Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University

Heidelberg

Presented by

Jonas Richard Bohn, M.Sc.

born in: Essen, Germany

Oral examination: February 10th, 2026

Towards Self-Configuring Radiomics for Robust and Reproducible Predictive Performance

Referees: Prof. Dr. Benedikt Brors

Prof. Dr. Klaus Maier-Hein

Abstract

Introduction:
Radiomics aims to extract quantitative features from medical images that capture underlying biological and clinical characteristics. Despite its promise for precision oncology, radiomics research continues to suffer from poor reproducibility and limited generalization across studies, software, and imaging modalities. This thesis addresses these fundamental limitations by systematically analyzing how methodological design choices—such as feature extraction, preprocessing, and model selection—affect the robustness and transferability of radiomic biomarkers. To enable this large-scale methodological investigation, I developed the Radiomics Processing Toolkit (RPTK), a fully automated and open-source framework that standardizes radiomics experimentation and benchmarking across heterogeneous datasets. Using RPTK, I conducted comprehensive evaluations on seven open-source cancer imaging cohorts and demonstrated the framework's applicability in two clinical studies on lung cancer immunotherapy response prediction and colorectal neoplasia detection.

Materials and Methods:
My work integrates radiomics analyses performed on retrospective data, including seven public datasets and two proprietary cohorts, comprising 3,189 Computer Tomography (CT) and Magnetic Resonance (MR) scans from 3,116 patients with a total of 3,273 segmented regions of interest (ROI). The seven open-source datasets include retrospective MR and CT cancer imaging data concerning different tasks for cancer classification from 931 patients. The proprietary data collection include a multi-timepoint (prior treatment and during treatment) CT dataset for lung cancer immunotherapy treatment response prediction (the Predict study) consisting of 73 patients and a large-scale CT liver imaging datasets with 1,997 patients investigating in colorectal neoplasia detection (LiverCRC study). The mean patient age was 62 $\pm$ 17 years, and 53.6 % of participants were male. Radiomics features were extracted using two independent feature-extraction tools, PyRadiomics and Medical Image Radiomics Processor (MIRP), enabling standardized cross-extractor comparisons in compliance with the Image Biomarker Standardisation Initiative (IBSI). The

RPTK framework integrates adaptive preprocessing, standardized feature extraction, and robust feature stability filtering to enhance reproducibility and robustness for subsequent model training. Six machine learning models were trained to predict tumor malignancy, treatment response, colorectal neoplasia, or cancer subtypes based on the selected feature sets from each extractor. The performance of RPTK was tested against a state-of-the-art radiomics tool (AutoRadiomics) and six different deep learning models.

Results:

Across seven open-source datasets, RPTK outperformed both AutoRadiomics and deep learning models (Residual Networks (ResNet) and Densely connected convolutional Networks (DenseNet)), achieving a mean test Area Under the Receiver Operating Characteristic curve (AUROC) of 0.81 ± 0.12 compared to 0.68 ± 0.15 and 0.60 ± 0.16, respectively. In the Predict study, longitudinal delta-radiomics analysis with RPTK improved early prediction of immunotherapy response compared to single-timepoint analyses with RPTK, and the inclusion of clinical variables further enhanced model performance in RPTK. RPTK achieved a test AUROC of 0.75 ± 0.10 using delta radiomics, outperforming AutoRadiomics (0.51 ± 0.14) and the best deep learning model (0.56 ± 0.14). In the LiverCRC study, RPTK reached a test AUROC of 0.86 ± 0.04, significantly exceeding AutoRadiomics (0.65 ± 0.03) and deep learning (0.60 ± 0.03), demonstrating scalability and generalization in large multi-thousand-sample datasets. Beyond these comparisons, RPTK also matched or outperformed 12 additional published test AUROC values reported on the integrated open-source datasets.

Conclusion:

Collectively, the results demonstrate that RPTK provides robust, state-of-the-art predictive performance across diverse imaging datasets and clinically relevant tasks. Its modular design enables fair cross-framework benchmarking while maintaining flexibility for clinical data integration and ensuring methodological transparency. The open-source release of RPTK fosters community-driven validation and supports future clinical implementation. This work thus represents both a methodological advancement and a step toward reliable, reproducible, and clinically translatable radiomics.

Zusammenfassung

Einleitung:
Radiomics zielt darauf ab, quantitative Merkmale aus medizinischen Bilddaten zu extrahieren, die zugrunde liegende biologische und klinische Charakteristika abbilden. Trotz des großen Potenzials für die Präzisionsonkologie leidet die Radiomics-Forschung weiterhin unter einer unzureichenden Reproduzierbarkeit und begrenzten Generalisierbarkeit über verschiedenste Studien, Softwarelösungen und Bildgebungsmodalitäten hinweg. Diese Arbeit adressiert diese grundlegenden Einschränkungen durch eine systematische Analyse, wie methodische Designentscheidungen, etwa in der Merkmalsextraktion, der Vorverarbeitung und der Modellauswahl, die Robustheit und Übertragbarkeit radiomischer Biomarker beeinflussen. Zur Ermöglichung dieser groß angelegten methodischen Untersuchung wurde das Radiomics Processing Toolkit (RPTK) entwickelt, ein vollständig automatisiertes und quelloffenes Framework, das Radiomics-Experimente und Benchmarking über heterogene Datensätze hinweg standardisiert. Mit RPTK wurden umfassende Evaluierungen auf sieben öffentlich verfügbaren Krebsbildgebungsdatensätzen durchgeführt und die Anwendbarkeit in zwei klinischen Studien zur Vorhersage des Ansprechens auf eine Immuntherapie bei Lungenkrebs sowie zur Detektion kolorektaler Neoplasien demonstriert.

Material und Methoden:
Die Arbeit integriert Radiomics-Analysen, die auf retrospektiven Daten basieren und sieben öffentliche Datensätze sowie zwei proprietäre Kohorten umfassen. Insgesamt wurden 3.189 CT- und MR-Aufnahmen von 3.116 Patientinnen und Patienten mit 3.273 segmentierten Regionen von Interesse (ROI) ausgewertet. Die sieben Open-Source-Datensätze enthalten retrospektive MR- und CT-Bilddaten zu unterschiedlichen Aufgaben der Krebs-Klassifikation von insgesamt 931 Patientinnen und Patienten. Die proprietären Datensätze umfassen eine longitudinale (vor und während der Behandlung erhobene) CT-Kohorte zur Vorhersage des Immuntherapieansprechens bei Lungenkrebs (Predict-Studie) mit 73 Patientinnen und Patienten sowie einen groß angelegten CT-Leberdatensatz mit 1.997 Fällen zur Untersuchung der kolorektalen Neoplasieerkennung (LiverCRC-Studie). Das Durchschnittsalter der Patientinnen und Pa-

tienten betrug 62 ± 17 Jahre, 53,6,% waren männlich. Radiomics-Merkmale wurden mit zwei unabhängigen Merkmalsextraktionswerkzeugen, PyRadiomics und MIRP, extrahiert, um standardisierte Vergleiche zwischen Extraktoren gemäß den Richtlinien der Image Biomarker Standardisation Initiative (IBSI) zu ermöglichen. Das RPTK-Framework integriert adaptive Vorverarbeitung, standardisierte Merkmalsextraktion und robuste Merkmalsstabilitätsfilterung, um die Reproduzierbarkeit und Robustheit für das anschließende Modelltraining zu verbessern. Sechs Machine-Learning-Modelle wurden trainiert, um auf Basis der extrahierten Merkmale Tumormalignität, Therapieansprechen, kolorektale Neoplasien oder Krebs-Subtypen vorherzusagen. Die Leistungsfähigkeit von RPTK wurde gegen ein aktuelles Radiomics-Tool (AutoRadiomics) sowie sechs verschiedene Deep-Learning-Modelle getestet.

Ergebnisse:

Über sieben Open-Source-Datensätze hinweg übertraf RPTK sowohl AutoRadiomics als auch die Deep-Learning-Modelle (ResNet und DenseNet) mit einer durchschnittlichen Test-AUROC von 0,81 ± 0,12 im Vergleich zu 0,68 ± 0,15 bzw. 0,60 ± 0,16. In der Predict-Studie verbesserte die longitudinale Delta-Radiomics-Analyse mit RPTK die frühe Vorhersage des Immuntherapieansprechens im Vergleich zu Einzelzeitpunktanalysen, und die Einbeziehung klinischer Variablen steigerte die Modellleistung weiter. Mit Delta-Radiomics erreichte RPTK eine Test-AUROC von 0,75 ± 0,10 und übertraf damit AutoRadiomics (0,51 ± 0,14) sowie das beste Deep-Learning-Modell (0,56 ± 0,14). In der LiverCRC-Studie erreichte RPTK eine Test-AUROC von 0,86 ± 0,04 und übertraf damit signifikant AutoRadiomics (0,65 ± 0,03) und Deep Learning (0,60 ± 0,03), was Skalierbarkeit und Generalisierbarkeit in großen Datensätzen mit mehreren tausend Fällen belegt. Darüber hinaus erreichte oder übertraf RPTK 12 weitere publizierte Test-AUROC-Werte, die für die integrierten Open-Source-Datensätze berichtet wurden.

Schlussfolgerung:

Insgesamt zeigen die Ergebnisse, dass RPTK eine robuste, moderne prädiktive Leistung über verschiedene Bildgebungsdatensätze und klinisch relevante Aufgaben hinweg bietet. Das modulare Design ermöglicht faire Cross-Framework-Benchmarks, gewährleistet methodische Transparenz und bietet gleichzeitig Flexibilität für die Integration klinischer Daten. Die frei-verfügbare Veröffentlichung von RPTK fördert eine gemeinschaftsgetriebene Validierung und unterstützt die zukünftige klinische Implementierung. Diese Arbeit stellt somit sowohl einen methodischen Fortschritt als auch einen Beitrag zu einer verlässlichen, reproduzierbaren und klinisch übertragbaren Radiomics dar.

Contents

Abstract **v**

Zusammenfassung **vii**

Contents **ix**

List of Abbreviations **xiii**

List of Figures **xix**

List of Tables **xxiii**

1 Introduction **1**

1.1 Motivation and Problem Statement 1

1.2 Research Objectives and Contributions 5

1.2.1 Objectives 5

1.2.2 Contribution 7

1.3 Outline 11

2 Background **13**

2.1 Imaging Background 13

2.1.1 Computed Tomography 14

2.1.2 Magnetic Resonance Imaging 18

2.1.3 Deep Learning Applications 20

2.1.4 Conclusion 24

2.2 Medical Background 25

2.2.1 Foundations of Cancer Biology 25

2.2.2 Advances in Cancer Treatment 27

2.2.3 Biomarker 29

2.2.4 Conclusion 30

2.3 Radiomics 31

2.3.1 Standardization in Radiomics . . . 32
2.3.2 The Radiomics Workflow . . . 35
2.3.3 Study Design & Image Acquisition . . . 37
2.3.4 Data Annotation & Image Segmentation . . . 38
2.3.5 Data Preprocessing . . . 39
2.3.6 Radiomics Feature Computation . . . 39
2.3.7 Machine Learning (Modeling) . . . 45
2.3.8 Clinical Integration & Reporting . . . 48
2.3.9 Radiomics Limitations & Future Directions . . . 50

3 State of the Art 52
3.1 Deep Learning . . . 54
3.2 Radiomics . . . 55
3.2.1 Customized Radiomics . . . 55
3.2.2 Workflow for Optimal Radiomics Classification (WORC) . . . 56
3.2.3 AutoRadiomics . . . 58
3.3 Conclusion and Outlook . . . 59

4 Data & Methods 62
4.1 Self-Configuring Radiomics Framework . . . 63
4.1.1 Data & Study Design . . . 64
4.1.2 The first RPTK Prototype . . . 68
4.1.3 Updates of the Proposed Approach . . . 70
4.1.4 Data Fingerprint . . . 70
4.1.5 Data Preprocessing . . . 71
4.1.6 Image and Segmentation Data Augmentation . . . 73
4.1.7 Radiomics Feature Computation and Reduction . . . 76
4.1.8 Model Training and Optimization . . . 78
4.1.9 Application of AutoRadiomics . . . 80
4.1.10 Application of Deep Learning Models . . . 80
4.1.11 Source Code Availability . . . 81
4.1.12 Used Computational Hardware . . . 82
4.1.13 Statistical Testing . . . 82
4.2 Predict Study - Predicting Immunotherapy Response . . . 83
4.2.1 Data & Study Design . . . 83
4.2.2 Image Segmentation . . . 85
4.2.3 RPTK Configuration in the Predict Study . . . 87
4.3 LiverCRC Study – Colorectal Neoplasia Prediction via Liver CT . . . 88

4.3.1 Data & Study Design 88
4.3.2 Image Segmentation 90
4.3.3 RPTK Configuration in the LiverCRC Study 90

5 Results 92
5.1 Self-Configuring Radiomics Pipeline 93
5.1.1 RPTK Handles a Variety of Different 3D Imaging Datasets . . . 93
5.1.2 RPTK Selects the Most Informative Radiomics Features 96
5.1.3 RPTK Selects the Best Performing Models 98
5.1.4 RPTK Outperforms Current State of the Art Methods 101
5.2 Predict Study - Predicting Immunotherapy Response 104
5.2.1 Longitudinal Imaging Improves RPTK Predictive Performance . 105
5.2.2 RPTK Selects Important Features for Treatment Response Prediction 107
5.2.3 RPTK Gains Performance with Additional Clinical Information 111
5.2.4 Clinical Potential and Decision Evaluation 113
5.2.5 Comparing RPTK Prediction Performance on Longitudinal Data 115
5.3 LiverCRC Study – Colorectal Neoplasia Prediction via Liver CT 116
5.3.1 The Selected Informative Features for Colorectal Noeplasia Prediction 117
5.3.2 RPTK Internal Model Performance Evaluation 121
5.3.3 RPTK Outperforms Other Tools Significantly on Larger Datasets 123

6 Discussion 126
6.1 Self-Configuring Radiomics Pipeline 126
6.1.1 The Datasets 126
6.1.2 Methodological Advances of the RPTK Framework 131
6.1.3 RPTK Model Prediction Performance 136
6.2 Predict Study - Predicting Immunotherapy Treatment Response 139
6.2.1 Data 139
6.2.2 Predict Study - Longitudinal Performance Impact 141
6.3 LiverCRC Study – Colorectal Neoplasia Prediction via Liver CT 148
6.3.1 Data 149
6.3.2 RPTK Performance on the LiverCRC Dataset 154
6.3.3 RPTK Performance Comparison on the LiverCRC Dataset . . . 155

7 Conclusion and Outlook 157
7.1 Summary of Key Findings 157

7.2 Methodological Contributions of RPTK . . . 158
7.3 Clinical and Translational Implications . . . 160
7.4 Limitations . . . 161
7.5 Outlook and Future Work . . . 162

8 Appendix **164**
8.1 Overview of all Datasets in this Thesis . . . 164
8.2 Self-Configuring Radiomics Pipeline . . . 166
8.2.1 Data Fingerprint . . . 166
8.2.2 RPTK Feature Extraction . . . 169
8.2.3 The RPTK Prototype . . . 171
8.3 Prediction Performance . . . 172
8.4 Predict Study - Predicting Immunotherapy Response . . . 178
8.5 LiverCRC Study – Colorectal Cancer Prediction via Liver CT . . . 185
8.6 Own Publications . . . 186
8.6.1 First Authorships . . . 187
8.6.2 Co-Authorships . . . 187

Bibliography **189**

Acknowledgements **227**

List of Abbreviations

ADASYN Adaptive Synthetic Sampling

ADC antibody-drug conjugates

AI Artificial Intelligence

ALK Anaplastic Lymphoma Kinase

ANOVA Analysis of Variance

ARISE Assessment for Radiomics Implementation Study Excellence

ASD Average Surface Distance

AUROC Area Under the Receiver Operating Characteristic curve

AutoML Automated Machine Learning

BiTEs Bispecific T-cell Engagers

BRAF B-Raf proto-oncogene serine/threonine kinase

CASH Combined Algorithm Selection and Hyperparameter

CD3 Cluster of Differentiation 3

CI Confidence Interval

CLAIM Checklist for Artificial Intelligence in Medical Imaging

CLEAR Checklist for Evaluation of Radiomics

CNN Convolutional Neural Network

CRC Colorectal Cancer

CRLM Colorectal Liver Metastasis

CRP C-Reactive Protein

CSF Cerebrospinal Fluid

CT Computer Tomography

CTC Circulating Tumour Cells

CTLA-4 Cytotoxic T Lymphocyte-associated Antigen 4

CV Cross Validation

DenseNet Densely connected convolutional Networks

DepEntropy Dependence Entropy

DICOM Digital Imaging and Communications in Medicine

DL Deep Learning

DNUNorm Dependence Non-Uniformity Normalized

DSC Dice Similarity Coefficient

DTF Desmoid Type Fibromatosis

ECOG Eastern Cooperative Oncology Group Performance Status

EGFR Epidermal Growth Factor Receptor

ET Echo Times

FCNs Fully Convolutional Networks

GANs Generative Adversarial Networks

GIST Gastro-Intestinal Stroma Tumor

GLCM Grey Level Co-occurrence Matrix

GLDZM Grey Level Distance Zone Matrix

GLNU Grey level non-uniformity

GLRLM Grey Level Run Length Matrix

GLSZM Grey Level Size Zone Matrix

HbA$_1c$ Glycated Hemoglobin

HDE High dependence emphasis

HU Hounsfield Units

IBSI Image Biomarker Standardization Initiative

ICC Intraclass Correlation Coefficient

ICI Immune Checkpoint Inhibitor

IDMN Inverse Difference Moment Normalized

IDRI Image Database Resource Initiative

IH Intensity histogram

IntegInt Integrated intensity

IQR Inter-quartile Range

iRECIST immune Response Evaluation Criteria In Solid Tumors

IS Intensity-based statistic

LASSO Least Absolute Shrinkage and Selection Operator

LBP2D Local Binary Pattern from 2D

LDA Linear Discriminant Analysis

LDLGE Low dependence low grey level emphasis

LGBM Light Gradient-Boosting Model

LGLRE Low Grey Level Run Emphasis

LGZE Low Gray Level Zone Emphasis

LIDC Lung Image Database Consortium

LoG Laplacian of Gaussian

LRHGE Long Run High Grey level Emphasis

MCC Maximum Correlation Coefficient

MICCAI Medical Image Computing and Computer Assisted Intervention

MIRP Medical Image Radiomics Processor

MITK Medical Imaging Interaction Toolkit

MONAI Medical Open Network for AI

MR Magnetic Resonance

MSD Medical Segmentation Decathlon

MTANN Massive-Training Artificial Neural Network

NGLDM Neighbouring Grey Level Dependence Matrix

NGTDM Neighbourhood Grey Tone Difference Matrix

NIFTI Neuroimaging Informatics Technology Initiative

NLR Neutrophil over Lymphocyte Ratio

nnU-Net no new U-Net

NSCLC Non-Small Cell Lung Cancer

OOI Origin of Information

PCA Principal Component Analysis

PD-1 Anti-Programmed cell Death protein-1

PD-L1 Anti-Programmed cell Death Ligand-1

PET Positron Emission Tomography

QDA Quadratic Discriminant Analysis

RECIST Response Evaluation Criteria In Solid Tumors

ResNet Residual Networks

RFE Recursive Feature Elimination

RMS Root Mean Square

ROC Receiver Operating Characteristic curve

ROI Region Of Interest

RPTK Radiomics Processing ToolKit

RQS Radiomics Quality Score

RT Repetition Time

SFS Sequential Feature Selection

SHAP Shapley Additive exPlanations

SMAC Sequential Model-based Algorithm Configuration

SMOTE Synthetic Minority Over-sampling Technique

SOTA State-Of-The-Art

SPP2177 Schwerpunktprogramm 2177

SRE Short runs emphasis

SVC Support Vector Classifier

SVM Support Vector Machine

TCIA The Cancer Imaging Platform

TMB Tumor Mutational Burden

TME Tumor Micro-Environment

TPE Tree-structured Parzen Estimator

TRIPOD Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis

VGG Visual Geometry Group

WDLPS Well Differentiated Lipo-Sarcoma

WORC Workflow for Optimal Radiomics Classification

WT Wild Type

XAI Explainable AI

XGBoost Extreme Gradient Boost

ZSEntr Zone Size Entropy

List of Figures

1.1 Radiomics Processing ToolKit (RPTK) workflow architecture and motivation 4

2.1 Hounsfield scale for characteristic values in CT images. 16
2.2 Overview of a General Radiomics Workflow 36
2.3 Motivation of Longitudinal Treatment Response Prediction 38
2.4 The Concept of Pixel discretizations and the Impact 42
2.5 The concept of using radiomics in a prospective usecase. 49

3.1 General domains of state of the art applications in medical image classification. 52

4.1 Demographics of the used open source datasets. 64
4.2 Overview comparison between the first RPTK prototype and the proposed RPTK. 69
4.3 The effect of connected component segmentation filtering. 73
4.4 The concept and overview of segmentation perturbation applied in RPTK. 74
4.5 The implementation of data augmentation for radiomics in RPTK. . . . 74
4.6 Experimental design of the Predict study. 84
4.7 Survival impact of treatment response in the Predict study. 85
4.8 Overview of the semi-automated segmentation generation for the Predict study. 86
4.9 Examples of automated (**a**) and corrected segmentations (**b**) in the Predict study. 86
4.10 Example images from responders and from non-responders in the Predict study. 87
4.11 STARD flow-chart of show the cohort generation of the LiverCRC study. 89
4.12 Example images of the liver from patients with colorectal neoplasia and without in the LiverCRC study. 90

5.1 Comparison of slice thickness distributions across open-source datasets. 94

5.2 Comparison of connected components distributions across open-source datasets. 95
5.3 Summary of selected features extracted with PyRadiomics across all open-source datasets. 97
5.4 Summary of selected features extracted with MIRP across all open-source datasets. 98
5.5 Summary plot of the validation performance (AUROC) of all models trained by RPTK across open-source datasets. 99
5.6 Validation AUROC comparison of RPTK, AutoRadiomics and deep learning models. 102
5.7 Test AUROC comparison of RPTK, AutoRadiomics, and deep learning models. 103
5.8 Comparison of RPTK test AUROC performance across the literature. . 104
5.9 Performance comparison of RPTK using single time-point verses multi time-point. 106
5.10 Heatmap of selected features from RPTK using PyRadiomics in the Predict study. 108
5.11 Heatmap of selected features from RPTK using MIRP in the Predict study. 109
5.12 SHAP values of best performing RPTK models based on the selected features in the Predict study. 110
5.13 SHAP values from the best RPTK models based on the clinical and combined models in the Predict study. 112
5.14 Survival impact of treatment response predictions in the Predict study. 114
5.15 AUROC performance comparison between RPTK, AutoRadiomics and deep learning models. 115
5.16 Heatmap of selected features from RPTK using PyRadiomics in the LiverCRC study. 118
5.17 Heatmap of selected features from RPTK using MIRP in the LiverCRC study. 119
5.18 SHAP values of best performing RPTK models based on the selected features in the LiverCRC study. 120
5.19 AUROC performance summary of all models applied in RPTK on the selected PyRadiomics features in the LiverCRC study. 121
5.20 AUROC performance summary of all models applied in RPTK on the selected MIRP features in the LiverCRC study. 123

5.21 Comparison of AUROC performance between RPTK, AutpRadiomics and deep learning models on the LiverCRC data. 124
5.22 ROC comparison between RPTK, AutoRadiomics and deep learning models. 125

8.1 Distribution of patient age at the time point of imaging over all datasets included in this thesis. 164
8.2 Distribution of imaging manufacturers of the scanners used to generate the 3D imaging data across all datasets. 165
8.3 Overview of imaging modalities as well as patient sex distribution across all used datasets in this thesis. 165
8.4 Boxplot for the size of the ROI measured by the number of voxels in the scans for all open-source datasets. 166
8.5 Boxplot for the number of bins calculated in the scans for all open-source datasets. 167
8.6 Boxplot for the number slices in the scans for all open-source datasets. 168
8.7 General IBSI feature profile for the PyRadiomics feature extraction. . . 169
8.8 General IBSI feature profile for the MIRP feature extraction. 170
8.9 Receiver operating characteristic (ROC) comparisons between AutoRadiomics and RPTK acrossthe open-source datasets. 172
8.10 Performance comparison between published results from AutoRadiomics and reproduced results. 174
8.11 Learning curve on selected PyRadiomics and MIRP Features on the Desmoid dataset. 174
8.12 Learning curve on selected PyRadiomics and MIRP Features on the CRLM dataset. 175
8.13 Learning curve on selected PyRadiomics and MIRP Features on the GIST dataset. 175
8.14 Learning curve on selected PyRadiomics and MIRP Features on the Liver dataset. 176
8.15 Learning curve on selected PyRadiomics and MIRP Features on the Lipo dataset. 176
8.16 Learning curve on selected PyRadiomics and MIRP Features on the Melanoma dataset. 177
8.17 Learning curve on selected PyRadiomics and MIRP Features on the LIDC-IDRI dataset. 177

8.18 Selected delta radiomics features heatmap (MIRP) in combination with clinical features in the Predict study. 180
8.19 Selected clinical features heatmap in the Predict study. 181
8.20 AutoRadiomics selected radiomics features heatmap in the Predict study. 182
8.21 Confusion matrix of the best clinical model and the best radiomics model in the Predict study. 183
8.22 Confusion matrix of the best combined (radiomics and clinical) model in the Predict study. 183
8.23 Learning curve on selected PyRadiomics and MIRP Features on the Predict dataset. 184
8.24 AutoRadiomics selected radiomics features heatmap in the LiverCRC study. 185
8.25 Learning curve on selected PyRadiomics and MIRP Features on the LiverCRC dataset. 186

List of Tables

2.1 The application of fixed bin size pixel discretizations in radiomics workflows. . . . 43
2.2 The application of fixed bin width pixel discretizations in radiomics workflows. . . . 43

4.1 Overview of datasets used in this thesis and their respective classification tasks. . . . 63
4.2 Overview of example images need to get classified across the open source datasets. . . . 65
4.3 Label distributions in training and test sets of the open-source datasets. 79

5.1 RPTK performance summary across all open-source datasets. 101
5.2 Imaging fingerprint of the Predict study. 105
5.3 Performance overview about longitudinal usage of RPTK. 111
5.4 Overview of the imaging fingerprint generated by RPTK on the LiverCRC study cohort. . . . 117
5.5 AUROC performance summary of all models applied in RPTK on the LiverCRC study cohort. . . . 122

8.1 Validation AUROC performance from the prototype RPTK application. 171
8.2 Deep learning performance summary across all open-source datasets. . 173
8.3 AutoRadiomics performance summary across all open-source datasets. . 173
8.4 Summary of clinical parameters used in the Predict study. 178
8.5 Performance comparison of radiomics, clinical, and delta features across time-points and modalities in the Predict study. . . . 179
8.6 AutoRadiomics performance summary for the Predict dataset. 179
8.7 Deep learning performance summary for the Predict dataset. 179
8.8 AutoRadiomics performance summary for the LiverCRC dataset. . . . 186
8.9 Deep learning performance summary for the LiverCRC dataset. 186

Chapter 1

Introduction

1.1 Motivation and Problem Statement

Radiomics is a quantitative image parameterization technique and has become an established approach in oncological radiology [1]. It enables the extraction of high-dimensional, disease-specific characteristics from standard radiological images, providing information that extends beyond human visual perception. These quantitative imaging biomarkers form the basis for machine learning applications that support a wide range of radiological tasks, including tumor classification, biomarker discovery, and treatment response prediction.

Despite its promise, the application of radiomics in diverse clinical settings remains challenging. Radiological imaging modalities such as CT and MR are based on different physical principles and therefore encode distinct information in their image contrast and gray-value distribution (see Section 2.1). Furthermore, acquisition parameters such as contrast agent use, MR sequence selection, or reconstruction kernel choice in CT can strongly influence image appearance and thus affect feature extraction (see Section 2.1.1 and 2.1.2). Without proper harmonization and standardization of image preprocessing and radiomics computation, the extracted features are not reproducible and are susceptible to such technical biases [2–4]. This sensitivity represents a fundamental scientific challenge in radiomics research, as it limits the comparability and generalizability of findings derived from quantitative image analysis. Therefore, robust radiomics pipelines must ensure feature stability and independence from modality- or protocol-specific variations.

The quality and amount of available imaging data often represent a bottleneck for radiomics research. Many studies rely on small, homogeneous datasets, which limits the generalizability of developed models [5–7]. Although data sharing initiatives are growing, issues related to data privacy, harmonization, and logistics still hinder the

creation of large, diverse multi-center cohorts [5, 6].

While deep learning approaches have demonstrated strong predictive power in medical image classification and segmentation [8], their black-box nature remains a major limitation for clinical translation. The lack of interpretability and explainability hampers clinical trust and adoption in oncological radiology [5, 9]. In addition, the application of deep learning models on small datasets often leads to non generalizable model predictions which are not applicable in clinical practice. In contrast, radiomics offers transparent and interpretable features and shows applicability on smaller datasets, but its inconsistent application across studies has led to poor reproducibility and limited generalization.

The absence of standardized processing led to the foundation of several initiatives, including the Image Biomarker Standardization Initiative (IBSI) and the German Research Foundation's Schwerpunktprogramm 2177 (SPP2177), which aim to define robust feature computation standards and highlight common methodological pitfalls [10, 11]. Furthermore, guideline frameworks such as Checklist for Evaluation of Radiomics (CLEAR), Assessment for Radiomics Implementation Study Excellence (ARISE), Checklist for Artificial Intelligence in Medical Imaging (CLAIM), and the Radiomics Quality Score (RQS) provide criteria to assess study design, reproducibility, and clinical applicability [12–14] (see Section 2.3.1). These developments have motivated the creation of software libraries such as PyRadiomics [15] and MIRP [16], which implement standardized feature definitions. However, these tools primarily focus on the extraction of statistical features derived from images and differ in extraction configurations, preprocessing options, and included different image transformation kernels (see Section 2.3.6). Importantly, they do not integrate feature selection, model training, or optimization, which are essential steps for building complete and reproducible radiomics workflows (see Section 2.3.2).

Radiomics outcomes are not only influenced by standardization but also by multiple experimental design decisions, including feature computation, post-processing, model selection, and optimization. Traditionally, these design steps have been manually tuned for specific datasets or clinical questions, a practice referred to as manually radiomics design [1, 17]. While such approaches can achieve high task-specific performance, their reliance on extensive manual parameterization and dataset-specific optimization increases the risk of overfitting and thus hampers reproducibility and generalizability across imaging modalities, tumor sites, and clinical endpoints.

As a response, automated radiomics frameworks such as AutoRadiomics and Workflow for Optimal Radiomics Classification (WORC) have been developed to provide generalizable workflows that reduce manual effort and bias [18, 19]. However, de-

spite improving reproducibility and usability, these frameworks often show limited predictive performance compared to expert-tuned pipelines and may not outperform alternative machine learning or deep learning–based approaches, which can offer better task-specific adaptability.

In summary, three major challenges characterize modern radiomics research:
First, the lack of reproducibility and generalization across imaging modalities, acquisition protocols, and institutions continues to limit clinical translation.
Second, radiomics workflows often depend on extensive manual design decisions and dataset-specific optimizations, which increase the risk of bias and overfitting.
Third, existing frameworks provide only partial standardization and rarely integrate all steps of the radiomics pipeline—from feature extraction to model optimization—into a unified and automated process.

This thesis aims to address these challenges by establishing a systematic and reproducible foundation for radiomics experimentation and benchmarking. Instead of designing each workflow manually for a specific clinical question, the goal is to create an automated framework that is self-configurable and optimizes model performance in a consistent and transparent way. To this end, the RPTK (Radiomics Processing Toolkit) was developed as a unified, reproducible, and performance-optimized framework for radiomics analysis. It integrates literature-based recommendations for optimal feature computation, preprocessing, feature selection, and model training into a standardized workflow. Furthermore, RPTK harmonizes radiomics extraction configurations for PyRadiomics and MIRP, ensuring comprehensive and comparable feature coverage [15,16], and introduces an ensemble optimization strategy that enhances predictive performance and robustness across diverse datasets.

The scientific hypothesis underlying this work is that a standardized, automated, and harmonized radiomics framework can achieve performance comparable to expert-tuned workflows while providing higher reproducibility, transparency, and scalability. Consequently, the Radiomics Processing ToolKit (RPTK) aims to bridge the gap between labor-intensive, non-generalizable customized radiomics workflows and automated but often suboptimal Automated Machine Learning (AutoML)-based solutions (see Figure 1.1). It provides an accessible, high-performance tool for reproducible binary classification tasks in oncological imaging, ultimately contributing to the clinical translation and reliability of radiomics applications.

The RPTK framework, illustrated in Figure 1.1 summarizes the conceptual differences between traditional customized radiomics pipelines, AutoML approaches, and the proposed RPTK framework. Each approach focuses on a Region Of Interest (ROI) segmented in a medical image, as depicted on the left side of the figure.

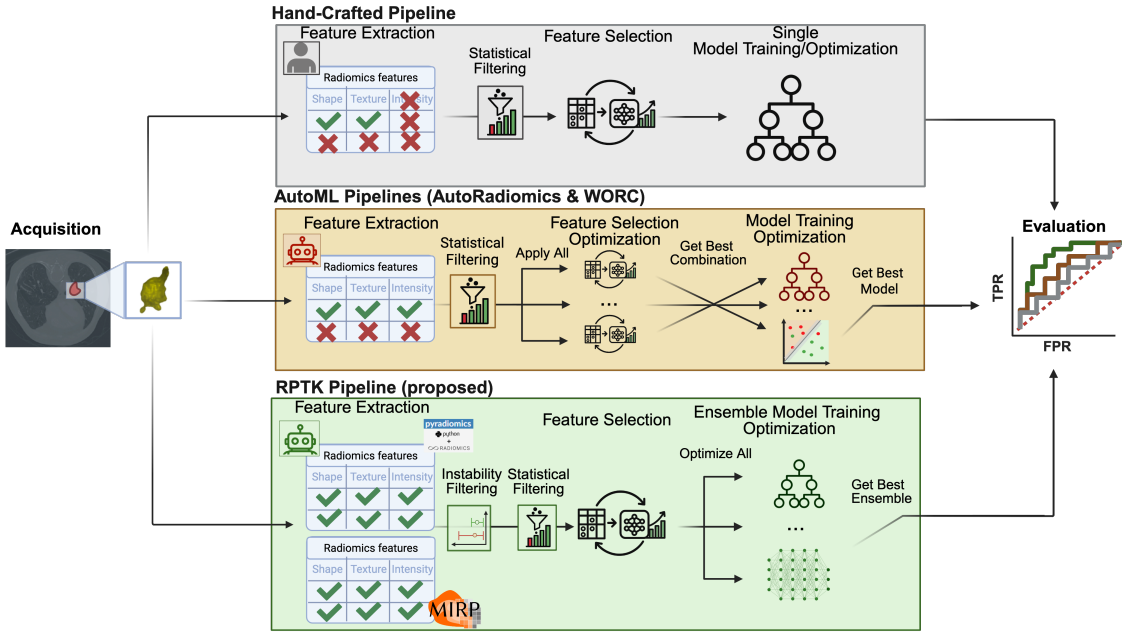


Figure 1.1. Overview of the RPTK workflow architecture, illustrating its contribution to unify traditional handcrafted (customized) radiomics and automated approaches, providing high-performance radiomics analysis with minimal manual intervention. Starting from the acquition of different imaging data, the different approaches rely on a sequence of feature extraction, filtering, feature selection and different training and optimization steps (TPR=True Positive Rate, FPR= False Positive Rate).

The upper part represents a typical customized radiomics workflow. These approaches often rely on a predefined subset of radiomics features selected based on prior knowledge or empirical experience, sometimes neglecting relevant ROI characteristics such as shape, texture, or intensity. Feature filtering and model training are manually tuned to a specific dataset, resulting in strong task-specific performance but poor generalizability. Such pipelines are rarely benchmarked across multiple feature extraction settings or modeling configurations, limiting their reproducibility and robustness.

The middle part illustrates AutoML-based frameworks, such as AutoRadiomics and WORC [18, 19]. These methods typically employ a standardized extractor, such as PyRadiomics, but with a reduced feature space due to limited coverage of Image Biomarker Standardization Initiative (IBSI)-defined feature classes and image transformations [15]. They apply initial statistical feature filtering followed by automated optimization of feature selection and model combination to identify the best-performing configuration for a given task. While these approaches improve reproducibility and enable systematic performance comparisons, their restricted feature

computation and optimization scope may limit their ultimate predictive performance.

The lower part of the figure introduces the proposed RPTK pipeline. This framework extends the radiomics workflow by including both tumor and peritumoral (margin) regions during feature extraction as well as integrating two complementary feature extractors, PyRadiomics and MIRP, to ensure comprehensive IBSI feature coverage and it synchronize the performed image transformations [15, 16]. Segmentation perturbations are introduced to evaluate the stability of extracted features, filtering out those that show high variability and thereby improving generalizability. Subsequent statistical filtering removes highly correlated and low-variance features, and a feature selection step reduces the feature space to a compact subset of maximally informative features (up to 20). These features are then used to train six machine learning models, each optimized in a five-fold cross-validation setting, with their predictions ensembled to obtain the final robust model.

Overall, the proposed RPTK pipeline bridges the gap between highly specialized customized workflows and generalizable but low-performing AutoML systems by combining comprehensive feature extraction, rigorous stability analysis, and ensemble model optimization into a unified and reproducible framework. Designed as an open-source toolkit, RPTK aims to make optimized state-of-the-art radiomics accessible to non-expert users while enhancing the performance and reproducibility of quantitative imaging studies in oncological radiology.

1.2 Research Objectives and Contributions

1.2.1 Objectives

The overarching objective of this thesis is to advance the scientific foundation of radiomics by improving the reproducibility, generalizability, and methodological transparency of quantitative image analysis. Radiomics has shown great potential in oncology research, yet its clinical translation is hindered by inconsistent workflows, non-reproducible feature definitions, and limited comparability across studies. This thesis addresses these issues by introducing RPTK as a unifying research framework that enables systematic, automated, and harmonized radiomics experimentation. The proposed framework aims to improve radiomics applications across diverse clinical tasks by fulfilling the following specific requirements:

- **Handling real-world data heterogeneity**: Processing heterogeneous imaging data originating from different scanners, imaging modalities, and acquisition

protocols. Robustly handle imaging and segmentation artifacts through standardized preprocessing and quality control procedures.

- **Extraction of stable and reproducible features**: Integrate IBSI-confirmed feature extraction methods, to identify and integrate important but missing radiomics features, and filter for feature robustness to specifically tackle inter-rater heterogeneity.

- **Performance and generalization**: Achieve competitive or superior predictive performance across multiple datasets, imaging modalities, and experimental designs without labor intensive work. This includes systematic comparison with state-of-the-art radiomics and deep learning approaches.

- **Integration of clinical covariates**: Facilitate seamless incorporation of additional clinical, demographic, or molecular variables into the radiomics workflow in order to gain performance from additional information. Include support for both purely clinical analyses and hybrid models by embedding clinical features within the radiomics feature space, thereby enabling joint modeling and comparative evaluation of imaging and non-imaging biomarkers.

- **Support for longitudinal analysis**: Enable application of radiomics in longitudinal study designs by providing dedicated functionality for the computation of delta-radiomics features, i.e., temporal changes of imaging biomarkers across multiple time points. This allows modeling of treatment response dynamics and performance enhancement through the integration of temporal information.

- **Accessibility for non-experts**: Design the framework as an end-to-end solution that can be operated without extensive knowledge of machine learning, programming, or radiomics. RPTK should provide default optimized configurations while maintaining flexibility for advanced customization.

- **Applicability in the hospital**: Ensure that RPTK can be executed efficiently on standard CPU-based clinical systems without requiring dedicated GPUs or internet connectivity, enabling potential use in routine clinical environments and data-secure infrastructures as an end-to-end framework.

Together, these objectives aimed to establish a radiomics processing framework that advances methodological robustness, performance, and clinical interpretability, while supporting multimodal and longitudinal data integration to democratize access to optimized quantitative imaging analysis within the radiomics community.

1.2.2 Contribution

This thesis advances the field of radiomics by strengthening its methodological robustness and clinical applicability across diverse oncological contexts, thereby promoting higher predictive performance and improved generalization of radiomics models. The contributions are organized according to the main studies presented in this work, which together establish new insights into optimal radiomics computation, integration, and performance on real-world medical imaging data as well as my contribution in these sections.

1. Advances in Self-Configuring Radiomics Pipelines

The first major contribution of this thesis is the development and systematic evaluation of an advanced self-configuring radiomics framework that extends existing automated approaches such as AutoRadiomics and WORC [18, 19]. In contrast to these frameworks, the proposed design integrates comprehensive feature extraction, extensive feature robustness filtering, and intensive model optimization with ensemble-based performance calibration included in a single reproducible workflow (see Sections 4.1 and 5.1). This advancement enables better predictive performance as well as more consistent and robust radiomics experimentation across heterogeneous imaging datasets and clinical tasks.

This study provides detailed methodological insights into how the configuration of feature extraction pipelines, including the selection of feature extractors and the proportion of integrated features defined by the IBSI (see Figure 8.7), affects model performance and reproducibility. Using multiple open-source benchmark datasets, I demonstrated that variations in feature computation and preprocessing have a stronger impact on prediction performance than the choice of feature selection or automated model optimization techniques.

Data acquisition: The seven open-source datasets include the images and the ground truth labels from the WORC radiomics benchmark database [7] and the Lung Image Database Consortium (LIDC)-Image Database Resource Initiative (IDRI) dataset [20] from the The Cancer Imaging Platform (TCIA) portal (see Section 4.1). The segmentations of the ROIs from the WORC database were part of the downloaded data whereas for the segmentation of the LIDC-IDRI dataset an open source available tool set was used (see Section 4.1.1).

Method development and application: This thesis introduces the Radiomics

Processing Toolkit (RPTK), a modular and extensible framework designed to advance methodological reproducibility and performance optimization in radiomics. The framework unifies the essential components of a radiomics workflow—image preprocessing, feature extraction, feature selection, model training, and ensemble optimization, within a single reproducible system. Building upon existing standardized libraries such as PyRadiomics and MIRP [15,21], RPTK harmonizes feature definitions and integrates them with automated data curation, quality assessment, and adaptive learning strategies. The framework further extends conventional approaches by introducing segmentation artifact filtering, peritumoral region analysis, and cross-validated ensemble learning to enhance robustness and reduce bias.

For comparative evaluation, established frameworks such as AutoRadiomics [18] and complementary deep learning models implemented using PyTorch and Medical Open Network for AI (MONAI) [22, 23] were applied to benchmark the proposed pipeline. All components of RPTK were version-controlled, validated, and released as open-source software to ensure transparency, reproducibility, and community accessibility (see Section 4.1.11).

Result generation and analysis: This work contributes to the methodological understanding of radiomics performance and reproducibility by systematically evaluating the proposed RPTK framework against existing automated radiomics solutions and deep learning approaches across seven benchmark datasets. Through quantitative comparisons and statistical validation using bootstrapped performance metrics, the study demonstrates how standardized and self-configuring radiomics workflows can achieve competitive predictive accuracy while offering improved reproducibility and transparency. The analyses provide empirical evidence on the trade-offs between automation, model performance, and generalization in radiomics, offering practical guidance for the design of reproducible imaging biomarker studies. The first prototype of the proposed framework and its benchmarking methodology were published as a peer-reviewed Medical Image Computing and Computer Assisted Intervention (MICCAI) conference paper, establishing the groundwork for this thesis [24].

2. Longitudinal and Multimodal Radiomics Integration in Immunotherapy Response Prediction - Predict Study

The second major contribution of this thesis lies in demonstrating the potential of radiomics to capture temporal and multimodal imaging biomarkers for predicting immunotherapy response in advanced-stage lung cancer (see Sections 4.2 and 5.2). Using a longitudinal cohort from the Thoraxklinik Heidelberg, the study evaluated whether delta-radiomics features—quantifying temporal changes in tumor phenotype—can en-

hance predictive performance compared to conventional single time-point analyses. The results provide empirical evidence that longitudinal radiomics modelling improves early treatment response prediction, supporting adaptive and personalized immunotherapy strategies. Moreover, the integration of radiomics with complementary clinical and molecular parameters highlights the benefit of multimodal data fusion for outcome modelling in oncology

Data acquisition: Imaging and clinical data were collected and pseudonymized by clinical collaborators at the Thoraxklinik Heidelberg under ethical approval. The longitudinal dataset comprised multiple time points from patients undergoing immunotherapy for advanced-stage lung cancer. Primary lung tumors were segmented using a pretrained no new U-Net (nnU-Net) model [25, 26], and all segmentations were reviewed and validated by expert radiologists before inclusion in the analysis (see Section 4.2.2).

Method application: The standardized radiomics workflow implemented in the RPTK framework (see Section 4.2.3) was applied to extract delta-radiomics features, integrate multimodal clinical data, and perform predictive modelling. This application demonstrates the framework's capacity to handle complex, multi-timepoint datasets and to model temporal changes in tumor characteristics.

Result generation and analysis: Comparative experiments were performed to evaluate RPTK against AutoRadiomics and deep learning approaches. The results were statistically assessed using bootstrapped performance metrics to determine the added predictive value of longitudinal and multimodal feature integration. Findings from this study establish the feasibility and clinical promise of longitudinal radiomics workflows for early immunotherapy response prediction.

3. Radiomics-Based Colorectal Disease Characterization on Large-Scale CT Data - LiverCRC Study

The third major contribution of this thesis investigates the scalability and translational potential of radiomics in large-scale abdominal imaging datasets (see Sections 4.3 and 5.3). Using a multi-center cohort of liver CT scans, this study demonstrates that radiomics can be effectively applied to non-primary tumor sites for indirect disease characterization. Specifically, radiomics features extracted from healthy liver parenchyma and perihepatic regions were shown to discriminate patients with colorectal neoplasia from healthy controls, suggesting that the liver may serve as a surrogate

imaging biomarker reflecting systemic tumor processes along the gut–liver axis. Beyond its clinical relevance, this work establishes the robustness and adaptability of standardized radiomics workflows for bigger datasets and complex pathophysiological questions, underlining their potential for non-invasive disease detection and patient stratification in precision oncology.

Data acquisition: Imaging and clinical data were collected and pseudonymized by clinical collaborators at the Hector Cancer Institute and the Department of Radiology and Nuclear Medicine, University Medical Center Mannheim, under institutional ethical approval. Automatic liver segmentations were generated using the MultiTalent tool [27] and reviewed for quality assurance by technicians from the Department of Medical Image Analysis at the German Cancer Research Center Heidelberg (see Section 4.3.2). The study leveraged a large and heterogeneous dataset, enabling an assessment of radiomics scalability and reproducibility under real-world clinical imaging conditions.

Method application: The RPTK framework was employed to perform standardized feature extraction, selection, and model optimization (see Section 4.2.3). This application illustrates the capacity of RPTK to process high-volume data efficiently and to integrate harmonized radiomics features across imaging centers and protocols. In contrast to the approach described in the accompanying manuscript currently under submission [22], which focuses primarily on demonstrating the clinical feasibility of liver-based colorectal disease detection, the present work emphasizes the methodological validation of radiomics scalability and the benchmarking of standardized workflows. Here, RPTK is applied as a unified analytical framework to directly compare radiomics, AutoRadiomics, and deep learning approaches under identical experimental conditions, highlighting methodological effects on generalization and predictive performance.

Result generation and analysis: Comparative experiments were conducted to benchmark RPTK against AutoRadiomics and deep learning–based approaches. Model performance was evaluated using bootstrapped statistical metrics to assess predictive accuracy and robustness. The results demonstrate that harmonized radiomics workflows can generalize effectively to large-scale datasets, supporting their use in translational imaging studies. A detailed account of the clinical findings and their translational interpretation is included in the related publication [22], while this thesis extends the methodological analysis and provides an integrated comparative

evaluation across modeling paradigms.

1.3 Outline

The structure of my thesis follows a logical progression from the theoretical foundations to the methodological developments, applications, and the overall discussion and outlook.

Chapter 2 provides the scientific background necessary to understand the concepts and data used in the thesis. It begins with an overview of medical imaging modalities, focusing on Computed Tomography (CT) and Magnetic Resonance Imaging (MRI), and introduces deep learning applications in medical image analysis. The following section covers the medical background, including the biological foundations of cancer, advances in cancer treatment, and the role of biomarkers. The chapter concludes with a detailed introduction to the field of radiomics, explaining the complete radiomics workflow, from image acquisition and segmentation to feature computation, model building, and clinical integration, while also addressing current challenges such as reproducibility, standardization, and reporting.

Chapter 3 reviews the state of the art in both deep learning and radiomics. It discusses existing radiomics frameworks and tools, including customized radiomics approaches, the Workflow for Optimal Radiomics Classification (WORC) [19], and the AutoRadiomics framework [18]. The chapter highlights the methodological developments and limitations of current approaches, motivating the need for the self-configuring and reproducible framework developed in this work.

Chapter 4 describes the data and methods used in this thesis. The first part presents the design and implementation of the RPTK framework, including data preprocessing, feature computation, model optimization, and integration of AutoRadiomics and deep learning approaches. It also introduces the concept of data fingerprints, describes hardware usage, and provides source code availability for reproducibility. The second and third parts of the chapter present two major applications of RPTK: (1) the *Predict Study*, which investigates radiomics-based prediction of immunotherapy response using longitudinal imaging data; and (2) the *LiverCRC Study*, which explores the prediction of colorectal neoplasia from liver CT images in a large-scale dataset.

Chapter 5 presents the results of the three main experiments: the evaluation of the RPTK framework, the Predict Study, and the LiverCRC Study. It details the performance of RPTK across diverse datasets, its ability to automatically select informative features and models, and its comparison to state-of-the-art tools. For each

study, model performance, feature relevance, and clinical implications are analyzed.

Chapter 6 discusses the findings in the same structure as the results section, relating the observed outcomes to existing literature and highlighting the methodological and clinical relevance. It reflects on the datasets, methodological advances of RPTK, and performance outcomes in each application. The discussion also addresses limitations, assumptions, and potential directions for improving reproducibility and clinical translation.

Finally, Chapter 7 concludes the thesis by summarizing the key findings and methodological contributions of RPTK. It discusses the clinical and translational implications of the work, identifies its limitations, and provides an outlook on future research directions to further advance reproducible and generalizable radiomics.

Chapter 2

Background

2.1 Imaging Background

Medical imaging encompasses a broad range of technologies in cancer research used to visualize internal anatomical structures and physiological processes, playing a pivotal role in cancer detection, diagnosis, staging, treatment planning, and monitoring. Each imaging modality offers distinct advantages depending on the biological characteristics of the tumor, the anatomical site, and the clinical question at hand. In my radiomics studies I used two radiological three-dimensional imaging data modalities, MR and CT. In addition, deep learning applications on medical images are aiming to improve the quality of annotations like segmentations or detections of cancer and enabling fast processing of large amounts of different medical images.

Comparing two-dimensional (2D) medical imaging techniques like microscopy, ultrasound, or X-Ray, to three-dimensional (3D) radiological imaging techniques like CT and MRI highlight significant advantages of 3D Imaging techniques in cancer research and associated radiomics studies. Non-invasive 2D imaging modalities like X-ray radiography and ultrasound are widely available and fast applicable, they are fundamentally limited by overlapping anatomical structures, lower spatial resolution, and a lack of depth information. X-ray images, for instance, compress complex anatomical volumes into a single projection, often obscuring critical details and reducing diagnostic sensitivity—particularly in regions with overlapping tissues [28, 29]. These images are stored in specific data structures for processing and documentation of imaging related parameters like the Digital Imaging and Communications in Medicine (DICOM) or Neuroimaging Informatics Technology Initiative (NIFTI) data format including 3D imaging data and additional information for the clinical usage as well as for technical data curation [30]. Similarly, 2D ultrasound provides real-time imaging but is highly operator-dependent and offers limited reproducibility and anatomical context. While

microscopy-based imaging delivers cellular-level resolution, it is inherently invasive, often limited to small tissue biopsies, and prone to structural distortion during preparation—factors that restrict its representativeness and clinical applicability [31]. In contrast, 3D radiological imaging such as the most widely used techniques, Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) refers to a class of advanced medical imaging techniques that acquire non-invasive volumetric datasets, enabling multi-planar visualization and comprehensive quantitative analysis of anatomical and pathological structures, superior tissue contrast, depict the whole-tumor, and detailed anatomical structures [32]. These qualities are essential for accurate tumor localization, characterization, and treatment planning and can be quantitatively analyzed using radiomics to capture tumor heterogeneity, classify cancer subtypes, and guide personalized treatment strategies [28, 31, 32].

In 3D radiological imaging, each modality offers unique strengths and is selected based on the clinical question, patient conditions, and tissue characteristics. CT is particularly well-suited for imaging bone, lungs, and calcified structures due to its high spatial resolution and rapid acquisition [33–35]. MRI provides superior soft tissue contrast, making it ideal for imaging the brain, liver, prostate, and musculoskeletal system [28]. In oncology, 3D imaging is indispensable for tumor detection, staging, treatment planning, and monitoring therapeutic response. It enables detailed characterization of complex biological tumor features such as heterogeneity, perfusion, and necrosis [32]. Furthermore, it supports image-guided interventions and radiotherapy through accurate anatomical localization. Recent advances in image analysis techniques, including radiomics, allow for the extraction of high-dimensional quantitative features, offering the potential to develop novel imaging biomarkers that may enhance personalized diagnosis and treatment strategies.

The sections that follow will provide detailed overviews of CT and MRI, the two primary imaging modalities employed in my studies as well as deep learning applications on medical images to perform image segmentation and classification.

2.1.1 Computed Tomography

Computed Tomography (CT) is one of the most widely used non-invasive imaging modalities in clinical practice and research. Following the discovery of X-rays by Wilhelm Conrad Roentgen in 1895 [36], the development of the first CT scanner by Godfrey Hounsfield in 1971 marked a significant advancement in medical imaging. While conventional X-ray imaging provides two-dimensional projection images with limited soft tissue contrast and overlapping anatomical structures, CT revolutionized the field by enabling three-dimensional visualization through the mathematical re-

construction of multiple X-ray projections acquired from different angles [28]. This volumetric approach significantly enhances soft tissue differentiation and spatial resolution, allowing for more accurate clinical interpretation.

Calculation of Grey Values in CT

Modern computed tomography (CT) systems comprise an X-ray tube and a detector array mounted on a rotating gantry that surrounds the patient, who lies on a motorized table that advances through the scanner. During rotation, the X-ray tube emits a cone-shaped beam of polychromatic X-rays that traverse the patient's body. As these X-rays propagate through various tissues, they undergo differential attenuation depending on the physical composition and density of the materials encountered.

This attenuation process is described by the Beer-Lambert law, which models the exponential reduction in transmitted X-ray intensity as a function of the product of the material thickness and its linear attenuation coefficient (μ), as shown in Equation 2.1 [37]. The coefficient μ itself depends on intrinsic properties of the material, including electron density, atomic number, and the energy of the incident photons. Given measured values of the incident and transmitted X-ray intensity and the thickness of the material, μ can be directly calculated using the logarithmic form in Equation 2.2. As a result, tissues with higher density or atomic number, such as bone or iodine-enhanced vasculature, exhibit greater attenuation compared to soft tissues or air.

$$I = I_0 \cdot e^{-\mu \cdot x} \tag{2.1}$$

$$\mu = \frac{\ln(I_0/I)}{x} \tag{2.2}$$

I	Transmitted X-ray intensity after passing through the material
I_0	Incident X-ray intensity (before attenuation)
μ	Linear attenuation coefficient [cm^{-1}]
x	Thickness of the material [cm]

The detector array opposite the X-ray source measures the transmitted radiation, converting it into electrical signals that represent the cumulative attenuation along each projection path. These signals are then processed using reconstruction algorithms such as filtered back projection or iterative reconstruction to generate cross-sectional images. The resulting pixel values in these images are mapped onto the Hounsfield Unit (HU) scale (see Figure 2.1), a quantitative measure that normalizes tissue attenuation relative to water (0 HU) and air (-1000 HU) [38]. This enables consistent

tissue characterization across scans and patients, supporting accurate diagnosis, segmentation, and quantitative analysis in radiology [37].
Calculating Hounsfield Units (HU) as grey scale values in CT scans:

$$\mathrm{HU}(x,y) = 1000 \cdot \frac{\mu(x,y) - \mu_{\mathrm{water}}}{\mu_{\mathrm{water}}} \tag{2.3}$$

$HU(x,y)$ Hounsfield Units of each image pixel (x,y)

$\mu(x,y)$ Physical attenuation values of each image pixel (x,y)

μ_{water} Physical coefficient of water

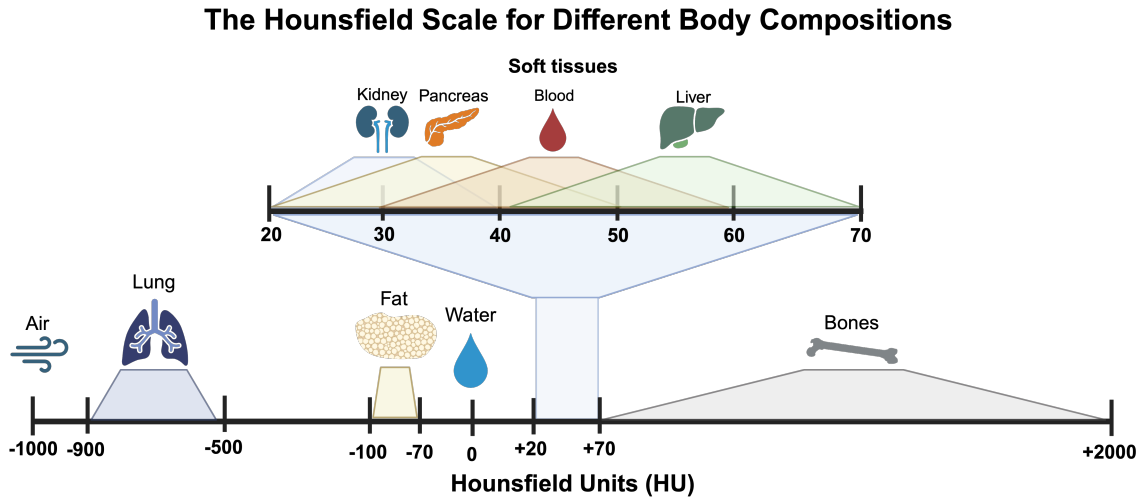


Figure 2.1. Hounsfield scale for characteristic values related to different human body compositions in a CT scan. HU values used for this categorization were extracted from [28].

Effect of Reconstruction Kernels on CT Gray Values

Although the HU scale provides a standardized quantitative measure of tissue attenuation, the observed gray values in CT images are not solely determined by physical attenuation coefficients. One important post-processing factor influencing HU values is the choice of reconstruction kernel, also referred to as convolution kernel or filter. These kernels are applied during the image reconstruction process to modulate spatial frequency content and balance spatial resolution against image noise [37].

Reconstruction kernels in CT are routinely classified into three categories based on their spatial frequency characteristics and clinical utility. Convolution kernels are developed by manufacturers like Siemens, Philips, or GE and therefore different between CT devices from these producers. Reconstruction kernels broadly classified by [39, 40]:

- **Smooth kernels** (e.g., Siemens kernels B10–B30, B40f) reduce image noise by suppressing high-frequency components, but result in lower spatial resolution. Typically used for soft tissue evaluation.

- **Medium/standard kernels** (e.g., Siemens kernels B40) balance spatial resolution and noise for general diagnostics.

- **Sharp kernels** (e.g., Siemens kernels B50–B80, B70f) accentuate edges and high-frequency content, making them ideal for bone, lung, and vascular imaging, albeit with increased noise.

The application of different reconstruction kernels can lead to systematic shifts in measured HU values, even when all other scan parameters remain constant. Studies have reported HU differences of up to 70 units between smooth and sharp kernels in musculoskeletal and phantom imaging contexts [41, 42]. This effect is particularly relevant for applications requiring quantitative interpretation, such as radiomics, bone mineral density assessment, and hepatic fat quantification.

The variability introduced by reconstruction kernel selection has emerged as a significant confounder in radiomics-based analyses, where the quantitative stability of image-derived features is essential. Several studies have shown that radiomic feature values can vary substantially depending on the applied convolution kernel, even when other acquisition parameters are held constant [43, 44].

To address the issue of kernel heterogeneity in retrospective or multi-center datasets, several correction strategies have been proposed. These include statistical harmonization methods such as ComBat [45] or Reconstruction Kernel Normalization (RKN) [46], which aim to align the statistical distributions of features across kernels, as well as data-driven approaches like CNN-based image translation [44]. Such techniques are particularly important when consistent kernel use cannot be guaranteed, as is often the case in large-scale multi-institutional studies or public imaging repositories.

Given the measurable impact of reconstruction kernels on image sharpness, noise levels, and even quantitative HU values, their selection must be carefully aligned with the clinical objective [40]. Kernel choice is not merely a technical preference but a critical decision that can influence diagnostic accuracy, reproducibility, and quantitative interpretation [42, 47]. Therefore, understanding and standardizing kernel use is essential when translating CT data into clinical decision-making tools or longitudinal assessments.

