Tumor Micro-Environment

TOST: Two One-Sided Test

TP: Time Point

UID: Unique Identifier

List of Figures

4.1 Standard PyCURT workflow 42
4.2 ROC curves BP-Class CT 3D vs 2D network architecture . . . 43
4.3 BP-Class CT saliency maps 45
4.4 Results for the head-neck BP-Class CT classifier 46
4.5 Results for the lung BP-Class CT classifier 47
4.6 Results for the abdominal-pelvic BP-Class CT classifier 48
4.7 ROC curves for the two BP-Class MR classifiers 49
4.8 BP-Class MR saliency maps 50
4.9 Results for the abdominal-pelvic and head-neck BP-Class MR classifiers 50
4.10 Radiotherapy data sorting overview in PyCURT 51
4.11 Region of interest extraction from the RT structure set in PyCURT 52
4.12 Session labelling flowchart 53
4.13 RADIANTS flowchart 58
4.14 Example of a features extraction workflow in RADIANTS . . 59
4.15 Example of a Nipype execution graph 60
4.16 GTV segmentation results 61
4.17 Example of low accuracy GTV segmentation 62
4.18 Comparison between GTV segmented using two channels (T1ce and FLAIR contrasts) vs one channel (T1ce) 63
4.19 Example of miss-segmented GTV in the follow up time point . 64
4.20 Mouse lung segmentation network architecture 66
4.21 Overview of the workflow used for training, validation and testing a U-net CNN for lung segmentation 67
4.22 Dice score coefficients and Hausdorff distances distributions in the test set 68

4.23 Representative segmentation results of the CNN for a random mouse from the test set 69
4.24 Overview of the transfer learning approach 72
4.25 Results of the re-trained network using a transfer learning approach 73
4.26 Pipeline used to quantify the spatial mismatch 74
4.27 Results of the spatial analysis of mismatches 75
4.28 Equivalence analysis results 76
4.29 Overview of the glioblastoma cohort and the inclusion criteria adopted 78
4.30 Analysis pipeline used to identify and evaluate the radiomics signature associated with PD-L1 in pre-surgery time points . . 79
4.31 PD-L1 R training and validation results 80
4.32 PD-L1 R results in the whole set 81
4.33 PD-L1 R results obtained with different MR contrasts 82
4.34 Association between PD-L1 R and overall survival 83
4.35 Association between PD-L1 R and molecular covariates 84
4.36 RNA sequencing-based PD-L1 distributions 84
4.37 PD-L1 R and MGMT interaction 85
4.38 Rationale and implementation of the dynamic PD-L1 R model 89
4.39 Overall survival between late and early switch PD-L1 R groups 90

List of Tables

4.1 AUC BP-Class CT classifiers 44
4.2 Median probabilities BP-Class CT 45
4.3 Summary of the images found in the different probability ranges for the BP-Class CT classifiers 54
4.4 AUC BP-Class MR classifiers 55
4.5 Median probabilities BP-Class MR 55
4.6 Summary of the images found in the different probability ranges for the BP-Class MR classifiers 56
4.7 Qualitative analysis of the GTV segmentation in follow up time points 64
4.8 Multi-raters vs network analysis 65
4.9 Inter-rater variability analysis 65
4.10 Segmentation results on an independent test set 68
4.11 PD-L1 R radiomics signature 77
4.12 Logistic regression model parameters 80
4.13 Radiomics signature associated to early/late switch in pre-surgical time point 87
4.14 PD-L1 R-Dyn validation results 87
4.15 Cox proportional hazard model results 88
4.16 Primary glioblastoma cohort overview 91
4.17 Recurrent high-grade glioma cohort overview 92
4.18 Pre-clinical mouse lung CT cohorts overview 92
4.19 Non-small-cell lung cancer cohort overview 93
4.20 Anal adenocarcinoma cohort overview 94
4.21 Head and Neck Squamous Cell Carcinoma cohort overview . . 95

Chapter 1

Introduction

1.1 Radiomics in medicine

Radiomics (Aerts et al., 2014) is a uprising field of research concerned with the extraction of quantitative metrics—the so-called radiomics features —from medical images. Radiomics changed the perception of medical images, which rapidly evolved from being a mere diagnostic tool to play a pivotal role in the context of personalized medicine (Goetz and Schork, 2018).
Radiomics features capture normal tissue and tumor characteristics, such as heterogeneity and shape, that may be used for clinical problem solving and to improve medical decision making (Lambin et al., 2017).
Radiomics features can be hand-crafted or automatically extracted from computed tomography (CT), magnetic resonance (MR) and/or positron emission tomography (PET). These features are combined with other pertinent data, such as medical records, pathology, biology, genomics or dose information from radiotherapy (dosiomics) to produce robust models. These models can then be used to tailor the treatment to the single patient and improve the final outcome.
Lately, deep learning-based radiomics started to be investigated. Using this approach, the whole image is fed into the network and it is automatically identify the most important features based on the outcome (Lohmann et al., 2020). These methods are promising since they require less human input, however the lack of interpretability currently limits their translation into clinical workflow.
Radiomics analysis has been already successfully applied to study different

type of cancer. For example, in one cornerstone radiomics paper from Aerts et al. (Aerts et al., 2014) authors were able to link radiomics features to overall survival in a cohort of patients diagnosed with lung cancer. Since then, several other papers used radiomics to study different types of cancer, including head and neck (Ou et al., 2017), esophageal (Van Rossum et al., 2016), rectal (Liu et al., 2017; Lovinfosse et al., 2018) and brain cancer. In the latter, efforts focused on differentiating between low and high grade glioma (Skogen et al., 2016; Tian et al., 2018), characterizing the tumour molecular sub-types (i.e. IDH1 mutant vs wild type) of low-grade gliomas (Jakola et al., 2018; Han et al., 2019) and discriminating between treatment effects (for example radiation necrosis) and tumour recurrence (Hu et al., 2012; Tiwari et al., 2016).
Despite the promising results obtained so far, a number of issues have to be solved before radiomics could be applied routinely in clinical environments.

1.2 Typical radiomics workflow

There are several steps to build a radiomics analysis workflow (Lambin et al., 2017; van Timmeren et al., 2020; Papanikolaou et al., 2020), and they can be summarize as following:

- Data collection.
- Data curation and pre-processing.
- Region of interest segmentation.
- Features extraction and selection.
- Model building and validation

1.2.1 Data collection

To build robust and reliable radiomics models, a large set of high quality images is needed. This is particularly true, for example, when an outcome of interest has to be model longitudinally over time. This requires a huge set of images (multiple subjects over multiple time points) to have enough statistical power to obtain meaningful results.
Given the vast number of images required, data are often collected from

multiple clinical institutions. This results in images coming from different vendors and/or with different acquisition parameters that, if not properly accounted, can impair the final radiomics results (Lambin et al., 2017; Mackin et al., 2015).

1.2.2 Data curation and pre-processing

This is one of the most crucial step and needs to be performed rigorously in order to create a relevant radiomics model. Imaging data are usually stored and shared as Digital Imaging and Communications in Medicine (DICOM), which is the standard format for the communication and management of medical imaging information and related data.
Data curation refers to the process of identifying several key elements from the DICOM header (usually called metadata); those elements include the image modality (CT, MR, PET) and, in case of MR images, the contrast (i.e. T1w or FLAIR), and the anatomical region scanned (extremely important to choose the best pre-processing pipeline). Other parameters, like slice thickness, matrix size and image resolution can also be extracted from the header and stored for subsequent processing. Furthermore, DICOM files' integrity has to be assessed to filter corrupted files.
Beyond the images, other sources of data can be leveraged to create more complex, heterogeneous models, such as clinical phenotypes from electronic health records (Bibault et al., 2016). Finally, the curated dataset needs to have a well-defined structure before the actual analysis can begin.

Data pre-processing comes immediately after curation and provide a way to reduce the impact of image variability in the radiomics workflow. There are multiple approaches that can be used, depending on the image modality and on the anatomical region. For example, if brain MR images are considered, brain extraction and magnetic field inhomogeneity correction are usually performed; while for CT images intensity thresholding can be applied to remove unwanted voxels from the images. Furthermore, if more than one modalities is combined together, image registration has to be performed accurately. In conclusion, it is evident that pre-processing plays an important role and currently there is no available tool that could help building a flexible and reproducible pre-processing pipeline.

1.2.3 Region of interest segmentation

Before any feature extraction could be performed, a volume of interest needs to be defined. In most cases, this volume is the tumor itself, although sometimes the normal tissue can be examined as well. Segmentation can be both manual or (semi-) automatic.

When using manual approach, at least two independent segmentations should be performed in order to identify robust features that do not depend on the operator (Haarburger et al., 2020). Nevertheless, this rarely happens, for example, during the radiotherapy planning phase when tumor delineation is usually done by only one radiographer.

When possible, automatic segmentation, for example using deep learning, should be preferred to minimize human rater variability, a well-known problem in different segmentation tasks (Kim et al., 2016; Riegel et al., 2006; Weltens et al., 2001; Bø et al., 2017). However, automatic methods have often specific requirements (for example the majority of the brain tumor segmentation algorithms require four MR contrasts to work (Menze et al., 2015)) that are not always easy to satisfy and/or they are confined to a limited number of segmentation tasks.

1.2.4 Features extraction and selection

Segmented regions are then used to extract the radiomics features from the pre-processed images. Those features can be classified into three categories:

- first order features: describing the distribution of voxel intensities within the region of interest (ROI).
- shape features: describing the three-dimensional size and shape of the ROI.
- texture features: describing the variation of the intensities within the ROI. They are related to intra-tumoral heterogeneity.

All feature classes, with the exception of shape, can be calculated on either the original image and/or on a derived image, obtained by applying one of several filters. For example, the original images can be decomposed using wavelet filter or Laplacian of Gaussian before running the features extraction. Despite there are multiple mathematical definitions, in the last years several efforts have been made to create a general consensus on how the features

should be calculated (Imaging Biomarker Standardization Initiative (IBSI) (Zwanenburg et al., 2020)).

The number of features that can be extracted is virtually unlimited. The inclusion of all possible features in a model would inevitably result in overfitting, undermining the generalizability of the model to patients not previously evaluated. For this reason, the next step is to identify robust and relevant features for the specific model under evaluation. This part is called features selection. Several dimensionality reduction techniques (like principal component analysis and clustering) are available and the best method depends on the analysis pipeline that will be later applied to create the model (Parmar et al., 2015).

1.2.5 Model building and validation

There are countless methodologies to create a model with radiomics. From regression to deep learning, each has its own advantages and pitfalls (Bibault et al., 2016). It has been shown that the model choice affects prediction performance in radiomics (Parmar et al., 2015). For this reason, multiple modelling methodology implementations should be ideally implemented. When this is not possible, the key aspect in the selection of a modelling methodology is that, when reported, the work must be entirely reproducible.
The last step is the model evaluation on unseen data. Preferably its generalizability should be assessed on external cohorts, i.e. on data coming from different centers and/or acquired with different protocols. If this is not possible, then other approaches like k-fold cross validation or leave-one-out, could also be used. The quality of the model is often reported using the receiver operating characteristic (ROC) curve or the area under the ROC curve (AUC), which quantifies the sensitivity and specificity of the model. AUC are typically reported with their confidence interval and are between 0 (worst) and 1 (best) (Lambin et al., 2017).

1.3 Current challanges in radiomics analyses

1.3.1 Data curation

As mentioned in section 1.2.2, data curation is one of the most crucial step in a radiomics analysis workflow. The dataset organization is essential in

order to maximize its shareability and integrity ensuring repeatability and reproducibility. This is even more critical when data are acquired/collected from multiple clinical institutions. Indeed, the organization of medical imaging data is not standardized, and the format in which a dataset is provided often differs between sources (Lambin et al., 2017; van Ooijen, 2019).

DICOM metadata-based curation

One common way of sorting the files is based on the DICOM metadata, i.e. using the information stored in the DICOM header. The header contains all the parameters related to a specific file, like series description, acquisition date, patient information and image parameters. Files are then organized in a structured manner based on those parameters. However, this approach made the following assumptions about the DICOM fields:

- they have to be always present.
- they have the same name across scanners and/or vendors.
- they have consistent naming conventions (e.g. the series description).
- they were inserted correctly by the human operator.

In reality, often this does not happen. For example, some metadata information might not always be exported or might be anonymized during the process (Moore et al., 2015). Series descriptions might be center specific (even with different language), and might not be representative of the specific acquisition (van Ooijen, 2019; Wang et al., 2011). Moreover, in some cases, this metadata is not consistently stored and might even be partially or entirely missing, like for example the anatomical part scanned or the contrast agent administration (Hirsch et al., 2015). As a result, the sorting has to be manually supervised, by looking at each scan and labeling it accordingly. This is extremely time consuming and prone to human error.

Deep learning-based curation

In the last decade, artificial intelligence techniques have revolutionized medical image processing. Machine learning, and in particular deep learning,

algorithms have been applied to several tasks including detection and classification of lesions, automated image segmentation, data analysis, prioritizing reporting and study triage, and image reconstruction (McBee et al., 2018; Liew, 2018; Thrall et al., 2018*a*,*b*).
In terms of data curation, all the efforts made until now focused mostly on MR contrast classification (van der Voort et al., 2021; Remedios et al., 2018). In the current approaches, convolutional neural networks, inspired by the VGG (Simonyan and Zisserman, 2015) or ResNet (He et al., 2016) architectures, are trained to distinguish between different MR contrasts, usually T1w pre and post contrast agent injection, FLAIR, T2w, proton density, diffusion and Apparent Diffusion Coefficients (ADC). Among them, only DeepDicomSort (van der Voort et al., 2021) provides additional folder sorting in addition to the MR contrast classification.

In conclusion, data curation is an open issue that has to be addressed before any meaningful radiomics analysis could be performed. Tedious and time-consuming manual curation is often needed in order to get reliable results. Efforts have been made to automatize the whole process but they are still limited to MR contrast classification. There are no existing methods dealing with the identification of the anatomical region scanned or that take into account radiotherapy data, hampering the application of radiomics analyses in radiation oncology.

1.3.2 Lack of standardization and reproducibility

One of the major drawback that prevents radiomics to be applied in clinical routines is the lack of standardization and reproducibility. In particular, as briefly introduced in the previous section (1.2), there is no clear indications on the pre-processing and on the region of interest segmentation.
Each center develops its own radiomics analysis pipeline, which is not always publicly available, thus hard to reproduce and reuse. This behaviour has the effect of slowing down the progresses in the whole radiomics research field.

The pre-processing steps are highly dependent on the dataset and the research question under investigation. For example brain and abdominal MR would undergo different pre-processing, as well as CT and MR brain images. The current path is to develop a study-specific pre-processing pipeline using different software. This is highly inefficient and prone to errors. Furthermore, by doing that, it is hard to keep track of the whole process and of the tools

used.

Another step in a radiomics pipeline that suffers from lack of standardization is the region of interest segmentation. As pointed out in section 1.2.3, manual segmentation is the most often encountered method but it has several drawbacks. Firstly, it is time-consuming, depending on how many images and datasets have to be segmented. Second, it introduces a considerable observer-bias, and studies have shown that many radiomics features are not robust against intra- and inter-observer variations concerning ROI delineation (Baeßler et al., 2019). Consequently, studies using manual segmentation should perform assessments of intra-and inter-observer reproducibility of the derived radiomic features and exclude non-reproducible features from further analyses.

Deep learning-based image segmentation (often using some sort of U-Net (Weng and Zhu, 2021)) is rapidly emerging and many different algorithms have already been trained for image segmentation tasks of various organs, several of them being published as open-source. Automated image segmentation certainly is the best option, since it avoids intra- and inter-observer variability. Nonetheless, generalizability of trained algorithms currently is a major limitation, and applying those algorithms on a different dataset often results in complete failure. nnUNet is a recently developed software that provides a flexible framework for medical image segmentation (Isensee et al., 2021). It can be trained to segment any region of interest with minimal manual intervention.

However, all the existing segmentation methods are standalone tools that need to be integrated manually within a radiomics pipeline.

To conclude, there is the urge of an abstract and flexible framework where it would be easy to add/change/remove the pre-processing steps to the analysis pipeline, based on the user application. A framework where existing or new automatic image segmentation methods could be integrated with minimal efforts. This would facilitate the whole radiomics processing and would allow to easily repeat the entire workflow.

1.4 Translation: immune monitoring using radiomics

Understanding the tumor immune micro-environment is extremely important given its central role in tumor response to anti-cancer therapies like immune checkpoint inhibitors. Inflammation of the tumor micro-environment (TME) can be pro-tumor or anti-tumor depending on the immune context (Fridman et al., 2017). The presence of certain surface markers and gene expression profiles also determines the tumor immune state. For instance, high expression of PD-L1 by cells in the TME associates with better response to PD-1 blockade, whereas tumors without expression of PD-L1 respond poorly (Tumeh et al., 2014; Herbst et al., 2014).

1.4.1 Link between PD-L1 and tumor response

Programmed cell death protein 1 (PD-1) and its ligand (PD-L1) play a fundamental role in controlling the induction and maintenance of immune tolerance within the tumor micro-environment. The activity of PD-1 and PD-L1 are responsible for T cell activation, proliferation, and cytotoxic secretion in cancer to degenerating anti-tumor immune responses (Han et al., 2020). Anti-PD 1 checkpoints blockade immunotherapies have been recently developed to increase anti-tumor immune responses. To date, PD-1/PD-L1 checkpoint blockade therapy is part of the standard therapy for multiple malignancies, including melanoma, non-small-cell lung cancer (NSCLC), head and neck squamous cell carcinoma (HNSCC), colorectal cancer (CRC) and is intensively being investigated in clinical trials for the treatment of additional malignant conditions (Wu et al., 2019). However, there are still some limitations. For example, the clinical and prognostic value of PD 1/PD-L1 checkpoint blockade therapy in glioblastoma remains controversial.

PD-L1 Expression in Glioblastoma

Glioblastoma (GBM) is the most aggressive brain tumor, with high morbidity and mortality. Despite current multi-modalities treatment efforts, including maximal surgical resection, followed by a combination of radiotherapy and/or chemotherapy, the median survival is only about 15 months after diagnosis. There is an increasing interest in the correlation between PD-L1 expression

and overall survival in GBM. Despite that, to date there are no clear evidences of the PD-L1 role in GBM prognosis. For example, Liu et al. reported that the expression of PD-L1 in patients with glioblastoma could be a possible indicator for poor prognosis (Liu et al., 2013). Similar results were showed in other studies (Lee et al., 2018; Pratt et al., 2019). However in a retrospective study, as well as a TCGA database analysis comprising 446 glioblastoma patients, researchers did not find a significant connection between PD-L1 and the overall survival (Berghoff et al., 2015). Furthermore, the majority of the current studies on the efficacy of the immune checkpoint inhibitors in GBM are conducted in recurrent setting (Brahm et al., 2020) and only one phase II clinical trial investigated their effect in newly diagnosed glioblastoma (Schalper et al., 2019).

1.4.2 Current pitfalls in PD-L1 expression quantification strategy

One of the most important open issue with immunotherapies targeting PD-L1 is the variability in test cutoffs and standards for PD-L1 testing. Immunohistochemistry tests are typically used to detect the presence of the PD-L1 protein (Udall et al., 2018). However, based on the tumor under consideration, the cutoff, i.e. the threshold to consider a tumor PD-L1 positive or not, varies considerable. For example, the cutoff for the non–small cell lung cancer (NSCLC) is 50%, whereas in the triple-negative breast cancer it drops to 1% cutoff (Walk et al., 2020). And those cutoffs depend also on which antibody chosen to perform the test.
In glioblastoma, in particular, assessment of the PD-L1 expression is still under investigation, as there is no approved cutoff yet (Hao et al., 2020).

Moreover, all the existing studies are predominantly based on one single time-point (usually pre-treatment) due to the scarcity of repeated biopsies, necessary to perform immunohistochemistry tests. This restricts the dynamic tracking of the tumor immune response.

1.4.3 Assess PD-L1 expression using radiomics

In the last few years, radiomics started to be considered a powerful resource to study the tumor biology, and in particular the tumor immune microenvironment (Leijenaar et al., 2018).

This was due to the increase number of clinical images available and because it can capture the full spatial organization of tumors, that could overcome the critical limitations of sampling from biopsies. In addition, imaging data have the potential to be used longitudinally to monitor the tumor response to treatments.

Radiomics has been already applied to develop a predictive model to identify patients who may benefit from immunotherapy in non-small cell lung cancer (Wen et al., 2021), or to stratify patients, diagnosed with hepatocellular carcinoma, based on the tumor immuno-oncological characteristics (Hectors et al., 2020).
However, to date no link has been found between imaging features and PD-L1 expression in glioblastoma, highlighting the urge of performing additional research in order to exploit the full radiomics potential.

To summarize, immunotherapy is an extremely promising treatment option in clinical oncology. Checkpoints blockade therapy targeting the PD 1/PD-L1 pathway has shown encouraging results in the treatment of different tumor types. However, one of the most common and aggressive brain tumor, primary glioblastoma, has shown little or no benefit from it.
Radiomics analysis could be applied to fill in the gap and finally disentangle the relationship between glioblastoma immune status and patient prognosis. Furthermore, given the large availability of medical images, radiomics could be used to non-invasively monitor the longitudinal changes of the tumor immune biology which might help to discover novel glioblastoma immune-subgroups.

Chapter 2

Aims

The aim of this thesis was to develop a robust and flexible infrastructure for radiomics analysis, that could automatically take care of the steps from data curation to image pre-processing and segmentation. Radiomics analysis was then applied to characterize the tumor immune biology and its response to radio-chemotherapy in primary glioblastoma.
In details, in terms of analysis infrastructures, the following contributions were made:

- Implementation of a python-based data curation tool to automatically sort the imaging data most commonly available in clinical radiation oncology studies. This tools made use of both content- and metadata-based information to:

    1. Identify the anatomical region scanned in both MR and CT images.
    2. Classify the MR image contrast.
    3. Sort and label the radiotherapy information.

- Development of a python-based abstract framework to easily handle the both pre-processing and region of interest segmentation. Two new image segmentation tools, for brain gross tumor volume and pre-clinical lungs, were developed and integrated in the framework.

The previously described software were then applied to:

- Link the PD-L1 expression to imaging features to non-invasively characterize the tumor immune micro-environment in primary glioblastoma.

- A feasibility study to use radiomics analysis to monitor the tumor immune response after radio-chemotherapy, to identify possible evolution trajectories that might help to discover novel glioblastoma immune-subgroups.

Chapter 3

Material & Methods

3.1 Cohorts

3.1.1 Primary GBM

A cohort of 320 patients diagnosed with primary glioblastoma (GBM) tumor has been collected from the Heidelberg Ion Therapy center (HIT) and the University Clinic Heidelberg (UKHD). This cohort contained CT, MR and radiotherapy data. CT and MR images from this cohort were used to train the head-neck body part classification networks described in section 3.2.2.
A sub-set of 142 patients had radiotherapy and pre-irradiation planning MR time-points with at least T1 contrast enhanced (T1ce) and FLAIR MR contrasts. This sub-set was used to develop the method in section 3.3.2.
A further sub-set of 121 patients, with GBM sub-type defined as RTK I, RTK II or mesenchymal, were used in the PD-L1 analyses (sections 4.3 and 4.4).
Finally, 115 subjects among the 121 had at least one follow up (FU) time point after the first irradiation. The total number of FU timepoints was 665 (median per subject=4, min=1, max=27). Those images were used in section 3.3.2 and section 4.4.
See Table 4.16 for a detailed description of the image parameters.

Molecular biomarkers

The sub-set of 121 subjects had additional information regarding the patient's molecular background. In particular, the following biomarkers were

collected:

- PD-L1 M: Methylation-based quantification of the Programmed death-ligand 1 (PD-L1) expression (Knoll et al., 2020).
- PD-1 M: Methylation-based quantification of the Programmed cell death protein 1 (PD 1) expression (Knoll et al., in preparation).
- SVZ: DNA Methylome-based classification of the subventricular zone (Knoll et al., in preparation).
- Prognosis: Methylome based classification of glioblastoma prognosis (Knoll et al., 2019).
- GBM sub-type: DNA methylation-based classification of glioblastoma tumor (Capper et al., 2018).
- MGMT: Methylation status of the O(6)-methylguanine-DNA methyltransferase (MGMT) gene, as assessed by the MGMT-STP27 model (Bady et al., 2016).

3.1.2 Recurrent high grade glioma (rHGG)

A cohort of 115 patients diagnosed with Recurrent high grade glioma tumor (rHHG, grade 3 and 4) has been collected from the Heidelberg Ion Therapy center (HIT). This cohort was used to develop the method in section 3.3.2. See Table 4.17 for a detailed description of the image parameters.

3.1.3 Pre-clinical mouse lung CT

Clinical CT

A total of 323 CT (cohort A) and 41 CT (cohort B) images were retrospectively collected from multiple existing projects at HIT. Those CT included both healthy and fibrotic mice (irradiated with photons or carbon). All the details regarding the cohort can be found in the original publications (Zhou et al., 2017). These cohorts were used to develop and validate the pre-clinical lung segmentation method described in section 3.3.3. A overview of the main CT features are summarized in Table 4.18.

High resolution micro-CT

16 micro-CT healthy mouse CT (cohort C) were provided by the University of Maastricht and used in the transfer learning re-training of the deep learning method described in section 3.3.3 to segment lungs from high resolution mouse CT images. This cohort was firstly described in (Granton et al., 2014). A summary of the main image parameters are summarized in Table 4.18.

3.1.4 Non-small-cell lung cancer

215 patients diagnosed with non-small-cell lung cancer (NSCLC) were collected from an existing project at HIT. The whole cohort was used to train the lung body part classification networks described in section 3.2.2. A subset of 10 patients was used in the transfer learning re-training of the deep learning method described in section 3.3.3 to segment lungs from human CT (cohort D in the method section). Cohort details are reported in Table 4.19.

3.1.5 Anal adenocarcinoma

238 patients diagnosed with anal adenocarcinoma were collected from an existing project at HIT. This cohort consisted of both MR and CT images that were used to train the abdominal-pelvic body part classification networks described in section 3.2.2. 1899 MR images were used for training the abdominal-pelvic MR classifier while 814 images (30% of the total 2713) were used for the final evaluation. Cohort details are reported in Table 4.20.

3.1.6 Head and neck squamous cell carcinoma

A cohort of 162 patients with Head and neck squamous cell carcinoma (HNSCC) was provided within the German cancer consortium initiative (DKTK). An overview in reported in Table 4.21.

3.1.7 TCGA cohorts

Four cohorts coming from the The Cancer Imaging Archive (TCIA) (Clark et al., 2013) were collected and used to evaluate the body part classification networks described in section 3.2.2. The cohorts used were:

1. TCIA-HNSCC (Grossberg et al., 2018). 252 patients (54 female and 198 male), for a total of 1121 CT images, were analyzed. Patients were acquired between 1998 and 2013. The median age was 58 y ($43 < 50y,\ 112\ 51y - 69y,\ 97 > 70y$). These images were used to evaluate the head-neck body part CT classifier.

2. TCIA-NSCLC Radiomics (Aerts et al., 2014). 422 patients (132 female and 290 male), for a total of 587 CT images, were analyzed. Patients were acquired between 2005 and 2014. The median age was 61 y ($18 < 50y,\ 200\ 51y - 69y,\ 204 > 70y$). These images were used to evaluate the lung body part CT classifier.

3. TCIA-CT Colonography (Johnson et al., 2008). 309 patients (165 female and 144 male), for a total of 469 CT images, were analyzed. Patients were acquired between 2005 and 2006. The median age was 58 y ($0 < 50y,\ 256\ 51y - 69y,\ 53 > 70y$). These images were used to evaluate the abdominal-pelvic body part CT classifier.

4. TCIA-GBM. 256 patients (155 female and 101 male), for a total of 2117 MR images, were analyzed. Patients were acquired between 2000 and 2006. The median age was 57 y ($50 < 50y,\ 145\ 51y - 69y,\ 61 > 70y$). These images were used to evaluate the head-neck body part MR classifier.

3.2 Data sorting and curation

All the cohorts were sorted and curated using PyCURT, a tool developed as part of this thesis.
PyCURT builds upon NiPype (Gorgolewski et al., 2011), a well-known Python package that is used to create complex analysis workflows. An overview of PyCURT's workflow can be seen in Figure 4.1.

3.2.1 Folder sorting

The first step in a PyCURT workflow is to re-arrange the DICOM files into a well-defined hierarchy, given an input folder with custom data structure. The session (or time point, TP) and subject ID (subID) are retrieved respectively from the StudyDate and PatientID DICOM attributes, if present

otherwise are extracted from the sub-folders in the input directory. Next, the DICOM objects are sorted based on their "Modality" attribute. The considered modalities are: Computed Tomography (CT), Magnetic Resonance (MR), Positron-emission tomography (PET), and the different radiotherapy (RT) objects, i.e., the RTSTRUCT, RTPLAN, and RTDOSE. Next, the different DICOM objects are uniquely identified and split into different object directories via the Study and the Series Instance Unique Identifiers (UID) attributes, and the Series Description attribute. DICOM file integrity checks are then performed to remove all corrupted and duplicate files, localizer scans, and the structures in the RT Structure Set (SS) are renamed to remove special characters.

3.2.2 BP-Class: Body part classification

Preliminary sorted images (both MR and CT) are then classified based on the scanned body part. Five one vs. all binary ResNet18 classifiers (3 for CT and 2 for MR images) were trained to distinguish between three different scanned entities: head and neck (HNC), lungs (only for CT data) and abdominal-pelvic (abd-pel). All other possible entities are labeled as 'unclassifiable'. BP-Class was trained on 11714 CT and MR images, and validated on multiple independent cohorts (5108 CT and MR images).

Data preparation

The data used to develop this classifier were collected from multiple existing cohorts, both coming from our institution and from publicly available databases. For a complete overview please refer to section 3.1. The body part scanned in each CT was manually labelled by an expert rater.
Both CT and MR images shared the pre-processing, with the exception of the final intensity normalization. The common steps encompassed image reorientation to radiological convention and Otsu thresholding to find a coarse mask for the foreground voxels. Then, the CT images were clipped to the 0.5 and 99.5 percentile to keep the quantitative information provided by the voxels intesities, which are linked to the physical properties of the tissue. On the other hand, MR images were z-scored by subtracting the mean of the non-zero voxels and dividing by their standard deviation.

Convolutional neural network architecture and training

The Convolutional neural network (CNN) architecture used for all the classifiers in BP-Class was as of the ResNet-18, where an extra SoftMax layer appended to the last layer to produce probabilistic predictions of the classes. These networks were trained using a Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a learning rate scheduler that start with 0.01 and decays by 0.5 every epoch. Because of the one vs. all classification approach, the classes are imbalanced. To deal with the imbalanced data in each CNN, a weighted binary categorical cross-entropy loss was used, where the weights of a class are equal to the size of the largest class divided by the size of that specific class.

Each network for the CT body part classification was trained in both a 2D and a 3D fashion, to find the best structure for this specific task.

For the 3D training, a random patch of 128x128x128 voxels was cropped in each image and fed to the network. For the 2D training, a random slice from one random direction (axial, sagittal or coronal) was extracted from the 3D image and then randomly cropped before feeding into the network. The patch size in this case was 400x400 voxels. In both cases, if the size of the image was smaller than the patch size, a zero padding was performed before the cropping.

Data augmentation was randomly performed on each image before cropping (or padding) using the *batchgenerators* package in Python (Isensee et al., 2020). Possible augmentations included spatial transformations (including rotation, elastic deformations, scaling and cropping), mirroring along all axes, non-linear intensity values transformation (using gamma function), adding of Gaussian noise and image blurring.

The batch size was 5 and 150 for the 3D and 2D training, respectively. The number of training steps was 200 for both (160 for the validation phase). The number of epochs was set to 50.

The two networks for the MR body part classification (MR are usually not acquired for lung examination) were trained only in a 2D architecture given the better results obtained in the CT image classification (see section 4.1.1). The training time was approximately 3 hours per network in the 3D case and 1.5 hours in the 2D.

Body part inference on unseen images

New unseen images were firstly pre-processed in the same way described in the previous section (3.2.2). The inference was then performed on each slice (for the 2D networks) or on each patch (for the 3D). Specifically, for the 2D architecture, each slice in between the 25^{th} and the 75^{th} percentile of the total number of slices in each direction was extracted and sequentially fed to each classifier. The slices were padded or cropped to match the patch size used during training (400x400 voxels). Every single classifier labelled the slice as the body part of interest or as "other". If more than one classifier labelled the same slice, then the network having the highest activation in the last layer was taken as the body part for that slice.
Subsequently, the probability of each class was calculated as the number of slices labelled as that class divided by the total number of the labelled slices. For example, if an image had 100 slices and 70 were labelled as lung and 30 as head and neck, that image would be finally classified as 70% lung and 30% hnc.
For 3D networks a similar concept was used, with the only difference that the input image was splitted into multiple 3D patches (of 128 voxels isotropic) before feeding into each network. The probability was then calculated in the same way, as the number of patches labelled as one class divided by the total number of labelled patches.

BP-Class performance evaluation

BP-Class was evaluated on 2177 CT images (hnc: 1121, abd-pel: 469 and lungs: 587) and 2931 MRI (hnc: 2117, abd-pel: 814). Each image were fed to the whole stack of classifiers (three for CT and two for MRI) as described above (section 3.2.2). At the end, three (two for MR) probabilities were assigned to each image. Those probabilities were used to plot a receiver operating characteristic (ROC) curve, to show the classifiers performance. For each classifier, the Area Under the Curve (AUC) was also calculated an reported together with its 95% DeLong Confidence Interval (CI).
Furthermore, the histograms of the probabilities, as well as the median probability (and the 25^{th} - 75^{th} interquartile range, IQR) were reported for each classifier.
Finally, the saliency map for one representative subject for each network was plotted. The saliency map is a smart way to visualize deep learning models

result based on gradients (Simonyan et al., 2014).