Application in clinical practice

In clinical practice, CT is widely used across a broad spectrum of diagnostic scenarios due to its speed, accessibility, and high-resolution imaging capabilities. It is particularly valuable in emergency settings, such as trauma, stroke, or suspected internal bleeding, where rapid and comprehensive visualization of both soft and hard tissues is critical [48]. CT is also routinely employed in oncological imaging for tumor detection, staging, and follow-up assessments [49]. In cases where MRI is contraindicated—such as in patients with metallic implants, pacemakers, or severe claustrophobia, CT offers a reliable alternative.

To enhance the visualization of vascular structures, lesions, organ perfusion, and changes in the Tumor Micro-Environment (TME), intravenous contrast agents based on iodine are often administered for an contrast-enhances CT [50, 51]. These agents increase the attenuation of X-rays in specific tissues, thereby improving contrast resolution and diagnostic accuracy, especially in CT angiography or tumor delineation [50,51]. Despite its reliance on ionizing radiation, the effective dose of a modern CT scan is relatively low and often comparable to the natural background radiation a person receives annually. Nonetheless, radiation exposure remains a concern, particularly for radiosensitive populations such as children and pregnant women, where alternative imaging modalities should be considered when possible. Dose optimization strategies, including automatic exposure control and iterative reconstruction algorithms, are implemented to minimize unnecessary radiation while maintaining image quality [52].

Overall, CT remains a cornerstone in diagnostic radiology, offering unparalleled speed, anatomical detail, and versatility for evaluating a wide range of pathologies throughout the human body.

2.1.2 Magnetic Resonance Imaging

Magnetic Resonance (MR) imaging is a medical imaging modality that has seen continuous development since its first applications in 1973 [53]. It exploits the fact that the human body is approximately 50–62% composed of water, depending on the age [54] and thus contains a high density of hydrogen atoms [54]. The nuclei of the hydrogen atoms (protons) are embedded within different molecular environments respond differently to magnetic fields and radiofrequency excitation [53, 55], enabling MRI to generate highly detailed images without relying on ionizing radiation [55]. MRI effectively generates grayscale maps of proton density and interactions within tissues [55] and therefore generates soft-tissue contrast and versatility stem from the

numerous contrast mechanisms available, most notably T_1, T_2, flow, diffusion, and fat/water separation, as well as physiological parametrization [53].

Clinically, MR Imaging is widely applied in cancer research and patient care, supporting diagnostic imaging, therapy monitoring, and screening efforts. Its non-ionizing nature makes it especially suitable for pediatric imaging, where minimizing radiation exposure is crucial; this, coupled with motion-reduction techniques, enhances image quality in young patients [56]. Additional advances, such as parallel imaging, compressed sensing, higher-field strengths, improved workflow, and AI integration have significantly reduced scan times and improved sensitivity and usability [53].

The MRI process begins with placing the patient within a strong static magnetic field. An electromagnetic pulse in the range of radio-frequencies is then applied. This exposed energy gets absorbed by the hydrogen protons, effecting a rise of the energetic state resulting in the spin changes of the protons which try to return to the natural state (B_0) by so called relaxation. Two different types of relaxations are known as Spin-Lattice (T_1) and Spin-Spin relaxations (T_2) which are occurring simultaneously at the same time:

1. **T_1 Relaxation (Spin–Lattice Relaxation):**
   The time it takes for longitudinal magnetization to recover to about 63% of its equilibrium state. This process involves energy exchange between protons and surrounding molecular structures [55].

2. **T_2 Relaxation (Spin–Spin Relaxation):**
   The time over which transverse magnetization decays to approximately 37%, due to loss of phase coherence among the protons is called T_2 [57].

Fat and protein molecules are large molecules and effective in absorbing energy resulting in a short T_1 and T_2. Smaller molecules like H_2O move quicker, which makes them more inefficient in absorbing the energy and result in a longer T_1 and T_2. However, T_2 relaxations are very much faster than T_1 as T_1 relaxation need to transfer energy to surrounding molecules, whereas T_2 relaxation is based on nearby spin interactions in the local magnetic field tending to faster energy transfer.
Signal timing is controlled via two key parameters:

- **Repetition Time (RT):** The interval between successive radiofrequency pulses targeting the same slice. A shorter TR emphasizes T_1 effects, while a longer TR reduces T_1 weighting and enables T_2 or proton-density contrasts [58].

- **Echo Time (ET):** The elapsed time between the radiofrequency pulse and the peak of the detected signal ("echo"). A longer TE accentuates T_2 contrast by allowing more dephasing, while a shorter TE minimizes it [57].

T_1-Weighted MRI

T_1-weighted MRI sequences are optimized to highlight differences in how quickly tissues recover their longitudinal magnetization after being disturbed by a radiofrequency pulse. These sequences use short Repetition Time (RT) and short Echo Times (ET), which emphasize tissues that recover quickly, like fat, which appears bright, while water-containing tissues like Cerebrospinal Fluid (CSF) appear dark due to their slower relaxation.

This imaging mode is particularly well-suited for visualizing anatomical structures and is widely used in clinical practice to evaluate normal tissue contrast and detect enhancing lesions, especially after the administration of gadolinium-based contrast agents, which selectively shorten T_1 values and increase signal intensity in abnormal tissues [55]. T_1-weighted images are considered reliable "baseline" structural references in deep learning–based image synthesis due to their stable anatomical representation and low variability [53].

T_2-Weighted MRI

T_2-weighted MRI sequences focus on how quickly tissues lose coherence (i.e., phase alignment) in the transverse plane after radiofrequency excitation, a process known as T_2 (spin–spin) relaxation. These sequences use long RT and long ET to allow the transverse signal to decay, making it possible to distinguish tissues based on their water content.

Tissues rich in free water, such as edema, inflammation, or CSF, retain signal longer and thus appear bright, while fat and denser structures lose signal more quickly and appear darker [57]. This makes T_2-weighted MRI particularly effective in identifying pathological processes involving increased fluid, such as tumors, infections, or white matter lesions. T_2-weighted images are especially valuable for detecting disease activity in the brain, as many conditions—including multiple sclerosis and gliomas present as T_2 hyper-intensities [53].

2.1.3 Deep Learning Applications

Deep learning has revolutionized the field of medical imaging by improving diagnostic accuracy and workflow efficiency. The rising trend in deep learning applications began

with the breakthrough success of Convolutional Neural Network (CNN), particularly following the ImageNet competition [59]. Early models like the Massive-Training Artificial Neural Network (MTANN) demonstrated promising results in tasks such as lesion detection and false positive reduction [59]. Since then, deep learning architectures have evolved rapidly, incorporating deeper, more complex models such as ResNet and DenseNet, which have set new standards for accuracy in diverse medical imaging tasks [60, 61]. State-of-the-art deep learning models in medical imaging have advanced significantly, incorporating Fully Convolutional Networks (FCNs) and their variants to enable high-resolution and precise image segmentation tasks [62]. Additionally, attention mechanisms have been increasingly adopted to enhance feature representation and model interpretability, while Generative Adversarial Networks (GANs) play a critical role in synthetic data augmentation, image-to-image translation, and modality synthesis, addressing data scarcity issues pervasive in medical imaging [63, 64]. These architectural innovations have expanded the reach of deep learning techniques beyond segmentation to include image classification, registration, and synthesis, thereby establishing deep learning as an indispensable framework for contemporary medical image analysis [64, 65].

In my thesis image segmentation and image classification were performed using deep learning models. Image segmentation was done to generate segmentations of the primary lung tumor in section 4.2 and generating segmentation of the liver in section 4.3. Image classification was done in order to compare the performance of trained models from RPTK based on radiomics features to the performance of deep learning models, on all datasets included in this thesis.

Medical Image Segmentation

Medical image segmentation is a fundamental annotation process in a radiomics workflow which aims to generate a mask with labels to differentiate specific objects or structures from the rest of the image (see Figure 2.2). This annotation is essential to define the ROI for extracting quantitative features in various medical imaging modalities. Segmentation facilitates precise diagnosis, treatment planning, and disease monitoring by enabling detailed analysis of anatomical structures and pathological regions [8,66].

Depending on the clinical objective and the level of detail required, various segmentation approaches are employed to address varying clinical and analytical needs, such as identifying general tissue classes or distinguishing between multiple occurrences of the same anatomical structure. Therefore, different segmentation strategies are used depending on the clinical task and imaging modality.

- **Semantic Segmentation:** Labels every pixel in an image according to class

(e.g., tumor with label 1 vs background with label 0). All pixels of the same class share the same label, but individual instances are not distinguished.

- **Instance Segmentation:** Distinguishes and labels each separate instance of a structure, even if instances are the same class (e.g., differentiating multiple nodules in the lung with labels 1 to 10), providing object-specific masks

Manual segmentation in 3D imaging is traditionally performed by radiologists or clinical experts iteratively on 2D slices, to segment the entire tumor in the 3D space, using specialized software tools such as the Medical Imaging Interaction Toolkit (MITK) [67]. This process, while accurate and informed by clinical knowledge, is labor-intensive, time-consuming, and subject to inter- and intra-observer variability [25]. Manual segmentations are often considered the "gold standard" or ground truth for training and validating automated methods.

In contrast, automated segmentation leverages machine learning, especially deep learning models like CNN, to generate segmentations with none to minimal human intervention. A notable example is the nnU-Net framework, which automatically configures its architecture and preprocessing steps to adapt to a given dataset [8]. More recently, foundation models like the Segment Anything Model (SAM) have shown promising generalizability across domains, although their effectiveness in medical imaging tasks still requires domain-specific fine-tuning and validation [66].

In order to provide quantitative evaluations of segmentation quality—such as comparing automated segmentations to expert-annotated ground truths, or assessing inter-annotator variability—several statistical metrics are commonly used. The Dice Similarity Coefficient (DSC) is the most frequently applied metric in this context, measuring the degree of spatial overlap between two segmentations (see equation 2.4). Other commonly used metrics include the Jaccard Index (also known as Intersection over Union) [68], the Hausdorff Distance for evaluating boundary discrepancies [69], and the Average Surface Distance (ASD), which quantifies the mean distance between corresponding surface points of two segmentation masks. These complementary metrics provide a more comprehensive assessment of segmentation performance by capturing different aspects such as overlap, boundary alignment, and surface agreement.

Calculating the Dice Similarity Coefficient (DSC) in order to compare different segmentations:

$$\mathrm{DSC}(A, B) = \frac{2 \cdot |A \cap B|}{|A| + |B|} \tag{2.4}$$

$\mathrm{DSC}(A, B)$ Dice Similarity Coefficient between segmentation masks A and B

$|A \cap B|$ Number of overlapping pixels (or voxels) in both segmentations

$|A|, |B|$ Number of pixels (or voxels) in each segmentation mask

The quality of image segmentation is critical for accurately capturing the full extent, shape, and internal heterogeneity of the ROI, which directly influences the reliability of radiomics feature extraction [21, 70]. Inaccurate or inconsistent segmentations can lead to biased or non-reproducible features, ultimately affecting the validity of downstream analyses and the clinical conclusions drawn from them [21]. Therefore, robust and precise segmentation is essential to ensure meaningful and trustworthy results in radiomics-based decision support.

Medical Image Classification

Deep learning has advanced the field of medical image classification, by offering improvements over traditional machine learning algorithms in terms of accuracy on big datasets in cancer research [71]. Among deep learning architectures, ResNet and DenseNet have become widely adopted due to their ability to extract hierarchical and complex patterns from medical imaging data [71–73]. In comparison, traditional machine learning models like random forest need image segmentation and extraction of radiomics features for training, whereas deep learning models can be applied directly on the images.

The depth (number of layers) of deep learning models plays a crucial role in determining their predictive performance for medical image classification. Deeper architectures (more layers) are generally capable of learning more abstract and complex features, which can improve classification accuracy on challenging medical imaging tasks [74]. However, increasing depth also raises the risk of overfitting, especially when training data are limited, making it essential to balance model complexity with generalization capability [75]. Architectures such as ResNet and DenseNet have been pivotal in addressing these challenges by introducing innovative connectivity patterns that facilitate training of deep models while mitigating issues like vanishing gradients and redundant feature learning [73, 76]. ResNet introduces residual connections that allow for effective gradient flow across many layers, enabling very deep networks to be trained successfully and yielding strong performance across various medical image classification tasks [73]. DenseNet enhances feature reuse via dense connections between all layers, reducing the number of parameters and encouraging richer feature propagation, which is especially advantageous in medical imaging where data can be limited [76]. Studies have also demonstrated that these architectures can achieve superior accuracy and robustness when compared to other convolutional neural networks, making them mainstays in current medical image analysis applications [72, 74, 76].

Deep learning models in medical imaging are often criticized as "black boxes" because their decision-making processes are not inherently transparent, limiting clinical trust and adoption. To address this, there has been significant research dedicated to developing Explainable AI (XAI) techniques aimed at unveiling how models arrive at predictions, thereby improving interpretability. Popular methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) generate heatmaps or saliency maps that highlight image regions influential to the model's decision, making it easier to verify clinical relevance [77,78]. However, deep learning explainability methods are facing several limitations. One major limitation is that many explainability approaches only provide post hoc interpretations, which may not fully reflect the actual decision-making process of the model, leading to potential misleading explanations [79]. Furthermore, interpretability methods often suffer from lack of consistency, where explanations generated for similar inputs may vary, reducing trust in model behavior [80]. Explainability techniques, such as saliency maps and heatmaps, can also be sensitive to noise and perturbations, making them sometimes unstable or hard to reproduce reliably [81]. Additionally, these methods generally do not guarantee clinical relevance, as highlighted regions might not correspond to medically meaningful features, which complicates validation by experts [80, 82]. Finally, deep learning models still pose a risk of bias and overfitting, and explainability alone cannot fully address these fundamental issues without careful model and data design. These limitations underline ongoing research needs to develop more reliable, stable, and clinically grounded explainability tools to support widespread adoption in medical imaging workflows.

2.1.4 Conclusion

The accurate selection and application of medical imaging modalities, along with rigorous image acquisition and reconstruction methods, form the cornerstone of quantitative imaging and subsequent radiomics analysis. In this chapter, key physical and technical principles underlying CT and MRI were presented, emphasizing their distinct strengths and limitations in clinical oncology. CT imaging offers exceptional spatial resolution and rapid volumetric acquisition, ideal for anatomical delineation of bone and calcified structures, yet it introduces inherent considerations such as ionizing radiation exposure and sensitivity to reconstruction kernel choices. As demonstrated, the choice of convolution kernel significantly impacts quantitative Hounsfield Unit values and derived radiomic features, necessitating careful harmonization strategies in multi-center studies or retrospective analyses. MRI, on the other hand, provides superior soft-tissue contrast without radiation exposure, leveraging tissue-specific magnetic resonance properties (T_1 and T_2 relaxations) to reveal tumor biology and tissue

heterogeneity.

Critically, the reliability and reproducibility of quantitative imaging features extracted for radiomics heavily depend on consistent imaging protocols, standardized reconstruction parameters, and robust image segmentation methods. Kernel heterogeneity, image noise, and reconstruction settings can introduce substantial variability in quantitative metrics. Recent methodological advances, including deep learning-based image normalization and statistical harmonization, offer promising solutions to mitigate these confounding factors.

Finally, the fundamental process of medical image segmentation, whether manual or automated, directly influences radiomic analyses. Automated methods leveraging convolutional neural networks, such as nnU-Net, enhance consistency and reduce observer variability, yet their accuracy depends strongly on robust validation using metrics such as the Dice Similarity Coefficient, Jaccard Index, and Hausdorff Distance. Thus, precise image acquisition, standardized processing pipelines, and accurate segmentation are indispensable for ensuring clinically meaningful radiomics features that support personalized oncology.

2.2 Medical Background

There are several clinical tasks where radiomics can get applied on like diagnosis, treatment decisions, and prognosis depend on a complex interplay of histological, molecular, and anatomical factors. The addressed classification tasks in this thesis reflect common clinical challenges where radiomics-based tumor characterization can contribute to patient care.

2.2.1 Foundations of Cancer Biology

Globally, one of five individuals will develop cancer between the ages of 0 and 74 [83]. In 2019 alone, there were an estimated 18 million new cancer cases, with lung ($\sim$2.09 million), breast ($\sim$2.09 million), and prostate ($\sim$1.28 million) cancers being the most commonly diagnosed mortality [83]. Cancer is a complex group of diseases characterized by uncontrolled cell growth, the invasion of surrounding tissues, and the potential to metastasize to distant organs [84]. It originates from normal cells that have undergone genetic and epigenetic alterations, disrupting critical regulatory processes such as cell cycle control, apoptosis, and DNA repair. These disruptions typically accumulate over time and may result from environmental exposures, inherited mutations, or stochastic errors in DNA replication. Mortality varies substantially by cancer type.

Pancreatic, liver, esophageal, and lung cancers are among the most lethal, often due to late diagnosis, fast tumor evolution, treatment resistance, spreading to different locations in the human body, and high cell proliferation activity. Pancreatic cancer, for instance, is typically diagnosed at advanced stages and exhibits rapid progression and resistance to therapy [85]. Similarly, hepatocellular carcinoma is known for its aggressive course, particularly in regions with a high prevalence of hepatitis infections [86].

At the molecular level, cancer arises through the activation of oncogenes, which promote proliferative signaling, and the inactivation of tumor suppressor genes, which normally function to restrain growth or trigger cell death in damaged cells [84, 87]. Epigenetic modifications, such as DNA methylation and histone acetylation, can also silence tumor suppressors or activate oncogenes without changing the DNA sequence. These molecular events contribute to genomic instability, a key enabling factor in tumorigenesis.

To understand the biological underpinnings of cancer, Hanahan and Weinberg proposed a conceptual framework known as the "Hallmarks of Cancer", describing the functional capabilities acquired during tumor development [88]. These include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, activating invasion and metastasis, reprogramming energy metabolism, and avoiding immune destruction [88]. These hallmarks are not isolated but interdependent and shaped by the TME, which includes stromal cells, immune infiltrates, and extracellular matrix components that collectively support tumor progression and therapeutic resistance [88].

Cancers are traditionally classified by their tissue of origin—such as carcinomas from epithelial cells, sarcomas from connective tissue, or hematological malignancies like leukemia and lymphoma—and increasingly by their molecular and histopathological subtypes [84]. This classification is crucial for prognosis and treatment. For example, lung cancers are divided into Small-Cell Lung Cancer (SCLC), an aggressive subtype with early metastatic potential (accounting for about 15% of lung cancers), and Non–Small-Cell Lung Cancer (NSCLC), which is more common (85%) and generally progresses more slowly [89, 90].

The lethality of cancer is often associated with its ability to metastasize, a multistep process involving local invasion, entry into the bloodstream (intravasation), survival in circulation, exit into distant tissues (extravasation), and colonization of secondary sites. This process is facilitated by epithelial-to-mesenchymal transition (EMT), enabling cancer cells to become motile and invasive. Furthermore, tumors evolve dynamically through clonal selection and adaptation, giving rise to intratumoral hetero-

geneity [91]. This heterogeneity spans genetic, epigenetic, and phenotypic differences between subclones and is a major obstacle to effective therapy [84].

The biological complexity of cancer spans multiple layers of the molecular landscape, from genomic and epigenomic alterations to transcriptomic shifts, proteomic remodeling, and metabolic reprogramming [92]. Each of these layers contributes uniquely to tumor initiation, progression, and therapy resistance. Considering only one molecular data layer limits the findings, can lead to wrong interpretations, or unseen effects in cancer research [92]. Capturing the complexity on cancer requires the integration of multi-omics data, which enables a more comprehensive and high-resolution characterization of cancer biology, identification of regulatory functions, and supports the identification of novel therapeutic targets and biomarkers [93]. Resulting multi omics profiles can be used for different tasks like patient stratification, biomarker discovers, pathway analysis, drug analysis, cancer subtype classification, or multi-omics data discovery [93].

Modern oncology aims to address these complexities by integrating insights from cancer biology into clinical practice. This includes refining diagnostic criteria, stratifying tumors by molecular features, and developing personalized treatment strategies. Therapies now extend beyond traditional surgery and chemotherapy to include targeted therapies, immune checkpoint inhibitors, and epigenetic drugs, many of which exploit vulnerabilities arising from the cancer's molecular makeup. A deep understanding of tumor biology, especially the interplay between genetic mutations, TME, and evolutionary dynamics—is essential to improve patient outcomes and guide future research directions [84, 94].

2.2.2 Advances in Cancer Treatment

Cancer remains a highly heterogeneous disease, characterized by diverse molecular and cellular alterations that enable immune evasion, tissue invasion, and therapeutic resistance. Over the past decades, cancer treatment has evolved significantly, from non-specific, broadly cytotoxic strategies to personalized and molecularly targeted approaches.

Historically, surgical resection was among the earliest and most effective interventions, especially for localized tumors [95]. Surgery remains a cornerstone in oncology for tumor removal, staging, and histopathological classification [95]. Radiation therapy and chemotherapy soon followed as systemic therapies. Radiation induces DNA damage to kill rapidly dividing cells but also damages surrounding normal tissue [95]. Chemotherapy involves low-molecular-weight cytotoxic agents targeting rapidly proliferating cells but also lacks specificity and can cause significant toxicity to normal

organs [95].

In hormone-sensitive malignancies such as certain breast and prostate cancers, hormone therapies like estrogen receptor blockers or androgen deprivation, have become standard-of-care [95]. These treatments block hormonal pathways that are critical for tumor progression.

More recently, precision oncology has shifted the focus toward minimizing systemic toxicity and maximizing tumor-specific efficacy. Targeted therapies have been developed to inhibit oncogene-driven pathways that are frequently mutated in cancer subtypes (e.g., genes like Epidermal Growth Factor Receptor (EGFR), Anaplastic Lymphoma Kinase (ALK), B-Raf proto-oncogene serine/threonine kinase (BRAF)) [96]. These therapies are tailored based on molecular profiling and are often more effective in tumors with specific actionable alterations [95].

Among the most transformative advances in recent years has been the development of immunotherapy, particularly Immune Checkpoint Inhibitor (ICI). These agents work by blocking regulatory pathways that suppress immune activation, thereby enabling cytotoxic T cells to recognize and destroy cancer cells. The most clinically successful ICIs target the Anti-Programmed cell Death protein-1 (PD-1) or its ligand Anti-Programmed cell Death Ligand-1 (PD-L1), and Cytotoxic T Lymphocyte-associated Antigen 4 (CTLA-4). Therapies targeting the PD-1/PD-L1 axis have demonstrated durable responses in multiple cancer types, including melanoma, Non-Small Cell Lung Cancer (NSCLC), and urothelial carcinoma but can also cause specifically immune system induced complications like hyper-progression or pseudo-progression [88, 97].

PD-L1 expression, as detected by immunohistochemistry, has emerged as a key biomarker for predicting response to these agents, although it remains imperfect due to tumor heterogeneity and dynamic expression. As such, efforts are underway to develop multi-dimensional predictive models that integrate PD-L1 status with Tumor Mutational Burden (TMB), immune infiltration, and other immune-oncology markers [98, 99].

Beyond checkpoint inhibition, novel immunotherapeutic strategies are being explored. These include Bispecific T-cell Engagers (BiTEs) that link Cluster of Differentiation 3 (CD3) on T cells to tumor-associated antigens, antibody-drug conjugates (ADC) that deliver cytotoxins to cancer cells via antigen-specific antibodies, and oncolytic virus therapies designed to selectively lyse tumor cells and stimulate anti-tumor immunity [95]. Photodynamic therapy, while niche, offers localized tumor destruction via photosensitizer activation [95].

In addition to molecular predictors, patient-related and systemic factors substan-

tially influence immunotherapy outcomes. The Eastern Cooperative Oncology Group (ECOG) Performance Status provides a concise measure of functional capacity and treatment tolerance, and remains a cornerstone parameter guiding prognosis and therapeutic decision-making in oncology [100]. Likewise, the serum concentration of C-Reactive Protein (CRP) reflects systemic inflammation and has been associated with both prognosis and treatment response in patients receiving PD-L1-directed immunotherapy [101].

In the context of immunotherapy, treatment response assessment is complicated by atypical tumor growth patterns that are not adequately captured by conventional criteria. Therefore, clinical trial design and response evaluation methods have also adapted to unique immunotherapy characteristics. Traditional criteria such as Response Evaluation Criteria In Solid Tumors (RECIST) focus on tumor size reduction as an indicator of efficacy. However, ICIs can induce atypical response patterns such as Pseudo-progression, where tumors appear to grow before shrinking due to immune cell infiltration [102]. On the other side immunotherapy is also known to cause Hyperprogression which indeed refers to a detrimental increase in tumor burden following therapy initiation. This necessitated the development of immune Response Evaluation Criteria In Solid Tumors (iRECIST), a modified guideline that accounts for immune-related responses and progression [102–104].

These developments underscore a paradigm shift in oncology—from a one-size-fits-all model to a nuanced approach incorporating genomics, immunology, and dynamic treatment adaptation. Within this context, evaluating and predicting patient response to PD-L1–targeted therapies has become a critical research priority [98]. The work presented in this thesis aims to contribute to this field by assessing outcome metrics and response predictors in the setting of PD-L1 immunotherapy.

2.2.3 Biomarker

Biomarkers are measurable indicators of normal biological processes, pathological conditions, or responses to therapeutic interventions and ideally anticipate clinically meaningful outcomes or endpoints that are otherwise challenging to directly assess [105]. They may originate from a wide range of biological and physiological sources, including nucleic acids, proteins, cells, metabolites, or imaging-derived features [106]. They play a central role in early disease detection, diagnosis, prognosis, risk stratification, therapeutic selection, and monitoring of treatment response [14, 106]. Their importance is especially evident in oncology and neuro-oncology, where disease heterogeneity necessitates individualized clinical approaches [107, 108].

Biomarkers can be categorized based on the biological level at which they are ex-

pressed. These include molecular biomarkers (e.g., enzyme activity), cellular biomarkers (e.g., serum electrolytes), tissue-level biomarkers (e.g., Glycated Hemoglobin (HbA_1c)), organ-level biomarkers (e.g., blood pressure, echocardiographic findings), and whole-body biomarkers (e.g., body weight, body size) [105]. From a functional perspective, biomarkers are commonly classified into three major categories: *diagnostic*, *prognostic*, and *predictive* biomarkers [106]:

- **Diagnostic biomarkers** are used to detect or confirm the presence of a specific disease or pathological condition (e.g. Circulating long non-coding RNAs (lncRNAs) in blood for early cancer detection) [109].
- **Prognostic biomarkers** provide information about the likely progression or outcome of a disease, regardless of treatment (e.g. increased Circulating Tumour Cells (CTC) expression in squamous cell carcinoma of the head and neck) [110].
- **Predictive biomarkers** help identify individuals who are more likely to benefit from or respond to a specific therapeutic intervention (e.g. PD-L1 expression for predicting the response to ICI) [111].

Ideal biomarkers are characterized by high sensitivity and specificity, reproducibility, and robust clinical relevance. In addition, they should ideally be accessible, cost-effective, and minimally invasive to facilitate widespread clinical implementation [106, 108].

Recent research emphasizes the integration of multiple biomarker types, especially combining molecular and imaging biomarkers as a promising strategy to enhance diagnostic precision and therapeutic decision-making [107, 108]. This multidimensional approach aligns with the goals of precision medicine by supporting more tailored and effective patient management [14].

In oncology, most biomarkers are acting on the molecular level, as cancer arises based on changes on this level. Molecular biomarkers typically encompass genomic, transcriptomic, proteomic, or metabolomic indicators obtained from biological specimens. In contrast, imaging-based biomarkers are derived from non-invasive modalities such as MRI, CT, PET, or ultrasound and provide spatial, morphological, or functional information about tissue and disease processes [107, 108].

2.2.4 Conclusion

In summary, cancer represents a biologically and clinically complex disease driven by multilayered genomic, epigenomic, transcriptomic, proteomic, and metabolic al-

terations. This heterogeneity poses significant challenges for diagnosis, prognosis, and treatment, necessitating precision strategies grounded in comprehensive molecular understanding. Over time, oncology has transitioned from broadly applied cytotoxic treatments toward personalized approaches leveraging targeted therapies and immunotherapies, particularly immune checkpoint inhibitors such as PD-1/PD-L1 blockade. These therapeutic innovations are underpinned by the identification and integration of molecular and imaging biomarkers that offer predictive and prognostic value. The interplay between tumor microenvironment, mutational landscape, and therapy response continues to fuel the need for high-resolution, non-invasive methods for patient stratification and response monitoring. Radiomics, a field focused on extracting quantitative imaging features, has emerged as a promising tool to bridge molecular biology and clinical imaging, offering novel insights for tumor characterization and treatment outcome prediction. This thesis builds upon these developments by exploring how radiomics can support clinically relevant classification tasks in oncology.

2.3 Radiomics

Radiomics is an emerging field at the intersection of medical imaging and computational analysis, where a large number of quantitative features are extracted from standard medical images using data-characterization algorithms [14, 112]. These features ranging from shape descriptors to texture patterns and statistical summaries aim to capture underlying tissue characteristics that are often imperceptible to the human eye. The fundamental idea behind radiomics is to convert images into mineable data and apply machine learning or statistical modeling to uncover relationships between image features and clinical outcomes to support comprehensive data integration (e.g. integration of genetic sequencing results, or clinical parameters like blood pressure) and explainability (showing explicit radiomics features describing reproducible characteristics of the tumor) [113, 114]. A typical radiomics workflow includes image acquisition, segmentation of regions of interest (ROIs), feature extraction, and modeling [1]. Radiomics thereby enables a non-invasive means of phenotyping disease, which can enhance diagnosis, predict prognosis, and guide treatment decisions. Although initially developed within the context of oncology, radiomics has been applied to a variety of medical domains. These include neurology, cardiology, pulmonology, and infectious diseases [115]. In neurology, radiomics has been extensively applied in glioma research, analyzing tumor heterogeneity, TME and molecular subtypes using MRI-derived features combined with advanced modeling approaches [116]. Beyond

oncology, radiomics is also investigating non-oncologic neurological conditions like stroke, aneurysms and demyelinating disease [117]. Radiomics has found its impact in widespread application in medicine and oncology. It has been used to characterize tumor heterogeneity, predict treatment response, assess prognosis, and even infer underlying genetic mutations, commonly referred to as Radiogenomics [14, 113]. In lung and head-and-neck cancers, radiomics features have demonstrated strong prognostic value independent of clinical factors, often outperforming them in predictive models [113]. Additionally, in radiotherapy, radiomics supports adaptive planning and early toxicity prediction [118]. Radiomics offers several advantages. It is non-invasive, leverages already available imaging data, and can provide high-dimensional insights without the need for biopsy or additional procedures [14]. Moreover, it enables personalized medicine by uncovering patient-specific imaging biomarkers, the ability to incorporate multi-domain information for more precise and comprehensive analysis, as well as explainable and reproducible features [1, 119]. Radiomics ultimately aims for prospective clinical application by establishing standardized processing of routine clinical care data, enabling complex and comprehensive analyses that integrate information from multiple domains. Integration of radiomics workflow into the clinic for prospective use needs proper evaluation but can increase quality and reduce labor intensive radiological evaluations (see Figure 2.5).

However, radiomics faces notable limitations. These include lack of standardization of radiomics terms, imaging protocols, variability in feature extraction across platforms, and insufficient external validation [1, 11]. Many studies remain retrospective and lack clinical integration, leading to a gap between research and real-world application. Systematic reviews have shown that only a limited number of radiomics approaches currently meet the threshold for high-quality clinical evidence [120].

2.3.1 Standardization in Radiomics

While radiomics promises non-invasive phenotyping and personalized treatment guidance, its clinical implementation remains impeded by variability and insufficient methodological transparency. Quantitative Radiomics imaging biomarkers are highly sensitive to variations in image acquisition, reconstruction, and preprocessing protocols. The multi-stage radiomics workflow is particularly susceptible to inconsistencies across institutions, software, and practices. Standardization is thus crucial to enable reproducibility, data comparability, and real-world deployment of imaging biomarkers [119, 121, 122].

Standardization ensures comparable radiomics features regardless of imaging protocols, vendors, or computational pipelines between studies. It enhances scientific

rigor, facilitates multi-center research, and builds confidence in radiomics-derived biomarkers for regulatory and clinical use. This need is especially urgent given the heterogeneity in current practices and the tendency for subtle technical differences to significantly impact results [122, 123].

Several expert-developed guidelines and checklists which are used in the radiomics community provide structured recommendations to improve quality, transparency, and clinical relevance of radiomics study consistencies:

- **IBSI (Image Biomarker Standardization Initiative)**: Aims to standardize the definition and calculation of radiomic features to enhance reproducibility and comparability across studies and software platforms. [119].

- **TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis)**: Provides a structured framework for transparent reporting of multivariable prediction models, including those based on radiomics. [124].

- **RQS (Radiomics Quality Score)**: A semi-quantitative scoring system to assess methodological rigor in radiomics studies including TRIPOD criteria, with the goal of encouraging reproducibility, transparency, and clinical relevance [14].

- **CLAIM (Checklist for Artificial Intelligence in Medical Imaging)**: Offers structured guidance for reporting standards for AI-driven medical imaging studies, especially those using deep learning. [125].

- **CLEAR (CheckList for EvaluAtion of Radiomics)**: A guideline designed specifically for the evaluation and reporting of radiomics studies, focusing on reproducibility, transparency, and interpretability [12].

- **ARISE (Assessment for Radiomics Implementation Study Excellence)**: A harmonized framework assessing radiomics studies from the perspective of clinical applicability and methodological integrity [13].

These guidelines and checklists have been introduced to improve the methodological rigor, transparency, and clinical relevance of radiomics studies. While some of these guidelines are domain-specific to radiomics (e.g., IBSI, RQS, CLEAR, ARISE), others originate from broader methodological domains (e.g., Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) for prediction models, CLAIM for Artificial Intelligence (AI) in imaging). The IBSI

provides standardized definitions of reproducible radiomics feature extraction, including formula, as well as general recommendations on the radiomics workflow design, and validate feature reproducibility using CT, Positron Emission Tomography (PET), and T_1 MR images of 51 patients [119]. It is the only guideline offering concrete implementation-level specifications for radiomics pipelines. The RQS attempts to quantitatively evaluate radiomics studies through a point-based system reflecting various methodological aspects—many of which are informed by general principles from the TRIPOD checklist, which itself offers a structured framework with 22 items for transparent reporting of multi-variable prediction models [14]. CLAIM provides a publication-focused checklist for AI applications in medical imaging, guiding researchers on proper terminology, dataset documentation, and evaluation reporting [125]. CLEAR is a domain-specific checklist for radiomics, aiming to promote transparency and reproducibility through 58 reporting items covering image processing, feature extraction, modeling, and interpretation [12]. ARISE contributes a translational lens by providing 13 high-level recommendations to evaluate whether a radiomics study adequately addresses clinical applicability, regulatory relevance, and workflow integration [13]. Although these frameworks differ in scope and granularity, many share overlapping reporting criteria particularly concerning data curation, validation, and reproducibility which may lead to partial redundancy across checklists.

Despite the growing number of guidelines, several practical and conceptual limitations remain. IBSI, while rigorous in its mathematical formalism, was validated using only 51 patient cases across CT, PET, and T_1-weighted MRI, its generalizability to other modalities like T_2-weighted MRI or advanced multi-parametric sequences and other entities remains uncertain and may require further harmonization or adaptation [119]. RQS, though widely cited, is constrained by its arbitrary and uneven point allocation: for instance, prospective data collection yields disproportionately high scores (+7), whereas critical methodological aspects such as feature selection or validation only yield marginal gains (+1), thereby distorting overall quality assessments [14]. Moreover, these weights lack empirical justification and can overemphasize study design over reproducibility. TRIPOD, being model-agnostic and domain-independent, does not address imaging-specific challenges (e.g., segmentation variability, acquisition harmonization), and is intended primarily for reporting purposes, not for guiding the methodological development of radiomics workflows [124]. CLAIM, similarly, offers little radiomics-specific guidance, and remains primarily focused on deep learning models rather than handcrafted radiomics features [125]. While CLEAR provides a comprehensive list of transparency criteria, it is based on limited consensus and offers no prescriptive guidance on best practices, challenging interpretation and implemen-

tation largely up to the authors [12]. ARISE, although innovative in its translational focus, operates at a high level of abstraction; it provides no concrete instructions for methodology, and is currently confined to documentation guidance rather than influencing the analytical pipeline itself [13]. As such, most guidelines except for IBSI are aiming for a better suited structuring and evaluating publications than for actively informing radiomics pipeline design or implementation.

Beyond the structural and conceptual limitations of current radiomics guidelines, reproducibility in radiomics is further challenged by a number of well-documented technical factors along the processing pipeline which are hindering clinical translation. Multiple studies have demonstrated that image acquisition and reconstruction parameters, such as slice thickness, dose, and reconstruction algorithm, have a significant impact on the stability of radiomics features—particularly for texture- and shape-based metrics [10, 126]. In addition, preprocessing steps such as voxel resampling, pixel discretization, intensity normalization, and filtering (e.g., wavelet or Laplacian of Gaussian) can drastically alter feature values and thus affect both intra- and inter-study reproducibility [119,127,128]. While some filtering strategies may improve predictive performance, they may also reduce reproducibility if not harmonized appropriately [129]. Moreover, segmentation variability, although generally less impactful than acquisition-related factors, remains a critical source of uncertainty, especially for smaller lesions or irregular structures [21, 130]. Differences in feature extraction software also contribute to inconsistencies, even when nominally compliant with standardized definitions [119]. These findings underscore the pressing need for not only reporting standards but also pipeline harmonization and empirical benchmarking of reproducibility across modalities, software environments, and clinical contexts.

2.3.2 The Radiomics Workflow

Radiomics has rapidly transitioned from a purely research-oriented concept to a valuable tool in clinical oncology, offering non–invasive, quantitative insights into tumor phenotypes that can enhance diagnosis, prognostication, and treatment planning. A typical radiomics pipeline encompasses sequential steps of study design, image acquisition, preprocessing (e.g., denoising, normalization), tumor segmentation, feature extraction, feature selection, model development, and validation [11, 119]. However, significant heterogeneity exists in how these phases are defined and executed across the literature. In response, the IBSI has proposed detailed guidelines for radiomics studies with the focus on feature computation [119], and the SPP2177 consensus statement has delineated a seven-phase workflow with 37 aspects to harmonize radiomics studies [11]. In the following sections, I introduce the general radiomics workflow and

then discuss the key differences observed among published protocols.

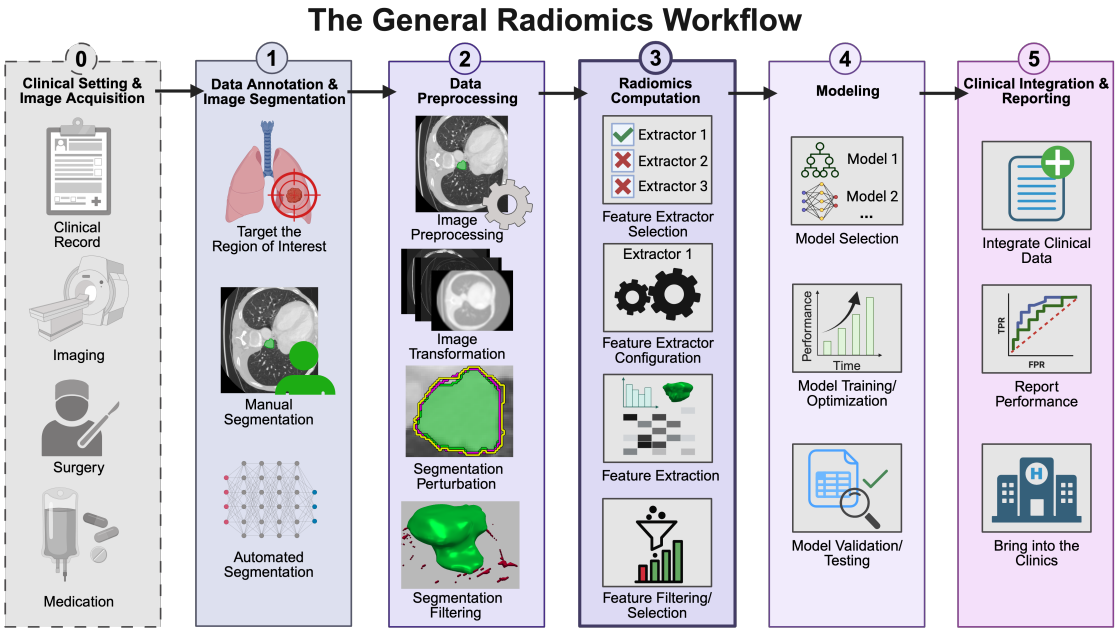


Figure 2.2. Overview on the general sequence of procedures covered by radiomics studies as the general radiomics workflow, starting at study design and data acquisition and ending with the integration of the models into the clinical setting. The clinical setting and image acquisition in the beginning is related to the circumstances and the starting point of the radiomics study and therefore not always necessary. The main part of radiomics processing in highlighted in the middle of the figure. This workflow is based on the consensus workflow previously defined by Floca et al. [11].

Whereas the IBSI workflow focuses narrowly on the steps leading up to feature calculation namely data conversion, post-acquisition processing, segmentation, interpolation, re-segmentation, ROI extraction and intensity discretisation [119], the SPP2177 consensus extends this pipeline through feature extraction into dedicated modeling and reporting phases based on consensus voting coming from a Delphi process of radiomics experts [11]. However, neither framework explicitly addresses the ultimate goal of clinical deployment: applying a validated radiomics model back onto new imaging studies to support real-time decision making. In routine practice, the completed radiomics pipeline is run on external patient cohorts, and the resulting risk scores or phenotypic biomarkers are integrated with clinical and pathological data to guide diagnosis, prognosis, or treatment selection. The actual goal of using the model which learns on selected features is to apply it in the clinics without redoing the entire validation and feature selection process. To capture this end-to-end continuum, Figure 2.2 presents my generalized radiomics workflow covering important aspects in

my thesis, which not only aligns with the IBSI and SPP2177 recommendations but also closes the loop by showing how model outputs are fed back into the clinical environment as decision support tools. However, prospective use of pretrained radiomics models on new data should be handled carefully to ensure the reliable application of the radiomics model to the predefined selected feature set. The model must be rigorously trained, tested and calibrated via validation on heterogeneous datasets to quantify and mitigate the effects of imaging-protocol variability, segmentation differences and technical noise on the feature selection and on the model performance in order to have trustworthy predictions on prospective data [131, 132]. Moreover, permissible clinical use cases should be explicitly defined by the original study design and patient recruitment criteria for the training, validation and test cohorts, and any subsequent alteration of the radiomics pipeline or feature space should prompt a full re-evaluation to confirm feature stability and clinical validity [131, 132].

2.3.3 Study Design & Image Acquisition

A robust radiomics study begins with a clearly articulated clinical question and testable hypotheses, such as predicting treatment response or patient survival, that guide all downstream decisions [11]. Prior to data collection, ethical approval must be secured, detailing patient consent procedures and data-protection measures. Key outcome measures (e.g., overall survival, recurrence-free interval) and strict inclusion/exclusion criteria (tumor size thresholds, prior therapies) are defined to minimize bias and confounding. Recognizing the impact of scanner and protocol variability, one should plan for data clustering and heterogeneity mitigation—either through stratified sampling or harmonization algorithms. The choice of imaging modality (CT, MRI, PET) must reflect the underlying biology of interest, and imaging parameters (slice thickness, reconstruction kernels, field of view) should be recorded meticulously as metadata. Finally, DICOM images are retrieved via secure pipelines, pseudonymized at import, and converted to standardized formats (e.g., NIfTI or DICOM-RT) to enable reproducible batch processing [114].

The study design is inherently tied to the primary aim and should balance technical feasibility with clinical applicability. In personalized oncology, careful design is vital for refining patient stratification and guiding treatment selection, with the ultimate goal of improving survival. In the context of the Predict approach, this means identifying non-responders at the earliest treatment stage to avoid ineffective and potentially harmful therapies, thereby preventing unnecessary tumor progression. Anticipating and addressing technical limitations—such as imaging quality or algorithm generalizability and clinical constraints, including cohort representativeness and

follow-up duration, is essential to ensure meaningful and clinically relevant outcomes. An illustrative example of patient stratification for treatment selection, and its role in advancing personalized oncology, is shown in Figure 2.3.

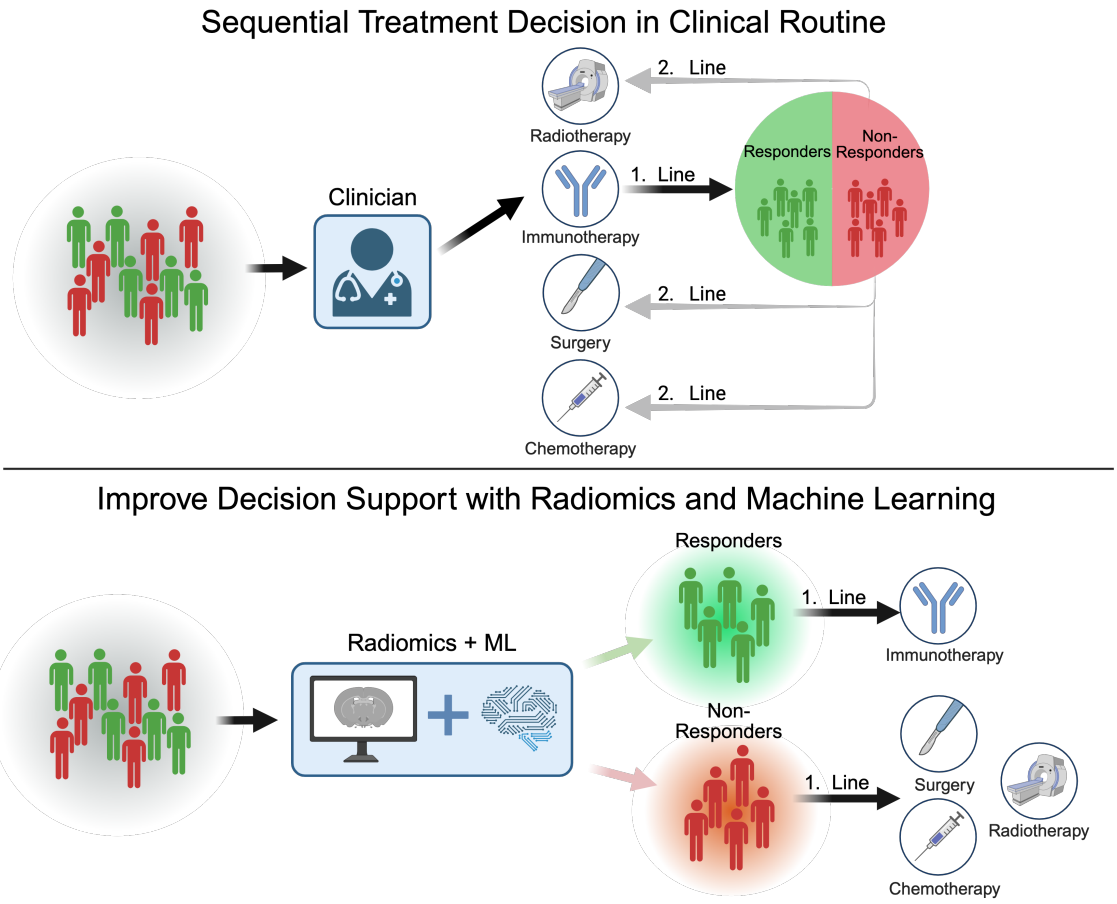


Figure 2.3. Conceptual overview of patient stratification for optimizing personalized oncological treatment design. The upper section illustrated the traditional sequential application of multiple treatments where non-responding patients of first line treatment get assigned to another treatment. The lower section illustrates the aim of of the Predict study (Section 4.2) as a decision support system for early patient treatment response stratification enabling timely therapy adjustments to avoid unnecessary tumor progression for increased patients survival in an earlier stage of the disease.

2.3.4 Data Annotation & Image Segmentation

Accurate delineation of the region of interest (ROI) is critical, as segmentation variability can substantially affect feature reproducibility (see section 2.1.3) [119]. Depending on resources and task complexity, expert radiologists may perform manual annotations on 2D slices or 3D volumes, which is labor intensive but remain the gold standard. Alternatively, automated methods ranging from classical thresholding and

region growing to modern deep-learning models (e.g. U-Net, nnU-Net) can accelerate this process, provided they are trained and validated on representative datasets [8,25]. All manual- and automatic segmentations should undergo quality checks against expert references using metrics such as Dice similarity coefficient, Jaccard index, or Hausdorff distance, supplemented by visual review and inter-observer concordance studies to ensure clinical plausibility [119]. The type (instance segmentation, or semantic segmentation) and meaning of the performed segmentation (e.g. label for tumor, label for background, label for organs) should be clearly defined. The label of the target object in the segmentation should be correctly documented as this influences the further processing of the data and helps to identify the correct ROI.

2.3.5 Data Preprocessing

Before feature computation, images and masks must be rendered comparable across patients and scanners. Volumetric resampling is performed to achieve isotropic voxel spacing [133, 134], and segmentation artifacts (small disconnected components) are corrected algorithmically. These segmentation artifacts occur after automated as well as manual segmentation and should be verified before extracting features. Intensity normalization (via z-score scaling or N4 bias field correction) is applied to mitigate scanner-specific variations, with MR data often requiring additional normalization, whereas CT scanners output calibrated Hounsfield units. When multi-modal data are used (e.g. CT or T_1/T_2 MRI), rigid or deformable registration aligns sequences to a common coordinate space. To assess feature robustness, it is recommended to apply image filters (Laplacian of Gaussian, wavelets) [10] and segmentation perturbations [21], while clinical covariates undergo imputation for missing values and removal of constant or highly correlated variables [119].

2.3.6 Radiomics Feature Computation

Radiomic features quantify tumor characteristics beyond human perception, categorised across shape, intensity, and texture, as defined by the Image Biomarker Standardization Initiative (IBSI), which specifies 11 feature classes and 169 benchmarked features [119].

- **Shape features**: Features about the segmentation and geometric descriptors (e.g. volume, surface area, sphericity).
- **First-order intensity (histogram) features**: Features describing general statistics about grey value distributions within the ROI (e.g. mean, variance,

skewness, percentiles, etc.).

- **Texture features**: Features describing spatial distribution of Grey-values within the ROI including different metrics capturing Grey-level patterns.

Radiomics Feature Extraction

Radiomic features fall into two broad categories: those standardized and validated by the IBSI, and non-IBSI descriptors which, although widely used, often lack cross-platform reproducibility and formal benchmarking [119]. IBSI features are grouped into 11 families, morphological, first-order intensity, and nine texture matrices to describe specific image properties without prescribing every individual metric. For example, shape features quantify geometric properties (volume, surface-to-volume ratio), first-order features summarize the overall intensity distribution (mean, variance, skewness), and texture families capture spatial grey-level patterns at varying scales (e.g. coarseness in NGTDM, zone size in GLSZM) [119].

The IBSI categorizes 169 radiomic features into 11 distinct families, each capturing a specific aspect of the image phenotype [119]:

- **Morphologic characteristics (26 features):** Geometric descriptors of the ROI—volume, surface area, compactness, sphericity, elongation, flatness—that quantify three-dimensional shape and size.
- **Local intensity (LI) (2 features):** Statistics computed from each voxel's immediate neighborhood (e.g. local mean and variance), reflecting fine-scale intensity variations.
- **Intensity-based statistics (IS) (18 features):** First-order descriptors summarizing the grey-level distribution within the ROI (mean, median, variance, skewness, kurtosis, percentiles).
- **Intensity histogram (IH) (23 features):** Histogram-derived metrics (energy, entropy, uniformity, histogram variance, percentile thresholds) capturing the overall grey-level frequency distribution.
- **Intensity–volume histogram (IVH) (5 features):** Features relating intensity thresholds to volumetric fractions (e.g. volume above threshold, area under the cumulative histogram), describing intensity–volume relationships.
- **Gray-level co-occurrence matrix (GLCM) (25 features):** Second-order texture features quantifying pairwise grey-level spatial dependencies at specified offsets (contrast, dissimilarity, homogeneity, energy, correlation, cluster metrics).

- **Gray-level run-length matrix (GLRLM) (16 features):** Texture metrics measuring the length of consecutive runs of identical grey-levels along given directions (short-run/long-run emphasis, run-length variance).
- **Gray-level size-zone matrix (GLSZM) (16 features):** Features quantifying the size distribution of connected zones of constant intensity in 3D (small-zone/large-zone emphasis, zone size variance).
- **Gray-level distance-zone matrix (GLDZM) (16 features):** Similar to size-zone but incorporating spatial distance between voxels, capturing texture scale and distance-dependent zone characteristics.
- **Neighborhood gray-tone difference matrix (NGTDM) (5 features):** Metrics that compare each voxel's intensity to the average of its neighborhood, capturing local texture coarseness, contrast, and busyness.
- **Neighboring gray-level dependence matrix (NGLDM) (17 features):** Measures of how a voxel's intensity depends on its neighbors at specified distances (low-dependence/high-dependence emphasis, dependence variance).

Second or higher order features are features describing matrices which are calculated based on the discretizised image. Pixel discretization is a crucial process in radiomics feature extraction which effects the signal to noise ratio (see Figure 2.4. It is important to find a good balance between the reduction of Grey values to compensate technical bias from the Grey values in the scan, but not losing important information which support predictive performance. Two major methods are applied in this regard depending on the influencing binning parameter of the Grey value discretization (number of bins or size/width of bins). The IBSI recommends pixel intensity discretizations prior to not all feature classes (e.g. local intensity features are based on non-discretizised images) which impacts features directly [119].

The configuration of the discretization process (fixed number of bins or fixed bin width) can therefore be fundamentally impact about 83 % of the IBSI features describing the textures of the tumor using the descritizied images shown in Table 2.2 and 2.1. In contrast, the most commonly used tool, PyRadiomics, implements pixel intensity discretization for all feature classes besides the shape features but is also missing about 37% of IBSI features [15, 24, 119]. This divergence not only affects feature values but also downstream model performance, underscoring the need to document and, where possible, harmonize discretization methods in any radiomics study. My recent work demonstrated that applying different extractors (like PyRadiomics [15] and MIRP [21]) with default configuration parameters can yield superior outcome

prediction compared to state-of-the-art pipelines that rely on a single toolkit [24]. Non-IBSI features include statistical features describing the dataset can enrich phenotypic characterization but suffer from inconsistent definitions and limited validation across sites.

Furthermore, the configuration of the extraction process should also be verified to consider on how to perform radiomics extraction and further processing on multiple ROIs per patients. Radiomics studies routinely extract hundreds to thousands of features per lesion, often exceeding the number of available patient samples. This high dimensionality greatly increases the risk of overfitting and degrades model generalization. [135, 136].

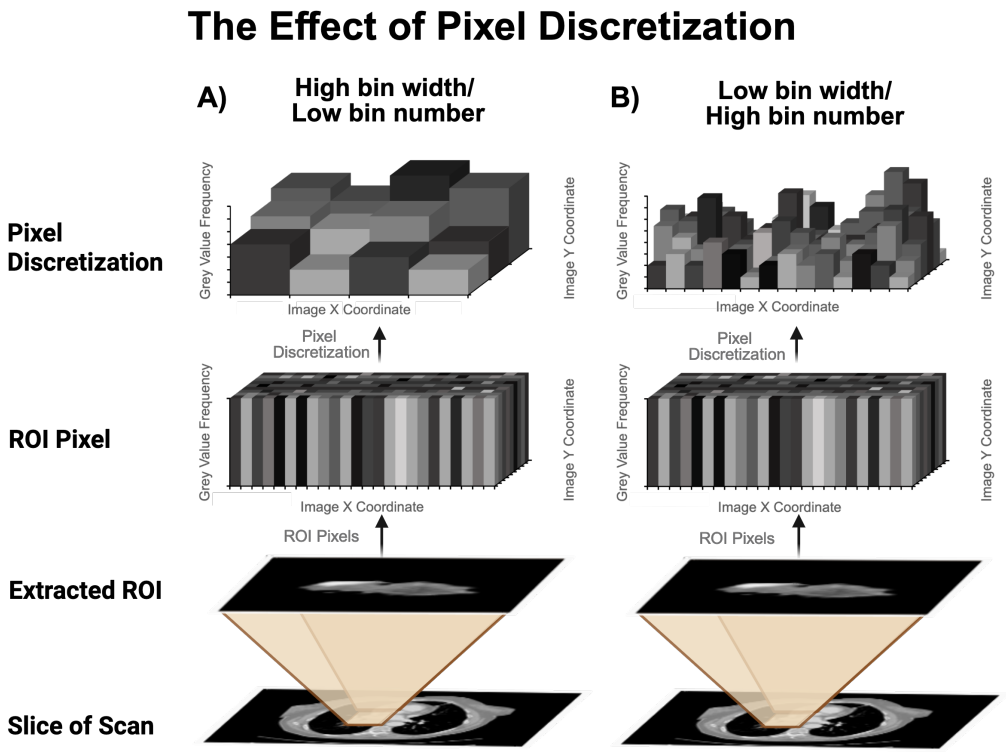


Figure 2.4. Impact of Pixel discretization on noise and resolution illustrated for A) high bin width or low bin number and B) low bin width and high bin number. Visualizing the tradeoff between losing important information (A) and not including imaging bias (B) from the ROI in the image.

Table 2.1. IBSI applied pixel discretization with Fixed Bin Size (FBS) on ROI extracted from CT and MR images to show the impact of the configuration of FBS discretization for both modalities. The CT Image comes from the CRLM dataset and the MR image comes from the Desmoid dataset.

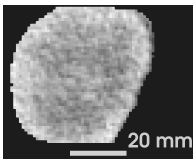
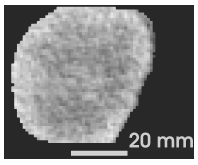
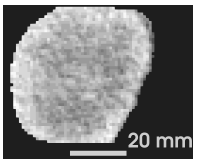
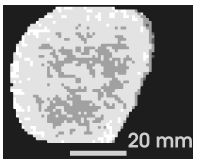
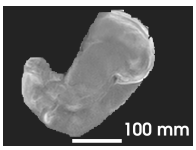
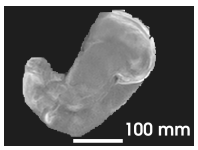
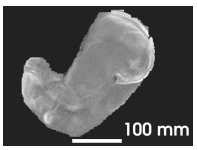
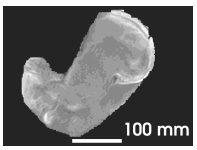
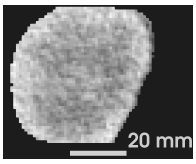
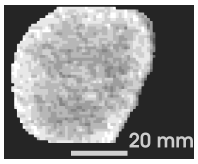
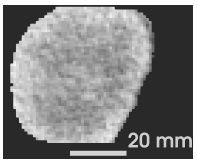
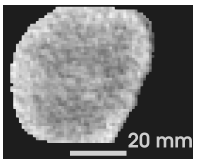
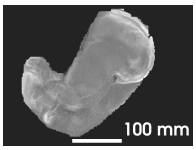
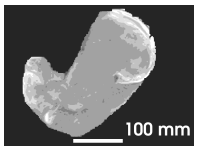
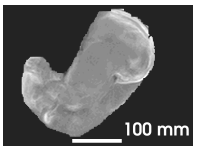
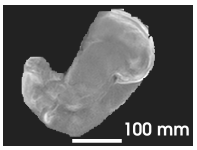
	Original	**Bin width: 10**	**Bin width: 25**	**Bin width: 100**
CT Image				
MR Image				

Table 2.2. IBSI applied pixel discretization with Fixed Bin Number (FBN) on ROI extracted from CT and MR images to show the impact of the configuration of FBN discretization for both modalities. The CT Image comes from the CRLM dataset and the MR image comes from the Desmoid dataset.

	Original	**Bin number: 10**	**Bin number: 25**	**Bin number: 100**
CT Image				
MR Image				

Feature Filtering

Before any model is trained, statistical filters remove features unlikely to contribute useful signal:

- **Low-variance removal:** Exclude features with near-zero variance across the cohort.

- **Correlation thresholding:** For any pair of features with Pearson $|r| > 0.9$, drop one to mitigate multicollinearity.

- **Stability assessment:** Retain only features with high intraclass correlation coefficients (ICC) under segmentation perturbations or phantom studies.

These steps are model-agnostic and efficiently prune gross redundancies and noise before machine learning models get applied [136].

Feature Selection

The feature selection step is a very critical step in the radiomics workflow, as it reduces the feature space very much to get only most informative radiomics features but also losing information which might harm the performance in the end. The feature selection is an optimization process and should therefore define a training and testing data. The same train and test data sets from the feature selection should also be used in the model training to avoid potential information leakage between the train and test sets [137]. Subsequent selection methods leverage model performance to identify the most predictive subset [135]:

- **Wrapper methods:** Recursive Feature Elimination (RFE) or Sequential Feature Selection (SFS) iteratively fits a classifier (e.g. random forest or Support Vector Machine (SVM)) and removes the least important features.

- **Embedded methods:** Regularized algorithms (e.g. LASSO) incorporate penalty terms during training to shrink irrelevant feature coefficients to zero.