3.2.3 MR-Class: MR image sequence classification

Similarly to BP-Class, MR image classifier (MR-Class) is a method based on Convolutional Neural Networks (CNN) that can automatically identify the contrast of a brain MR scan (Salome et al., in preparation). MR-Class consists of 5 one vs. all binary ResNet18 classifiers, where each classifier can identify a single sequence. The classifiable MR sequences are T1w, T2w, FLAIR, apparent diffusion coefficient (ADC), and susceptibility-weighted imaging (SWI). Furthermore, a binary classifier for T1w was trained to distinguish between pre- vs. post-contrast agent images.

3.2.4 DICOM metadata-based radiotherapy sorting and final session labelling

The DICOM-RT objects considered in PyCURT are the radiotherapy (RT) plan, which summarizes the RT treatment delivered, RT dose distribution (DD) and RT structure set (SS), that includes contoured structures of the patient anatomy (for example, the Gross tumor volume (GTV) and the organs at risk).
As first step, all the RT plans for one subjects were examined by PyCURT to identify the used plan. This was done by checking the DICOM metadata to find the most recent plan that had a "curative" plan intent with an "approved" status.
After the extraction of the delivered RT plan, the RT SS instance UID was retrieved from the Referenced SOP Instance UID attribute in the Referenced Structure Set Sequence attribute. Then the planning CT series instance UID, was identified from the RT SS and used to label the RT-CT.
Finally, the actual delivered dose distributions were identified using the corresponding series instances retrieved from the RT plan. For more information regarding the relationships between UIDs in RT-objects and the RT CT used in PyCURT refer to Figure 4.10.
In addition, the contoured structures in the RT SS can be extracted and stored separately as NiFTI files. Structure extraction was done in PyCURT by *plastimatch*. Before structure extraction, non-ASCII characters present in the SS contour names were removed from the SS DICOM file, to avoid

character encoding errors. There is also the option to only extract the structure with the highest overlap with the dose distribution (when present).

Session labelling

As final step, each time point (or session) for a subject was assigned to one of the following five groups: pre-RT, MR-RT, RT, post-RT and FU (follow up). This labelling is based on the guidelines provided by the German Cancer Consortium (DKTK). In order to achieve this subdivision, PyCURT started from labelling as "RT" the time point that contains the DICOM-RT objects (see above). If there was no DICOM RT, then the labelling stopped.
Then, all the sessions in a pre-defined temporal window from the first detected radiotherapy treatment are labelled as "post-RT". The default window used was 31 days, but this can be overwritten by the user. The "MR-RT" time point was defined as the nearest time point to the RT that included MR sequences (often acquired due to their superior tumor/ normal tissue contrast). The maximum allowed time difference between the two was set to 15 days, but, again, it can be changed based on the considered cohort. The RT and the RT-MR time points can also overlap. Finally, any session before the RT (or RT-MR) time point was labelled as "pre-RT", and all the others after 31 days from the radiotherapy were "FU". A detailed flowchart can be found in Figure 4.12.

3.3 Data pre-processing and radiomics features extraction

The next part of this thesis focused on developing a tool to make radiomics analysis robust and reproducible, RADIANTS.
RADiomics Image ANalysis for radioTherapy Studies (RADIANTS) is a Python-based framework, built upon Nipype, that enables easy, robust and reproducible radiomics analysis, with a focus on radiotherapy studies. It expects as input a PyCURT sorted folder, to rely on well curated datasets.
RADIANTS main part is the $BaseWorkflow$ class. $BaseWorkflow$ is an abstract class that manages all the processes within RADIANTS. Based on the requested analysis workflow (see below for those currently supported), the $BaseWorkflow$ checks whether the required inputs for one specific workflow are already present in the input directory or not. If not, it will check

the workflow's inputs dependencies to establish if the requested inputs can be generated by another RADIANTS workflow. If this can be done, they will be computed first and then the original workflow will start; if inputs cannot be generated the whole process will stop throwing a "missing input" error. The summary of how RADIANTS manage the workflows can be seen in Figure 4.13.

3.3.1 RADIANTS workflows

RADIANTS currently supports the following workflows:

- MR image pre-processing, involving brain extraction (using HD-BET (Isensee et al., 2019)) and N4 bias field correction (using ANTs (Tustison et al., 2010)).
- Multi-contrasts MR image registration (using ANTs (Avants et al., 2008)).
- Brain gross tumor volume (GTV) segmentation (using in-house trained nnUNet, see section 3.3.2).
- Pre-clinical lung segmentation (using in-house developed method, see section 3.3.3).
- Radiomics features extraction (using Pyradiomics).

However, given its modular structure, it is straightforward to change/add new workflow(s) into RADIANTS and extend it based on the user applications. An example of the code used to define the pre-processing workflow within RADIANTS can be found in Listing 3.1.

```
"Pre-processing workflow"
import nipype
from radiants.interfaces.mic import HDBet
from nipype.interfaces.fsl.utils import Reorient2Std
from radiants.workflows.base import BaseWorkflow
from nipype.interfaces.ants import N4BiasFieldCorrection


TON4 = ['T1ce', 'T1']
POSSIBLE_SEQUENCES = ['T1ce', 'T1', 'FLAIR', 'SWI', 'T2',
```

```
                      'ADC']


class PreprocWorkflow(BaseWorkflow):

    def workflow_inputspecs(self):
        #Define the workflow's inputs and their dependencies
        #here
        input_specs = {}
        input_specs['inputs'] = {
            '': {'mandatory': True,
                 'format': '.nii.gz',
                 'dependency': None,
                 'possible_sequences': POSSIBLE_SEQUENCES,
                 'multiplicity': 'all',
                 'composite': None}}
        input_specs = self.generate_inputspec(input_specs)

        return input_specs

    #The actual workflow goes here, based on Nypipe
    conventions
    def workflow(self):

        #call a RADIANTS base function to source the data.
        #This function assumes that the input directory is
        #the output of a previously executed PyCURT workflow
        datasource = self.data_source

        #assign some parameters here
        dict_sequences = self.dict_sequences
        nipype_cache = self.nipype_cache
        result_dir = self.result_dir
        sub_id = self.sub_id

        #Only MR images are used as input for this workflow,
        #CT and RT are excluded
        toprocess = {**dict_sequences['MR-RT'],
                     **dict_sequences['OT']}
        workflow = nipype.Workflow('preprocessing_workflow',
                                   base_dir=nipype_cache)
        datasink = nipype.Node(nipype.DataSink(
            base_directory=result_dir), "datasink")
        substitutions = [('subid', sub_id)]
        substitutions += [('results/', '{}/'.format(
```

```
            self.workflow_name))]
        substitutions += [('_preproc_corrected.',
                           '_preproc.')]
        datasink.inputs.substitutions = substitutions

        for key in toprocess:
            files = []
            if toprocess[key]['scans'] is not None:
                files = files + toprocess[key]['scans']
            for el in files:
                el = el.strip(self.extention)
                node_name = '{0}_{1}'.format(key, el)
                bet = nipype.Node(
                    interface=HDBet(),
                    name='{}_bet'.format(node_name),
                    serial=True)
                bet.inputs.save_mask = 1
                bet.inputs.out_file = '{}_preproc'.format(el)
                reorient = nipype.Node(
                    interface=Reorient2Std(),
                    name='{}_reorient'.format(node_name))
                if el in TON4:
                    n4 = nipype.Node(
                        interface=N4BiasFieldCorrection(),
                        name='{}_n4'.format(node_name))
                    workflow.connect(
                        bet, 'out_file', n4, 'input_image')
                    workflow.connect(
                        bet, 'out_mask', n4, 'mask_image')
                    workflow.connect(
                        n4, 'output_image', datasink,
                        ('results.subid.{0}.@{1}_preproc'
                        .format(key, el)))
                else:
                    workflow.connect(
                        bet, 'out_file', datasink,
                     'results.subid.{0}.@{1}_preproc'.format(
                         key, el))
                workflow.connect(
                    bet, 'out_mask', datasink,
                    ('results.subid.{0}.@{1}_preproc_mask'
                     .format(key, el)))
                workflow.connect(reorient, 'out_file',
                                 bet, 'input_file')
                workflow.connect(datasource, node_name,
```

```
                                    reorient, 'in_file')

        return workflow

    def workflow_setup(self, create_database=True,
                       dict_sequences=None, **kwargs):

        return BaseWorkflow.workflow_setup(
            self, create_database=create_database,
            dict_sequences=dict_sequences, **kwargs)
```

Listing 3.1: Python code for the pre-processing workflow in RADIANTS

When adding a new workflow, the first thing to do is to define the *workflow_inputspecs*. The most important part are the *inputs*. There the user can list all the inputs required by the workflow. This is a dictionary where the keys (when present) represent the suffix that RADIANTS is going to append to the *possible_sequences* (if provided). The *possible_sequences* are especially useful when a particular workflow has to be applied only to specific inputs (for example, in Lst. 3.1 the pre-processing workflow can only be applied to the "POSSIBLE_SEQUENCES" images). If no *possible_sequences* are provided (i.e. an empty list is provided) then the workflow will run using all the images found in the input directory (if the key is present, only the images that end with that string will be considered). The *multiplicity* parameter tells RADIANTS in which sub-folder, for a given subject, look for the specified image. In particular, this parameter can be "mrrt", "rt" or "all". This distinction is based on the labelling performed by PyCURT (see section 3.2.4). If "mrrt" is specified, only the inputs in the "MR-RT" labelled folder (if present) will be considered; if "rt" then only the files in the "RT" folder will be checked. Finally, if "all" is selected, then all the images in all the sub-folders are checked.
Mandatory informs RADIANTS whether the specified inputs must be present or not; if mandatory is True and the input cannot be found or generated, then the process will stop. *Format* specifies the format of the input (for example ".nii.gz" for compressed Nifti). *Composite* has to be set when the considered key needs more than one input in order to be generated. For example, the gross tumor volume segmentation (section 3.3.2) requires both T1 post-contrast agent injection and FLAIR contrasts in order to work, so in this case $composite = [T1ce, FLAIR]$.
Lastly, in the *dependency* the user can specify which other RADIANTS workflow can be used to generate the key in case the file is not available in the

input directory. In case the input needs to be computed, RADIANT will check the dependency's *workflow_inputspecs* to determine whether it can be generated or not.

Pre-processing workflow

As seen in the previous section, one of the workflow implemented in RADIANTS is the one for image pre-processing.
In particular, it takes all the MR images (that match the "POSSIBLE_SEQUENCES") in one session and performs first the brain extraction using HD-BET (Isensee et al., 2019), then it corrects for inhomogeneities of the main magnetic field using N4 bias field correction in ANTs (Tustison et al., 2010). The latter is only performed on T1 weighted images as it might distort the signal from white matter hyper-intensities in T2w FLAIR images (Valdés Hernández et al., 2016).

Multi-contrast image registration workflow

After pre-processing, the images are registered together. Specifically, all the images in one session are linearly registered to the corresponding T1ce image; if T1ce is not present, the other two possible references are T1 and T2, in this order.
In addition, if the "MR-RT" time point is available for one subject, the user can choose to non-linearly register each reference image for all the other sessions to the T1ce (or T1, or T2) from the MR-RT. All the other images are subsequently normalized to the MR-RT reference.
Finally, if the "RT" and the "MR-RT" time points are both accessible, the reference in the MR-RT time point can be linearly registered to the planning CT in the RT session. Then all the images can also be normalized to the planning CT space.
All the registrations are performed using ANTs (Avants et al., 2008).

Radiomics imaging features calculation using Pyradiomics

Radiomics imaging features are extracted using Pyradiomics (Van Griethuysen et al., 2017).
Images and masks are first resampled to have 1 mm isotropic resolution, using a third order B-Spline as interpolation for the images (nearest neighbour for the masks). To discretize the images, a bin-width of 10 is used.

By default, features are extracted from the original image and from the image filtered using both the wavelet and the Laplacian of Gaussian (LoG, with sigma equal to 1, 2, 3, 4 mm) functions.

Another pivotal part in a standard radiomics workflow is the region of interest segmentation. The segmentation must be accurate in order to extract meaningful imaging features.
As part of RADIANTS, two type of segmentation workflows have been developed using artificial intelligence:

- Gross tumor volume (GTV) segmentation from T1ce and FLAIR MR contrasts
- Lung segmentation from multi-resolution CT images

3.3.2 Gross tumor volume segmentation in high grade gliomas

The 257 subjects with pre-irradiation MR time points used to train and test this methods came from the cohorts descried in 3.1.1 and 3.1.2. The subset of 115 subjects that had a total of 665 follow up (FU) MR time points after the first irradiation came entirely from the primary GBM cohort (3.1.1).

Data curation and pre-processing

All the imaging data were curated using PyCURT (section 3.2). PyCURT returned all the MR images labelled as "brain", as well as all the radiotherapy planning CT and the linked structure set and dose distribution. The gross tumor volumes (GTV) used to train the network were automatically extracted from the structure set, using the dose distribution (when present) to identify the one used to plan the irradiation. All the images were then converted to Nifti.
RADIANTS took then care of the pre-processing and the image registration as described in sections 3.3.1 and 3.3.1. For this application, only T1ce and FLAIR images were used. In addition, the planning CT was rigidly registered to the corresponding T1ce (not skull stripped) for all the subjects and the affine matrix was applied to the GTV to normalize it to the T1ce (used as reference here) space, using the nearest neighbor interpolation.

nnUNet training and inference

nnU-Net (Isensee et al., 2021) is a powerful framework, based on a U-Net structure (Weng and Zhu, 2021), that enables fast and accurate out-of-the-box biomedical image segmentation. It automatically takes care of the pre-processing steps (like image resampling) and parameters tuning. For the training, 85% of the total number of images (220) were randomly selected from the whole set (balanced between GBM and rHHG); 37 images were left out for the final evaluation. Both T1ce and FLAIR images were used as input for the network. By default, nnUNet trains all U-Net configurations in a 5-fold cross-validation, which enables nnUNet to determine the optimal post-processing on the training dataset. For the inference, both T1ce and FLAIR contrasts, coming from the 37 subjects left out from the pre-irradiation cohort and from the FU time points, underwent the same pre-processing described above and were then fed into the network.
The same process described above was repeated using only T1ce for training/inference, to investigate whether the use of T1ce alone would have provided similar results.

Network performance evaluation

The performance on the 37 pre-irradiation time points was assessed by using both the Dice Score coefficient (DSC) and the 95^{th} percentile Hausdorff distance (HD95) metrics (see section 3.3.4 for their definitions).
The performance on the follow up time points was both quantitatively and qualitatively evaluated.
First, one human rater segmented 50 randomly selected time points among the 665 that were then compared with those provided by the network using DSC and HD95. In addition, all the segmented FU GTV were manually checked and classified into 3 different classes: "good", "usable" and "wrong".
Finally, to test the inter-rater variability, two additional raters segmented the same 50 subjects and they were compared both to each other and to the network.

3.3.3 Deep learning-based automatic lung segmentation for healthy and fibrotic lungs in mice from multi-resolution CT

The main cohorts used to train and evaluate this method (cohort A: 323 CT and cohort B: 41 CT) are described in the first sub-section in 3.1.3; the cohorts used in the transfer learning approach are described in the second sub-section in 3.1.3 (cohort C) and in section 3.1.4 (cohort D).

Data partition and preparation

291 CT images (90%) from cohort A were used for training/validation. These 291 images were split into five different folds (Kfold algorithm, training/validation ratio 80%/20%). The remaining 32 CT were used for evaluation (testing). Cohort B was used exclusively for testing. See Figure 4.21 for a graphical overview.
80% of the CT in cohort C (n=13) and cohort D (n=10) were used to re-train the model in the transfer learning approaches. The remaining 3 CTs from cohort C (and the 3 from cohort D) were used for testing. See Figure 4.24.

All the CT images were converted from DICOM to NRRD format using MITK (Wolf et al., 2004).
CT images in cohorts A and B contained 4 to 6 mice each, therefore the first step was to crop each CT scan to obtain CT image data from single mice. Coordinates of the lungs' edges were identified from the reference mask and used to find the x, y, z coordinates to crop the corresponding CT image. This resulted in a total of 1654 mice in cohort A (1500 mice for training/validation and 154 for testing) and 237 mice in cohort B. Cropped images were resized to have a 0.35 mm isotropic resolution.
The reference lung masks (ground truth) were segmented semi-automatically, as described in (Zhou et al., 2017), for both fibrotic and normal lungs. Briefly, each CT was thresholded between -900 and -100 HU to obtain an approximate lung segmentation. Subsequently, wrongly assigned voxels, mostly in the trachea and primary bronchi areas, were removed manually.
CT images in cohort C contained 1 or 2 mice each. Those with 2 mice were cropped as described before. The total number of mice in cohort C was 20 (16 used to re-train the model and 4 for testing). Each cropped CT was then

resampled to 0.2 mm isotropic. Ground-truth lung masks were segmented semi-automatically.
Human subjects in cohort D were resempled to 1 mm isotropic and the ground-truth lung masks were again segmented semi-automatically.
Finally, given the 2D architecture of the proposed method, each axial slice in the 3D images coming from all the cohorts was extracted and the intensity was normalized between 0-1.

Convolutional neural network architecture and training

The Convolutional Neural Network (CNN) was implemented with a 2D U-net architecture consisting of one "encoder", which allows capturing the global structure of the data, and one "decoder", which allows a fine-grained localization. The details of the network structure are reported in Figure 4.20. The network was trained 5 times, each time on a different fold, using Adam optimization algorithm and a learning rate of 0.0001. Binary cross-entropy was used as loss function and Dice score as metric. The batch size was fixed at 50 for training and validation. The number of epochs was set to 30. The inference was run 5 times, using one different set of weights each time. The final probability lung mask was calculated as median. Both hard thresholding (0.5) and Otsu were tested to obtain the final binary lung mask.

Transfer learning on high-resolution mouse CT and human CT images

To apply this network to segment lungs from mouse in cohort C and human subjects in cohort D, a transfer learning approach was implemented. In particular, the encoder part was frozen during the network re-training and only the decoder was further trained, to enable an accurate lung segmentation. The network was re-trained independently for each of the two cohorts.

Network performance evaluation

The Dice Score Coefficient (DSC) and the 95^{th} percentile Hausdorff distance (HD95) were used as metrics to quantify the segmentation accuracy, both in terms of spatial overlap and shape mismatch (see section 3.3.4 for their definitions).

The next part of the evaluation focused on the spatial mismatch quantification.
To find areas in the lung more sensitive to different classification, all 154 mice in the test set from cohort A were registered to a common space and then tested for deviation.
Each voxel in the difference image between the reference lung mask and the CNN segmented mask was z-scored across mice to find where the CNN segmentation were more consistently deviating from the manual segmentation. For an overview of the process refer to Figure 4.26.
To perform this analysis, a mouse template was first created. This reference template provided a standardized space, reflecting the anatomical variability within the present cohort (healthy and fibrotic mice). This step was crucial to improve the following image registration. Indeed, by simply registering to a random mice within the same cohort resulted in a complete failure, given the huge variability between the two mice.
To create the template, 20 random mice were firstly linearly registered to one common reference using FSL FLIRT with 6 degrees of freedom, corratio as cost function and trilinear interpolation. The average of these 20 registered mice was used to initialize the template calculation. The actual template was built by feeding 60 random mice, from the remain set of 134 mice, into the antsMultivariateTemplateConstruction2.sh script (Avants et al., 2010), within ANTs. Very briefly, this function first warped all the images to the initial estimate of the template using the Bspline symmetric normalization (BsplineSyN) method and a cross-correlation metric. Every registration proceeded in a multi-resolution scheme (six steps with a ratio of 1/16, 1/12, 1/8, 1/4, 1/2, 1 of the original resolution respectively). Afterwards, all the warped images were voxel-wise averaged to create a temporary template. The final template was then obtained by iterating this process several times.
All the mice in the test set were then registered to the template using ANTs with a symmetric normalization (SyN) cost function. The warps and affine matrices obtained from the previous step were then applied to the difference image between each reference lung mask and the CNN segmented mask, to bring them into the template space. The image of the difference between the two masks consisted of 1 (when the voxel was segmented in the reference mask and not in the CNN mask), -1 (when the voxel was labelled as lung by the CNN and not by the human rater) and 0 (when the voxel was segmented as lung in both masks).
Finally, the normalized difference images were merged into a 4D file and the

average across mice was performed in each voxel and then divided by the standard deviation.

Lastly, an equivalence analysis was performed. In particular, to justify the usage of the lung mask segmented using the proposed approach instead of those semi-automatically contoured, we compared the Fibrotic Indices (Zhou et al., 2017) and the histograms of the Hounsfield units within the lungs calculated using both masks. A cohort of 106 mice collected 24 weeks after irradiation were used in both analyses. The two-one-sided t-test (TOST) procedure (Lakens, 2017) was used to test for equivalence. In both cases, the standard deviation of the semi-automatically segmented group was used as the magnitude of the region of similarity (equivalence margin epsilon).

3.3.4 Metrics used for quantitative segmentation evaluation

Dice Score Coefficient (DSC)

The Dice score coefficient measures the spatial overlap between the network-segmented lung masks and the reference. The DSC was defined as:

$$DSC = \frac{2 * (Voxel_{GT} \bigcap Voxel_{SEG})}{|Voxel_{GT}| + |Voxel_{SEG}|}$$

where $Voxel_{GT}$ are the voxels in the ground truth mask, $Voxel_{SEG}$ are the voxels in the mask segmented by the proposed approach and $\bigcap$ is the intersection operator between two spatial masks.
The Dice score can range from 0 s(no spatial overlap between two images) to 1 (complete spatial overlap). However, it does not give much information about volumetric under-/over estimation. Shape mismatch is only captured if the deviation is volumetrically impactful: a thin 'hole' will not result in a large deviation from 1.

Hausdorff Distance (HD)

The Hausdorff distance (HD) between the CNN-segmented mask and the reference has been calculated as follows:

$$HD = \max(HD(GT \rightarrow SEG), HD(REF \rightarrow GT))$$

where HD is defined for two general datasets A and B as:

$$HD(A, B) = \max_{a \in A}(\min_{b \in B}(d(a, b)))$$

Where a and b are single points in the dataset A and B respectively and d is the Euclidean distance between each pair of points.
This metric is particularly sensitive to strong local deviations which do not necessarily take up large volumes, but rather result in a large shape difference.
In this work, the 95^{th} percentile of the HD was taken (instead of the maximum), as it allows to overcome the high sensitivity of the Hausdorff distance to outliers (Isensee et al., 2019; Menze et al., 2015).

3.4 PD-L1 R: a MR based surrogate for PD-L1 expression in Glioblastoma multiforme

The next part of this thesis focused on applications. In particular, the tools developed in the first part were applied to link imaging features to primary glioblastoma tumor immune response, described in terms of PD-L1 expression, and its evolution under radio-chemotherapy.

3.4.1 Data collection and pre-processing

The 121 subjects enrolled in this study are described in section 3.1.1, while the inclusion criteria are shown in Figure 4.29.
All the subjects were sorted and curated using PyCURT 3.2. After sorting, only T1ce and FLAIR MR contrasts from the MT-RT time points were kept. Those images were fed into RADIANTS that performed pre-processing (section 3.3.1) and image registration (3.3.1, in this case only FLAIR to T1ce). Then, the gross tumor volume was segmented using the network described in 3.3.2. Finally, radiomics features were calculated as described in section 3.3.1.
Each subjects had also a methylation-based quantification of PD-L1 (PD-L1 M), which has been demonstrated to be highly correlated to PD-L1 in different tumor types (Knoll et al., 2020).

3.4.2 Features selection and PD-L1 R model training

The whole cohort of 121 subjects was randomly splitted into training (86 subjects) and validation (35).
Feature selection was performed on the training set. The following steps were performed:

1. 80% of the training subjects (69) were randomly selected.

2. Linear regression fit was calculated between each feature and PD-L1 M and compared against the mean of each feature using an anova test. Only the features with a p value less than 0.05 were kept.

3. Each significant feature found in 2) was linearly regressed against PD-L1 M using a 10-folds cross-validation scheme and the mean squared error (ms) was saved. At the end, only the feature with the lowest ms and all the features in a range of 1% were kept.

4. A multivariate regression model was then fitted using all the features from step 3) and backward step-wise selection was used to find the best model using AIC. The features used in the best model were finally saved.

5. All the steps from 1 to 4 were repeated 100 times, saving each time the output of step 4.

6. Finally, the feature with the highest frequency across the 100 runs and any features in a range of 30% were chosen as the radiomics signature associated to PD-L1 M.

The selected radiomics signature was then applied to split the training set into two groups. In particular, a t-distributed stochastic neighbor embedding (t-SNE) (Van Der Maaten and Hinton, 2008) approach was used to create a low-dimensional representation of the data space, to immediately visualize any clustering structure in the data. Then, K-Means was used to assign each subject to one of the possible two clusters. Two was selected as the best number of clusters based on the silhouette method. Finally, a logistic regression model was trained to classify between the two groups. This model was called PD-L1 R. A t-test was used to assess any difference in PD-L1 M in the two identified groups.

3.4.3 PD-L1 R model evaluation

PD-L1 R was then applied to the validation set (35 subjects) and the whole set (121) to split them into two groups. A t-test was again used to assess any difference in PD-L1 M.
Furthermore, features coming from other MR modalities (T1w, FLAIR, T2w) were used to investigate the ability of the model to identify two groups with different PD-L1 M expression, even with features coming from different contrasts. A total of 67 T1w, 100 T2w and 121 FLAIR images were pre-processed as described above and then rigidly registered to the corresponding T1ce image. Then, imaging features were extracted from each image using the segmented GTV and fed into the logistic regression model to assign it to one of the two groups. For each modality, a t-test was calculated between PD-L1 M values in the two groups.

Survival analysis

Difference in overall survival between the two groups identified in the whole set was further analyzed. The Kaplan-Meier estimator was used to estimate the survival function. The p value was calculated using the likelihood ratio. The Cox proportional hazard model was used to investigate the association between the survival time and the group splitting, using the PD-L1 R "High" group as reference.

Association with molecular covariates

In addition, the association between the group splitting provided by PD-L1 R and some molecular covariates has been explored. Specifically, the two groups identified in the whole set were tested for differences in PD-1, MGMT promoter methylation status, subventricular zone (SVZ), GBM tumor subtype and prognosis. Barnard's (for binary covariates) and Fisher's test (for 3-classes covariates) were used to assess the association.
Finally, to elucidate the interaction between tumor response and DNA repair mechanism, PD-L1 R was used to further stratify MGMT methylated/unmethylated patients. The Kaplan-Meier estimator and the Cox proportional hazard model were then used to study the overall survival.

3.5 Tumor immune evolution determined by dynamic PD-L1 R in Glioblastoma patients treated with post-surgical radiochemotherapy

The main idea behind this analysis and an overview of the steps performed can seen in Figure 4.38.

3.5.1 Dataset description and pre-processing

The dataset used in this longitudinal analysis consisted of a sub-set of 115 subjects, among the 121 patients previously described (section 3.4). Those subjects had at least one follow up imaging time point after first surgery with T1ce and FLAIR MRI. Furthermore, only the time points with one GTV segmented by the network described in 3.3.2 were included in this analysis. These inclusion criteria resulted in a total of 557 time points, ranging from 1 to 46; the median follow up time was 46 months.
Data curation, pre-processing and features extraction were performed in PyCURT and RADIANTS and were already described in in the previous section. Imaging features were then used to compute PD-L1 R scores, ranging between 0 and 1, the cutoff for PD-L1 R "High" (H) and "Low" (L) was set to 0.5.

3.5.2 PD-L1 R "High" longitudinal evolution as prognostic biomarker

In order to assess the prognostic value of the longitudinal evolution of PD-L1 R, the time between the pre-surgery time-point (TP0) and the time of the first switch from H to L was evaluated. Only the subjects labelled as PD-L1 R "High" in TP0 were included (n=58).
In particular, the median time, across subjects, between TP0 and the first switch was used to split the cohort of 58 patients based on their PD-L1 R dynamics:

- High-early (h-e), those subjects switching before the median time.
- High-late (h-l), those subjects switching after the median time.

- High-miss (h-m), those subjects that did not switch group in the considered time frame.

H-e and h-m were merged based on their similar overall survival (see result section 4.4), and they will be indicated only as h-e in the following sections. The Kaplan-Meier estimator and the Cox proportional hazard ratio model were used to assess any difference in the overall survival between the h-l group (used as reference in the Cox model) and the h-e group.

3.5.3 PDL1R-Dyn model training and evaluation

A treebag model, called PDL1R-Dyn, was trained on TP0 image features to differentiate between h-e and h-l.
This was done to find a radiomics signature able to predict PD-L1 R evolution using pre-surgical features. The whole cohort of 58 PD-L1 R "High" subjects was divided into training (70%, n=41) and validation (20%, n=16). Least absolute shrinkage and selection operator, LASSO (Tibshirani, 1996), was used to select the most robust features; 10-folds cross validation was also applied to stabilize LASSO and to find the optimal lambda with the smallest error. Finally, those features, i.e. the radiomics signature, was used to train a treebag model to distinguish between h-e and h-l.
PDL1R-Dyn was then tested on the validation set of 16 subjects. Area under the curve (AUC), accuracy, sensitivity and specificity were reported.

3.5.4 PDL1R-Dyn correlation with overall survival

Both univariate (using only PDL1R-Dyn to predict survival) and multivariate (using both PDL1R-Dyn and PD-L1 R) analyses were performed to investigate which model would provide the best survival predictor. The best model was selected based on the Akaike information criterion (AIC).

Chapter 4

Results

Building a robust and meaningful radiomics model usually requires to collect huge amount of data which are only available from multiple clinical centers. This lead to several issues, mostly related to:

1. Imaging data curation, often requiring manual input to handle data variability.
2. Data pre-processing, often study-dependent and difficult to reproduce.

The following python-based software were developed as part of this work to facilitate the data curation and pre-processing:

- PyCURT, a robust data curation tool, able to automatically sort and curate all the relevant imaging data.
- RADIANTS, a flexible framework that enables building custom radiomics pre-processing workflows (including region of interest segmentation) in an easy and reproducible way.

Those tools were then applied to develop an imaging-based radiomics signature to monitor the tumor immune status in a primary glioblastoma.

As summarized above, this thesis focused on two main parts: the building of two analysis infrastructures and the their application to answer to a

relevant clinical question.
To address the first part, two separate tools were developed: PyCURT and RADIANTS. PyCURT is a framework to sort and curate images coming from multiple modalities (MRI, CT and RT) and/or institutions. RADIANTS build upon PyCURT and provide an abstract environment to encapsulate all the major processing steps to run a radiomics workflow, with a focus on radiotherapy-based studies, in robust and reproducible manner.

For the second part, a cohort of 121 patients diagnosed with a primary Glioblastoma tumor was used to identify an imaging-based radiomics signature linked to the PD-L1 expression in pre-surgery T1ce MR contrast. In addition, further analyses in follow up time-points showed that the tumor immune evolution, determined by PD-L1 dynamics, is prognostic in patients treated with post-surgical radio-chemotherapy.

4.1 PyCURT: Python-based tool for radiotherapy imaging data curation

Clinical data curation, for radiomics analysis, is a challenging task. The main issues raise from different factors:

- Heterogeneous data exported from multiple centers, with different naming scheme and acquisition parameters.
- Unreliable DICOM metadata, especially regarding anatomical information and image contrast.
- Unstructured radiotherapy data often unlabelled and with center-specific names in the RT structure set.

PyCURT addresses most of those issue in a automatic and robust fashion. In particular, it provides:

- Two content-based convolutional neural networks classification methods to:
  - Identify the anatomical part images in each scan (BP-Class).
  - Identify the contrast in each brain MR scan (MR-Class).
- DICOM metadata-based radiotherapy sorting.
  - Extraction of the regions of interest from the structure set.
  - Session labelling based on the RT time-point.

A graphical representation of the different steps in a standard PyCURT pipeline can be found in Figure 4.1

4.1.1 CT body part classification (BP-Class CT)

To automatically differentiate between head-neck, lung and abdominal-pelvic CT images, a stack of three one-vs-all convolutional neural networks (CNNs) were trained.

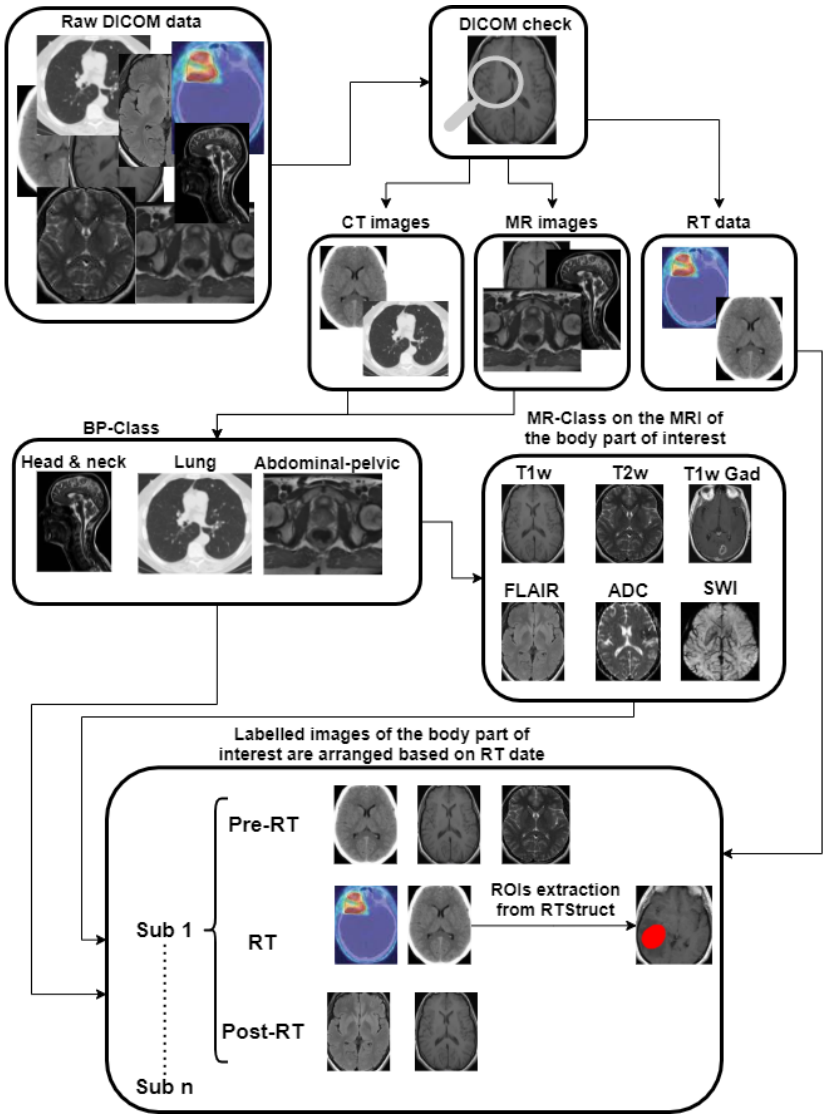


Figure 4.1: Representation of the steps included in a standard PyCURT workflow. Briefly, PyCURT takes as input a folder with unstructured DICOM data. It then checks the integrity of each file and preliminarily sort them based on their modality. Subsequently, the body part imaged and the MR contrast are classified using BP-Class and MR-Class, respectively. Finally, all the session are re-arranged in a well-defined structure and labelled based on the radiotherapy information (if present).

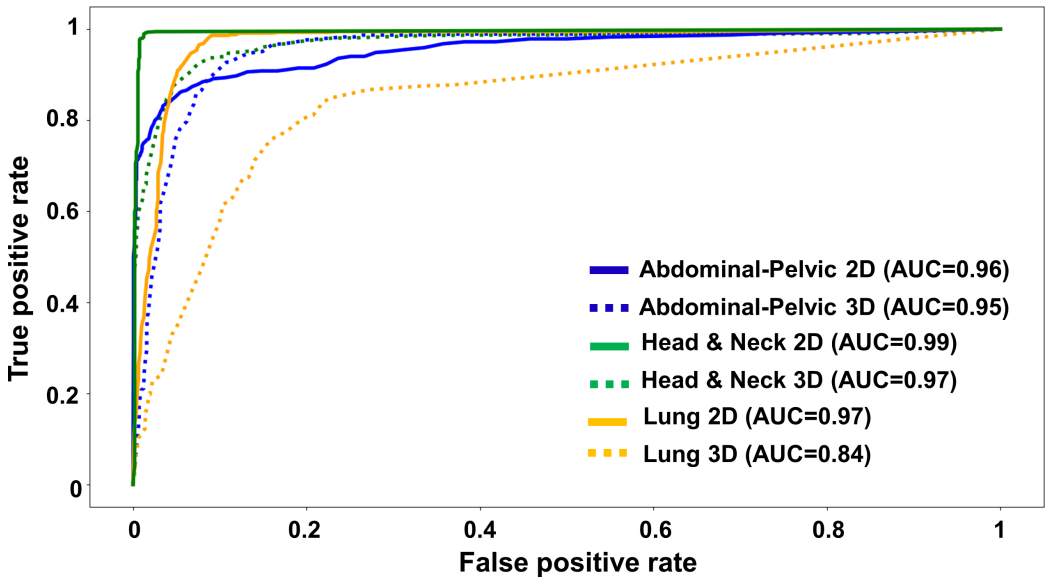


Figure 4.2: ROC curves for the three one-vs-all CNNs trained using a 2D network architecture (solid lines) and for those with a 3D architecture (dotted lines). Only the CNN trained using CT images are shown. Area Under the Curve (AUC) is also reported.

Comparison between 2D and 3D network architectures

As first step, a comparison between 2D and 3D network structure was performed in order to find the best architecture for this particular task. The 3 CNNs for the CT images were trained and tested two times on the same images. Receiver Operating Characteristic (ROC) curve was used to assess the networks performance. Results for the 2D (solid lines) and the 3D (dotted lines) testing are reported in Figure 4.2. Area under the curve (AUC), with 95% confidence interval (CI, DeLong), was used to quantify the networks accuracy. All three classifiers trained in a 2D fashion had higher AUC. Head & neck and abdominal-pelvic CNNs had similar AUC, 0.99 and 0.96 (2D case) vs 0.97 and 0.95 (3D), while the 3D-trained lung classifier had a substantial drop in accuracy (AUC=0.84) respect to the 2D scenario (AUC=0.97). For the details refers to Table 4.1.
Given the overall better performance of the networks with 2D structure, this was the chosen architecture for all the following analyses.

CNN	AUC	95%CI
abd-pel 2D	0.9558	0.9440 - 0.9678
lung 2D	0.9721	0.9649 - 0.9793
head-neck 2D	0.9940	0.9905 - 0.9976
abd-pel 3D	0.9540	0.9442 - 0.9637
lung 3D	0.8432	0.8240 - 0.8624
head-neck 3D	0.9686	0.9613 - 0.9760

Table 4.1: Results for the 3 BP-Class CT classifiers trained using a 2D (top) and 3D (bottom) structure. AUC: Area under the curve; CI: Confidence interval (DeLong).

Results BP-class CT

To characterize the three 2D CNNs for CT images, both the saliency maps and the histograms of the probabilities are reported. The saliency maps for one representative subject for each CNN are reported in Figure 4.3. As can be noted, the voxels which are more important to classify an image as lung are mainly within the lung area, while those that are contributing more to the head and neck classification lies on the skull region. Interestingly, the salient voxles for abdominal-pelvic classification are those in the interfaces between two different tissues, here between fat and organs.
The histograms of the probabilities for each CNN are reported in Figures 4.4 to 4.6. The median probabilities for the three classifiers lied between 0.84 and 0.9, with the abdominal pelvic CNN having the highest one. Each median value, with the corresponding $25^{th} - 75^{th}$ interquartile range, are reported in Table 4.2. A summary of the CT images found in the three probability ranges (0-0.33, 0.34-0.66 and 0.67-1), with a description of their main characteristics, for each BP-class CT classifier is reported in Table 4.3. Interestingly, the classifiers proved to be able to correctly classify images with uncommon features, not encountered during the training. For example, the CT with only one lung was seen as lung with a probability of 10%, while the colonography CT was classified as abdominal-pelvic (probability 12%) despite its completely different image contrast.

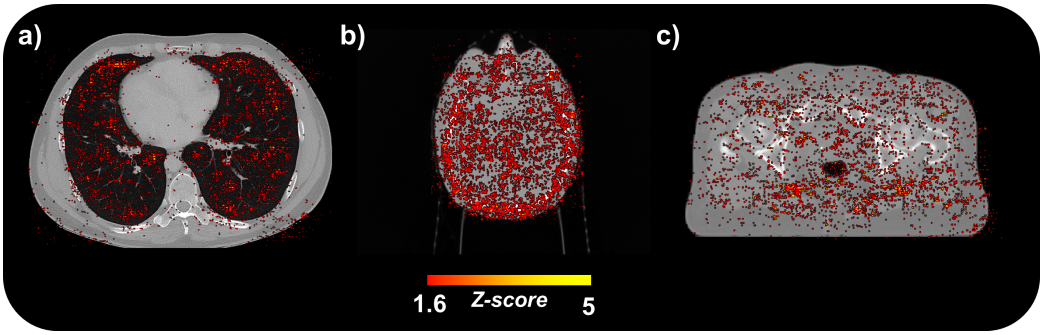


Figure 4.3: Saliency maps showing the contribution of each pixel to the final classification, in the three 2D BP-Class CT CNNs: lung (a), head-neck (b) and abdominal-pelvic(c). Brighter colors mean higher contribution.

CNN	**Mdn-P**	**IQR (25 - 75)**
head-neck	0.84	0.65 - 0.93
lung	0.84	0.74 - 0.94
abd-pel	0.9	0.52 - 1

Table 4.2: Median probabilities for the three 2D BP-Class CT CNNs. Mdn-P: median probability; IQR: interquartile range.

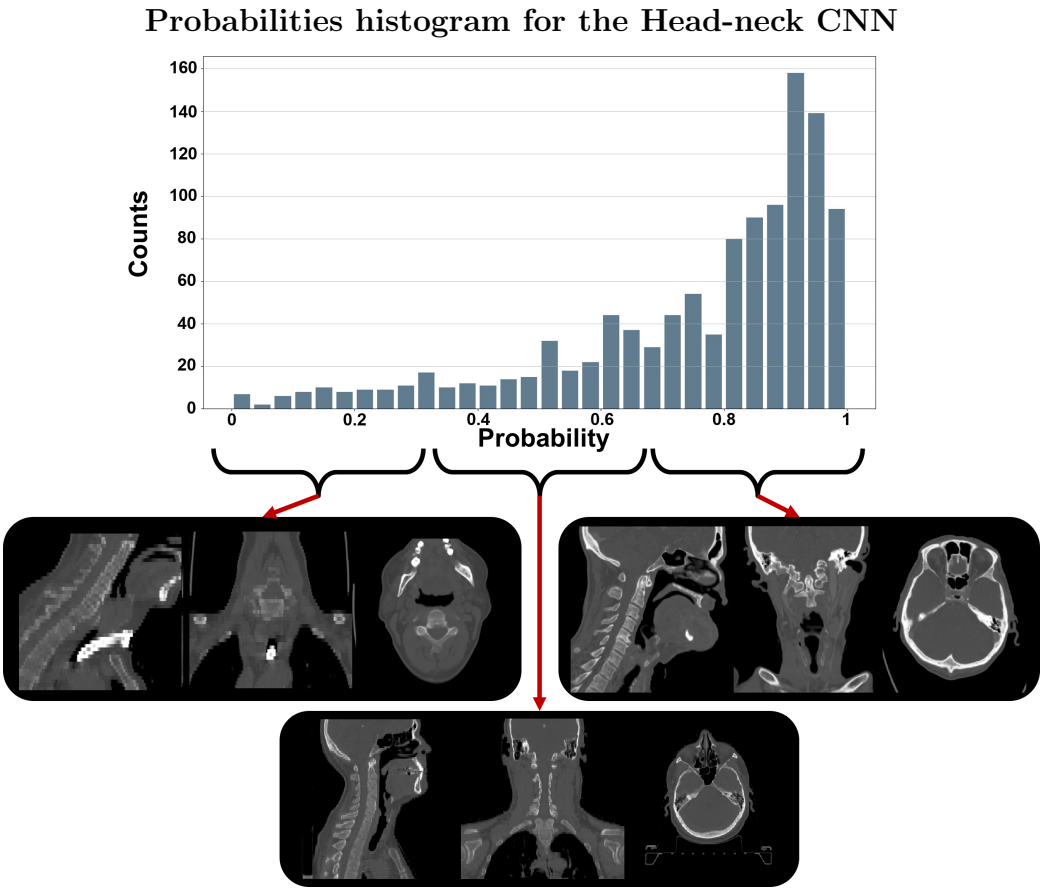


Figure 4.4: The histogram in the top graph displays the probabilities for 1121 head-neck images as inferred by the corresponding classifier. The median probability was 0.84, with 0.65-0.93 interquartile range ($25^{th} - 75^{th}$ percentile); see Table 4.2 for more information. The images in the bottom part show a representative subject belonging to the corresponding probability section (highlighted by the curly braces). The probabilities of being head-neck for the three images shown were (from left to right): 0.16 (lung 0.37), 0.57 (lung 0.22) and 1.

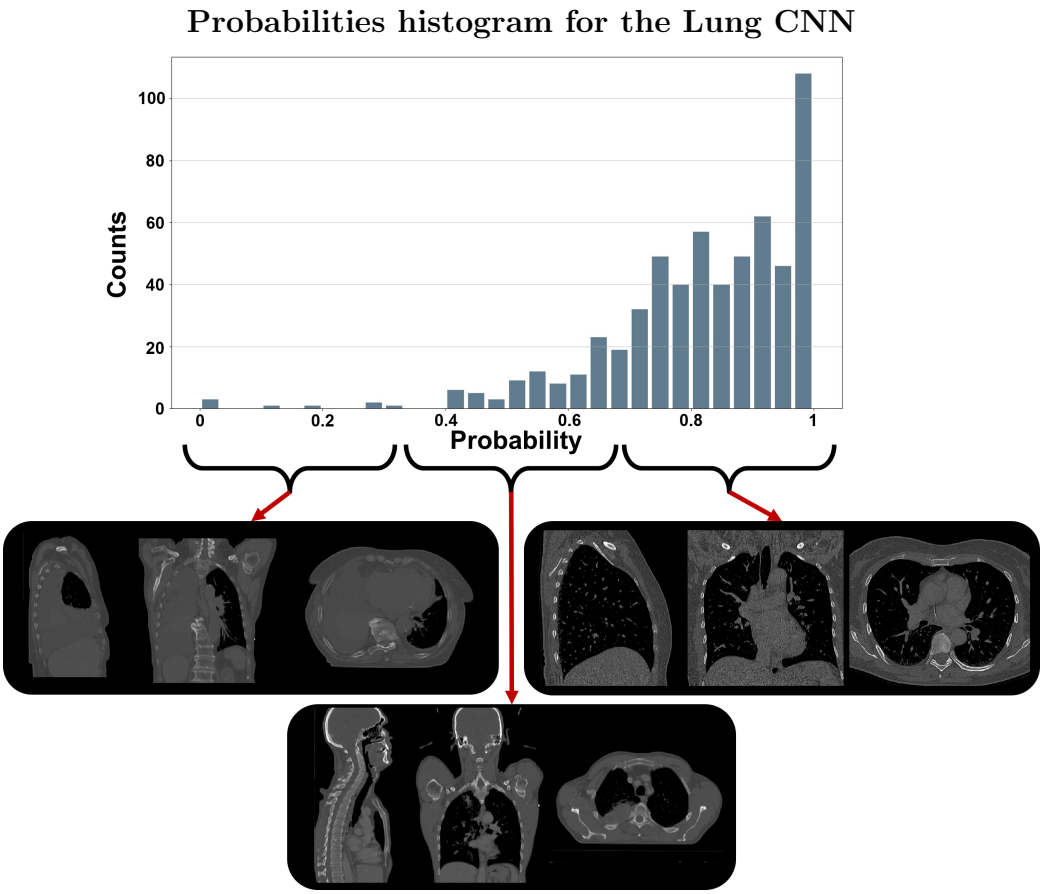


Figure 4.5: The histogram in the top graph displays the probabilities for 587 lung images as inferred by the corresponding classifier. The median probability was 0.84, with 0.74-0.94 interquartile range ($25^{th} - 75^{th}$ percentile); see Table 4.2 for more information. The images in the bottom part show a representative subject belonging to the corresponding probability section (highlighted by the curly braces). The probabilities of being lung for the three images shown were (from left to right): 0.1, 0.62 (head-neck 0.18) and 1.

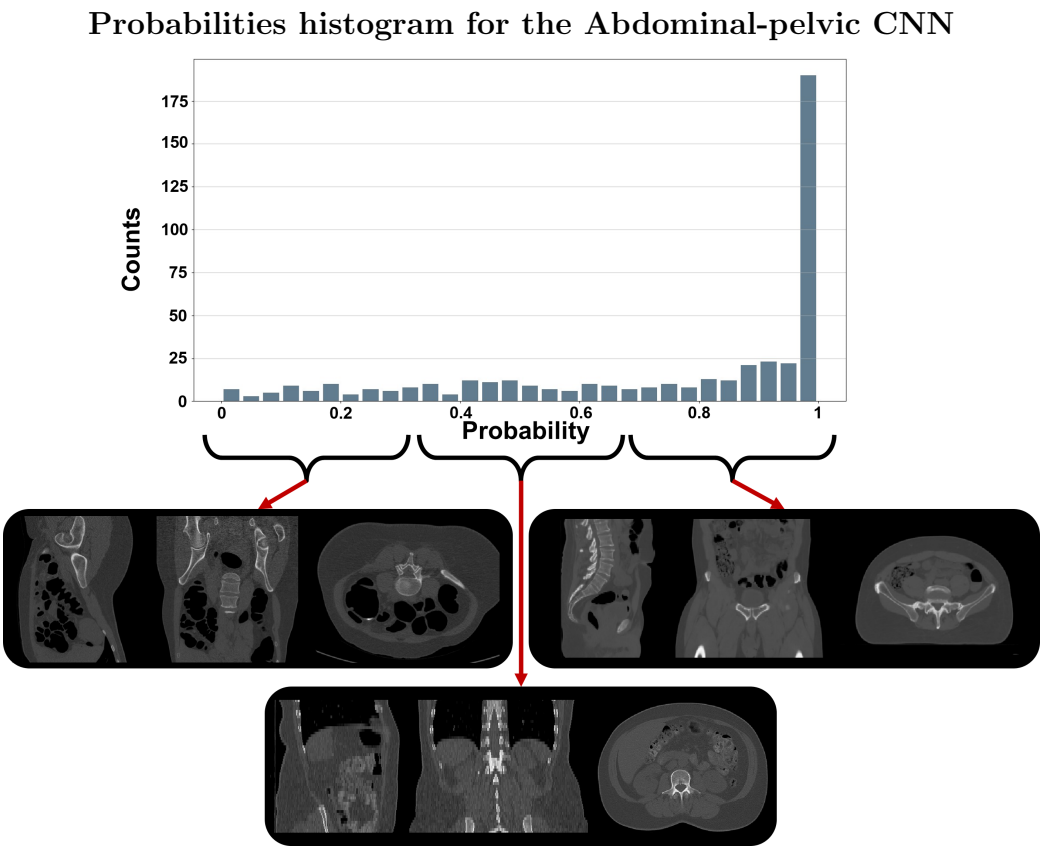


Figure 4.6: The histogram in the top graph displays the probabilities for 469 abdominal-pelvic images as inferred by the corresponding classifier. The median probability was 0.9, with 0.52-1 interquartile range ($25^{th} - 75^{th}$ percentile); see Table 4.2 for more information. The images in the bottom part show a representative subject belonging to the corresponding probability section (highlighted by the curly braces). The probabilities of being abdominal-pelvic for the three images shown were (from left to right): 0.12, 0.59 (lung 0.19) and 1.

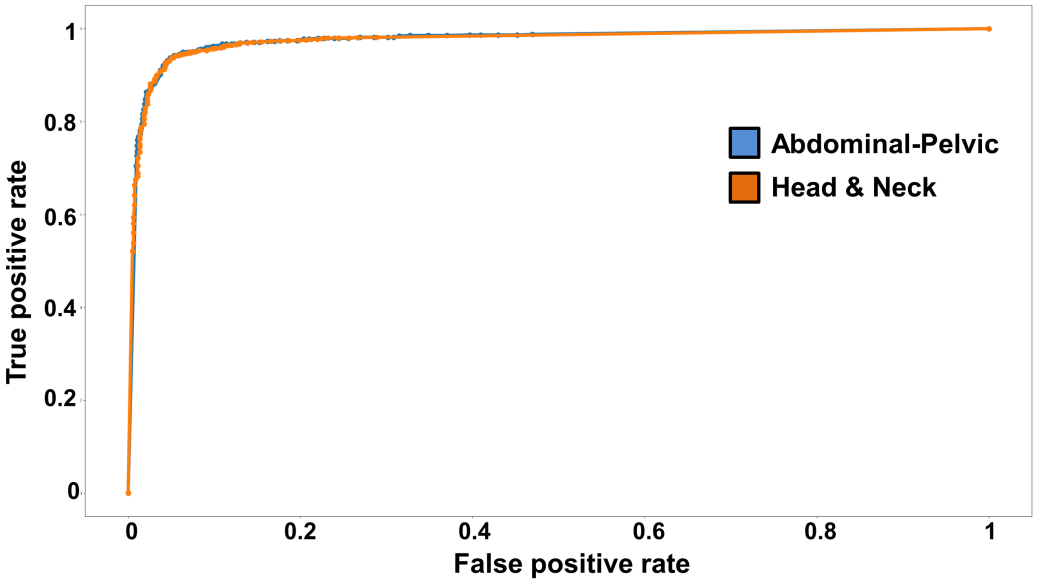


Figure 4.7: ROC curve for the two one-vs-all CNNs trained to distinguish between abdominal-pelvic and head-neck MR images.

4.1.2 MR body part classification (BP-class MR)

Two additional CNNs were trained in a 2D fashion to distinguish between head-neck and abdominal-pelvic MR images. Results are reported in this section. The ROC curves for both classifiers are shown in Figure 4.7. Both CNNs had high AUC, 0.9755 for the abdominal-pelvic and 0.9748 for the head-neck, as shown in Table 4.4.
The saliency maps depicted in Figure 4.8, show that the most important voxels when classifying a head-neck MR image lie in the ventricles region and close to the brain/skull interface, while those highly involved in the prediction of the abdominal-pelvic class are moslty following the soft tissues boundaries.
Finally, the histograms of the probabilities and their median values are illustrated in Figure 4.9 and Table 4.5. A summary of the MR images found in the three probability ranges (0-0.33, 0.34-0.66 and 0.67-1), with a description of their main characteristics, is reported in Table 4.6.

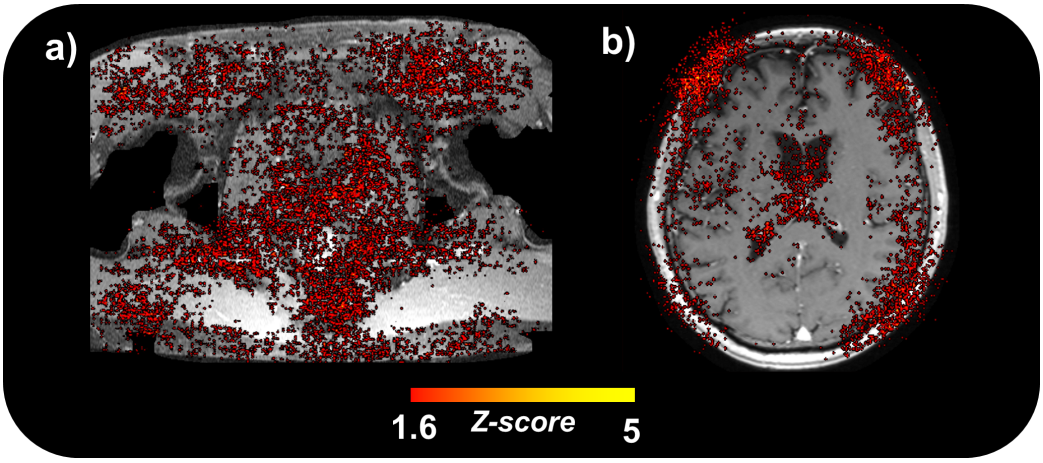


Figure 4.8: Saliency maps for the abdominal-pelvic (a) and the head-neck (b) body part classifiers.

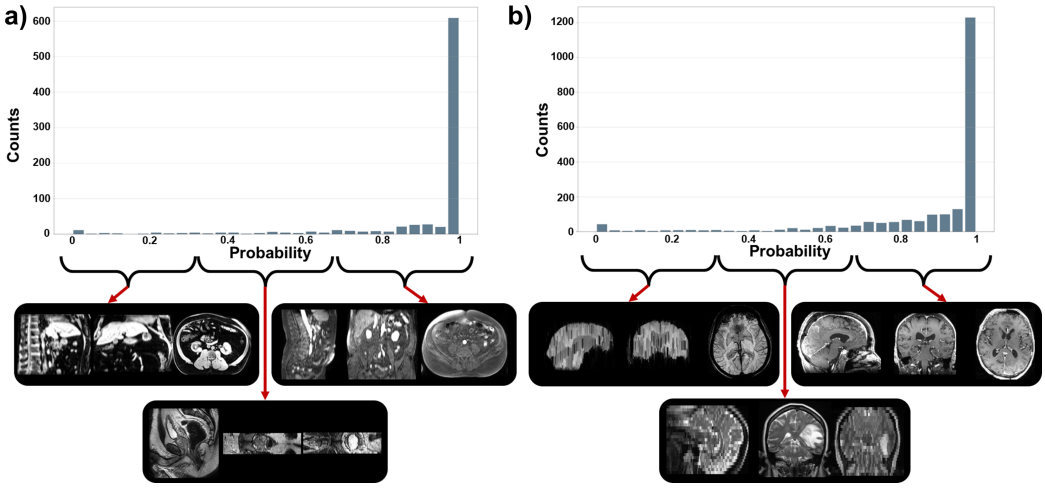


Figure 4.9: The histograms in the top graphs display the probabilities for 814 abdominal-pelvic (a) and 2117 head-neck (b) MR images, as inferred by the corresponding classifiers. The median probability was 1 for both of them, with an interquartile range ($25^{th} - 75^{th}$ percentile) of 0.96-1 for the abdominal-pelvic and 0.85-1 for the head-neck (see Table 4.5 for more information). The images in the bottom part show a representative subject belonging to the corresponding probability section (highlighted by the curly braces).

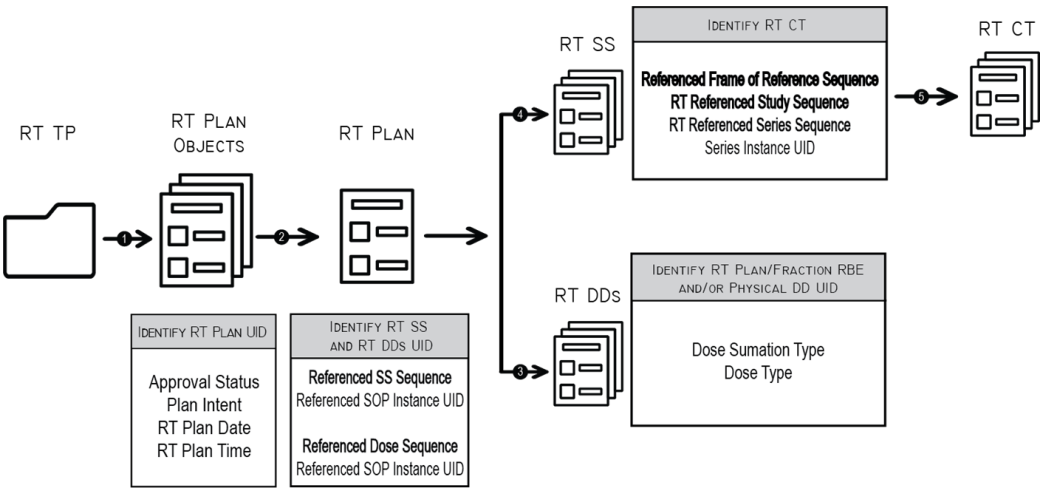


Figure 4.10: Overview of the steps implemented in PyCURT to sort the radiotherapy data. RT TP: radiotherapy time point; RT SS: structure set; RT DDs: dose distributions.

4.1.3 MR image contrast classification (MR-Class)

The MR images labelled by BP-class MR as the body part of interest (chosen by the user during the PyCURT setup) are then further classified based on their contrast using MR-Class (Salome et al., in preparation). In particular, each MR image goes into a stack of six one-vs-all classifiers able to differentiate between T1w (pre- or post contrast agent injection), T2w, FLAIR, Apparent Diffusion Coefficient (ADC) or Susceptibility weighted imaging (SWI).

4.1.4 Radiotherapy data curation and sessions labelling

The approach implemented in PyCURT to sort the radiotherapy data is reported in Figure 4.10. PyCURT's RT sorting was run on a cohort of 152 patients with head-neck squamous carcinoma (see cohort description in section 3.1.6). It correctly identified all the used RT plan objects for 102 of them. Furthermore, it linked together all the planning CT (92), the structure sets (79) and the dose distributions (64).
For the RT time point with both planning CT and RT structure set, PyCURT can extract all the manually contoured structures and save them as NIfTI. If the dose distribution is also present, then it can be used to identify

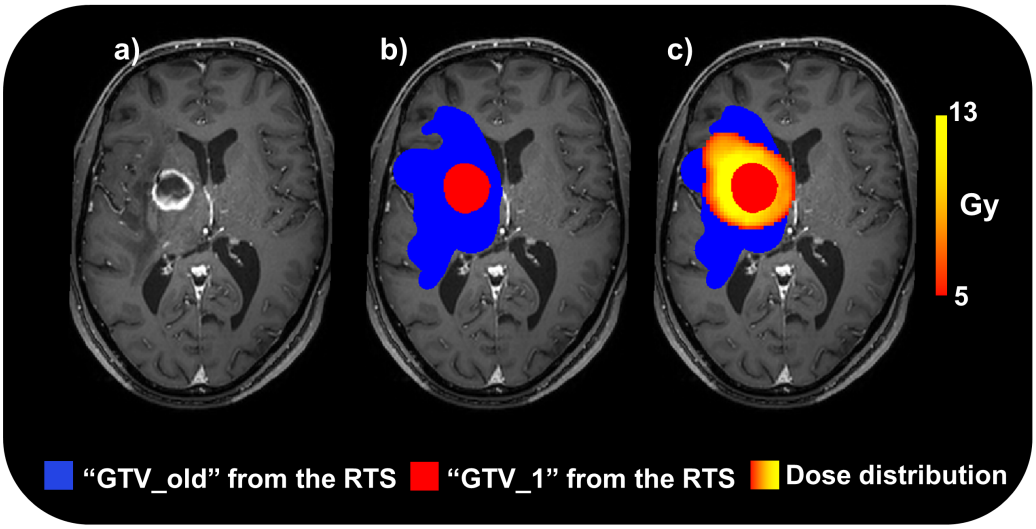


Figure 4.11: Example on how the information stored in the dose distribution can be useful to identify the structure, among those stored in the RT structure set (RTS), with the highest probability to be the one used during the radiotherapy treatment planning.

and extract the structure within the RT SS with the highest overlap with it. This can be useful when multiple structures with similar names are saved in the same RT SS and the user only need the one used to plan the dose delivery to the patient. A visual example can be found in Figure 4.11.
The final step in a PyCURT workflow is the sessions labelling based on the radiotherapy time point. A detailed flowchart is shown in Figure 4.12.

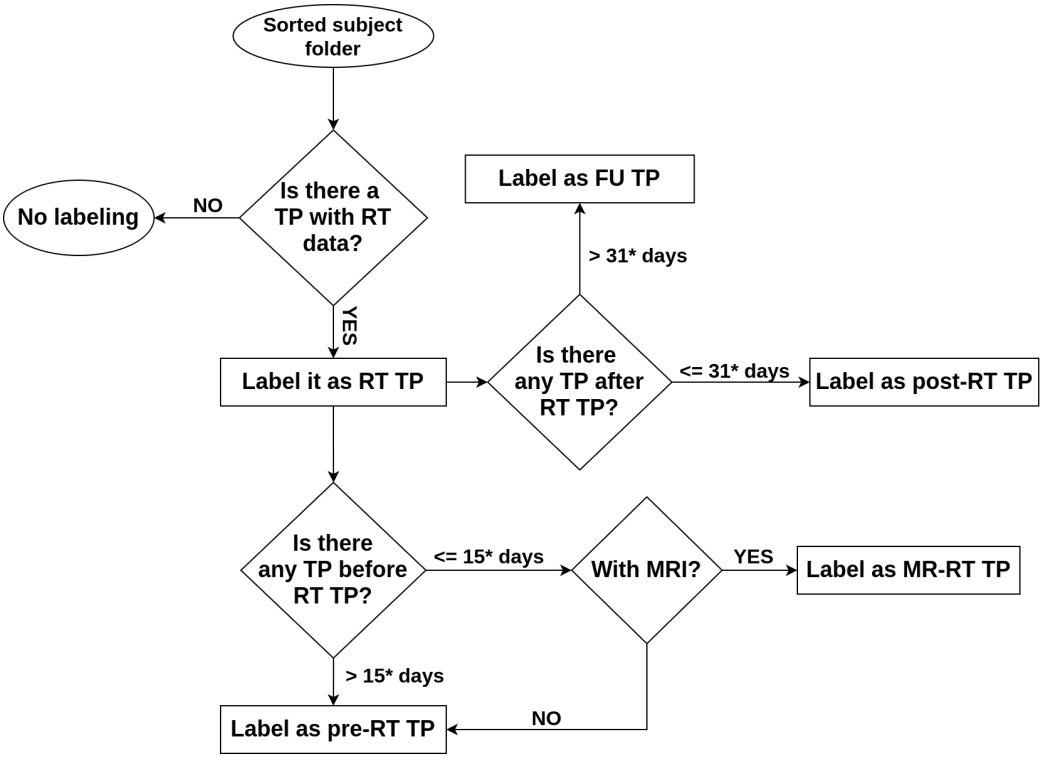


Figure 4.12: Time point (TP) labelling based on radiotherapy (RT) as implemented in PyCURT.
$*$ the number of days can be changed by the user based on the project under investigation.

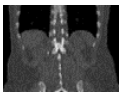
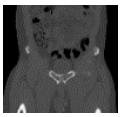
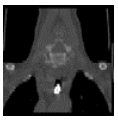
Probability range	**Visual example**	**Description**
0 - 0.33		CT images with only one lung
0.34 - 0.66		CT images with both lung and head-neck
0.67 - 1		Standard lung CT images
0 - 0.33		Colonography CT images with different patient position
0.34 - 0.66		CT images with both abdomen and lung
0.67 - 1		Standard abdominal-pelvic CT images
0 - 0.33		CT images with zoomed neck, no head
0.34 - 0.66		CT images with both head-neck and lung
0.67 - 1		Standard head-neck CT images

Table 4.3: Summary of the images found in the different probability ranges for the lung (top), abdominal-pelvic (middle) and head-neck(bottom) BP-Class CT classifiers.

CNN	AUC	95%CI
abd-pel	0.9755	0.9689 - 0.9822
head-neck	0.9748	0.9690 - 0.9807

Table 4.4: Results for the abdominal-pelvic and head-neck MR image classifiers trained with a 2D structure. CNN: Convolutional neural network; AUC: Area under the curve; CI: Confidence interval (DeLong).

CNN	Mdn-P	IQR (25 - 75)
abd-pel	1	0.96 - 1
head-neck	1	0.85 - 1

Table 4.5: Median probabilities for the two BP-Class MR CNNs. Mdn-P: median probability; IQR: interquartile range.

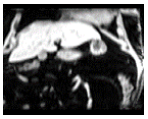
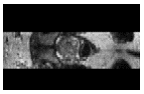
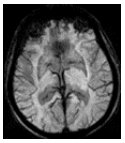
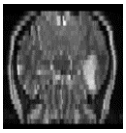
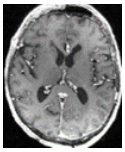
Probability range	**Visual example**	**Description**
0 - 0.33		Abdominal MR images with unusual tissue contrast
0.34 - 0.66		Abdominal MR with very small field of view in one direction
0.67 - 1		Standard abdominal-pelvic MR images
0 - 0.33		MR images with reconstruction artefacts
0.34 - 0.66		MR images with poor resolution in one direction
0.67 - 1		Standard head-neck MR images

Table 4.6: Summary of the images found in the different probability ranges for the abdominal-pelvic (top) and head-neck(bottom) BP-Class MR classifiers.