- **Hybrid approaches:** Combine filtering and wrapper stages to balance computational cost with predictive accuracy.

By integrating feature importance into the selection process, these techniques yield parsimonious feature sets that improve prediction and reduce overfitting [135].

These steps for feature reduction are essential and very important as they define the information where the predictions are based on and which parameters are important for the specific clinical task. Each method comes with specific limitations. These processes can trigger model overfitting as well as underfitting and should be defined also in regard to the data size (not more features as samples) to overcome these problems and not overrate potential important features [136].

2.3.7 Machine Learning (Modeling)

Within the radiomics workflow, machine learning (ML) plays a central role in transforming extracted imaging features into clinically meaningful predictions. After careful preprocessing and feature selection, the resulting radiomic signature is modeled to capture associations with biological characteristics, treatment outcomes, or clinical endpoints. ML methods are particularly well suited for radiomics because of the typically high-dimensional but limited sample size settings, where conventional statistical approaches often fail to provide robust generalization [136, 138]. By systematically learning patterns from the data, ML enables both classification tasks (e.g. tumor subtyping, mutation status prediction) and regression tasks (e.g. survival estimation, biomarker quantification) [139, 140].

The design of ML experiments in radiomics must carefully balance model complexity with the risk of overfitting and should integrate strategies for rigorous validation. Overfitting occurs when models adapt too closely to training data, leading to over interpretation of data specific technical bias (imaging parameters like convolution kernel or contrast agent) and poor generalization to unseen and external data (e.g. data acquired from a different institute or hospital as the training data) [141]. This challenge is particularly pronounced in radiomics due to the imbalance between feature dimensionality and cohort size. Experimental design therefore typically involves splitting data into training, validation, and test sets, or applying cross-validation to ensure that performance estimates reflect true generalizability [142]. Importantly, ML is not only a predictive tool but also a methodological framework that structures how models are trained, tuned, and evaluated in a reproducible way [143, 144]. Properly designed modeling therefore provides the bridge between abstract radiomic features and clinically actionable insights, ensuring that reported models are both technically sound and clinically interpretable.

Subsequent sections detail the two main pillars of this process: the choice and training of models, and the strategies for their evaluation.

Models and Training

The training of predictive models in radiomics must follow a carefully designed experimental setup to ensure valid and reproducible results. Importantly, models must be trained and validated using the same data partitions that were defined during feature selection, so that all steps of the pipeline are consistently applied on identical train–test splits [137]. This consistency guarantees that performance estimates are not biased by differences in data usage.

Cross-validation is one of the most widely applied strategies to maximize data efficiency and obtain robust estimates of generalizability. In k-fold cross-validation, the training data are split into k subsets (folds), and the model is trained iteratively on $k-1$ folds while being validated on the remaining fold. Variants such as stratified k-fold (maintaining class proportions), shuffle-split, or repeated random subsampling introduce different trade-offs between bias, variance, and computational cost [142]. Nested cross-validation is a more rigorous extension, designed to avoid information leakage during hyperparameter optimization. In this approach, an outer loop is used to split the data into training and test folds, while an independent inner loop is applied within each training fold to tune hyperparameters and select models. This design ensures that the test data in the outer loop remain completely unseen during model selection and optimization, thereby providing an unbiased estimate of the true generalization error [142, 143]. Nested cross-validation further reduces the risk of information leakage when hyperparameters are optimized, providing a less biased estimate of the true model performance [143]. However, recent studies have cautioned that nested CV is not always the optimal choice for small sample sizes [145–148]. The repeated partitioning of limited data can result in unstable estimates, inflated variance, and poor hyperparameter tuning due to the small number of training examples available in inner folds [145–148]. In such scenarios, repeated stratified k-fold cross-validation or carefully designed resampling (e.g., bootstrapping) may offer more stable and interpretable performance estimates than single split validation, especially in small datasets [149, 150].

Hyperparameter optimization has a major impact on model performance. The selection of hyperparameters, the definition of their search ranges, and the number of iterations tested all influence final results. Classical optimization strategies such as grid search and random search are widely used; however, more advanced techniques like Bayesian optimization or the Tree-structured Parzen Estimator (TPE) often allow more efficient exploration of large hyperparameter spaces, particularly when many hyperparameters exist, or when computational cost must be constrained [151–153]. The chosen optimization procedure must be integrated within the cross-validation loop to avoid overfitting, as tuning hyperparameters outside of the validation folds can lead to optimistic bias [141].

The selection of train/test splits is particularly critical for small datasets ($n < 100$). In such scenarios, a single patient included or excluded from the test set may substantially change the measured performance, underlining the importance of repeated cross-validation or resampling to achieve stable estimates [142]. To ensure reproducibility, randomness in all steps of training (e.g., fold assignment, initialization of

optimization routines) should be controlled, and identical data splits should be reused for all models under comparison.

Different cross-validation strategies can be employed depending on the data characteristics. Standard k-fold splits respect the natural order of the data, while shuffle splits randomize sample allocation, and repeated random splits allow samples to appear in different validation sets across iterations. Illustrations of these schemes are often provided to clarify their differences and to justify the chosen validation design.

Overfitting remains one of the most important risks in radiomics modeling. It occurs when the model adapts too closely to the peculiarities of the training or validation data, thereby losing the ability to generalize to unseen data [141]. This problem is exacerbated in situations with imbalanced class distributions, where the model may achieve deceptively high accuracy by preferentially predicting the majority class. Mitigation strategies include careful resampling, oversampling methods such as Synthetic Minority Over-sampling Technique (SMOTE), or penalization techniques that adjust decision thresholds for minority classes.

Finally, ensemble methods can further improve robustness and reduce variance by aggregating predictions from multiple models. In hard voting ensembles, the majority prediction across models determines the output, whereas in soft ensembles, class probabilities are averaged or combined through softmax operations, often leading to improved calibration and smoother decision boundaries. By combining complementary learners, ensemble strategies help counteract instability caused by small data sizes and heterogeneous imaging cohorts [143].

Evaluation

The evaluation of predictive models is a critical step in the radiomics workflow and determines whether a developed model is likely to generalize to unseen data and provide clinically useful information. Model evaluation is closely linked to the study design and must be performed with methods that are statistically rigorous and clinically meaningful. Recently, initiatives such as *Metrics Reloaded* have provided comprehensive recommendations for selecting problem-aware metrics and reporting them consistently in image analysis studies [144]. These guidelines stress that no single metric is universally appropriate, but instead the evaluation strategy must be aligned with the clinical use case and the characteristics of the prediction task.

In practice, statistical methods such as permutation testing or bootstrapping are commonly applied to assess whether model performance exceeds what would be expected by chance, thereby providing statistical significance testing of the results [154]. Beyond significance testing, it is increasingly recognized that clinically relevant eval-

uation goes beyond global metrics such as accuracy or Area Under the Receiver Operating Characteristic curve (AUROC). For threshold-based clinical decision making, sensitivity (true positive rate) and specificity (true negative rate) are of particular importance because they directly reflect the trade-off between detecting disease and avoiding false alarms. Several studies highlight that reporting these values together with 95% confidence intervals is crucial for assessing the robustness and reproducibility of model predictions in clinical contexts [154, 155].

Optimizing threshold-based metrics often involves the use of the Youden Index, which defines the threshold that maximizes the sum of sensitivity and specificity. The Youden Index has been widely adopted in biomedical research as a principled way of identifying clinically optimal decision thresholds [156, 157]. In radiomics, it provides an interpretable and reproducible strategy for threshold selection when the balance between sensitivity and specificity is critical.

Finally, evaluation practices differ somewhat between radiomics and deep learning. Radiomics studies typically emphasize a broad set of metrics (e.g., AUROC, sensitivity, specificity, concordance index) to characterize performance across tasks such as classification and survival prediction. Deep learning studies, in contrast, often prioritize AUROC or accuracy as headline results, although calibration curves, decision-curve analysis, and external validation are increasingly recommended for both paradigms to assess clinical utility [144, 154]. Regardless of the modeling approach, careful selection and transparent reporting of evaluation metrics are essential to ensure that models can be meaningfully compared and eventually integrated into clinical workflows.

2.3.8 Clinical Integration & Reporting

The final phase of the radiomics workflow involves the integration of radiomics signatures with clinical and pathological data, aiming to provide comprehensive decision support and risk stratification. This integration seeks to bridge the translational gap between promising computational biomarkers and their adoption in routine clinical practice [158, 159].

Reliable clinical integration necessitates robust validation, ideally via prospective trials, multi-center cohorts, and reader studies, which collectively help ascertain generalizability and real-world efficacy [158, 159]. Integrating radiomics into prospective clinical application enables standardized processing of routine clinical care data for complex and comprehensive analyses with information from multiple domains (see Figure 2.5). Before a radiomics workflow can be integrated for a specific clinical task, it must first be rigorously validated for that purpose to identify the most rele-

vant information, distinguish true signal from bias, and enhance both the quality and quantity of data processing. Selected radiomics features and trained models can be used after validation for the prospective use case but might need constant reevaluation as the clinical setting develops and data quality and data size change over time. Moreover, strong emphasis is placed on harmonization of image acquisition, standardized analysis pipelines, and transparent model sharing, all of which accelerate regulatory acceptance and clinical translation [12, 160].

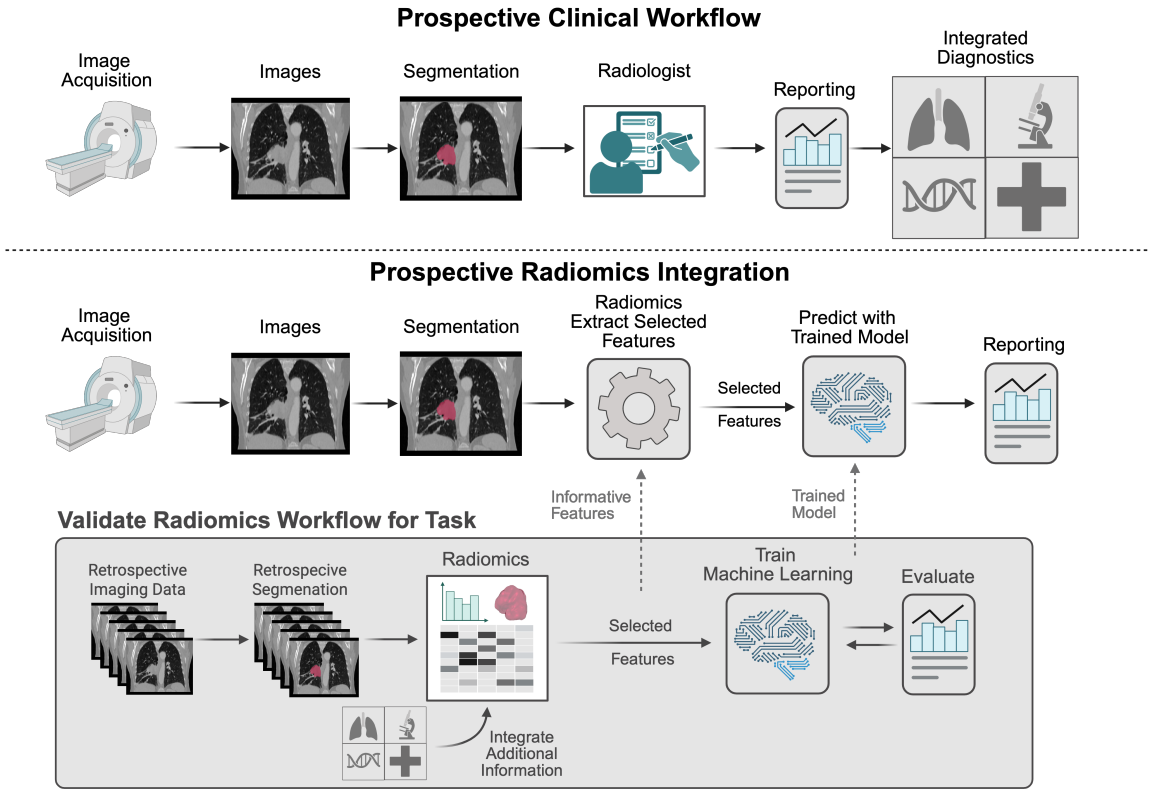


Figure 2.5. Integration of radiomics in a prospective clinical setting to handle live data processing. The standard clinical workflow includes the radiologist, who evaluates the tumor based on the radiological images and sub sequential integrates the findings into the clinical reportings. The radiomics pipeline needs to be validated on retrospective data to select informative radiomics features and integrate additional clinical parameters for the subsequent machine learning training and validation. The optimized radiomics workflow can be integrated into the prospective radiomics workflow for fast and precise data processing.

Reporting radiomics studies according to comprehensive guidelines such as the RQS, TRIPOD, or CLEAR (see section 2.3.1) is currently considered best practice [12,160]. Adherence to these frameworks improves reporting quality, facilitates critical appraisal, and expedites integration into clinical workflows and regulatory pathways

[12, 161].

Recent systematic reviews, however, highlight ongoing challenges in reporting consistency and methodological rigor, underscoring the importance of continued community adoption of such guidelines for radiomics to have a measurable impact at the bedside [120, 159]. However, the integration of AI tailored system for decision making or decision support in the clinical practice remains doubts about who is responsible if the integrated model does not give the correct suggestion and could harm the patient [162]. Therefore, the clinician who applies these models need to stay responsible and check critically any outcome from these models where the models need to be understandable for the users, which is true for radiomics features and models trained on these features (see Section 2.3).

2.3.9 Radiomics Limitations & Future Directions

Radiomics offers considerable promise for the extraction of high-dimensional quantitative features from medical images, enabling non-invasive assessment of tissue heterogeneity and supporting clinical decision-making. However, substantial limitations remain, many of which restrict its current clinical application.

First, radiomics features are fundamentally phenotype-descriptive and do not replace molecular measurements for biological characteristics that are not expressed phenotypically. Therefore, radiomics alone cannot capture all relevant (molecular or cellular) tumor or tissue biomarker [161, 163].

Second, the majority of radiomic features, while numerically standardized by initiatives such as IBSI [119], are not always directly interpretable or descriptive for the diverse range of clinical and research problems in which they are applied. Their clinical significance may thus vary, and many features remain mathematically abstract rather than biologically meaningful [163, 164].

Third, the defined set of features were defined through incremental attempts to capture ever more known phenotypic characteristics but may leak some important information and thus could not be seen as the final set of features to capture all important information as the definition of important information remains a challenge itself and might be different for every clinical application. There is a lack of systematic assessment regarding the added value or redundancy of many features within different imaging and clinical contexts [164]. The application and selection of different imaging filters like Wavelet, Laplacian of Gaussian (LoG), Square, or Logarithm are not standardized [10]. The feature extraction on these filtered images multiply the feature space [138] and could support redundant and non-informative feature inclusion but also serve as an augmentation technique or highlight substantial information in the

image for more comprehensive and stable feature calculations [138].
Fourth, radiomics is highly sensitive to variations in imaging protocols, scanner parameters, and segmentation practices [138]. Even minor changes in image acquisition or reconstruction can dramatically affect measured feature values, and aggressive correction or normalization approaches, while intended to enhance reproducibility, may inadvertently obscure or eliminate meaningful biological signals [138, 161].
Best practice for feature calculation should further be investigated for reliable and reproducible radiomics studies. It should further be determined by consensus for each imaging modality, grounded in evaluations across large, heterogeneous datasets with a wide spectrum of pathologies, imaging settings, and research aims [161, 164].

Chapter 3

State of the Art

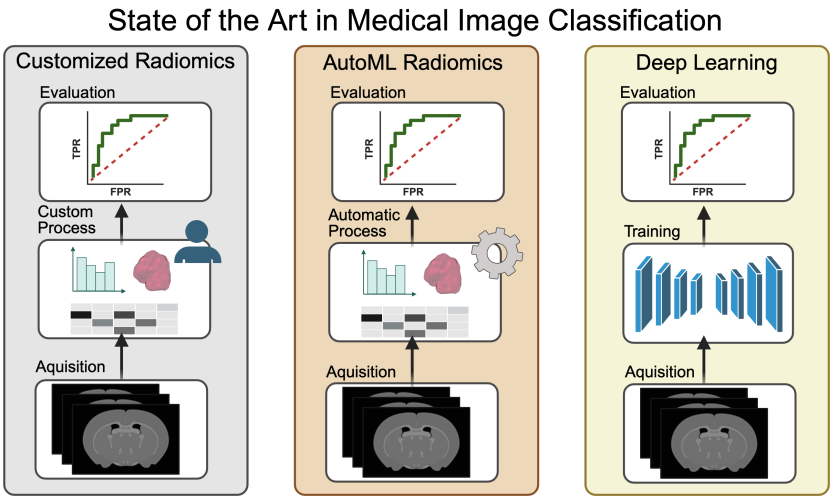


Figure 3.1. Domains of state of the art methods for 3D medical image classification include deep learning approaches as well as radiomics based approaches. Radiomics based approaches are either customized which is lacking in generalization whereas the automated radiomics applies a comprehensive optimization and aims to achieve generalizable and reproducible results.

This chapter gives an overview of the related state of the art methods for medical 3D image classification and the role of radiomics in this context. In general state of the art methods for medical 3D image classification can be clustered in 3 classes (see Figure 3.1): deep learning approaches, customized radiomics approaches, and AutoML Radiomics approaches. The selection of the method relays on several factors and requirements. For training deep learning models high amounts of data (1000+ samples) and GPU hardware support are required to get good performing models for the specific task [165]. In contrast, radiomics approaches can be applied on less data but sufficient data size is related to the task and the data heterogeneity [166].

Radiomics approaches have showed good performance on phenotypic related tasks like tumor subtype classification or tumor malignancy classification, but also for prognosis and treatment response prediction [1].

The domain of 3D medical image classification has been dominated by methods based on deep learning, in particular customized CNN models are dominating the literature but also more standardized model architectures like ResNet, Visual Geometry Group (VGG)-16 , EfficientNet and DenseNet have been very frequently applied to different medical image datasets for classification [167]. These networks are capable of exploiting spatial context in all three dimensions and have consistently outperformed slice-based or customized approaches when sufficient annotated data are available [165, 167].

At the same time, radiomics remain widely applied, especially in settings with limited data, heterogeneous acquisition protocols, or when interpretability is prioritized. The landmark work of Aerts et al. introduced the concept of the radiomic signature, demonstrating prognostic value in lung and head-and-neck cancer and establishing radiomics as a methodology for quantitative imaging research [113]. Customized radiomics approaches process radiomics features for a specific task on a specific dataset which can be seen in several studies like [168–170]. In such pipelines, quantitative features describing tumor shape, texture, or intensity are extracted and modeled with classical machine learning classifiers such as support vector machines or random forests solving this specific task mostly very accurate [171, 172].

Building customized designs is very labor intensive and fails for generalization on other datasets, more recent frameworks automate preprocessing, feature selection, model training, and evaluation to improve reproducibility and generalizability. AutoML-based pipelines such as AutoRadiomics [18] and WORC [19] have been applied successfully across multiple heterogeneous datasets. These automated radiomics approaches are increasingly considered the state of the art within the AutoML Radiomics paradigm but remain underperformed customized radiomics on some datasets [18].

In this chapter, Section 3.2 reviews radiomics methods, from customized feature pipelines to automated frameworks, with particular focus on reproducibility and generalizability. Section 3.1 surveys deep learning models for 3D classification, including CNN-based architectures and recent extensions. Together, these subsections provide the context against which the proposed RPTK framework is positioned.

3.1 Deep Learning

Deep learning (DL) has become a cornerstone of medical imaging, consistently surpassing traditional machine-learning pipelines based on handcrafted features and classifiers such as Random Forest or XGBoost [61,64,173]. Core application areas include image segmentation, with U-Net and its derivatives as the canonical architectures [61]; image generation and phantom/synthetic data creation, where diffusion models have recently demonstrated strong performance for augmentation, reconstruction, and simulation [174–177]; image classification, where residual and densely connected convolutional networks are widely applied in clinical tasks [178,179]; and image regression, in which CNNs predict continuous outcomes such as biological age or mortality risk directly from radiographs [180–183].

Residual and densely connected CNNs have become standard backbones in medical imaging reviews and benchmarks [61,173]. Comparative analyses in chest radiography report competitive performance across variants of ResNet and DenseNet, with smaller models often offering favorable accuracy–efficiency trade-offs [178,179]. In our work, multiple configurations (ResNet-18/200; DenseNet-121/169/201/264) were assessed to characterize this balance empirically.

Segmentation remains the most mature area, where encoder–decoder CNNs underpin radiological and histopathological pipelines [61]. In parallel, diffusion models have emerged as generative priors for medical imaging. Surveys demonstrate their applications in modality translation, denoising, and synthetic CT generation [174–177], which in turn facilitate phantom data creation for algorithm development.

Deep Learning (DL) regression models extend beyond classification by estimating continuous variables directly from images. Chest X-ray–based age estimation and mortality risk prediction exemplify how image-derived quantitative phenotypes can provide clinically actionable biomarkers [180–182].

All deep learning experiments in this work relied on the MONAI framework [23], an open-source PyTorch-based ecosystem specifically designed for healthcare imaging. MONAI provides standardized data loaders, preprocessing pipelines, and reference implementations of widely used architectures such as ResNet and DenseNet, which were adopted in our classification experiments. In addition, MONAI includes modules for model evaluation, reproducibility, and deployment, as well as extensions such as MONAI Label for interactive annotation and integration into clinical workflows [184]. The use of MONAI ensures that our implementation adheres to community standards and facilitates reproducibility and future extension of our work.

Medical imaging is undergoing a paradigm shift toward foundation models trained

with large-scale self-supervision or multimodal objectives. For example, self-supervised chest X-ray models achieve radiologist-level pathology detection without manual labels [185], and RETFound leverages 1.6 million unlabeled retinal images for generalizable ocular disease detection [186]. Vision–language models trained on millions of biomedical image–text pairs, such as BiomedCLIP, further expand adaptation capabilities [187]. Domain-specific pretraining sets like RadImageNet highlight the value of radiology-focused initialization for downstream tasks [188].

3.2 Radiomics

It is difficult to define the state of the art method for radiomics as the field is very broad and there are no benchmarks with sufficient sample size to systematically evaluate the state-of-the-art radiomics tool [115]. It is very common in radiomics that pipelines for radiomics processing get created and designed for a single study on a specific research question and therefore are not generalizable to be applied to other clinical tasks (customized/ahdncrafted approaches). Anyhow, according to the literature we can see commonly used techniques and tools in the field of radiomics applied in radio-oncology like PyRadiomics for feature extraction [19]. Designing frameworks for optimizing radiomics workflows to be generalizable and adaptable to different clinical tasks require extensive evaluations on diverse datasets. In the literature, two notable frameworks for optimizing the radiomics workflow AutoRadiomics [18] and WORC [19] aim to improve predictive performance and have been evaluated on diverse datasets.

3.2.1 Customized Radiomics

Customized radiomics approaches refer to pipelines that are specifically designed and tuned for a single clinical question or dataset. In these studies, researchers typically construct an end-to-end workflow comprising image preprocessing, feature extraction, feature selection, and machine learning classification. The design is usually highly task-specific, including human knowledge on feature extraction (extracting only predefined feature classes) or feature selection (select features which are hypothesized to be very relevant), reflecting the clinical endpoint and modality at hand, but as a result such pipelines often lack generalizability to external datasets or other clinical problems.

Several clinical use cases illustrate the potential of customized radiomics. For example, Aerts et al. demonstrated that handcrafted quantitative descriptors of tumor shape, intensity, and texture could be combined into a radiomic signature predictive of

patient survival in lung and head-and-neck cancer, establishing radiomics as a quantitative imaging biomarker paradigm [113]. Since then, numerous single-application studies have appeared, such as the differentiation of low- from high-grade gliomas on MRI [169], prediction of treatment response in non-small cell lung cancer [168], or gastric cancer characterization based on CT imaging [170]. These studies typically employ conventional classifiers such as support vector machines, random forests, or gradient boosting, trained on a tailored feature set chosen for the specific imaging task [171, 172].

A notable limitation of many customized pipelines is the lack of standardized feature extraction. Prior to the introduction of the Image Biomarker Standardisation Initiative (IBSI) guidelines [115, 119], radiomics features were often computed with in-house or non-standardized software, raising concerns about reproducibility across institutions. Even widely used tools such as PyRadiomics, while representing the de facto standard for feature extraction, do not yet cover the complete IBSI feature set [119]. Other packages like MIRP address a more complete implementation of the IBSI-defined features, but require integration into custom pipelines. Consequently, differences in feature definitions, discretization schemes, or image filters across studies complicate direct comparison of results and limit clinical translation.

Despite these challenges, customized radiomics studies have played a crucial role in demonstrating the feasibility and clinical potential of radiomics across diverse applications. They provide evidence that carefully engineered features, coupled with classical machine learning models, can yield accurate predictions for highly specific tasks. However, their bespoke nature highlights the need for more standardized and generalizable frameworks, which motivates the development of AutoML radiomics approaches discussed in the subsequent section.

3.2.2 WORC

The WORC tool optimizes radiomics workflows aiming to replace the common labor intensive trial-and-error construction of radiomics pipelines—where researchers manually choose among many preprocessing, feature extraction, selection, and learning options with a reproducible, application-specific AutoML procedure that jointly selects algorithms and hyperparameters across the full workflow and validates this end-to-end strategy across diverse clinical problems [19]. WORC formalizes the construction of a radiomics workflow as a Combined Algorithm Selection and Hyperparameter (CASH) problem, aiming to (i) reduce researcher degrees of freedom and overfitting to validation quirks, (ii) standardize and automate per-dataset choices (including modality-aware defaults) to improve generalizability, and (iii) provide open

software + released data so that results are auditable and repeatable rather than bespoke to a single study [19]. The WORC tool tries to streamline and de-bias biomarker discovery by automating the search for an optimal radiomics workflow, demonstrating competitive or superior performance to handcrafted baselines and Bayesian optimizers while emphasizing reproducibility [19].

WORC optimizes a weighted F1 objective on the training folds, sampling complete pipelines via random search and then forming a simple top-N ensemble by averaging posterior probabilities of the best pipelines; this design was benchmarked against Sequential Model-based Algorithm Configuration (SMAC) Bayesian optimization [189] and more elaborate ensembling, with top-N chosen for robustness on test sets [19]. Concretely, model selection and tuning occur inside a k_{training}=5 random-split CV on the training data; final performance is reported from repeated random-split evaluation (typically 80/20, k_{test}=100) or bootstrap on a held-out test set [19].

The search space is modular and standardized. A light fingerprinting step uses prior knowledge to (i) enable z-score normalization only for qualitative modalities (e.g., T1-w MRI, US) but not for quantitative ones (e.g., CT, T1-mapping), (ii) choose fixed bin count (qualitative) versus fixed bin width (quantitative) discretization for texture computation, (iii) decide between 2D/2.5D/3D features from voxel spacing and slice thickness, and (iv) omit resampling when classes are near-balanced (<=60/40) [19].

For feature extraction, WORC uses PyRadiomics [15] and PREDICT [190] in combination to compute 564 radiomics features for each sample. Scale-dependent filters (e.g., LoG, Gabor) are precomputed over predefined ranges; the downstream selection stages decide which instances survive [15].

WORC processes features in a fixed, "safe" order. Optional steps are controlled by a simple on/off switch (an activator) that is tuned during optimization. The sequence is: (i) a group-wise filter that can keep or drop entire families (intensity, shape, texture) and a small variance threshold filter (threshold ¡ 0.01; both always applied); (ii) robust z-scaling; (iii) optionally RELIEF; (iv) optionally model-based selection (LASSO, logistic regression, or random forest as selectors); (v) optionally Principal Component Analysis (PCA) (retain 95% variance or a fixed number of components variance or $n \in 10, 50, 100$; (vi) optionally a univariate Mann–Whitney U filter with a tunable p-value cut-off; and (vii) class-imbalance handling via resampling methods (e.g., random under/over-sampling, SMOTE, Adaptive Synthetic Sampling (ADASYN)), with the method and its settings included in the search. During WORC's default random search, each optional selector (RELIEF, model-based, PCA, univariate) is included with probability 0.275, which encourages diversity without letting combinations explode [15].

The classifier is chosen jointly with the selection settings from a curated set: logistic regression, SVM (linear/polynomial/RBF), random forest, Gaussian naive Bayes, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), AdaBoost, and XGBoost. For each, WORC tunes the standard hyperparameters (e.g., SVM kernel; random-forest trees and depth; logistic-regression penalty; AdaBoost estimators and learning rate; XGBoost rounds, depth, learning rate), alongside the on/off activators of the upstream steps so that the entire pipeline is optimized as one unit [15].

On the six public datasets released with the WORC study [7, 15], AutoRadiomics [18] reported higher test AUROCs on most of these datasets, indicating that WORC is not the top performer on most of these datasets. More broadly, WORC's own multi-application evaluation shows substantial variability (AUROCs from 0.45 to 0.87 across 12 tasks), underscoring that performance is highly dataset-dependent rather than uniformly strong [15].

3.2.3 AutoRadiomics

The AutoRadiomics framework was developed as a modular, intuitive, and reproducible AutoML platform to address persistent challenges in radiomics research, particularly the lack of standardization and reproducibility across studies [18]. Built on top of PyRadiomics and scikit-learn [15, 191], AutoRadiomics automates all essential steps of a radiomics workflow, including feature preprocessing, feature selection, model training, and evaluation, within a single end-to-end pipeline. By design, the framework targets usability for both technical researchers and clinicians, offering a streamlined web interface for non-programmers while ensuring reproducibility and transparency.

AutoRadiomics formalizes the pipeline optimization problem as a categorical hyperparameter search, embedding multiple workflow components directly into the optimization space. Specifically, preprocessing employs Min–Max scaling of radiomics features, while optional data balancing methods (SMOTE or ADASYN) are included as tunable categorical variables. Feature selection is performed prior to model training using one of several supported algorithms (Analysis of Variance (ANOVA), Least Absolute Shrinkage and Selection Operator (LASSO), or Boruta), with the choice of selector added to the search space. Classifier selection is similarly treated as a categorical hyperparameter, drawn from a curated set of models (logistic regression, support vector machines, random forest, and Extreme Gradient Boost (XGBoost)). The joint space of preprocessing, feature selection, oversampling, and classifier hyperparameters is explored using Optuna's Tree-structured Parzen Estimator (TPE) algorithm [153],

with up to 200 optimization iterations per run. Each candidate pipeline is evaluated using stratified 5-fold cross-validation, and final model performance is assessed with 95% confidence intervals estimated from 1000 bootstrap replicates of the held-out test set [18]. At present, AutoRadiomics is restricted to binary classification tasks, although it has been benchmarked across diverse clinical applications.

In contrast to WORC, which tackles a broader and more complex CASH formulation with modality-aware defaults, extensive fingerprinting, and a large search space spanning many optional selectors, AutoRadiomics deliberately simplifies the workflow into a smaller but more structured optimization problem. Rather than employing random search with on/off activators for optional steps, AutoRadiomics exposes key categorical design choices directly to the optimizer, which increases the degrees of freedom in a controlled manner. This design yields a more accessible and reproducible pipeline while reducing the barrier for non-expert users. Furthermore, AutoRadiomics integrates a voxel-based feature map generator using PyRadiomics, enabling voxel-wise feature extraction for improved interpretability in the imaging context [15, 18].

Empirical results show that AutoRadiomics achieves superior performance to WORC on most benchmark datasets. On the six public datasets included in the WORC release, the AutoRadiomics study reported higher test AUROCs in the majority of tasks, demonstrating that despite its more compact search space, AutoRadiomics is more effective across typical radiomics applications [18]. This suggests that a leaner and modular optimization strategy, coupled with strong defaults and a transparent implementation, may be preferable to a broader but more complex search. While AutoRadiomics is currently limited to binary tasks and applies PyRadiomics for feature extraction which does not extract the entire IBSI feature space [119], it represents the current state of the art in reproducible and accessible AutoML for radiomics classification.

3.3 Conclusion and Outlook

This chapter surveyed contemporary methods for 3D medical image classification across three paradigms: deep learning, customized radiomics, and AutoML radiomics. Deep learning models—particularly residual and densely connected CNNs—have become the default choice when large, well-annotated datasets and suitable compute are available, delivering strong performance by leveraging end-to-end representation learning. At the same time, radiomics remains widely used in settings with limited data, heterogeneous acquisition, or heightened interpretability requirements. Customized (handcrafted) pipelines can achieve high task-specific accuracy but are time-

consuming to design and difficult to generalize beyond the originating study. To address these issues, AutoML radiomics frameworks have emerged, standardizing preprocessing, feature selection, model training, and evaluation in reproducible, end-to-end workflows.

Within AutoML radiomics, WORC and utoRadiomics exemplify two complementary design philosophies. WORC tackles a broad CASH formulation with modality-aware defaults and a rich search space spanning multiple optional selectors and classifiers, emphasizing standardization and robustness across diverse clinical problems. AutoRadiomics adopts a leaner, more structured optimization space (categorical choices for selector, oversampling, and classifier) explored via modern hyperparameter optimization, with an emphasis on usability and reproducibility. Empirically, AutoRadiomics reports higher test AUROC on most of the public datasets released with the WORC study, indicating that a compact yet well-curated search space can be competitive or superior in typical radiomics settings. Nonetheless, both approaches inherit limitations of handcrafted features and, in practice, show dataset-dependent performance.

Beyond the broad paradigms, it is important to note that radiomics tools vary substantially in scope. Specialized tools such as PyRadiomics and MIRP focus primarily on feature extraction, leaving users to design their own preprocessing and modeling frameworks. PyRadiomics in particular has become the de facto standard for radiomics feature extraction but remains incomplete with respect to the IBSI standard, currently omitting roughly one-third of defined features. Other tools provide end-to-end pipelines from acquisition to prediction, but these are often validated only on synthetic data or tailored to narrow use cases, limiting their generalizability. In contrast, AutoML frameworks such as WORC and AutoRadiomics aim for broader applicability by automating end-to-end workflow construction and validating across diverse datasets without dataset-specific tailoring.

Deep learning methods continue to dominate medical image classification, offering strong predictive performance when sufficient annotated data and computational resources are available. Their limitations—high GPU demands, large data requirements, and restricted interpretability—have spurred increasing use of transfer learning and fine-tuning, which allow pretrained networks to be adapted to smaller datasets. While this mitigates overfitting, challenges remain in achieving interpretability and in matching radiomics approaches on small, heterogeneous cohorts. Hybrid pipelines combining handcrafted radiomics features with deep learning representations, as well as self-supervised and multimodal foundation models, represent promising directions to bridge the gap between interpretability, efficiency, and scalability.

Taken together, the literature suggests a set of persistent challenges that shape current practice:

- **Data scale and heterogeneity:** Deep models excel with abundant, consistent data; radiomics is more forgiving at smaller scales but sensitive to acquisition variability.
- **Generalizability and reproducibility:** Customized pipelines often overfit study-specific quirks; AutoML frameworks mitigate this but still show variable transfer across tasks and centers.
- **Search design and complexity:** Rich search spaces increase flexibility but complicate optimization and reproducibility; leaner spaces improve usability but may miss beneficial configurations.
- **Scope of tasks:** Many automated radiomics evaluations focus on binary classification and handcrafted descriptors, leaving multi-class and multimodal problems comparatively underexplored.
- **Evaluation rigor:** Fair, transparent comparisons (fixed data splits, nested Cross Validation (CV), calibrated uncertainty, and confidence intervals) remain essential yet inconsistently applied.

Overall, the state of the art demonstrates a spectrum of solutions, from specialized radiomics feature extractors to general-purpose AutoML frameworks and data-intensive deep learning approaches. Each class of methods offers strengths and weaknesses depending on data scale, task heterogeneity, and interpretability requirements. This landscape motivates the development of new frameworks that combine reproducibility and accessibility with broader feature completeness, flexible algorithmic scope, and transparent evaluation—goals pursued in the RPTK framework presented in the following chapter.

Chapter 4

Data & Methods

The Data and Methods chapter outlines the datasets, computational frameworks, and experimental setups that form the foundation of my thesis. The central objective is to advance radiomics analysis by developing and validating RPTK, a self-configuring framework designed to address challenges of data heterogeneity, pipeline variability, lack of reproducibility and performance in current radiomics workflows.

To this end, RPTK was applied across nine diverse datasets in total (see Table 4.1), comprising 3,116 patients, 2,685 CT and 504 MR scans, and 3,273 segmented three-dimensional regions of interest. From the total amount of CT scans, 1,997 originated from the large-scale LiverCRC project (see Section 4.3). In addition, two datasets (Desmoid and Lipo) contributed 318 T1-weighted MRI scans and one dataset (Liver) provided 186 T2-weighted MRI scans (see Table 4.1). Demographically, approximately 75% of patients were older than 50 years, while age information was unavailable for 115 patients (see Figure 8.1). Around 54% of patients were male, with sex information missing in about 4% (see Figure 8.3b). From a technical perspective, roughly 80% of the scans were acquired on Siemens scanners, complemented by data from Philips, GE Healthcare, and Toshiba systems (see Figure 8.2), and 84% of the imaging data consisted of CT (see Figure 8.3a).

Six of the nine datasets, including Colorectal Liver Metastasis (CRLM), Melanoma, Gastro-Intestinal Stroma Tumor (GIST), Desmoid, Liver, Lipo, and LIDC-IDRI, are open-source and were used for systematic adaptation and evaluation of the pipeline (see Section 4.1). The remaining two datasets demonstrate specific clinical applications of RPTK: the Predict dataset for immunotherapy response prediction in NSCLC patinets (see Section 4.2), and the large-scale LiverCRC project for colorectal cancer risk stratification from liver imaging (see Section 4.3).

Table 4.1. Overview of datasets used in this thesis and their respective classification tasks.

Dataset	Patients	Images	ROIs	Modality	Task
CRLM	77	77	93	CT	Classification of desmoplastic (n=37) against replacement growth pattern (n=40) in colorectal liver metastases [192]
Melanoma	103	103	169	CT	Classification of BRAF-mutated (n=51) against BRAF-wild type (n=52) in lung metastases of melanoma [193]
GIST	247	247	248	CT	Classification of gastrointestinal stromal tumor (GIST) (n=125) against intra-abdominal gastrointestinal tumors (n=121) [194]
Desmoid	203	203	203	T1w MR	Classification of desmoid-type fibromatosis (n=72) against extremely soft-tissue sarcoma (n=131) [195]
Liver	186	186	186	T2w MR	Classification of malignant (n=94) against benign (n=93) primary solid liver tumor [196]
Lipo	115	115	116	T1w MR	Classification of well-differentiated liposarcoma (n=58) against lipoma (n=58) [197]
LIDC-IDRI	115	115	115	CT	Classification of malignant (n= 79) against benignant (n=36) lung nodules [20]
Predict	73	146	146	CT	Longitudinal prediction of PD-L1 treatment response (n=38) against non-response (n=35) on primary NSCLC tumor
LiverCRC	1997	1997	1997	CT	Classification of patients with colorectal neoplasia (n= 808) or liver pathology against healthy colon (n= 1189) [22]

This chapter is organized as follows:
Section 4.1 introduces the RPTK framework, its self-configuring principles, and its evaluation across multiple datasets. Section 4.2 presents an clinical application of RPTK, the Predict Study for immunotherapy response prediction in advanced stage lung cancer patients. Section 4.3 details the second clinical application of RPTK, the LiverCRC study on colorectal cancer prediction using liver imaging. Each section highlights the datasets, preprocessing, radiomics calculation procedures, model training and optimization strategies, and evaluation procedures specific to the respective application. For all the following sections, we show the comparison of RPTK performance to State-Of-The-Art (SOTA) radiomics tool (AutoRadiomics) and deep learning models (DenseNet and ResNet).

4.1 Self-Configuring Radiomics Framework

The development of RPTK was guided by general challenges known in the community as well as encountered in the analysis of the WORC database [7, 11, 24]. The WORC database includes six publicly available datasets: CRLM, Melanoma, GIST, Desmoid, Liver, and Lipo [7]. These datasets revealed a variety of data-related issues that motivated the integration of adaptive preprocessing, feature extraction, and feature filtering strategies into the framework. After configuring the framework on the WORC datasets, RPTK was applied to the LIDC-IDRI dataset to assess its generalizability

beyond the initial development cohort [198].

The following subsections detail the methodological components, covering dataset and study design, preprocessing, feature extraction and filtering, feature selection, model training and optimization, and the comparative evaluation of RPTK against AutoRadiomics and deep learning baselines (DenseNet and ResNet).

4.1.1 Data & Study Design

The evaluation of the RPTK was based on seven open-source datasets. Six datasets from the WORC database (CRLM, Melanoma, GIST, Desmoid, Liver, and Lipo) were used to identify and address data-related challenges and to configure the framework. The LIDC-IDRI dataset was subsequently employed as an external test case to assess the generalizability of the configured pipeline.

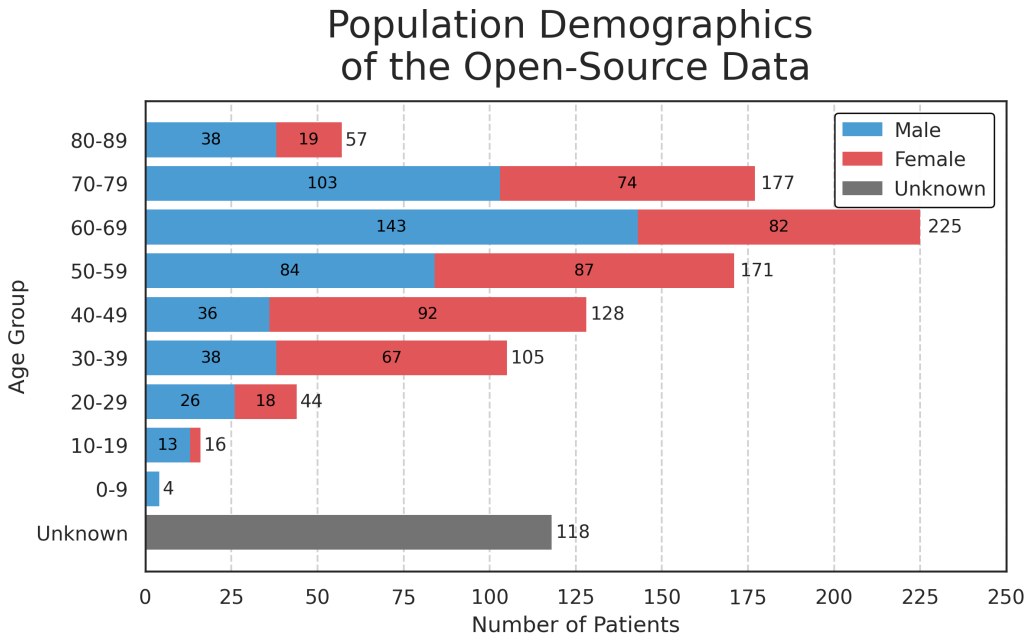
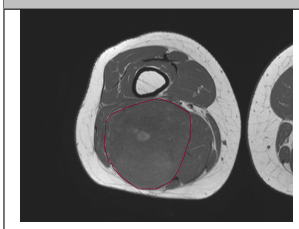
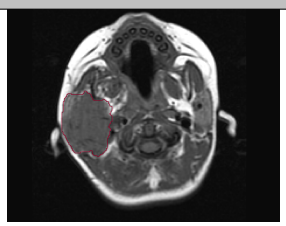
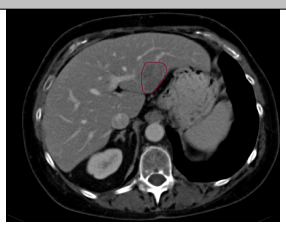
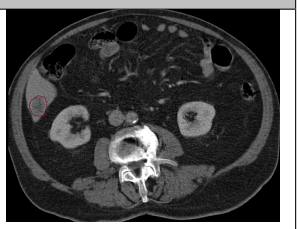
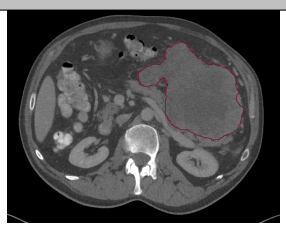
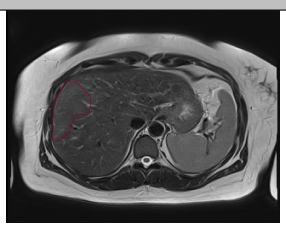
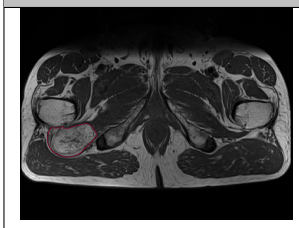
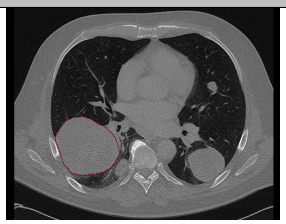
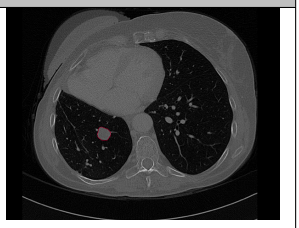
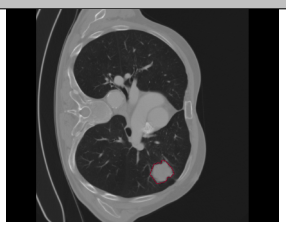


Figure 4.1. Age and Sex distribution of the seven datasets (WORC database + LIDC-IDRI) used in this section. Unknown refers to the patients either having no sex parameter or having no age parameter.

These datasets represent a broad spectrum of clinical indications, imaging modalities, and classification tasks (Table 4.1). Their heterogeneity is reflected not only in the sample size, which ranges from 77 patients in the CRLM dataset to 247 patients in the GIST dataset, but also in the mix of imaging modalities (CT, T1w MRI, T2w MRI) and segmentation entities (see Table 4.2). This diversity makes the WORC datasets particularly suited for methodological development, and they have

been widely used as benchmarks for AutoRadiomics, WORC, and other published radiomics approaches.

Table 4.2. Visual examples of the open-source datasets used in this study, illustrating the class labels image examples (Gastro-Intestinal Stroma Tumor (GIST), Desmoid Type Fibromatosis (DTF), B-Raf proto-oncogene serine/threonine kinase (BRAF), Well Differentiated Lipo-Sarcoma (WDLPS)).

Desmoid		**CRLM**	
			
Soft-tissue Sarcomas	DTF	Replacement growth pattern	Desmoplastic growth pattern
GIST		**Liver**	
			
non-GIST	GIST	Benign	Malignant
Lipo		**Melanoma**	
			
Lipoma	WDLPS	BRAF-wt	BRAF
LIDC-IDRI			
			
Benign	Malignant		

A summary of patient demographics across the seven datasets used in this section is provided in Figure 4.1. Approximately, 54% of patients were older than 50 years, within the annotated population of 927 patients, a close to equal sex distribution is recognizable (52% male, 48% female), although sex or age information was missing for about 14% of cases. Such demographic and imaging heterogeneity provided a realistic test bed for building a framework that can adapt to varying conditions without manual reconfiguration.

The WORC Data Collection

The WORC data collection was developed as an open-source resource to benchmark and compare radiomics pipelines in a reproducible manner [7]. It combines six radiomics studies from the Erasmus Medical Center, covering a total of 930 patients across CT and MR imaging. Unlike many radiomics datasets, WORC was specifically curated to reflect the heterogeneity of clinical practice and imaging was acquired across multiple centers on different scanners, with varying voxel sizes, protocols, and modalities (CT, T1w MRI, and T2w MRI).

Each dataset addresses a clinically relevant binary classification problem (see Table 4.2): growth pattern prediction in colorectal liver metastases (CRLM), BRAF mutation status (Wild Type (WT) vs BRAF mutated) in lung metastases of melanoma (Melanoma), differentiation of gastrointestinal stromal tumors (GIST), distinction between Desmoid Type Fibromatosis (DTF) and extremity soft-tissue sarcomas (Desmoid), diagnosis of benign versus malignant primary liver tumors (Liver), and discrimination of lipoma versus Well Differentiated Lipo-Sarcoma (WDLPS) (Lipo). Together, these tasks span mutation status prediction, tumor subtype classification, and benign–malignant differentiation, thereby covering a broad spectrum of oncological use cases. For the CRLM dataset multiple segmentation's from different raters were provided. we used the segmentation's from the radiologist with the highest experience as the ground truth segmentation.

All tumor segmentations were obtained semi-automatically and subsequently reviewed and corrected by trained observers under radiological supervision to ensure high-quality reference regions [7]. Ground truth labels were primarily derived from pathology or biopsy; for a subset of benign cases (e.g., focal nodular hyperplasia in the Liver dataset), diagnosis was based on radiologically typical appearances, reflecting common clinical practice [7].

The WORC datasets were accessed from the Health-RI XNAT repository*. Their

*WORC datasets available at `https://xnat.health-ri.nl/data/projects/worc` (accessed May 2023)

open availability, heterogeneous nature, and prior use in benchmark studies such as AutoRadiomics make them particularly suitable for the methodological development and evaluation of the RPTK.

The LIDC-IDRI Dataset

The Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) was used in order to validate performance of RPTK without any adaptation to the framework configuration. This dataset is one of the mostly used publicly available repositories for thoracic CT imaging, hosted by TCIA [198]. However, previous studies have shown that the malignancy annotations in LIDC-IDRI are largely based on subjective radiologist assessments rather than pathological confirmation, which introduces ambiguity and label noise [20, 199, 200]. Inter-observer variability and inconsistent ratings have repeatedly been reported, and in some cases radiologist-derived malignancy scores were found to deviate substantially from pathological ground truth.

The clinical diagnosis labels provided directly within the TCIA repository were used in this thesis instead of external malignancy scores [198]. These labels are based on explicit diagnostic procedures, including biopsy, surgical resection, radiological review of two-year stability, and evidence of progression or treatment response [201]. In the benign cases, the majority were confirmed by long-term radiological stability, with a smaller number validated by biopsy or resection. Malignant cases, both primary lung cancers and metastatic lesions, were predominantly confirmed histologically by biopsy or surgical resection, with some additional confirmation through progression or response during follow-up. This reliance on established clinical diagnostic methods ensures that the labels are less ambiguous and more clinically meaningful than the subjective radiologist malignancy scores. Although this restricted the dataset to a smaller subset, the use of unambiguous clinical diagnoses was preferred over extended radiologist-derived labels from external resources, thereby reducing label uncertainty and improving the reliability of the external validation for the RPTK. The malignancy annotations were available for a subset of 157 patients. From these, 17 patients were excluded because images or segmentations were missing or empty, and an additional 25 patients were excluded due to an unknown diagnosis of the lung nodules. This resulted in a final cohort of 115 patients.

Segmentations were processed using the E2MIP repository, which provides automated pipelines for the preprocessing and conversion of LIDC-IDRI images and masks [202]. The final patient cohort included CT scans acquired on scanners from three different manufacturers (GE Healthcare, Philips, and Siemens). Based on the TCIA clinical annotations, 79 cases labeled as “malignant primary lung cancer” or

"malignant metastatic" were classified as malignant, while 36 cases labeled as "benign or non-malignant disease" were classified as benign.

The LIDC-IDRI dataset was accessed via TCIA[†]. Its combination of multi-vendor acquisition, curated clinical diagnoses, and detailed radiological annotations provides a heterogeneous external test set to evaluate the generalizability of the RPTK.

4.1.2 The first RPTK Prototype

This part is based on the development and application of the first RPTK prototype on the WORC datasets at the Medical Image Computing and Computer Assisted Intervention (MICCAI) Conference and has been published in the Conference Proceedings 2023 at Springer Nature [24]. The methodological design of RPTK evolved from this initial prototype to the extended version of RPTK presented in this thesis. The initial implementation was conceived as a simplified and robust framework to investigate the influence of missing IBSI-defined features in the widely used PyRadiomics library [15] discovered and evaluated by [119]. For this purpose, I compared the predictive performance of features extracted with PyRadiomics against those obtained from MIRP, which implements the complete IBSI feature space [21]. Both feature extractors were applied with their full range of image transformations to avoid constraining the potential feature space and to ensure that all relevant sources of image information were included.

The feature extraction outputs were subjected to standardized preprocessing, statistical filtering, and feature selection before training and optimizing a Random Forest classifier for binary classification tasks on the six WORC datasets. To ensure reproducible and efficient model performance, a pre-training stage was introduced to estimate the optimal model capacity prior to hyperparameter optimization. The model size was incrementally increased, and validation performance was monitored until convergence, defined as a plateau over at least three iterations. Fixing the model size at this stage prevents the optimization routine from producing excessively large or over-parameterized models, which can increase computational cost and risk overfitting without improving generalisation. By decoupling model capacity estimation from hyperparameter optimization, this procedure ensures a more stable, efficient, and interpretable model development process. Similar effects have been observed in ensemble learning, where increasing the number of base learners yields diminishing improvements once a performance plateau is reached. Empirical studies have demonstrated that reducing ensemble size can maintain accuracy while improving

[†]LIDC-IDRI dataset available at `https://www.cancerimagingarchive.net/collection/lidc-idri/` (accessed May 2024)

computational efficiency and generalization stability [203, 204]. As the evaluation of the prototype was limited to the WORC datasets, the prototype was also applied to the LIDC-IDRI dataset to establish a reference for external generalization for this thesis. While this initial study demonstrated promising predictive performance, its design relied on a single classifier and non-harmonized configurations between the two feature extractors. Differences in the number of image filters, parameterization, and feature definitions between PyRadiomics and MIRP limited the comparability of the extracted feature spaces. As a result, the prototype served primarily as a feasibility study for automated feature extraction and selection rather than a comprehensive framework for robust radiomics evaluation.

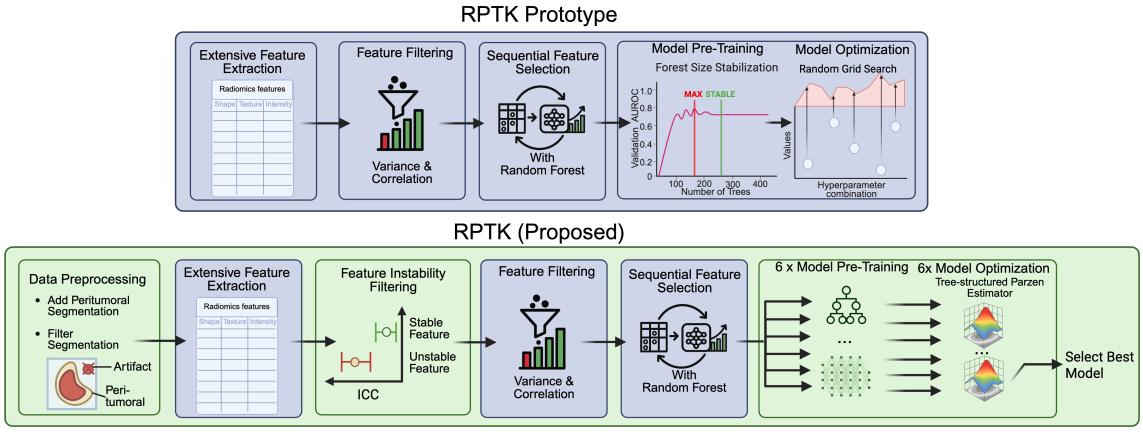


Figure 4.2. Overview of the experimental design of the first RPTK Prototype (upper) comparing to the proposed version of RPTK (lower) highlighting common structures of the workflow (in blue) and added functionalities (in green) (this Figure is partially based on my MICCAI conference paper [24]). (ICC = Intraclass Correlation Coefficient)

The prototype RPTK tool was deliberately kept simple (see Figure 4.2): (i) feature extraction with PyRadiomics (v. 3.0.1) and MIRP (v. 1.3.0), (ii) correlation filtering and variance thresholding were applied to reduce redundancy, (iii) sequential feature selection with a random forest classifier (with n_estimators=100 and default model configuration from sklearn v. 1.2.2) was performed to identify informative features, and (iiii) model training used a reduced hyperparameter space optimized (including the hyperparameters: ccp_alpha, max_depth, max_samples, max_features, and criterion) via 5 fold cross validated randomized grid search (from sklearn v. 1.2.2). This configuration allowed for a first assessment of whether differences in feature extraction (PyRadiomics vs. MIRP) translate into measurable performance differences when evaluated under a standardized yet lightweight radiomics workflow.

4.1.3 Updates of the Proposed Approach

The current proposed version of RPTK presented in this thesis extends the foundation of the prototype (see Section 4.1.2) with several methodological enhancements (see Figure 4.2). This Framework follows a structured workflow, adapted from the general radiomics framework (Figure 2.2), and consists of four main stages: data preprocessing, data augmentation, radiomics feature computation, and model training and optimization. Each stage integrates methodological extensions to improve robustness, reproducibility, and clinical relevance.

Updates to the current propose RPTK tool are: First, the framework incorporates further data preprocessing including the addition segmentation artifact filtering by detection and deletion of connected components. In addition, a perturbation-based stability assessment of features extracted from slightly modified segmentation similarly performed in [21] as well as the generation of the peritumoral margin for extracting additional radiomics features from the surrounding region of the ROI. This allows the identification and removal of features that are sensitive to small artifacts of the ROI, reducing segmentation-related bias as well as using information of the TME or potential related information which is not included in the segmentation directly. Second, the configurations between the extractors where synchronized in the way that both tools use the same image transformations and pixel discretizations settings. In addition, the framework integrates a broad range of six different machine-learning algorithm, each combined with a per-fold hyperparameter optimization routine, applying a TPE optimization technique (see Section 4.1.8). Finally, the optimization is resulting in a model ensemble classifier for each machine learning algorithm containing the five optimized fold-models. This expansion moves beyond the single Random Forest approach and provides a systematic comparison of model families while maintaining a consistent preprocessing and feature-selection pipeline.

4.1.4 Data Fingerprint

A first requirement for a self-configuring radiomics pipeline is the ability to process heterogeneous data sources in a reproducible manner. To evaluate this, I applied RPTK to seven openly available 3D imaging datasets covering a wide spectrum of clinical applications and imaging characteristics (see Table 4.1). These datasets represent typical challenges encountered in radiomics research, such as small sample sizes, class imbalance, varying imaging protocols (e.g. reconstruction kernels, slice thickness), and data quality (e.g. segmentation artifacts). Demonstrating consistent applicability across such diverse conditions establishes the foundation for subsequent

analyses of the pipeline.

The data fingerprint step provides a systematic overview of the imaging and segmentation characteristics of the datasets. It summarizes technical descriptors that capture the heterogeneity of acquisition protocols, image geometry, intensity distributions, and segmentation properties. This overview enables transparent reporting of data variability and supports the interpretation of downstream preprocessing and modeling steps.

For each scan and its corresponding ROIs, the following descriptors are extracted:

- **Acquisition geometry:** number of slices per scan, slice thickness and in-plane resolution.
- **ROI size and topology:** voxel count (ROI volume), number of ROIs per scan, number of connected components for each segmentation per scan.
- **ROI intensity profile:** number of gray values present in the ROI, minimum, maximum, and mean intensity values.
- **Discretization summary:** number of bins obtained when applying fixed bin width discretization (25 bin width).
- **Radiomics fingerprint:** all first-order, shape, and texture features computed with PyRadiomics on the original images (without image transformations, feature filtering, or perturbations).

Together, these descriptors provide a compact yet comprehensive "fingerprint" of the imaging and segmentation data, enabling a reproducible characterization of dataset heterogeneity at the technical level.

4.1.5 Data Preprocessing

RPTK can process tabulated clinical data within the input csv file. Missing clinical data were handled using a robust imputation strategy implemented in *sklearn* (v. 1.5.0), applying a K-nearest-neighbor imputer (*KNNImputer* with *n_neighbors*=2, using the euclidean matrix) for continuous variables and a most-frequent-value imputer (*SimpleImputer* with the *most_frequent* strategy) for categorical or ordinal variables. These clinical features will then be handed into the feature selection process together with the radiomics data in order to investigate into the predictive power of the constellation by including clinical data into the radiomics feature space.

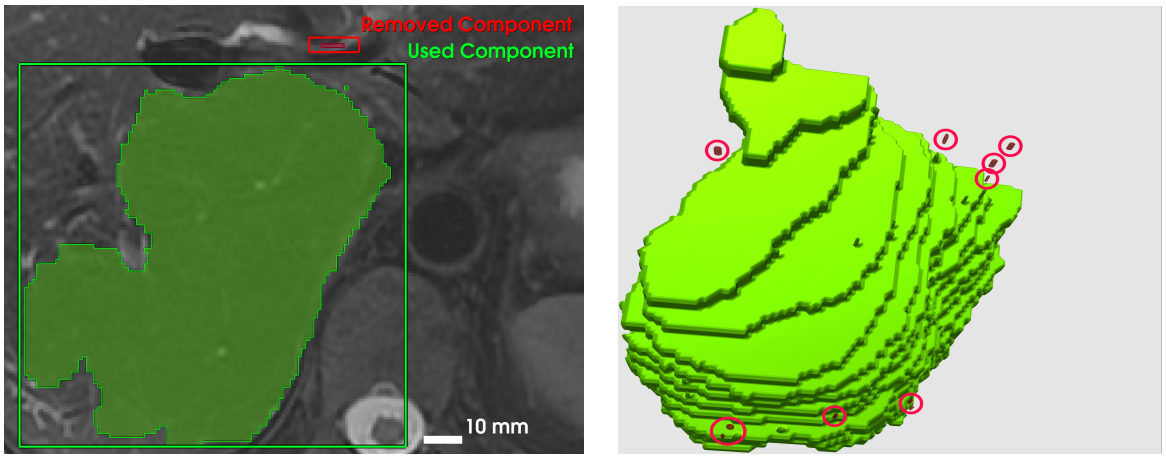
The preprocessing pipeline in the RPTK is designed to standardize and harmonize imaging data and segmentation masks, ensuring consistency and robustness prior to radiomics feature extraction. This pipeline addresses inter-scanner variability, segmentation inconsistencies, and the impact of heterogeneous acquisition protocols on

feature reproducibility. By implementing these standardized steps, the RPTK provides a reliable foundation for reproducible and generalizable radiomics analyses.

The first step, resampling of images and segmentations, ensures that all scans are standardized to isotropic voxel spacing and aligned to a uniform orientation which effects radiomics features and therefore also the final model performance [133, 134]. Segmentations were resampled with K-Nearest Neighbor (*sitkNearestNeighbor*) and images were resampled with the B-Spline (*sitkBSpline*) algorithm from *SimpleITK* (v. 2.5.2) to 1 mm^3 isotropic voxel spacing [205]. This guarantees comparability of radiomics features across heterogeneous datasets. Next, z-score image normalization is performed on MR data to compensate for non-normalized intensity inhomogeneities this has been done inside the feature extraction procedure included in PyRadiomics (v. 3.0.1) and MIRP (v. 1.3.0) (see Section 2.1.2). Image normalization is not performed on CT images based on their standardized calibration (see Section 2.1.1). According to influencing factors in CT scan protocol variations, z-score normalization might also get applied for high variations in convolution kernels, or contrast agent applications (see Section 2.1.1), but this was not done within this thesis.

Segmentation filtering is then applied to refine the ROIs. Segmentations that do not span multiple slices in any orientation are excluded, as well as masks which are smaller than three voxels. Connected component filtering has been performed by using the *label* function from the *scitkit-image* library (v. 0.25.2) with a connectivity of 1 voxel performing a direct connectivity filtering (components need to be connected by direct neighboring voxels) [206]. For multi-component segmentations, only the largest connected component is retained, while small isolated regions are removed (Figure 4.3). This connected component filtering step ensures that only the clinically relevant lesion is preserved. RPTK can be set up to process multiple ROIs within a single sample.

For each ROI a surrounding segmentation is additionally generated in order to add the peritumoral region of a tumor or the surrounding region of the ROI. The surrounding region has been generated by morphological dilation using the *dilation* function from *scikit-image* (v. 0.25.2). The peritumoral region involves the TME which has been identified as containing important information for several clinical tasks [207–209]. Therefore, a surrounding region of 3 voxels around the ROI has been defined to additionally compute features from this region and further extend the radiomics analysis and gain additional information.



(a) Connected component segmentation filtering to remove segmentation artifacts (red) and use a filtered segmentation for further processing (green).

(b) 3D View of filtered connected component segmentation in top down view with segmentation artifacts (red circled) and filtered segmentation (green).

Figure 4.3. Connected component segmentation filtering applied across all datasets to filter segmentation artifacts (pictures are generated with MITK v. 2024.12) [67]. Examples display a T2w MRI from a liver tumor from the Liver dataset.

4.1.6 Image and Segmentation Data Augmentation

Radiomics features are sensitive to variations in both segmentation definitions and imaging protocols. Even subtle changes in region delineation can substantially alter feature values, leading to reduced reproducibility and generalization. Prior work has demonstrated that segmentation variability among raters can strongly influence radiomics signatures, with especially pronounced effects for texture features [70, 210]. Similarly, systematic reviews highlight that inter-rater segmentation differences remain a major source of instability across radiomics studies [121].

To address this, segmentation perturbation strategies were introduced in this study to mimic inter-observer variability and thereby filter out unstable features. Three perturbation strategies are employed:

- **Random contour change perturbation** - randomly added/removed voxels at the mask boundary from [191]
- **Random Supervoxel perturbation** - randomly added/removed voxels based on supervoxels from [21]
- **Morphological Dilation perturbation** - add voxels by morphological dilation from [206]

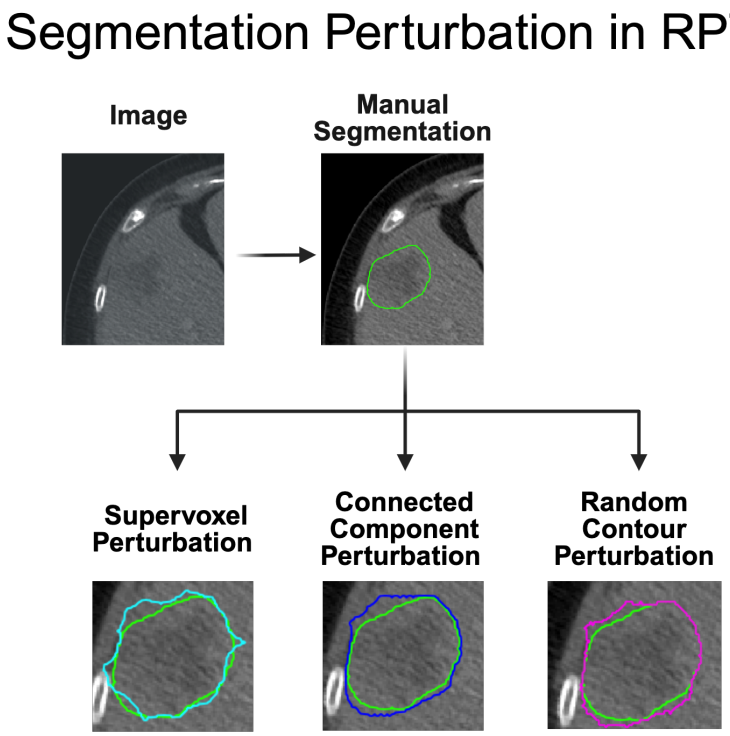


Figure 4.4. Simulation of interrater segmentation variability. Controlled perturbations are generated from the original mask to mimic real-world segmentation differences between segmentators, which are then used to identify unstable features. Randomized supervoxel change has been used from the MIRP implementation [21]

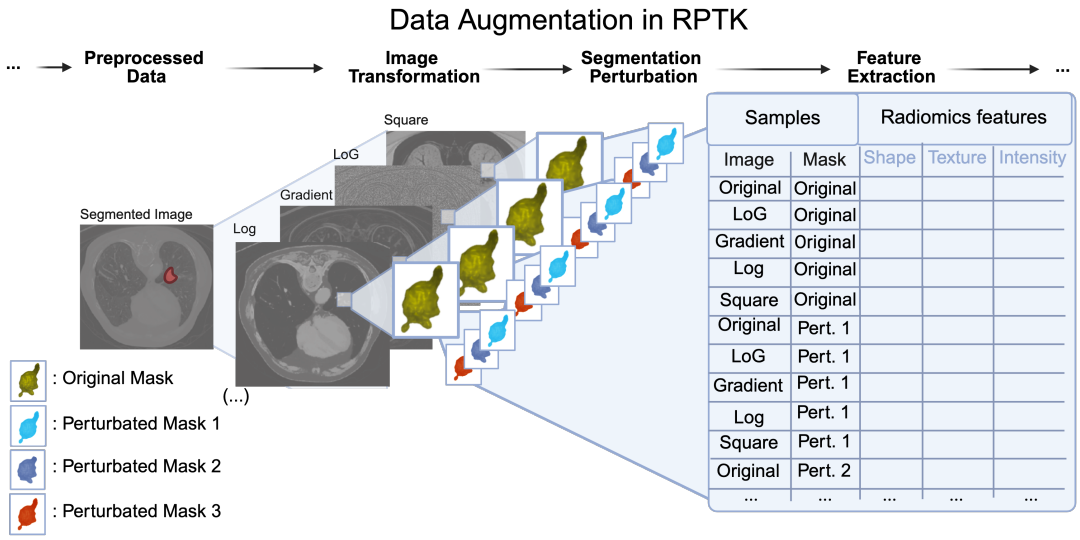


Figure 4.5. Implementation of data augmentation in RPTK includes image transformation and Segmentation perturbations. The data augmentation multiplies the extracted feature space.

In comparison to labor intensive and expensive generation of multiple manual seg-

mentations, segmentation perturbation generations arbitration from one single manual segmentations (see Figure 4.4). To ensure that only realistic segmentation perturbations were included in the feature robustness analysis, the spatial overlap between each perturbed and the corresponding manual segmentation was evaluated using the Dice similarity coefficient (cf. Eq. 2.4). Only perturbations with a Dice value greater than 0.85 were retained for subsequent instability filtering. This threshold was chosen to exclude unrealistic or artifact-prone segmentations while preserving minor, anatomically plausible variations. The Dice computation was implemented in Python (v. 3.10) using *NumPy* (v. 1.26.4) for voxel-wise logical operations. The Features that consistently showed instability across perturbations were excluded from further analysis. Feature robustness can be systematically evaluated under segmentation variability [70].

In addition to segmentation variability, imaging protocol heterogeneity—such as differences in convolution kernels, voxel sizes, and contrast administration—also impacts feature distributions. Radiomics features can vary significantly across CT reconstruction kernels [43], while acquisition-related variability, particularly voxel size and slice thickness, is a dominant factor influencing feature stability [211]. To account for such heterogeneity, a set of image transformations either from PyRadiomics [15] or form the MIRP [21] tool were applied in this work:

- 3D wavelets including High or Low pass filter in different combinations (LLL, LLH, LHH, HHH, HLH, HHL, HHH, LHL, HLL) from PyRadiomics (v. 3.0.1) [15]
- Laplacian-of-Gaussian (LoG) filters from PyRadiomics (v. 3.0.1) [15]
- Local Binary Pattern from 2D (LBP2D) from PyRadiomics (v. 3.0.1) [15]
- Mean from MIRP (v. 1.3.0) [21]
- Expotential from PyRadiomics (v. 3.0.1) [15]
- Gradient from PyRadiomics (v. 3.0.1) [15]
- Gaussian smoothing filters from MIRP (v. 1.3.0) [21]
- Laws' texture filter from MIRP (v. 1.3.0) [21]
- Gabor filters from MIRP (v. 1.3.0) [21]
- Logarithmic from PyRadiomics (v. 3.0.1) [15]
- Polynomial transforms (square, square-root) from PyRadiomics (v. 3.0.1) [15]

These transformations emulate protocol-related differences in intensity distributions and texture characteristics, thereby improving robustness against acquisition variability.

Together, segmentation perturbations and image transformations form a complementary augmentation pipeline (see Figure 4.5), designed to systematically expose the feature space to realistic variability during preprocessing. Radiomics features for every

transformed image as well as for every perturbated segmentation in combination need to get extracted. This generated a up to 63 times bigger input data handed to the feature extraction. However, this added information increases the needed computational resources and also comes with additional bias and redundancy. Therefore, filtering of non necessary, repetitive, or non-beneficial information reduces the amount of needed computational resources for model training. By filtering unstable features and training models on data augmented for both segmentation and imaging heterogeneity, the resulting radiomics signatures are expected to achieve higher reproducibility, stability, and ultimately clinical transferability.

4.1.7 Radiomics Feature Computation and Reduction

Radiomics features were extracted using both PyRadiomics (v. 3.0.1) and MIRP (v. 1.3.0). Harmonized extractor settings were applied to maximize overlap with IBSI definitions while exploiting complementary feature spaces. Feature computation included intensity, shape, and texture features derived from original and transformed images.

Although both extractors follow IBSI conventions, they differ in their feature coverage and composition. MIRP provides a complete set of IBSI-defined features, whereas PyRadiomics lacks a subset of IBSI features and includes additional non-IBSI features. Moreover, the relative distribution of feature classes (e.g. Grey Level Co-occurrence Matrix (GLCM), Grey Level Run Length Matrix (GLRLM), Grey Level Size Zone Matrix (GLSZM), morphological features) varies between the two extractors. These differences are illustrated in Appendix 8.7 and Appendix 8.8. In addition, RPTK does not extract morphological radiomics features from the peritumoral margin as the morphology characteristic of the surrounding region does not related to the morphological constitution of the ROI itself and should concentrate on structural and intensity features. Therefore, the resulting initial feature space extracted by these tools differ including feature augmentations from Section 4.1.6. For MIRP the initial feature space result in about 6,766 features, whereas the initial feature space of PyRadiomics results in about 3,546 features.

After feature extraction, all features were standardized using z-score normalization, subtracting the mean and dividing by the standard deviation using the *Pandas* library (v. 2.2.1) with Python (v. 3.10). Afterwards, duplicated features were removed in order to remove redundant information.

To reduce dimensionality and enhance robustness, several filtering steps were applied: (i) variance filtering to discard near-constant features, (ii) correlation filtering to eliminate redundant features, and (iii) instability filtering with perturbation based

Intraclass Correlation Coefficient (ICC) calculations. This procedure is motivated from repetitively, successfully performed approaches for radiomics feature reduction in the literature [212–214].

To remove non-informative features which are showing very small variance across the dataset, variance filtering was applied (i). This step was performed using the *VarianceThreshold* function from the *scikit-learn* library (v. 1.5.0) [191], which excludes all features with a variance below a specified threshold. A threshold of 0.1 was applied, resulting in the removal of features exhibiting minimal variability across samples, thereby reducing noise and improving the stability of subsequent analyses.

To identify and remove highly correlated features (ii), a correlation-based filtering procedure was applied prior to model training. Pairwise Pearson correlation coefficients were computed between all numeric features using the *corr* function from the *Pandas* library (v. 2.2.1), which internally relies on *NumPy* (v. 1.26.4) for numerical operations. Absolute correlation values ($|r|$) were considered to detect both positive and negative dependencies. From each pair of features with an absolute correlation exceeding the threshold of $|r| > 0.90$, only the one feature was retained, while the other was excluded from the dataset. This deterministic filtering approach reduces feature redundancy and mitigates potential multicollinearity effects, ensuring a more stable and interpretable feature set for subsequent modeling steps.

To assess the robustness of radiomics features with respect to segmentation variability, an intraclass correlation analysis was performed across all perturbed segmentations of the dataset (iii). For each feature individually, the one-way random-effects, single-measurement model ICC$(1, 1)$ was computed [215, 216], treating the subjects as fixed targets and the segmentation perturbations as random raters. This formulation quantifies the proportion of total variance in feature values that is attributable to inter-subject differences relative to variance induced by segmentation perturbations. The ICC(1,1) statistic was derived from an ANOVA model using the ratio of between-subject to within-subject mean squares, and a one-sided F-test ($H_0 : \text{ICC} = 0$; $H_A : \text{ICC} > 0$) was used to test for nonzero reliability. Approximate 95% confidence intervals were obtained using the F-distribution–based transformation proposed by McGraw and Wong [216]. All computations were implemented in Python using *Pandas* (v. 2.2.1) for data handling, *NumPy* (v. 1.26.4) for numerical operations, and *SciPy* (v. 1.13.1) for statistical calculations. Features showing low reliability, i.e. an ICC below the threshold 0f 0.9, were excluded from subsequent modeling steps to ensure that only stable and reproducible features were retained. This approach follows established radiomics robustness analyses, where ICC(1,1) is commonly used to evaluate feature stability across image or segmentation perturbations [21, 70].

From the remaining feature set, a Sequential Feature Selection (SFS) procedure was applied to identify the most informative subset of predictors. The implementation followed the *SequentialFeatureSelector* from the *mlxtend* library (v. 0.23.4) [217], using a random forest classifier from *scikit-learn* (v. 1.5.0) with 100 estimators and default hyperparameters as the applied model. SFS was executed in both forward and backward directions: the forward selection iteratively added features that maximized the model performance, while the backward selection iteratively removed the least informative features. In each direction, feature subsets were evaluated based on the AUROC using fivefold cross-validation, and the ten best-performing features were retained per direction. Features selected in both forward and backward passes were merged into a final candidate set which consists of up to 20 radiomics features. The use of the SFS algorithm showed promising performance in published radiomics studies such as [218, 219] and was therefore implemented in the RPTK workflow.

4.1.8 Model Training and Optimization

Before model training, the dataset was examined for class imbalance, a common challenge in radiomics where unequal representation of clinical outcomes can bias learning toward the majority class and impair model generalization [220]. To mitigate this effect, the Synthetic Minority Over-sampling Technique (SMOTE) [221] was applied to the training data whenever one class comprised $\geq 65\%$ of the samples. SMOTE generates synthetic minority samples by interpolating between each minority observation and its k-nearest neighbors ($k = 5$ by default), thereby balancing class frequencies in feature space without simple duplication. This oversampling strategy has been successfully adopted in several recent radiomics studies, demonstrating improved classification performance and more stable model behavior in imbalanced datasets [220, 222]. In this work, oversampling was implemented in Python using the *SMOTE* function from the *imbalanced-learn* library (v. 0.12.3) [223], applied exclusively to the training split to prevent information leakage into validation or test data. Newly generated synthetic samples were flagged with an identifier prefix (*simu-*) to ensure full traceability. Table 4.3 represent the distribution of the binary classification labels across the open-source datasets.

Selected features were used to train six different classifiers, covering a diverse spectrum of modeling paradigms for binary classification tasks:
(i) Random Forest Classifier from *scikit-learn* (v. 1.5.0) [191],
(ii) Gradient Boosting Classifier from *scikit-learn* (v. 1.5.0) [191],
(iii) XGBoost Classifier from *xgboost* (v. 2.0.3) [224],
(iv) Light Gradient-Boosting Model (LGBM) Classifier from *lightgbm* (v. 4.6.0) [225],

(v) TabNet, a deep learning architecture for tabular data from *pytorch-tabnet* (v. 4.1.0) [226],
(vi) Support Vector Classifier (SVC) with a linear kernel for linear decision boundaries from *scikit-learn* (v. 1.5.0) [191]
This set of models was deliberately chosen to ensure high heterogeneity in performance comparison, spanning ensemble tree-based methods, gradient boosting approaches, deep learning, and linear margin-based classification.

Table 4.3. Label distributions of the training and test splits of the open-source datasets across imaging modalities and tumor types. The table reports the number of patients (and percentages) per class.

Dataset	Label	Training (n, %)	Test (n, %)
Lipo (T1w MRI)	Well-differentiated liposarcoma	46 (49%)	11 (48%)
Lipo (T1w MRI)	Lipoma	45 (51%)	12 (52%)
Desmoid (T1w MRI)	Desmoid-type fibromatosis	57 (35%)	15 (37%)
Desmoid (T1w MRI)	Extremity soft-tissue sarcoma	105 (65%)	26 (63%)
Liver (T2w MRI)	Malignant primary solid liver tumor	75 (51%)	19 (50%)
Liver (T2w MRI)	Benign primary solid liver tumor	73 (49%)	19 (50%)
GIST (CT)	Gastrointestinal stromal tumor	98 (51%)	25 (51%)
GIST (CT)	Other intra-abdominal tumors	97 (49%)	24 (49%)
CLRM (CT)	Colorectal liver metastases	29 (48%)	7 (50%)
CLRM (CT)	Other colorectal tumors	32 (52%)	7 (50%)
Melanoma (CT)	Lung metastases of melanoma	38 (50%)	9 (47%)
Melanoma (CT)	Other lung tumors	38 (50%)	10 (53%)
LIDC-IDRI (CT)	Benign lesion	29 (32%)	7 (30%)
LIDC-IDRI (CT)	Malignant lesion	63 (68%)	16 (70%)

Each model underwent a pre-training step to calibrate its complexity (e.g., number of estimators for tree-based models (i - v), margin parameters for SVM) to ensure stable convergence (see Section 4.1.2 for details). Hyperparameter optimization was then performed within a five-fold stratified cross-validation framework, with 200 optimization iterations per fold. Optimization was guided by the Tree-structured Parzen Estimator (TPE) algorithm as implemented in *optuna* (v. 3.6.1) [153]. The cross-validation algorithm selects randomly samples from the training set and assigns them into five equally sized parts, whereas one part gets left out and used for validation of the model trained on the other four parts. This procedure get applied 5 times where the validation part is always another fold and the training folds are always the remaining data in the training set.

The best configuration for each fold was selected based on validation AUROC, and

fold-specific models were ensembled using the *EnsembleVoteClassifier* with soft voting from *mlxtend* (v. 0.23.4) to generate the final predictions. Final performance was evaluated by the ensemble model on the held-out test sets, reporting mean AUROC with bootstrapped 95% confidence intervals. Bootstrapping has been performed on the prediction outcomes and then subsequent calculation of the evaluation matrix (AUROC, F1, Sensitivity, or Specificity) from the models on the respective data split 1,000 times (this procedure has been also applied in [18] and [19] to compute the Confidence Interval (CI)95). In addition, threshold-based metrics optimized with the Youden Index, including sensitivity and specificity, were reported to provide clinically relevant model evaluation.

4.1.9 Application of AutoRadiomics

AutoRadiomics was applied following the default configuration provided in the official repository[‡]. With this setup, I was able to reproduce the published results on the WORC database [18], confirming that the default pipeline yields consistent performance across datasets (see Figure 8.10).

After completing the AutoRadiomics runs, the training–testing splits generated by AutoRadiomics were reused for the evaluation and training of RPTK and Deep Learning. Specifically, the same sample identifiers were selected for training (80% of the data) and testing (20%) to ensure a fair comparison. In the RPTK pipeline, these splits were introduced after the feature filtering stage and were subsequently applied for both feature selection and model training/optimization, and finally model evaluation and testing.

It should be noted that while the outer train–test splits were synchronized between AutoRadiomics and RPTK, the internal cross-validation procedures were not. AutoRadiomics uses its own internal cross-validation strategy, while RPTK applies a five-fold stratified cross-validation with hyperparameter optimization as described in Section 4.1.8. Preserving this independence ensures that the performance of RPTK reflects its original configuration, while still allowing a fair comparison on identical held-out test sets.

4.1.10 Application of Deep Learning Models

To benchmark radiomics-based classification against deep learning, several 3D convolutional neural networks were trained using the *MONAI* framework [23], specifically ResNet18, ResNet200, DenseNet121, DenseNet169, DenseNet201, and DenseNet264.

[‡]`https://github.com/pwoznicki/AutoRadiomics/tree/main`, accessed April 2024

All models were implemented in *Python* (v. 3.10) using *PyTorch* (v. 2.2.2), *MONAI* (v. 1.3.0), and *TorchIO* (v. 0.19.6). Experiment tracking was performed via *Weights & Biases* (*wandb* (v. 0.16.6)), and medical image I/O relied on *SimpleITK* (v. 2.5.2) and *NiBabel* (v. 5.2.1).

All models were trained on single-channel volumetric images to ensure comparability with the RPTK and AutoRadiomics experiments. Training and test splits were kept identical across all approaches to enable a fair performance comparison.

Preprocessing: Each image was cropped around the segmentation mask (if available), resampled to isotropic $1\,\mathrm{mm}^3$ voxel spacing, and resized to $32 \times 32 \times 32$ voxels when cropping was applied (otherwise 96^3). Intensities were scaled to the range $[0, 1]$ and optionally normalized via z-score standardization.

Training setup: Models were trained for up to 200 epochs with a batch size of 15 using the Adam optimizer from *PyTorch* (v. 2.2.2) (initial learning rate 10^{-4}, minimum learning rate 10^{-6}) and a cosine annealing learning rate schedule with warm restarts [227]. The cross-entropy loss function was minimized. Early stopping with a patience of 3 epochs and $\Delta_{\min} = 0.1$ was enabled to prevent overfitting. Batch normalization layers were included in all models. Random seeds were fixed (seed = 1234) to ensure deterministic and reproducible results across runs.

Data augmentation: To increase robustness, random flips along all three spatial axes, random 90° rotations, and random intensity scaling were applied for training samples with a probability of $p = 0.2$ for each transform from the *MONAI* (v. 1.3.0) *transforms* collection, the *Compose* function was applied.

Validation and model selection. A five-fold cross-validation was performed using the same predefined splits as in the AutoRadiomics experiments. For each fold, the model checkpoint with the highest validation AUROC was retained. Test set predictions were ensembled by averaging class probabilities across folds to obtain final test scores.

Evaluation: Independent test sets were used for the final evaluation. Model performance was quantified using the area under the receiver operating characteristic curve (AUROC), reported alongside AutoRadiomics and RPTK results for direct comparison (see Section 4.1.9). All experiments were executed on NVIDIA GPUs using CUDA acceleration, and the training pipeline was made fully reproducible by saving preprocessing metadata, model weights, and experiment configurations.

4.1.11 Source Code Availability

The RPTK framework has been developed as an open-source project and is made publicly available for reproducibility and further research. The source code, configuration

files, and scripts for all experiments described in this thesis can be accessed via the Zenodo reference [228] or via GitHub[§]. The Zenodo reference shows the exact version of the release of the software used in this thesis whereas the GitHub repository can be viewed in addition for documentation purposes.

The repository includes the complete RPTK pipeline, including data preprocessing, feature extraction, feature filtering and selection, model training and optimization, and evaluation modules. In addition, the source code and configurations for the deep learning experiments (ResNet, DenseNet) are provided. All dependencies and library versions are documented in the repository.

For user convenience, a Docker container has been prepared, enabling direct deployment of RPTK on arbitrary datasets without requiring manual installation or environment configuration. This container includes all required libraries, and preconfigured workflows, thereby lowering the barrier for application in both research and clinical settings.

4.1.12 Used Computational Hardware

All experiments were performed on a machine with Intel(R) Xeon(R) W-2145 CPU @ 3.70GHz CPU including 16 Cores with 125 GB of Memory and 9 TB of disk space. Calculation time for RPTK applications took around 72 hours for the smaller datasets and up to 96 hours for the larger datasets. Application of AutoRadiomics needed 28 until 42 hours for calculation with the same hardware.

4.1.13 Statistical Testing

To evaluate whether differences in model performance were statistically significant, the AUROC values obtained on the test set were compared using the nonparametric DeLong test for *paired* and *correlated* receiver operating characteristic (ROC) curves [229]. As both models were evaluated on the same test samples, a paired formulation of the test was used. The analysis was performed in R using the *pROC* (v. 1.19.0.1) package in R (v. 4.5.1) [230]. A two-sided hypothesis test was used to assess whether the observed difference in AUROC values deviated significantly from zero ($H_0 : \mathrm{AUC}_1 = \mathrm{AUC}_2$, $H_A : \mathrm{AUC}_1 \neq \mathrm{AUC}_2$). The corresponding p-value was computed at a significance level of $\alpha = 0.05$.

Overall survival was analysed with Kaplan–Meier (KM) estimators [231] for the Predict study in order to show survival differences stratified by response groups. To

[§]Access the RPTK GitHub Repository: `https://github.com/MIC-DKFZ/RPTK`

test for differences between the group-specific survival functions, I used the (multi-sample) log-rank test [232, 233] with a two-sided alternative and significance level $\alpha = 0.05$. KM curves (with 95% confidence intervals from Greenwood's variance) and the global log-rank p-value were computed in Python using *lifelines* (v. 0.27.8; *multivariate_logrank_test*) and *KaplanMeierFitter*.

4.2 Predict Study – Predicting Immunotherapy Treatment Response in Lung Cancer Patients

This section documents how the Predict cohort was used to evaluate our radiomics pipeline. First, we summarize the dataset (patients, imaging, and clinical variables used in modeling). Second, we specify the retrospective study design, including eligibility, imaging time points, and the training–testing protocol; for comparability, the train/test splits are synchronized with AutoRadiomics. Third, we detail the RPTK application to this cohort, covering image/ROI handling and preprocessing, feature extraction and parameterization, model selection and hyperparameter search, and the evaluation metrics used. Together, these elements define a reproducible setup for a fair comparison between RPTK, AutoRadiomics, and the deep learning baselines. Each patient contributed two CT scans from the earliest time points of the treatment.

4.2.1 Data & Study Design

The ethics application for this this retrospective study was approved by the ethics committee of the Heidelberg University Hospital based on the national laws and the Declaration of Helsinki (S-145/2017). This section specifies (i) the data used from the Predict cohort, (ii) the retrospective study design and outcome definition, and (iii) how these data were prepared for analysis with RPTK and compared against AutoRadiomics and deep learning models. An overview of the workflow and timeline is shown in Figure 4.6.

Clinical response at the first on treatment evaluation followed iRECIST categories (partial response, stable disease, progressive disease). For modeling, we defined a binary endpoint: responders = partial response or stable disease; non-responders = progressive disease (n=38 vs. n=35). Overall survival differed significantly between responders and non-responders (log-rank $p < 0.01$), see Figure 4.7. The Predict dataset was split into training and testing subsets, the training set contains 28 non responders (48 %) and 30 responders (52 %), where the test set includes 7 non-responders (47 %) and 8 responders (53 %). This splitting was done by AutoRadiomics

and has been used by RPTK to select the features and train predictive models for this study.

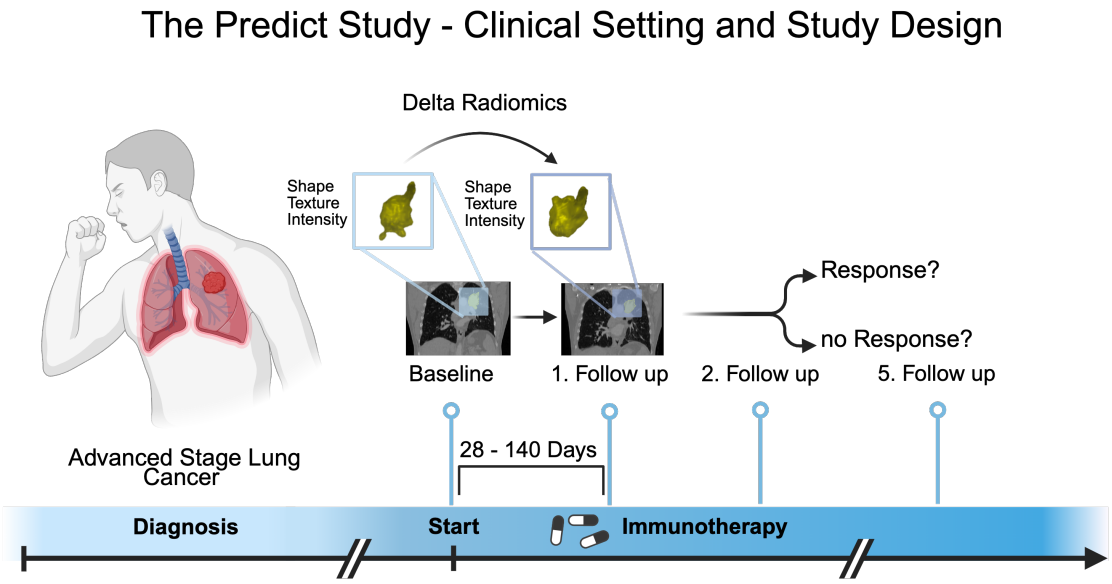


Figure 4.6. Experimental design of the Predict study for immunotherapy treatment response prediction. Included are patients with diagnosed NSCLC lung cancer, where a CT image was acquired at the start of the treatment and 28 - 140 days afterwards. RPTK was applied on the baseline and the 1. follow-up CT scans. The treatment response was evaluated during the treatment by multiple experts based on comprehensive data (see Table 8.4).

The Predict data comprises additional 28 clinical parameters (see Table 8.4), including demographic information, laboratory values, tumor staging, and therapy-related descriptors. This extensive clinical setting equals to a real clinical evaluation of patients for immunotherapy treatment response and therefore include a comprehensice clinical evaluation of the patients. Among these parameters, 14 variables are continuous (e.g., PD-L1 expression, Neutrophil over Lymphocyte Ratio (NLR), CRP, Albumin, Hemoglobin, Age, Weight, and Tumor size), 9 are categorical (e.g., Sex, Smoking status, Contrast phase, Therapy class, Pleural effusion), and 5 are ordinal or staging-related (e.g., Eastern Cooperative Oncology Group Performance Status (ECOG) performance state, T-, N-, and M-staging, and overall disease stage). Most parameters describe macroscopic or physiological conditions rather than molecular biomarkers; PD-L1 represents the only molecular parameter directly linked to immunotherapy response.

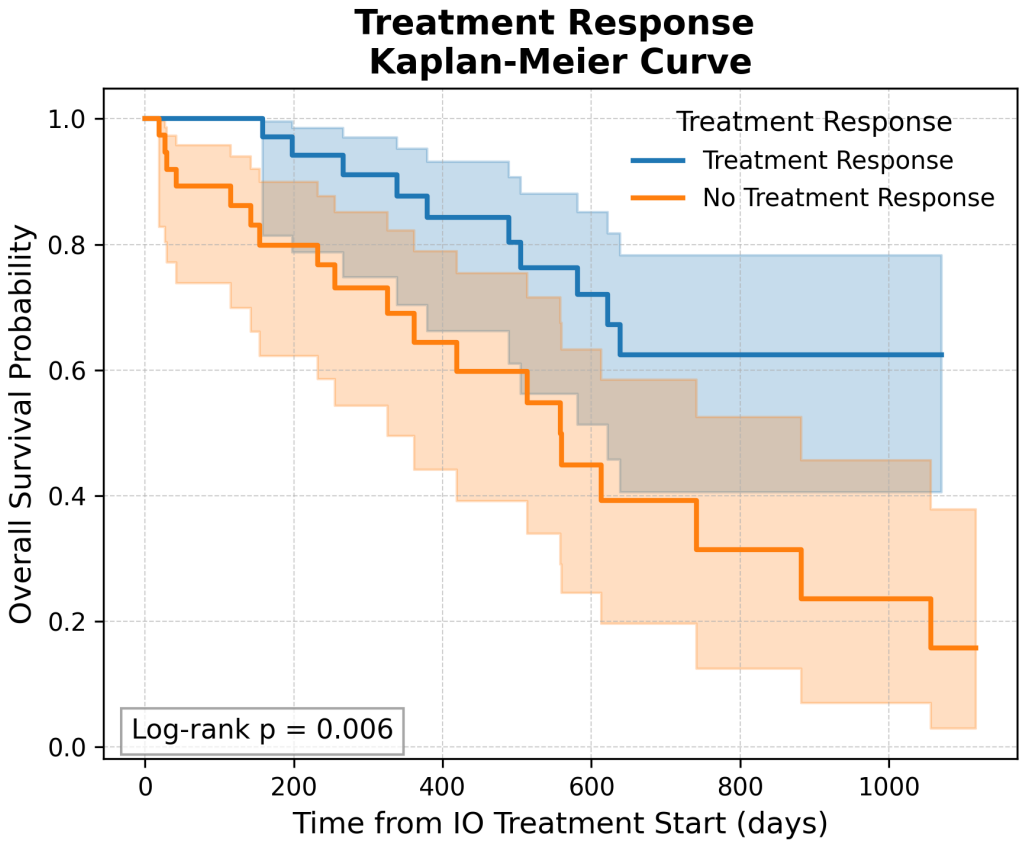


Figure 4.7. Kaplan–Meier plot for survival including binary treatment response evaluation where stable disease and partial response are handled as response with Log-rank test for significance evaluation (log-rank p-value = 0.0057) . The y axis shows the overall survival probability and the x axis the time in days after immunotherapy treatment start. This plot was generated by using the *KaplanMeierFitter* from *lifelines* (v. 0.30.0) with alpha = 0.05

4.2.2 Image Segmentation

Initial primary tumor masks were generated semi-automatically with a pretrained nnU-Net[¶] (version 2.1) [8] (see Figure 4.8), using the 3D full-resolution configuration trained on the Medical Segmentation Decathlon (MSD) Lung task (MSD Task06) (Creative Commons Attribution Non Commercial 4.0 International) [25, 26].

[¶]nnU-net pretrained model download available at `https://zenodo.org/records/3734294` (accessed June 2023)

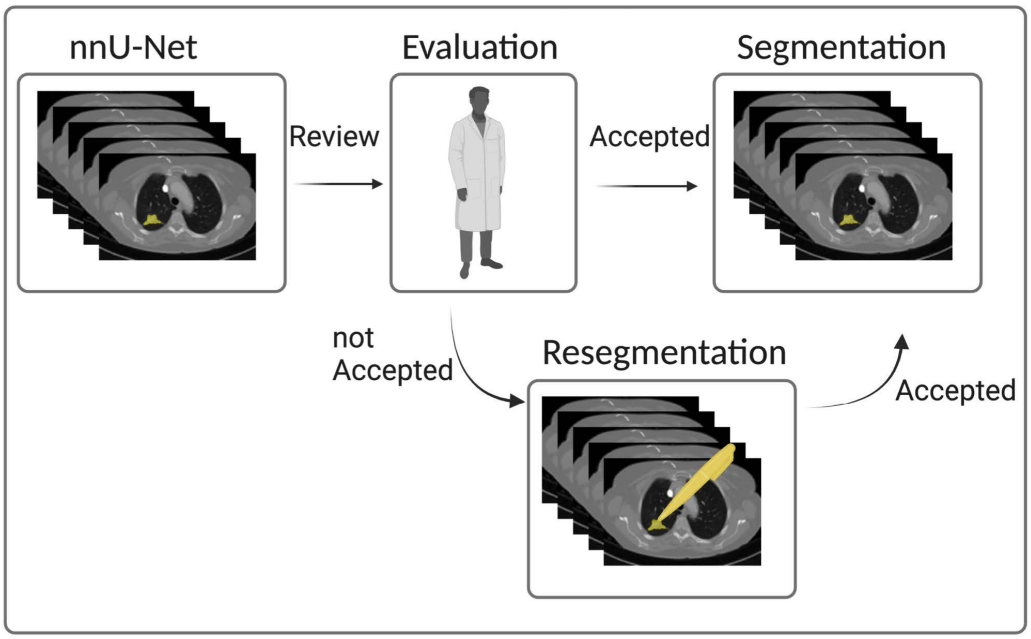
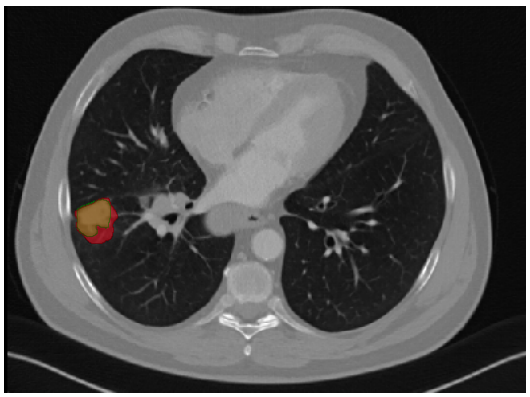
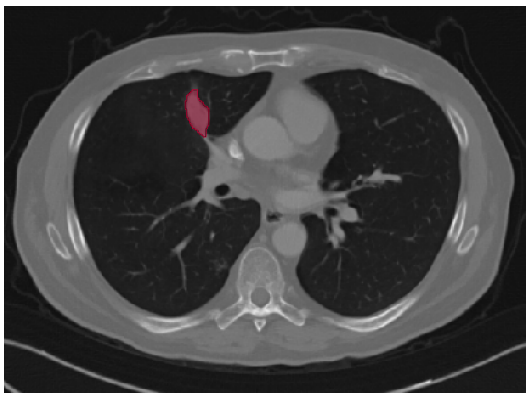


Figure 4.8. Semi-automatic segmentation workflow for reviewing and correcting automated generated segmentations by pretrained nnU-Net segmentation model for lung cancer segmentation in the Predict data [8]. This workflow generated the segmentations of the primary tumor included in this study.



(a) Not accepted nnU-Net segmentation (red) and manually performed segmentation (yellow).



(b) Accepted automated nnU-Net segmentation without manually performed segmentation correction.

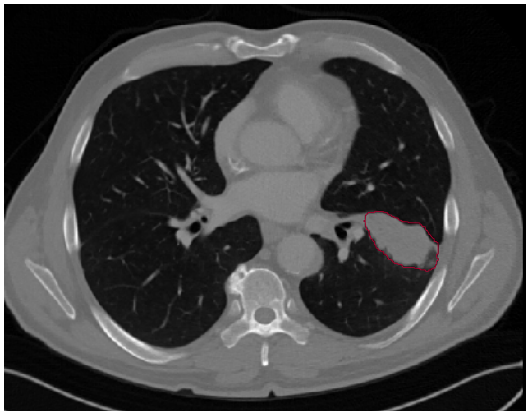
Figure 4.9. Re-segmentation performed by a radiologist based on an automated generated nnU-Net segmentation for lung cancer in the predict data (see Figure 4.8).

Inference followed the nnU-Net defaults defined by the model plans (spacing/orientation harmonization, intensity normalization, sliding-window inference with mirrored test

time augmentation, and post-processing). A radiologist (> 5 years experience) verified the primary tumor in all cases and corrected or fully resegmented 71/146 studies using MITK (v. 2022.04) [67]. A second radiologist (> 8 years experience) reviewed the full set and introduced 4 additional corrections. Representative examples of an unaccepted nnU-Net mask with manual re-segmentation and an accepted automated mask are shown in Figures 4.9a and 4.9b. All accepted (automated or corrected) masks were used for subsequent application of RPTK and AutoRadiomics.

Correct segmentation examples which got integrated into the radiomics diagnostics can be reviewed in Figure 4.10 for responding and non-responding labels.

(a) CT slice image with segmentation of a primary NSCLC not responding to immunotherapy.



(b) CT slice image with segmentation of a primary NSCLC responding to immmunotherapy.

Figure 4.10. CT images of segmented responding and not responding lung cancer in the Predict study. **a.** Example of an non-responding primary lung tumor. **b.** Example of an responding primary lung tumor.

4.2.3 RPTK Configuration in the Predict Study

The RPTK framework has been applied in the same configuration as described in Section 4.1, with the adaptation to logitudinal data. Beyond single time point radiomics, we leveraged the longitudinal design to compute delta radiomics, i.e., changes in image-derived features between the two longitudinal scans, providing a within patient readout of early therapy effects.

Delta features were computed inside the feature-extraction pipeline in three steps:
(i) extract the full radiomics feature set separately at each time point (baseline t_0 and first on-treatment t_1)
(ii) normalize features for each time point individually
(iii) subtract the normalized features per sample to obtain the delta feature vector

following the formula 4.1

Concretely, for patient i and feature k we define:

$$\Delta f_{i,k} \;=\; f_{i,k}^{(t_1)} \;-\; f_{i,k}^{(t_0)} \text{for } t \in \{t_0, t_1\} \tag{4.1}$$

$f_{i,k}^{(t)}$ Radiomic feature k for patient i at time point t (baseline t_0, on-treatment t_1).

$\Delta f_{i,k}$ Delta (change) of feature k used for modeling.

As AutoRadiomics does not include delta radiomics calculations, the same procedure has been applied to AutoRadiomics features after extraction following the formula 4.1. RPTK was then compared to the performance of AutoRadiomics and the deep learning baselines (see Section 4.1.9 and 4.1.10). The training set includes 28 non-responding patients and 30 responding patients, the test set includes 7 non-responding and 8 responding patients.

4.3 LiverCRC Study – Colorectal Neoplasia Prediction via Liver CT

The LiverCRC project investigates whether liver-derived radiomic features can serve as non-invasive biomarkers for colorectal neoplasia, thereby exploiting the biological link of the gut–liver axis. This retrospective proof-of-concept study was motivated by persistently low Colorectal Cancer (CRC) screening participation rates and aimed to explore abdominal CT scans as an opportunistic screening tool to identify patients at risk.

4.3.1 Data & Study Design

The ethics application for this retrospective study was approved by a local ethics committee based on the national laws and the Declaration of Helsinki (EK II 2023-887-AF 11) [22]. The study cohort consisted of 1,997 patients who had undergone both colonoscopy and contrast-enhanced abdominal CT. Based on colonoscopy results, 1,189 patients had no colorectal neoplasia (CRN), while 808 patients had confirmed neoplasia (adenomas, $n = 423$; CRC, $n = 385$). The liver was automatically segmented in all cases using the *MultiTalent* framework [27], and three-dimensional liver segmentations were processed with the RPTK.

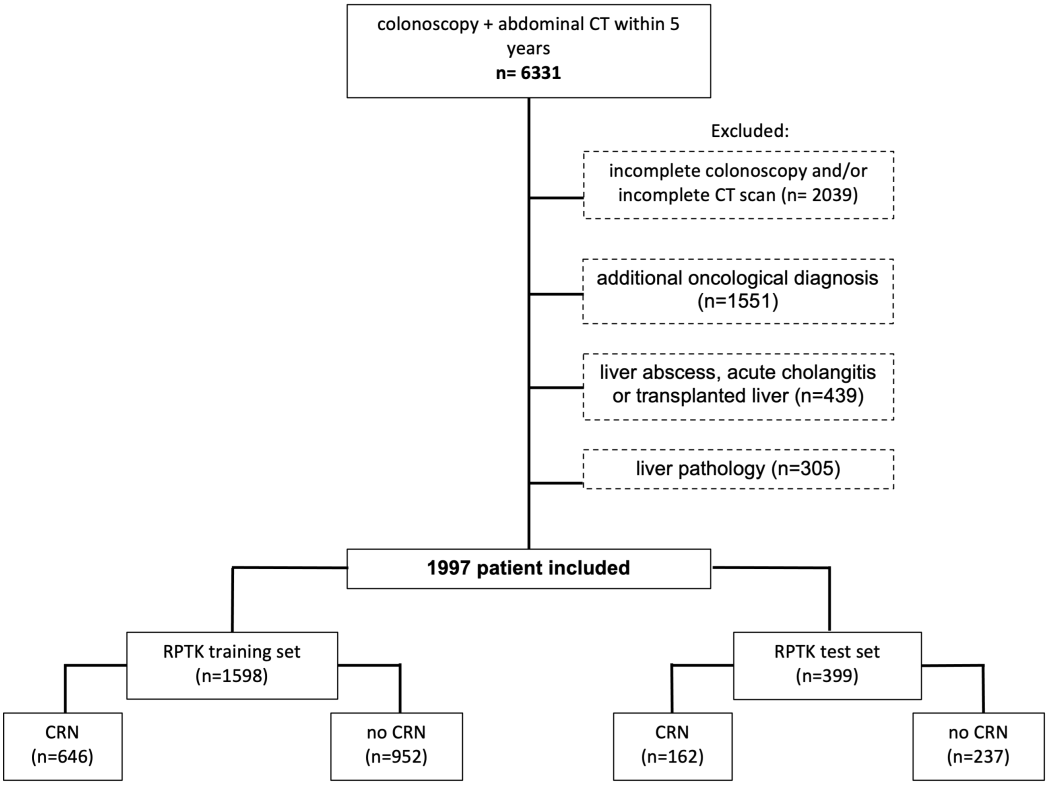


Figure 4.11. Standards for The Reporting of Diagnostic accuracy flow-chart of the LiverCRC cohort. Out of 6,331 patients with colonoscopy and abdominal CT within 5 years, 1,997 were included after excluding patients with incomplete data, additional oncological diagnosis and liver related diseases. The dataset was split into a training set ($n = 1,598$) and an independent test set ($n = 399$), with Colorectal Neoplasia confirmed by colonoscopy serving as the reference standard. (We adapted this figure from the manuscript [22]).

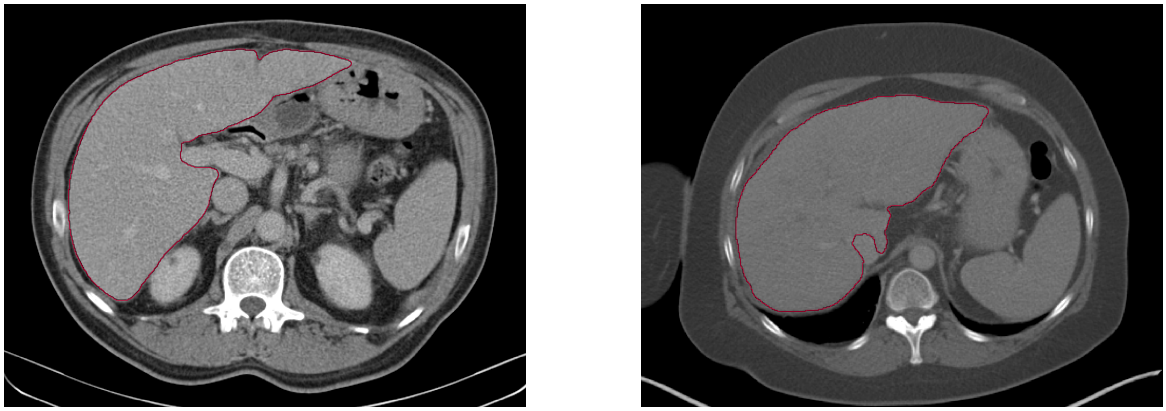
The dataset was randomly divided into a training set ($n = 1,598$) and an independent test set ($n = 399$) by AutoRadiomics. The same train/test split was also used by the feature selection procedure. Five different machine learning classifiers were trained using five-fold cross-validation on the training data, restricted to the 20 most informative features selected by the RPTK. Ensemble models were generated from the cross-validation folds. The final models were evaluated on the held-out test set using AUROC with bootstrapped 95% confidence intervals. In addition, threshold-based performance metrics, including sensitivity and specificity, were optimized using the Youden Index to provide clinically interpretable results. In addition to that we also compared the performance of RPTK with AutoRadiomics and deep learning models (see Section 4.1.9 and 4.1.10) (non-published results).

This design provides a technical foundation for investigating liver-derived radiomics as biomarkers for colorectal neoplasia. While the single-center and retrospective na-

ture of the study limits generalizability, the setup demonstrates feasibility and informs subsequent validation in prospective, multi-center cohorts.

4.3.2 Image Segmentation

Liver segmentations for this study were not manually generated but were obtained using the MultiTalent framework [‖], which is based on the nnU-Net architecture [27].



(a) CT slice image with segmentation of a the liver without colorectal neoplasia.

(b) CT slice image with segmentation of a the liver with colorectal neoplasia.

Figure 4.12. CT images of segmented liver with or without colorectal neoplasia in the Predict study. **a.** Example CT slice of a liver with no colorectal neoplasia. **b.** Example CT slice of a liver with colorectal neoplasia.

MultiTalent was trained on 13 publicly available abdominal CT datasets comprising more then 1,000 images and about 50 different anatomical classes [22, 27]. The pretrained model was applied to generate binary liver masks [22]. The resulting segmetations were used for the prediction of colorectal neoplasia (see Figure 4.12). The application of the tool as well as the generation of the segmentations were done by a colleague in the division of Medical Image Computing at the DKFZ, Heidelberg.

4.3.3 RPTK Configuration in the LiverCRC Study

The LiverCRC data were processed using the RPTK framework. In contrast to the smaller open-source datasets, no additional data augmentation techniques were applied, as the sample size of 1,997 patients was sufficiently large to support robust model training.

All images and corresponding liver segmentations were resampled to isotropic voxel spacing. Radiomics features were extracted with both PyRadiomics and MIRP, ensur-

[‖] MultiTalent available at `https://github.com/MIC-DKFZ/MultiTalent` (accessed May 2024)

ing coverage of IBSI-compliant and extended feature sets. Standard feature filtering steps were applied, including variance filtering and correlation filtering, followed by sequential feature selection using a random forest classifier to select up to 20 informative features.

Instability filtering based on segmentation perturbations was not performed in this study, as no perturbation masks were generated for the LiverCRC cohort. The resulting feature sets were subsequently used for model training and optimization as described in Section 4.1.8.

In addition, RPTK was applied to allow a direct performance comparison with AutoRadiomics (non-published data). For this purpose, the extracted and filtered radiomics features were used, while feature selection and model training/optimization were performed on the training sets (80%) defined by AutoRadiomics. Final predictions were then obtained on the corresponding AutoRadiomics-defined test sets (20%), ensuring a synchronized evaluation protocol (see Section 4.1.9).

Chapter 5

Results

This chapter presents the results obtained in this thesis, organized into three interconnected parts that follow the methodological structure described in Chapter 4.

The first part, *Self-Configuring Radiomics Pipeline*, reports on the benchmarking of the Radiomics Pipeline Toolkit (RPTK) across seven publicly available datasets. These experiments demonstrate the reproducibility and robustness of RPTK on heterogeneous imaging data and enable direct comparisons against established frameworks such as WORC and AutoRadiomics (see Section 4.1).

The second part, *Predict Study – Predict Immunotherapy Response in Lung Cancer Patients*, evaluates the clinical applicability of RPTK in a retrospective setting. Here, RPTK is applied to the longitudinal *Predict* study to forecast treatment response in patients undergoing immunotherapy for primary lung cancer. The study demonstrates how temporal radiomics features can contribute to early stratification of responders and non-responders, supporting clinical decision-making in advanced-stage cases (see Section 4.2).

The third part, *LiverCRC Study – Colorectal Neoplasia Detection on the Liver*, explores the scalability of RPTK when applied to a large dataset of nearly 2,000 patients. In this setting, radiomics features of the liver are investigated for their potential to non-invasively detect colorectal neoplasia, thereby addressing challenges related to colon segmentation and small-lesion detection. RPTK has been applied to the data in an non-syncronized way in a mansucript which is currently in submission [22]. My contribution in [22] lies in the application of RPTK, as well as the analysis and presentation of the results. In contrast to the manuscript, I focus on the comparison of the RPTK performance to AutoRadiomics and deep learning models on the imaging data from [22] in my thesis (see Section 4.3).

Across all three sections, RPTK is systematically compared to the SOTA AutoRadiomics framework (see Section 4.1.9 and 3.2.3) and SOTA deep learning models

(DenseNet, ResNet) (see Section 4.1.10 and 3.1), providing a comprehensive evaluation of its methodological advances, clinical relevance, and scalability.

The following sections present quantitative and qualitative results for each component, accompanied by comparisons to established methods and statistical analyses of performance differences.

5.1 Self-Configuring Radiomics Pipeline

This section presents the results of applying the RPTK framework to seven publicly available imaging datasets (see Table 4.1 and Section 4.1 for dataset details).

To ensure comparability with published methods and reproducibility on open-access data, the experiments evaluate how RPTK performs across heterogeneous 3D datasets that represent clinically relevant tasks and challenging small-sample size scenarios.

My specific contributions to this work include data acquisition and preprocessing, the development and implementation of the RPTK framework, and the quantitative analysis of the benchmarking results. The datasets and their characteristics are summarized in Chapter 4 and Table 4.1.

The following subsections present the results in four parts:
(i) an analysis of dataset heterogeneity captured by the automatically generated RPTK fingerprint
(ii) an overview of feature selection outcomes, including the number and origin of selected features from intra- and peritumoral regions and the distribution of feature classes
(iii) the automated selection of the best-performing models trained on these selected features and their corresponding predictive performance
(iv) a comparison of RPTK with the current state-of-the-art radiomics frameworks, deep learning models, and published benchmark approaches

Together, these results illustrate the adaptability, robustness, and benchmarking performance of RPTK across diverse imaging datasets.

5.1.1 RPTK Handles a Variety of Different 3D Imaging Datasets

To systematically capture heterogeneity across datasets, an exploratory data fingerprint was implemented in RPTK. The fingerprint provides descriptive metrics of imaging protocols, segmentation characteristics, and ROI properties that may influence ra-

diomics analyses. These metrics serve as an overview of dataset-specific characteristics and support quality assessment and pipeline adaptation.

One representative acquisition parameter included in the fingerprint is the slice thickness, which showed substantial variation across datasets (Figure 5.1). The distributions of slice thickness differ both in shape and in the median among the datasets. The *GIST* dataset exhibits a categorical distribution of slice thickness, whereas the remaining datasets display broader or more continuous distributions. In *Desmoid* and *CRLM*, a small number of scans with higher slice thickness values are visible.

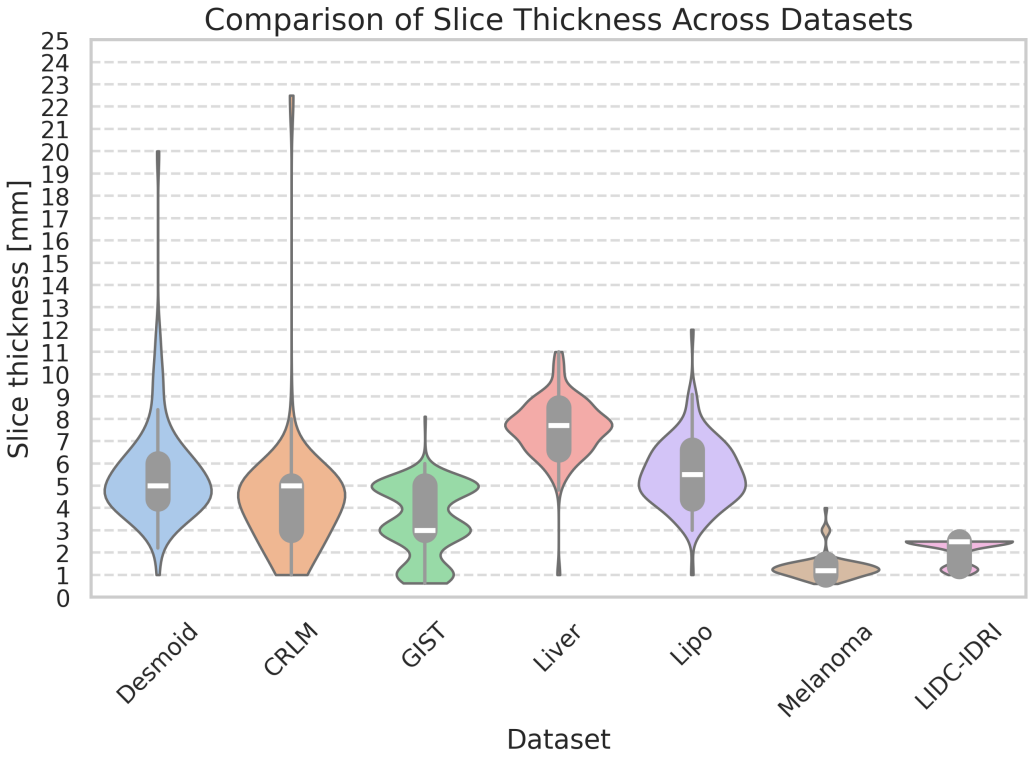


Figure 5.1. Comparison of slice thickness distributions across datasets. Each violin plot represents the slice thickness variability within each dataset, illustrating the heterogeneity in acquisition protocols. In the center of each violin is a small box plot, showing the ends of the first and third quartiles and a horizontal line showing the median. The plot was generated with the *Seaborn* (v. 0.13.2) python library by using the *violinplot* function.

The fingerprint also quantifies the number of connected components detected within each segmentation mask (Figure 5.2). This parameter serves as an internal quality indicator for segmentation integrity. Ideally, a single connected component corresponds to one contiguous lesion, whereas a higher number of components often reflects small isolated regions, noise, or annotation artifacts that require filtering (often coming from automated generated segmentations). In RPTK, connected component

filtering is applied systematically to all segmentations to remove such spurious regions before feature extraction. Importantly, the framework can differentiate between cases with multiple user-defined ROI (for example, patients with several lung lesions) and performs connected component filtering separately for each instance. When multiple semantic or instance segmentations are provided, each segmentation mask is processed individually to preserve user-defined structures. An example of this procedure is illustrated in Figure 4.3 in Section 4.1.5. The distribution of connected components across datasets (Figure 5.2) thus provides an overview of segmentation complexity and potential artifact prevalence prior to filtering. The GIST dataset as well as the Lipo dataset show more and higher fragmented segmentations compared to the other datasets.

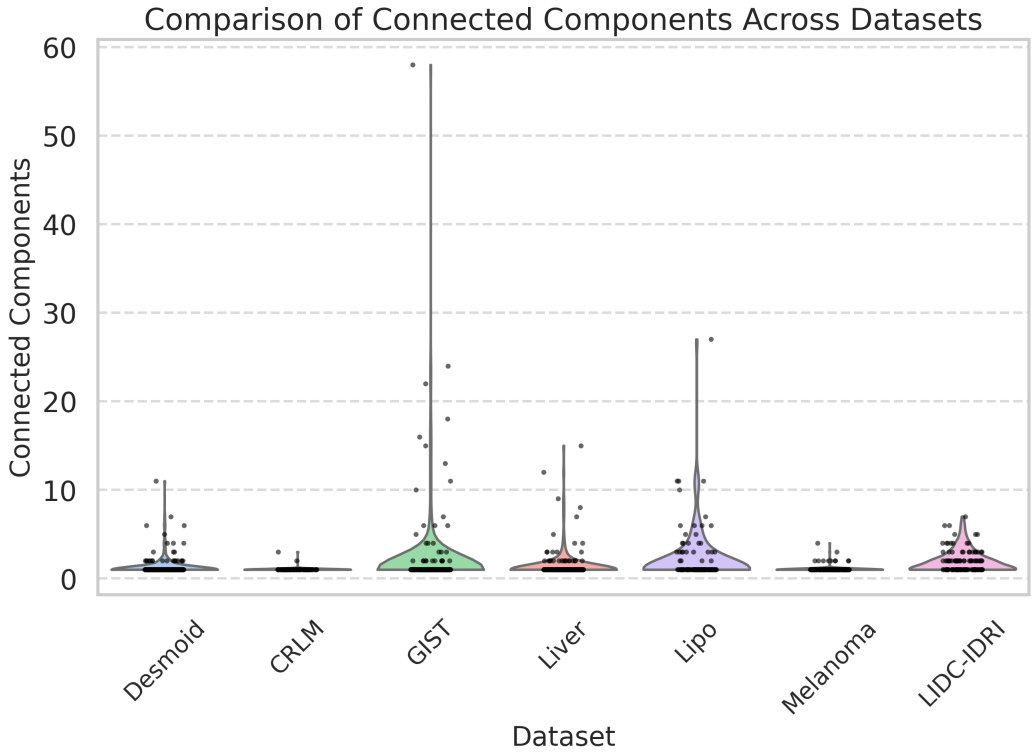


Figure 5.2. Distribution of connected components across all segmentations in the datasets. Datasets with larger variance indicate higher occurrence of multiple or fragmented segmentations. Shown is a violin plot generated with the *Seaborn* (v. 0.13.2) python library by using the *violinplot* function. The dots represent the number of connected components of each sample from the datasets.