4.2 RADIANTS: Radiomics analysis framework for radiotherapy studies

RADIANTS is an abstract framework to implement radiomics analysis workflows in radiotherapy studies. The main features implemented in RADIANTS are:

- It glues together the main steps for radiomics image analysis, including pre-processing, registration and features extraction.
- Given its modular structure it is easy to customize adding/changing/removing new processing steps.
- Two independent deep learning-based region of interest segmentation algorithms were developed and incorporated in RADIANTS:
    1. One method to segment the gross tumor volume in high grade gliomas, in both pre-irradiation and follow up imaging time points.
    2. One method to segment the lungs from multi-resolution CT images.

The second tool developed during this thesis work was RADIANTS. RADIANTS builds upon Nipype and allows an easy and reproducible execution of all the necessary steps to perform a radiomics analysis, with a focus on radiotherapy application.
An overview of the abstract idea behind RADIANTS can be seen in flowchart 4.13. Figure 4.14 shows all the steps involved in a features extraction workflow in RADIANTS, while Figure 4.15 depicts the Nipype's execution graph for the image pre-processing workflow.

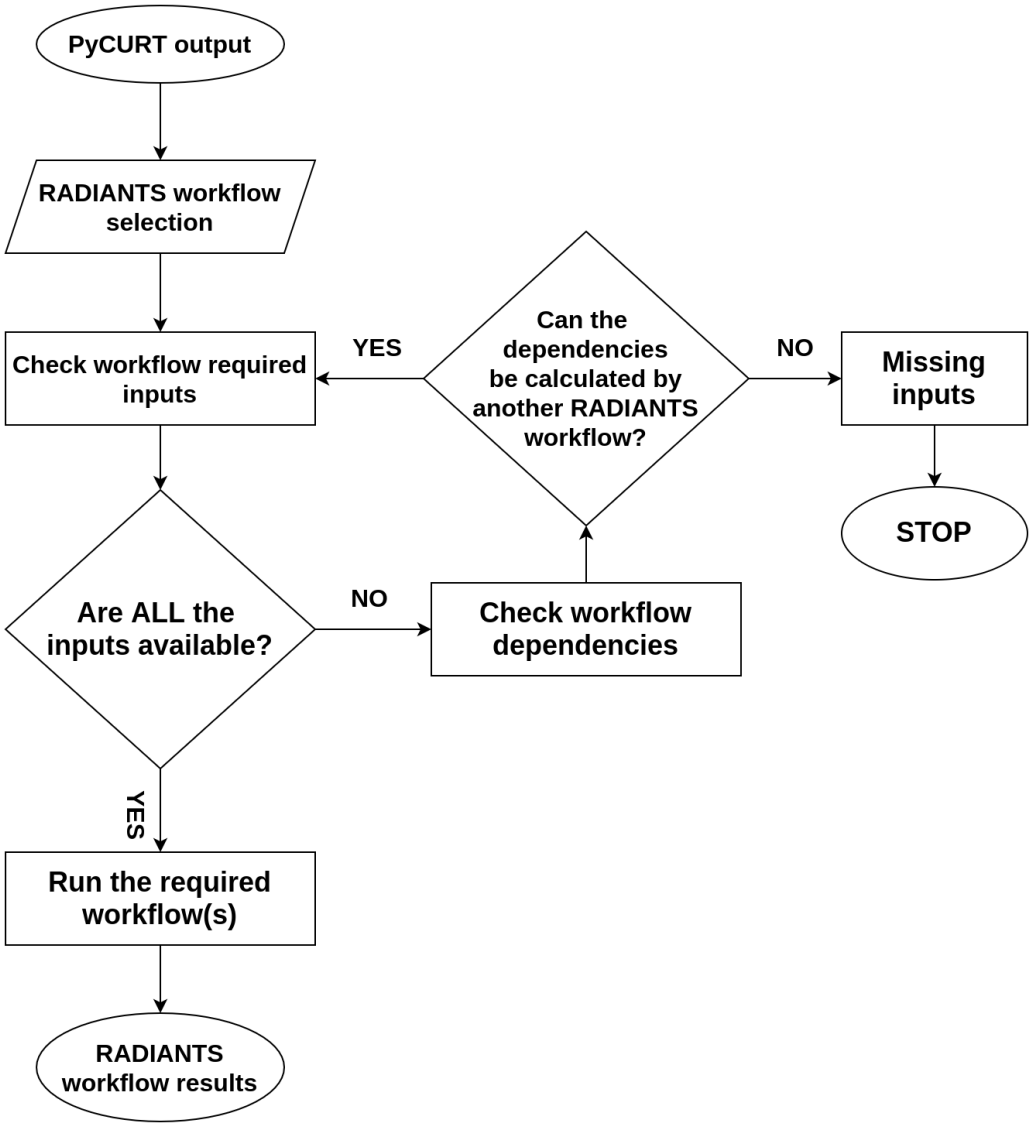


Figure 4.13: Flowchart of the abstract behaviour behind RADIANTS.

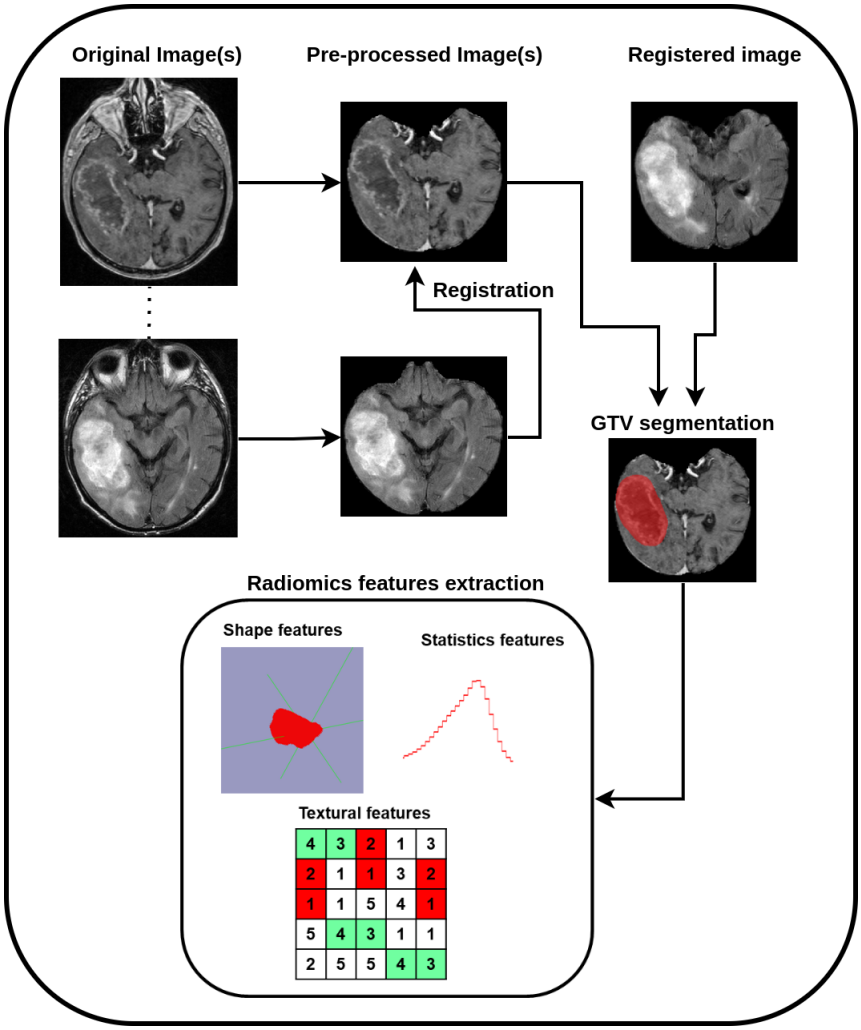


Figure 4.14: Steps run automatically by RADIANTS when a "features extraction" workflow is selected.

4.2.1 Gross tumor volume segmentation in high grade gliomas, from pre-surgical and follow up imaging time points

RADIANTS supports the extraction of radiomics features from the automatically segmented gross tumor volumes (GTV). This GTV segmentation has been developed as part of this work using a deep learning approach. In

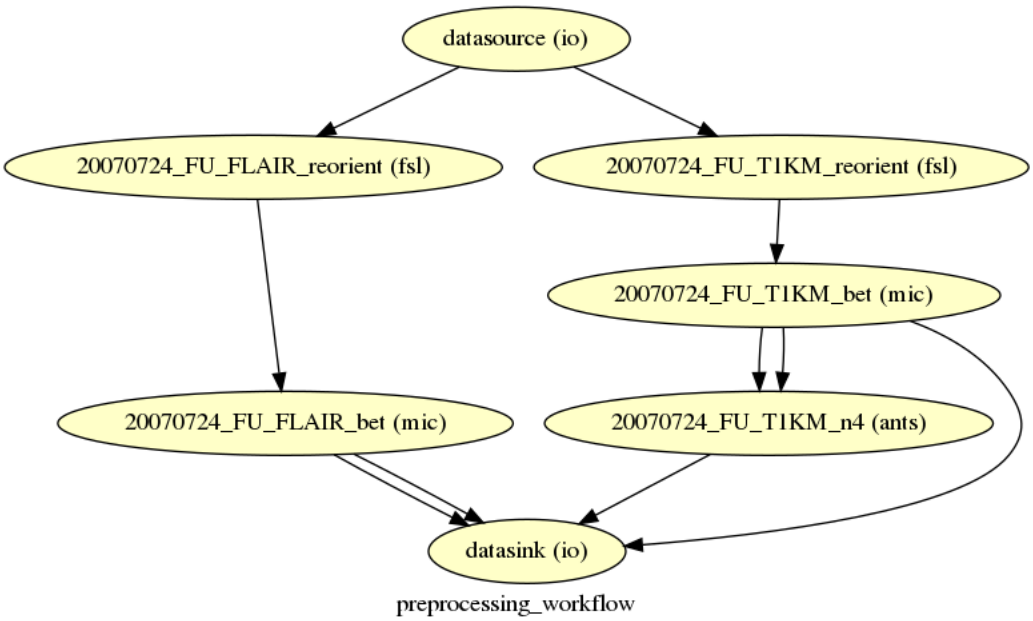


Figure 4.15: Nipype's execution graph for the image pre-processing workflow in RADIANTS.

particular, a nnUNet framework has been trained on pre-irradiation imaging time points, using T1ce and FLAIR MR contrasts and manually contoured GTV.
Figure 4.16 shows the results on a test set of 37 subjects. An example of a randomly selected patient can be found in panel a). There, the manually segmented GTV is portrayed in blue while the one automatically segmented is red. Their overlap is magenta coloured. The distributions of the Dice score coefficients (DSC) and the 95^{th} percentile Hausdorff distances (HD95) in the test set can be found in graphs b) and c), respectively. The median DSC was 0.91, with an $25^{th} - 75^{th}$ interquartile range (IQR) of 0.8-0.94, while the median HD95 was 3.46 mm (IQR: 1.62-5.62 mm).
Figure 4.17 highlights the subjects with the highest Hausdorff distance (a) and the lowest Dice score (b). As can be noted in Fig. 4.17a, the proposed approach failed in segmenting tumor infiltration along the white matter fibers. In addition, it seemed to be more conservative in segmenting the resection cavity with respect to the manual delineation (Fig. 4.17b).

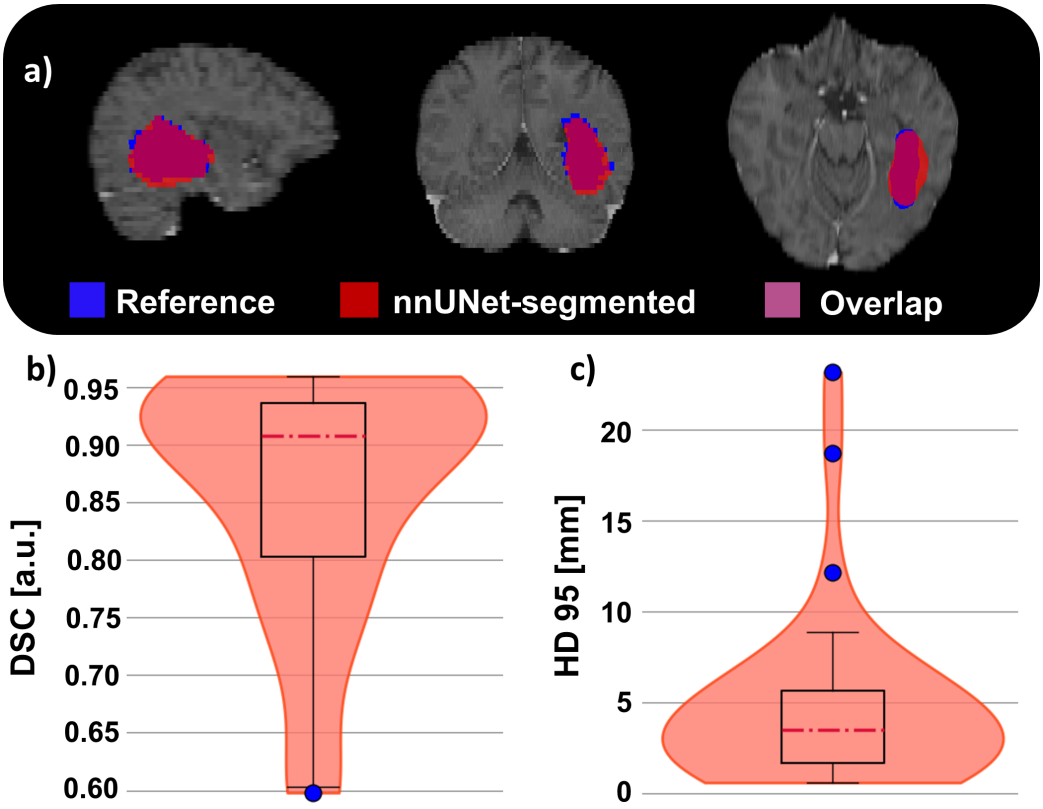


Figure 4.16: GTV segmentation results on a test set of 37 subjects. Panel a) shows the reference GTV (in blue) and the one segmented by the trained nnUNet (in red) and their overlap (magenta) for a representative subject. The distributions of the the Dice scores and the Hausdorff distances are reported in b) and c), respectively.

Performance comparison between networks trained with T1ce and FLAIR and T1ce only.

In order to evaluate whether the usage of the FLAIR MR contrast was beneficial or not to obtain an accurate GTV segmentation, the nnUNet framework was also trained using only T1ce images and tested on the same 37 subjects. Figures 4.18a and 4.18b report the Dice scores and the Hausdorff distances distributions for the nnUNet trained using only T1ce. As can be noted, the results are significantly worst than using both T1ce and FLAIR. In particular, the median DSC was 0.77 (IQR: 0.4-0.85) against 0.91, and the HD95

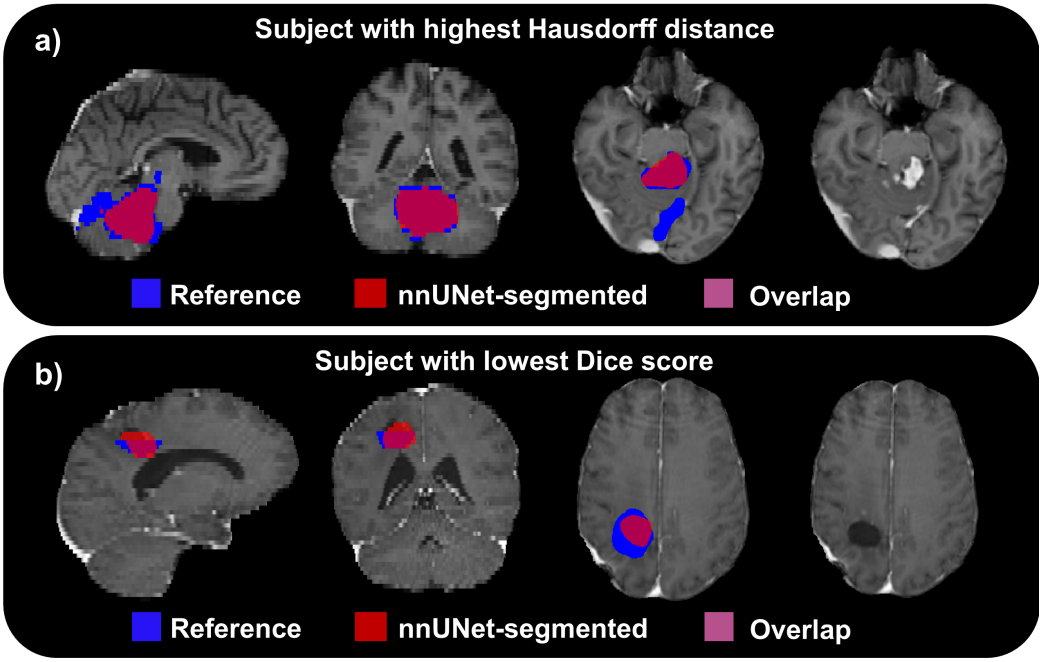


Figure 4.17: The subject with the highest Hausdorff distance (a) and the one with the lowest Dice score (b) are shown here. In both panels, the manually delineated GTV (blue), the automatically segmented (red) and their overlap (magenta) are reported.

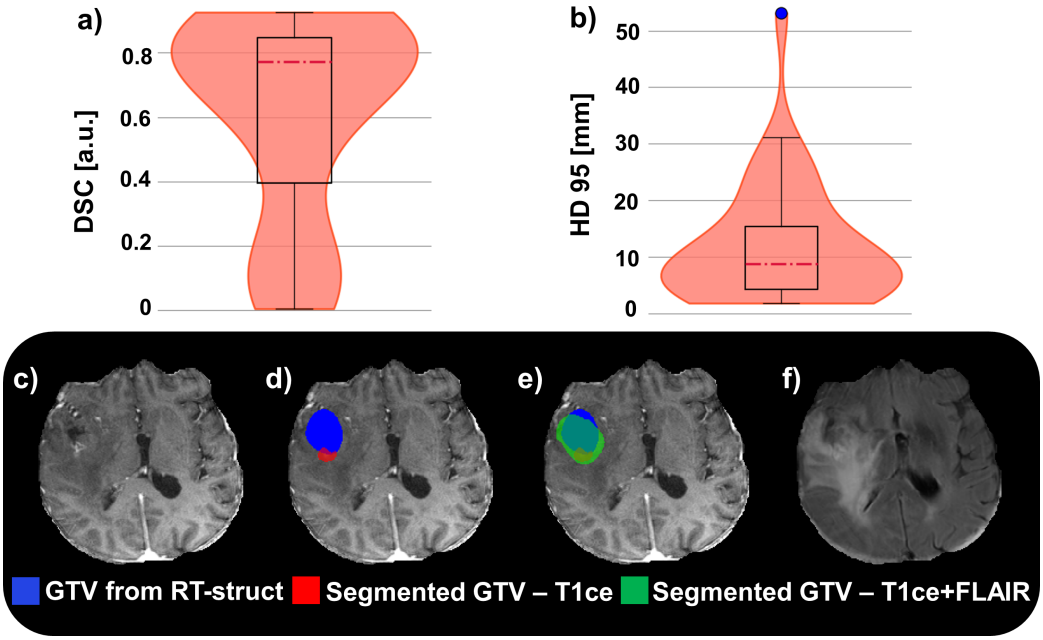


Figure 4.18: GTV segmentation results on a test set of 37 subjects, using the network trained with only T1ce images. The distributions of the the Dice scores and the Hausdorff distances are reported in a) and b), respectively. The bottom panel shows the reference GTV (in blue), the one segmented by nnUNet trained with T1ce and FLAIR contrasts (in red) and with only T1ce (green) for a representative subject.

was 8.8 mm (before 3.46 mm). The bottom panel in Fig. 4.18 shows an example of the GTV segmentation for a representative subject. These results support the usage of both contrasts to enhance GBM gross tumor volume segmentation.

GTV segmentation in follow up time points

The trained nnUNet was also tested to segment GTV in 665 follow up time points (from 115 different patients). Out of the total number, the network segmented at least one GTV in 553 time points. These GTV were qualitatively analyzed by one human rater to classify them as "good", "usable" or "wrong". Results are reported in Table 4.7. Only 15 GTV were labelled as "wrong", proving the robustness of this network to longitudinally segment

	Good	Usable	Wrong
Human rating	502	36	15

Table 4.7: Results of the qualitative analysis on the 553 GTV segmented in the follow up time-points by the proposed method.

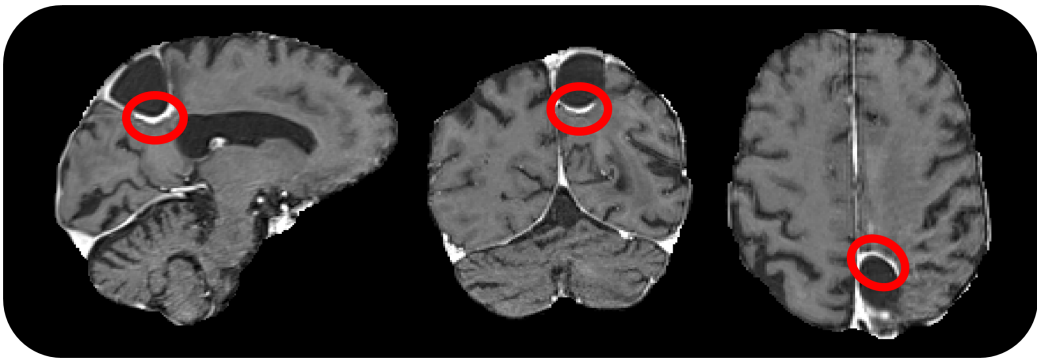


Figure 4.19: Example of one image where the human rater identified a tumor (red circles) that was not segmented by the network.

the GTV. The remaining 112, where no GTV was segmented by the network, were manually checked to look for false negative. In 48 out of 112, at least one GTV was identified while in the remaining 64 no GTV was seen by the human rater. Most of the tumors that were seen by the human rater and not by the network looked like the one showed in Figure 4.19. To quantitatively analyze the follow up segmentation, the same human rater segmented 50 randomly chosen time points among the 553 segmented by the network. This analysis resulted in a median Dice score of 0.76, (IQR: 0.6-0.8) and in a median 95^{th} Hausdorff distance of 8.78 mm (IQR: 6-17).

Inter-raters variability assessment

To finally corroborate the usage of automatic GTV segmentation algorithm over manual contouring, an inter-rater variability analysis was performed on the same 50 follow up time points as in section 4.2.1. Three independent human raters (one was the same who performed the analyses in the previous section) segmented the same time points and DSC and HD95 were used

to quantify their agreement. Results are reported in Tables 4.8 and 4.9. The comparison between the network and the three raters resulted in very similar median DSC and HD95 (0.76-0.77-0.71 and 8.78mm-6.28mm-8.06mm, respectively (Tbl. 4.8)). On the other hand, inter-rater analysis showed higher variability, with the DSC oscillating between 0.63 and 0.77 and the HD95 between 5 and 11.36 mm (Tbl. 4.9).

	Rater #1	**Rater #2**	**Rater #3**
DC ($25^{th} - 75^{th}$ percentile)	0.76 (0.6 - 0.8)	0.77 (0.7 - 0.8)	0.71 (0.61 - 0.72)
HD95 ($25^{th} - 75^{th}$ percentile)	8.78 mm (6 - 17)	6.28 mm (4.44 - 8.2)	8.06 mm (6 - 14.1)

Table 4.8: Table with the comparison between the 50 randomly chosen GTV in the follow up time-points segmented by the neural network and by three independent human raters. DC: Dice coefficient, HD95: 95^{th} percentile Hausdorff distance.

	Rater #1 vs Rater #2	**Rater #1 vs Rater #3**	**Rater #2 vs Rater #3**
DC ($25^{th} - 75^{th}$ percentile)	0.66 (0.54 - 0.73)	0.77 (0.71 - 0.81)	0.63 (0.5 - 0.71)
HD95 ($25^{th} - 75^{th}$ percentile)	11.36 mm (7.28 - 19.44)	5 mm (3 - 9.54)	10.98 mm (8.02 - 16.67)

Table 4.9: Inter-rater manual segmentation comparison on the 50 randomly chosen GTV in the follow up time-points. DC: Dice coefficient, HD95: 95^{th} percentile Hausdorff distance.

4.2.2 Deep learning-based automatic lung segmentation for healthy and fibrotic lungs in mice from multi-resolution CT

Another segmentation implemented within RADIANTS is the pre-clinical lung segmentation from multi-resolution CT images, developed as part of this thesis.
This method builds upon a well-established convolutional neural network architecture, called UNet. The details of the network structure can be found in Figure 4.20. This network has been trained/tested on around 2000 mice images. Figure 4.21 shows an overview of the pipeline used to train and test the method using two independent mouse cohorts.

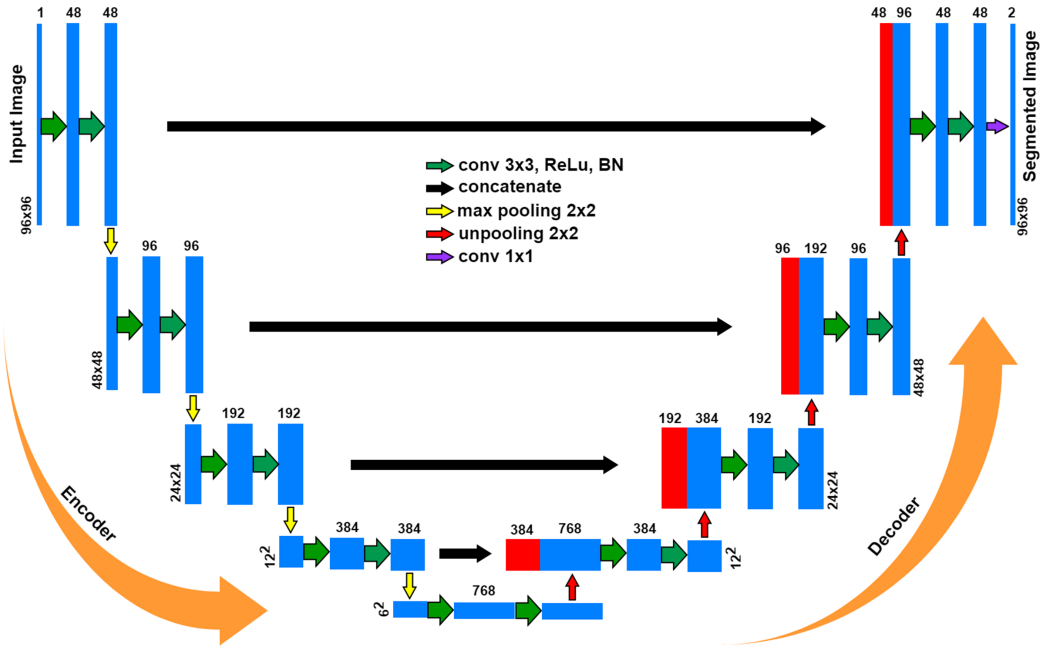


Figure 4.20: Mouse lung segmentation network architecture used in the current method.

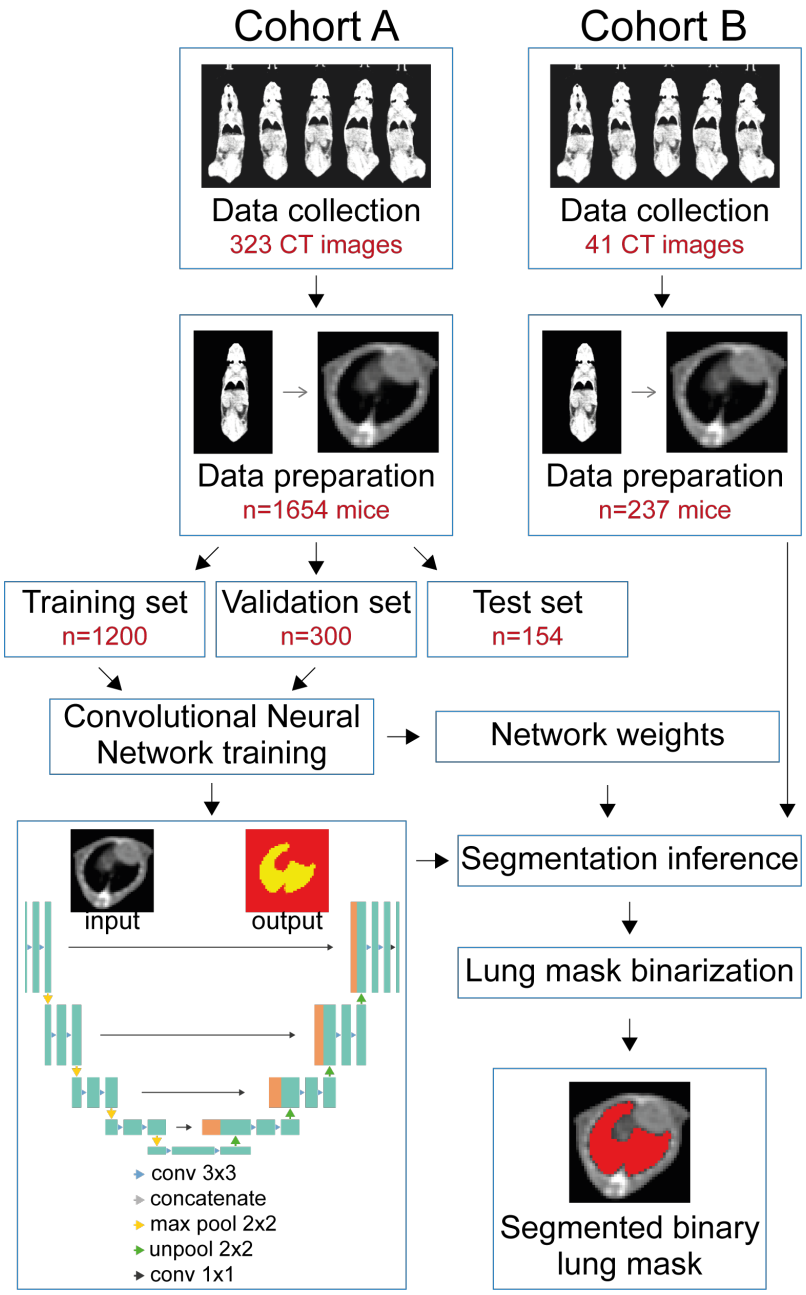


Figure 4.21: Overview of the workflow used for training, validation and testing a U-net CNN for lung segmentation. Both the main mouse-CT cohort (cohort A, left) and cohort B (right) are shown.

Network performance

The total model training time was around 15 hours on a workstation equipped with Ubuntu 18.04, Intel Xeon R processor with 8 cores and 16 Gb of RAM and a graphics card NVIDIA GeForce GTX 1060 (6 Gb). The average inference time for one mouse was 0.31 seconds.
A first evaluation of the proposed CNN was performed on a cohort of 154 mice (test set cohort A). On this test set, the proposed algorithm yielded a median Dice score coefficient of 0.9842 and a median Hausdorff distance (95th percentile) of 0.4707 mm (see also Table 4.10). Distributions of the Dice scores and Hausdorff distances are shown in Figure 4.22.

Metric	**Median**	**IQR** ($25^{th} - 75^{th}$ **percentile**)
Dice Score Coefficient	0.9842	0.0117 (0.9765 - 0.9882)
Hausdorff distance 95^{th} [mm]	0.4707	0.5117 (0 - 0.5117)

Table 4.10: Results on a test set of 154 mice from cohort A. IQR: Interquartile range.

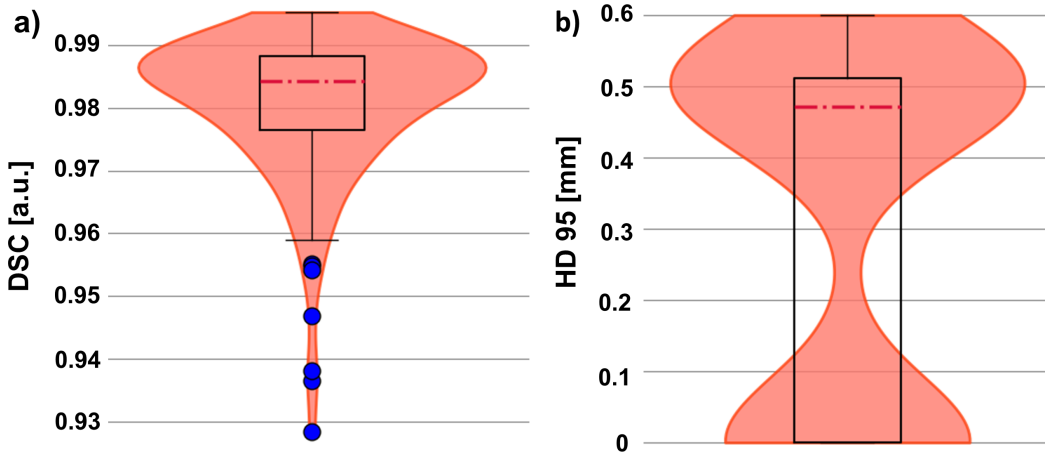


Figure 4.22: Dice score coefficients (a) and Hausdorff distances (b) distributions in the 154 mice test set from cohort A.

DSC and HD95 were also calculated using CNN lung masks binarized with a hard probability threshold of 0.5. In this case the median DSC was 0.9821

(IQR, 0.9759-0.9882) and a HD95 of 0.4707 mm (IQR, 0 – 0.5117). Given the slightly better results, Otsu thresholding was chosen as binarization method. A representative lung segmentation from the test set is shown in Figure 4.23. The DSC, in this case, was 0.9812. Voxels that were exclusive to the reference lung mask are marked red; those present only in the CNN-segmented mask are marked blue. As can be noted, almost all voxels in the lungs are colored purple, indicating a high overlap between the reference and the CNN-segmented masks. Single discordant voxels were observed on the lungs' margins.