Additional parameters captured by the fingerprint include the ROI size, number of bins, and number of slices. ROI size provides an overview of the tumor-volume distribution within each dataset and enables detection of atypical or artificially large and small segmentations (see Appendix 8.2.1 Figure 8.4). The ROIs from

the datasets CRLM, Melanoma, and LIDC-IDRI contain less voxels compared to the others, whereas the Lipo dataset contains the most voxels on average. The number of bins, computed using a fixed bin width of 25, reflects the image intensity distribution and serves as a measure of grey-level heterogeneity, indicating whether the standard discretization setting is appropriate (e.g., avoiding very few bins < 10 or excessively many bins > 100) (see Figure 8.5). The segmentations from the CRLM dataset contain less then 10 bins on average whereas the segmentations from the other dataset contain bins between 10 and 100. The number of slices describes the axial extent of each 3D volume and is influenced by the underlying clinical acquisition protocol (e.g., whole-body versus regional scans). It is important to consider the number of slices parameters together with the slice thickness, as it serves an an indicator for the ROI resolution and can be used to get details from the radiological phenotype and morphology (see Figure 8.6). The number of slices for the datasets CRLM, GIST, Melanoma, and LIDC-IDRI are on average above 100 whereas Desmoid, Liver, and Lipo do have less then 100 slices.

Beyond acquisition and segmentation descriptors, the fingerprint also includes distributions of extracted radiomics features. These allow the exploration of potential covariations between features, clinical parameters, and acquisition-related variables. Such information can serve as a quality control step to identify technical or clinical biases prior to downstream modeling. In this way, the fingerprint extends beyond simple descriptive metrics and establishes a first-level assessment of how radiomics features interact with dataset-specific conditions.

The descriptive results presented in this subsection demonstrate that RPTK accommodates datasets with pronounced heterogeneity in acquisition and segmentation properties. Their implications for harmonization, segmentation filtering, and downstream analyses are discussed in Section 6.1.

5.1.2 RPTK Selects the Most Informative Radiomics Features

In the following feature selection part, I want to show that RPTK identifies the most informative radiomics features for each dataset and extractor. The selection process reduces the dimensionality of the feature space to a subset that optimally contributes to model performance while minimizing overfitting, particularly important in small-sample scenarios where the number of features exceeds the number of available cases [10, 122]. These selected features serve as the final input for subsequent model training and evaluation.

The composition of the selected feature space differs across datasets and between the two feature extractors, PyRadiomics and MIRP. Figures 5.3 and 5.4 illustrate the distribution of selected features by IBSI feature class and by Origin of Information (OOI), distinguishing between features derived from the intratumoral and the peritumoral regions.

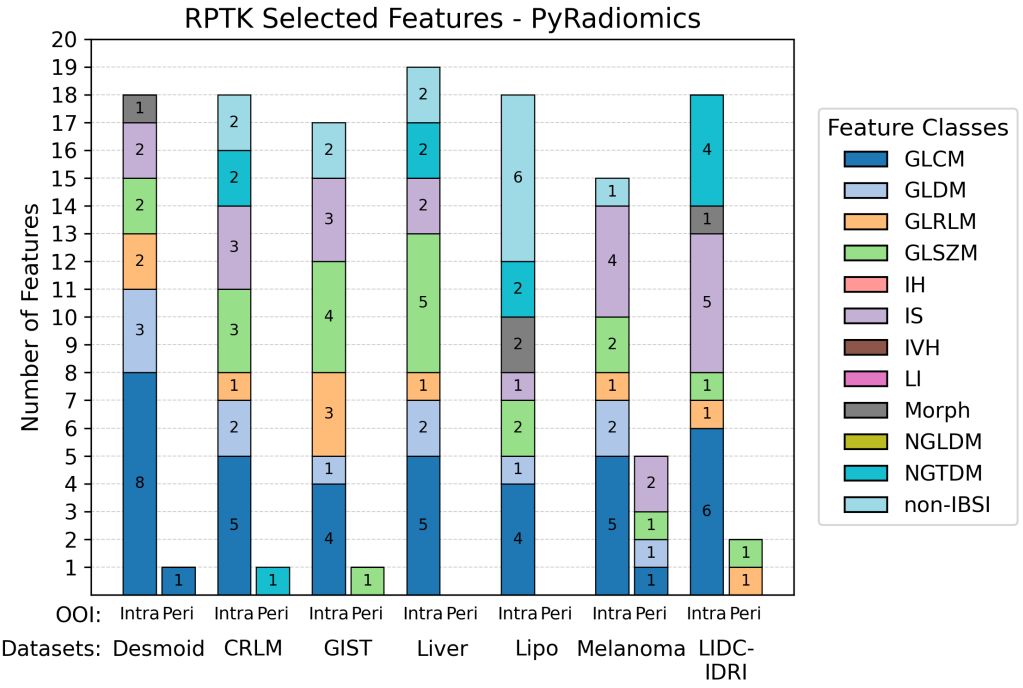


Figure 5.3. Summary of selected features extracted with PyRadiomics across all datasets. Bars indicate the number of selected features per IBSI feature class and the corresponding region of origin (OOI: intratumoral vs. peritumoral). The distribution highlights dataset-specific differences and the relative importance of different feature categories.

Across both extractors, texture-based feature classes such as GLCM, GLRLM, and GLSZM were most frequently represented in the final feature sets, followed by first-order intensity-based and morphological descriptors. These feature classes include texture information.

The peritumoral margin, was implemented as it represents the tissue surrounding the annotated lesion, capturing contextual tissue information related to tumor–host interactions [234, 235]. Features from the peritumoral regions are present in about 70% of the datasets in the selected feature space from Pyradiomics and inabout 60% of datasets in the selected feature space from MIRP. The majority of selected features extracted by MIRP for the Lipo dataset come from the peritumoral region. In other datasets, the intratumoral region dominated.

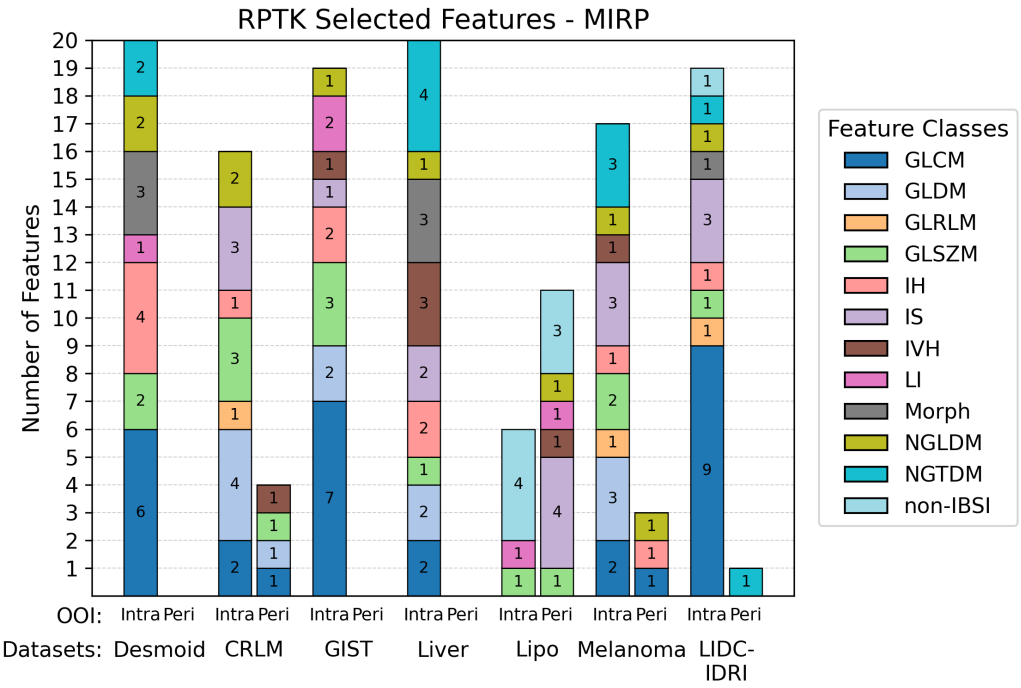


Figure 5.4. Summary of selected features extracted with MIRP across all datasets. Bars indicate the number of selected features per IBSI feature class and the corresponding region of origin (OOI: intratumoral vs. peritumoral). Compared to PyRadiomics, MIRP exhibits a broader feature-class composition due to its additional support for three-dimensional texture features and non IBSI extensions.

In total, RPTK selected up to 20 features per extraction, representing less than 5% of the initially extracted features. This drastic reduction ensures model generalizability and computational efficiency while preserving predictive signal. The resulting feature subsets form the basis for automated model selection and benchmarking presented in the next subsection.

5.1.3 RPTK Selects the Best Performing Models

The performance of radiomics pipelines is strongly influenced by both the extracted features and the configuration of the feature extraction process as well as the data size and quality. To illustrate this, RPTK integrates two widely used feature extractors, PyRadiomics and MIRP, to evaluate the effect of feature definitions and implementation details on downstream predictive performance (see Section 4.1.7 for extractor details). For smaller datasets, the train/test splitting makes also a great effect and was therefore synchronized to the AutoRadiomics splits (see Table 4.3).

To estimate the relationship between dataset size, task complexity, and the influence on model performance variance, a learning curve analysis was performed using

the *LearningCurveDisplay* function from *scikit-learn*. Each dataset was randomly subsampled in increasing proportions, and model performance was evaluated in a 5-fold cross-validation setting (see Appendix, Figure 8.11 - 8.17). The resulting curves illustrate how the AUROC evolves with increasing training size for a representative random forest classifier. The validation AUROC variances for the datasets CRLM, Lipo, Melanoma, and LIDC-IDRI are more wide and do not reach a plateau compared to the learning curves of Liver, GIST, and Desmoid where the performance variation between the iterations are not as high.

For model optimization, RPTK trains and validates six different machine learning algorithms using a five-fold cross-validation strategy (see Section 4.1.8). In each round, models are trained on four folds and validated on the remaining fold, resulting in one optimized model per fold. This setup allows performance assessment during training (validation folds) to determine whether models effectively learn from the data. Subsequently, the optimized models are evaluated on the independent test set, which remains unseen during training, to assess generalization capability. The final ensemble model is derived by aggregating predictions from the five optimized models per algorithm.

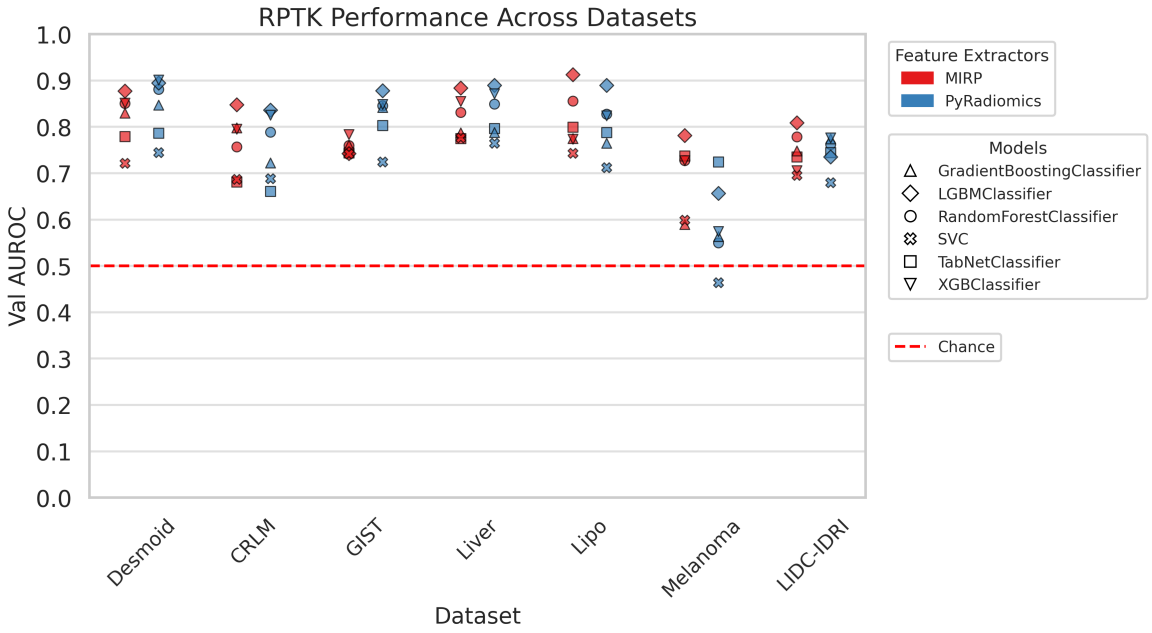


Figure 5.5. Validation performance (AUROC) of all models trained by RPTK across datasets. Each point represents the mean validation AUROC from five-fold cross-validation for one model–extractor combination. The red dashed line indicates random classification performance (AUROC = 0.5). Substantial differences between extractors and algorithms highlight the influence of feature definitions and model choice on overall performance.

Figure 5.5 summarizes the validation performance of all trained models across datasets. Each point represents the mean AUROC from the five-fold cross-validation for one model–extractor combination. Substantial variation in performance can be observed between the feature extractors, particularly for datasets such as GIST, where models based on MIRP features achieved markedly higher AUROC values compared to those based on PyRadiomics features. This highlights that the choice of feature extractor and corresponding feature definitions has a considerable impact on downstream model performance.

Across all datasets, ensemble-based tree models, specifically the XGBoost and LGBM, achieved the highest validation AUROC scores, consistently ranking among the best-performing models. In contrast, the Support Vector Machine (SVC) classifier showed the lowest and most variable performance across datasets, confirming its limited suitability for high-dimensional radiomics feature spaces. Overall, the best-performing models for each dataset achieved mean AUROC values between 0.8 and 0.9, indicating robust predictive discrimination despite the heterogeneity of the datasets.

To enhance clinical interpretability, threshold-based performance metrics were also computed. While AUROC provides a threshold-independent measure of discrimination, clinical decision-making typically depends on specific cutoff values. Therefore, sensitivity and specificity were calculated after applying the Youden correction to identify the optimal decision threshold. Table 5.1 summarizes the validation and test performance of all models across datasets and feature extractors, highlighting the best-performing configurations in bold.

Across all datasets, RPTK consistently selected ensemble-based gradient boosting models, such as LGBM and XGBoost, as the top-performing approaches. These models outperformed other evaluated architectures in most cases. Models trained on MIRP-derived features achieved the highest validation performance in four datasets, whereas models based on PyRadiomics features performed best in three datasets. The largest discrepancy between the two feature extractors was observed for the GIST dataset, with a validation AUROC difference of 0.096. Smaller differences were found for Melanoma (0.05), LIDC-IDRI (0.03), and the remaining datasets (≈ 0.02). Notably, the models trained on features from CRLM and Melanoma exhibited higher performance standard deviations compared to the others, indicating greater variability in these datasets.

The trained and optimized models identified in this section form the basis for the comparative benchmarking of RPTK against state-of-the-art frameworks presented in the next subsection.

Table 5.1. Performance metrics of best performing models in bold across datasets and feature extractors based on the mean validation AUROC metric. Best performing approach per dataset is bold. Test AUROC, F1, Sensitivity and Specificity are displayed with 95% CIs after 1000x bootstrapping as described in Section 4.1.8.

Dataset	Extractor	Model	Val AUROC	Test AUROC	Youden	Test F1	Test Sensitivity	Test Specificity
Desmoid	MIRP	LGBM	0.874 (± 0.056)	0.941 [0.867, 0.994]	0.928	0.956 [0.913, 0.991]	0.949 [0.883, 1.000]	0.981 [0.951, 1.000]
Desmoid	**PyRadiomics**	**XGBoost**	**0.901 (± 0.070)**	**0.936 [0.849, 0.994]**	**0.917**	**0.939 [0.891, 0.981]**	**0.965 [0.912, 1.000]**	**0.952 [0.908, 0.990]**
CRLM	**MIRP**	**LGBM**	**0.853 (± 0.070)**	**0.893 [0.679, 1.000]**	**0.868**	**0.769 [0.500, 0.952]**	**0.882 [0.625, 1.000]**	**0.574 [0.200, 1.000]**
CRLM	PyRadiomics	LGBM	0.834 (± 0.122)	0.842 [0.600, 1.000]	0.934	0.812 [0.545, 1.000]	1.000 [1.000, 1.000]	0.574 [0.167, 1.000]
GIST	MIRP	XGBoost	0.782 (± 0.047)	0.780 [0.640, 0.915]	0.897	0.751 [0.594, 0.880]	0.680 [0.476, 0.857]	0.877 [0.739, 1.000]
GIST	**PyRadiomics**	**LGBM**	**0.878 (± 0.021)**	**0.835 [0.709, 0.948]**	**0.826**	**0.717 [0.565, 0.850]**	**0.677 [0.500, 0.857]**	**0.791 [0.615, 0.947]**
Liver	MIRP	LGBM	0.883 (± 0.056)	0.809 [0.663, 0.937]	0.770	0.765 [0.592, 0.894]	0.792 [0.588, 0.950]	0.739 [0.538, 0.923]
Liver	**PyRadiomics**	**LGBM**	**0.891 (± 0.051)**	**0.859 [0.729, 0.970]**	**0.906**	**0.805 [0.647, 0.927]**	**0.789 [0.600, 0.947]**	**0.842 [0.667, 1.000]**
Lipo	**MIRP**	**LGBM**	**0.920 (± 0.033)**	**0.886 [0.712, 1.000]**	**0.890**	**0.686 [0.421, 0.880]**	**0.730 [0.444, 1.000]**	**0.666 [0.400, 0.917]**
Lipo	PyRadiomics	LGBM	0.891 (± 0.024)	0.909 [0.746, 1.000]	0.759	0.793 [0.533, 0.960]	0.728 [0.429, 1.000]	0.919 [0.727, 1.000]
Melanoma	**MIRP**	**LGBM**	**0.777 (± 0.048)**	**0.611 [0.318, 0.881]**	**0.973**	**0.987 [0.957, 1.000]**	**0.974 [0.917, 1.000]**	**1.000 [1.000, 1.000]**
Melanoma	PyRadiomics	TabNet	0.727 (± 0.082)	0.622 [0.333, 0.885]	0.395	0.760 [0.652, 0.852]	0.973 [0.913, 1.000]	0.420 [0.267, 0.581]
LIDC-IDRI	**MIRP**	**LGBM**	**0.804 (± 0.074)**	**0.705 [0.421, 0.950]**	**0.936**	**0.732 [0.500, 0.903]**	**0.622 [0.350, 0.846]**	**0.863 [0.500, 1.000]**
LIDC-IDRI	PyRadiomics	XGBoost	0.776 (± 0.056)	0.750 [0.500, 0.960]	0.949	0.820 [0.667, 0.944]	0.876 [0.687, 1.000]	0.429 [0.000, 0.833]

5.1.4 RPTK Outperforms Current State of the Art Methods

To assess its relative performance, RPTK was compared with the automated radiomics framework AutoRadiomics, a deep learning model trained on the same data splits, and results from previously published studies. All approaches were evaluated on synchronized dataset partitions to ensure a consistent comparison (see Section 4.1.9 for details). Performance was evaluated on both validation and test sets using AUROC as the primary metric, supplemented by confidence intervals to illustrate performance variability.

Figure 5.6 shows the validation performance across datasets for RPTK, AutoRadiomics, and the deep learning models. RPTK consistently achieved higher validation AUROC scores than AutoRadiomics, while deep learning models frequently exhibited much higher validation than test performance, indicating overfitting to the training data. This effect is particularly visible for datasets with limited sample size, where the variance in validation AUROC is large. In contrast, RPTK demonstrates sta-

ble and reproducible performance across folds, reflecting its effective regularization and feature-selection strategy. AutoRadiomics validation AUROC performance for LIDC-IDRI goes below the 0.5 AUROC.

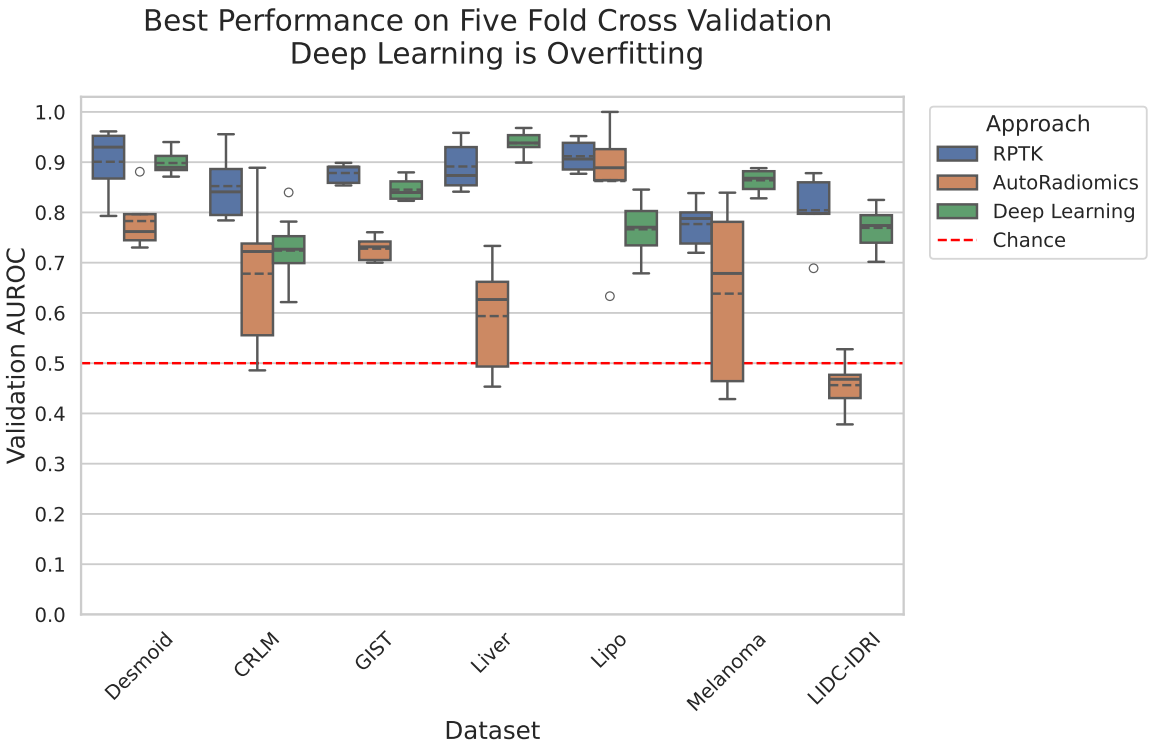


Figure 5.6. Validation AUROC across datasets using five-fold cross-validation for the best-performing models from RPTK (see Table 5.1), AutoRadiomics, and the trained deep learning model. The horizintal dotted line inside the boxplot refers to the mean whereas the solid line refers to the median. RPTK consistently outperforms AutoRadiomics, while the deep learning approach shows signs of overfitting, achieving higher validation AUROC but poor generalization on test data (see Figure 5.7). The red dotted line shows the threshold for performance with random guessing (0.5 AUROC).

Test performance is shown in Figure 5.7, where the 95% confidence intervals (CI) illustrate variability due to small dataset sizes. RPTK achieves the highest or near-highest test AUROC values across most datasets, outperforming AutoRadiomics in all but one case (Lipo) and clearly exceeding the deep learning models on every dataset. While deep learning occasionally surpasses AutoRadiomics on individual datasets (Liver and LIDC-IDRI), three deep learning models perform worse than random classification (AUROC < 0.5), further highlighting the limited generalization ability of these models in small, heterogeneous datasets. AutoRadiomics performs equal or worse then random guessing (Test AUROC $= 0.5$) on the Melanoma and the

LIDC-IDRI datasets.

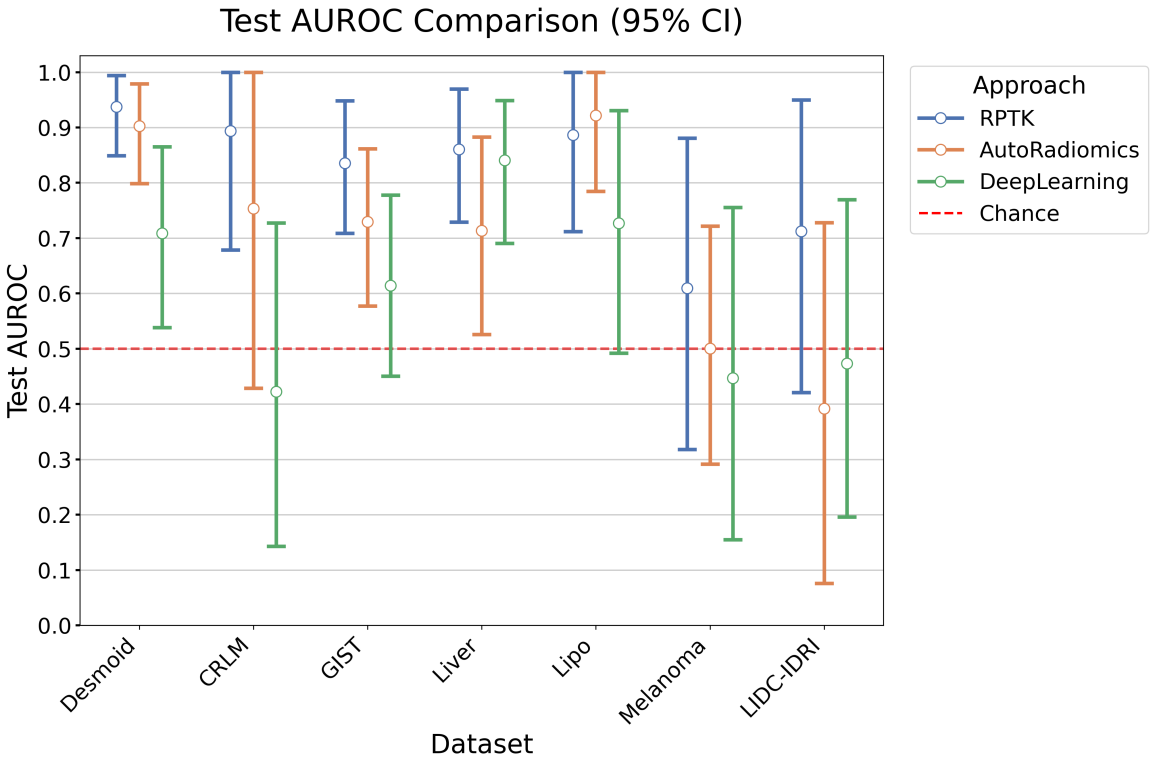


Figure 5.7. Test AUROC comparison of the best models from RPTK, AutoRadiomics, and deep learning approaches, with 95% confidence intervals. RPTK consistently achieves higher test AUROC across datasets and demonstrates more stable generalization compared to AutoRadiomics and deep learning. The red dotted line shows the performance by random guessing (AUROC = 0.5).

Finally, RPTK was benchmarked against results from the literature, including classical radiomics studies, deep learning publications, and radiologist assessments (Figure 5.8). Across all datasets, RPTK ranked among the best-performing frameworks and often achieved the highest AUROC values. In addition to outperforming AutoRadiomics and most deep learning approaches, RPTK also exceeded or matched the diagnostic accuracy reported for radiologists performing the same classification tasks. The distance of the Test AUROC performance compared to other approaches is especially visible on the Melanoma and the LIDC-IDRI datasets. This demonstrates the framework's potential to achieve expert-level or superior performance in quantitative image analysis.

In summary, RPTK demonstrates robust and generalizable performance across heterogeneous datasets, outperforming existing automated radiomics frameworks and deep learning models.

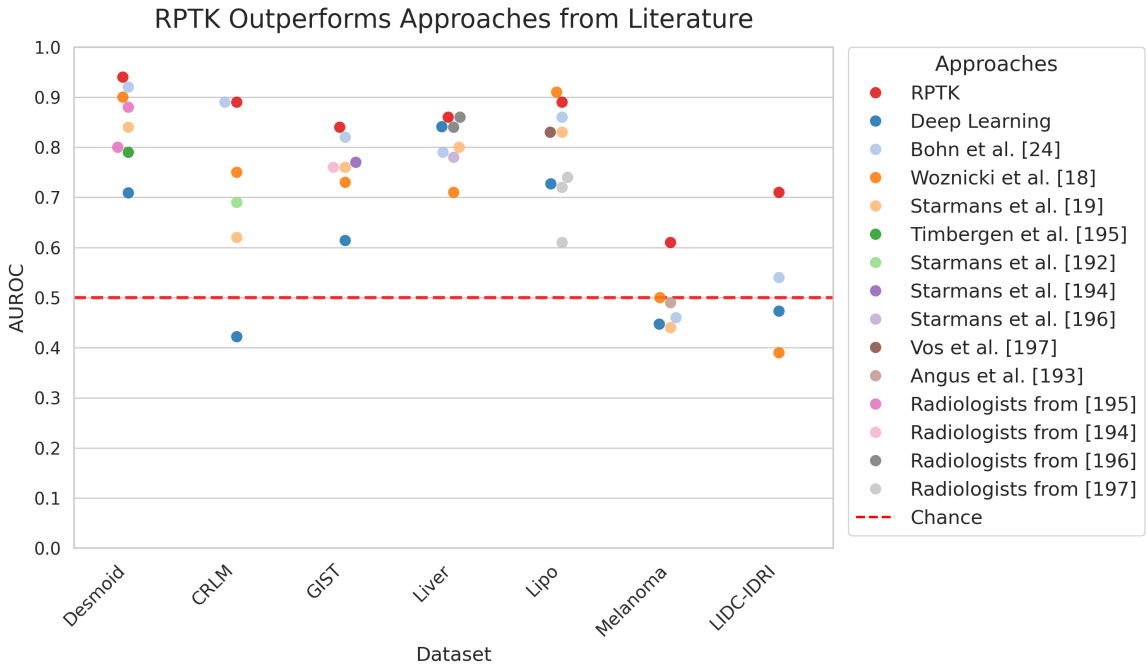


Figure 5.8. Comparison of RPTK test AUROC with published results from the literature, including radiomics frameworks, deep learning models, and radiologist assessments. RPTK consistently ranks among the top-performing approaches across datasets, surpassing human-level performance in several tasks. The red dashed horizontal line indicates random classification performance (AUROC = 0.5).

5.2 Predict Study – Predicting Immunotherapy Treatment Response in Lung Cancer Patients

This part of my thesis reflects one of two clinical applications where I used RPTK for applying the framework in a non open source context on real world clinical data how they would be used in the clinics to tackle the problem at hand. This section of my thesis is about the longitudinal prediction of early Immunotherapy response from advanced stage lung cancer patients treated at the thorax clinic Heidelberg. The image acquisition as well as the acquisition of the clinical data was done by clinicians of the thorax clinic Heidelberg. My contribution starts at the data curation and includes the application of an pretrained nnU-net model for automated segmentation. The review of the automated segmentations as well as the correction of these were done by two radiologists of the thorax clinic. The sub-sequential analysis of the data, the application of my tool (RPTK) as well as the experimental design and the analysis of the results were done by me.

Here I apply RPTK on real-world, non-public, retrospective, longitudinal, radio-

logical and clinical data. The Predict Study investigates the longitudinal prediction of early immunotherapy response in patients with advanced-stage lung cancer treated at the Thoraxklinik Heidelberg. This section presents unpublished results.

Table 5.2. Imaging fingerprint characteristics of the *Predict* dataset. Median, mean, and standard deviation (std) values are shown for key image properties.

Parameter	Median	Mean	Std
Number of slices	280.5	294.16	46.18
Slice thickness (mm)	3.00	2.85	0.52
ROI size (voxels)	56 480.0	136 765.51	272 559.57
Number of connected components	1.00	1.78	2.12
Number of bins	50.00	52.45	14.58

To assess the imaging characteristics of the Predict dataset in relation to the previously analyzed open-source datasets, the data fingerprint data descriptive parameters (see Table 5.2). Compared to the open-source datasets, the Predict dataset has a relative big mean ROI size with 136,765 and a very high std. The number of bins and the number of connected components are within the distributions of the open-source datasets. The mean slice thickness for the Predict study is 3 mm with a small std. of 0.52. Consistent with the results presented in Section 5.1.1, the Predict dataset displays moderate variability in imaging parameters. Specifically, the slice thickness tends to be lower, reflecting higher-resolution clinical acquisition, while the number of slices is higher due to extended volumetric coverage. The number of bins, ROI size, and connected components remain close to the overall mean across datasets.

Having established the imaging characteristics and overall data quality of the Predict cohort, the following subsections present the radiomics-based analysis of longitudinal imaging data and clinical variables for early immunotherapy treatment response prediction.

5.2.1 Longitudinal Imaging Improves RPTK Predictive Performance

The Predict dataset was split into training and testing subsets, this splits were generated by AutoRadiomics beforehand and used in the RPTK framework for feature selection and model training (see Section 4.2.1). To determine the optimal use of longitudinal imaging data for treatment response prediction, RPTK was applied to three different dataset configurations:
(i) features extracted from baseline CT images only (T0)

(ii) features extracted from the first follow-up CT images after treatment initiation (T1)
(iii) delta radiomics features computed as the difference between T1 and T0 feature values according to Equation 4.1 in Section 4.2.3
Independent evaluation of all dataset configurations was performed on the same held-out test set.

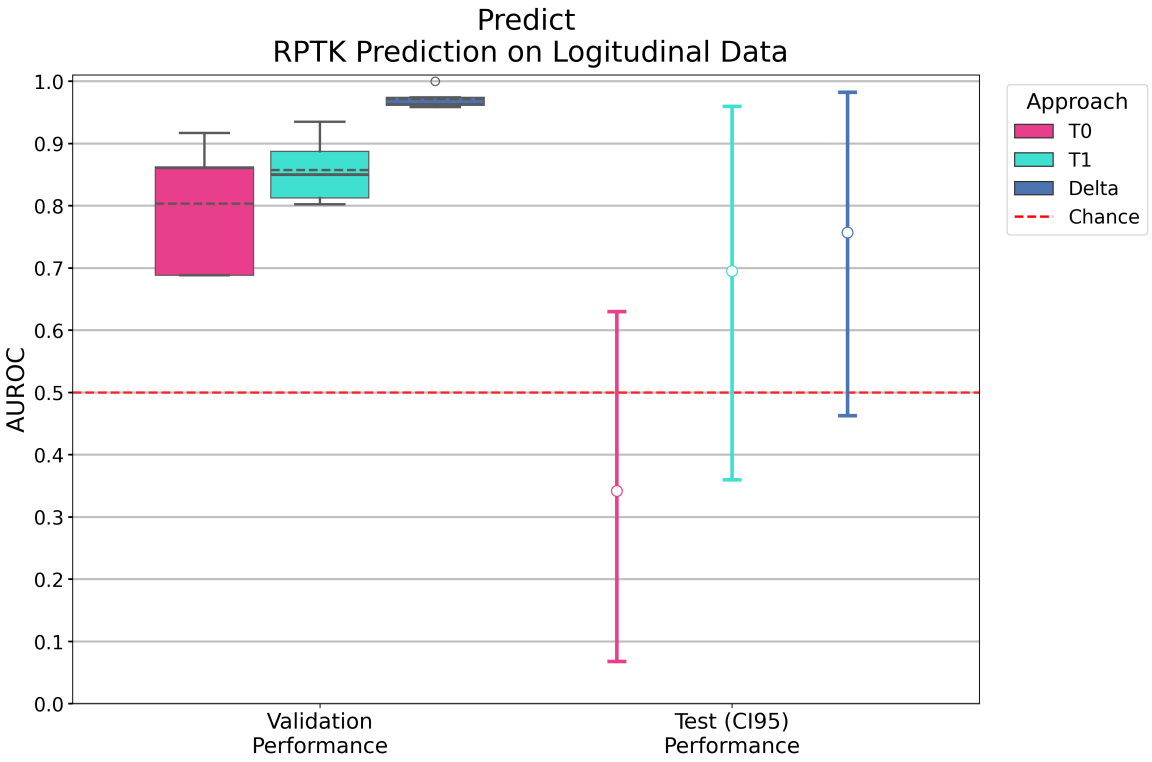


Figure 5.9. Comparison of RPTK performance across longitudinal data configurations for the Predict study. The boxplots on the left show validation AUROC values obtained from five-fold cross-validation with standard deviation, while the points with error bars on the right show test AUROC performance with 95% confidence intervals estimated from 1,000 bootstrap iterations. T0 corresponds to baseline CT scans, T1 to first follow-up scans, and Delta to radiomics features derived as the difference between T1 and T0 feature values (Equation 4.1). The horizontal dashed line within the boxplot represents the mean validation AUROC, and the solid line indicates the median validation AUROC. For performance values see Table 8.5. The red dashed horizontal line indicates random classification performance (AUROC = 0.5)

As shown in Figure 5.9, RPTK performance was compared across the three longitudinal configurations. The boxplots illustrate the distribution of validation AUROC values across cross-validation folds, while the markers with vertical error bars represent the test AUROC values together with the corresponding 95% confidence intervals

derived from bootstrap resampling. Across all configurations, validation AUROC values were consistently above the chance level. The delta configuration yielded the highest validation and test AUROC values, followed by the T1 and T0 configurations. Validation performance variability was lowest for the T1 and delta configurations, whereas T0 showed a wider spread. Test AUROC confidence intervals overlapped between T1 and delta, while T0 exhibited the largest uncertainty.

A complementary analysis including models based on clinical parameters and combined clinical–radiomics data is presented in Section 5.2.3.

5.2.2 RPTK Selects Important Features for Treatment Response Prediction

To characterize the image-derived biomarkers contributing to the predictive performance on the longitudinal Predict dataset, the selected delta radiomics feature spaces were examined for both extraction frameworks, PyRadiomics and MIRP. Figures 5.10 and 5.11 display heatmaps of the selected features, where each row corresponds to one selected feature and each column to one patient. Feature values are z-score normalized across patients, with higher and lower values shown in red and blue, respectively. Rows are grouped by IBSI feature class (color bar on the right), and patients are ordered by treatment response (non-responders to the left, responders to the right). This visualization allows the inspection of value distributions across feature classes and response groups.

The selected feature space from PyRadiomics extraction comprises 19 features. More than half of these belong to the Grey Level Distance Zone Matrix (GLDZM) and GLCM feature classes. With respect to image transformations, six features originate from different kernel combinations of 3D wavelet transformations, and one feature was derived from the peritumoral margin. Repeated occurrences of the GLDZM feature Low dependence low grey level emphasis (LDLGE) were observed under different image transformations. Among the texture features, the GLCM Inverse Difference Moment Normalized (IDMN) and the first-order Root Mean Squared feature showed distinct value patterns between the treatment response groups.

The MIRP-based feature space comprises 20 selected features. The most frequently represented feature classes are GLSZM, GLCM, and Intensity histogram (IH). Approximately half of the features were computed from 3D wavelet transformations. One feature originated from the peritumoral margin, eight were computed directly in 3D, and five were derived from 2D slice-wise computations averaged across the region of interest. The Neighbourhood Grey Tone Difference Matrix (NGTDM) Complex-

ity and Zone Size Entropy (ZSEntr) features appeared multiple times across distinct image transformations or dimensions. Pronounced value variations between response groups can be observed for the GLCM IDMN (Gabor-transformed) and GLRLM Long Run High Grey level Emphasis (LRHGE) features.

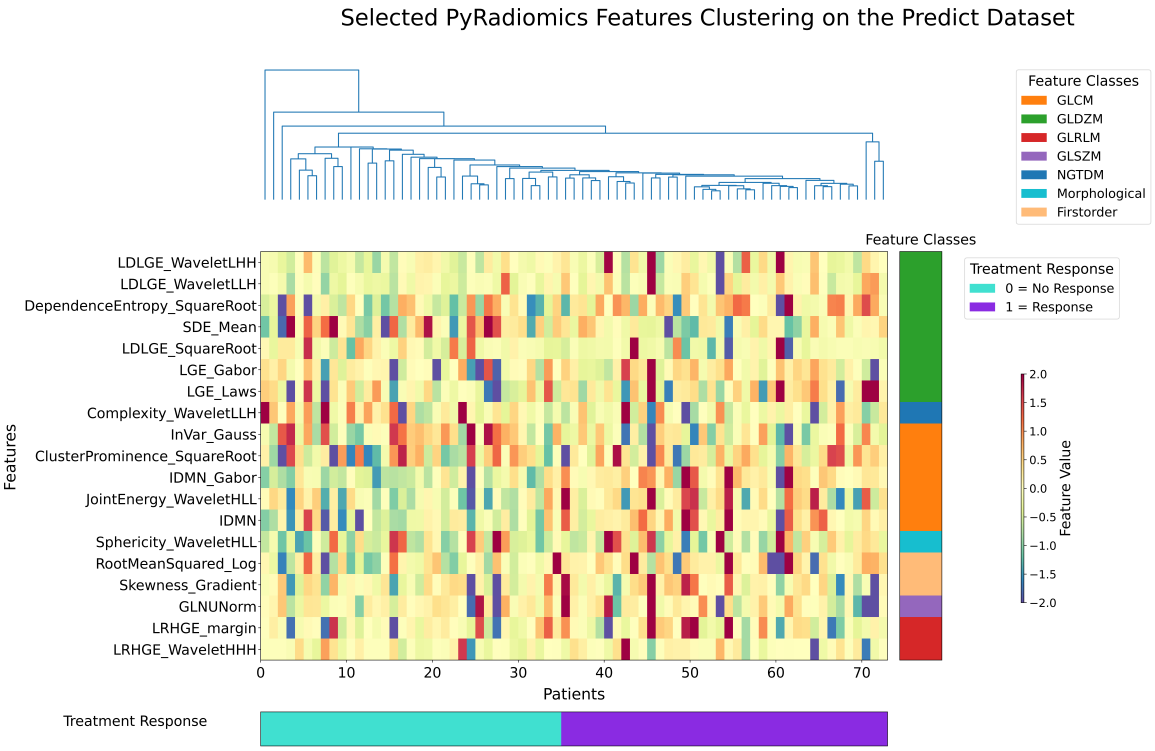


Figure 5.10. Selected delta radiomics features (PyRadiomics) from RPTK. Rows correspond to features ordered by IBSI class (color bar at right), and columns correspond to patients ordered by treatment response. Cell colors represent z-score normalized feature values. The plot visualizes the distribution of selected features across feature classes and treatment response groups.

Several feature types occur in both PyRadiomics and MIRP extractions. Both extraction frameworks included the GLCM IDMN feature and the morphological descriptor Sphericity. Differences between the two frameworks are primarily related to implementation: PyRadiomics computes most texture features (GLCM, GLSZM, GLDZM, NGTDM, GLRLM) in 3D, whereas MIRP includes additional 2D implementations. Furthermore, MIRP provides features from the Neighbouring Grey Level Dependence Matrix (NGLDM) class, which is not available in PyRadiomics (see Figure 8.7 in Section 8.2.2).

In addition to the RPTK feature selection, I analyzed the feature selection generated by AutoRadiomics for comparison (see Figure 8.20). The heatmap of the selected features from AutoRadiomics displays a total of ten selected features, all derived from

wavelet-transformed images. Only two IBSI feature classes are represented, namely GLDZM and GLRLM. Within these classes, features such as Large Dependence Emphasis and Run Variance appear multiple times under different wavelet filter combinations. Despite the limited diversity of selected feature types, the AutoRadiomics feature matrix shows a recognizable separation between responders and non-responders, indicating distinct value distributions between treatment response groups.

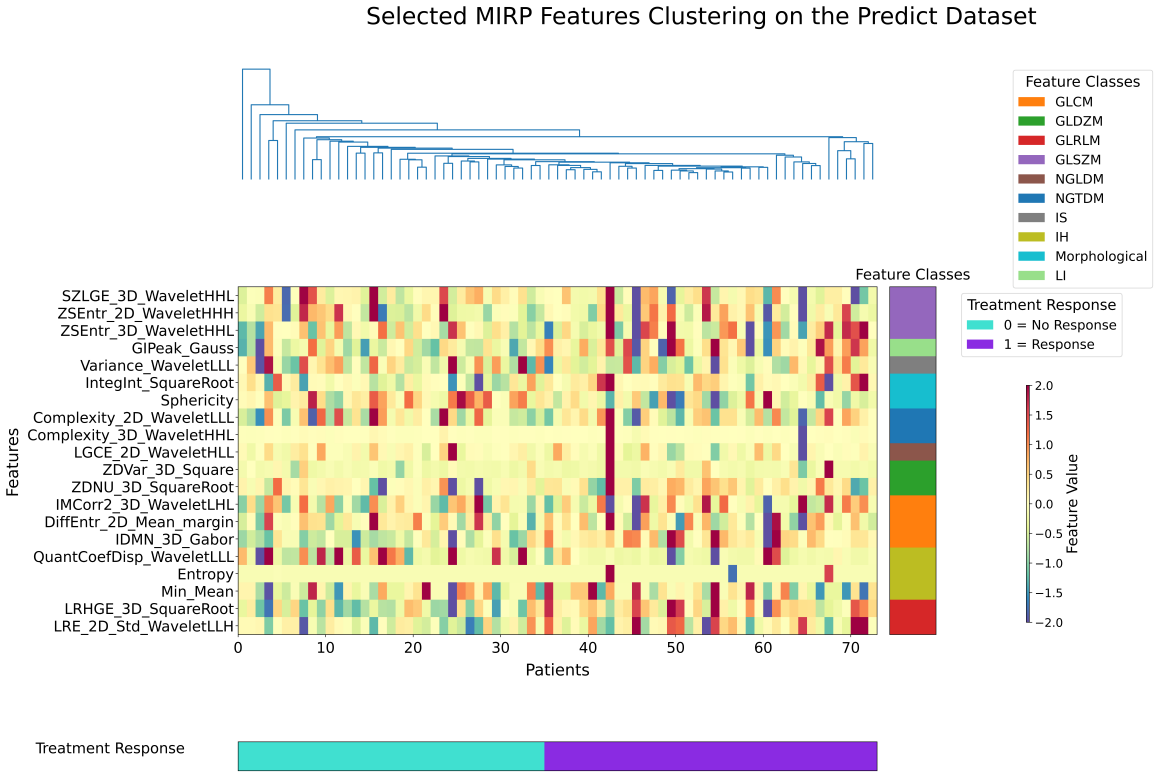


Figure 5.11. Selected delta radiomics features (MIRP) from RPTK. Rows correspond to features ordered by IBSI class (color bar at right), and columns correspond to patients ordered by treatment response. Cell colors represent z-score normalized feature values. The plot visualizes the distribution of selected features across feature classes and treatment response groups.

No direct overlap in specific features was observed between the AutoRadiomics and RPTK-selected feature spaces. However, both methods emphasize similar texture-based feature classes, with GLDZM and GLRLM being highly represented across both approaches. This overlap at the feature class level suggests that these texture families consistently contribute to treatment response modeling, regardless of the feature selection strategy applied.

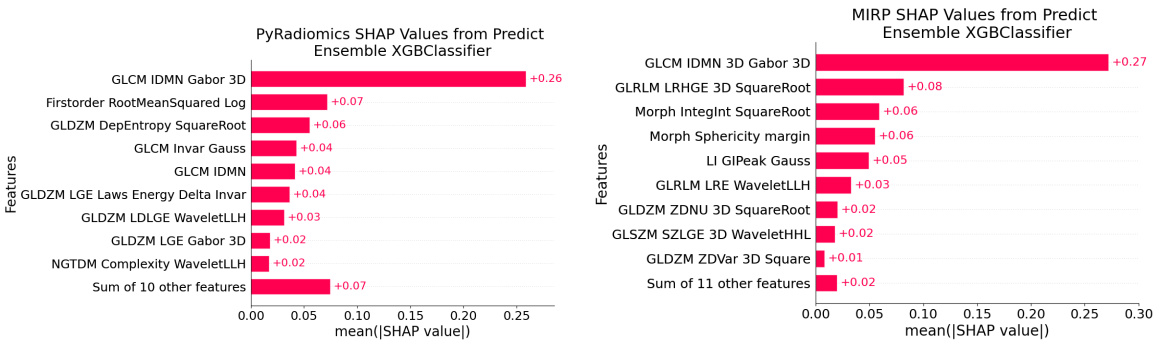
The selected features from RPTK on the Predict dataset correspond to those ranked as most important by the best selected predictive models (XGBoost model fro MIRP and PyRadiomics) in subsequent SHAP analyses (see Figure 5.12). Features

displaying strong value differentiation between response groups in the heatmaps, such as GLCM IDMN and GLRLM LRHGE, are among the top-ranking features in model-based importance estimates.

As shown in Figure 5.12, the SHAP value summary plots visualize the relative contribution of each selected feature to the output of the best-performing RPTK models for both extraction frameworks. The bar lengths correspond to the mean absolute SHAP value, representing the average feature impact on model predictions across all training samples.

For the PyRadiomics-based model (Figure 5.12a), feature SHAP values are distributed across multiple feature classes, with the GLCM IDMN feature extracted from the Gabor-transformed image showing the highest SHAP value. Other features with noticeable contributions include first-order (Root Mean Square) and texture-based features from the GLDZM class (Dependence Entropy (DepEntropy), LDLGE), as well as additional GLCM descriptors.

The MIRP-based model (Figure 5.12b) exhibits a similar distribution of feature importance, with the GLCM IDMN feature from the Gabor-transformed image again showing the highest contribution. Additional features with comparatively high SHAP values belong to the GLRLM (LRHGE), morphological (Sphericity, Integrated intensity (IntegInt)), and GLDZM feature classes.



(a) SHAP values summary bar plot of best model from RPTK based on PyRadiomics features of the Predict dataset.

(b) SHAP values summary bar plot of best model based from RPTK based on MIRP features of the Predict dataset.

Figure 5.12. SHAP values plots from best model on the Predict dataset to show feature impact on model performance. **a.** The SHAP values displaying the feature impact of the best performing model based on PyRadiomics features. **b.** The SHAP values displaying the feature impact of the best performing model based on MIRP features.

Across both extraction frameworks, the overall distribution of SHAP values shows

that a few dominant features contribute most strongly to the model predictions, while the remaining features have smaller but cumulative effects. The overlap in top-ranked features, particularly the GLCM IDMN descriptor, indicates that both extraction pipelines highlight similar feature types among the most influential predictors.

5.2.3 RPTK Gains Performance with Additional Clinical Information

In addition to imaging-based radiomics features, RPTK was extended to incorporate structured clinical data routinely collected in daily clinical practice. These parameters go beyond patient demographics and include disease-specific characteristics, such as tumor staging, laboratory values, and radiological assessments, reflecting both systemic and morphological aspects of disease status. Together, they provide complementary information to the imaging-derived radiomics features used in the Predict study.

To evaluate the contribution of these data sources, RPTK was trained and validated on three dataset configurations: (i) delta radiomics features only, (ii) clinical features only, and (iii) a combined dataset containing both clinical and delta radiomics parameters. All models were trained and evaluated on identical training and test partitions to ensure comparability. Table 5.3 summarizes the validation and test performance of the best-performing models for each configuration, including the mean validation AUROC, test AUROC with 95% confidence intervals, and threshold-based metrics (F1-score, Sensitivity, and Specificity) after Youden correction.

Table 5.3. Performance metrics across datasets, feature extractors, and models on the Predict dataset. The metrices on the test set like AUROC, F1, Sensitivity and Specificity are represented as mean and CI95 range. Displayed threshold dependent matrices (F1, Sensitivity and Specificity) have been optimized via Youden beforehand.

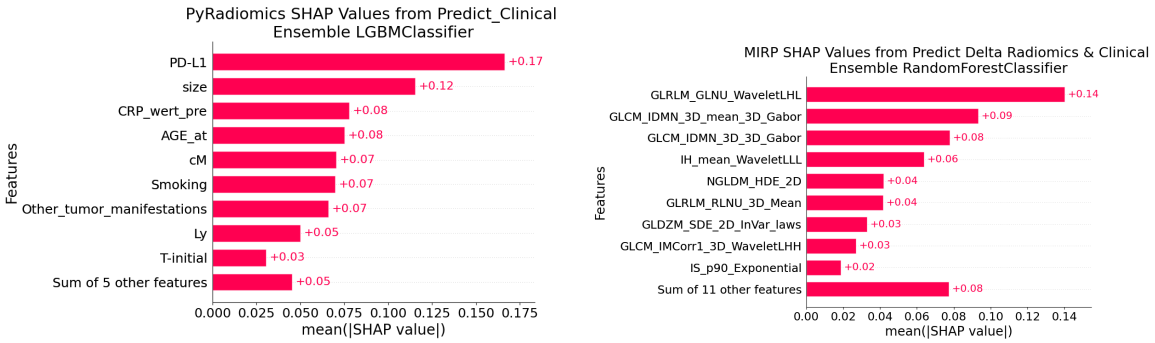
Predict data	Extractor	Model	Val AUROC	Test AUROC	Youden	Test F1	Test Sensitivity	Test Specificity
Delta Radiomics	MIRP	XGBoost	0.971 (+/- 0.017)	0.750 [0.536, 0.982]	0.798	0.647 [0.286, 0.889]	0.621 [0.222, 0.900]	0.707 [0.333, 1.000]
Clinical	Clinical	LGBM	0.817 (+/- 0.085)	0.786 [0.518, 1.0]	0.967	0.731 [0.429, 0.941]	0.740 [0.400, 1.000]	0.710 [0.333, 1.000]
Clinical and Delta Radiomics	Clinical and MIRP	Random Forest	0.948 (+/- 0.055)	0.767 [0.464, 0.964]	0.833	0.743 [0.461, 0.941]	0.760 [0.429, 1.000]	0.712 [0.333, 1.000]

As shown in Table 5.3, the performance of RPTK was evaluated across three configurations: delta radiomics features only, clinical features only, and a combined dataset containing both feature types. All models achieved validation AUROC values

above 0.80, with test AUROC values ranging from 0.75 to 0.79. The model trained on delta radiomics features reached the highest validation AUROC (0.97), followed by the combined configuration (0.95) and the clinical model (0.82).

In the test evaluation, the combined configuration achieved the highest F1-score (0.74) together with the highest sensitivity (0.76) and specificity (0.71). The clinical model showed comparable AUROC but slightly lower F1 and sensitivity values, while the delta radiomics model reached the lowest threshold-based metrics. Across all configurations, the 95% confidence intervals of the test AUROC overlapped, suggesting similar performance ranges.

Model interpretability based on Shapley Additive exPlanations (SHAP) values is illustrated in Figures 5.13a and 5.13b, which display the ranked feature contributions for the clinical-only and combined models, respectively. The bar lengths correspond to the mean absolute SHAP values, representing the average impact of each feature on the model output.



(a) SHAP values summary bar plot of best ensemble model based on clinical features (see Table 8.4).

(b) SHAP values summary bar plot of best ensemble model based on clinical and delta radiomics features.

Figure 5.13. SHAP values plots from best model on the Predict dataset to show feature impact on model performance on clinical and delta radiomics features. **a.** The SHAP values displaying the feature impact of the best performing model based on clinical features (see Table 8.4). **b.** The SHAP values displaying the feature impact of the best performing model based on clinical and delta radiomics features.

In the clinical-only model (Figure 5.13a), PD-L1 expression showed the highest contribution, followed by tumor size, CRP level, patient age, and metastatic stage (cM). Additional clinical factors, including smoking status, presence of other tumor manifestations, and lymphatic invasion (Ly), also contributed to the model predictions but with lower mean SHAP values.

In the combined clinical–radiomics model (Figure 5.13b), radiomics features were

among the top-ranked predictors. The most prominent contributors included texture-based descriptors such as GLRLM Grey level non-uniformity (GLNU), GLCM IDMN, and GLRLM LRHGE. Clinical parameters such as ECOG performance status and smoking status were also present in the overall feature set but were ranked below the top ten by mean SHAP value.

To further examine the selected features of the clinical-only and combined models, the corresponding feature value distributions are visualized in Figures 8.19 and 8.18 in the Appendix. Each heatmap displays the z-score–normalized feature values for all patients (columns) and selected features (rows). Patients are ordered by treatment response (non-responders to the left, responders to the right). In both figures, the color scale indicates relative feature intensity, with higher and lower values represented by red and blue, respectively.

The clinical feature heatmap (Figure 8.19) shows that several variables, including PD-L1 expression, tumor size, and CRP levels, exhibit visible value shifts between response groups. Features such as ECOG performance status, age, and smoking status show weaker but consistent variation across patients, suggesting a heterogeneous contribution of clinical parameters to treatment response representation.

The combined clinical–radiomics heatmap (Figure 8.18) includes both MIRP-derived radiomics descriptors and clinical variables. Radiomics features from texture-based classes (GLCM, GLRLM, GLSZM) dominate the upper rows of the heatmap, displaying structured intensity differences between non-responders and responders. Among the clinical variables, ECOG performance status, smoking status, and pack-years are visible within the lower section of the feature matrix. The distribution of values across both feature domains demonstrates that the selected clinical and radiomics features capture distinct but complementary signal patterns related to treatment response.

5.2.4 Clinical Potential and Decision Evaluation

To evaluate the clinical potential of the predictions generated by the delta radiomics model, I analyzed whether the predicted treatment response groups show a similar survival distribution to the ground truth response labels. The survival data used for this analysis were provided by the Thoraxklinik Heidelberg and were not part of the clinical data included in the clinical performance evaluation of RPTK. I performed the survival analysis using the ground truth and model-predicted classifications to assess their correspondence on a cohort level.

For this purpose, I calculated Kaplan–Meier survival curves for both the ground truth and the model-based classifications across the Predict cohort (see Figure 5.14).

The plot displays the survival distributions for responders and non-responders according to the true clinical labels and the classifications predicted by the delta radiomics model.

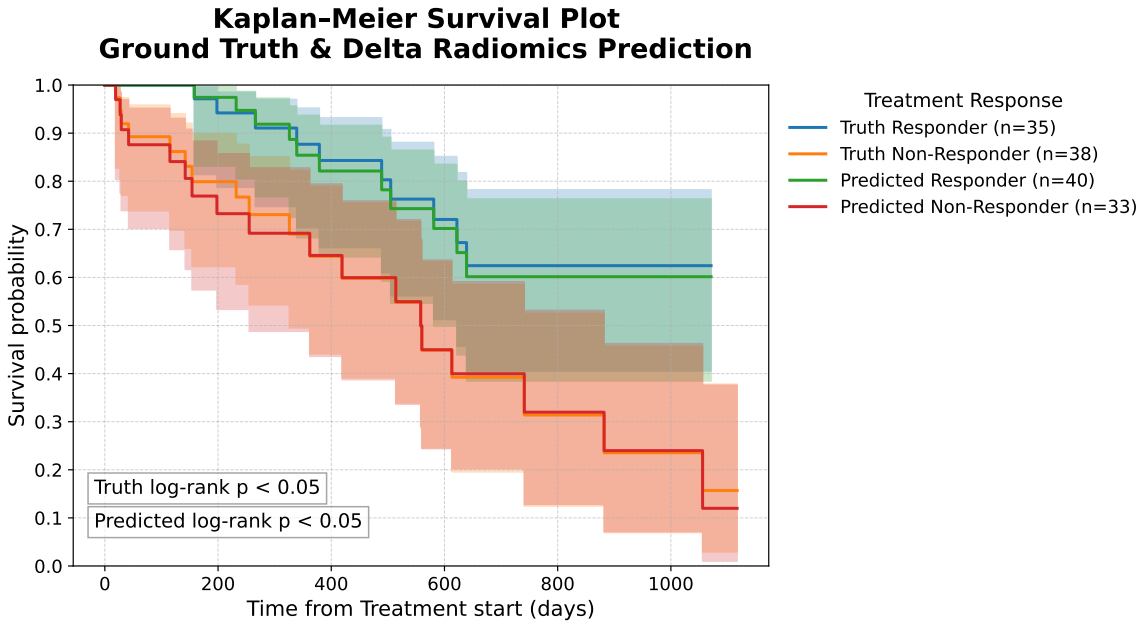


Figure 5.14. Kaplan–Meier survival curves comparing the ground truth (log-rank p-value = 0.0057) (see Figure 4.7) and delta radiomics–based predicted response (log-rank p-value = 0.0058) groups in the Predict cohort. The curves show the survival distributions for the responder and non-responder classes according to the true clinical labels and the model predictions in CI 95 distribution. The plot was generated by using the *KaplanMeierFitter* and the log-rank p-value was generated with alpha = 0.05 from the *lifelines* (v. 0.30.0) library (see Section 4.1.13).

Within the test set, I observed five misclassified patients, of which three were false positives (predicted as non-responders but were responders) and two were false negatives (predicted as responders but were non-responders) (see Figure 8.21b) showing very similar results to the clinical model performance (see Figure 8.21a).

When applying the trained delta radiomics model to the complete cohort, 40 patients were classified as responders and 33 as non-responders (see Figure 5.14). These numbers represent the model output across all available cases and are included here to provide an overview of the overall classification distribution, but the independent evaluation and performance assessment are based solely on the test set results. The corresponding Kaplan–Meier survival plot and confusion matrices summarize these classification outcomes.

5.2.5 Comparing RPTK Prediction Performance on Longitudinal Data

In order to compare the prediction performance of RPTK on longitudinal data to other approaches, I evaluated three methods using the same training and test splits of the Predict dataset: RPTK on delta radiomics features, AutoRadiomics on delta radiomics features, and a deep learning model trained on cropped CT images (see Section 4.2.3 for details). For comparability, I generated the delta radiomics features for AutoRadiomics in the same way as for RPTK by extracting features from both time-points and calculating the difference between the follow-up and baseline features per patient.

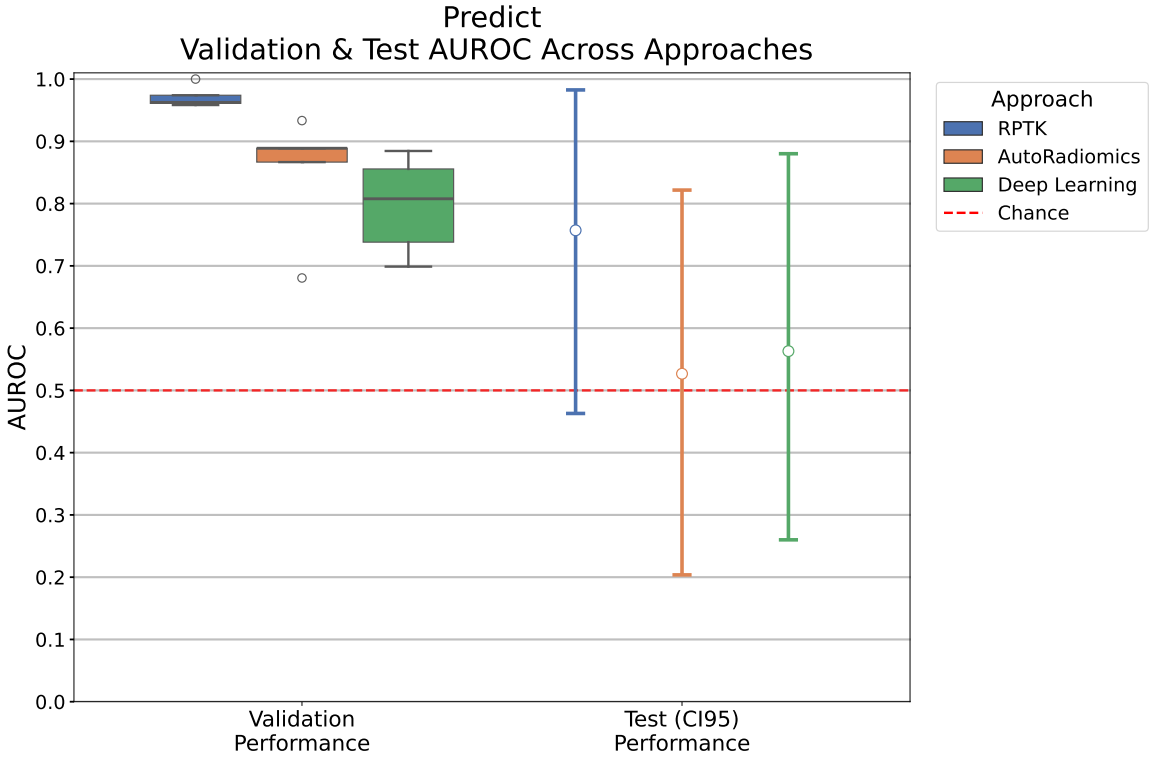


Figure 5.15. Validation and test AUROC performance of the Predict dataset obtained from three approaches: RPTK on delta radiomics, AutoRadiomics on delta radiomics, and Deep Learning on cropped CT images. The bars represent the mean validation AUROC, and the points indicate the test AUROC. All models were trained and evaluated on identical train/test splits. For performance values see Tables 8.5, 8.6, and 8.7.

Among all tested deep learning architectures, the ResNet-18 model achieved the best performance with a mean validation AUROC of 0.80 and a test AUROC of 0.55. For AutoRadiomics, the best-performing configuration was a random forest classifier

trained on delta radiomics features, reaching a validation AUROC of 0.88 and a test AUROC of 0.52. For RPTK, the best-performing model was an XGBoost classifier trained on delta radiomics features, achieving a validation AUROC of 0.97 and a test AUROC of 0.76. Figure 5.15 summarizes the validation AUROC and test AUROC performance for all three approaches on the Predict data.

5.3 LiverCRC Study – Colorectal Cancer Prediction via Liver CT

The results presented in this section are part of the LiverCRC project, which has been submitted as a manuscript entitled *Gut decisions based on the liver: A radiomics approach to boost colorectal cancer screening* [22]. I contributed to the project as a shared first author, being responsible for the data analysis, framework application, and evaluation of results. I performed the data preprocessing, radiomics feature extraction, and the subsequent analysis and interpretation of the results. The liver segmentation was performed by colleagues at the German Cancer Research Center (DKFZ), while data acquisition, pseudonymization, ethics approval, and data transfer were handled by the clinical partners at the University Medical Center Mannheim.

The results presented in this section were generated using synchronized experimental settings and data splits to ensure a fair comparison between RPTK, AutoRadiomics, and deep learning approaches. These experiments were performed independently of the submitted manuscript and are therefore unpublished. While the manuscript primarily focused on the feasibility and clinical implications of liver-based radiomics for colorectal cancer screening, the results shown here emphasize the methodological evaluation of RPTK under harmonized and reproducible conditions.

This section presents the second clinical application of my developed framework, RPTK. The study was designed to investigate whether liver-derived radiomics features can provide non-invasive biomarkers for colorectal neoplasia by exploiting biological relationships along the gut–liver axis. The approach also addresses current limitations of direct colorectal lesion segmentation, which is often constrained by the limited resolution of routine CT imaging and the anatomical variability of the colon. By focusing on the liver as a systemic organ, I aimed to identify radiomics patterns indirectly associated with colorectal tumorigenesis.

In the following subsections, I first present the radiomics features selected by RPTK from the liver segmentations, providing an overview of the feature composition and distribution across feature classes. Subsequently, I describe the performance of

the best-performing RPTK model and its evaluation on the independent test set. Finally, I compare the performance of RPTK to other automated radiomics frameworks and deep learning approaches to assess its relative performance and generalizability.

Table 5.4. Imaging fingerprint characteristics of the LiverCRC. Median, mean, and standard deviation (std) values are shown for key image properties. The mean and std values are copied from [22].

Parameter	Median	Mean	Std
Number of slices	99.0	111.27	53.26
Slice thickness (mm)	5.00	4.56	0.78
ROI size (voxels)	513 046.0	602 858.49	441 907.66
Number of connected components	1.00	1.43	2.14
Number of bins	12.00	12.89	6.53

5.3.1 The Selected Informative Features for Colorectal Noeplasia Prediction

The selected radiomics features provide insight into which image-derived characteristics contributed most to model performance. They reflect informative patterns within the liver that are associated with colorectal neoplasia and indicate how both intrahepatic and perihepatic signal variations influence classification. Examining these features helps to understand the image-based representation learned by RPTK and highlights which texture and intensity measures are most predictive within this cohort.

In the MIRP-based feature space, 19 features were selected (see Figure 5.17). Most of these belong to the GLCM, GLRLM, IH, and NGLDM feature classes. A large proportion of the selected features originate from 2D computations and margin-based extractions, indicating that both intrahepatic and perihepatic regions contributed relevant information to the classification. Recognizable clustering of feature values between patients with and without colorectal neoplasia is visible for the GLCM JointMax_Std feature extracted from the 2D space, as well as for the IH Median feature derived from the perihepatic region. Additional separation between neoplasia and non-neoplasia groups can be observed for features such as Joint Energy 2D, GLNU 2D Std from the margin, and High dependence emphasis (HDE) 3D from margin, where higher z-score–normalized values tend to correspond to neoplasia cases. In contrast, diagnostic and volume-related features, including V50_margin, V75_margin, Bounding Box Dimension Y, Low Grey Level Run Emphasis (LGLRE) 2D Mean from the margin, and Cluster Prominence 2D Mean, show less class-specific variation and

appear more uniformly distributed across the cohort.

The PyRadiomics-based model selected 15 features in total, with a higher representation of Diagnostics, GLRLM, Firstorder, and GLCM feature classes. A notable portion of the selected features originated from the perihepatic margin, such as Short runs emphasis (SRE), LRHGE, Dependence Non-Uniformity Normalized (DNUNorm), and Inter-quartile Range (IQR), indicating that textural heterogeneity around the liver boundary contributes to the classification of colorectal neoplasia. Among the texture-related features, Low Gray Level Zone Emphasis (LGZE) from the perihepatic region and Grey Value minimum feature show very few variance and less separation between patient groups, with increased normalized feature values in patients with colorectal neoplasia.

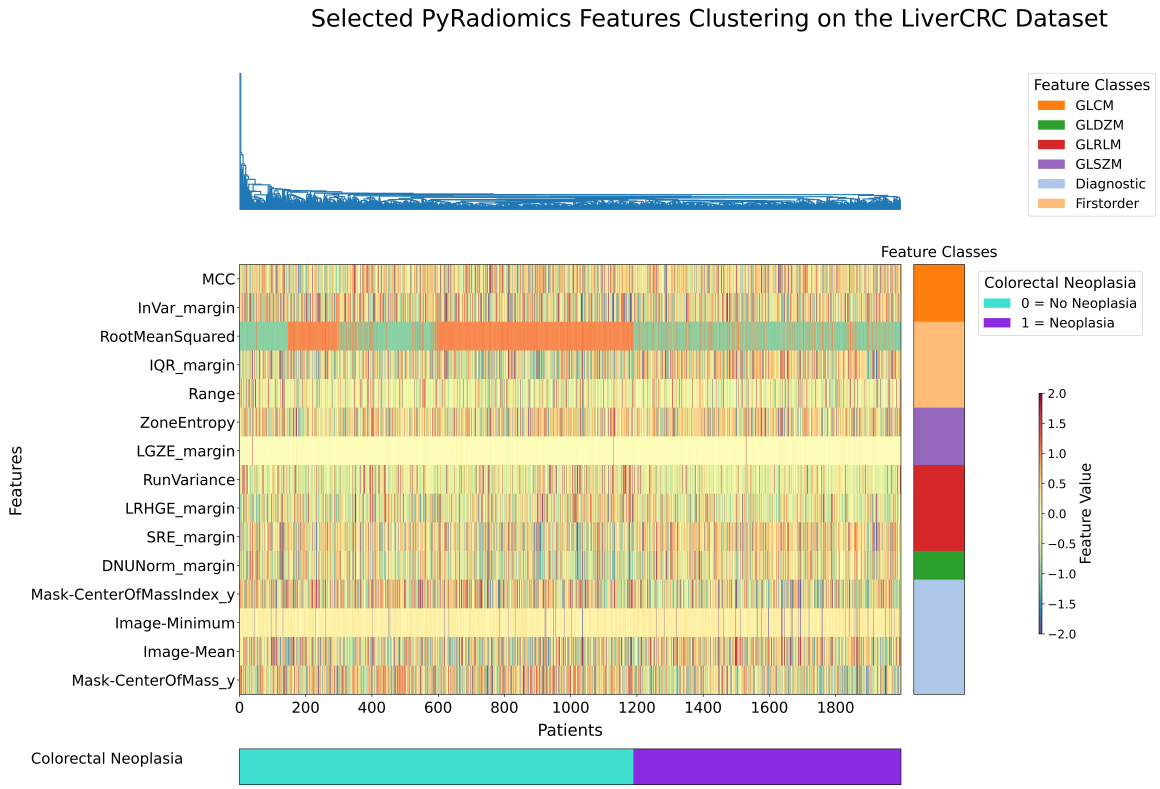


Figure 5.16. Heatmap of selected PyRadiomics features showing z-score–normalized feature values across patients of the LiverCRC cohort. Columns correspond to patients ordered by colorectal neoplasia label (non-neoplasia on the left, neoplasia on the right), and rows represent the selected features grouped by IBSI feature class (color bar on the right). Higher and lower normalized values are represented by warm and cool colors, respectively. The plot highlights clustering of the Firstorder Root Mean Squared feature and texture features such as GLCM Maximum Correlation Coefficient (MCC) between the two patient groups.

A pronounced clustering pattern is visible for the Firstorder Root Mean Squared

feature, which also achieved the highest SHAP importance in the model, underlining its strong impact on the overall prediction performance (see Figure 5.18). This feature reflects the overall magnitude of voxel intensities within the region of interest and may capture general differences in tissue composition or contrast uptake within the liver.

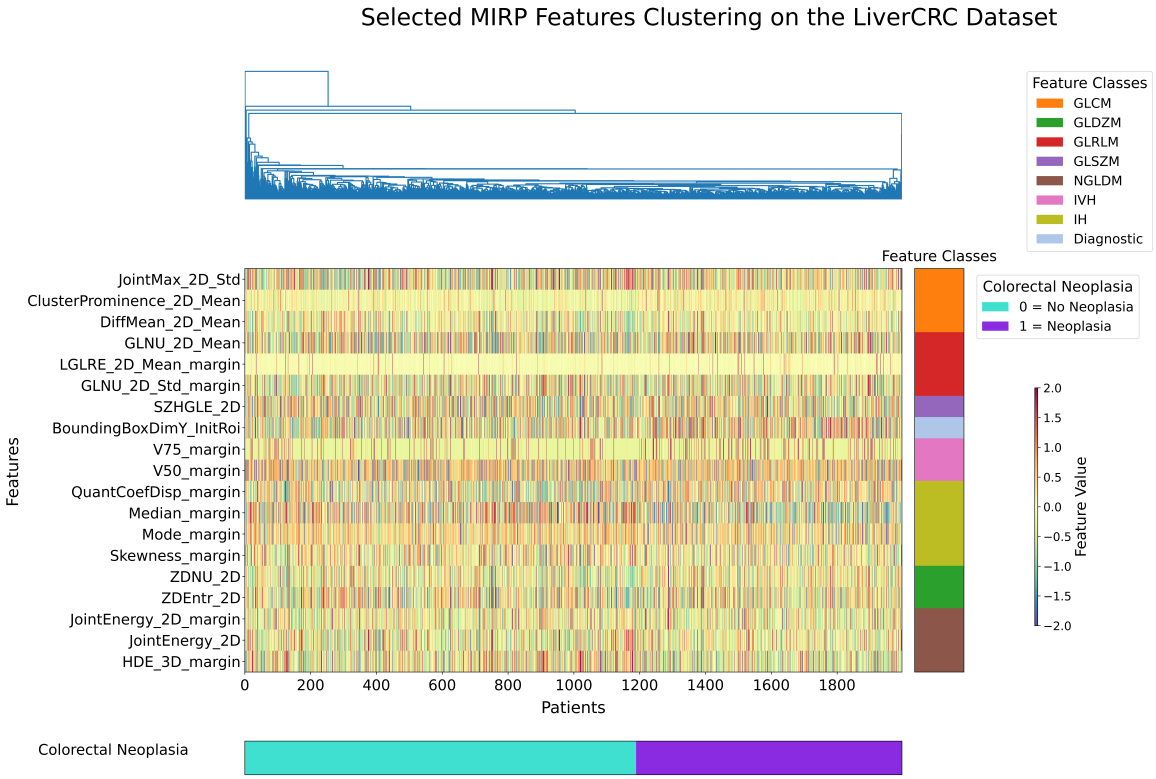
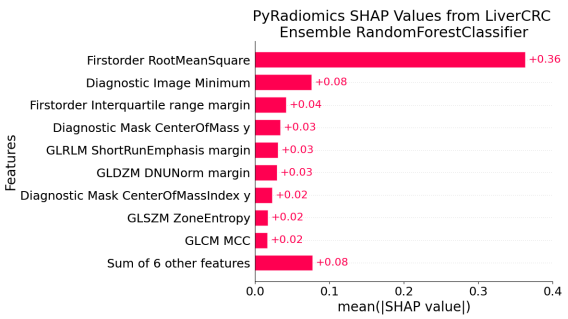


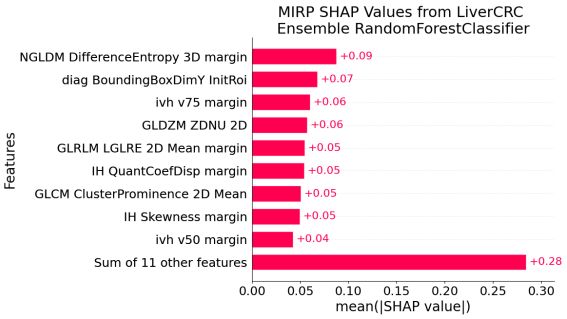
Figure 5.17. Heatmap of selected MIRP features showing z-score–normalized feature values across patients of the LiverCRC cohort. Columns represent patients ordered by colorectal neoplasia label (non-neoplasia on the left, neoplasia on the right), and rows represent selected features grouped by IBSI feature class (color bar on the right). Warm and cool colors indicate higher and lower normalized feature values, respectively. Visible clustering patterns can be observed for texture features such as GLCM JointMax_Std and IH Median, illustrating differences between neoplasia and non-neoplasia groups.

As a volume-confounded feature, its value partially depends on the size of the segmented region, which should be considered when comparing across subjects with heterogeneous liver volumes. Other first-order features, including Energy and Range, and texture features such as the GLCM Maximum Correlation Coefficient (MCC) and GLRLM Run-Entropy, further contributed to the model output with lower but consistent relevance scores. Additionally, several diagnostic features related to segmentation geometry and image characteristics, such as Mask-Center-of-Mass on the vertical dimension and Image-Maximum, were included in the selected feature space.

Comparing the feature spaces extracted by MIRP and PyRadiomics shows that both frameworks identified texture-related features, particularly from the GLCM and GLRLM classes, as most informative for the classification of colorectal neoplasia. However, differences in feature composition reflect the methodological design of each extractor. MIRP produced a higher proportion of 2D and margin-based features, emphasizing perihepatic and surface-related texture characteristics, whereas PyRadiomics selected a more diverse set of first-order and diagnostic features in addition to texture descriptors. PyRadiomics was able to capture clustering patterns between neoplasia and non-neoplasia cases, whereas MIRP features do not show such clear patterns.



(a) SHAP values summary bar plot of best model from RPTK based on PyRadiomics features from the LiverCRC dataset.



(b) SHAP values summary bar plot of best model from RPTK based on MIRP features from the LiverCRC dataset.

Figure 5.18. SHAP values plots from best model on the LiverCRC dataset to show feature impact on model performance. **a.** The SHAP values displaying the feature impact of the best performing model based on PyRadiomics features. **b.** The SHAP values displaying the feature impact of the best performing model based on MIRP features.

In addition to the RPTK-derived feature spaces, the feature selection from AutoRadiomics was analyzed on the LiverCRC dataset for comparison (see Figure 8.24). In contrast to RPTK, which was applied to features extracted from the original images, AutoRadiomics was used in its default configuration without modification to the extraction settings. AutoRadiomics selected 35 features, whereas RPTK based on PyRadiomics identified 15 features. The AutoRadiomics feature space is dominated by first-order statistics and GLSZM features, with the majority of features derived from image transformations such as wavelet or Laplacian-of-Gaussian filters. Only two features originate from the original image domain.

The AutoRadiomics heatmap shows that most selected features originate from first-order and the GLSZM classs, including RootMeanSquared, ZoneEntropy, and

GLNU. Features are repeated whereas they roiginate from different image transformatins (e.g. first-order features Minimum, Maximum, RootMeanSquared are included from wavelet and logarthmic image transformations). No distinct clustering structure is visible between patients with and without colorectal neoplasia, although localized intensity differences can be observed within subgroups of patients. Several features, such as RootMeanSquared, occur in both AutoRadiomics and RPTK selections, but AutoRadiomics selected the transformed versions (e.g., wavelet-based) rather than those from the original image. Overall, AutoRadiomics produced a broader feature set with greater emphasis on transformed intensity statistics, while RPTK selected a more compact and diverse set of features emphasizing both texture and margin-related descriptors.

5.3.2 RPTK Internal Model Performance Evaluation

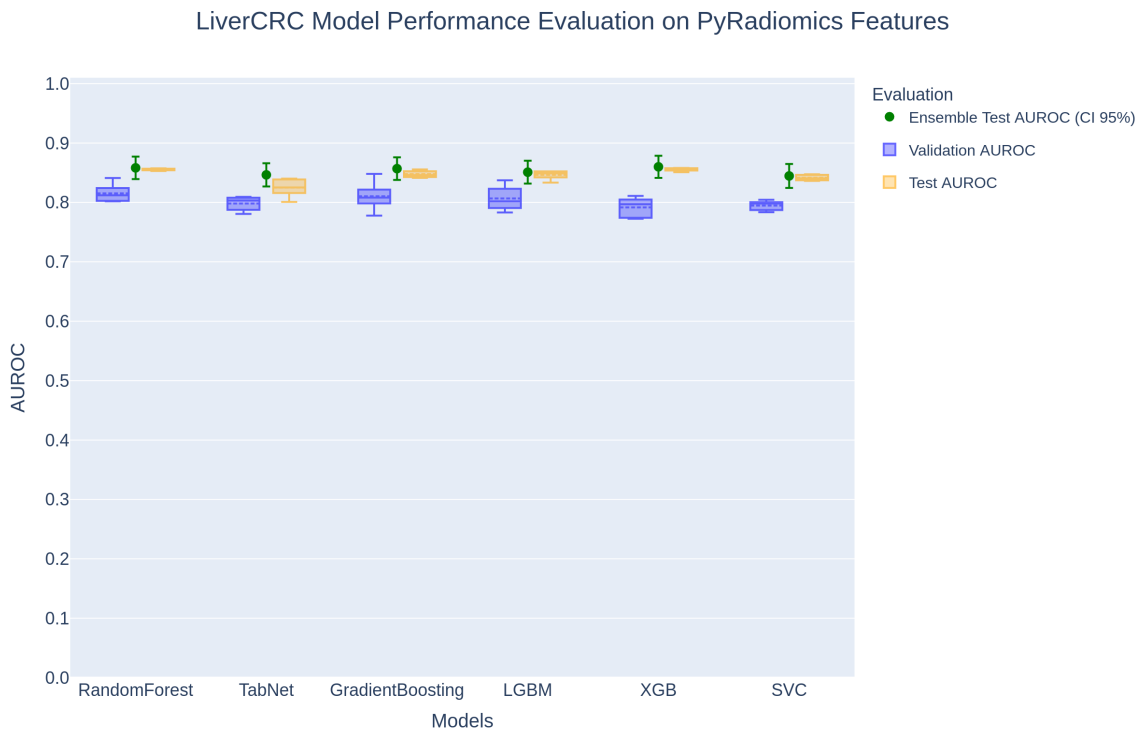


Figure 5.19. Summary of validation and test AUROC performance for all models trained on PyRadiomics-derived features. Boxplots represent validation and test AUROC distributions from five-fold cross-validation with one standard deviation, while the green dots and error bars indicate the ensemble model test AUROC performance and its 95% confidence interval. The boxplots represent the performance distribution of the five fold models (not ensembled). The horizontal dotted line in the boxplot refer to the mean, the solid line refer to the median.

This section summarizes the internal performance evaluation of the models trained with RPTK on the LiverCRC dataset using selected features extracted with PyRadiomics and MIRP shown in Section 5.3.1. Each plot displays the validation AUROC distributions obtained during five-fold cross-validation, the corresponding test AUROC values for each fold model, and the ensemble test AUROC with 95% confidence intervals based on bootstrap resampling of the ensemble of all fold models. This setup allows for a direct comparison of model generalization performance between folds, model types, and feature extraction configurations.

Table 5.5. Performance metrics across datasets, feature extractors, and models on the LiverCRC dataset. The metrices Test-AUROC, F1, Sensitivity and Specificity are represented as mean and CI95 range. Displayed threshold dependent matrices (F1, Sensitivity and Specificity) have been optimized via Youden beforehand.

Dataset	Extractor	Model	Val AUROC	Test AUROC	Youden	Test F1	Test Sensitivity	Test Specificity
LiverCRC	MIRP	XGBoost	0.662 (+/- 0.020)	0.638 [0.584, 0.689]	0.673	0.551 [0.490, 0.611]	0.584 [0.503, 0.659]	0.634 [0.571, 0.694]
LiverCRC	PyRadiomics	Random Forest	0.815 (+/- 0.016)	0.859 [0.819,0.893]	0.757	0.757 [0.704, 0.806]	0.780 [0.716, 0.844]	0.807 [0.756, 0.858]

Across all models and both feature extraction frameworks from RPTK, validation and test performances were relatively stable on the LiverCRC dataset, indicating consistent training behavior across folds (see Figure 5.19 and 5.20). Detailed values of the best RPTK model performance are integrated in Table 5.5, best performance of AutoRadiomics models on the LiverCRC data are integrated in Table 8.8 and the performance of the best deep learning model are in Table 8.9. However, clear differences were observed between the two extractors. Models trained on PyRadiomics features achieved higher AUROC values on both validation and test data compared to those trained on MIRP features. For PyRadiomics, mean validation AUROC values were consistently around 0.80, while mean test AUROC values ranged between 0.83 and 0.86 across the applied classifiers. In contrast, MIRP-based models reached validation AUROC values between 0.60 and 0.70 and slightly lower test values, with most models performing around 0.63.

For both extractors, the ensemble test performance closely matched the average test results from the individual fold models, as visualized by the overlapping 95% confidence intervals. Among the MIRP-based models, the XGBoost classifier achieved the highest validation performance (AUROC = 0.66), while for PyRadiomics, the random forest classifier achieved the highest validation and test performance (AUROC = 0.82 and 0.86, respectively; see Table 5.5). Threshold-based evaluation metrics followed

the same trend, with higher F1-scores, sensitivities, and specificities observed for the PyRadiomics-based configurations compared to those derived from MIRP features.

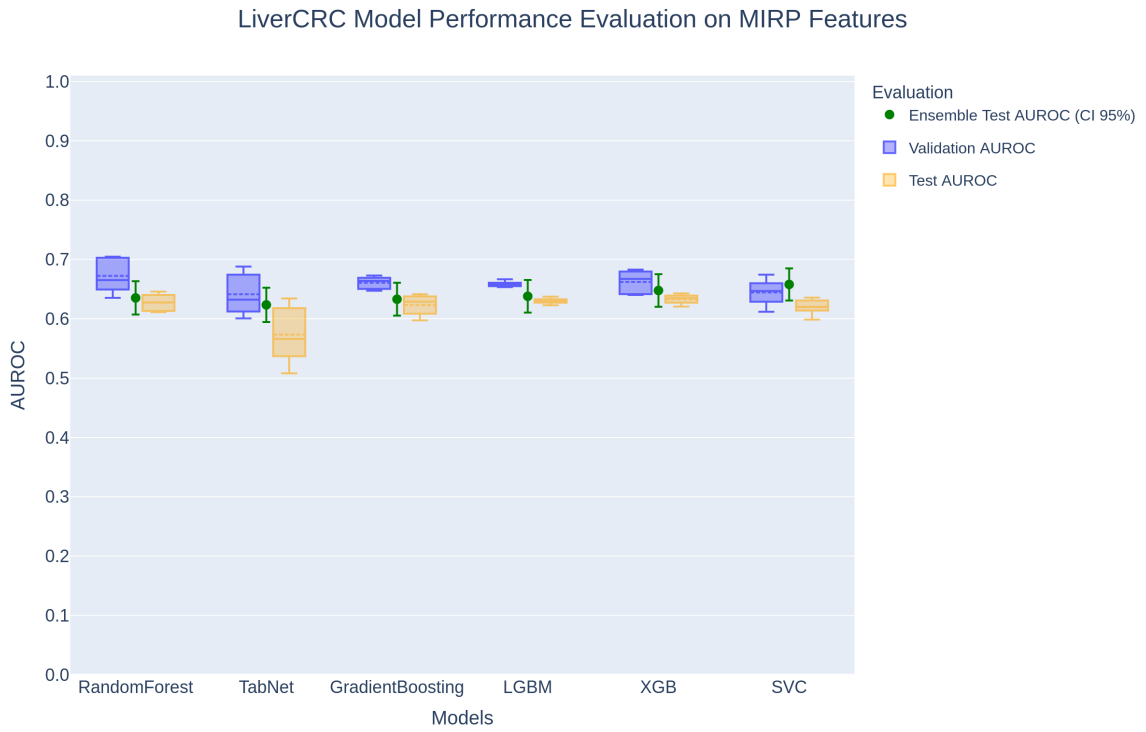


Figure 5.20. Summary of validation and test AUROC performance for all models trained on MIRP-derived features. Boxplots represent validation and test AUROC distributions from five-fold cross-validation with one standard deviation, while the green dots and error bars indicate the ensemble test AUROC and its 95% confidence interval. The boxplots represent the performance distribution of the five fold models (not ensembled). The horizontal dotted line in the boxplot refer to the mean, the solid line refer to the median.

5.3.3 RPTK Outperforms Other Tools Significantly on Larger Datasets

To evaluate the generalization performance of RPTK on large datasets, I collected the results and compared them against the best-performing AutoRadiomics configuration and the optimized deep learning model. Figure 5.21 summarizes validation and test AUROC values, including statistical comparisons based on the paired DeLong test (see Section 4.1.13).

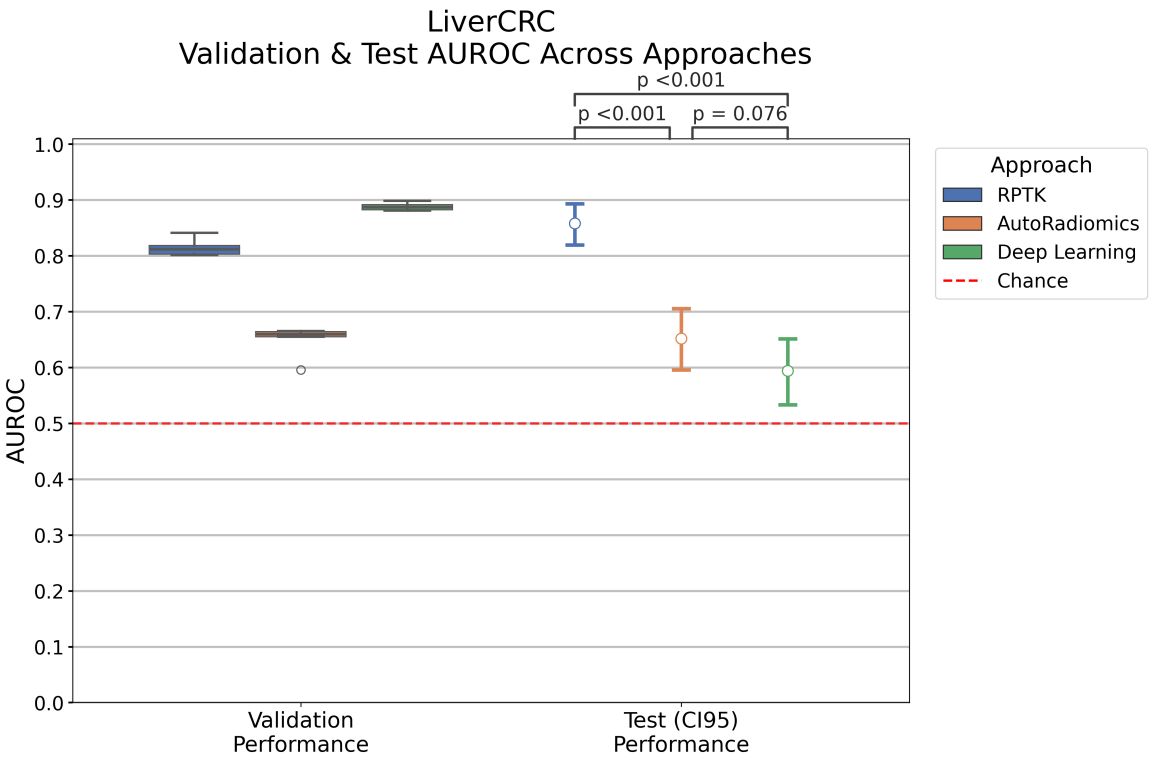


Figure 5.21. AUROC performance on validation folds (with standard deviation) and test set (with CI 95) of the LiverCRC dataset from the models out of the three approaches: RPTK, AutoRadiomcis and deep learning. The paired DeLong significance test was performed on the test predictions to compare the approaches (RPTK vs. AutoRadiomics p-value=$1.751 \cdot 10^{-13}$, RPTK vs. deep learning p-value= $4.441 \cdot 10^{-16}$). For detailed performance values on AutoRadiomics performance see Table 8.8 and for deep learning performance see Table 8.9.

The resulting p-values for comparing RPTK with AutoRadiomics on the test AUROC performance is $1.751 \cdot 10^{-13}$ and the p-value comparing between RPTK and deep learning equals $4.44 \cdot 10^{-16}$ whereas the p-value between the best deep learning model and the best AutoRadiomics model is not significant with a p-value of 0.106. Additionally, the corresponding Receiver Operating Characteristic curve (ROC) curves on the test set are shown in Figure 5.22.

Across all evaluated approaches, the RPTK framework achieved the highest test performance (Figure 5.21). The mean test AUROC reached 0.859 for RPTK using a Random Forest classifier, compared to 0.654 for AutoRadiomics (Random Forest) and 0.598 for the deep learning model (ResNet18). Validation AUROC values showed a compact distribution across folds for RPTK, AutoRadiomics and the deep learning approach. The validation–test performance difference was smallest for AutoRadiomics (ΔAUROC = 0.02) and RPTK (ΔAUROC = 0.04), while the deep learning model

showed the largest drop between validation and test AUROC (ΔAUROC = 0.35). All three approaches performed above random classification level (AUROC = 0.5).

The paired DeLong test confirmed statistically significant differences between RPTK and the other applied approaches, with $p < 0.001$ for RPTK versus AutoRadiomics and as well as for RPTK versus deep learning. The performance difference between AutoRadiomics and deep learning was not significant ($p = 0.078$). The test ROC curves shown in Figure 5.22 visualize these differences in predictive performance, where RPTK exhibits the steepest ascent and the highest area under the curve, followed by AutoRadiomics and deep learning.

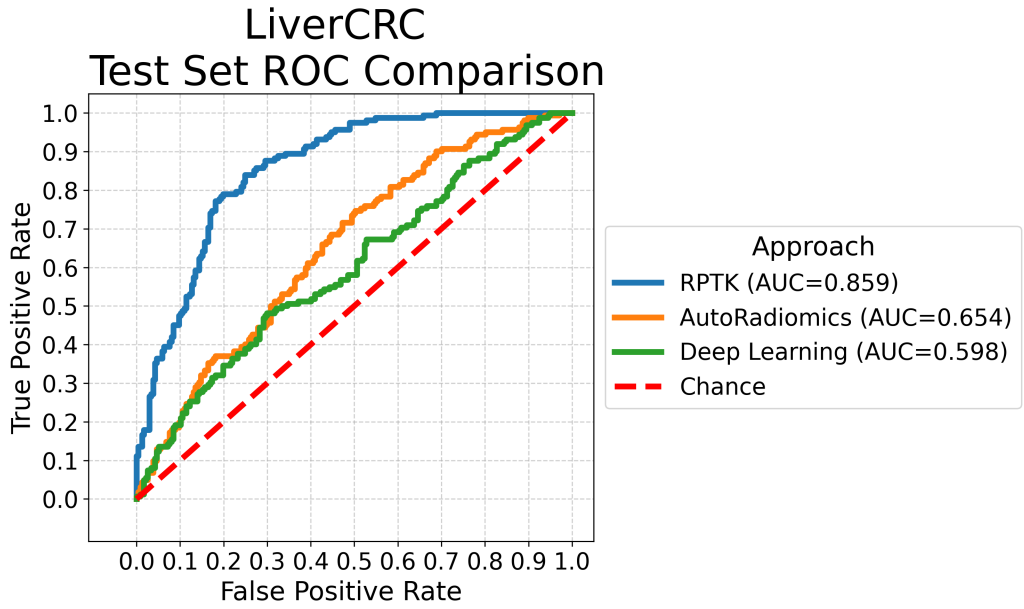


Figure 5.22. ROC on test set of the LiverCRC dataset from the best models out of the three approaches: RPTK, AutoRadiomcis and Deep Learning.

The corresponding Test ROC curves (Figure 5.22) illustrate the classification behavior of all approaches. For RPTK, a sensitivity of 0.90 was achieved at a false positive rate of 0.35. For AutoRadiomics and deep learning, sensitivities and specificities at their respective optimal thresholds were lower, consistent with overall AUROC values.

Chapter 6

Discussion

This thesis investigates the development, validation, and clinical application of a self-configuring radiomics pipeline for automated and reproducible quantitative imaging analysis. The work is structured around three major parts that together demonstrate the methodological advancements, generalization capacity, and translational potential of the proposed approach.

In the first part, the *Self-Configuring Radiomics Pipeline* introduced the RPTK, an open and modular framework enabling standardized and automated radiomics experiments. In the second part, the *Predict Study – Immunotherapy Response in Lung Cancer* illustrated how RPTK can be applied to longitudinal imaging and comprehensive clinical data to predict response to immunotherapy. The third part, the *LiverCRC Study – Colorectal Neoplasia Prediction via Liver CT*, explored the application of RPTK to large-scale cohorts and evaluates how imaging-derived features from the liver can reveal systemic disease patterns.

Across all studies, RPTK was systematically compared to AutoRadiomic, an automated radiomics pipeline, and to multiple deep learning models, highlighting differences in performance, interpretability, and generalization. Together, these investigations provide a comprehensive perspective on the potential and limitations of automated radiomics frameworks in medical imaging research.

6.1 Self-Configuring Radiomics Pipeline

6.1.1 The Datasets

The evaluation of the RPTK framework was based on seven open-source datasets originating from the WORC collection (CRLM, Melanoma, GIST, Desmoid, Liver, and Lipo) and the LIDC-IDRI dataset accessed via TCIA. Together, these datasets span

a broad spectrum of oncological tasks, imaging modalities, and acquisition protocols (see Table 4.1). Their heterogeneity provided a suitable benchmark for testing the self-configuring properties of RPTK, but also introduced several data-related limitations that affect generalization, interpretability, and statistical power.

The datasets included in the WORC database originate from multiple clinical studies conducted across several institutions, imaging centers, and scanner manufacturers, and were subsequently curated and standardized under the coordination of Erasmus MC [7]. This multi-institutional composition introduces substantial heterogeneity in acquisition protocols, reconstruction kernels, field strengths, and imaging vendors, further amplified by the diversity of underlying disease types. While such variability complicates preprocessing and harmonization, it provides an essential test basis for evaluating the robustness of self-configuring radiomics workflows such as RPTK. The inclusion of the LIDC-IDRI dataset further extends the level of heterogeneity, as it encompasses CT scans from numerous hospitals and scanner types worldwide. The combination of multi-center MRI and CT data from WORC and globally sourced thoracic CT data from LIDC-IDRI thus offers a broad spectrum of acquisition conditions. This diversity strengthens the assessment of the generalization capability of RPTK, demonstrating that its adaptive preprocessing and configuration principles can handle data from heterogeneous origins without manual adjustment.

While the combined dataset covers more than 1,000 patients, most individual cohorts are relatively small ($n < 250$), which restricts the effective training size for cross-validation and independent testing. Consequently, model evaluation suffers from limited statistical power, and small differences between methods do not reach significance. The wide confidence intervals observed in several test results (see Figure 5.7) reflect this uncertainty. Small sample sizes also increase the risk of data-partition bias, where individual cases can disproportionately influence validation results, particularly in heterogeneous imaging settings.

To overcome these sensitive data bias, I used the same data from training and testing in order to provide models which are trained and tested on the same data across the performed approaches (RPTK, AutoRadiomics, and Deep Learning; see Section 4.2.1). However, I used the splitting technique from AutoRadimics to synchronize the train and test sets but did not adapt the five fold cross validation algorithm of RPTK to the cross-validation technique from AutoRadiomics. To control random splitting in the cross-validation subsets, I used a seed, as done in AutoRadiomics. However, as the seeds between RPTK and AutoRadiomics differ, resulting splits also vary. Experiments are designed like this in order to show the performance of RPTK "as is" without further adaptation of the framework to reduce the impact of small

sample size to performance variance influenced by adaptations to the AutoRadiomics algorithm.

To assess the relationship between dataset size, task complexity, and model variance, I performed a learning-curve analysis using the *LearningCurveDisplay* function from *scikit-learn*. Each dataset was randomly subsampled in increasing proportions, and model performance was evaluated in a five-fold cross-validation setting (see Appendix, Figure 8.11 - 8.17). This analysis quantifies how the area under the AUROC evolves with growing training size and provides an estimate of the variance and stability of model performance.

The results confirm that smaller datasets exhibit higher variance in validation performance, indicating greater instability due to limited data (see Appendix, Figure 8.11 –8.17). The variance decreased as the proportion of training samples increased, reflecting the expected convergence of the learning process. Pronounced fluctuations were observed for the CRLM (n=77), Lipo(n=115), Melanoma(n=103), and LIDC-IDRI (n=115) datasets, where standard deviations remained large even at full data utilization compared to the larger LiverCRC (n=1,997) dataset (see Figure 8.25). This behavior illustrates that the available sample size was not always sufficient to fully stabilize model performance, consistent with the limited number of cases. For every machine learning project it should be notable that not the entire data can be used for training the model, as I also need to evaluate the models ability to predict the target label on unseen data which reflects the actual use case of machine learning models and needs to be evaluated. The effective learning data is usually, 70 - 80 % of the data where respectively 30 - 20 % are the testing data. This limitation increases the cross-validation performance variance especially for small datasets like CRLM, Lipo, Melanoma, and LIDC-IDRI (see Appendix, Figure 8.12, 8.15, 8.16, 8.17). In addition, I could only perform an internal validation and testing of the data (splitting the data into training and testing from the same data source) instead an external testing of the data from a different institute would increase the confidence of the model to perform on data which is independent from the training data source to get a better estimate of the models generalization to different data sources as well as technical (related to the image acquisition and applications of MR and CT) and clinical (including different treatment effects and interventions which are viable in the images) bias.

The Melanoma dataset, in particular, demonstrated a typical signature of high task difficulty—low and unstable validation AUROC that improved only marginally with additional training data, suggesting that the underlying imaging phenotype is either weakly discriminative or masked by label noise. As stated in the dataset description the BRAF mutation status of the lung metastasis was assumed to be the same for all

segmented lung metastasis on a patient level whereas I included multiple segmented lung metastases which might not all follow the same BRAF mutation status [7, 193]. Therefore, the model might recognize radiomics patterns in the data where the label is incorrectly assigned, which reduces the predictive power of the radiomics approach for classification between BRAF mutated lung metastasis. RPTK extracted features from multiple ROIs per sample and averaged the radiomics features afterwards for further filtering. Together, these findings highlight that adequate dataset size relative to task complexity and precise label evaluation is crucial for reliable radiomics benchmarking.

All open-source datasets used in this section are purely imaging-based, providing no additional clinical information such as comorbidities, treatment regimens, or laboratory values. This deliberate focus on radiological information strengthens comparability between frameworks and isolates the performance of imaging biomarkers. However, it limits the assessment of multimodal integration, as real-world clinical decision-making typically relies on both imaging and clinical context. Hence, the imaging-only results presented here reflect a technically controlled evaluation scenario, but they likely underestimate the complexity of clinical application.

Although the WORC database provides high-quality, expert-reviewed segmentation masks, visual inspection and quantitative analysis revealed pronounced segmentation fragmentation in several datasets. Figure 5.2 shows the number of connected components across datasets in the segmentations. GIST and Lipo exhibited the highest occurrence of small, disconnected components, while CRLM and Melanoma showed compact, single-lesion segmentations. These findings indicate considerable heterogeneity in segmentation integrity and lesion morphology across datasets and tumor biology.

The high number of disconnected components in GIST and Lipo likely reflects segmentation challenges rather than annotation errors. Lipo contains liposarcomas, which often have diffuse boundaries and inhomogeneous textures, making precise delineation difficult even for experts. GIST lesions show similar variability in size and shape and often present low contrast against adjacent soft tissue. In contrast, CRLM (liver metastases) and Melanoma (lung metastases) lesions are typically well circumscribed, leading to fewer segmentation artifacts. Importantly, all WORC segmentations were manually or semi-automatically generated under radiological supervision [7]. Thus, fragmentation likely stems from manual variability, morphological complexity, or algorithmic artifacts. While RPTK mitigates these effects through systematic connected-component filtering and quality control, segmentation heterogeneity remains a potential confounder. Future work should evaluate how segmentation uncertainty and automated contouring affect downstream model performance.

The datasets also differ substantially in slice thickness (Figure 5.1), ranging from sub-millimeter slices in Melanoma and LIDC-IDRI to slices above 10 mm in Desmoid and CRLM. These differences reflect distinct acquisition protocols and lesion characteristics. Small lung metastases and nodules in Melanoma and LIDC-IDRI require fine-resolution protocols to capture detail, whereas abdominal or soft-tissue lesions are often scanned with thicker slices for full-organ coverage. All images were resampled to isotropic spacing of $1 \times 1 \times 1$ mm to standardize voxel geometry across datasets. Although the chosen resampling method minimizes interpolation artifacts according to prior literature, residual effects are unavoidable. In datasets with coarse original spacing, resampling introduces artificial smoothness and may alter textural or morphological features. Consequently, radiomic feature distributions can vary not only due to biological differences but also due to technical interpolation effects.

The LIDC-IDRI dataset served as an extension of the open-source data cohort to test the generalizability of RPTK. However, its use required substantial preprocessing and curation. Previous studies have shown that the standard LIDC-IDRI malignancy annotations are based on subjective radiologist scores rather than pathological confirmation [20, 199, 200]. To reduce label noise, I restricted the dataset to cases with explicit clinical diagnostic confirmation (biopsy, resection, or long-term radiological stability). This decision improved label confidence but reduced the dataset size from the commonly reported 1,018 scans to 115 scans. Although this restriction enhances reliability, it limits comparability with literature. The smaller sample size therefore improves label precision at the expense of statistical power and benchmarking consistency.

The age and sex distributions of the combined open-source datasets (see Figure 4.1) closely resemble those of the entire cohort analyzed in this thesis (see Figure 8.3b and Figure 8.1). Both distributions are dominated by older patients, particularly males above the age of 60, who constitute the majority of cases in these populations. This predominance reflects the typical risk group for the studied diseases and thus represents the most clinically relevant patient population. However, the underrepresentation of younger individuals limits the model's ability to generalize across diverse demographic subgroups. Future studies would benefit from a more balanced inclusion of younger patients and female participants to enhance model robustness, incorporate age- and sex-related biological variability, and reduce demographic bias in predictive performance. Additionally, the presence of missing demographic information (e.g., age or sex) introduces further uncertainty and should be minimized to strengthen the interpretability of future analyses.

Together, these data characteristics and limitations define the context in which

the RPTK results should be interpreted. Small sample sizes and missing clinical metadata restrict statistical inference, while segmentation fragmentation, limitation to internal evaluation, and resampling heterogeneity introduce potential bias. Nevertheless, these same challenges underline the necessity of a self-configuring framework: RPTK is specifically designed to adapt to variable input conditions and to ensure consistent preprocessing and feature extraction without manual reconfiguration. The systematic characterization of these data limitations through automated fingerprinting and harmonization steps represents one of the core advantages of RPTK over conventional radiomics workflows.

6.1.2 Methodological Advances of the RPTK Framework

The underlying design philosophy of RPTK differs from that of AutoRadiomics. Whereas AutoRadiomics explores combinations of multiple feature selection methods and predictive models, RPTK emphasizes exhaustive feature extraction, data preprocessing, and stability-based feature filtering, followed by a uniform feature-selection strategy and extensive model optimization. This simplified and controlled design minimizes the methodological bias introduced by varying feature-selection methods and isolates model performance as a relevant factor (see inter model performance variations on different datasets in Figure 5.5). Each predictive model in RPTK is optimized for each fold using 200 iterations, whereas AutoRadiomics uses 200 iterations for optimization of the best configuration and selection of the feature selection algorithm in combination to the predictive model and the optimization. This addition of AutoRadiomics includes further parameters to optimize and may need more iterations for a better performance. This also results in a very reduced calculation time of AutoRadiomics compared to RPTK. RPTK needed 72 to 96 computation hours on the same datasets AutoRadiomics needed 28 until 42 hours (see Section 4.1.12).

A further improvement implemented in the current version is the inclusion of a pre-training stage to determine model capacity. By analyzing performance saturation with respect to model size, RPTK prevents unnecessary model complexity and reduces overfitting risk, especially in small datasets. Together, these enhancements align with recommendations from the radiomics literature for improving reproducibility and model interpretability.

A direct comparison between the prototype introduced at the MICCAI Conference and the extended version presented here shows higher or similar test AUROC values across the benchmark datasets (see Figure 5.8). The improvement can be attributed to the combination of feature stability filtering, expanded model diversity, and ensemble-based aggregation. However, these results should be interpreted with caution, as

they may also partially reflect dataset-specific factors or random variation in cross-validation rather than intrinsic methodological superiority.

Overall, the RPTK framework represents a structured and transparent methodological design that focuses on reproducibility, systematic evaluation, and control of bias. Its modular structure enables consistent preprocessing, comprehensive feature assessment, and standardized model optimization, providing a reproducible foundation for benchmarking radiomics workflows. Nevertheless, further work could explore the integration of additional feature-selection techniques or automated model-selection strategies to potentially enhance predictive performance while maintaining methodological transparency.

RPTK Radiomics Feature Extraction

The configuration of the feature extraction step represents one of the most critical components in radiomics workflows. A key parameter influencing feature stability and interpretability is the pixel discretization of image intensities prior to texture computation. As introduced in Section 2.3.6, discretization reduces the number of grey levels in the radiological image and thereby lowers the noise-to-signal ratio. However, the choice of discretization parameters also constrains the information available to texture-based feature classes such as GLCM, GLRLM, GLSZM, GLDZM, NGTDM, and NGLDM, which rely on grey-level co-occurrence or dependency statistics. The recommended fixed bin width of 25, as proposed by [1], provides a balance between noise suppression and signal preservation. Nevertheless, the optimal discretization may vary between modalities and datasets, and future extensions of RPTK could include adaptive discretization based on dataset-specific intensity statistics captured in the data fingerprint.

The number of bins calculated from the fixed bin width discretization is included as a quantitative fingerprint parameter. This parameter reflects the granularity of grey-level encoding within the region of interest (ROI) and therefore provides indirect information on image contrast and texture richness. As shown in Figure 8.5, substantial variation in the number of bins was observed across datasets. MRI-based datasets such as Desmoid, Liver, and Lipo exhibited higher variability due to the absence of grey-value normalization, whereas CT-based datasets (CRLM, GIST, Melanoma, and LIDC-IDRI) displayed more homogeneous distributions. For CRLM in particular, the number of bins occasionally fell below 10, suggesting that the applied discretization might have been too coarse to capture fine texture variations. Consequently, implementing an adaptive discretization strategy that maintains at least 10 bins per ROI could improve the information content of texture matrices and potentially enhance

model performance.

Both feature extraction backends integrated in RPTK, PyRadiomics and MIRP, adhere to the feature definitions of the BSI, which promotes reproducibility and cross-platform comparability. However, PyRadiomics currently lacks 63 IBSI-defined features (see Appendix Figures 8.7 and 8.8), including the entire NGLDM class describing texture coarseness. While some missing features have been reported as mathematically redundant by the PyRadiomics developers, others represent morphological descriptors such as volume and area density that may carry complementary information about lesion geometry. The absence of these features alters the relative weighting of feature classes within the extracted space and increases the proportional contribution of well-represented classes such as GLCM and GLDZM. This may explain the stronger influence of texture-based features observed in the PyRadiomics-derived feature subsets (see Figures 5.3 and 5.4).

In addition to the IBSI-defined features, both extractors include a set of image-descriptive and metadata-related parameters that capture technical and contextual properties of the ROI but also of the image, such as voxel spacing, mask interpolation, bounding box geometry, and tool version information. While certain spatial descriptors can contain useful information about acquisition characteristics or lesion position within the image volume, others—such as software version or configuration identifiers—are not biologically meaningful. RPTK systematically filters these non-relevant descriptors during feature preselection to ensure that only biologically and technically interpretable features contribute to downstream analysis.

The comparison between the two extractors underscores the methodological importance of feature-space completeness and harmonization in radiomics. Features that are absent in one extractor cannot be recovered through feature selection or model optimization, which could limit the predictive power of the derived models. Consequently, RPTK emphasizes comprehensive feature extraction rather than extensive feature-selection diversity, ensuring that the full range of potentially informative descriptors is represented prior to dimensionality reduction. These design choices highlight the significance of extractor configuration as a primary determinant of radiomics model performance and reproducibility.

RPTK Radiomics Feature Filtering and Selection

The reduction of the feature space is a necessary step in RPTK based on the fact that I do not reduce the initial extracted feature space in order to not limit the number of potentially informative features. This step needs to follow a systematic procedure. Therefore, I implemented several steps which are applied in a specific

order to reduce the amount of uninformative features and potential bias as well as reduce the calculation time for the feature selection. The feature filtering is the step of reducing the feature space from the feature extraction based on heuristic rules and statistical characteristics of the features to remove non-informative and redundant features (see Section 4.1.7). As these procedures have been widely applied in radiomics studies but with slightly differences, there is no common recommendation for defining the variation nor correlation thresholds for feature filtering [212–214]. Therefore, I decided to use thresholds to filter for features showing very low variance (threshold = 0.1) and very high correlation (threshold = 0.9) in order to minimize the risk of dropping important features which can not get selected anymore for the predictive task.

Feature selection plays a crucial role in radiomics workflows, as it determines which information is retained for model training and directly affects the balance between predictive power and overfitting. In most radiomics datasets, the number of extracted features greatly exceeds the number of available samples, increasing the risk of spurious correlations and loss of generalizability. Therefore, an effective selection process must condense the feature space to its most informative components while maintaining sufficient diversity to capture relevant image-derived information.

The feature selection results shown in Figures 5.3 and 5.4 illustrate the distribution of selected features across IBSI feature classes and their respective origins of information, distinguishing between intra- and peritumoral regions. For each dataset, up to 20 features were retained, 10 from forward sequential selection and 10 from backward selection. Features appearing in both subsets were counted only once, resulting in final feature spaces of fewer than 20 features in some cases. In addition, previous performed correlation- and variance-based filtering ensured that highly redundant or invariant features were excluded prior to the selection process.

Across datasets and both extractors, texture-based feature classes dominated the selected feature sets, particularly the GLCM features. In the PyRadiomics-derived selections, GLCM features accounted for roughly half of the total selected features in the Desmoid dataset and about 20% in the Lipo dataset. For MIRP, GLCM features represented between 50% (LIDC-IDRI) and 10% (Liver). This dominance is partly attributable to the high representation of GLCM features in the original extracted feature space of both tools (see Appendix, Figure 8.7 and 8.8) and is consistent with findings from previous studies that highlight their relevance in diverse clinical applications [236]. Nevertheless, other texture-based feature classes such as NGTDM and Intensity-based statistic (IS) were also recurrently selected across datasets, despite their comparatively smaller representation in the initial feature space. This indicates

that even less frequent feature types can carry complementary predictive information and get selected during feature selection.

Overall, texture-describing feature classes form the majority of selected features, suggesting that spatial intensity relationships and heterogeneity metrics are consistently informative across datasets and modalities. The presence of non-IBSI features in several PyRadiomics-selected subsets (five out of seven datasets) and the increased occurrence of these features in the Lipo dataset for both extractors indicate that features outside the current IBSI definition may contribute additional relevant information. However, the reproducibility and generalizability of these non-standard features require further validation, particularly since they may depend on specific algorithmic implementations.

The inclusion of the peritumoral region as an additional origin of information extends the radiomics analysis beyond the lesion boundary. Previous studies have reported that peritumoral tissue characteristics may carry prognostic information, particularly in the context of treatment response and tumor–host interactions [207–209]. In the present results, MIRP-derived features included peritumoral descriptors in four datasets, whereas PyRadiomics selected them in five datasets. The extent and composition of selected peritumoral features varied between extractors and datasets, reflecting both segmentation definitions and feature extraction strategies. Notably, for the MIRP feature extraction in the Lipo dataset, peritumoral features were selected more frequently than intratumoral ones, while PyRadiomics selected exclusively intratumoral features for the same dataset.

These differences demonstrate that the selection outcome is strongly influenced by the underlying feature extractor and its implementation details. MIRP, for example, computes texture features in both two and three dimensions, whereas PyRadiomics limits certain feature classes to 3D, leading to divergent representations of spatial relationships. Consequently, differences in feature-space coverage and extraction methodology contribute directly to performance variations observed in the trained classifiers (see Figure 5.5).

In summary, the feature selection procedure within RPTK effectively reduces feature dimensionality while retaining a balanced set of texture- and intensity-based and also includes morphological descriptors. The resulting feature profiles highlight both commonalities and divergences between extractors and confirm that the most predictive information in the examined datasets arises from texture heterogeneity metrics. At the same time, the observed differences between intra- and peritumoral contributions, and between IBSI and non-IBSI features, emphasize the importance of standardized and comprehensive feature definitions for reproducible radiomics mod-

eling.

6.1.3 RPTK Model Prediction Performance

To identify the most suitable predictive model for each dataset, the RPTK framework trains six distinct machine learning algorithms and automatically selects the configuration with the highest validation AUROC (see Figure 5.5). This model selection strategy ensures consistent evaluation across extractors and datasets, relying on the same validation criterion used in AutoRadiomics and WORC benchmarking studies [18, 19].

The performance comparison between models trained on features extracted by PyRadiomics and MIRP revealed notable dataset-dependent differences. Substantial performance gaps were observed for GIST, Melanoma, and LIDC-IDRI (delta Val AUROC about 0.05–0.10), while results for Desmoid, Liver, and CRLM were highly similar (delta Val AUROC <= 0.03) (see Table 5.1). These differences reflect the influence of extractor-specific feature definitions and dimensional computation strategies. Nevertheless, considering the standard deviations across folds, most AUROC differences between extractors fall within the expected variability range and are not statistically significant. On the LIDC-IDRI dataset, RPTK applied the SMOTE oversampling on the training data to to compensate the class imbalance, which reached the SMOTE activation threshold (see Section 4.1.8, see Table 4.3).

The exclusive reliance on validation AUROC as the model selection criterion in RPTK may not always yield the most robust or generalizable model. Small differences in validation AUROC can lead to the selection of suboptimal configurations when evaluated across other performance metrics or on the test set. For instance, in the Lipo dataset, the best-performing RPTK model during validation did not outperform AutoRadiomics on the test data (delta AUROC 0.036), the only dataset where this occurred. Nevertheless, as the model trained on PyRadiomics features achieved a test performance nearly equivalent to that of AutoRadiomics, with only a 0.013 AUROC difference, this discrepancy is likely attributable to the internal model selection process within RPTK rather than to the framework's overall capability.

To assess the contribution of the additional procedures implemented in the current RPTK version compared to the previously developed RPTK prototype, I included a direct comparison under identical best-model selection criteria (see Table 8.1). As shown in Figure 5.8, the enhanced RPTK consistently improved test performance across all datasets, except for CRLM, where it performed equally to the prototype. This observation indicates that the newly integrated components of the RPTK framework, as described in Section 4.1.3, contribute to improved model performance and

robustness in most studies.

It is important to note that model selection based solely on validation AUROC does not guarantee the best test-set performance, as this metric does not directly capture generalization capability. For several datasets (Desmoid, Lipo, Melanoma, and LIDC-IDRI), the best-performing model on the test set originated from the alternative extractor. This finding underscores the sensitivity of small-sample datasets to cross-validation variance and highlights the necessity of independent testing for reliable performance assessment.

Across all datasets, RPTK achieved test AUROC values above 0.70, except for Melanoma, which has been consistently challenging for both classical and deep learning approaches (see Figure 5.7). Threshold-based performance metrics, calculated using the Youden correction, further support these findings (Table 5.1). The average test F1-score across datasets was 0.805, with mean sensitivity and specificity of 0.805 and 0.813, respectively. According to the performance scale proposed by [237], these values correspond to a "good" or "useful" level of predictive accuracy, indicating reliable discrimination for most tasks. Only isolated cases, such as the Lipo dataset (F1 = 0.686) or datasets with low sensitivity or specificity (GIST, LIDC-IDRI, CRLM, and Lipo), fall below thresholds considered clinically relevant. Future versions of RPTK may incorporate threshold-dependent optimization criteria during training to minimize the gap between AUROC and clinically interpretable metrics such as sensitivity and specificity.