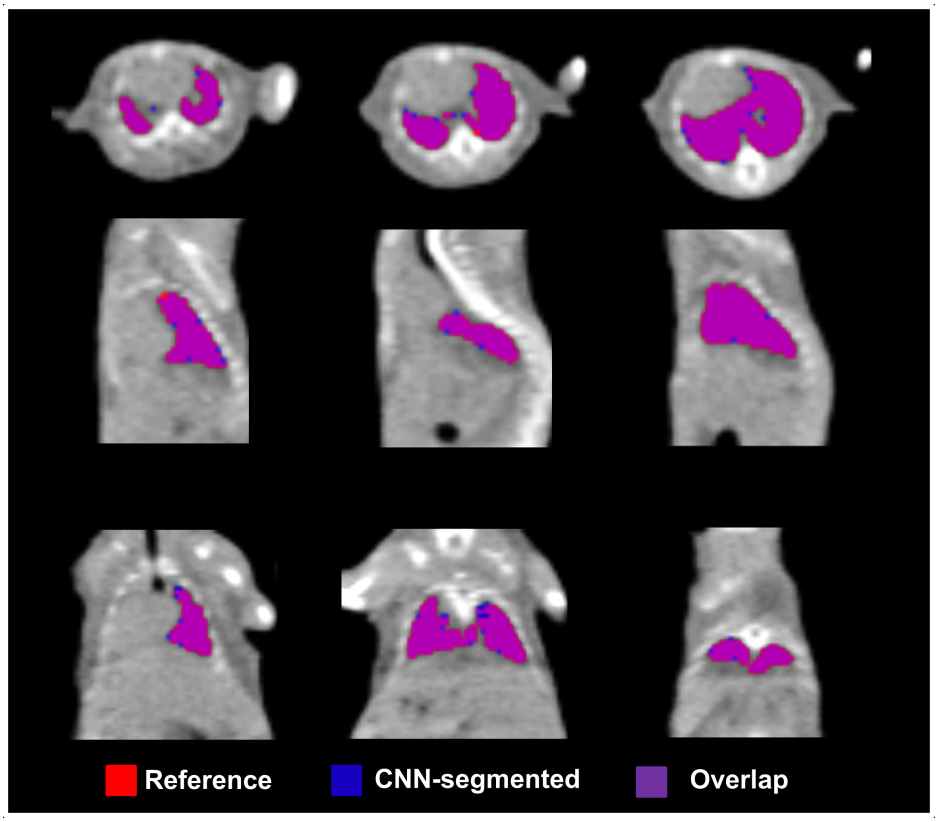


Figure 4.23: Representative segmentation results of the CNN for a random mouse from the test set (cohort A). The reference lung mask (red) is overlapped with the CNN-segmented mask (blue). Their intersection is depicted in purple

For additional validation, an independent set of 237 mice was evaluated (co-

hort B). These mice were acquired with different resolutions as compared to those used to train the network (see section 3.1.3). Observed median DSC and HD95 in this validation dataset were 0.9661 (IQR, 0.9553 – 0.9724) and 0.3066 mm (0.2964 – 0.3242mm), respectively.

High-resolution mouse CT and human lung segmentation

This network was then re-trained, using a transfer learning approach, to investigate its ability to segment lungs from higher resolution mouse CT images and from human CT. See Figure 4.24 for an overview of the transfer learning approach.
Results for the high-resolution mice are shown in Figure 4.25a that reports the overlap between the CNN and the semi-automatically segmented lung mask, showing a high overlap. The median DSC for the four tested mice was 0.9047 and the 95^{th} percentile Hausdorff distance was 0.3317 mm. The results for the human lung segmentation are reported in Figure 4.25b. The median DSC across the three subjects was 0.9651 and the Hausdorff distance was 3 mm.

Spatial mismatch quantification

To further evaluate the performance of the proposed approach, areas in the lungs that were more prone to diverging segmentation between CNNs and reference were highlighted. For this, data from the test set of cohort A (n=154) were used. See Figure 4.26 for the complete analysis pipeline.
Figure 4.27 shows three representative views of the mouse template overlaid with z-scores. The blue voxels are those where the CNN over-segmented the lungs, while the red voxels represent the area of under-segmentation with respect to reference masks. As can be noted, most of the voxels classified divergently were restricted to two anatomical regions (red and blue hot spots in Figure 4.27).

Equivalence analysis

Finally, the results for the equivalence analysis are reported in Figure 4.28. Testing for equivalence using a TOST approach of FI values extracted from both masks yielded a $p - value$ of 2.13×10^{12}, thus rejecting the Null hypothesis of a relevant difference in means of both groups (the distributions of the Fibrotic indices are shown in Figure 4.28b). Results for the histogram

comparison are reported in Figure 6a. Again, results confirmed the equivalence of using any of the two sets of masks, as an almost perfect overlap of the two histograms was observed (TOST $p - value : 0.002$).

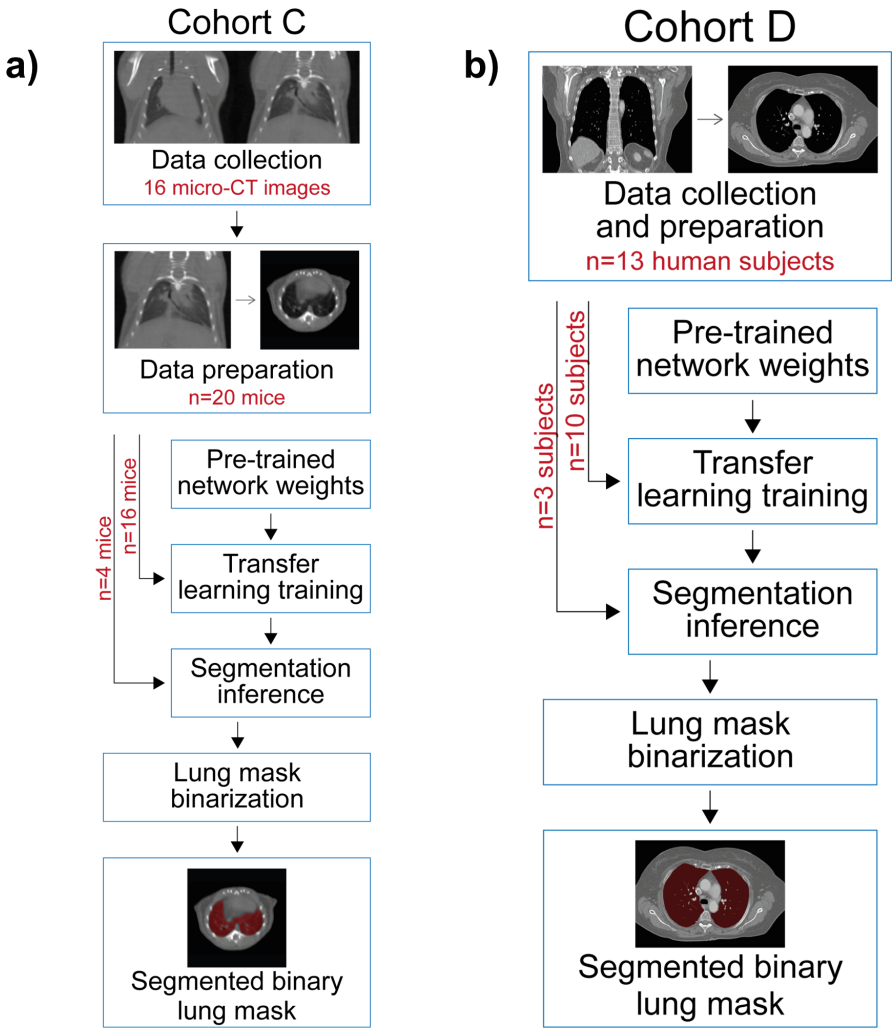


Figure 4.24: Overview of the transfer learning approach used to re-train the network to segment lungs from high resolution mouse CT (a) and clinical human CT (b).

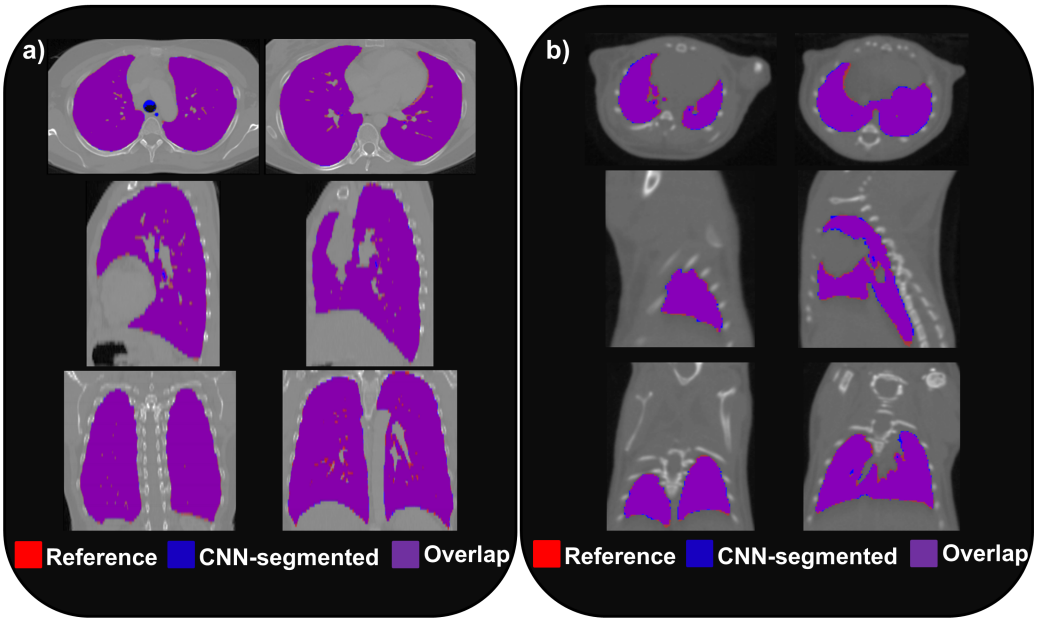


Figure 4.25: Results of the re-trained network using a transfer learning approach. Segmentation results for both a representative human (a) and high-resolution mouse (b) are presented. The reference lung mask (red) is overlapped with the CNN-segmented mask (blue). Their intersection is depicted in purple.

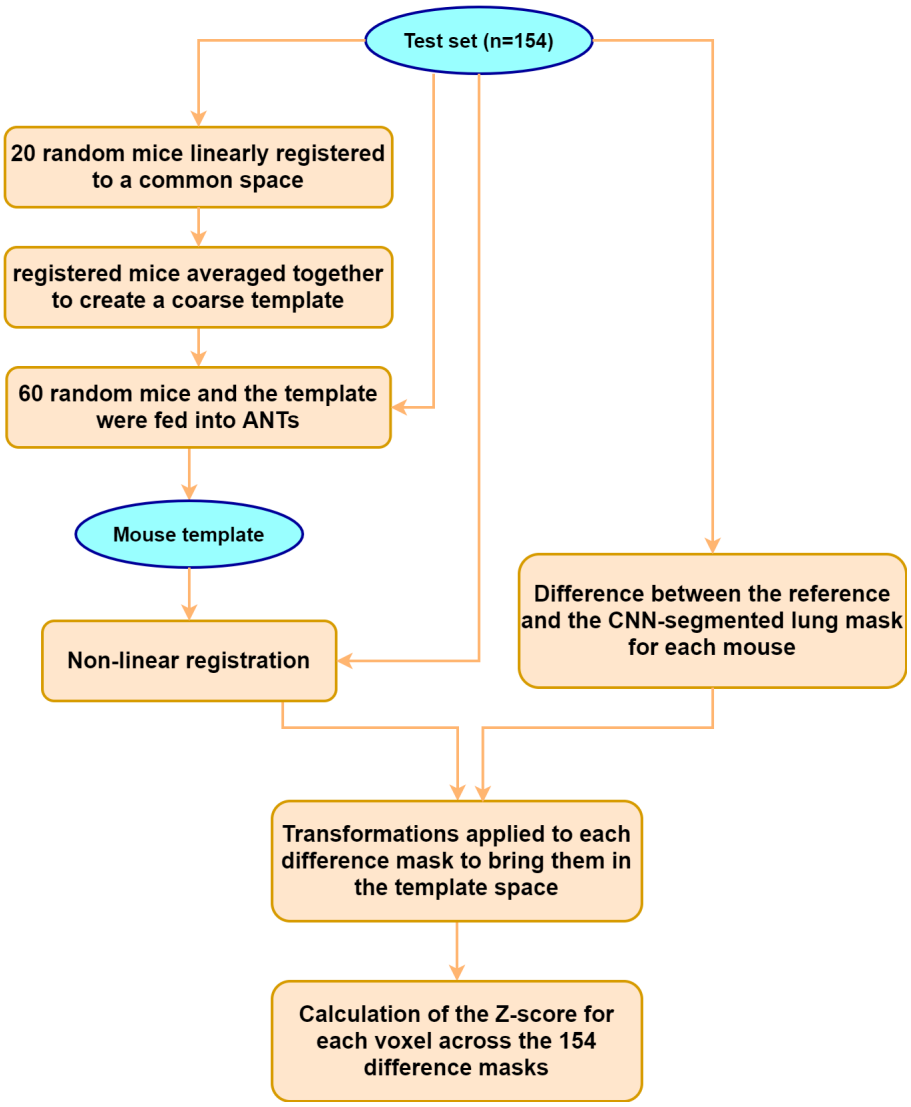


Figure 4.26: Analysis pipeline used to identify areas of diverging segmentation between the ground truth and the automatic segmentation in the test set from cohort A.

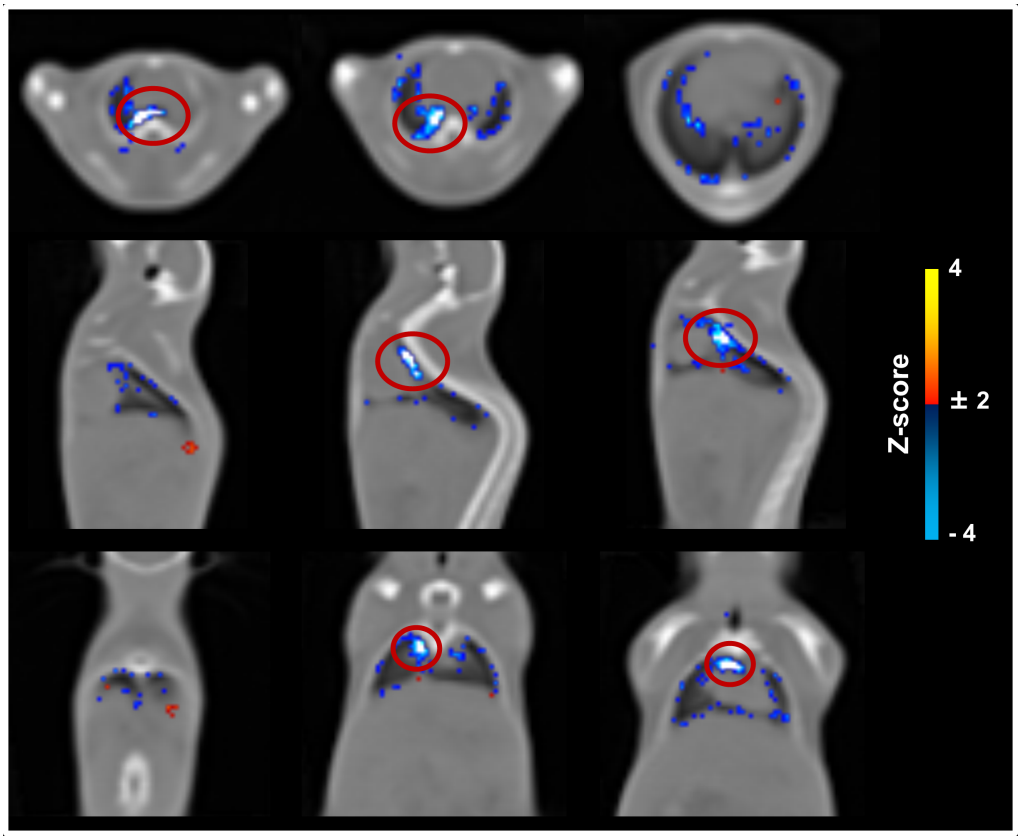


Figure 4.27: Spatial analysis of mismatches between the reference and the CNN-segmented lung masks across the mice in the test set (cohort A). The voxels over-segmented by the CNN are colored blue, while those under-segmented are shown in red.

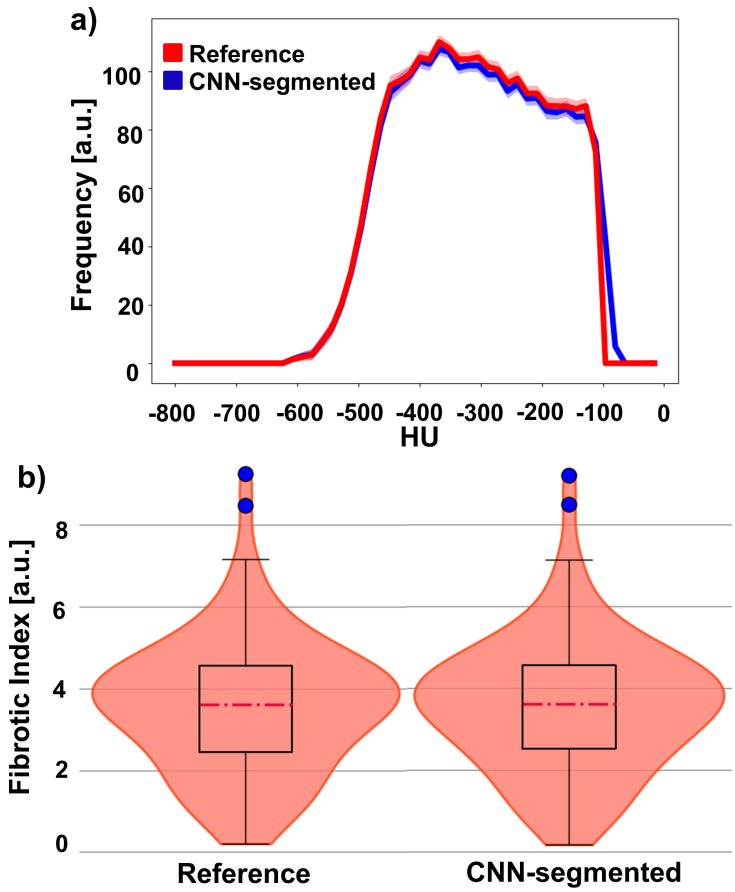


Figure 4.28: Comparison of CNN vs. reference masks using the Fibrotic Index and overall distributions of HU included in masks. a) Result for the histograms comparison. The average Hounsfield Unit (HU) values histogram across mice extracted using the semi-automatically segmented mask is shown in red, while those calculated using the CNN segmented mask is colored blue. The shadow represents the standard error of the mean. TOST $p - value$: 0.002. b) Violin plots showing the distributions of the Fibrotic Indices calculated using the semi-automatically segmented lung masks (reference) and the masks segmented by the proposed method (CNN-segmented). TOST $p - value$: 0.002

4.3 PD-L1 R: a MR based surrogate for PD-L1 expression in Glioblastoma multiforme

- Identification of an imaging-based radiomics signature, PD-L1 R, linked to PD-L1 expression in primary glioblastoma.
- Patients labelled as PD-L1 R High showed a better prognosis.
- Assessment of the interaction between PD-L1 R and major molecular covariates, including PD-1, SVZ± and MGMT promoter methylation status.

The second main part of this thesis work focused on applications. In particular, the developed infrastructures were applied to link imaging features to tumor immune environment and its evolution under radio-chemotherapy. The first application focused on the identification of key radiomics features involved in predicting PD-L1 expression in glioblastoma.
A total of 121 subjects were enrolled in this study (for an overview of the inclusion criteria refer to Figure 4.29). The whole pipeline implemented to find and evaluate the radiomics signature associated with PD-L1 can be found in Figure 4.30.

4.3.1 PD-L1 R results

The most robust imaging features, the so-called PD-L1 R radiomics signature, encompassed 3 features, which are reported in Table 4.11. PD-L1 R

Feature	Image Type	Class	Name
F1	wavelet.LHH	first order	median
F2	wavelet.HHL	GLCM	IDMN
F3	original	shape	sphericity

Table 4.11: PD-L1 R radiomics signature. L: Low; H: High; IDMN: mean inverse difference moment normalized; GLCM: gray-level co-occurrence matrices.

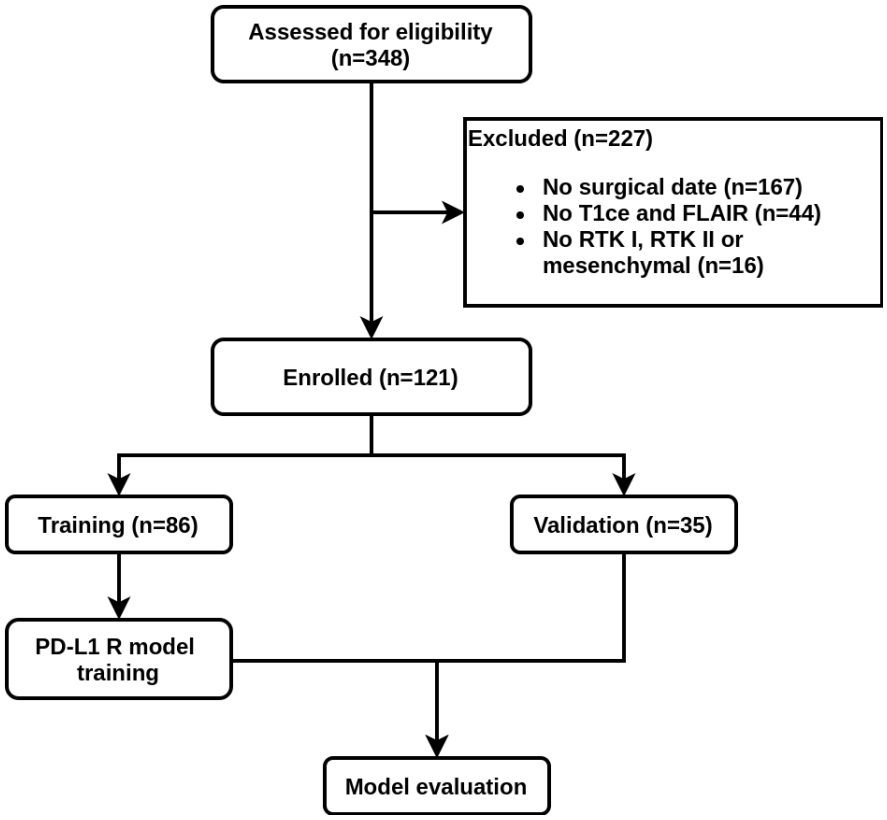


Figure 4.29: Overview of the cohort and the inclusion criteria adopted in this study.

was then used to divide the train set into 2 groups with different PD-L1 M expression; a logistic regression model was then trained to classify between those groups. The model parameters can be seen in Table 4.12.
The logistic regression model was then applied to split the validation sets into 2 groups. The PD-L1 M distributions in the two groups, for both the training and the validation set, can be found in Figure 4.31.
The logistic regression model trained using PD-L1 R was also applied to split the whole set of 121 subjects into 2 groups (60 labelled as PD-L1 R high and 61 as low). Again, the identified groups were statistically significantly different ($p = 0.00023$) in terms of PD-L1 M, as shown in Figure 4.32a. The black panels on the right of Fig. 4.32 show the T1ce image of the subject with the highest PD-L1 M value in the "High" group (b) and the one with

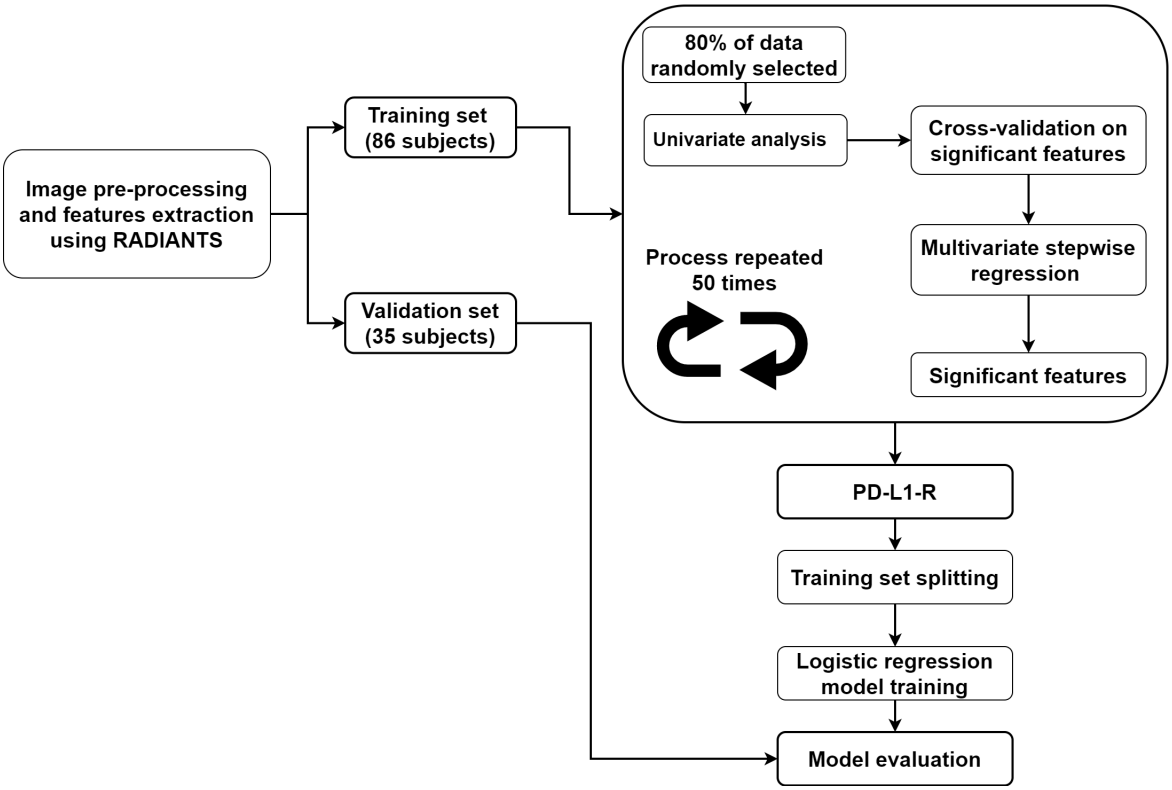


Figure 4.30: Complete analysis pipeline used to identify and evaluate the radiomics signature associated with PD-L1 in pre-surgery imaging time points.

the lowest PD-L1 M in the "Low" group (c). To further test the robustness of the proposed radiomics signature, it was also applied to split the whole cohort using features extracted from T1w, T2w and FLAIR MR contrasts. The PD-L1 M distributions are shown in Figure 4.33a-c. Interestingly, there is at least a clear trend demonstrating the different PD-L1 M distribution between the high and low group in all the three modalities. For T2w contrast (Fig. 4.33b) the difference is significant ($p = 0.037$). Images in panel d) show the T1w, T2w and FLAIR image of the subject with the highest PD-L1 M value in the "High" group, respectively; while those in Figure 4.33e for the subject with the lowest PD-L1 M in the "Low" group.

Coefficients	Estimate	SE	$Zvalue$	$Pr(>\|z\|)$
Intercept	0.44	0.60	0.73	0.46
wavelet.HHL_glcm_Idmn	-6.64	1.86	-3.57	**0.00035**
wavelet.LHH_firstorder_Median	2.41	1.07	2.25	**0.025**
original_shape_Sphericity	-4.19	1.21	-3.47	**0.00053**

Table 4.12: Logistic regression model parameters. SE: Standard Error; L: Low; H: High; IDMN: mean inverse difference moment normalized; GLCM: gray-level co-occurrence matrices.

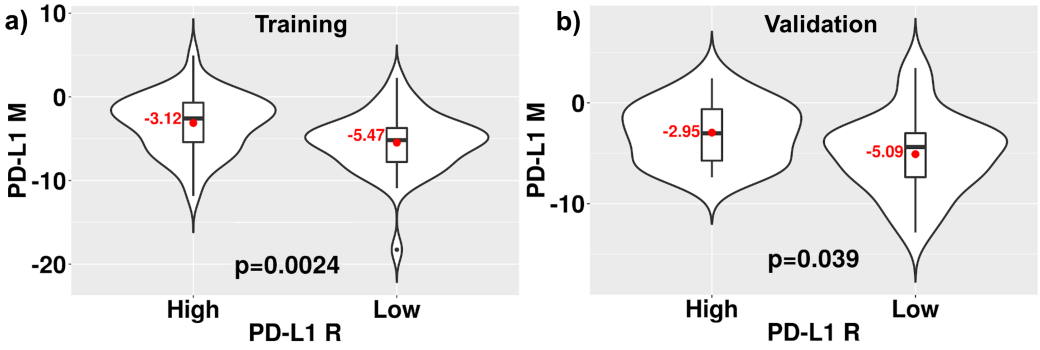


Figure 4.31: PD-L1 M distributions in the two groups ("High" and "Low"), identified using the proposed PD-L1 R signature, for both the training (a) and the validation (b) set. The mean PD-L1 M value is reported in red. p: t-test $p-value$ between the PD-L1 M values in the 2 groups.

4.3.2 Association between PD-L1 R and overall survival

The association between the group splitting provided by PD-L1 R and the overall survival was assessed in the whole cohort. As can be noted in Figure 4.34, the Kaplan-Meier plot of the two groups showed a significant different survival (likelihood ratio test $p-value$: $p = 0.048$), with the PD-L1 R "High" group having a better prognosis. In addition, the Cox proportional hazard ratio analysis resulted in a hazard ration "High" vs "Low" of 0.69 ($95\%CI : 0.46 - 0.98$).

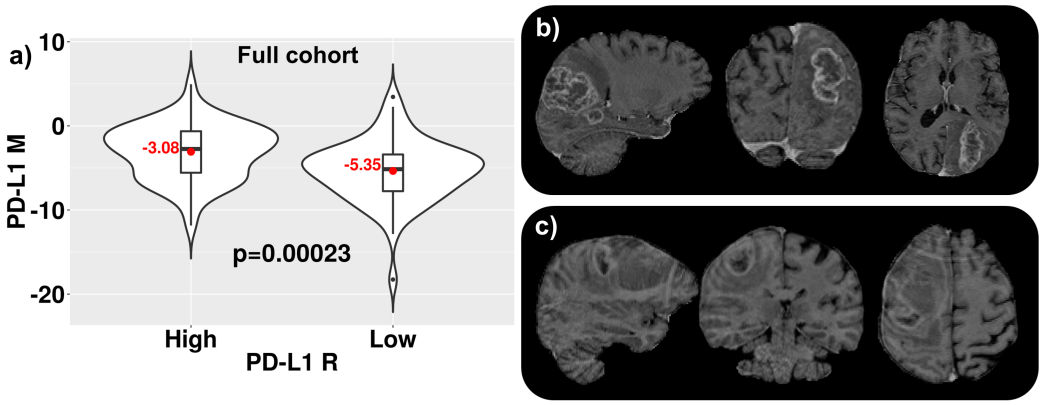


Figure 4.32: PD-L1 M distributions in the two groups ("High" and "Low"), identified using the proposed PD-L1 R signature, for the whole set (a). The mean PD-L1 M value is reported in red. The black panels on the right show the T1ce image of the subject with the highest PD-L1 M value in the "High" group (b) and the one with the lowest PD-L1 M in the "Low" group (c). p: t-test $p - value$ between the PD-L1 M values in the 2 groups.

4.3.3 Association between PD-L1 R and additional molecular covariates

To further characterize the molecular background of the two PD-L1 R high and low groups, the association between PD-1, SVZ, glioblastoma sub-type and prognosis was evaluated. Results reported in Figure 4.35 highlighted that two groups had statistically different PD 1 expression ($p - value = 0.015$), prognosis ($p - value = 0.014$) and GBM sub-types distribution ($p - value = 0.015$). In addition, for a subset of 61 patients the PD-L1 quantification based on the standard RNA sequencing was also available (instead of the PD-L1 M surrogate used for modelling). As demonstrated in Figure 4.36, the two groups identified with the same PD-L1 R signature, showed a statistically significant difference in PD-L1 RNA-seq ($p - value = 0.047$), corroborating the previous findings.

PD-L1 R and O^6-methylguanine-DNA methyltransferase (MGMT) promoter methylation status.

As final remark, the interaction between PD-L1 R and MGMT was studied. Figure 4.37a show the percentage of subjects belonging to PD-L1 R high and

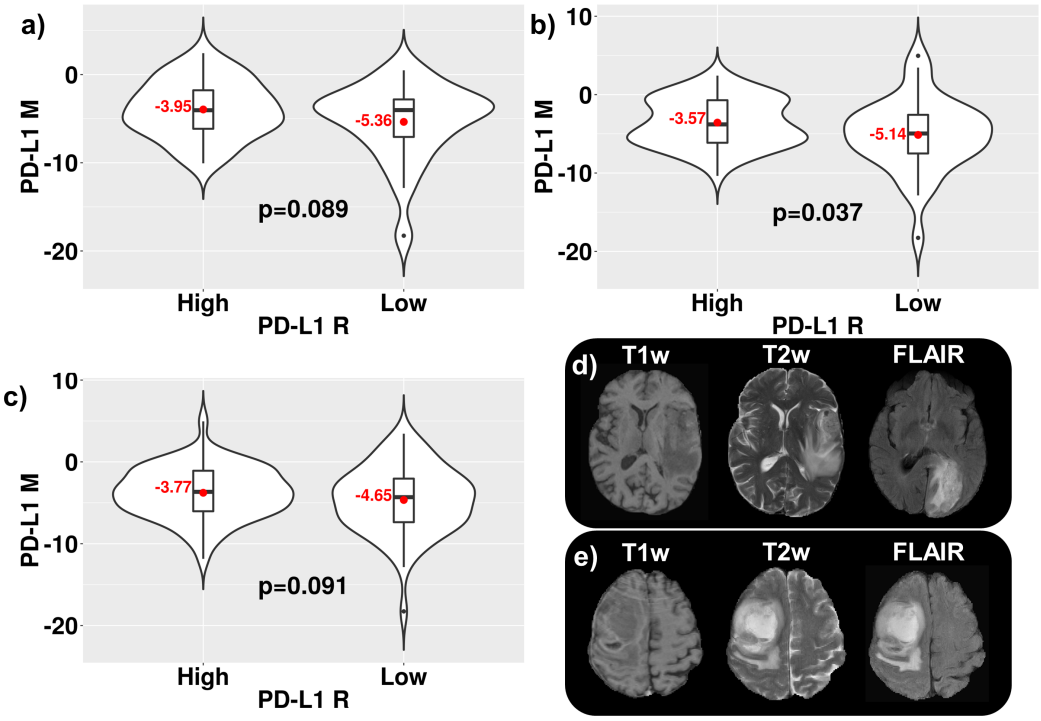


Figure 4.33: PD-L1 M distributions in the two groups ("High" and "Low"), identified using the proposed PD-L1 R signature, for the whole set (a). The mean PD-L1 M value is reported in red. The black panels on the right show the T1ce image of the subject with the highest PD-L1 M value in the "High" group (b) and the one with the lowest PD-L1 M in the "Low" group (c). p: t-test $p - value$ between the PD-L1 M values in the 2 groups.

low for both MGMT methylated (left) and unmethylated (right) groups.
The MGMT methylated group had a higher percentage of PD-L1 R "Low" subjects, while the MGMT unmethylated showed an opposite trend, having 53% of the subjects belonging to the PD-L1 R "High" group.
Of notice is the ability of the proposed PD-L1 R signature to stratify the MGMT unmethylated subjects into 2 groups with different survival ($p = 0.047$), with the PD-L1 R "high" group having again better prognosis (Fig. 4.37b).

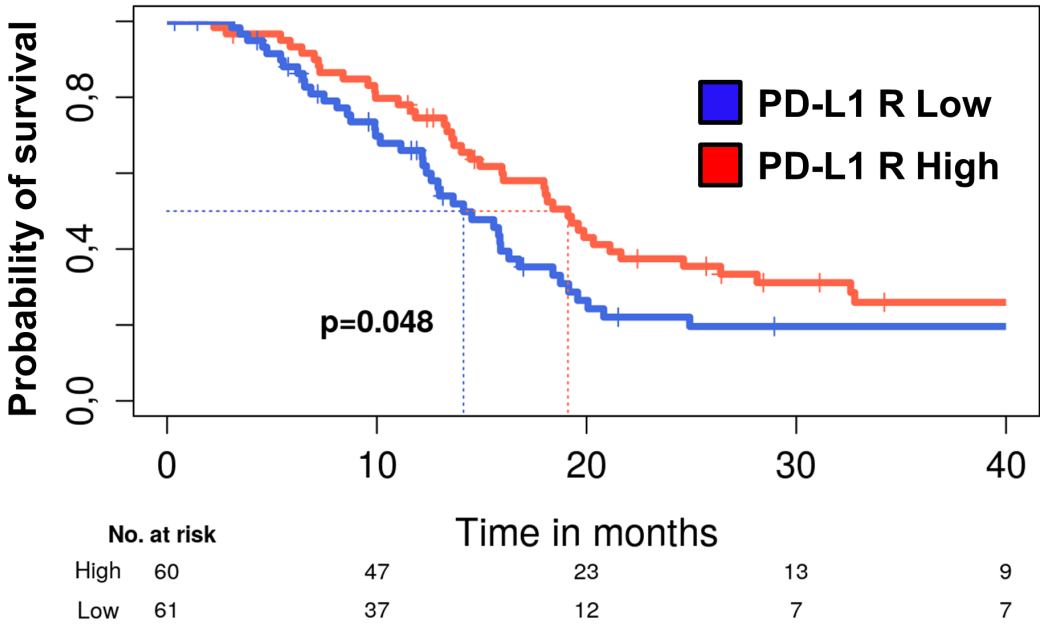


Figure 4.34: Kaplan-Meier plot showing the different overall survival between the "High" and "Low" group identified by the proposed radiomics signature. The likelihood ratio test $p - value$ was $p = 0.048$.

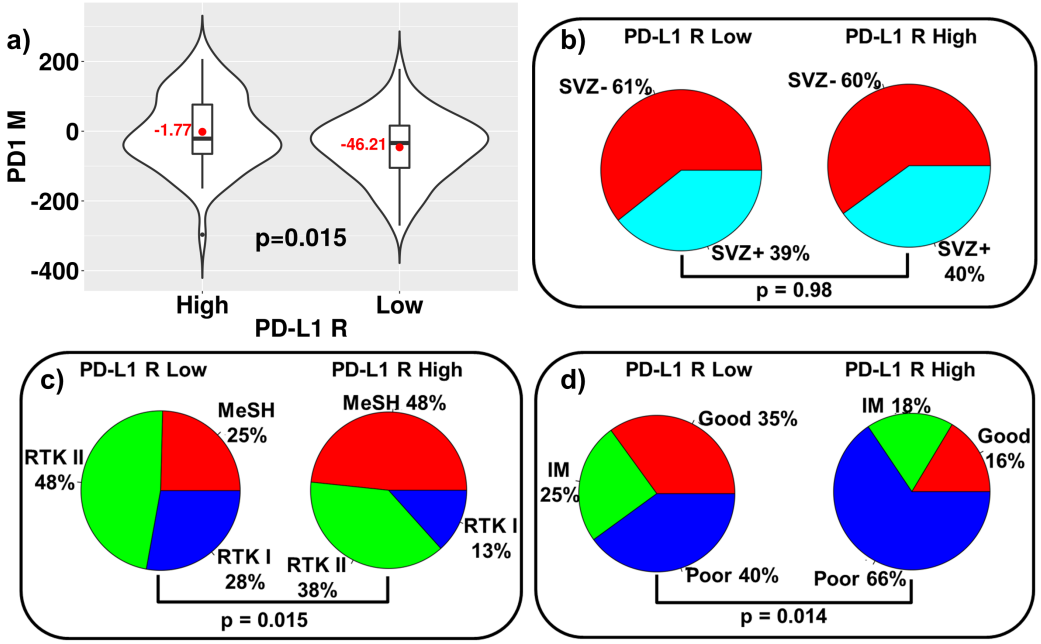


Figure 4.35: PD-1 distributions int the two PD-L1 R groups is reported in graph a). Pie charts in b) show the SVZ± distribution in the two groups, while those in panels c) and d) display, respectively, the distribution of the GBM sub-types and the prognosis in PD-L1 R Low and High. The $p-values$ are reported on the bottom of each sub-figure. MeSH: Mesenchymal; IM: Intermediate.

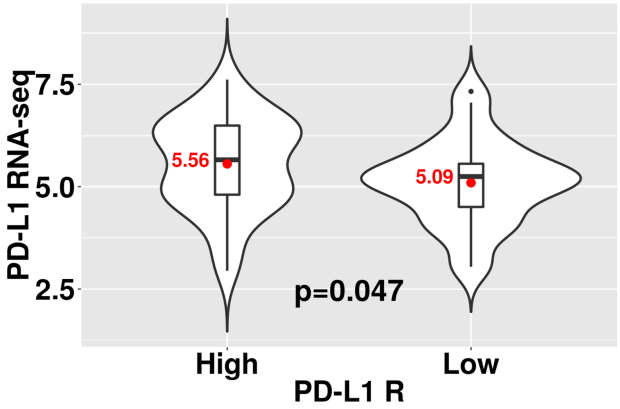


Figure 4.36: RNA sequencing-based PD-L1 distribution in a subset of 61 subjects, splitted using the proposed PD-L1 R signature.

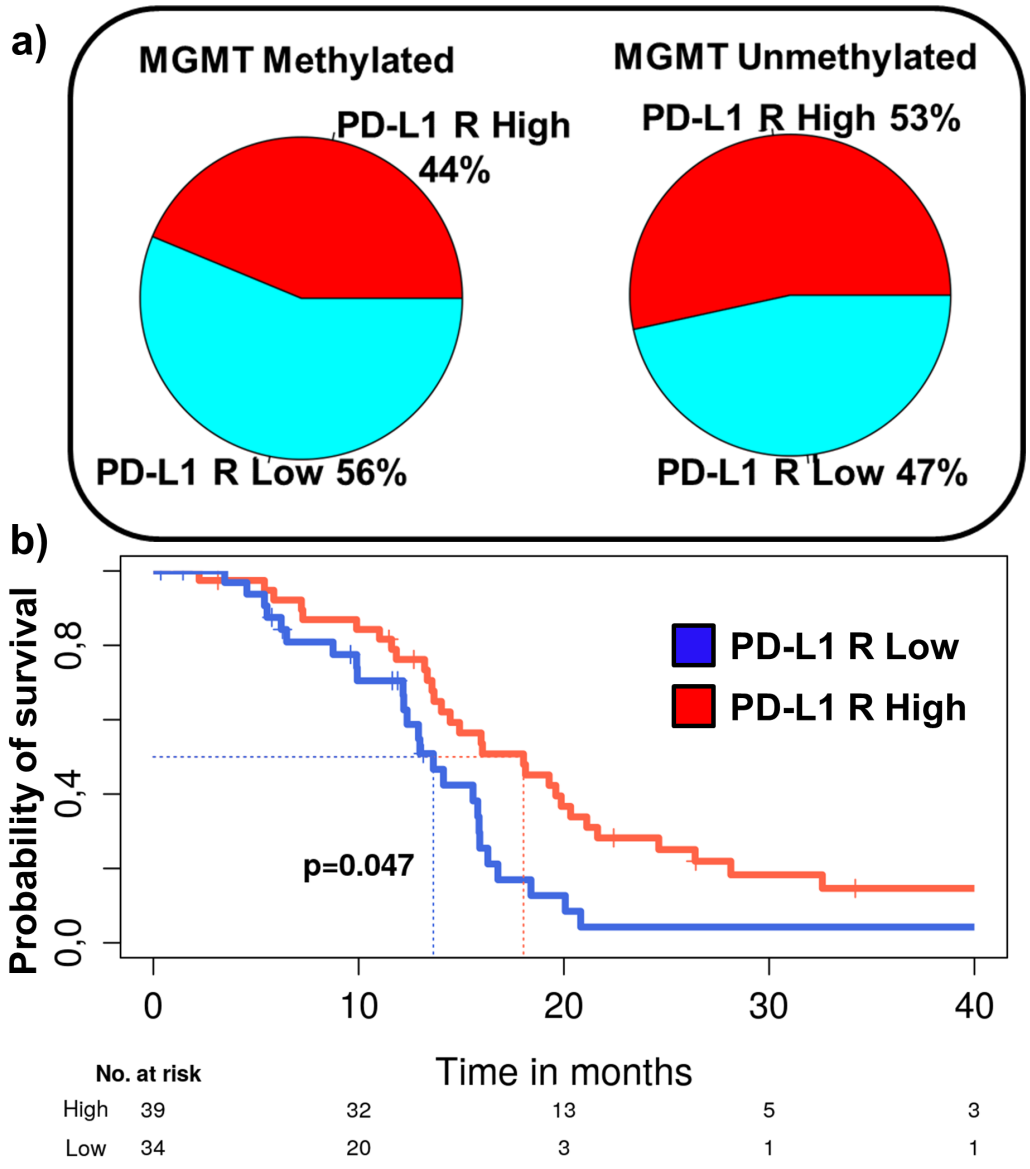


Figure 4.37: Panel a) shows the PD-L1 R distribution in both MGMT methylated (left) and unmethylated (right) subjects. The Kaplan-Meier plot (b) highlights the different overall survival between PD-L1 R high and low among the MGMT unmethylated subjects.

4.4 Tumor immune evolution determined by dynamic PD-L1 R in Glioblastoma patients treated with post-surgical radio-chemotherapy

- Assessment of the evolution of the previously identified PD-L1 R signature in follow up time points.
- High PD-L1 R status and delayed switch to PD-L1 R Low phenotype were found to be associated with better prognosis.
- PD-L1 R-Dyn was trained on pre-surgery time points yielding a better survival model.

The main idea behind this analysis and its implementation can be found in Figure 4.38. Briefly, in this feasibility study the aim was to monitor the GBM tumor immune response after radio-chemotherapy to identify possible evolution trajectories that might help to discover novel GBM immune-subgroups. This analysis studied how the immune-status in initially favorable tumors (PD-L1 R "High" group, as highlighted in the previous section) changed over time, to identify specific evolutionary pattern that might be associated with prognosis and help in further GBM tumor stratification.
In pre-surgergical time point (TP0), 49% of the tumors were assigned to PD-L1 R High. Post surgery and radio-chemotherapy (RCHT) 65% switched from high to low. The high group was stratified based on the median time to switch (high-early switch, h-e, $<$ 3.5 months, n=18, high-late switch, h-l, $>=$ 3.5 months). For n=20, the group threshold was not reached (h-miss) within median of 1.5 time points vs 7.5 and 6.5 for h-e and h-l. The groups h-miss, h-e and h-l showed median overall survival of 16, 16 and 28 months, respectively. Figure 4.39 shows the Kaplan-Meier plot between h-e and h-l subjects. As can be noted, the h-l group that showed a significantly better survival ($p = 0.00016$) compared to the early switch group.
The Cox proportional hazard model results confirmed this finding, with the Hazard ratio for Late vs Early switch being 0.3544 ($95\%CI$: $01748 - 0.7184, p = 0.004$).

4.4.1 PD-L1 R-Dyn model results

The next step of this analysis was to find a radiomics signature able to differentiate between early and late switch, using features extracted only from the pre-surgical time point. To this end, h-e and h-miss were merged based on their similar survival time to train PD-L1 R-Dyn. The identified radiomics signature consisted of seven features, which are reported in Table 4.13.

Feature	Image Type	Class	Name
F1	wavelet.LHH	first order	mean
F2	wavelet.LLH	first order	skewness
F3	wavelet.LLH	GLDM	LargeDependenceHighGrayLevelEmphasis
F4	wavelet.HHH	GLRLM	LongRunHighGrayLevelEmphasis
F5	log.sigma.2.0.mm	GLDM	LargeDependenceHighGrayLevelEmphasis
F6	log.sigma.4.0.mm	GLSZM	SizeZoneNonUniformityNormalized
F7	original	first order	minimum

Table 4.13: Radiomics signature associated to early/late switch in pre-surgical time point. L: Low; H: High; LOG: Laplacian of Gaussian; GLRLM: Gray Level Run Length Matrix; GLSZM: Gray Level Size Zone Matrix; GLDM: Gray Level Dependence Matrix.

PD-L1 R-Dyn performed well on a validation set of 16 subjects as demonstrated by an Area under the curve (AUC) of 0.84 (95%CI: 0.64-1) and an accuracy of 0.82 (see Table 4.14).

	AUC	Accuracy	Sensitivity	Specificity
PD-L1 R-Dyn	0.84	0.82	0.83	0.75

Table 4.14: PD-L1 R-Dyn results on a validation set of 16 subjects. AUC: Area Under the Curve.

Univariate analysis of PD-L1 R-Dyn in the full cohort yielded a hazard ratio (HR) of 1.71 (95%CI: 1.09-2.66). Multivariate survival analysis (MV) of PD-L1 R and PD-L1 R-Dyn yielded an AIC of 711.4, slightly outperforming

univariate survival models (AIC PD-L1 R: 715.0, PD-L1 R-Dyn: 712.2). HR of PD-L1 R-Dyn in MV was 1.70 (95%CI: 1.09-2.65). See also Table 4.15.

	PD-L1 R	**PD-L1 R-Dyn**	**PD-L1 R-Dyn + PD-L1 R**
Hazard Ratio	0.69	1.71	1.7
AIC	715.0	712.2	711.4

Table 4.15: Cox proportional hazard model results for the univariate (first two columns) and multi-variate (last column) analyses. The reference group for the PD-L1 R analysis was the "High" group while the on the for the PD-L1 R-Dyn was the "Early switch" group. AIC: Akaike information criterion.

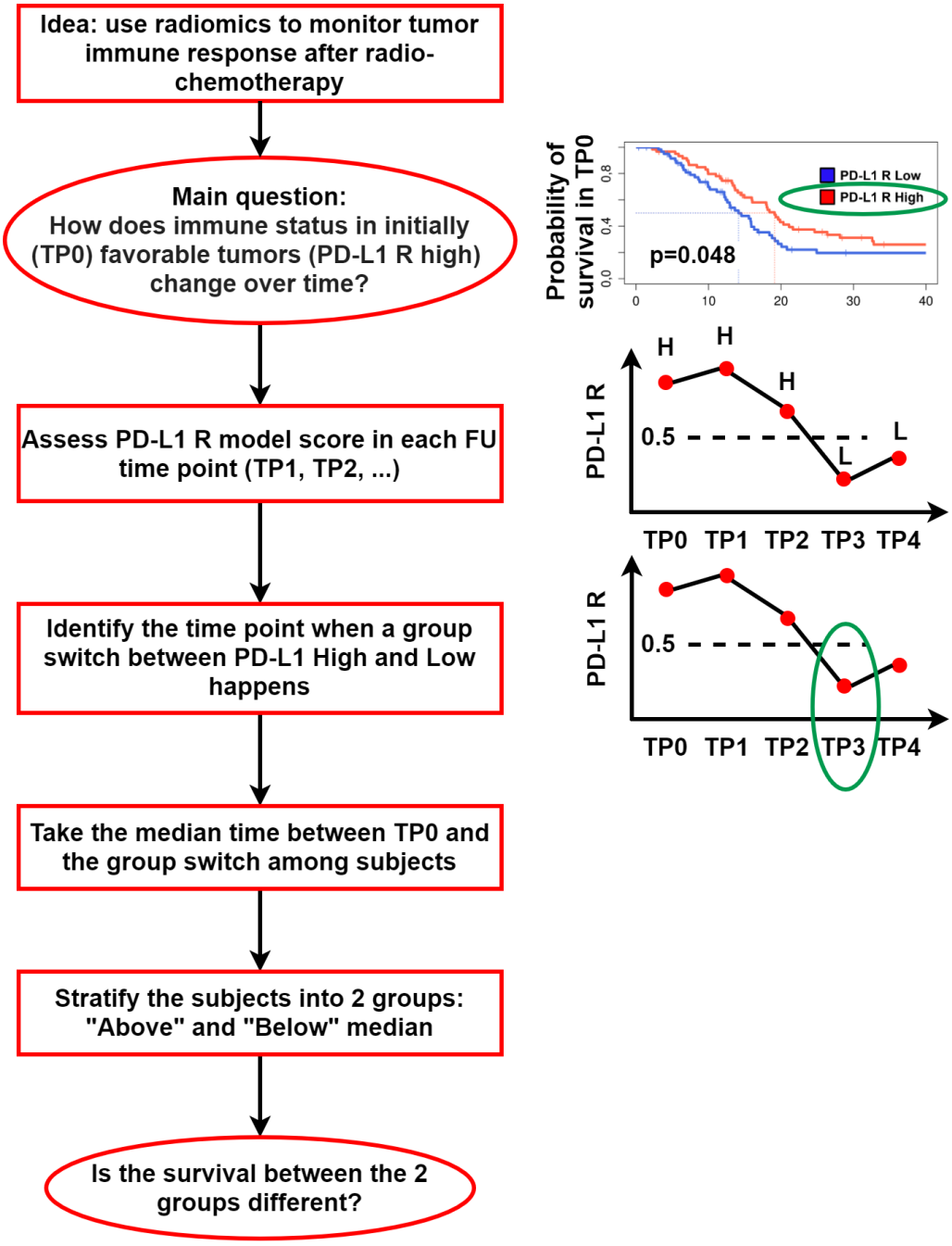


Figure 4.38: Rationale and implementation of the dynamic PD-L1 R model. H: PD-L1 R High; L: PD-L1 R Low; TP: time point

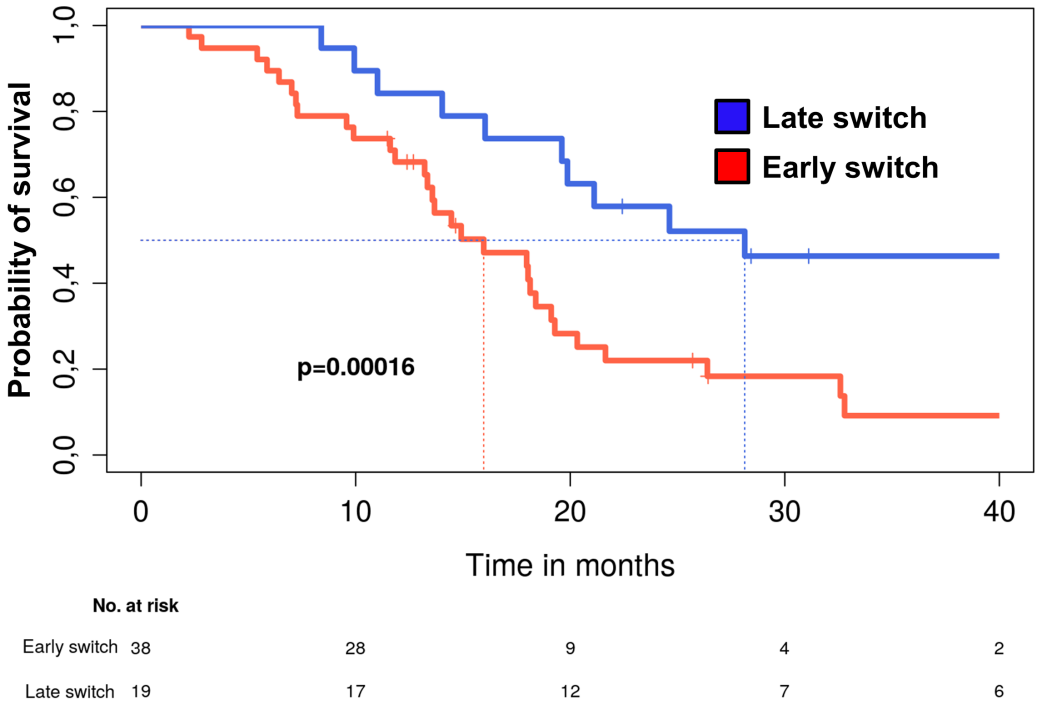


Figure 4.39: Kaplan-Meier plot for the patients labelled as PD-L1 R High in TP0, stratified based on the median time between TP0 and the first group switching (from high to low). Late switch means a transition from high to low occurred after the median switching time across subjects; early switch means before median.

4.5 Cohorts description

4.5.1 Primary glioblastoma

	Value		
Total Patients	320		
Sub-set with RT and MR-RT TP	142		
Sub-set labelled as MeSH or RTK I/II	121		
Sub-set with FU-TP (total number of FU)	115 (665)		
Total number of CT	5022		
Total number of MRI	3149		
Gender			
Male	196		
Female	124		
Age			
< 50	104		
50 - 70	167		
> 70	49		
Median Age (std)	57 ($\pm$14.14)		
Years of treatment	2006 - 2018		
Image parameters			
	TR $[ms]$	**VS** $[mm^2]$	**ST** $[mm]$
MRI-T1	1770	0.78x0.78	1
MRI-T1ce	1770	0.78x0.78	1
MRI-FLAIR	8500	0.89x0.89	5.5
MRI-T2	5500	0.57x0.57	5.5
MRI-ADC	8500	1.64x1.64	6.5
MRI-SWI	30	0.9x0.9	3
CT	-	0.32x0.32	1

Table 4.16: Primary glioblastoma (GBM) cohort overview. MeSH: mesenchymal; FU-TP: follow up time points; std: standard deviation; TR: Repetition Time; T1ce: T1 contrast enhanced; SWI: susceptibility Weighted Imaging; ADC: Apparent Diffusion Coefficient; ST: Slice Thickness; VS: Voxel Spacing; MRI: Magnetic Resonance Imaging; CT: Computed Tomography; ms: milliseconds; mm: millimeters.

4.5.2 Recurrent high-grade glioma

		Value	
Total Patients		115	
Gender			
Male		72	
Female		43	
Age			
< 50		48	
50 - 70		63	
> 70		4	
Median Age (std)		52 ($\pm$13.34)	
Years of treatment		2011 - 2018	
Image parameters			
	TR $[ms]$	**VS** $[mm^2]$	**ST** $[mm]$
MRI-T1ce	1740	0.42x0.42	1.3
MRI-FLAIR	9000	1x1	6

Table 4.17: Recurrent high-grade glioma (rHGG) cohort overview. std: standard deviation; TR: Repetition Time; T1ce: T1 contrast enhanced; ST: Slice Thickness; VS: Voxel Spacing; MRI: Magnetic Resonance Imaging; ms: milliseconds; mm: millimeters.

4.5.3 Pre-clinical mouse lung CT

Cohort	**CT**	**# mice**	**VS** $[mm^2]$	**ST** $[mm]$
A	323	1654	0.31x0.31	0.6
B	41	237	0.26x0.26	0.6
C	16	20	0.1x0.1	0.1

Table 4.18: Pre-clinical mouse lung CT cohorts overview. ST: Slice Thickness; VS: Voxel Spacing; mm: millimeters.

4.5.4 Non-small-cell lung cancer

	Value		
Total Patients	215		
Total number of CT	1052		
Gender			
Male	134		
Female	81		
Age			
< 50	4		
50 - 70	68		
> 70	143		
Median Age (std)	66 ($\pm$9.67)		
Years of treatment	2008 - 2019		
Image parameters			
	TR $[ms]$	**VS** $[mm^2]$	**ST** $[mm]$
CT	-	1x1	2

Table 4.19: Non-small-cell lung cancer (NSCLC) cohort overview. std: standard deviation; ST: Slice Thickness; VS: Voxel Spacing; mm: millimeters; TR: Repetition Time; ms: milliseconds; CT: Computed Tomography.

4.5.5 Anal adenocarcinoma

	Value		
Total Patients	238		
Total number of MR	2713		
Total number of CT	1905		
Gender			
Male	134		
Female	81		
Age			
< 50	46		
50 - 70	78		
> 70	114		
Median Age (std)	59 ($\pm$11.83)		
Years of treatment	2005 - 2019		
Image parameters			
	TR $[ms]$	**VS** $[mm^2]$	**ST** $[mm]$
MRI-T1ce	630	1.25x1.25	3
MRI-T1	630	1.25x1.25	4
MRI-T2	5000	0.98x0.98	6
CT	-	0.57x0.57	3

Table 4.20: Anal adenocarcinoma cohort overview. std: standard deviation; ST: Slice Thickness; VS: Voxel Spacing; mm: millimeters; TR: Repetition Time; ms: milliseconds; CT: Computed Tomography.

4.5.6 Head and Neck Squamous Cell Carcinoma

	Value
Total Patients	166
Gender	
Male	127
Female	39
Age	
< 50	28
50 - 70	131
> 70	7
Median Age (std)	58 ($\pm$8.42)

Table 4.21: Head and Neck Squamous Cell Carcinoma (HNCSCC) cohort overview. std: standard deviation.

Chapter 5

Discussion

Radiomics image analysis is a promising non-invasive tool that could help physicians in the clinical oncology decision making process. However, before it could be routinely applied in the clinical domain, some limitations have to be carefully addressed.

Data sorting and curation is the very first step in a radiomics workflow (or, more in general, in any medical imaging analysis workflow). It is a tedious and time consuming task that often needs to be performed manually by the researchers. Its automation is critical in order to allow the whole radiomics field to move into clinical environments.

Therefore, a new fully-automated data sorting and curation method, PyCURT, was developed in this work. It fuses together content-based CT and MR image classification and metadata-based radiotherapy (RT) data sorting and labelling, saving the final output in a well-structured format. To the best of our knowledge, this is the first tool merging both anatomical and contrast image classification and that takes into consideration RT data. This tool could facilitate analyses in distributed environments by automatically handling imaging data variability.

As part of PyCURT, a deep learning-based approach, BP-Class, was built up to identify the anatomical region scanned in an image (CT or MR). Very recently, efforts have been made to highlight the importance of body part labeling. Currently, the body part labelling depend on the information stored in the DICOM header, which often is missing or incorrect (for instance non-brain scan labelled as "brain"). Having a correct and consistent labelling would greatly simplify multi-center imaging studies. One proposed solution

was to adopt an universal naming scheme for the anatomical regions scanned that each institution should report in the DICOM header (Towbin et al., 2021). However, all the solutions currently proposed rely on the user to assign metadata data elements to newly created data, which is an error-prone and inefficient process. BP-Class solves the problem by automatically labelling an image based on the body part(s) scanned, enabling the users to automatically focus their analysis on the region(s) of interest. This would help reducing the inter-center bias and improving standardization and reproducibility.

The BP-Class classifiers (3 for CT and 2 for MR images) were trained in a one-vs-all manner to solve the open set recognition problem (Scheirer et al., 2013). Thereby, images that do not belong to any of the training classes will not be classified, making PyCURT able not only to predict the body part/ MR scan type of an image but also able to indicate if a scan belongs to an unknown anatomical region and/or MR contrast. This represents a step forward from previous approaches van der Voort et al. (2021), where classifiers were usually constrained to a specific number of classes, forcing the network to assign a new image to one class, even though that image did not belong to any of the classes used for training.
This in extremely useful when dealing with clinically exported data. When exporting data for a patient the images are dumped based on the patient ID. This results in having all the images associated with that ID, which can include, among other, knee or spinal examination, as well as image localizers. Those data are usually not of interest for a radiomics study and having a tool that automatically recognize them as "unknown" might be of great help for the researcher.
A very comprehensive analysis was also performed to find the best network structure for anatomical region classification. In particular, both 2D and 3D architectures were investigated to classify the body part in the CT images. Results (section 4.1.1) showed that, for this particular task, a 2D structure provides the best results in terms of classification accuracy. This can be explained by the fact that using a 2D network allows to extract a larger region of the image to be used as input for the classifier. Additionally, this region could encompass a full slice of the scan, enabling the network to learn features that capture the relative differences in appearance of the various anatomical regions (lung, thoracic cage, skull, ventricles, abdomen, etc.).

In addition, using a 2D approach multiple labels were available for the same 3D scan, resulting in each image to have assigned a continuous probability between 0 and 1, instead of a hard label. This is more suitable for

real-case applications, where acquired images do not encompass only one macro-region (i.e. or lung or abdomen) but are constituted of parts that come from different regions. The proposed method would assign various probabilities to each image, to reflect its anatomical parts distribution. For example, CT images acquired in the non-small cell lung cancer cohort contained at least a partial part of the head-neck region. In this example, two probabilities, one for head-neck and one for lung, would get assigned to one image.
By doing this, users can choose the body part of interest and a probability threshold specific for their study and only retain the images that PyCURT classified as the chosen part with that probability (or higher). As a consequence, even those images that are extremely different from those used for the network training can be still classified by just adjusting the threshold. This was the case, for instance, of the CT containing only one lung; they are still classified as lung even though the probability was lower than a standard lung CT.

A further central part of PyCURT is the RT data sorting, indeed it is the first tool that provides a way to automatically handle RT data. PyCURT performs multiple checks in order to find the best RT data coupling. It first checks the DICOM information stored in RT plan, as described in the methods; however, some times the acquisition or study date in the DICOM header could be missing or wrong, leading to a inaccurate sorting. To avoid this, PyCURT compares the RT plan UID with all the other scans in one subject's folder and force scans with the same UIDs to be in the same RT session. Furthermore, if planning CT for one RT session cannot be extracted from the RT plan, PyCURT first extracts all the series descriptions from the identified planning CT within the considered cohort, then compares those names with the series description in every other CT. If it finds a match, that CT is then labelled as planning CT.
PyCURT supports also the extraction of the various ROIs stored in the RT structure set. More interestingly, it implements a dose-based extraction of the ROIs. This is particularly useful in real-case scenario when unwanted structures might be saved in the same structure set. For example, it happens that multiple radiographers delineate the same contour (i.e. gross tumor volume) but then only one is used to plan and delivery the irradiation. For radiomics analysis, it is important to use the GTV segmented with the highest confidence (i.e. the one used for planning) and PyCURT can extract it automatically.

The RT data sorting was developed based on particle irradiation but was also successfully applied to sort cohorts with photon-based RT, proving its robustness.
One future PyCURT improvement will be to include the support for RT boost plans that are currently ignored.

Finally, PyCURT labels the imaging sessions for one subject based on the RT time-point (if present). By looking at the sessions' content and their temporal distance from (to) the RT, it automatically assigns them to one of the following groups: MR-RT, pre-RT, post-RT, FU (follow up). This scheme is based on the guidelines provided by the German Cancer Consortium (DKTK). This labelling make the subsequent analyses much more easy to handle, since the users can promptly access to the data of their interests.

Once the data are properly sorted into a well-defined structure, it is easier to then use them to build predictive model. This can be done using end-to-end method, like deep learning, and/or using engineered features, i.e. using radiomics. Deep learning methods have found applications in many research areas of medical image analysis, ranging from image acquisition to image registration, segmentation, and classification [8,9]. Convolutional neural networks (CNN) can automatically learn discriminative features directly from medical images and use them to build model for cancer diagnosis (Sun et al., 2020; De Fauw et al., 2018; Kermany et al., 2018). However, the current lack of interpretability of deep learning models limits their applicability in clinical context (Parekh and Jacobs, 2019).
On the other hand, radiomics approaches are much more transparent and interpretable and could be easier to translate into clinical domain.

In order to apply radiomics analysis in clinical contexts, both the lack of standardization and reproducibility during image preparation, have to be tackled. Image preparation includes all the steps before the final modelling, going image pre-processing to region of interest segmentation (Lambin et al., 2017).
In the last years, most of the efforts to account for the lack of standardization in radiomics focused on features definition (Zwanenburg et al., 2020) and to provide guidelines on how to implement a complete workflow (Lambin et al., 2017). However, there are no infrastructures available to help researchers in building the whole radiomics pipeline in a structured, easy and reproducible way.