To contextualize these results, RPTK was benchmarked against two alternative approaches AutoRadiomics and deep learning models on synchronized data splits. Validation AUROC comparisons (Figure 5.6) show that RPTK consistently outperforms AutoRadiomics and achieves similar or higher scores than deep learning models across most datasets. The only exceptions occur for Liver and Melanoma, where deep learning exhibits slightly higher validation AUROC but fails to generalize on the test set, indicating overfitting. AutoRadiomics, in contrast, produced consistently lower validation and test AUROC scores, with the single exception of the Lipo dataset, where it slightly outperformed RPTK (delta Test AUROC = 0.02). The performance differences between AutoRadiomics and RPTK are also viable by comparing the ROC curves of both approaches across the open-source datasets (see Figures 8.9).

The analysis of model frequencies across datasets (see Tables 8.2 and 8.3) reveals distinct selection patterns between the deep learning and AutoRadiomics approaches. Among the deep learning models, DenseNet-based architectures were most frequently selected as best performers compared to the ResNets. The validation AUROC of the deep learning models ranged from 0.724 (CRLM) to 0.938 (Liver), with corresponding

test AUROC values between 0.422 and 0.841, resulting in a mean validation AUROC of approximately 0.83 and a mean test AUROC of around 0.61 across all open-source datasets. The test AUROC performance decrease compared to the validation AUROC reflects the weak generalization performance of the approached deep learning models.

In contrast, the AutoRadiomics optimization framework selected a broader variety of model–feature selection combinations, reflecting its search across multiple pipeline configurations (see Section 3.2.3 for details). Logistic Regression was most frequently chosen (2×), followed by SVM (2×), XGBoost (2×), and Random Forest (1×). Regarding feature selection, ANOVA dominated (4×) followed by Boruta (2×). The AutoRadiomics models exhibited validation AUROC values ranging from 0.456 (LIDC-IDRI) to 0.862 (Lipo) and test AUROC values between 0.392 and 0.922, with a mean validation AUROC of approximately 0.67 and a mean test AUROC of about 0.70.

Overall, while deep learning models achieved higher validation performance on average, AutoRadiomics demonstrated more consistent and often superior generalization on the test sets, suggesting that the integrated feature selection and model optimization pipeline may better capture dataset-specific radiomic patterns under limited sample conditions.

In some datasets, notably Melanoma and LIDC-IDRI, both deep learning and AutoRadiomics achieved near-chance or below-chance AUROC values (AUROC < 0.5). Such below-chance results for AutoRadiomics indicate unstable or misaligned feature selection, likely due to excessive automation and the limited size of the datasets. This effect was particularly pronounced in LIDC-IDRI, where AutoRadiomics validation AUROC ranged between 0.38 and 0.53, consistent with underfitting and class-imbalance sensitivity despite the use of internal SMOTE balancing. Inverting predictions would mathematically yield AUROC > 0.5 but would constitute post-hoc bias and was therefore avoided. These results instead indicate a lack of a generalizable signal, as the same dataset yielded AUROC > 0.7 under RPTK, confirming that its adaptive preprocessing and feature extraction more effectively captured relevant imaging information. Deep learning models, conversely, achieved moderate validation performance but exhibited substantial performance collapse on test data, a typical manifestation of overfitting in small and heterogeneous cohorts.

When compared to published literature (Figure 5.8), RPTK consistently ranked among the top-performing approaches across all datasets, matching or exceeding the AUROC values reported for radiologists performing the same classification tasks. The most pronounced performance differences were observed in the Melanoma and LIDC-IDRI datasets, where many prior approaches struggled to surpass random perfor-

mance. RPTK clearly outperformed these baselines, further supporting its ability to generalize across diverse and complex imaging conditions.

Across datasets, the width of the 95% confidence intervals (CI) provides additional insight into performance stability. Datasets with larger sample sizes, such as Desmoid ($n \approx 200$) and GIST ($n \approx 247$), exhibited narrower CIs, reflecting more consistent model estimates. In contrast, datasets with fewer cases (CRLM, Lipo, Melanoma, LIDC-IDRI; $n < 120$) showed broader confidence intervals, indicating greater variability. These patterns align with the learning curve analyses (see Appendix, Figures 8.11–8.17), which demonstrate that CRLM, Melanoma, and LIDC-IDRI exhibit continuing performance gains with increased data, whereas Desmoid, GIST, and Liver appear to approach a performance plateau.

On average, RPTK achieved 0.818 test AUROC, 0.805 test F1, 0.805 sensitivity, and 0.813 specificity across the seven datasets, representing robust performance and effective generalization. The comparative evaluation against AutoRadiomics, deep learning, and previously published literature establishes RPTK as a reliable and high-performing self-configuring radiomics framework capable of handling heterogeneous datasets and achieving expert-level or superior predictive accuracy.

6.2 Predict Study - Predicting Immunotherapy Treatment Response in Lung Cancer Patients

6.2.1 Data

The Predict study cohort consists of 73 patients with advanced non-small cell lung cancer (NSCLC) treated at the Thoraxklinik Heidelberg with immunotherapy as a mono therapy. For each patient, two thoracic CT scans were acquired—one at treatment initiation and a follow-up scan after the first administration cycle—capturing the early response to immunotherapy. All patients received the same anti–PD-L1 drug, Pembrolizumab, administered at a standard dose of 200 mg in a continues interval of about three weeks. This uniform therapeutic protocol ensures consistency of treatment-related factors across the cohort and minimizes variability in systemic exposure that could confound imaging-based outcome modeling.

Despite this controlled treatment setting, the dataset size remains a central limitation. With only 73 patients and two imaging time-points per case, the cohort provides a valuable but statistically constrained sample for training data-driven models. Small sample sizes restrict the stability of cross-validation and limit the ability to perform independent testing, which may result in high variance of model estimates and lower

statistical power for detecting true predictive signals. Performed learning curve analysis underlay this limitation (see Figure 8.23). Increasing usage of the data shows high variation in the cross-validation performance, even using the total dataset for training, this suggests an increase of the data size would benefit further prediction quality and model certainty.

Clinically, this cohort represents a population with advanced disease burden: over 50% of patients presented with stage IV NSCLC at inclusion (Table 8.4). Furthermore, 88% exhibited metastatic disease, either local or distant resulting from previously applied treatments which did not reduce the tumor progression or got a relapse. As a result, radiomics features may be influenced by cumulative treatment effects, such as post-therapeutic inflammation or fibrosis, which are difficult to disentangle from true tumor response patterns. Pleural effusion, reported in 30% of patients, exemplifies such secondary effects that may alter image characteristics without directly reflecting tumor burden. Hence, confounding from prior treatments and disease progression must be considered when interpreting radiomics-based response predictions in this cohort.

Demographically, the cohort shows a pronounced predominance of older male patients with a history of heavy smoking. The mean age was approximately 66 years, 62% were male, and 92% were current or former smokers with an average exposure of 37 pack-years. This demographic profile aligns with established epidemiological risk patterns for NSCLC , where older male smokers represent the highest-risk subgroup [238]. However, such homogeneity limits generalizability, as the underrepresentation of younger women and non-smokers may lead to demographic bias in trained models. Consequently, predictive performance for underrepresented subgroups should be interpreted with caution and validated on more diverse populations.

From a radiological perspective, all scans were acquired using thoracic CT protocols with limited protocol variability. The main acquisition difference relates to contrast administration, encompassing arterial, venous, and non-contrast phases. The imaging fingerprint (Table 5.2) shows that slice thickness, number of slices, and discretizations-related parameters are comparable to those of other thoracic CT datasets used in this thesis, such as Melanoma and LIDC-IDRI. The mean ROI size, however, was substantially larger—exceeding 100,000 voxels on average and approximately 36,000 voxels more than the largest open-source dataset (see Figure 8.4). This large ROI variation corresponds to the high variability in tumor volumes expected for late-stage lung cancer. The standard deviation of ROI sizes was also the highest among all datasets, reflecting the clinical heterogeneity of tumor burden within this population.

Segmentation characteristics were generally consistent and of high quality. The

number of connected components per segmentation was close to one for all cases, similar to the thoracic open-source datasets (Melanoma and LIDC-IDRI). This consistency indicates minimal segmentation fragmentation, an important factor for ensuring stable radiomics feature extraction.

Another limitation arises from the restriction to the primary tumor as the sole region of interest. Given that most patients had metastatic disease, secondary lesions were excluded from analysis to maintain consistency. While this decision simplifies the modeling setup, it likely under-represents the systemic disease state and spatial heterogeneity of the tumor burden. Future work could address this limitation by incorporating multi-lesion or whole-body imaging data to capture global disease dynamics.

In summary, the Predict dataset offers a clinically well-characterized but statistically limited cohort with homogeneous treatment, high disease stage, and restricted demographic diversity. These characteristics provide a controlled environment for exploring radiomics-based response prediction but also impose constraints on generalization. The strong demographic and clinical bias toward older male smokers with advanced NSCLC, combined with small sample size and prior treatment heterogeneity, represents a key challenge for robust model development. Nevertheless, the dataset's consistent imaging protocol, controlled therapy regimen, and well-defined clinical endpoints make it a valuable foundation for investigating radiomics-based biomarkers of immunotherapy response in a real-world clinical context.

6.2.2 The Performance impact of longitudinal Data and Delta Radiomics

To evaluate the potential benefit of longitudinal image analysis for predicting treatment response, RPTK was applied to three different input configurations of the Predict dataset: baseline CT scans only, follow-up CT scans only, and the combination of both time points using delta radiomics. The corresponding validation and test performances are summarized in Figure 5.9. Across all configurations, a consistent performance gain was observed from the baseline to the follow-up setting, with the highest predictive accuracy achieved when both time points were combined through delta feature computation.

This result aligns with previous studies demonstrating that delta radiomics, capturing changes in quantitative image features over time, can enhance the sensitivity of predictive models to therapy-related alterations [239, 240]. By explicitly modeling temporal feature differences, delta radiomics highlights changes in radiomics features

as treatment effects that may remain obscured in static single-time-point analyses. In contrast, radiomics models trained solely on baseline images rely exclusively on pre-treatment morphology and texture, which can not contain information to predict treatment response before therapy onset as the effect of treatment initiation did not occur. Consequently, the lower performance observed for baseline-only models is likely attributable to the absence of early treatment-induced changes in tumor phenotype (see Figure 5.9, and Table 8.5).

The performance improvement from baseline to follow-up and ultimately to delta radiomics configurations suggests that the discriminative signal related to treatment response becomes more pronounced over time and is best captured by quantifying feature differences between time-points. Although the small cohort size limits statistical power and prevents the demonstration of significance—reflected by wide 95% confidence intervals, the observed increasing performance trend in both validation and test AUROC supports the added value of longitudinal modeling. This improvement was consistent across model folds, indicating that the observed gain is systematic rather than random variation.

Overall, the findings confirm that delta radiomics provides a valuable means of integrating temporal information into radiomics-based prediction frameworks. Within the context of RPTK, the implementation of delta feature computation enables the detection of early imaging-based treatment response patterns, even in small and heterogeneous datasets. These results highlight the importance of incorporating longitudinal feature representations in radiomics studies addressing therapeutic response, and they justify the continued use of delta radiomics in the subsequent analyses presented in this study.

Radiomics Feature Processing on the Predict Dataset

Radiomics feature extraction in the Predict study was conducted using two independent extraction frameworks, PyRadiomics and MIRP, integrated within the RPTK framework. Both extractors produced a comparable number of selected features (19 and 20, respectively), with strong overlap in the types of image characteristics identified as predictive. Texture-related feature classes dominated both selections, confirming that spatial grey-level heterogeneity remains the most informative image descriptor for assessing immunotherapy response in advanced-stage lung cancer.

For the PyRadiomics-based feature set, more than half of the selected features originated from the GLDZM and GLCM classes, reflecting the importance of distance- and co-occurrence–based grey-level dependencies. Several features, including GLCMIDMN and the first-order *RootMeanSquared*, exhibited a separation between responders and

non-responders in feature-value distributions. These features displayed higher normalized values in responders, suggesting increased textural uniformity associated with effective treatment response. The repeated selection of features such as GLDZM LDLGE under multiple wavelet transformations indicates that similar structural information was consistently captured across different filtered representations, highlighting the robustness of these textural patterns.

The MIRP-derived feature space showed a comparable dominance of texture-based metrics, with the highest representation from GLSZM, GLCM, and intensity histogram (IH) classes. Approximately half of the selected MIRP features originated from 3D wavelet transformations, while the remainder included both 2D slice-aggregated and 3D-computed descriptors. The NGTDM Complexity and ZSEntr features recurred under different transformation settings, suggesting that MIRP's inclusion of 2D and 3D feature variants enhances the capture of complementary spatial information. Similar to PyRadiomics, distinct response-related clustering was observed for the GLCM IDMN feature and the GLRLM LRHGE feature, where higher values corresponded to treatment responders, further underscoring their predictive potential.

Cross-extractor comparison revealed notable methodological but not conceptual differences. PyRadiomics computes texture features primarily in 3D, whereas MIRP includes additional 2D implementations and the NGLDM feature class, which is not available in PyRadiomics (see Appendix Figure 8.7). MIRP also extracts a larger fraction of intensity-based descriptors, while PyRadiomics places greater emphasis on co-occurrence and zone-based metrics. Despite these structural differences, both frameworks converged on key predictive features, most prominently the GLCM IDMN and morphological Sphericity, which were independently selected and ranked as highly important in the SHAP feature importance analysis (see Figures 5.12a and 5.12b).

The agreement between independent extraction pipelines in identifying overlapping predictive features highlights the reproducibility and robustness of RPTK's feature selection process. The observed consistency in the predictive relevance of GLCM-, GLDZM-, and GLSZM-based features across extractors indicates that radiomics biomarkers derived from tumor texture heterogeneity may generalize across different software implementations. At the same time, extractor-specific variations in dimensionality and feature coverage emphasize the importance of methodological transparency and standardized feature definitions in radiomics. Together, these results demonstrate that RPTK effectively integrates distinct feature extraction strategies and yields stable, biologically interpretable feature representations for treatment-response prediction.

A comparison between the feature spaces derived by RPTK and AutoRadiomics on

the Predict dataset further illustrates the influence of feature selection strategy on the resulting model characteristics. Although both approaches rely on the same extraction backend (PyRadiomics), the composition and diversity of the selected features differed markedly. AutoRadiomics identified a compact subset of ten features, all originating from wavelet-transformed representations and restricted to two IBSI feature classes (GLDZM and GLRLM). In contrast, RPTK selected a broader and more heterogeneous feature set encompassing multiple texture families, first-order descriptors, and morphological parameters. Despite this methodological difference, both approaches converged on similar feature classes—particularly GLDZM and GLRLM—which were consistently found to be associated with treatment-response prediction. However, no direct overlap in specific features was observed between the two methods, reflecting that AutoRadiomics tends to favor transformed versions of texture features, while RPTK selects a more diverse combination of untransformed and transformed descriptors. The stronger response-related clustering observed in the RPTK feature heatmap suggests that its sequential feature selection procedure may better preserve complementary information across multiple feature domains, whereas AutoRadiomics optimization routine emphasizes compactness and redundancy reduction within single feature families. The absence of reparative clusters in the radiomics heatmap is based on the intense statistical feature filtering of RPTK which eliminates highly correlating features before feature selection and impacts the available features handed to the selection process in RPTK.

Integration of Clinical and Radiomics Features

The Predict dataset provides a comprehensive collection of clinical and demographic parameters in addition to imaging data, enabling the evaluation of whether radiomics can contribute added predictive value within an already information-rich clinical context. To systematically assess this, three models were trained: a clinical-only model, a delta radiomics model, and a combined clinical–radiomics model. Their comparative test performances are summarized in Table 5.3. The combined model achieved the highest overall performance, particularly for threshold-based metrics such as F1-score, sensitivity, and specificity, followed by the clinical-only model and then the delta radiomics model. This trend indicates that clinical and radiomics features provide complementary rather than redundant information. While clinical data encode established prognostic and demographic factors, radiomics captures quantitative imaging dynamics related to therapy response, thereby improving individualized prediction when combined.

Model interpretability based on SHAP values supports this conclusion (see Figures

5.13a and 5.13b). In the clinical-only model, PD-L1 expression—known to be a strong predictive biomarker for immunotherapy response—was ranked as the most influential variable, followed by tumor size, CRP level, age, and smoking status. These findings align with current clinical evidence for prognostic biomarkers in NSCLC immunotherapy [97, 241, 242]. The clinical model thus reflects real-world decision-making, where PD-L1, inflammatory markers, and patient performance metrics are the primary indicators of likely response to anti–PD-L1 therapy. This parameter also generates a clear pattern viable in the heatmap of the clinical feature selection for training a model only on the most informative clinical features (see Figure 8.19).

The combined model, by contrast, demonstrated a more balanced distribution of feature importance across radiomics and clinical predictors. Among the ten most important features, radiomics-based descriptors such as GLCM IDMN and GLRLM LRHGE ranked highest, while clinical features such as ECOG performance status, smoking behavior, and pack-years contributed complementary information (see Figure 8.18). This integration highlights that radiomics features provide orthogonal, image-derived information that refines predictions beyond traditional clinical variables. In particular, the inclusion of radiomics features allowed the model to capture subtle treatment-related changes not directly visible in aggregate clinical descriptors.

The performance advantage of the combined model compared to the clinical model was moderate but consistent, reflecting the strong baseline predictive power of the clinical data. The clinical dataset itself is inherently multimodal, integrating demographic, laboratory, and radiological information such as tumor size and location. Nonetheless, the improvement in test performance and the redistribution of feature importance toward radiomics descriptors suggest that quantitative image information adds fine-grained complementary value to existing clinical predictors.

Missing clinical data were handled using a robust imputation strategy implemented in SimpleITK, applying a K-nearest-neighbor imputer for continuous variables and a most-frequent-value imputer for categorical or ordinal variables (see Section 4.1.5). Missingness did not exceed 50% for any feature. The variable with the highest missing rate was serum albumin (42%), followed by metastasis manifestation (12%), whereas all other parameters exhibited less than 10% missingness (Table 8.4). Whereas, data imputation includes data-related bias by simulating data, given these small proportions of values and missingness, the impact of imputation on model bias is expected to be minimal, and the retained variables maintain sufficient integrity for reliable modeling.

To explore the potential clinical relevance of the model predictions, a Kaplan–Meier survival analysis was performed based on the predicted responder and non-responder

classifications (Figure 5.14). The stratified survival curves revealed a significant survival benefit for patients predicted as responders by the combined model. Although five patients were misclassified as responders, the model's early response prediction—performed within the first 28 to 140 days of treatment—could still provide actionable clinical insight. If used prospectively, such predictions could help identify non-responding patients early, potentially allowing therapeutic reassignment at a stage when survival probability remains around 0.75, compared to approximately 0.15 after 1,000 days of continued ineffective treatment. This result demonstrates that integrating radiomics features into clinical prediction frameworks may not only improve quantitative accuracy but could also enhance clinical decision-making and patient outcomes through earlier identification of non-responders.

To further assess the reliability and clinical applicability of the predictive models, a confidence-based evaluation was performed using the confusion matrices shown in Figures 8.21a, 8.21b, and 8.22. These matrices illustrate the number of correctly and incorrectly classified patients with respect to their true treatment response. The model based solely on clinical variables and the model integrating both clinical and radiomics features each misclassified two responders as non-responders and two non-responders as responders, whereas the model relying exclusively on radiomics features misclassified one additional responder as a non-responder. The detailed performance metrics corresponding to these models are summarized in Table 5.3. Notably, the inclusion of clinical data affected only a single patient prediction in the test set, suggesting that while clinical variables already provide substantial prognostic information, larger datasets are needed to further explore these subtle differences, uncover potential imaging-related biases, and better quantify model uncertainty.

In summary, integrating delta radiomics with clinical parameters improved predictive performance and interpretability in a dataset already rich in prognostic information. While the incremental improvement was modest due to the strong baseline informativeness of the clinical data, the complementary role of radiomics features was evident. Radiomics contributed fine-grained, image-based descriptors of treatment-related heterogeneity that strengthened individualized response prediction. Overall, these results support the combined use of clinical and quantitative imaging data as a viable strategy for developing robust, explainable, and clinically meaningful prediction models in the context of immunotherapy response assessment.

RPTK Performance Comparison for Immunotherapy Response Prediction

To ensure a fair comparison across frameworks, all experiments were conducted on synchronized training and test partitions, identical to those used in the AutoRadiomics

evaluation. Delta radiomics features were computed for AutoRadiomics using the same procedure applied in RPTK, thereby eliminating potential biases arising from differing preprocessing or data split configurations. This alignment allows a direct assessment of methodological rather than data-related performance differences.

Across both validation and test sets, RPTK achieved the highest overall performance among the compared approaches. The best RPTK configuration based on MIRP extracted delta radiomics features—employed an XGBoost classifier, while the AutoRadiomics framework identified a Random Forest model combined with an ANOVA feature-selection scheme as its optimal pipeline (see Table 8.6). Despite this, the RPTK model consistently outperformed AutoRadiomics, achieving superior AUROC values in both cross-validation and held-out testing. The improvement was most evident on the test set, underscoring the greater generalization stability of RPTK's self-configuring optimization procedure.

In addition to the performance comparison, the feature selection behavior of AutoRadiomics was further examined to identify potential causes for its limited generalization (see Figure 8.20). The heatmap of the selected features shows that AutoRadiomics primarily relied on ten wavelet-transformed features, belonging exclusively to the GLDZM and GLRLM feature classes. Within these classes, descriptors such as *LargeDependenceEmphasis* and *RunVariance* were repeatedly selected under different wavelet decompositions. While this feature composition produced discernible response-related patterns in the training data, it also introduced high redundancy, as multiple correlated versions of the same texture measures were included. In contrast, RPTK explicitly filters correlated features, thereby preventing the inclusion of redundant transformations that contribute little new information. This decorrelation step promotes a more compact and generalizable feature representation.

Consequently, these observations suggest that AutoRadiomics captures meaningful but repetitive texture characteristics, which appear insufficient to sustain predictive performance on unseen data. The improved generalization of RPTK therefore likely arises from its stricter feature filtering and optimization strategy, which balances the inclusion of informative radiomics descriptors with the exclusion of redundant patterns.

Deep learning models were also trained on the same synchronized data splits for comparison. Among the tested architectures, ResNet-18 achieved the best validation performance with a mean AUROC of 0.80; however, its test AUROC dropped markedly to 0.56, indicating substantial overfitting (see Table 8.7). Similar validation–test performance gaps were observed in other small-sample datasets, reflecting the disproportionate impact that individual misclassified samples can exert on evalu-

ation metrics when cohort size is limited. The wide confidence intervals obtained in these experiments are therefore not indicative of methodological instability but rather of inherent statistical variance associated with small-sample machine learning [145]. Increasing dataset size or using larger multi-institutional cohorts remains the most effective strategy to reduce this uncertainty.

To prevent any form of data leakage, feature selection and model optimization in all approaches were strictly confined to the training folds, using the identical split definitions for both RPTK and AutoRadiomics. This ensured that performance differences arose exclusively from differences in pipeline design, not from discrepancies in data handling.

Overall, the results demonstrate that the self-configuring RPTK framework achieves higher robustness and generalization than both the AutoRadiomics and deep learning baselines when applied to longitudinal immunotherapy response prediction.

When comparing the Predict cohort to recently published studies on radiomics-based immunotherapy response prediction, it becomes evident that most external investigations relied on substantially larger and often multi-institutional datasets. For instance, Han et al. (2024) analyzed 179 patients from two hospitals in their longitudinal NSCLC study [243], while other single- and multi-center investigations typically included between 100 and 200 patients [244, 245]. These larger and more heterogeneous cohorts provide stronger statistical power, more stable performance estimation, and improved generalization across clinical environments.

Although the Predict dataset used in this thesis comprises only 73 patients, the consistent performance improvements achieved with RPTK—particularly when incorporating delta radiomics and clinical variables—suggest that the framework scales well with increasing data diversity and size. Given its self-configuring design and robust preprocessing, applying RPTK to multi-institutional immunotherapy cohorts could further enhance prediction accuracy and reduce variance. Future work should therefore focus on external validation in larger, independent datasets to assess the reproducibility and clinical transferability of the developed models across scanners, institutions, and treatment regimens.

6.3 LiverCRC Study – Colorectal Cancer Prediction via Liver CT

The LiverCRC study presented in this thesis is based on results that are also included in the joint manuscript currently under review [22]. However, the objectives and

methodological scope differ between the two works. In the manuscript, the primary focus lies on the clinical interpretation and validation of RPTK in the context of colorectal neoplasia prediction, whereas the present thesis emphasizes a methodological comparison between RPTK, AutoRadiomics, and deep learning approaches.

Accordingly, the dataset and experimental configuration used in this thesis were adapted to enable a standardized comparison across all frameworks. Specifically, the AutoRadiomics-defined data splits were applied for both training and testing, whereas the manuscript used the default RPTK split configuration. In the thesis experiments, 1,598 CT scans (80%) were allocated for training and 399 CT scans (20%) for testing. In contrast, the manuscript employed a 70/30 split, resulting in 1,397 training and 600 test scans. The larger training proportion used here increases the number of samples available for model optimization but slightly reduces the statistical confidence of the generalization estimate due to the smaller test set. Both strategies are valid and commonly used in machine learning research, reflecting a trade-off between training stability and evaluation robustness.

Another key difference concerns the inclusion of clinical covariates. While the manuscript incorporated selected clinical parameters alongside imaging data, the present thesis intentionally restricts the analysis to CT-based radiomics features only. This design isolates the contribution of image-derived biomarkers and facilitates a fair methodological comparison to other purely imaging-based frameworks. As reported in the manuscript, the inclusion of additional clinical variables did not substantially improve predictive performance for colorectal neoplasia detection, suggesting that radiomics features capture the dominant discriminative signal in this context. The results presented in this thesis are therefore fully consistent with the manuscript findings while reflecting a distinct experimental focus on the technical evaluation of framework performance.

6.3.1 Data

The LiverCRC dataset represents the largest cohort analyzed within this thesis and forms the basis for investigating the potential of RPTK to detect colorectal neoplasia using liver imaging. The study aims to identify colorectal neoplasia encompassing the spectrum from benign adenomatous polyps to malignant colorectal carcinomas through quantitative features extracted from liver CT scans. The rationale for this approach builds on the established clinical understanding that most colorectal cancers originate from benign polyps and can be prevented through early detection and removal during screening colonoscopy [246]. However, colonoscopy participation rates among high-risk populations remain suboptimal due to procedural anxiety, discomfort,

and logistical barriers [247]. By leveraging routinely acquired abdominal CT imaging as a non-invasive alternative or adjunct, the LiverCRC study aims to improve early risk assessment and promote screening adherence.

The initial cohort comprised 6,331 patients who had undergone both CT imaging and colonoscopy within the same clinical setting (see Figure 4.11). A series of rigorous inclusion and exclusion criteria were applied to ensure data integrity and clinical relevance. Patients were excluded if their colonoscopy was incomplete or missing, resulting in the absence of a definitive diagnostic label for colorectal neoplasia, or if the corresponding CT scan was unavailable or did not sufficiently cover the liver region. Additional exclusions were made for patients with oncological diseases unrelated to the colon, prior liver transplantation, or liver pathologies secondary to alcohol abuse or systemic treatments, which could confound hepatic texture and morphology. These criteria eliminated 2,295 patients, resulting in a final study population of 1,997 patients, 808 with confirmed colorectal neoplasia and 1,189 without. This sample size represents an order of magnitude increase compared to the smaller open-source datasets analyzed earlier in this thesis (ranging from 73 to 247 patients), providing a substantially stronger statistical basis for model training and evaluation.

The unique design of this study introduces an inherent cross-organ modeling challenge: the ROI for radiomics feature extraction is the liver, whereas the target condition—the presence or absence of colorectal neoplasia—originates in the colon. This indirect relationship introduces potential bias, as hepatic characteristics are influenced by a wide range of systemic, metabolic, and inflammatory factors not exclusively linked to colorectal pathology. Nonetheless, emerging evidence supports the relevance of the gut–liver axis in colorectal cancer pathophysiology. Studies have demonstrated that hepatic necrosis, metabolic dysfunction, and microenvironmental alterations are associated with increased risk of colorectal carcinoma and preferential metastatic spread to the liver [248]. Therefore, the liver provides a plausible systemic biomarker site for non-invasive colorectal disease risk assessment.

Segmentation of the liver was performed by colleagues using the MultiTalent segmentation tool [27], which yielded high-quality, organ-level masks with minimal fragmentation. The number of connected components was close to one across all cases, indicating a low prevalence of segmentation artifacts (see Table 5.4). The mean slice thickness was 5 mm, with low standard deviation, reflecting the use of a standardized abdominal CT protocol. The mean ROI size was approximately 602,858 voxels, substantially larger than the tumor-based ROIs in the open-source datasets, as expected for a whole-organ segmentation. High variance in liver volume was observed, consistent with physiological variability and pathological enlargement (hepatomegaly),

which can occur as a secondary manifestation of systemic or colorectal disease [249]. The average number of discretization bins after applying the fixed bin width of 25 was 13 (standard deviation 6.5), following the bin width recommendations from [1]. This distribution suggests sufficient grey-level diversity for texture analysis without excessive discretization noise.

In contrast to conventional liver imaging studies that typically employ magnetic resonance imaging (MRI), the LiverCRC dataset is based entirely on CT scans. While this choice improves grey-level standardization and harmonization across patients (see Section 2.1.1), it may reduce sensitivity to soft-tissue contrast variations that are better captured via MRI [15]. Nevertheless, the consistent CT acquisition protocol ensures high reproducibility of radiomic features and supports cross-patient comparability.

Feature extraction for this dataset was intentionally limited to the original CT images without additional image transformations or segmentation perturbation for feature-stability filtering. This decision was made for two main reasons: (i) the large dataset already provides substantial statistical power, reducing the necessity for data augmentation, and (ii) the computational cost of performing full feature extraction with multiple transformations and perturbations would exceed practical resource limits (estimated to require over twenty times more CPU hours and thirty times more storage). While the omission of feature-stability filtering may slightly reduce robustness, the extensive sample size compensates by providing a strong empirical basis for model training. Future iterations of the framework could incorporate more efficient processing of data augmentations for application to big datasets like the LiverCRC dataset.

In summary, the LiverCRC dataset provides a large, high-quality foundation for evaluating the generalizability of RPTK. Its design bridges radiomics and clinical screening by linking liver imaging to colorectal neoplasia risk, a task that is both unconventional and clinically relevant within the context of preventive oncology. The combination of large sample size, standardized imaging, and organ-level segmentation establishes this dataset as a valuable resource for assessing scalability, reproducibility, and clinical applicability of radiomics-based prediction frameworks.

Radiomics Feature Processing on the LiverCRC Dataset

The analysis of the selected feature spaces from PyRadiomics and MIRP in the LiverCRC study reveals a consistent dominance of texture-related descriptors, particularly from the GLCM and GLRLM feature classes. These feature families quantify spatial relationships between grey levels and are commonly associated with tissue hetero-

geneity, fibrosis, or microstructural irregularities—factors that may indirectly reflect systemic alterations linked to colorectal neoplasia. The convergence of both extraction frameworks on texture-based features therefore supports their biological plausibility for this task.

Despite this agreement, distinct methodological differences between the extractors influence the specific feature composition. MIRP produced a larger fraction of 2D and margin-based features, highlighting perihepatic and surface-related intensity variations, whereas PyRadiomics yielded a broader spectrum of texture, first-order, and shape descriptors. The higher concentration of high-impact SHAP values within a small subset of PyRadiomics features indicates a more compact but stronger predictive signal, whereas MIRP showed a more distributed importance pattern, suggesting a wider but less focused representation of relevant image characteristics. These differences reflect not only extractor implementation choices—such as the inclusion of 2D versus 3D feature formulations—but also potential differences in how each framework handles discretization, normalization, and ROI boundary effects.

Among all selected features, the first-order Root Mean Square (RMS) feature emerged as the most influential predictor across both frameworks (see Equation 6.1). RMS measures the average magnitude of voxel intensities within the ROI, corrected for negative Hounsfield Unit (HU) values, and can thus be interpreted as a proxy for the mean tissue density of the liver. Its consistent selection across multiple models and datasets indicates that overall hepatic intensity levels, rather than higher-order texture patterns, may carry significant diagnostic information in detecting systemic effects of colorectal neoplasia. This aligns with findings from the joint manuscript [22], where RMS and related first-order features were also among the top-ranked predictors.

Following the definition from the official PyRadiomics documentation [15], the RMS feature is calculated as:

$$\text{RMS} = \sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}\left(X(i)+c\right)^2} \tag{6.1}$$

$X(i)$ Intensity value of voxel i within the ROI, expressed in HU for CT images.

N_p Total number of voxels within the ROI.

c Intensity correction factor added to handle negative HU values (applied by PyRadiomics when enabled for CT data).

RMS Root Mean Squared intensity value, representing the mean energy or magnitude of voxel intensities within the ROI.

Notably, approximately half of the selected features overlap between the present analysis and the manuscript results, despite differences in training and test splits. This

overlap demonstrates the reproducibility of feature selection in RPTK and indicates that the observed feature relevance is not purely data-partition dependent. At the same time, the remaining variation between feature subsets highlights the sensitivity of wrapper-based selection methods to data sampling, an inherent limitation in radiomics analyses that rely on small or imbalanced datasets.

A direct comparison between the feature spaces identified by RPTK and AutoRadiomics on the LiverCRC dataset highlights the influence of optimization strategy and extraction settings on the resulting feature composition. While both approaches were based on PyRadiomics features, RPTK was applied to untransformed images, whereas AutoRadiomics used the default extraction configuration including multiple image filters and transformations. Consequently, AutoRadiomics selected a considerably larger and more transformation-dominated feature space, comprising 35 features, primarily first-order and GLSZM descriptors derived from wavelet and Laplacian-of-Gaussian filtered images. In contrast, RPTK identified only 15 features from the original images, mainly texture descriptors from first-order, GLCM and GLRLM classes. The overlap between both feature spaces was limited to a small number of shared feature concepts such as *RootMeanSquared*, which AutoRadiomics selected in its wavelet-transformed form. In addition, RPTK includes features from the surrounding region which are substantially included in the selected PyRadiomics feature space from RPTK and therefore also extents the source of information compared to AutpRadiomics.

Despite the differing feature selection breadth, both methods converged on texture- and first-order-related image descriptors as the dominant predictive factors. The broader and transformation-rich selection of AutoRadiomics may reflect its built-in hyperparameter search strategy, which explores redundant variants of similar image patterns, whereas RPTK's sequential selection procedure yields a more compact feature subset emphasizing complementary and non-redundant characteristics, based on intensive feature filtering prior to the selection step. The absence of distinct clustering patterns in the AutoRadiomics heatmap compared to the clearer class separation observed for the RPTK PyRadiomics feature matrix further supports that the latter approach captures a more discriminative but less redundant representation of the underlying radiomic signal. Together, these findings underline that the choice of feature selection framework substantially affects the balance between feature diversity, redundancy, and interpretability, even when identical extraction libraries are used.

Overall, the selected features from RPTK emphasize the diagnostic importance of both global intensity measures and fine-grained texture characteristics for capturing systemic alterations of the liver associated with colorectal neoplasia. Their reproducibility across independent extraction tools and experiments supports the method-

ological robustness of the RPTK framework in identifying stable and interpretable imaging biomarkers.

6.3.2 RPTK Performance on the LiverCRC Dataset

The evaluation of the RPTK framework on the LiverCRC dataset demonstrates the scalability and stability of the self-configuring radiomics pipeline when applied to a large clinical cohort. Figure 5.19 and Figure 5.20 display the validation and test AUROC distributions across all trained models for both PyRadiomics and MIRP feature extractors, based on the synchronized AutoRadiomics train/test splits. Each point represents the average performance from five cross-validation folds, while the final ensemble model performance on the independent test set is indicated as a single aggregated measure.

Overall, the ensemble model predictions exhibit narrow confidence intervals for both validation and test performance, indicating robust model convergence and stable predictions across folds. This stability suggests that the available training data are sufficient to capture the variance within the dataset and to achieve consistent generalization. The corresponding learning curve (see Figure 8.25) confirms this interpretation, showing that the model performance plateaus as the training data proportion increases, with minimal variance across repetitions.

Across the six model architectures evaluated within RPTK, the validation AUROC for PyRadiomics-based features reached approximately 0.80, with the Random Forest classifier achieving the highest mean validation score of 0.815 ± 0.016, slightly outperforming the other models. The MIRP-based models yielded similar validation results, although their variance across folds was marginally higher, consistent with the broader feature space composition of MIRP.

Interestingly, for both extractors, the test AUROC values exceeded the validation scores, with the best-performing PyRadiomics model achieving a mean test AUROC of 0.859 compared to a validation score of 0.815. This finding contrasts with the corresponding results reported in the LiverCRC manuscript, where the best model, an XGBoost classifier trained using the default RPTK split, achieved a slightly lower test performance of 0.805 against a validation score of 0.831 ± 0.014. The difference arises primarily from the proportion of data allocated to training and testing: the AutoRadiomics setting applies an 80/20 split, providing more training samples but a smaller and potentially less representative test set, whereas the RPTK default uses a 70/30 split that better captures population heterogeneity in the test data.

In practical terms, this indicates that the AutoRadiomics splitting strategy may underestimate model performance on unseen data due to reduced test-set variance,

while the RPTK setting offers a more conservative but realistic generalization estimate. These differences are consistent with known effects of train/test proportion on performance metrics: larger test sets yield more reliable generalization estimates, whereas larger training sets reduce bias but increase the risk of optimistic performance evaluation [145]. Given the large sample size of the LiverCRC cohort, both strategies remain statistically valid, but the RPTK split provides better alignment between validation and test results, suggesting stronger internal consistency.

A comparison of sensitivity and specificity metrics supports this interpretation. Using the AutoRadiomics split, the best-performing RPTK model achieved a sensitivity of 0.780 and a specificity of 0.807 (see Table 5.5), outperforming the RPTK-trained model based on its native split configuration, which achieved 0.741 and 0.723, respectively. This corresponds to a delta of +0.039 in sensitivity and +0.084 in specificity, demonstrating that the additional training samples from the AutoRadiomics split improved the model's discriminative ability. Taken together, these findings underline the robustness of RPTK in large-scale clinical data applications and its capacity to maintain stable performance across different train/test partition strategies. The consistent validation-to-test agreement, narrow confidence intervals, and high reproducibility across feature extractors indicate that the framework can generalize effectively while remaining resilient to data-split variability, a critical property for future multi-institutional extensions of this study.

6.3.3 RPTK Performance Comparison on the LiverCRC Dataset

Performance comparison of the best RPTK model to the best AutoRadiomics and deep learning models is shown in Figure 5.21. Both RPTK and AutoRadiomics demonstrate consistent performance between the validation and test sets, with RPTK achieving slightly higher AUROC values as discussed above. In contrast, the best deep learning model shows strong overfitting, achieving a validation performance of approximately 0.95 AUROC but dropping to 0.6 on the test set. RPTK outperformed both AutoRadiomics and deep learning significantly according to the paired DeLong test ($p < 0.001$), while the difference between AutoRadiomics and deep learning was not significant ($p = 0.106$).

Figure 5.22 illustrates the corresponding ROC curves for the best-performing models on the LiverCRC test set, highlighting characteristic differences in classification behavior between RPTK, AutoRadiomics, and the deep learning baseline. While the quantitative AUROC values already demonstrate a clear performance advantage of RPTK (0.859 ± 0.02) compared to AutoRadiomics (0.654 ± 0.03) and deep learning (0.598 ± 0.04), the qualitative shape of the ROC curves provides further insight into

model behavior.

Across all operating points, the RPTK ROC curve consistently lies above those of AutoRadiomics and deep learning, indicating improved sensitivity for nearly all false-positive rates. At low false-positive regions (< 0.2), RPTK achieves a considerably higher true-positive rate, demonstrating its ability to detect positive cases with minimal false alarms—a key advantage in clinical screening tasks where specificity is critical (see also Tables 8.8 and 8.9). Moreover, the RPTK curve exhibits a steeper initial rise near the origin, reflecting superior discriminative power for highly confident predictions. In contrast, the flatter slopes observed for AutoRadiomics and deep learning suggest less distinct class separation. Finally, the RPTK curve approaches the ideal upper-left corner (0,1) more closely than the alternatives, visually confirming a better balance between sensitivity and specificity.

Taken together, these results emphasize that RPTK not only achieves higher overall discrimination as measured by AUROC, but also maintains a more favorable balance between true- and false-positive rates across clinically relevant thresholds. This robustness suggests that the combination of feature selection, model optimization, and ensemble aggregation implemented in RPTK contributes to more stable and generalizable predictions compared to conventional automated radiomics or deep learning baselines.

To contextualize these findings within the current literature, it is noteworthy that no peer-reviewed study was identified that applies radiomics on liver imaging to predict colorectal disease in the reverse (liver-to-colon) direction. Existing research instead focuses on the opposite relationship—predicting liver metastases or liver-specific outcomes based on colorectal cancer data. For example, studies by Yu et al. [250], Tang et al. [251], and Devoto et al. [252] investigate radiomics-based prediction of metachronous or synchronous liver metastases from colorectal primaries, typically using CT or MRI imaging of the primary tumor or existing liver lesions, with dataset sizes ranging from 80 to 250 patients and target labels linked to metastatic progression. These works differ fundamentally in both the region of interest (focusing on colorectal lesions or liver metastases rather than the healthy liver) and the prediction objective (metastatic risk or treatment outcome rather than initial colorectal pathology).

In contrast, the LiverCRC study employs a substantially larger cohort ($n = 1{,}997$) and investigates systemic image-derived biomarkers from the liver parenchyma to infer colorectal neoplasia risk. This approach has, to the best of current knowledge, not been previously reported in the literature. This further underscores the methodological novelty and translational potential of using liver-based radiomics as a non-invasive biomarker source for colorectal disease screening.

Chapter 7

Conclusion and Outlook

7.1 Summary of Key Findings

The results of this thesis demonstrate that the developed RPTK (Radiomics Pipeline Toolkit) constitutes a robust, self-configuring, and task-agnostic framework for automated radiomics analysis. RPTK was designed to improve the reproducibility and generalizability of radiomics workflows while maintaining high predictive performance across diverse clinical applications.

RPTK was successfully applied to seven heterogeneous open-source datasets, covering both CT and MRI modalities and a wide range of oncological classification tasks. Across these datasets, RPTK consistently outperformed or matched both AutoRadiomics and deep learning baselines, with particularly major advantages in small-sample and heterogeneous imaging settings. This emphasizes its capability to generalize effectively even under conditions of data scarcity and domain variability.

Overall, the framework bridges an important methodological gap between highly customized radiomics pipelines, often tailored for specific studies and therefore difficult to reproduce, and generalized AutoML approaches such as AutoRadiomics, which prioritize automation at the expense of optimization. By combining adaptive preprocessing, standardized feature extraction, and automated model optimization, RPTK achieves a balance between generalizability and methodological transparency (see Section 1.1 and Figure 1.1).

The feature extraction and selection mechanisms of RPTK consistently identified robust and biologically meaningful radiomic descriptors, particularly texture-based classes such as GLCM and GLSZM features. These findings demonstrate that RPTK captures reproducible imaging biomarkers across datasets and imaging modalities.

A comprehensive cross-validation and ensemble learning strategy ensured stability and generalization of the trained models, minimizing variance between validation

and test performance. In contrast, the learning-curve analysis revealed that several datasets were too small to reach performance plateaus or highly variable performance, highlighting the inherent data limitations of current radiomics benchmarks rather than weaknesses of the framework itself.

The integration of delta radiomics in the longitudinal Predict study confirmed the predictive benefit of modeling temporal feature changes over static single-time-point features in RPTK. Furthermore, combining clinical covariates with radiomics features yielded additional performance improvements, underscoring the complementary nature of quantitative imaging and clinical information in treatment-response prediction.

Finally, the large-scale LiverCRC study extended RPTK's application to a dataset of approximately 2,000 CT scans, demonstrating the scalability and transferability of the framework to real-world clinical imaging cohorts. In this study, RPTK achieved significantly higher performance than both AutoRadiomics and deep learning approaches, confirming its robustness and adaptability in large, heterogeneous datasets.

7.2 Methodological Contributions of RPTK

The methodological innovations of this thesis are consolidated in the development of RPTK (Radiomics Pipeline Toolkit), a modular and reproducible framework designed for automated radiomics experimentation across diverse imaging modalities and clinical tasks. RPTK combines adaptive preprocessing and feature extraction with robust feature filtering and ensemble-based modeling in a transparent and extensible architecture.

A key contribution is the introduction of a self-configuring feature extraction and preprocessing system. Rather than applying a static configuration across all studies, RPTK automatically adapts feature-extraction parameters and preprocessing settings to the image modality (e.g., CT vs. MRI) and uses literature based recommendations for robust data processing. This includes modality-specific resampling, normalization as well as feature extraction tool recommendation for re-segmentation and Grey value correction (see e.g. Equation 6.1), which are selected according to modality-dependent best practices collected from the literature (see Section 4.1.5). This targeted form of self-configuration ensures optimal use of modality-specific information while maintaining comparability and reproducibility across datasets.

To quantitatively describe the imaging data and support these adaptive steps, RPTK integrates a data fingerprinting module that automatically computes descriptive metrics of the input data, including voxel spacing, slice thickness, number of bins

after pixel discretizations, and segmentation properties. These fingerprints serve as quality indicators and allow objective assessment of dataset heterogeneity, supporting reproducible preprocessing decisions and model interpretation.

Another major methodological contribution is the implementation of a robust feature stability filtering mechanism designed to mitigate segmentation-related bias. RPTK simulates inter-rater variability by introducing controlled perturbations to segmentation masks and evaluating feature stability across these perturbations. Features that show high sensitivity to small segmentation variations are excluded from downstream modeling, thereby improving feature reproducibility and reducing dependence on specific segmentation practices or annotators.

In addition, RPTK provides a standardized feature-extractor integration layer that supports two different existing feature extraction tools called PyRadiomics and MIRP. This dual integration enables systematic cross-tool benchmarking and facilitates evaluation of feature completeness and consistency with the Image Biomarker Standardisation Initiative (IBSI) recommendations. By comparing the outputs of both extractors within the same preprocessing and modeling framework, RPTK enables reproducible feature-space analyses and enhances methodological transparency as well as incorporates performance benefits by the inclusion of different feature aggregation strategies.

For predictive modeling, RPTK implements a cross-validation model optimization and ensemble strategy. Multiple machine learning algorithms are independently optimized on features from both integrated extractors through cross-validation. The fold models are ensembled for each machine learning algorithm to improve robustness and minimize overfitting. Subsequently, the best performing machine learning algorithm was selected for final performance evaluation. This design leads to more stable and generalizable models compared to single-model approaches, particularly in heterogeneous or small-sample datasets.

All components of RPTK follow transparent and reproducible design principles aligned with current radiomics reporting and methodological guidelines. Every preprocessing, feature extraction, and modeling step is explicitly logged and reproducible, ensuring methodological traceability across studies.

The framework further facilitates controlled cross-framework benchmarking, enabling fair comparisons with alternative pipelines such as AutoRadiomics and deep learning models using identical data splits and evaluation metrics.

Finally, RPTK is open-source and publicly available via a dedicated GitHub repository (see Section 4.1.11). This open-access release promotes transparency, reproducibility, and collaborative extension by the wider research community.

7.3 Clinical and Translational Implications

The results presented in this thesis demonstrate that the RPTK framework provides a robust foundation for the clinical translation of radiomics-based decision-support systems. Its reproducible processing design, optimized preprocessing, and robust feature engineering collectively support the development of predictive models that are not only statistically sound but also potentially deployable in clinical workflows.

RPTK has shown the ability to extract quantitative imaging biomarkers that correlate with disease characteristics and treatment response, thereby supporting the concept of imaging as a non-invasive biomarker source. By standardizing the feature extraction and selection process across heterogeneous datasets and imaging modalities, the framework minimizes methodological bias and enhances reproducibility which is a prerequisite for clinical implementation.

The Predict study demonstrated the potential of radiomics to provide early indicators of immunotherapy response in patients with non-small cell lung cancer. Through the integration of delta radiomics, temporal changes in radiomic features between baseline and follow-up CT scans were successfully modeled, allowing the prediction of therapy response at an early treatment stage. Such early response assessment could enable timely therapy adaptation, reducing unnecessary exposure to ineffective treatments and improving patient outcomes and quality of live.

The LiverCRC study extended this concept to a large-scale screening scenario and illustrated the feasibility of using systemic imaging biomarkers from non-disease-target organs to detect remote pathologies. By leveraging liver CT scans to predict colorectal neoplasia, the study demonstrated that radiomics information from secondary or indirectly affected organs can reflect systemic disease manifestations. This approach offers a promising path toward non-invasive and opportunistic cancer screening strategies, potentially complementing or guiding standard diagnostic procedures such as colonoscopy.

Furthermore, the integration of radiomics features with clinical variables improved both model interpretability and predictive performance, as demonstrated in the Predict study. The combination of quantitative imaging biomarkers with established clinical indicators provides a more comprehensive representation of disease status and enhances the clinical utility of radiomics-based models. In this regard, RPTK supports the integration of radiomics into multimodal clinical decision-making by enabling reproducible model development, validation, and interpretation.

7.4 Limitations

Despite the promising results and methodological advances presented in this thesis, several limitations have to be acknowledged. These limitations, however, not only define areas for improvement but also highlight the robustness and adaptability of the RPTK framework and the opportunities it create for future research and clinical translation.

A key limitation lies in the restricted size of some datasets, particularly in the Predict study focusing on immunotherapy response prediction. The acquisition of longitudinal imaging data is inherently expensive, time-consuming, and suitable open-source datasets are scarce. As a result, the statistical power for detecting small but clinically relevant effects is limited, and model generalization may be constrained. Nonetheless, this limitation underscores one of RPTK's core strengths, its ability to operate robustly and deliver stable results even under data-scarce conditions, where deep learning and traditional radiomics pipelines often fail to generalize.

Another important limitation concerns the lack of external multi-center validation. All datasets used in this thesis were derived from open-source repositories or single-institution sources. External, multi-center testing would provide stronger evidence for generalizability and ensure that the developed models are not biased toward specific institutional imaging characteristics or clinical practices. Integrating RPTK into future multi-center collaborations would therefore be a critical step toward clinical translation and the development of radiomics models applicable across diverse imaging environments.

Variability in segmentation quality and acquisition protocols represents an additional challenge, as it introduces technical bias that can influence feature stability and model performance. However, this variability also demonstrates RPTK's capability to handle heterogeneous data sources. Through systematic data fingerprinting and segmentation-perturbation-based feature stability filtering, the framework actively detects and mitigates these artifacts, ensuring that the extracted features remain robust to segmentation and acquisition differences.

A further methodological limitation arises from the dependence on the Image Biomarker Standardisation Initiative (IBSI)-defined feature space. While this ensures high reproducibility and standardization, the results also indicate that certain non-IBSI features may carry task-specific predictive value. However, their reproducibility remains uncertain. Future work should therefore include a systematic extension of the feature stability analysis to these non-IBSI features to evaluate their robustness and potential inclusion in standardized radiomics feature sets.

Finally, while the current implementation of RPTK focuses on computationally efficient batch processing and large-scale benchmarking, its prospective clinical application will require additional optimization for integration into routine workflows. As outlined in Section 2.3.8, RPTK could be applied in a pre-optimized form to individual patient cases after task-specific model training and validation. This would enable real-time or near-real-time decision support in clinical environments without requiring extensive retraining.

In summary, while these limitations outline areas for improvement, they simultaneously illustrate the flexibility, methodological rigor, and clinical readiness of RPTK. Future research should focus on large-scale multi-institutional validation, prospective clinical deployment, and the expansion of feature-space reproducibility testing to further advance radiomics toward routine clinical utility.

7.5 Outlook and Future Work

The results of this thesis establish the RPTK framework as a robust foundation for reproducible, interpretable, and clinically relevant radiomics research. Building on these developments, several avenues for future work can further enhance its methodological scope, generalizability, and translational applicability.

A primary objective for future studies is the multi-center validation of RPTK on large-scale, multi-institutional datasets. Such validation would confirm the framework's robustness across different scanner types, acquisition protocols, and patient populations, providing essential evidence for clinical generalization. This step is crucial for regulatory acceptance and real-world deployment of radiomics-based decision-support models.

Further progress should focus on multimodal integration. Diseases such as cancer are complex and acting on multiple different molecular layers such as genomics, proteomics, and transcriptomics. By extending the framework to incorporate complementary data sources from genomic, transcriptomic, or proteomic features, as well as structured clinical data, RPTK could support holistic modeling of such complex disease phenotypes and improve the personalization of predictive models. Such integration could bridge the gap between image-derived and molecular biomarkers, contributing to precision medicine.

An additional direction is the development of automated parameter adaptation mechanisms. Currently, discretizations and resampling parameters are determined based on best-practice recommendations from literature. Future work could employ data-driven optimization strategies to automatically infer optimal parameter settings

from dataset characteristics. However, this should avoid over-automation, as demonstrated by performance declines in fully AutoML-based workflows, which can overfit combined decision layers and reduce performance across datasets.

From a methodological standpoint, deep learning hybridization represents an exciting opportunity. Combining RPTK's interpretable, feature-based approach with learned representations from convolutional neural networks (CNNs) or transformer architectures could leverage the strengths of both paradigms interpretability and non-linear pattern recognition, while preserving robustness and transparency.

For clinical implementation, future efforts should focus on integrating RPTK into prospective studies and clinical infrastructures. The framework is conceptualized for non-specialists by incorporating a standardized workflow application without manual intervention and therefore, it is suited for clinical application by non-experts. Developing intuitive user interfaces for radiologists and oncologists would facilitate practical adoption in daily workflows even more. In particular, technical integration with the Kaapana platform for federated learning [253] would enable deployment within clinical networks such as RACOON [254], supporting privacy-preserving distributed analysis. The containerized design of RPTK, distributed via Docker through its GitHub repository, already provides a straightforward path for such integration.

Additional optimization of feature selection strategies could further simplify model architectures by reducing the inclusion of marginally contributing features, thereby improving interpretability and computational efficiency.

RPTK will continue to contribute to the principles of open science and reproducibility. The open-source release of the framework encourages community participation, benchmarking collaborations, and transparent methodological comparison. Ongoing collaborative efforts can help establish community-wide standards for radiomics evaluation and foster collective progress toward clinically validated, reproducible imaging biomarkers.

In summary, RPTK provides a scalable and extensible foundation for future developments in radiomics research and clinical translation. Its continued methodological refinement, clinical integration, and community-driven validation will be key to realizing the full potential of radiomics in precision oncology.

Chapter 8

Appendix

8.1 Overview of all Datasets in this Thesis

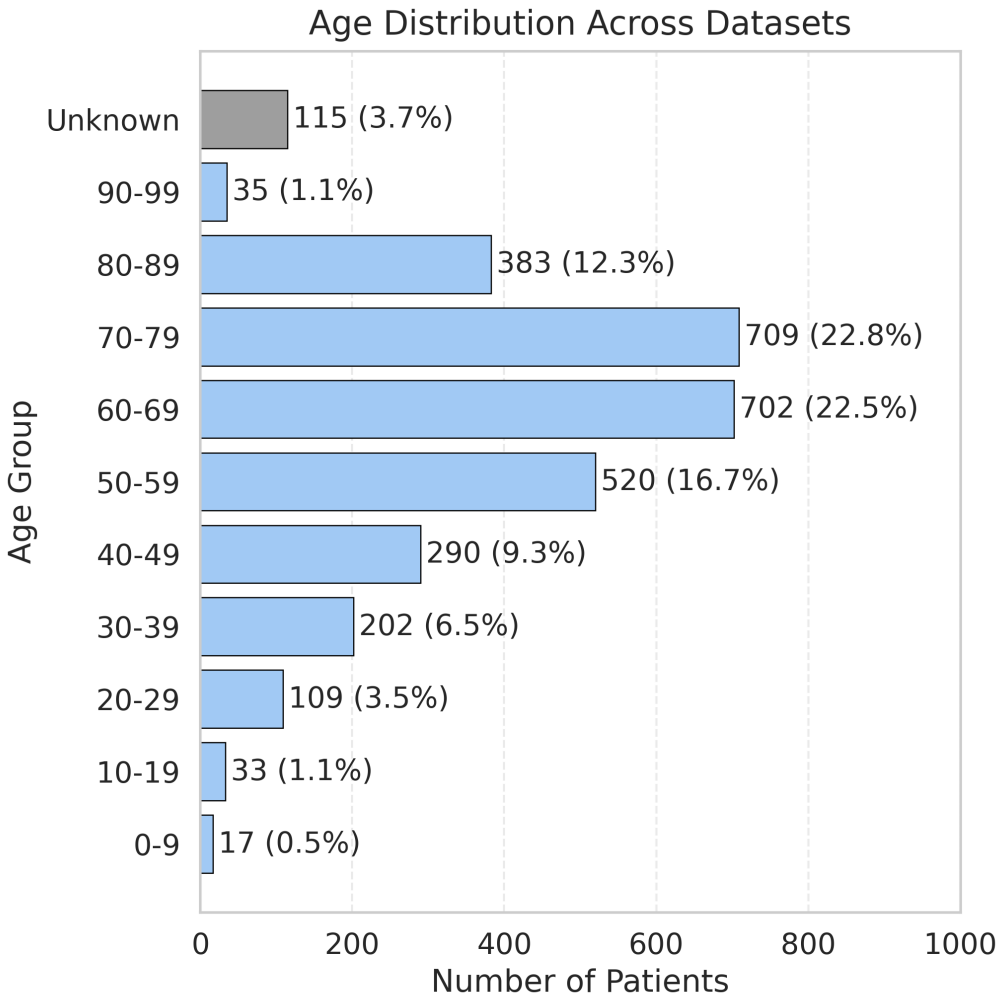


Figure 8.1. Distribution of patient age at the time point of imaging over all datasets included in this thesis.

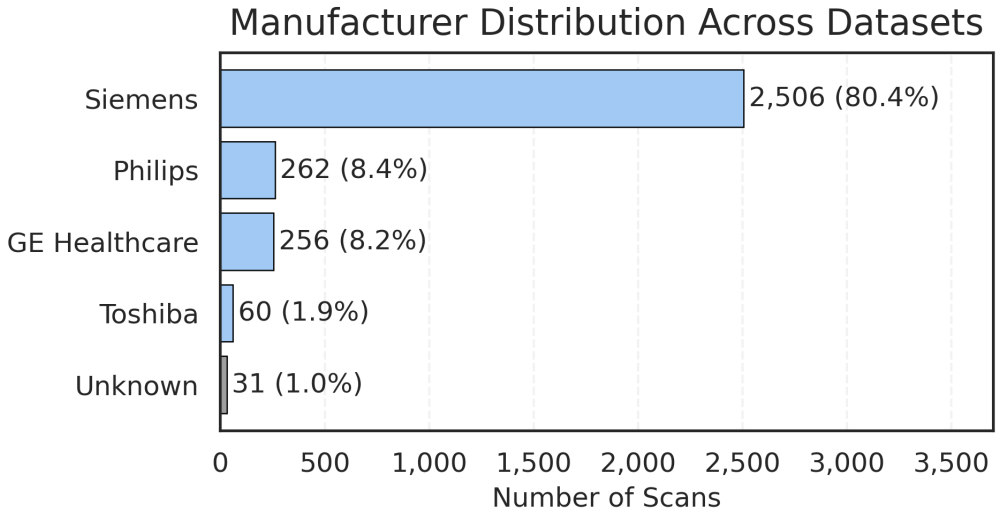
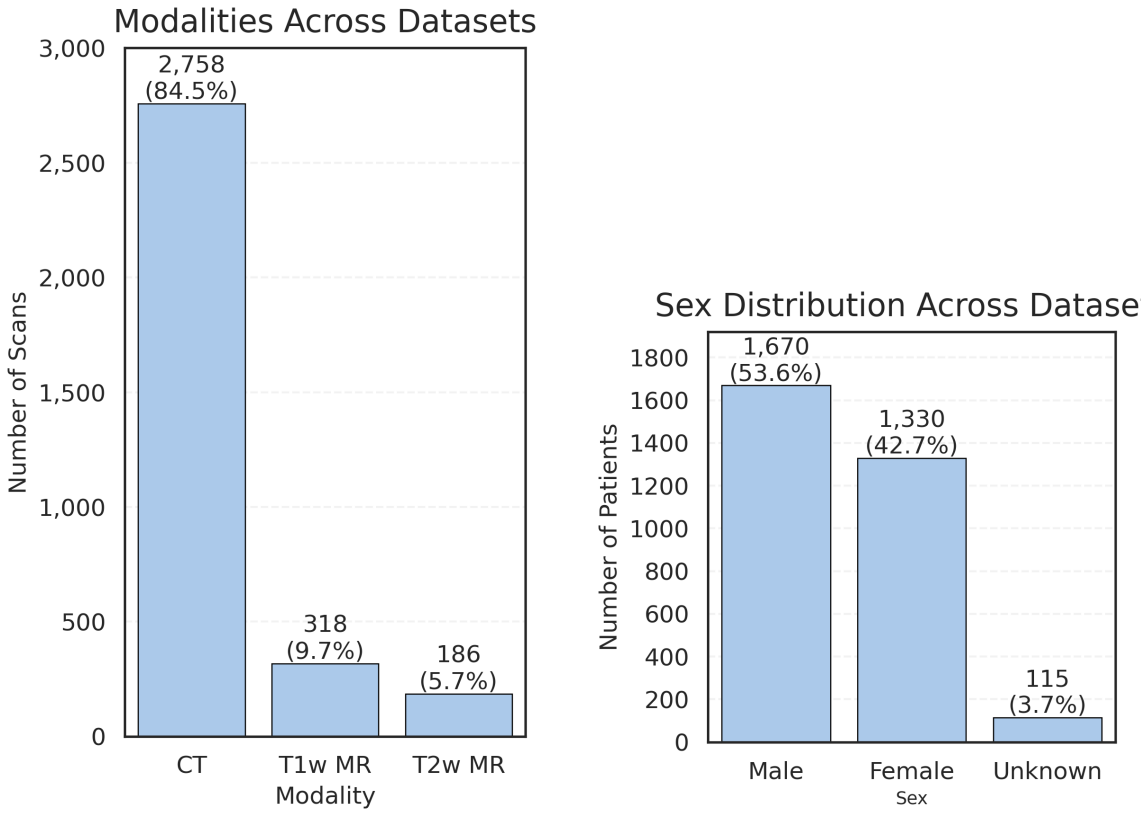


Figure 8.2. Distribution of imaging manufacturers of the scanners used to generate the 3D imaging data across all datasets.