To overcome the previous issue, a new abstract python-based software,

RADIANTS, has been developed during this thesis. It provides a framework to assemble any custom radiomics analysis. It comes with few workflow already available to perform standard processes included in a radiomics workflow. For example, it includes a basic pre-processing workflow (using a state-of-the-art brain extraction tool (Isensee et al., 2019)), a multi-contrast image registration workflow (Avants et al., 2008), and two in-house developed deep learning-based segmentation methods (for brain gross tumor volume and pre-clinical lung segmentation).
As previously mentioned, there are no current tools that provide exhaustive image processing pipelines for radiomics analysis. There are general guidelines on what to perform (Lambin et al., 2017) and several independent methods that can be used to perform image pre-processing (Tustison et al., 2010; Isensee et al., 2019), registration (Avants et al., 2008; Jenkinson and Smith, 2001), segmentation (Wolf et al., 2004; Zhou et al., 2019) and features extraction (Van Griethuysen et al., 2017; Wolf et al., 2004). However, linking all of them together to obtain the best processing workflow for a new radiomics study is not straightforward. For example, different tools might require distinct input format (i.e. Nifti or Nrrd) or rely on slightly different preliminary steps (for instance different image normalization procedures) or that are built using a specific number of contrast as input (for instance, most of the deep learning-based methods require 4 MR contrast to provide tumor segmentation).
RADIANTS helps researchers in combine all the discussed steps together, with minimal efforts and in a reproducible manner.

RADIANTS builds upon Nipype (Gorgolewski et al., 2011). It extends Nipype by adding an additional layer to source the data based on a standard PyCURT sorted output. In this way, it can manage a more general data structure. For example, Nipype assumes that all the sessions within a subject have the same scan(s). This might be true for research applications but does not apply to clinical cohorts. Moreover, it exploits the session labelling provided by PyCURT to figure out what kind of processes have to be run. For instance, if the MR-RT time-point is present, RADIANTS will automatically add a non-linear registration step to normalize all the MR images in the subsequent time-points to the planning MR.

RADIANTS has a modular structure that makes it easy to add/change/remove individual process(es). Users can customize it with their own or preferred tools with minimal efforts. Furthermore, it keeps track of the already processed inputs to avoid recomputing and save time. As before, it extends

Nipype capabilities by looking at pre-computed inputs in the working directory, while Nipype only checks the cache directory.
In case a process has a missing input and it is dependent on another RADIANTS node, it will automatically create the execution graph to run all the requested dependencies.

As previously mentioned, RADIANTS also incorporates two in-house developed tools for image segmentation. This is extremely helpful since, to date, region of interest delineation are mostly performed by hand or using (semi-)automatic standalone tools that have to be manually incorporated in the radiomics workflow.
The first method was built to segment the gross tumor volume (GTV) in high grade gliomas. A nnUNet framework (Isensee et al., 2021) was trained using T1 post contrast agent injection (T1ce) and FLAIR from both recurrent high-grade glioma and primary glioblastoma tumor cohorts. The method was thoroughly evaluated for segmenting GTV in both pre-surgery and follow up imaging time-point.
The nnU-Net framework has been proven able to outperform several other approaches in the image segmentation domain (Isensee et al., 2021). Here, it was trained using both T1ce and FLAIR MR contrasts, as they are now the recommended images to use for target volume segmentation in radiotherapy (Niyazi et al., 2016). In contrast to most of the state-of-the-art deep learning based methods (Kickingereder et al., 2019), which rely on four MR contrasts (T1, T1ce, T2 and FLAIR) in order to segment the tumor, the proposed method requires only two channels as input. This is particularly useful since requiring only two contrasts, instead of four, would allow to include much more imaging time points, increasing the power of the model.
In addition, it was also investigated whether the inclusion of the FLAIR contrast would provide a significant contribution to the overall accuracy of the segmentation. Results suggest that using FLAIR contrast as second channel would highly improve the final segmentation, as demonstrated by the median Dice coefficient of 0.91 compared to 0.77 when using only T1ce. This supports the current standard of using both MR contrasts to delineate the GTV during the radiotherapy treatment planning. However, this also demonstrates that even one single channel could provide an indication about the presence of a tumor and facilitate the manual delineation.

To further elucidate the ability of this method to provide robust GTV segmentation in multiple scenarios, it was tested on 665 follow up time-points of the primary glioblastoma cohort. The network segmented at least

a GTV in 553 time-points, and only 51 of them were manually labelled as usable or wrong (9%). By taking a closer look at those GTV, most of the T1ce images used for the inference showed artefacts (mostly due to motion or aliasing) and/or a poor image resolution, especially in the slice direction. This supports the urge of a standardized MR image acquisition protocol, with minimum parameters requirements (Kaufmann et al., 2020). The 112 time-points that were not segmented by the network were also manually checked. In a total of 48 time-points (43%), the human rater identified at least one GTV. Most of them appeared like a thin and long infiltration close to the resection cavity. This shape might be the explanation why the network was not able to properly segment them.

Finally, both a quantitative assessment of the network performance in FU time-points and an inter-rater variability evaluation, were performed. For those purposes, three independent raters, with different level of expertise, were asked to segment the same randomly selected FU time-points. All of them were compared against the network segmentation and to each other, using Dice score and Hausdorff distance. Results of the quantitative analysis between the manual and the automatic segmentation showed a good agreement across raters. Conversely, inter-rater analysis revealed a high variability, with Dice coefficients ranging between 0.63 and 0.77. This is a well-known issue, present in several tumor entities (Kim et al., 2016; Riegel et al., 2006; Weltens et al., 2001) that encourages the need of automatic and bias-free method for GTV contouring.

Despite good overall performance of the proposed method, some limitations remain. The main limitation is the inability of the network to properly segment thin tumor infiltration. This was present in the follow up time-points segmentation, as already discussed above, as well as in the pre-surgery time-points. This is a huge problem also for human operator, since most of the times infiltration are not easy to discriminate. By comparing the GTV segmentation provided by the network and the human contouring could help understanding the features that are important to differentiate between different tumor compartments.

Another limitation is the segmentation of resection cavity. By looking at the results, it seems that the network tends to provide a more conservative GTV delineation, not including the whole cavity. This might be the result of another well-known problem, the missing of clear and standard guidelines on how to contour resection cavity (Minniti et al., 2021). This lack of standardization might reflect to high heterogeneity in the GTV used to train the

method, causing the network to do not handle properly the resection cavity delineation. One possible approach to alleviate this issue could be to train a network to distinguish between GTV including the resection cavity and GTV without it. Then classify new GTV based on the new network and then repeat the GTV segmentation training using only GTV without (or with, if of interest) the cavity.
As a final general note on segmentation tasks, caution should be taken during network training/evaluation. Indeed, what it usually adopted as "reference" always come from manual or semi-automatic segmentation, that, as it has been demonstrated by the inter-rater analysis, is highly dependent on the operator. One possible solution could be to train/validate automatic methods using multiple segmentation coming from different raters and/or method, with the goal to let the network learn the optimal features from them. Also, using additional metrics to the standard Dice score and Hausdorff distance, might help to better characterize the segmentation accuracy.

The second segmentation tool available in RADIANTS provides multi-resolution lung segmentation from CT images.
To the best of our knowledge, this is the first fully automatic method for lung segmentation that can be applied to low- and high-resolution CT mouse data of both healthy and fibrotic lungs.
In two independent mouse CT datasets it achieved high precision, indicated by Dice scores above 96% and Hausdorff distances (HD) below 0.5mm, confirming both almost perfectly overlap and no significant shape mismatches. Here, the 95^{th} percentile of the HD was taken, since is widely used for example in the evaluation of brain tumor segmentation (Menze et al., 2015), and brain extraction (Isensee et al., 2019) as it allows to overcome the high sensitivity of the maximum Hausdorff distance to outliers.

In contrast to the method developed in the paper from Schoppe et al. (Schoppe et al., 2020), where the authors presented a deep learning based method to segment multiple organs in healthy mice, the proposed method can be successfully applied to segment lungs from mice with different level of fibrosis. This can be extremely useful in translational research on, for example, the radiation induced side effects after thoracic radiotherapy. Furthermore, it can be applied to CT images with different resolution, and not only to high resolution micro-CT scans.

More detailed analyses were performed to identify areas in which the CNN might yield less reliable results. These analyses revealed that voxels close to the borders of the lungs were more consistently misclassified. Interestingly,

two hot-spots of discordant segmentation were observed. In particular, the area often over-segmented by the network is close to the air-tissue interface where there is no clear separation between contiguous voxels (for example due to partial volume effects).
To minimize any possible registration error, a population-representing mouse template was built. This template provides a meaningful reference for the registration of mice with both fibrotic and non-fibrotic lungs. The resulting non-linear registrations were then carefully checked, and no major failure was noticed. In contrast, selecting a random mouse from the test set as a reference, yielded poor registration results, probably due to substantial anatomical differences.

A proof of concept analysis was performed to investigate the ability of this network to segment lungs from high resolution mouse and human lung CT images.
It has been trained twice using a transfer learning approach on only 20 mice and 10 human subjects.
On a small testing cohort of 4 high resolution mice it yielded a median DSC of 0.91, which is slightly better than the 0.90 reported in Yan et al. (Yan et al., 2017), using dynamic contrast-enhanced micro-CT images having resolution of 0.15 mm isotropic (which is comparable to that used here which is 0.1 mm). Good results were also achieved by the second re-trained network that was tested on 3 human subjects, comparable with other methods available in the literature (Ayad et al., 2020).
Further evaluation is anyhow needed in order to fully validate these additional segmentations.

Fibrotic Index and the histogram of the grey values within the lungs are often reported in studies dealing with fibrosis development (Zhou et al., 2017). For this reason, to justify the usage of our method over the semi-automatic segmentation, those two features were calculated using both masks and the results were compared to check for any statistically significant similarity. Both comparisons showed equivalence as inferred by the TOST approach. These results corroborate the usage of this method to segment the lungs and to use the extracted mask for any further analysis.

One limitation of this method is the 2D network structure used in the current approach. This architecture inherently fails to capture global context from adjacent slices, which might explain the over-segmented spots in the mismatch analysis. This issue might be (at least partially) solved using a 3D network architecture.

The transfer learning approach also showed some drawback. For example, the network over-segmented the trachea in the human test set, that might depend still on the 2D architecture and/or on a heterogeneous training set. The lung edges were slightly eroded in both high resolution and human test sets, probably due to the huge difference in image size between the clinical mouse CT and the other two datasets.

The second part of this thesis dealt with applications, to demonstrate the powerfulness of a robust radiomics workflow.
The rationale behind these applications was to show how to establish a radiomics workflow as potential tumor immune-monitoring tool under radiotherapy.
However, thanks to its flexibility, it is straightforward to adapt the developed radiomics infrastructures to study different cancer types and, in general, different research question(s).
For the applications in this thesis, the infrastructures previously developed were applied to characterize the glioblastoma (GBM) tumor immune biology, here described based on PD-L1, and to monitor its evolution after radiochemotherapy.

Despite PD-L1 expression has been linked to patient prognosis in different cancer types, its role in GBM has remained elusive. While a positive link was found between PD-L1 expression and better prognosis in several tumors (colorectal cancer (Li et al., 2016), breast carcinomas (Schalper et al., 2014), pulmonary squamous cell carcinoma (Yang et al., 2016) and non-small cell lung cancer (Velcheti et al., 2014)) in glioblastoma this does not hold true. Indeed, part of the studies trying to address this issue did not find any correlation, while some others observed a negative impact of higher PD-L1 expression over survival. However, no study tried to apply radiomics to unravel this relationship.
In this work, an imaging-based radiomics signature (PD-L1 R), built on T1ce MR images, was linked to PD-L1 M expression in a primary glioblastoma cohort. This signature was able to stratify the patients into two separate groups, PD-L1 R "High" and "Low", with statistically different PD-L1 expression and, interestingly, a different overall survival, with "High" group exhibiting a favorable prognosis, in line with other cancer entities.

Features coming from other three MR contrasts (T1, T2 and FLAIR) were used to calculate PD-L1 R scores, to evaluate its robustness. Remarkably, the two groups identified in each modality showed different PD-L1 M distribution

demonstrating that the PD-L1 R feature components are highly robust and do not depend on the MR image contrast.

The PD-L1 surrogate used here for modelling (PD-L1 M) is a methylation-based PD-L1 quantification that has been demonstrated to be highly correlated to PD-L1 in several tumors, including head and neck squamous cell carcinoma and low grade glioma (Knoll et al., 2020). For sake of robustness, the distributions in the two groups of the PD-L1 RNA-based quantification scores, available for a subset of the whole cohort, were also calculated. The PD-L1 RNA distributions in the two groups were also significantly different, corroborating the findings.

Molecular associations between PD-L1 R "High" and "Low" also revealed some noteworthy results. For example, GBM sub-type distribution, based on DNA methylation (Capper et al., 2018), showed an enrichment of mesenchymal tumor sub-type (48%) in the PD-L1 "High" group, while the "Low" group had a predominance of RTK II tumors (48%). The relative abundance of mesenchymal in the group with better prognosis might indicate a favourable response to immunotherapy and intensive radio- and chemotherapy, as previously described (Behnan et al., 2019).
Patient prognosis, as defined in Knoll et al. (2019), and tumor evolution-derived prognostic groups were also found to have statistically different distributions in the two PD-L1 R groups. On the other hand, sub-ventricular zone (SVZ) classification did not show any difference in the PD-L1 R groups. Finally, the interaction between O^6-methylguanine-DNA methyltransferase (MGMT) and PD-L1 R was investigated. Notably, the stratification of the MGMT unmethylated subjects using PD-L1 R resulted in two groups with different survival, with, again, the PD-L1 R "High" group having better prognosis. On the contrary, stratification of the methylated subjects did not resulted in any difference in survival. Those results might suggest a interplay between DNA repair mechanisms and tumor immune micro-environment, particularly prominent in tumors with higher gene mutation.
To fully explain the relationship between PD-L1 and molecular covariates associated with tumor characteristics is beyond the scope of this work, however, using the developed radiomics model, some extremely promising indications were unveiled, which might shed light over the prognostic value of PD-L1 in glioblastoma tumors.

To provide further evidence of the flexibility of a radiomics model, the PD-L1 R signature described earlier was applied to monitor the tumor im-

mune response after treatment. In particular, the PD-L1 R model scores were calculated in all the follow up time-points after the surgical resection of the tumor and the radio-chemotherapy.
Remarkably, longitudinal changes of the immune-surrogate PD-L1 R contained prognostic information, with the patients that had a prolonged post-treatment PD-L1 R "High" status over time having a better prognosis. This analysis demonstrates that using only information encapsulated in MR images may allow deciphering the dynamic of tumor immune status and help in glioblastoma tumor stratification.
It is important to highlight that this kind of longitudinal analysis is only possible using an imaging-based radiomics approach. Indeed, it is impractical to rely on repeated biopsies to monitor PD-L1 expression over time. Here, without any further burden for the patient and additional costs, all the images acquired routinely during the treatment follow up were used to non-invasively study the temporal evolution of PD-L1 and correlate its dynamic with clinical outcome.

Summarized, a complete new framework was developed to help automatize and standardize all the major steps included in an imaging-based radiomics workflow. It was then applied to study the glioblastoma tumor immune status and its evolution after treatment. Overall, the work performed in this thesis could help fostering translational research in clinical radiation oncology.

Chapter 6

Summary

Radiomics is an evolving research field where spatially resolved quantitative features, extracted from standard-of-care medical images, are applied within clinical-decision support systems for precision diagnosis and treatment adaptation, providing a powerful tool in modern medicine. However, this field still lacks standardized and reproducible analysis workflows which currently limits its implementation into clinical routine. Therefore, this work focuses on developing a flexible and robust analysis pipeline to enable radiomics to be reliably and robustly applied in clinical decision making. The discovery of radiomics features correlating with tumor immune (PD-L1) status and its 4D longitudinal evolution after radiotherapy in patients with glioblastoma tumors in this thesis emphasizes the power of radiomics and the devised platform for identification of new imaging-based biomarker and therapy surrogates.

To facilitate data curation for images coming from different clinical institutions, a new python-based tool has been developed, PyCURT, which exploits both content and metadata-based data sorting techniques. In particular, a novel deep learning-based approach, BP-Class, was implemented as part of PyCURT to automatically classify the anatomical regions scanned in each MR or CT images, respectively. It can differentiate between head-neck, abdominal-pelvic and lung regions with high accuracy. Furthermore, PyCURT uses DICOM metadata to detect and link together the radiotherapy data. By doing that, PyCURT allows a comprehensive and robust multicentric medical image curation.

To alleviate the issue of reproducibility and workflow standardization, a second tool was built, RADIANTS, that provides a flexible and easy to use

framework to configure custom radiomics analyses in a reproducible setting. It comes with standard radiomics workflow for pre-processing and registration. In addition, two deep learning-based segmentation methods were developed and added into RADIANTS. The first method was able to accurately segment lung from CT images of both fibrotic and healthy mice, as demonstrated by high Dice score and low Hausdorff distance. After re-training using a transfer learning approach, it was also able to segment lungs from high resolution mice and human CT images. The second approach was trained to segment gross tumor volumes (GTV) from pre-treatment MR images in high-grade gliomas, where it achieved accurate results both in terms of Dice score and Hausdorff distance. It was then investigated whether it could provide reliable GTV segmentation also in follow up time points. Finally, delineations from the proposed method were compared with those manually contoured by three independent human raters with results showing similar variability between network and manual segmentation as compared to higher inter-rater heterogeneity.

The previously established tools were then translated to address a clinically relevant question. Could tumor PD-L1 status be correlated with radiomics features to identify and track tumor immune status? A radiomics-derived score (PD-L1 R score) was identified from pre-treatment MR images enabling prognostic stratification of patients into two distinct cohorts. Patients with high PD-L1 expression tumors had a better prognosis. Finally, a feasibility study showed promising results demonstrating that the dynamic change of PD-L1 R radiomics-surrogate over time could provide further prognostic stratification of patients with glioblastoma.

In summary, a comprehensive analysis framework was built to solve current radiomics limitations, paving the way towards the radiomics breakthrough into clinical applications, where radiomics might have the ability to serve as an attractive mean for tumor immune monitoring.

Kapitel 7

Zusammenfassung

Radiomics ist ein sich entwickelndes Forschungsgebiet, in dem räumlich aufgelöste quantitative Merkmale aus medizinischen Standardbildern extrahiert werden. Diese werden in klinischen Entscheidungsunterstützungssystemen für die Präzisionsdiagnose und die Anpassung der Behandlung eingesetzt, was ein leistungsstarkes Instrument in der modernen Medizin darstellt. Allerdings fehlt es in diesem Bereich noch an standardisierten und reproduzierbaren Analyseabläufen, was die Anwendung in der klinischen Routine derzeit einschränkt. Daher konzentriert sich diese Arbeit auf die Entwicklung einer flexiblen und robusten Analysepipeline, die es ermöglicht, Radiomics zuverlässig und robust in der klinischen Entscheidungsfindung einzusetzen. Die Entdeckung von Radiomics-Merkmalen, die mit dem Tumor-Immunstatus (PD-L1) und dessen 4D-Longitudinalentwicklung nach Strahlentherapie bei Patienten mit Glioblastom-Tumoren in dieser Arbeit korrelieren, unterstreicht die Leistungsfähigkeit von Radiomics und der entwickelten Plattform zur Identifizierung neuer bildgebender Biomarker und Therapiesurrogate.

Zur Erleichterung der Datenkuration für Bilder, die aus verschiedenen klinischen Einrichtungen stammen, wurde ein neues Python-basiertes Tool, PyCURT, entwickelt, das sowohl inhalts- als auch metadatenbasierte Datensortierungstechniken nutzt. Insbesondere wurde ein neuartiger, auf Deep Learning basierender Ansatz, BP-Class, als Teil von PyCURT implementiert, um automatisch die anatomischen Regionen zu klassifizieren, die in den einzelnen MR- bzw. CT-Bildern gescannt wurden. Er kann mit hoher Genauigkeit zwischen Kopf-Hals-, Bauch-Becken- und Lungenregionen unterscheiden. Darüber hinaus nutzt PyCURT DICOM-Metadaten, um die Strahlentherapie-Daten zu erkennen und miteinander zu verknüpfen. Auf die-

se Weise ermöglicht PyCURT eine umfassende und robuste multizentrische medizinische Bildkuration.

Um das Problem der Reproduzierbarkeit und der Standardisierung von Arbeitsabläufen zu lösen, wurde ein zweites Tool, RADIANTS, entwickelt, das ein flexibles und einfach zu bedienendes Framework für die Konfiguration benutzerdefinierter Radiomicsanalysen in einer reproduzierbaren Umgebung bietet. Es enthält einen Standard-Radiomics-Workflow für die Vorverarbeitung und Registrierung. Darüber hinaus wurden zwei auf Deep Learning basierende Segmentierungsmethoden entwickelt und in RADIANTS integriert. Die erste Methode war in der Lage, die Lunge aus CT-Bildern sowohl von fibrotischen als auch von gesunden Mäusen genau zu segmentieren, was durch einen hohen Dice-Score und eine niedrige Hausdorff-Distanz belegt wurde. Nach einem erneuten Training mit Hilfe eines Transfer-Learning-Ansatzes war sie auch in der Lage, die Lungen von hochauflösenden Mäusen und menschlichen CT-Bildern zu segmentieren. Der zweite Ansatz wurde für die Segmentierung von Gross Tumor Volume (GTV) aus MR-Bildern von hochgradigen Gliomen vor der Behandlung trainiert, wo er sowohl in Bezug auf den Dice-Score als auch auf die Hausdorff-Distanz genaue Ergebnisse erzielte. Anschließend wurde untersucht, ob die GTV-Segmentierung auch in der Nachbeobachtungsphase zuverlässig ist. Schließlich wurden die mit der vorgeschlagenen Methode erstellten Abgrenzungen mit denen verglichen, die von drei unabhängigen menschlichen Gutachtern manuell konturiert wurden.

Die zuvor etablierten Werkzeuge wurden dann zur Beantwortung einer klinisch relevanten Frage eingesetzt. Kann der PD-L1-Status des Tumors mit radiologischen Merkmalen korreliert werden, um den Immunstatus des Tumors zu identifizieren und zu verfolgen? Aus MR-Bildern vor der Behandlung wurde ein aus der Radiomics abgeleiteter Score (PD-L1 R-Score) ermittelt, der eine prognostische Stratifizierung der Patienten in zwei unterschiedliche Kohorten ermöglicht. Patienten mit Tumoren mit hoher PD-L1 Expression hatten eine bessere Prognose. Schließlich zeigte eine Machbarkeitsstudie vielversprechende Ergebnisse, die belegen, dass die dynamische Veränderung des PD-L1 R-Radiomics-Surrogats im Laufe der Zeit eine weitere prognostische Stratifizierung von Patienten mit Glioblastom ermöglichen könnte.

Zusammenfassend lässt sich sagen, dass ein umfassender Analyserahmen geschaffen wurde, um die derzeitigen Beschränkungen der Radiomics zu überwinden und den Weg für den Durchbruch der Radiomics in der klinischen Anwendung zu ebnen. Hierbei könnte Radiomics als attraktives Mittel zum Monitoring des Tumorimmunsystems dienen.

Chapter 8

References List

Aerts, H. J., Velazquez, E. R., Leijenaar, R. T., Parmar, C., Grossmann, P., Cavalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., Hoebers, F., Rietbergen, M. M., Leemans, C. R., Dekker, A., Quackenbush, J., Gillies, R. J. and Lambin, P. (2014), **Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach**, Nature Communications *5*(1), 1–9. URL: https://www.nature.com/articles/ncomms5006

Avants, B. B., Epstein, C. L., Grossman, M. and Gee, J. C. (2008), **Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain**, Medical Image Analysis *12*(1), 26–41. URL: https://pubmed.ncbi.nlm.nih.gov/17659998/

Avants, B. B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J. and Gee, J. C. (2010), **The optimal template effect in hippocampus studies of diseased populations**, NeuroImage *49*(3), 2457–2466.

Ayad, H., Ghindawi, I. W. and Kadhm, M. S. (2020), **Lung Segmentation Using Proposed Deep Learning Architecture**, International journal of online and biomedical engineering *16*(15), 141–147. URL: https://www.online-journals.org/index.php/i-joe/article/view/17115

Bady, P., Delorenzi, M. and Hegi, M. E. (2016), **Sensitivity Analysis of the MGMT-STP27 Model and Impact of Genetic and Epigenetic**

Context to Predict the MGMT Methylation Status in Gliomas and Other Tumors, Journal of Molecular Diagnostics *18*(3), 350–361. URL: https://pubmed.ncbi.nlm.nih.gov/26927331/

Baeßler, B., Weiss, K. and Santos, D. P. D. (2019), **Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study**, Investigative Radiology *54*(4), 221–228.

Behnan, J., Finocchiaro, G. and Hanna, G. (2019), **The landscape of the mesenchymal signature in brain tumours**, Brain *142*(4), 847–866. URL: https://academic.oup.com/brain/article/142/4/847/5420896

Berghoff, A. S., Kiesel, B., Widhalm, G., Rajky, O., Ricken, G., Wohrer, A., Dieckmann, K., Filipits, M., Brandstetter, A., Weller, M., Kurscheid, S., Hegi, M. E., Zielinski, C. C., Marosi, C., Hainfellner, J. A., Preusser, M. and Wick, W. (2015), **Programmed death ligand 1 expression and tumor-infiltrating lymphocytes in glioblastoma**, Neuro-Oncology *17*(8), 1064–1075. URL: /pmc/articles/PMC4490866/ /pmc/articles/PMC4490866/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4490866/

Bibault, J. E., Giraud, P. and Burgun, A. (2016), **Big Data and machine learning in radiation oncology: State of the art and future prospects**, Cancer Letters *382*(1), 110–117.

Bø, H. K., Solheim, O., Jakola, A. S., Kvistad, K. A., Reinertsen, I. and Berntsen, E. M. (2017), **Intra-rater variability in low-grade glioma segmentation**, Journal of Neuro-Oncology *131*(2), 393–402. URL: https://pubmed.ncbi.nlm.nih.gov/27837437/

Brahm, C. G., van Linde, M. E., Enting, R. H., Schuur, M., Otten, R. H., Heymans, M. W., Verheul, H. M. and Walenkamp, A. M. (2020), **The current status of immune checkpoint inhibitors in neuro-oncology: A systematic review**, Cancers *12*(3), 12. URL: /pmc/articles/PMC7139638/ /pmc/articles/PMC7139638/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7139638/

Capper, D., Jones, D. T., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D. E., Kratz, A., Wefers, A. K., Huang, K., Pajtler, K. W., Schweizer, L., Stichel, D., Olar, A.,

Engel, N. W., Lindenberg, K., Harter, P. N., Braczynski, A. K., Plate, K. H., Dohmen, H., Garvalov, B. K., Coras, R., Hölsken, A., Hewer, E., Bewerunge-Hudler, M., Schick, M., Fischer, R., Beschorner, R., Schittenhelm, J., Staszewski, O., Wani, K., Varlet, P., Pages, M., Temming, P., Lohmann, D., Selt, F., Witt, H., Milde, T., Witt, O., Aronica, E., Giangaspero, F., Rushing, E., Scheurlen, W., Geisenberger, C., Rodriguez, F. J., Becker, A., Preusser, M., Haberler, C., Bjerkvig, R., Cryan, J., Farrell, M., Deckert, M., Hench, J., Frank, S., Serrano, J., Kannan, K., Tsirigos, A., Brück, W., Hofer, S., Brehmer, S., Seiz-Rosenhagen, M., Hänggi, D., Hans, V., Rozsnoki, S., Hansford, J. R., Kohlhof, P., Kristensen, B. W., Lechner, M., Lopes, B., Mawrin, C., Ketter, R., Kulozik, A., Khatib, Z., Heppner, F., Koch, A., Jouvet, A., Keohane, C., Mühleisen, H., Mueller, W., Pohl, U., Prinz, M., Benner, A., Zapatka, M., Gottardo, N. G., Driever, P. H., Kramm, C. M., Müller, H. L., Rutkowski, S., Von Hoff, K., Frühwald, M. C., Gnekow, A., Fleischhack, G., Tippelt, S., Calaminus, G., Monoranu, C. M., Perry, A., Jones, C., Jacques, T. S., Radlwimmer, B., Gessi, M., Pietsch, T., Schramm, J., Schackert, G., Westphal, M., Reifenberger, G., Wesseling, P., Weller, M., Collins, V. P., Blümcke, I., Bendszus, M., Debus, J., Huang, A., Jabado, N., Northcott, P. A., Paulus, W., Gajjar, A., Robinson, G. W., Taylor, M. D., Jaunmuktane, Z., Ryzhova, M., Platten, M., Unterberg, A., Wick, W., Karajannis, M. A., Mittelbronn, M., Acker, T., Hartmann, C., Aldape, K., Schüller, U., Buslei, R., Lichter, P., Kool, M., Herold-Mende, C., Ellison, D. W., Hasselblatt, M., Snuderl, M., Brandner, S., Korshunov, A., Von Deimling, A. and Pfister, S. M. (2018), **DNA methylation-based classification of central nervous system tumours**, Nature *555*(7697), 469–474. URL: https://www.nature.com/articles/nature26000

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., Tarbox, L. and Prior, F. (2013), **The cancer imaging archive (TCIA): Maintaining and operating a public information repository**, Journal of Digital Imaging *26*(6), 1045–1057. URL: https://link.springer.com/article/10.1007/s10278-013-9622-7

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D.,

Karthikesalingam, A., Hughes, C. O., Raine, R., Hughes, J., Sim, D. A., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees, G., Back, T., Khaw, P. T., Suleyman, M., Cornebise, J., Keane, P. A. and Ronneberger, O. (2018), **Clinically applicable deep learning for diagnosis and referral in retinal disease**, Nature Medicine *24*(9), 1342–1350. URL: https://www.nature.com/articles/s41591-018-0107-6 https://www.nature.com/articles/s41591-018-0107-6)

Fridman, W. H., Zitvogel, L., Sautès-Fridman, C. and Kroemer, G. (2017), **The immune contexture in cancer prognosis and treatment**, Nature Reviews Clinical Oncology *14*(12), 717–734. URL: https://pubmed.ncbi.nlm.nih.gov/28741618/

Goetz, L. H. and Schork, N. J. (2018), **Personalized medicine: motivation, challenges, and progress**, Fertility and Sterility *109*(6), 952–963. URL: https://pubmed.ncbi.nlm.nih.gov/29935653/

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L. and Ghosh, S. S. (2011), **Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in Python**, Frontiers in Neuroinformatics *5*. URL: https://zenodo.org/record/50186

Granton, P. V., Dubois, L., Van Elmpt, W., Van Hoof, S. J., Lieuwes, N. G., De Ruysscher, D. and Verhaegen, F. (2014), **A longitudinal evaluation of partial lung irradiation in mice by using a dedicated image-guided small animal irradiator**, International Journal of Radiation Oncology Biology Physics *90*(3), 696–704. URL: www.redjournal.org.http://dx.doi.org/10.1016/j.ijrobp.2014.07.004

Grossberg, A. J., Mohamed, A. S., El Halawani, H., Bennett, W. C., Smith, K. E., Nolan, T. S., Williams, B., Chamchod, S., Heukelom, J., Kantor, M. E., Browne, T., Hutcheson, K. A., Gunn, G. B., Garden, A. S., Morrison, W. H., Frank, S. J., Rosenthal, D. I., Freymann, J. B. and Fuller, C. D. (2018), **Data descriptor: Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy**, Scientific Data *5*(1), 1–10. URL: https://www.nature.com/articles/sdata2018173

Haarburger, C., Müller-Franzes, G., Weninger, L., Kuhl, C., Truhn, D. and Merhof, D. (2020), **Radiomics feature reproducibility under inter-rater variability in segmentations of CT images**, Scientific Reports *10*(1). URL: https://pubmed.ncbi.nlm.nih.gov/32728098/

Han, L., Wang, S., Miao, Y., Shen, H., Guo, Y., Xie, L., Shang, Y., Dong, J., Li, X., Wang, W. and Song, Q. (2019), **MRI texture analysis based on 3D tumor measurement reflects the IDH1 mutations in gliomas – A preliminary study**, European Journal of Radiology *112*, 169–179. URL: https://pubmed.ncbi.nlm.nih.gov/30777207/ https://pubmed.ncbi.nlm.nih.gov/30777207/?dopt=Abstract

Han, Y., Liu, D. and Li, L. (2020), **PD-1/PD-L1 pathway: current researches in cancer.**, American journal of cancer research *10*(3), 727–742. URL: /pmc/articles/PMC7136921/ /pmc/articles/PMC7136921/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7136921/ http://www.ncbi.nlm.nih.gov/pubmed/32266087

Hao, C., Chen, G., Zhao, H., Li, Y., Chen, J., Zhang, H., Li, S., Zhao, Y., Chen, F., Li, W. and Jiang, W. G. (2020), **PD-L1 expression in glioblastoma, the clinical and prognostic significance: A systematic literature review and meta-analysis**, Frontiers in Oncology *10*, 1015.

He, K., Zhang, X., Ren, S. and Sun, J. (2016), **Deep residual learning for image recognition**, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition *2016-Decem*, 770–778. URL: https://arxiv.org/abs/1512.03385v1

Hectors, S. J., Lewis, S., Besa, C., King, M. J., Said, D., Putra, J., Ward, S., Higashi, T., Thung, S., Yao, S., Laface, I., Schwartz, M., Gnjatic, S., Merad, M., Hoshida, Y. and Taouli, B. (2020), **MRI radiomics features predict immuno-oncological characteristics of hepatocellular carcinoma**, European Radiology *30*(7), 3759–3769. URL: https://pubmed.ncbi.nlm.nih.gov/32086577/

Herbst, R. S., Soria, J. C., Kowanetz, M., Fine, G. D., Hamid, O., Gordon, M. S., Sosman, J. A., McDermott, D. F., Powderly, J. D., Gettinger, S. N., Kohrt, H. E., Horn, L., Lawrence, D. P., Rost, S., Leabman, M.,

Xiao, Y., Mokatrin, A., Koeppen, H., Hegde, P. S., Mellman, I., Chen, D. S. and Hodi, F. S. (2014), **Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients**, Nature *515*(7528), 563–567. URL: https://pubmed.ncbi.nlm.nih.gov/25428504/

Hirsch, J. D., Siegel, E. L., Balasubramanian, S. and Wang, K. C. (2015), **We Built This House; It's Time to Move in: Leveraging Existing DICOM Structure to More Completely Utilize Readily Available Detailed Contrast Administration Information**, Journal of Digital Imaging *28*(4), 407–411. URL: https://link.springer.com/article/10.1007/s10278-015-9771-y

Hu, L. S., Eschbacher, J. M., Heiserman, J. E., Dueck, A. C., Shapiro, W. R., Liu, S., Karis, J. P., Smith, K. A., Coons, S. W., Nakaji, P., Spetzler, R. F., Feuerstein, B. G., Debbins, J. and Baxter, L. C. (2012), **Reevaluating the imaging definition of tumor progression: Perfusion MRI quantifies recurrent glioblastoma tumor fraction, pseudoprogression, and radiation necrosis to predict survival**, Neuro-Oncology *14*(7), 919–930. URL: https://pubmed.ncbi.nlm.nih.gov/22561797/ https://pubmed.ncbi.nlm.nih.gov/22561797/?dopt=Abstract

Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J. and Maier-Hein, K. H. (2021), **nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation**, Nature Methods *18*(2), 203–211. URL: https://www.nature.com/articles/s41592-020-01008-z

Isensee, F., Jäger, P., Wasserthal, J., Zimmerer, D., Petersen, J., Kohl, S., Schock, J., Klein, A., Roß, T., Wirkert, S., Neher, P., Dinkelacker, S., Köhler, G. and Maier-Hein, K. (2020), **batchgenerators - a python framework for data augmentation**, Zenodo . URL: https://zenodo.org/record/3632567

Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H. P., Heiland, S., Wick, W., Bendszus, M., Maier-Hein, K. H. and Kickingereder, P. (2019), **Automated brain extraction of multisequence MRI using artificial neural networks**, Human Brain Mapping *40*(17), 4952–4964. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/hbm.24750 https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.24750 https://onlinelibrary.wiley.com/doi/10.1002/hbm.24750

Jakola, A. S., Zhang, Y. H., Skjulsvik, A. J., Solheim, O., Bø, H. K., Berntsen, E. M., Reinertsen, I., Gulati, S., Förander, P. and Brismar, T. B. (2018), **Quantitative texture analysis in the prediction of IDH status in low-grade gliomas**, Clinical Neurology and Neurosurgery *164*, 114–120. URL: https://pubmed.ncbi.nlm.nih.gov/29220731/ https://pubmed.ncbi.nlm.nih.gov/29220731/?dopt=Abstract

Jenkinson, M. and Smith, S. (2001), **A global optimisation method for robust affine registration of brain images**, Medical Image Analysis *5*(2), 143–156. URL: http://www.ncbi.nlm.nih.gov/pubmed/11516708

Johnson, C. D., Chen, M.-H., Toledano, A. Y., Heiken, J. P., Dachman, A., Kuo, M. D., Menias, C. O., Siewert, B., Cheema, J. I., Obregon, R. G., Fidler, J. L., Zimmerman, P., Horton, K. M., Coakley, K., Iyer, R. B., Hara, A. K., Halvorsen, R. A., Casola, G., Yee, J., Herman, B. A., Burgart, L. J. and Limburg, P. J. (2008), **Accuracy of CT Colonography for Detection of Large Adenomas and Cancers**, New England Journal of Medicine *359*(12), 1207–1217. URL: https://www.nejm.org/doi/full/10.1056/NEJMoa0800996

Kaufmann, T. J., Smits, M., Boxerman, J., Huang, R., Barboriak, D. P., Weller, M., Chung, C., Tsien, C., Brown, P. D., Shankar, L., Galanis, E., Gerstner, E., Van Den Bent, M. J., Burns, T. C., Parney, I. F., Dunn, G., Brastianos, P. K., Lin, N. U., Wen, P. Y. and Ellingson, B. M. (2020), **Consensus recommendations for a standardized brain tumor imaging protocol for clinical trials in brain metastases**, Neuro-Oncology *22*(6), 757–772.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., Shi, A., Zhang, R., Zheng, L., Hou, R., Shi, W., Fu, X., Duan, Y., Huu, V. A., Wen, C., Zhang, E. D., Zhang, C. L., Li, O., Wang, X., Singer, M. A., Sun, X., Xu, J., Tafreshi, A., Lewis, M. A., Xia, H. and Zhang, K. (2018), **Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning**, Cell *172*(5), 1122–1131.e9.

Kickingereder, P., Isensee, F., Tursunova, I., Petersen, J., Neuberger, U., Bonekamp, D., Brugnara, G., Schell, M., Kessler, T., Foltyn, M., Harting, I., Sahm, F., Prager, M., Nowosielski, M., Wick, A., Nolden, M.,

Radbruch, A., Debus, J., Schlemmer, H. P., Heiland, S., Platten, M., von Deimling, A., van den Bent, M. J., Gorlia, T., Wick, W., Bendszus, M. and Maier-Hein, K. H. (2019), **Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study**, The Lancet Oncology *20*(5), 728–740.

Kim, Y. S., Kim, J. W., Yoon, W. S., Kang, M. K., Lee, I. J., Kim, T. H., Kim, J. H., Lee, H. S., Park, H. C., Jang, H. S., Kay, C. S., Yoon, S. M., Kim, M. S. and Seong, J. (2016), **Interobserver variability in gross tumor volume delineation for hepatocellular carcinoma: Results of Korean Radiation Oncology Group 1207 study**, Strahlentherapie und Onkologie *192*(10), 714–721. URL: https://link.springer.com/article/10.1007/s00066-016-1028-2

Knoll, M., Balermpas, P., von der Grün, J., Tawk, B., Adeberg, S., Tinhofer, I., Budach, V., Linge, A., Krause, M., Stuschke, M., Grosu, A., Zips, D., Combs, S., Belka, C., Kriegsmann, M., Weichert, W., Baumann, M., Roedel, C., Debus, J. and Abdollahi, A. (2020), **Superior Prognostic Performance of an Immunohistochemistry Trained DNA-Methylation Based PD-L1 Score in Patients with HNSCC Treated with Radiochemotherapy: A Multicenter Study of the German Cancer Consortium Radiation Oncology Group (DKTK-ROG).**, International Journal of Radiation Oncology*Biology*Physics *108*(3), S161. URL: http://www.redjournal.org/article/S0360301620323440/fulltext http://www.redjournal.org/article/S0360301620323440/abstract https://www.redjournal.org/article/S0360-3016(20)32344-0/abstract

Knoll, M., Debus, J., Furkel, J., Warta, R., Bougatf, N., Rapp, C., Brors, B., Wick, W., Unterberg, A., Herold-Mende, C. and Abdollahi, A. (2019), **Glioblastoma evolution pattern under surgery and radio(chemo)therapy (RCHT) to identify novel methylome based glioma subtypes.**, Journal of Clinical Oncology *37*(15_suppl), 2012–2012.

Lakens, D. (2017), **Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses**, Social Psychological and Personality Science *8*(4), 355–362. URL: https://journals.sagepub.com/doi/full/10.1177/1948550617697177

Lambin, P., Leijenaar, R. T., Deist, T. M., Peerlings, J., De Jong, E. E., Van Timmeren, J., Sanduleanu, S., Larue, R. T., Even, A. J., Jochems, A., Van Wijk, Y., Woodruff, H., Van Soest, J., Lustberg, T., Roelofs, E., Van Elmpt, W., Dekker, A., Mottaghy, F. M., Wildberger, J. E. and Walsh, S. (2017), **Radiomics: The bridge between medical imaging and personalized medicine**, Nature Reviews Clinical Oncology *14*(12), 749–762. URL: http://dx.doi.org/10.1038/nrclinonc.2017.141

Lee, K. S., Lee, K., Yun, S., Moon, S., Park, Y., Han, J. H., Kim, C. Y., Lee, H. S. and Choe, G. (2018), **Prognostic relevance of programmed cell death ligand 1 expression in glioblastoma**, Journal of Neuro-Oncology *136*(3), 453–461. URL: https://pubmed.ncbi.nlm.nih.gov/29147863/

Leijenaar, R. T., Bogowicz, M., Jochems, A., Hoebers, F. J., Wesseling, F. W., Huang, S. H., Chan, B., Waldron, J. N., O'Sullivan, B., Rietveld, D., Leemans, C. R., Brakenhoff, R. H., Riesterer, O., Tanadini-Lang, S., Guckenberger, M., Ikenberg, K. and Lambin, P. (2018), **Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: A multicenter study**, British Journal of Radiology *91*(1086). URL: https://pubmed.ncbi.nlm.nih.gov/29451412/

Li, Y., Liang, L., Dai, W., Cai, G., Xu, Y., Li, X., Li, Q. and Cai, S. (2016), **Prognostic impact of programed cell death-1 (PD-1) and PD-ligand 1 (PD-L1) expression in cancer cells and tumor infiltrating lymphocytes in colorectal cancer**, Molecular Cancer *15*(1). URL: https://genome-cancer.ucsc.edu/

Liew, C. (2018), **The future of radiology augmented with Artificial Intelligence: A strategy for success**, European Journal of Radiology *102*, 152–156.

Liu, Y., Carlsson, R., Ambjørn, M., Hasan, M., Badn, W., Darabi, A., Siesjö, P. and Issazadeh-Navikas, S. (2013), **PD-L1 expression by neurons nearby tumors indicates better prognosis in glioblastoma patients**, Journal of Neuroscience *33*(35), 14231–14245. URL: /pmc/articles/PMC6618508/ /pmc/articles/PMC6618508/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6618508/

Liu, Z., Zhang, X. Y., Shi, Y. J., Wang, L., Zhu, H. T., Tang, Z., Wang, S., Li, X. T., Tian, J. and Sun, Y. S. (2017), **Radiomics analysis for evaluation of pathological complete response to neoadjuvant chemoradiotherapy in locally advanced rectal cancer**, Clinical Cancer Research *23*(23), 7253–7262. URL: https://pubmed.ncbi.nlm.nih.gov/28939744/

Lohmann, P., Bousabarah, K., Hoevels, M. and Treuer, H. (2020), **Radiomics in radiation oncology—basics, methods, and limitations**, Strahlentherapie und Onkologie *196*(10), 848–855. URL: /pmc/articles/PMC7498498/ /pmc/articles/PMC7498498/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7498498/

Lovinfosse, P., Polus, M., Van Daele, D., Martinive, P., Daenen, F., Hatt, M., Visvikis, D., Koopmansch, B., Lambert, F., Coimbra, C., Seidel, L., Albert, A., Delvenne, P. and Hustinx, R. (2018), **FDG PET/CT radiomics for predicting the outcome of locally advanced rectal cancer**, European Journal of Nuclear Medicine and Molecular Imaging *45*(3), 365–375. URL: https://pubmed.ncbi.nlm.nih.gov/29046927/

Mackin, D., Fave, X., Zhang, L., Fried, D., Yang, J., Brian Taylor, Rodriguez-Rivera, E., Dodge, C., Jones, A. K. and Court, L. (2015), **Measuring computed tomography scanner variability of radiomics features**, Investigative Radiology *50*(11), 757–765.

McBee, M. P., Awan, O. A., Colucci, A. T., Ghobadi, C. W., Kadom, N., Kansagra, A. P., Tridandapani, S. and Auffermann, W. F. (2018), **Deep Learning in Radiology**, Academic Radiology *25*(11), 1472–1480. URL: http://www.academicradiology.org/article/S1076633218301041/fulltext http://www.academicradiology.org/article/S1076633218301041/abstract https://www.academicradiology.org/article/S1076-6332(18)30104-1/abstract

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M. A., Arbel, T., Avants, B. B., Ayache, N., Buendia, P., Collins, D. L., Cordier, N., Corso, J. J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ç., Durst, C. R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A.,

Iftekharuddin, K. M., Jena, R., John, N. M., Konukoglu, E., Lashkari, D., Mariz, J. A., Meier, R., Pereira, S., Precup, D., Price, S. J., Raviv, T. R., Reza, S. M., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H. C., Shotton, J., Silva, C. A., Sousa, N., Subbanna, N. K., Szekely, G., Taylor, T. J., Thomas, O. M., Tustison, N. J., Unal, G., Vasseur, F., Wintermark, M., Ye, D. H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M. and Van Leemput, K. (2015), **The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)**, IEEE Transactions on Medical Imaging *34*(10), 1993–2024.

Minniti, G., Niyazi, M., Andratschke, N., Guckenberger, M., Palmer, J. D., Shih, H. A., Lo, S. S., Soltys, S., Russo, I., Brown, P. D. and Belka, C. (2021), **Current status and recent advances in resection cavity irradiation of brain metastases**, Radiation Oncology *16*(1), 1–14. URL: https://ro-journal.biomedcentral.com/articles/10.1186/s13014-021-01802-9

Moore, S. M., Maffitt, D. R., Smith, K. E., Kirby, J. S., Clark, K. W., Freymann, J. B., Vendt, B. A., Tarbox, L. R. and Prior, F. W. (2015), **De-identification of medical images with retention of scientific research value**, Radiographics *35*(3), 727–735. URL: https://pubs.rsna.org/doi/abs/10.1148/rg.2015140244

Niyazi, M., Brada, M., Chalmers, A. J., Combs, S. E., Erridge, S. C., Fiorentino, A., Grosu, A. L., Lagerwaard, F. J., Minniti, G., Mirimanoff, R. O., Ricardi, U., Short, S. C., Weber, D. C. and Belka, C. (2016), **ESTRO-ACROP guideline "target delineation of glioblastomas"**, Radiotherapy and Oncology *118*(1), 35–42. URL: http://www.thegreenjournal.com/article/S0167814015006611/fulltext http://www.thegreenjournal.com/article/S0167814015006611/abstract https://www.thegreenjournal.com/article/S0167-8140(15)00661-1/abstract

Ou, D., Blanchard, P., Rosellini, S., Levy, A., Nguyen, F., Leijenaar, R. T., Garberis, I., Gorphe, P., Bidault, F., Ferté, C., Robert, C., Casiraghi, O., Scoazec, J. Y., Lambin, P., Temam, S., Deutsch, E. and Tao, Y. (2017), **Predictive and prognostic value of CT based radiomics signature in locally advanced head and neck cancers patients treated with concurrent chemoradiotherapy or bioradiotherapy and its added**

value to Human Papillomavirus status, Oral Oncology *71*, 150–155. URL: https://pubmed.ncbi.nlm.nih.gov/28688683/

Papanikolaou, N., Matos, C. and Koh, D. M. (2020), **How to develop a meaningful radiomic signature for clinical use in oncologic patients**, Cancer Imaging *20*(1), 1–10. URL: https://cancerimagingjournal.biomedcentral.com/articles/10.1186/s40644-020-00311-4

Parekh, V. S. and Jacobs, M. A. (2019), **Deep learning and radiomics in precision medicine**, Expert Review of Precision Medicine and Drug Development *4*(2), 59–72. URL: https://www.tandfonline.com/doi/abs/10.1080/23808993.2019.1585805

Parmar, C., Grossmann, P., Bussink, J., Lambin, P. and Aerts, H. J. (2015), **Machine Learning methods for Quantitative Radiomic Biomarkers**, Scientific Reports *5*(1), 1–11. URL: https://www.nature.com/articles/srep13087

Pratt, D., Dominah, G., Lobel, G., Obungu, A., Lynes, J., Sanchez, V., Adamstein, N., Wang, X., Edwards, N. A., Wu, T., Maric, D., Giles, A. J., Gilbert, M. R., Quezado, M. and Nduom, E. K. (2019), **Programmed Death Ligand 1 Is a Negative Prognostic Marker in Recurrent Isocitrate Dehydrogenase-Wildtype Glioblastoma**, Neurosurgery *85*(2), 280–288. URL: /pmc/articles/PMC7137460/ /pmc/articles/PMC7137460/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7137460/

Remedios, S., Roy, S., Pham, D. L. and Butman, J. A. (2018), **Classifying magnetic resonance image modalities with convolutional neural networks**, https://doi.org/10.1117/12.2293943 *10575*, 89. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10575/105752I/Classifying-magnetic-resonance-image-modalities-with-convolutional-neural-networks/10.1117/12.2293943.full https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10

Riegel, A. C., Berson, A. M., Destian, S., Ng, T., Tena, L. B., Mitnick, R. J. and Wong, P. S. (2006), **Variability of gross tumor volume delineation in head-and-neck cancer using CT and PET/CT fusion**,

International Journal of Radiation Oncology Biology Physics *65*(3), 726–732. URL: https://pubmed.ncbi.nlm.nih.gov/16626888/

Schalper, K. A., Rodriguez-Ruiz, M. E., Diez-Valle, R., López-Janeiro, A., Porciuncula, A., Idoate, M. A., Inogés, S., de Andrea, C., López-Diaz de Cerio, A., Tejada, S., Berraondo, P., Villarroel-Espindola, F., Choi, J., Gúrpide, A., Giraldez, M., Goicoechea, I., Gallego Perez-Larraya, J., Sanmamed, M. F., Perez-Gracia, J. L. and Melero, I. (2019), **Neoadjuvant nivolumab modifies the tumor immune microenvironment in resectable glioblastoma**, Nature Medicine *25*(3), 470–476. URL: https://pubmed.ncbi.nlm.nih.gov/30742120/

Schalper, K. A., Velcheti, V., Carvajal, D., Wimberly, H., Brown, J., Pusztai, L. and Rimm, D. L. (2014), **In situ tumor PD-L1 mRNA expression is associated with increased tils and better outcome in breast carcinomas**, Clinical Cancer Research *20*(10), 2773–2782. URL: https://clincancerres.aacrjournals.org/content/20/10/2773 https://clincancerres.aacrjournals.org/content/20/10/2773.abstract

Scheirer, W. J., De Rezende Rocha, A., Sapkota, A. and Boult, T. E. (2013), **Toward open set recognition**, IEEE Transactions on Pattern Analysis and Machine Intelligence *35*(7), 1757–1772.

Schoppe, O., Pan, C., Coronel, J., Mai, H., Rong, Z., Todorov, M. I., Müskes, A., Navarro, F., Li, H., Ertürk, A. and Menze, B. H. (2020), **Deep learning-enabled multi-organ segmentation in whole-body mouse scans**, Nature Communications *11*(1). URL: /pmc/articles/PMC7648799/ /pmc/articles/PMC7648799/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7648799/

Simonyan, K., Vedaldi, A. and Zisserman, A. (2014), **Deep inside convolutional networks: Visualising image classification models and saliency maps**, 2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings . URL: https://arxiv.org/abs/1312.6034v2

Simonyan, K. and Zisserman, A. (2015), **Very deep convolutional networks for large-scale image recognition**, 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings . URL: https://arxiv.org/abs/1409.1556v6

Skogen, K., Schulz, A., Dormagen, J. B., Ganeshan, B., Helseth, E. and Server, A. (2016), **Diagnostic performance of texture analysis on MRI in grading cerebral gliomas**, European Journal of Radiology *85*(4), 824–829. URL: https://pubmed.ncbi.nlm.nih.gov/26971430/ https://pubmed.ncbi.nlm.nih.gov/26971430/?dopt=Abstract

Sun, Q., Lin, X., Zhao, Y., Li, L., Yan, K., Liang, D., Sun, D. and Li, Z. C. (2020), **Deep Learning vs. Radiomics for Predicting Axillary Lymph Node Metastasis of Breast Cancer Using Ultrasound Images: Don't Forget the Peritumoral Region**, Frontiers in Oncology *10*, 53.

Thrall, J. H., Li, X., Li, Q., Cruz, C., Do, S., Dreyer, K. and Brink, J. (2018*a*), **Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success**, Journal of the American College of Radiology *15*(3), 504–508. URL: https://doi.org/10.1016/j.jacr.2017.12.026

Thrall, J. H., Li, X., Li, Q., Cruz, C., Do, S., Dreyer, K. and Brink, J. (2018*b*), **Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success**, Journal of the American College of Radiology *15*(3), 504–508.

Tian, Q., Yan, L. F., Zhang, X., Zhang, X., Hu, Y. C., Han, Y., Liu, Z. C., Nan, H. Y., Sun, Q., Sun, Y. Z., Yang, Y., Yu, Y., Zhang, J., Hu, B., Xiao, G., Chen, P., Tian, S., Xu, J., Wang, W. and Cui, G. B. (2018), **Radiomics strategy for glioma grading using texture features from multiparametric MRI**, Journal of Magnetic Resonance Imaging *48*(6), 1518–1528. URL: https://pubmed.ncbi.nlm.nih.gov/29573085/ https://pubmed.ncbi.nlm.nih.gov/29573085/?dopt=Abstract

Tibshirani, R. (1996), **Regression Shrinkage and Selection Via the Lasso**, Journal of the Royal Statistical Society: Series B (Methodological) *58*(1), 267–288. URL: https://www.jstor.org/stable/2346178

Tiwari, P., Prasanna, P., Wolansky, L., Pinho, M., Cohen, M., Nayate, A. P., Gupta, A., Singh, G., Hatanpaa, K. J., Sloan, A., Rogers, L. and Madabhushi, A. (2016), **Computer-extracted texture features to distinguish cerebral radionecrosis from recurrent brain tumors on multiparametric mri:**

A feasibility study, American Journal of Neuroradiology *37*(12), 2231–2236. URL: https://pubmed.ncbi.nlm.nih.gov/27633806/ https://pubmed.ncbi.nlm.nih.gov/27633806/?dopt=Abstract

Towbin, A. J., Roth, C. J., Petersilge, C. A., Garriott, K., Buckwalter, K. A. and Clunie, D. A. (2021), **The Importance of Body Part Labeling to Enable Enterprise Imaging: A HIMSS-SIIM Enterprise Imaging Community Collaborative White Paper**, Journal of Digital Imaging *34*(1), 1–15. URL: https://link.springer.com/article/10.1007/s10278-020-00415-0

Tumeh, P. C., Harview, C. L., Yearley, J. H., Shintaku, I. P., Taylor, E. J., Robert, L., Chmielowski, B., Spasic, M., Henry, G., Ciobanu, V., West, A. N., Carmona, M., Kivork, C., Seja, E., Cherry, G., Gutierrez, A. J., Grogan, T. R., Mateus, C., Tomasic, G., Glaspy, J. A., Emerson, R. O., Robins, H., Pierce, R. H., Elashoff, D. A., Robert, C. and Ribas, A. (2014), **PD-1 blockade induces responses by inhibiting adaptive immune resistance**, Nature *515*(7528), 568–571. URL: https://pubmed.ncbi.nlm.nih.gov/25428505/

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A. and Gee, J. C. (2010), **N4ITK: Improved N3 bias correction**, IEEE Transactions on Medical Imaging *29*(6), 1310–1320. URL: https://pubmed.ncbi.nlm.nih.gov/20378467/

Udall, M., Rizzo, M., Kenny, J., Doherty, J., Dahm, S. A., Robbins, P. and Faulkner, E. (2018), **PD-L1 diagnostic tests: A systematic literature review of scoring algorithms and test-validation metrics**, Diagnostic Pathology *13*(1). URL: /pmc/articles/PMC5807740/ /pmc/articles/PMC5807740/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5807740/

Valdés Hernández, M. d. C., González-Castro, V., Ghandour, D. T., Wang, X., Doubal, F., Muñoz Maniega, S., Armitage, P. A. and Wardlaw, J. M. (2016), **On the computational assessment of white matter hyperintensity progression: difficulties in method selection and bias field correction performance on images with significant white matter pathology**, Neuroradiology *58*(5), 475–485. URL: /pmc/articles/PMC4846712/ /pmc/articles/PMC4846712/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4846712/

Van Der Maaten, L. and Hinton, G. (2008), **Visualizing data using t-SNE**, Journal of Machine Learning Research *9*, 2579–2625.

van der Voort, S. R., Smits, M. and Klein, S. (2021), **DeepDicomSort: An Automatic Sorting Algorithm for Brain Magnetic Resonance Imaging Data**, Neuroinformatics *19*(1), 159–184.

Van Griethuysen, J. J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G., Fillion-Robin, J. C., Pieper, S. and Aerts, H. J. (2017), **Computational radiomics system to decode the radiographic phenotype**, Cancer Research *77*(21), e104–e107. URL: https://pubmed.ncbi.nlm.nih.gov/29092951/

van Ooijen, P. M. (2019), **Quality and curation of medical images and data**, Springer, Cham. URL: https://link.springer.com/chapter/10.1007/978-3-319-94878-2_17

Van Rossum, P. S., Fried, D. V., Zhang, L., Hofstetter, W. L., Van Vulpen, M., Meijer, G. J., Court, L. E. and Lin, S. H. (2016), **The incremental value of subjective and quantitative assessment of 18F-FDG PET for the prediction of pathologic complete response to preoperative chemoradiotherapy in esophageal cancer**, Journal of Nuclear Medicine *57*(5), 691–700. URL: https://pubmed.ncbi.nlm.nih.gov/26795288/

van Timmeren, J. E., Cester, D., Tanadini-Lang, S., Alkadhi, H. and Baessler, B. (2020), **Radiomics in medical imaging—"how-to" guide and critical reflection**, Insights into Imaging *11*(1), 1–16. URL: https://insightsimaging.springeropen.com/articles/10.1186/s13244-020-00887-2

Velcheti, V., Schalper, K. A., Carvajal, D. E., Anagnostou, V. K., Syrigos, K. N., Sznol, M., Herbst, R. S., Gettinger, S. N., Chen, L. and Rimm, D. L. (2014), **Programmed death ligand-1 expression in non-small cell lung cancer**, Laboratory Investigation *94*(1), 107–116. URL: https://pubmed.ncbi.nlm.nih.gov/24217091/

Walk, E. E., Yohe, S. L., Beckman, A., Schade, A., Zutter, M. M., Pfeifer, J. and Berry, A. B. (2020), **The cancer immunotherapy biomarker**

testing landscape, Archives of Pathology and Laboratory Medicine *144*(6), 706–724.

Wang, S., Pavlicek, W., Roberts, C. C., Langer, S. G., Zhang, M., Hu, M., Morin, R. L., Schueler, B. A., Wellnitz, C. V. and Wu, T. (2011), **An automated DICOM database capable of arbitrary data mining (Including Radiation Dose Indicators) for quality monitoring**, Journal of Digital Imaging *24*(2), 223–233. URL: https://link.springer.com/article/10.1007/s10278-010-9329-y

Weltens, C., Menten, J., Feron, M., Bellon, E., Demaerel, P., Maes, F., Van Den Bogaert, W. and Van Der Schueren, E. (2001), **Interobserver variations in gross tumor volume delineation of brain tumors on computed tomography and impact of magnetic resonance imaging**, Radiotherapy and Oncology *60*(1), 49–59.

Wen, Q., Yang, Z., Dai, H., Feng, A. and Li, Q. (2021), **Radiomics Study for Predicting the Expression of PD-L1 and Tumor Mutation Burden in Non-Small Cell Lung Cancer Based on CT Images and Clinicopathological Features**, Frontiers in Oncology *11*, 3082.

Weng, W. and Zhu, X. (2021), **INet: Convolutional Networks for Biomedical Image Segmentation**, IEEE Access *9*, 16591–16603. URL: http://arxiv.org/abs/1505.04597

Wolf, I., Vetter, M., Wegner, I., Nolden, M., Bottger, T., Hastenteufel, M., Schobinger, M., Kunert, T. and Meinzer, H.-P. (2004), **The medical imaging interaction toolkit (MITK): a toolkit facilitating the creation of interactive software by extending VTK and ITK**, Medical Imaging 2004: Visualization, Image-Guided Procedures, and Display *5367*, 16. URL: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5367/0000/The-medical-imaging-interaction-toolkit-MITK–a-toolkit-facilitating/10.1117/12.535112.full https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5367/0000/The-medical

Wu, X., Gu, Z., Chen, Y., Chen, B., Chen, W., Weng, L. and Liu, X. (2019), **Application of PD-1 Blockade in Cancer Immunotherapy**, Computational and Structural Biotechnology Journal *17*, 661–674. URL: /pm-

c/articles/PMC6558092/ /pmc/articles/PMC6558092/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6558092/

Yan, D., Zhang, Z., Luo, Q. and Yang, X. (2017), **A novel mouse segmentation method based on dynamic contrast enhanced micro-ct images**, PLoS ONE *12*(1).

Yang, C. Y., Lin, M. W., Chang, Y. L., Wu, C. T. and Yang, P. C. (2016), **Programmed cell death-ligand 1 expression is associated with a favourable immune microenvironment and better overall survival in stage i pulmonary squamous cell carcinoma**, European Journal of Cancer *57*, 91–103. URL: https://pubmed.ncbi.nlm.nih.gov/26901614/

Zhou, C., Jones, B., Moustafa, M., Schwager, C., Bauer, J., Yang, B., Cao, L., Jia, M., Mairani, A., Chen, M., Chen, L., Debus, J. and Abdollahi, A. (2017), **Quantitative assessment of radiation dose and fractionation effects on normal tissue by utilizing a novel lung fibrosis index model**, Radiation Oncology *12*(1).

Zhou, T., Ruan, S. and Canu, S. (2019), **A review: Deep learning for medical image segmentation using multi-modality fusion**, Array *3-4*, 100004.

Zwanenburg, A., Vallières, M., Abdalah, M. A., Aerts, H. J., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R. J., Boellaard, R., Bogowicz, M., Boldrini, L., Buvat, I., Cook, G. J., Davatzikos, C., Depeursinge, A., Desseroit, M. C., Dinapoli, N., Dinh, C. V., Echegaray, S., El Naqa, I., Fedorov, A. Y., Gatta, R., Gillies, R. J., Goh, V., Götz, M., Guckenberger, M., Ha, S. M., Hatt, M., Isensee, F., Lambin, P., Leger, S., Leijenaar, R. T., Lenkowicz, J., Lippert, F., Losnegård, A., Maier-Hein, K. H., Morin, O., Müller, H., Napel, S., Nioche, C., Orlhac, F., Pati, S., Pfaehler, E. A., Rahmim, A., Rao, A. U., Scherer, J., Siddique, M. M., Sijtsema, N. M., Socarras Fernandez, J., Spezi, E., Steenbakkers, R. J., Tanadini-Lang, S., Thorwarth, D., Troost, E. G., Upadhaya, T., Valentini, V., van Dijk, L. V., van Griethuysen, J., van Velden, F. H., Whybra, P., Richter, C. and Löck, S. (2020), **The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping**, Radiology *295*(2), 328–338. URL: https://arxiv.org/abs/1612.07003v11

Chapter 9

Personal Publications

This work was funded by the European Union's Horizon 2020 research and innovation program (MSCA-ITN-PREDICT) and by the German Cancer Research Center (DKFZ). Patient data originated from patients treated at the Heidelberg Ion-Beam Therapy Center (HIT) and the University Hospital Heidelberg and was used with consent.
The proposed methodologies, both in terms of the development of the radiomics analysis infrastructures and the tumor immune status modelling, presented in chapter III, were developed by myself.
The generation of results presented in chapter IV was fully carried out on my own and represents the core results of this dissertation.

Parts of the results presented in this thesis were submitted in advance in peer reviewed journals:

1. Sforazzini F., Salome P., Moustafa M., Zhou C., Schwager C., Rein K., Bougatf N., Kudak A., Woodruff H., Dubois L., Lambin P., Debus J., Abdollahi A. and Knoll M. (2021) **Deep learning-based automatic lung segmentation for healthy and fibrotic lungs in mice from multi-resolution CT**. Under revision at: Radiology Artificial Intelligence.

Publication 1 is based on the results in chapter 4.2.2 of this dissertation, as well as part of the discussion section. My personal contribution to this publication consisted in the development and testing of the segmentation method and the drafting of the manuscript.

Conferences:

1. Sforrazini, F., Salome, P., Kudak, A., Ulrich, M., Bougatf, N., Debus, J., Knoll, M. and Abdollahi, A. (2020) **pyCuRT: An Automated Data Curation Workflow for Radiotherapy Big Data Analysis using Pythons' NyPipe.** International Journal of Radiation Oncology, Biology, Physics 108(3) (Supplement E772).

2. Sforazzini F., Salome P., Kudak A., Ulrich M., Bougatf N., Herold-Mende C., Warta R., Debus J., Knoll M. and Abdollahi A. (2021) **PD-L1-R: a MR based surrogate for PD-L1 expression in Glioblastoma multiforme.** Journal of Clinical Oncology 39:15_suppl, 2041-2041

3. Sforazzini F., Salome P., Warta R., Kudak A., Ulrich M., Bougatf N., , Vollmuth P., Wick W., König L., Herold-Mende C., Debus J., Knoll M. and Abdollahi A. (2021) **Tumor immune evolution determined by dynamic PD-L1 Radiomics signature (PD-L1-R) is prognostic in Glioblastoma patients treated with postsurgical radiochemotherapy** accepted for presentation at Radiological Society of North America (RSNA).

Curriculum Vitae

PERSONAL INFORMATION

Name:	Francesco Sforazzini
Date of Birth	01/06/1986
Place of Birth	Siena, Italy

RESEARCH EXPERIENCE AND EDUCATION

October 2018 Present Day	PhD student, **DKFZ and University of Heidelberg** *Heidelberg, Germany* - Development of radiomics analysis infrastructures and their application to radiation oncology.
March 2015 July 2018	Research Assistant, **Monash University** *Melbourne, Australia* - Analysis of functional MRI image and development of a new PET-MR head motion correction method.
May 2012 September 2014	Research fellow, **Italian Institute of Technology** *Rovereto, Italy* - Development of methods to study the brain functional connectivity in mouse models.
August 2011 May 2012	Master thesis internship, **IRCCS Stella Maris** *Pisa, Italy* - Implementation of methods for mapping the B1 field inhomogeneity at high magnetic field.
October 2009 May 2012	Master degree, **University of Physics** *Pisa, Italy* - Medical Physics focusing medical image processing.
October 2005 July 2009	Bachelor degree, **University of Physics** *Siena, Italy* - Physics and advanced technologies Emphasis on electronics and software development.

Eidesstattliche Versicherung

1. Bei der eingereichten Dissertation zu dem Thema:

   Development of Machine Learning-based Spatially and Temporally Resolved 4D Radiomics in Radiation Oncology

   handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort, Datum Unterschrift Doktorand