(a) Distribution of imaging modalities across all datasets.

(b) Distribution of patient sex across all datasets.

Figure 8.3. Overview of imaging modalities as well as patient sex distribution across all used datasets in this thesis.

8.2 Self-Configuring Radiomics Pipeline

8.2.1 Data Fingerprint

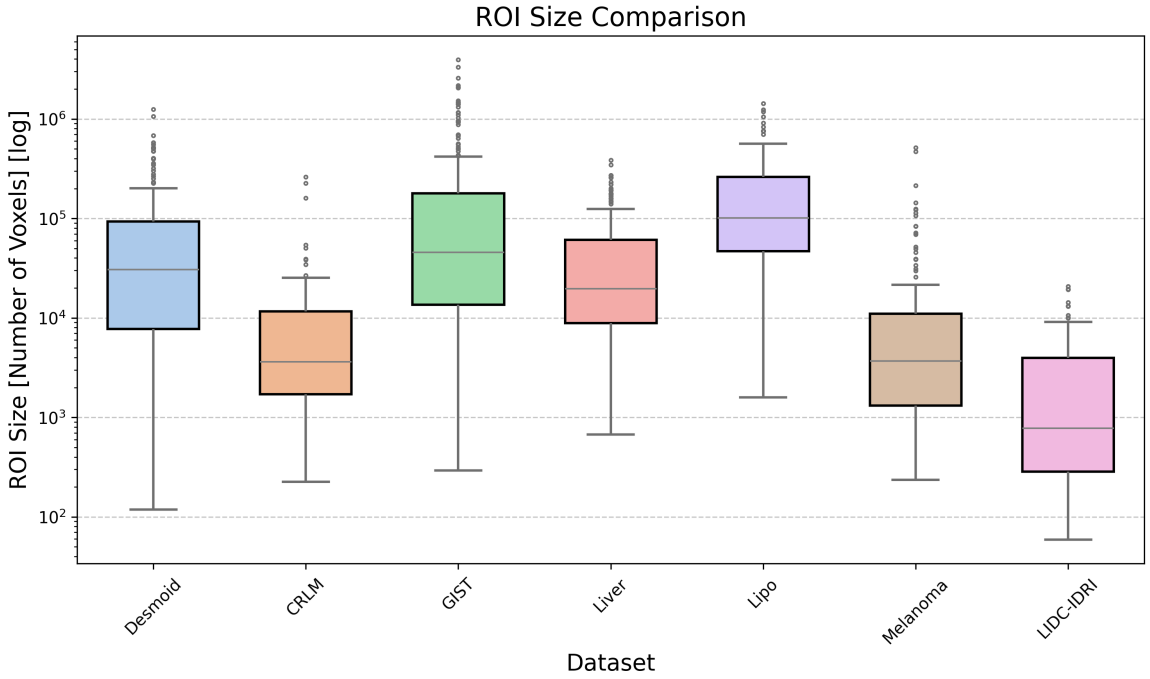


Figure 8.4. Size of the ROI measured by the number of voxels in the scans, displayed in logarithmic scale. This parameter shows how much the size of the ROI varies within the dataset in order to investigate artificially big or small ROIs but also investigate the biological heterogeneity.

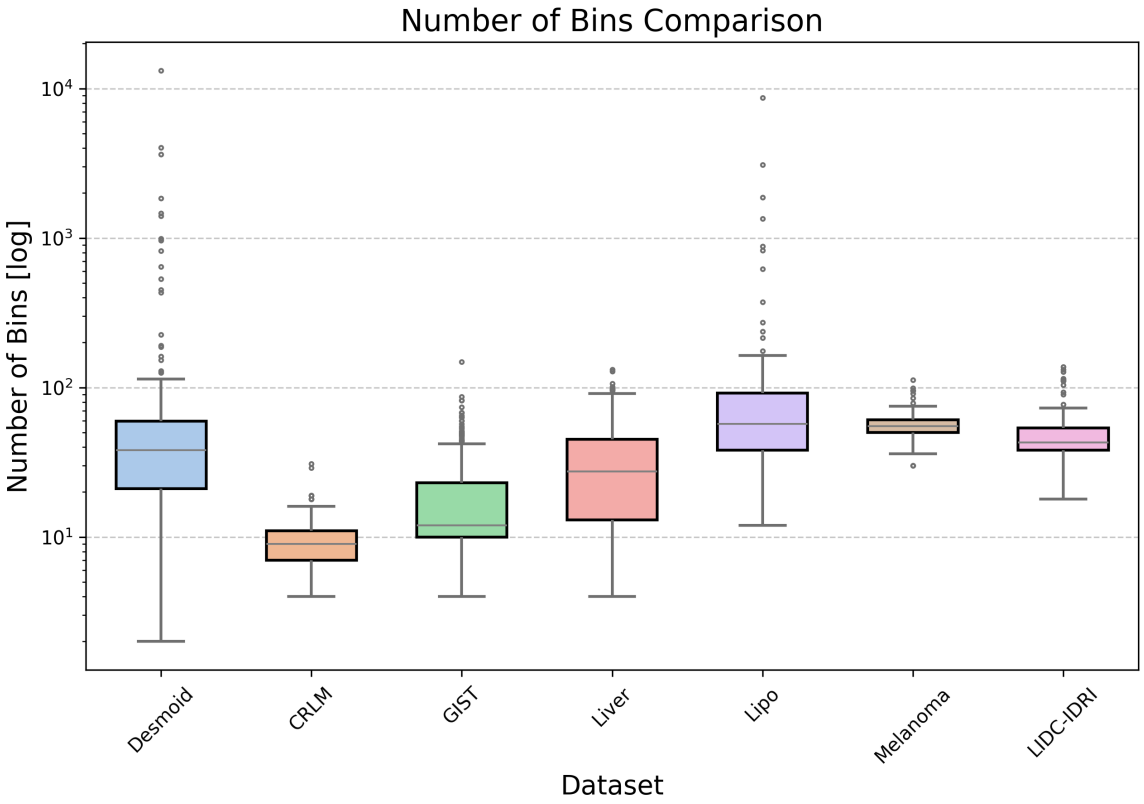


Figure 8.5. Boxplot showing the number of bins calculated with default pixel discretization settings (fixed bin width = 25) recommended by Timmeren et al. [1] in logarithmic scale. This gives a hint to the user of how heterogen the region of interest is (tumor heterogeneity) and if the standard pixel discretization setting is applicable (not very few bins (< 10) but also not too many (> 200).

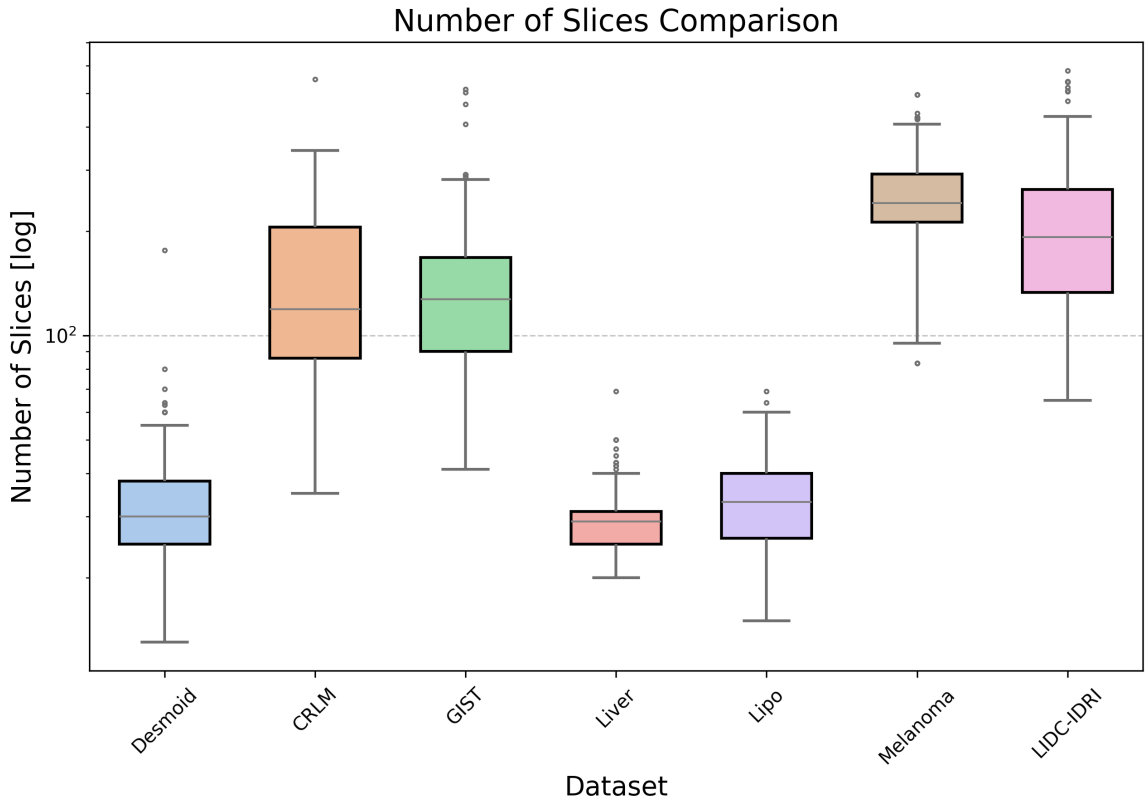


Figure 8.6. Boxplot showing the number slices in logarithmic scale between the open source datasets applied in the section for self-configuring radiomics framework. This parameter can be systematically dependent on clinical settings for different diagnostic purposes (whole body Scan vs. specifically regional scan).

8.2.2 RPTK Feature Extraction

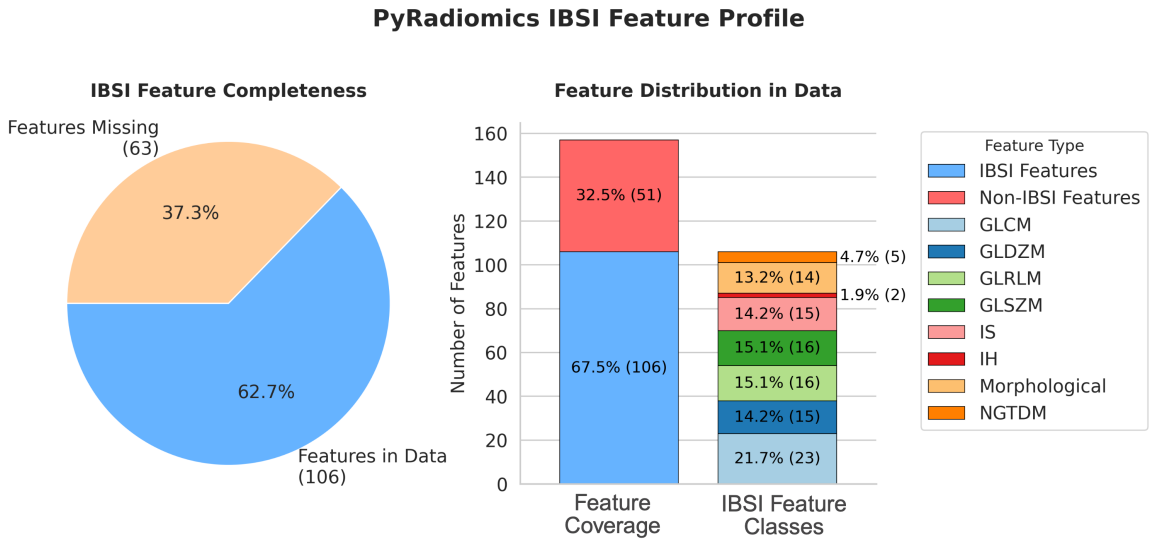


Figure 8.7. This plot shows the IBSI feature composition of the PyRadiomics features space after extraction in RPTK. The IBSI feature completeness plot (left) shows the amount of missing IBSI features in the PyRadiomics feature space with the number of features in brackets. The feature distribution plot (right) shows the detailed composition of the extracted features and the representation of IBSI feature classes in the feature space with the number of features in brackets.

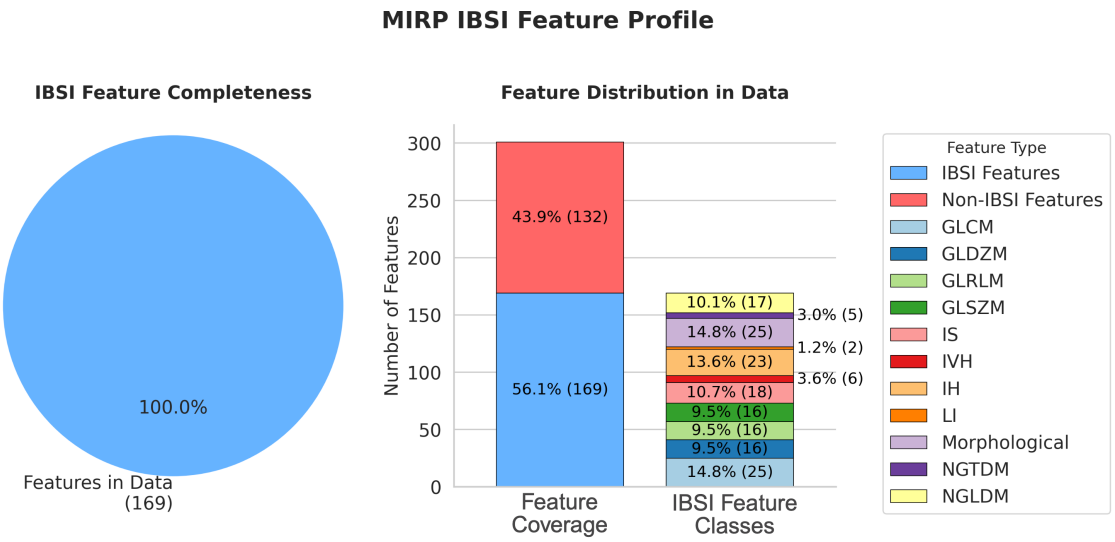


Figure 8.8. This plot shows the IBSI feature composition of the MIRP features space after extraction in RPTK. The IBSI feature completeness plot (left) shows the amount of missing IBSI features in the MIRP feature space with the number of features in brackets. The feature distribution plot (right) shows the detailed composition of the extracted features and the representation of IBSI feature classes in the feature space with the number of features in brackets.

8.2.3 The RPTK Prototype

Table 8.1. Validation AUROC performance to select the best performing models across datasets and feature extractors for the baseline RPTK approach. Bold values demonstrate better performance on validation folds for each dataset, across extractors (the validation AUROC values include unpublished data). This data was copied from my previous conference paper [24].

Dataset	Extractor	Val AUROC	Test AUROC
Desmoid	MIRP	0.695 (+/- 0.072)	0.93 [0.82-1.00]
Desmoid	PyRadiomics	**0.827 (+/- 0.053)**	**0.92 [0.82-0.99]**
CRLM	MIRP	**0.772 (+/- 0.096)**	**0.89 [0.66-1.00]**
CRLM	PyRadiomics	0.738 (+/- 0.077)	0.61 [0.28-0.89]
GIST	MIRP	0.768 (+/- 0.087)	0.79 [0.65-0.91]
GIST	PyRadiomics	**0.835 (+/- 0.021)**	**0.82 [0.69-0.92]**
Liver	MIRP	**0.805 (+/- 0.046)**	**0.79 [0.63-0.92]**
Liver	PyRadiomics	0.782 (+/- 0.027)	0.89 [0.75-0.98]
Lipo	MIRP	0.853 (+/- 0.091)	0.95 [0.82-1.00]
Lipo	PyRadiomics	**0.880 (+/- 0.106)**	**0.86 [0.67-1.00]**
Melanoma	MIRP	0.615(+/- 0.095)	0.63 [0.36-0.87]
Melanoma	PyRadiomics	**0.693 (+/- 0.116)**	**0.46 [0.18-0.73]**
LIDC-IDRI	MIRP	**0.589 (+/- 0.133)**	**0.41 [0.12-0.76]**
LIDC-IDRI	PyRadiomics	0.518 (+/- 0.190)	0.55 [0.22-0.82]

8.3 Prediction Performance

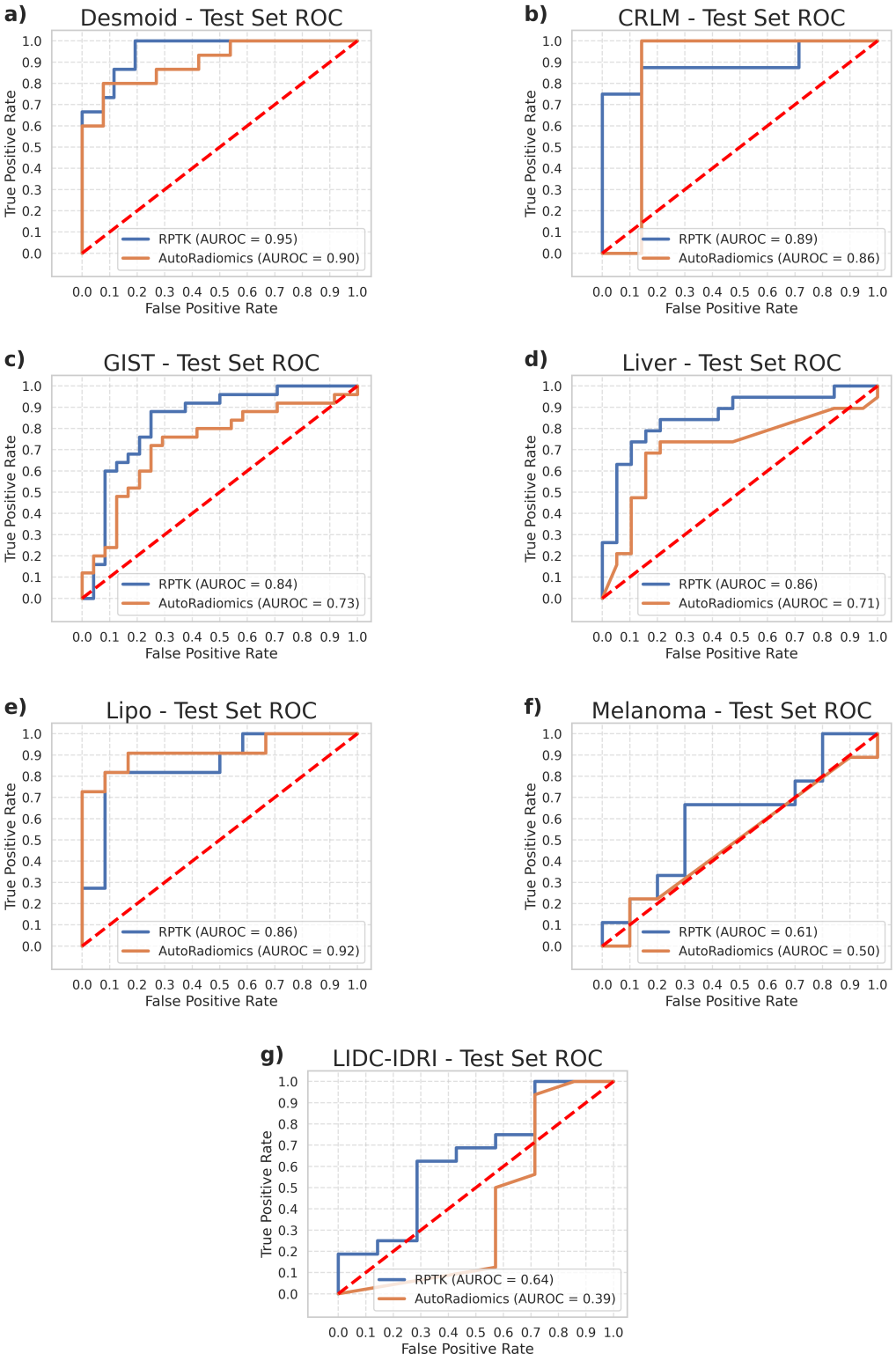


Figure 8.9. Receiver operating characteristic (ROC) comparisons between AutoRadiomics and RPTK across all seven datasets applied in section 5.1.

Table 8.2. Performance metrics of best performing deep learning models across datasets, selected based on the best mean validation AUROC (with std). Test AUROC, F1, Sensitivity and Specificity are displayed with 95% CIs after 1000x bootstrapping as described in Section 4.1.8.

Dataset	Best Model	Val AUROC	Test AUROC	Youden	Test F1	Test Sensitivity	Test Specificity
Desmoid	ResNet18	0.898 (± 0.022)	0.709 [0.538, 0.865]	0.828	0.356 [0.095, 0.600]	0.269 [0.059, 0.500]	0.885 [0.750, 1.000]
CRLM	DenseNet201	0.724 (± 0.061)	0.422 [0.143, 0.727]	0.747	0.357 [0.000, 0.632]	0.373 [0.000, 0.750]	0.369 [0.000, 0.728]
GIST	DenseNet169	0.845 (± 0.020)	0.614 [0.451, 0.778]	0.580	0.569 [0.368, 0.741]	0.483 [0.276, 0.667]	0.792 [0.619, 0.952]
Liver	ResNet18	0.938 (± 0.021)	0.841 [0.690, 0.949]	0.795	0.715 [0.545, 0.840]	0.946 [0.818, 1.000]	0.314 [0.111, 0.529]
Lipo	DenseNet121	0.766 (± 0.047)	0.723 [0.492, 0.931]	0.823	0.627 [0.333, 0.842]	0.551 [0.250, 0.846]	0.842 [0.600, 1.000]
Melanoma	DenseNet264	0.863 (± 0.020)	0.447 [0.155, 0.756]	0.658	0.492 [0.222, 0.714]	0.669 [0.364, 1.000]	0.104 [0.000, 0.333]
LIDC-IDRI	DenseNet201	0.769 (± 0.039)	0.473 [0.196, 0.769]	0.833	0.617 [0.414, 0.800]	0.623 [0.375, 0.857]	0.141 [0.000, 0.500]

Table 8.3. Performance metrics of best performing AutoRadiomics models across datasets, selected based on the best mean validation AUROC (with std). Test AUROC, F1, Sensitivity and Specificity are displayed with 95% CIs after 1000x bootstrapping as described in Section 4.1.8. Included is also the selected feature selection methods.

Dataset	Best Model	Feature Selection	Val AUROC	Test AUROC	Test F1	Test Sensitivity	Test Specificity
Desmoid	Logistic Regression	Boruta	0.782 (± 0.054)	0.902 [0.799, 0.979]	0.731 [0.526, 0.897]	0.664 [0.421, 0.875]	0.922 [0.808, 1.000]
CRLM	Logistic Regression	ANOVA	0.678 (± 0.143)	0.753 [0.429, 1.000]	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]	1.000 [1.000, 1.000]
GIST	SVM	ANOVA	0.727 (± 0.022)	0.729 [0.577, 0.862]	0.731 [0.571, 0.857]	0.721 [0.535, 0.897]	0.752 [0.565, 0.923]
Liver	XGBoost	ANOVA	0.594 (± 0.105)	0.713 [0.526, 0.883]	0.701 [0.485, 0.870]	0.633 [0.400, 0.833]	0.845 [0.667, 1.000]
Lipo	Random Forest	Boruta	0.862 (± 0.123)	0.922 [0.785, 1.000]	0.840 [0.615, 1.000]	0.735 [0.444, 1.000]	1.000 [1.000, 1.000]
Melanoma	SVM	ANOVA	0.638 (± 0.165)	0.501 [0.292, 0.722]	0.602 [0.348, 0.800]	0.890 [0.625, 1.000]	0.097 [0.000, 0.333]
LIDC-IDRI	XGBoost	ANOVA	0.456 (± 0.050)	0.392 [0.076, 0.728]	0.817 [0.647, 0.930]	1.000 [1.000, 1.000]	0.000 [0.000, 0.500]

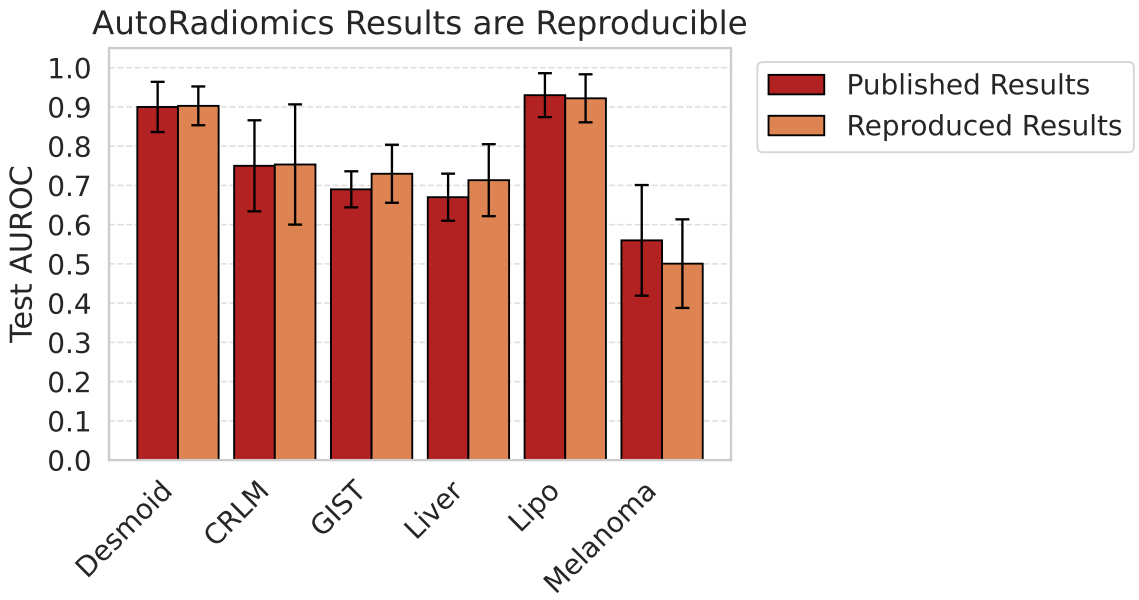
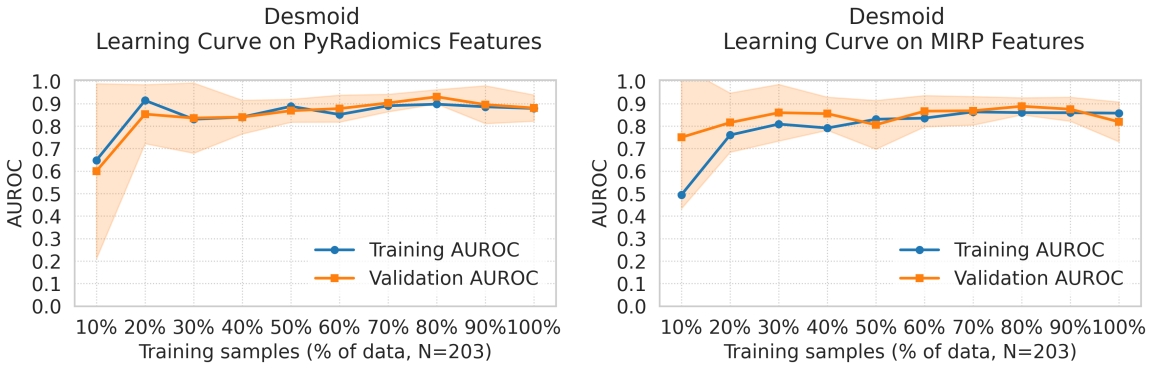


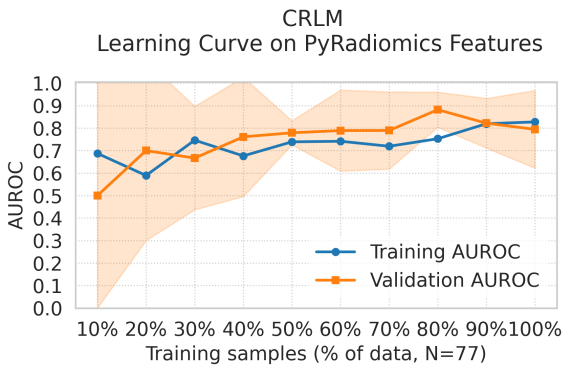
Figure 8.10. Performance comparison between published results from AutoRadiomics [18] and my reproduced results by applying AutoRadiomics on the same test set samples (the br represents the mean and the error bars are showing one standard deviation).



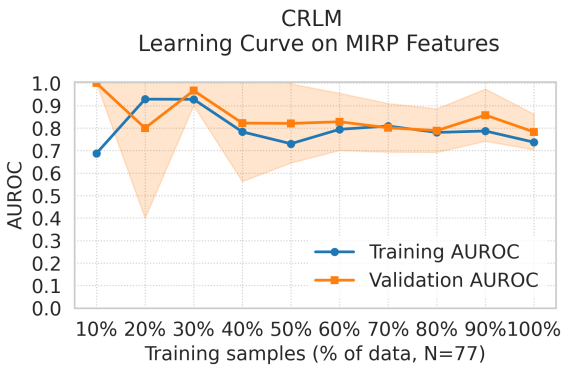
(a) Learning curve on selected PyRadiomics Features.

(b) Learning curve on selected MIRP Features.

Figure 8.11. Learning curve on selected PyRadiomics Features (a) and selected MIRP Features (b) from RPTK run. I trained a random forest classifier with 5 fold cv on the Desmoid data. I generated this figure by using the LearningCurveDisplay function from *scikit-learn* (v 1.5.0) [191]. This figures show the predictive model performance variances related to training size.

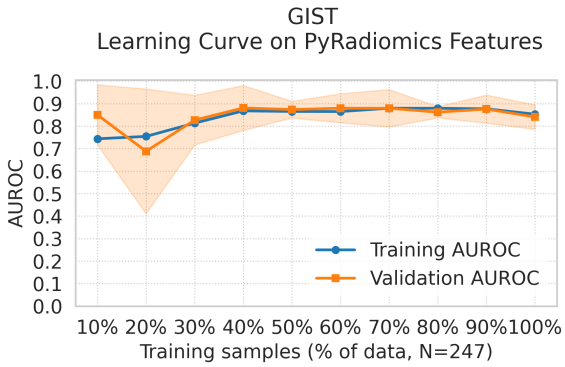


(a) Learning curve on selected PyRadiomics Features.

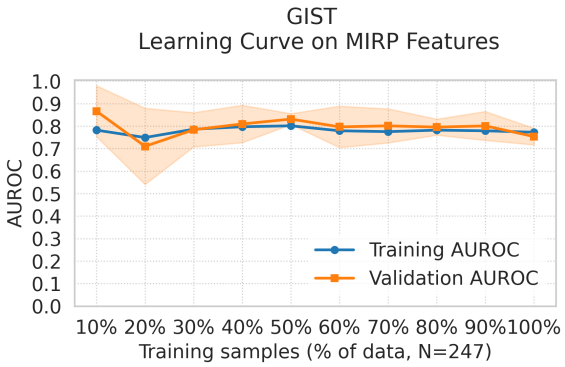


(b) Learning curve on selected MIRP Features.

Figure 8.12. Learning curve on selected PyRadiomics Features (a) and selected MIRP Features (b) from RPTK run. I trained a random forest classifier with 5 fold cv on the CRLM data. I generated this figure by using the LearningCurveDisplay function from *scikit-learn* (v 1.5.0) [191]. This figures show the predictive model performance variances related to training size.

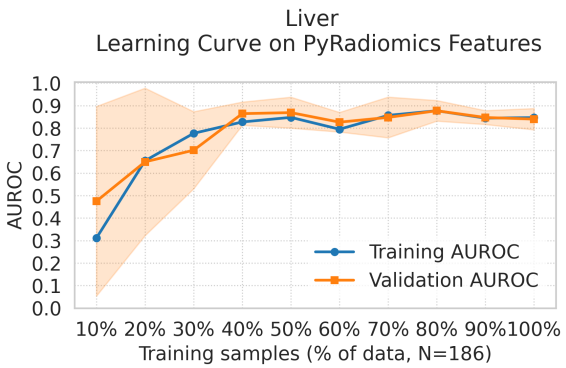


(a) Learning curve on selected PyRadiomics Features.

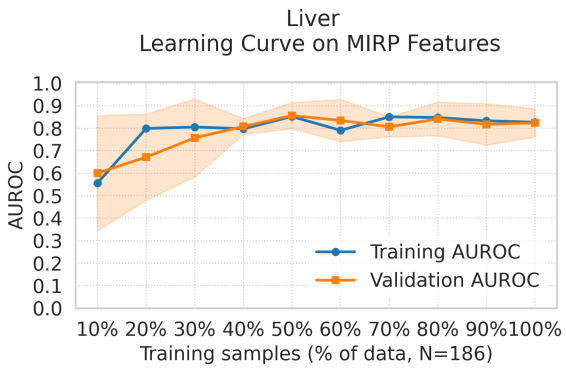


(b) Learning curve on selected MIRP Features.

Figure 8.13. Learning curve on selected PyRadiomics Features (a) and selected MIRP Features (b) from RPTK run. I trained a random forest classifier with 5 fold cv on the GIST data. I generated this figure by using the LearningCurveDisplay function from *scikit-learn* (v 1.5.0) [191]. This figures show the predictive model performance variances related to training size.

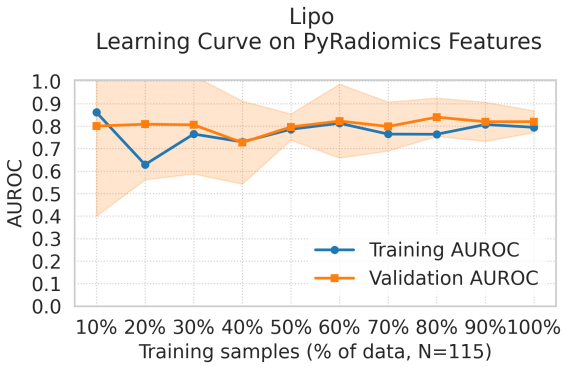


(a) Learning curve on selected PyRadiomics Features.

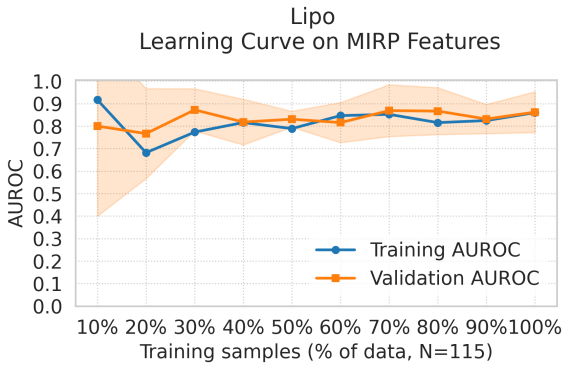


(b) Learning curve on selected MIRP Features.

Figure 8.14. Learning curve on selected PyRadiomics Features (a) and selected MIRP Features (b) from RPTK run. I trained a random forest classifier with 5 fold cv on the Liver data. I generated this figure by using the LearningCurveDisplay function from *scikit-learn* (v 1.5.0) [191]. This figures show the predictive model performance variances related to training size.

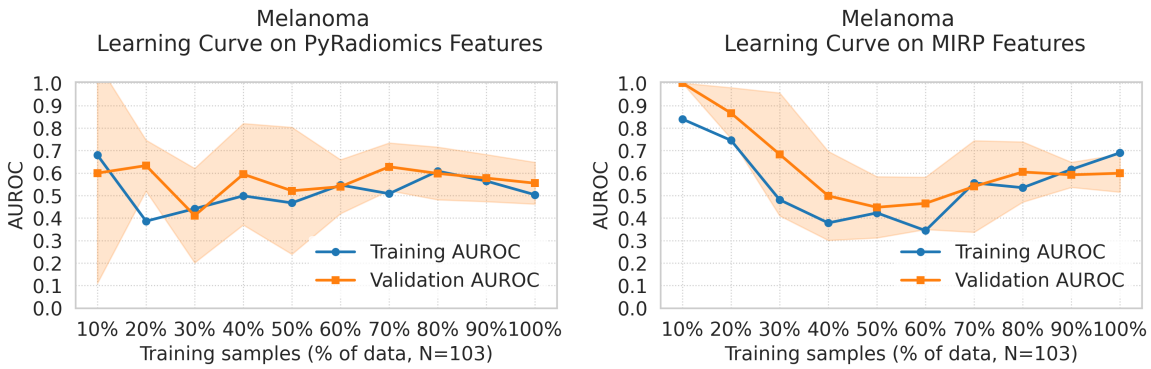


(a) Learning curve on selected PyRadiomics Features.



(b) Learning curve on selected MIRP Features.

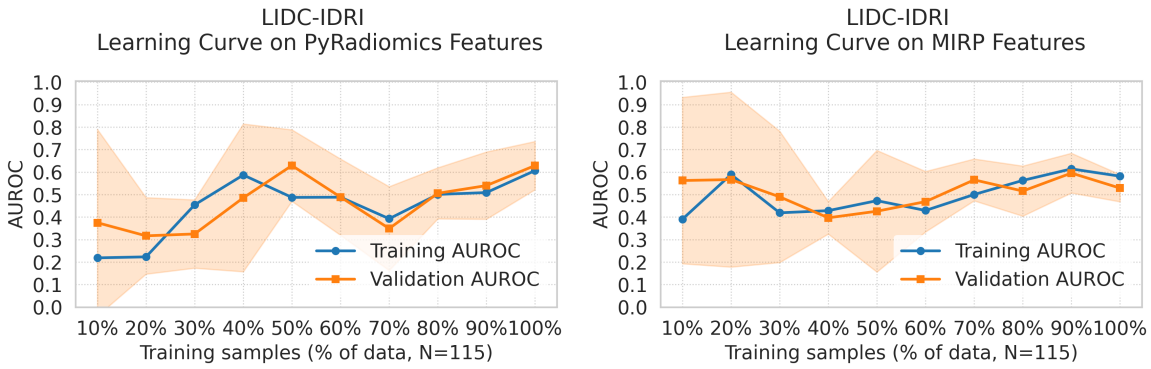
Figure 8.15. Learning curve on selected PyRadiomics Features (a) and selected MIRP Features (b) from RPTK run. I trained a random forest classifier with 5 fold cv on the Lipo data. I generated this figure by using the LearningCurveDisplay function from *scikit-learn* (v 1.5.0) [191]. This figures show the predictive model performance variances related to training size.



(a) Learning curve on selected PyRadiomics Features.

(b) Learning curve on selected MIRP Features.

Figure 8.16. Learning curve on selected PyRadiomics Features (a) and selected MIRP Features (b) from RPTK run. I trained a random forest classifier with 5 fold cv on the Melanoma data. I generated this figure by using the LearningCurveDisplay function from *scikit-learn* (v 1.5.0) [191]. This figures show the predictive model performance variances related to training size.



(a) Learning curve on selected PyRadiomics Features.

(b) Learning curve on selected MIRP Features.

Figure 8.17. Learning curve on selected PyRadiomics Features (a) and selected MIRP Features (b) from RPTK run. I trained a random forest classifier with 5 fold cv on the LIDC-IDRI data. I generated this figure by using the LearningCurveDisplay function from *scikit-learn* (v 1.5.0) [191]. This figures show the predictive model performance variances related to training size.

8.4 Predict Study – Predicting Immunotherapy Treatment Response in Lung Cancer Patients

Table 8.4. Clinical parameters and description of the Predict cohort including the parameter type as well as the percentage of how often this parameter was missing from the treatment start (RFA=Radiofrequenzablation, Radio.=Radiotherapy, Chemo.=Chemotherapy, Immuno.=Immunotherapy).

Parameter	Description	Type	Distribution	Missing Rate
Sex	Sex	categorical	male (62%), female (38%)	0 %
Smoking	Smoking category	categorical	ex (52%), current (40%), never (8%)	0 %
Packyears	How many packs of cigarettes were smoked per year	continuous	Mean: 36.87, Std: 22.66, Median: 40.00	2.74 %
PD-L1	PD-L1 expression [%]	continuous	Mean: 79.84, Std: 13.19, Median: 80.00	0 %
NLR	Neutrophil over Lymphocyte ratio at therapy start	continuous	Mean: 6.28, Std: 8.20, Median: 4.79	0 %
Neut	Neutrophil granulocytes concentration at therapy start [cell/nl]	continuous	Mean: 7.45, Std: 4.44, Median: 6.42	0 %
Ly	Lymphocyte concentration at therapy start [cell/nl]	continuous	Mean: 1.67, Std: 1.99, Median: 1.33	0 %
ECOG	ECOG Performance State	ordinal	0 (42%), 1 (55%), 2 (3%)	0 %
KM-Phase	Contrast Agent Application	categorical	arterial (89%), native (8%), venous (3%)	0 %
T-initial	Initial T-staging of primary tumor	ordinal	T1 (1%), T1b (3%), T1c (5%) T2 (1%), T2a (7%), T2b (7%), T2c (1%) T3 (16%), T4 (51%), TX (7%)	0 %
Primary tumor location	Location of primary tumor	categorical	Upper lobe right (27%), Lower lobe right (23%) Lower lobe left (18%), Upper lobe left (12%) Central right (10%), Lingula (4%), Middle lobe (3%) Central left (3%)	0 %
Max. Diameter Primary tumor	Maximum diameter of pulmonal primary tumor [mm]	continuous	Mean: 53.86, Std: 32.51, Median: 46.00	1.37 %
Other tumor manifestations initially	Other tumor manifestations at the initial stage	categorical	Pulmonary (19%), Breast (11%), Lymph nodes (5%) Breast & Pulmonary (3%), other (62%)	12.33 %
Pleural effusion	Binary pleural effusion status	categorical	no (70%), yes (30%)	0 %
Stage	General disease stage	ordinal	IA (3%), IA1 (1%), IB (3%), IIIA (3%) IVB (45%), IIIB (3%), IVA (38%) VIA (3%), VIB (1%)	0 %
Age at diagnosis	Age at diagnosis	continuous	Mean: 65.82, Std: 10.99, Median: 67.00	0 %
cT	Clinical T-staging	ordinal	T1a (1%), T1b (8%), T1c (4%) T2a (10%), T2b (7%), T3 (11%), T4 (59%)	0 %
cN	Clinical N-staging	ordinal	N0 (16%), N1 (12%), N2 (37%), N3 (34%)	0 %
cM	Clinical M-staging	ordinal	M0 (12%), M1 (4%), M1a (22%), M1b (26%), M1c (36%)	0 %
Weight	Weight [kg]	continuous	Mean: 74.64, Std: 16.06, Median: 72.00	0 %
Size	Size [m]	continuous	Mean: 170.86, Std: 8.27, Median: 170.00	0 %
CRP_wert_pre	C-reactive protein (CRP) measured pre-therapeutic [mg/L]	continuous	Mean: 35.79, Std: 55.51, Median: 15.05	1.37 %
HB_pre	Hemoglobin measured pre-therapeutic [g/dl]	continuous	Mean: 13.15, Std: 1.79, Median: 13.30	0 %
ALBQ_pre	Albumin measured pre-therapeutic [g/L]	continuous	Mean: 38.63, Std: 4.96, Median: 39.50	42.47 %
LYMPHOA_pre	Lymphocytes measured pre-therapeutic [cell/nl]	continuous	Mean: 1.57, Std: 0.57, Median: 1.47	6.85 %

Table 8.5. Performance comparison of radiomics, clinical, and delta features across time-points and modalities. Tha validation AUROC is displayed as the mean of the five fold validation AUROC plus standart deviation, the Test AUROC is the performance of the Ensembled model from the five fold models displayed as the mean and CI 95% range after 1000x bootstrapping as described in Section 4.1.8. Bold are the best performing models based on the mean validation AUROC.

Data	Framework	Extractor	Best Model	Mean Val AUROC	CI 95% Test AUROC
Timepoint 0 - Radiomics	**RPTK**	**MIRP**	**LGBM**	**0.804 (+/- 0.102)**	**0.339 [0.068 – 0.629]**
Timepoint 0 - Radiomics	RPTK	PyRadiomics	TabNet	0.749 (+/- 0.095)	0.440 [0.115 – 0.796]
Timepoint 1 - Radiomics	RPTK	MIRP	LGBM	0.833 (+/- 0.042)	0.513 [0.196 – 0.818]
Timepoint 1 - Radiomics	**RPTK**	**PyRadiomics**	**LGBM**	**0.857 (+/- 0.047)**	**0.696 [0.360 – 0.960]**
Delta - Radiomics	**RPTK**	**MIRP**	**XGBoost**	**0.971 (+/- 0.017)**	**0.750 [0.536 – 0.983]**
Delta - Radiomics	RPTK	PyRadiomics	XGBoost	0.927 (+/- 0.049)	0.729 [0.519 – 0.963]
Clinical	**RPTK**	**Clinical**	**LGBM**	**0.817 (+/- 0.085)**	**0.786 [0.518 – 1.000]**
Delta Radiomics & Clinical	**RPTK**	**MIRP & Clinical**	**Random Forest**	**0.949 (+/- 0.055)**	**0.768 [0.464 – 0.964]**
Delta Radiomics & Clinical	RPTK	PyRadiomics & Clinical	Random Forest	0.942 (+/- 0.047)	0.588 [0.273 – 0.889]

Table 8.6. Performance metrics of best performing AutoRadiomics models for the Predict dataset, selected based on the best mean validation AUROC (with std). Test AUROC, F1, Sensitivity and Specificity are displayed with 95% CIs after 1000x bootstrapping as described in Section 4.1.8. Included is also the selected feature selection method.

Dataset	Best Model	Feature Selection	Val AUROC	Test AUROC	Test F1	Test Sensitivity	Test Specificity
Predict	Random Forest	ANOVA	0.851 (± 0.088)	0.526 [0.204, 0.822]	0.378 [0.000, 0.667]	0.371 [0.000, 0.700]	0.426 [0.000, 0.800]

Table 8.7. Performance metrics of best performing deep learning models for the Predict dataset, selected based on the best mean validation AUROC (with std). Test AUROC, F1, Sensitivity and Specificity are displayed with 95% CIs after 1000x bootstrapping as described in Section 4.1.8.

Dataset	Best Model	Val AUROC	Test AUROC	Youden	Test F1	Test Sensitivity	Test Specificity
Predict	ResNet18	0.799 (± 0.068)	0.563 [0.260, 0.880]	0.833	0.310 [0.000, 0.625]	0.245 [0.000, 0.571]	0.711 [0.333, 1.000]

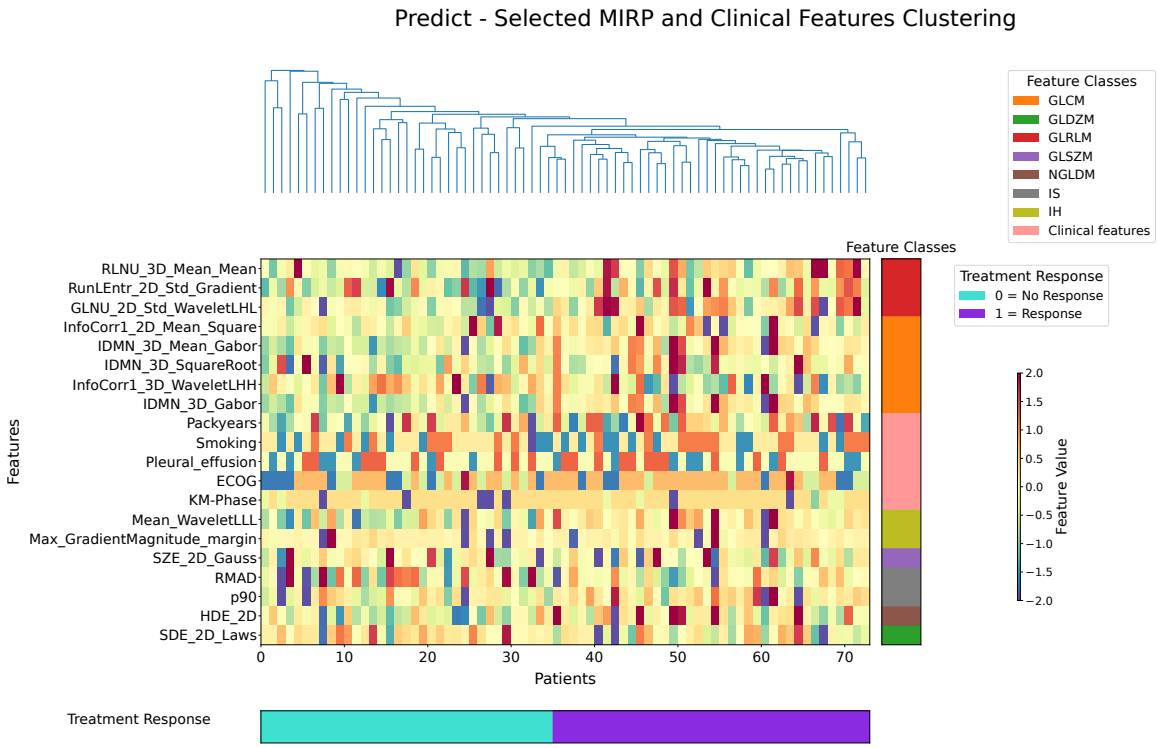


Figure 8.18. Selected delta radiomics features (MIRP) in combination with clinical features. Rows are features ordered by IBSI feature class (color bar at right); columns are patients ordered by treatment response. Cell colors show z-score–normalized feature values. This plot should give an impression on the feature clusters from the selected feature space which show separation of the treatment response label.

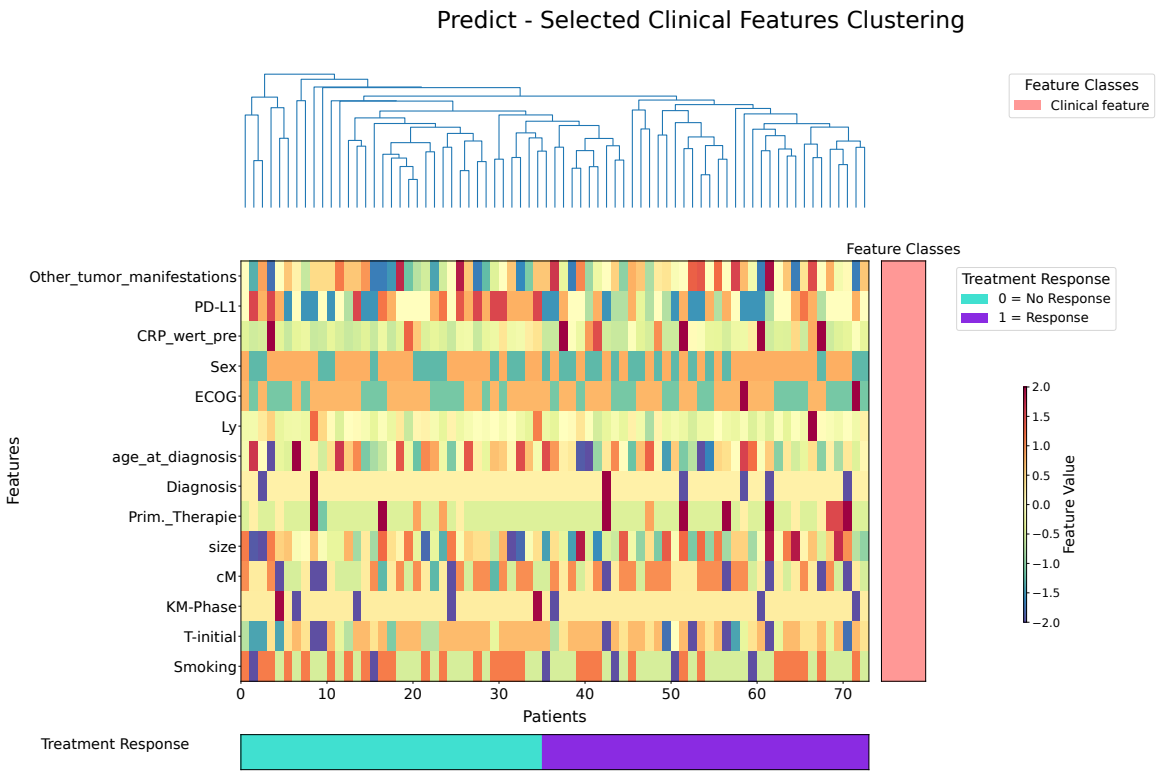


Figure 8.19. Selected clinical features. Columns are patients ordered by treatment response. Cell colors show z-score–normalized feature values. This plot should give an impression on the feature clusters from the selected feature space which show separation of the treatment response label. These features have been selected from all clinical features shown in Table 8.4.

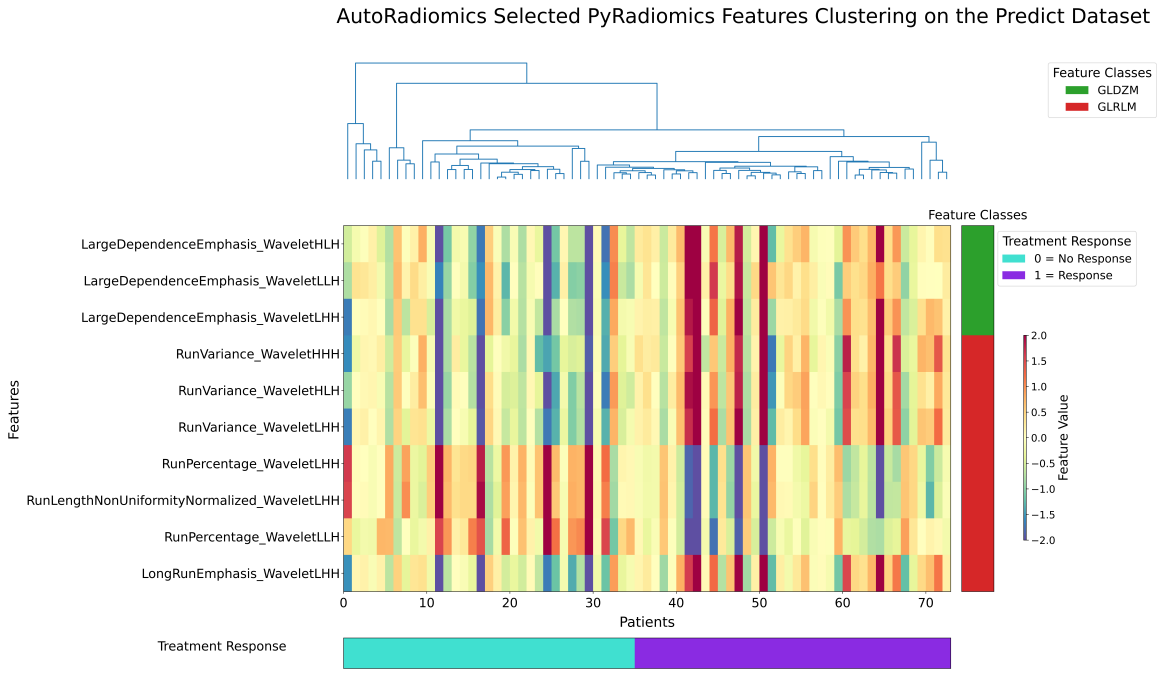
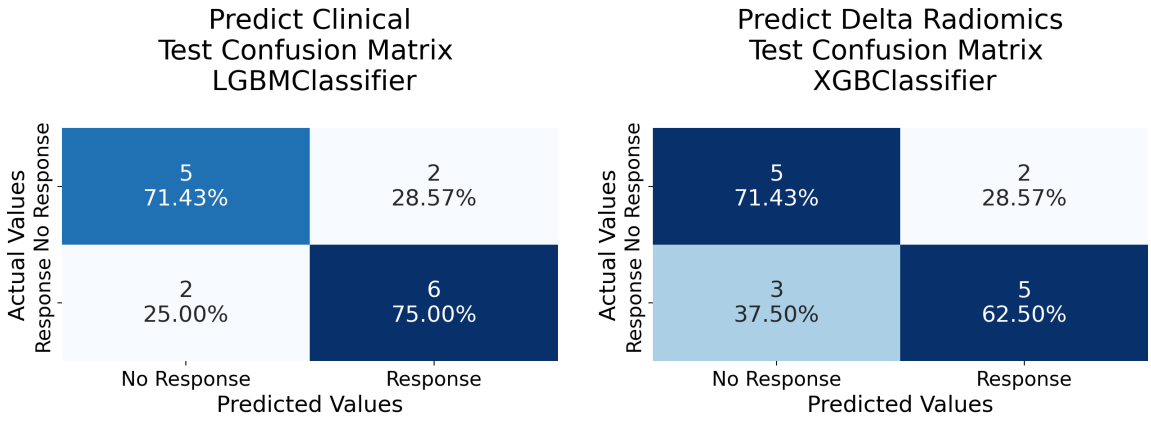


Figure 8.20. AutoRadiomics selected radiomics features. Columns are patients ordered by treatment response. Cell colors show z-score–normalized feature values. This plot should give an impression on the feature clusters from the selected feature space from AutoRadiomics which show separation of the treatment response label. These features were selected from the AutoRadiomics framework based on delta radiomics features.



(a) Clinical data confusion matrix of the best RPTK model after Youden correction.

(b) Delta radiomics confusion matrix of the best RPTK model after Youden correction.

Figure 8.21. Confusion matrix after Youden correction of the best RPTK model performance on the test set of **a.** clinical features from Table 8.4 and **b.** delta radiomics features. See Table 8.5 for details.

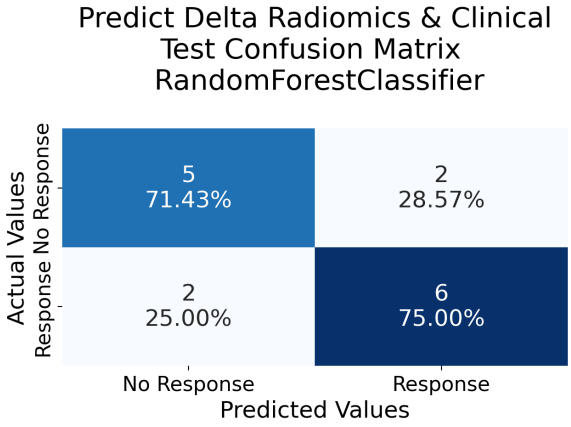
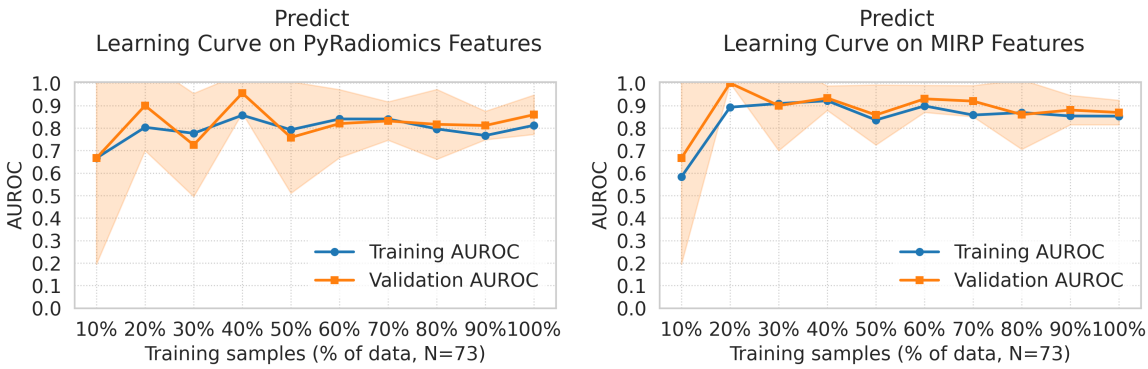


Figure 8.22. Confusion matrix of the best RPTK model ensemble on delta radiomics and clinical data after Youden correction based on selected clinical and delta radiomics features, displayed predictions are based on the test set (see Table 8.5 for details).



(a) Learning curve on selected PyRadiomics Features.

(b) Learning curve on selected MIRP Features.

Figure 8.23. Learning curve on selected PyRadiomics Features (a) and selected MIRP Features (b) from RPTK run. I trained a random forest classifier with 5 fold cv on the Predict data. I generated this figure by using the LearningCurveDisplay function from *scikit-learn* (v 1.5.0) [191].

8.5 LiverCRC Study – Colorectal Cancer Prediction via Liver CT

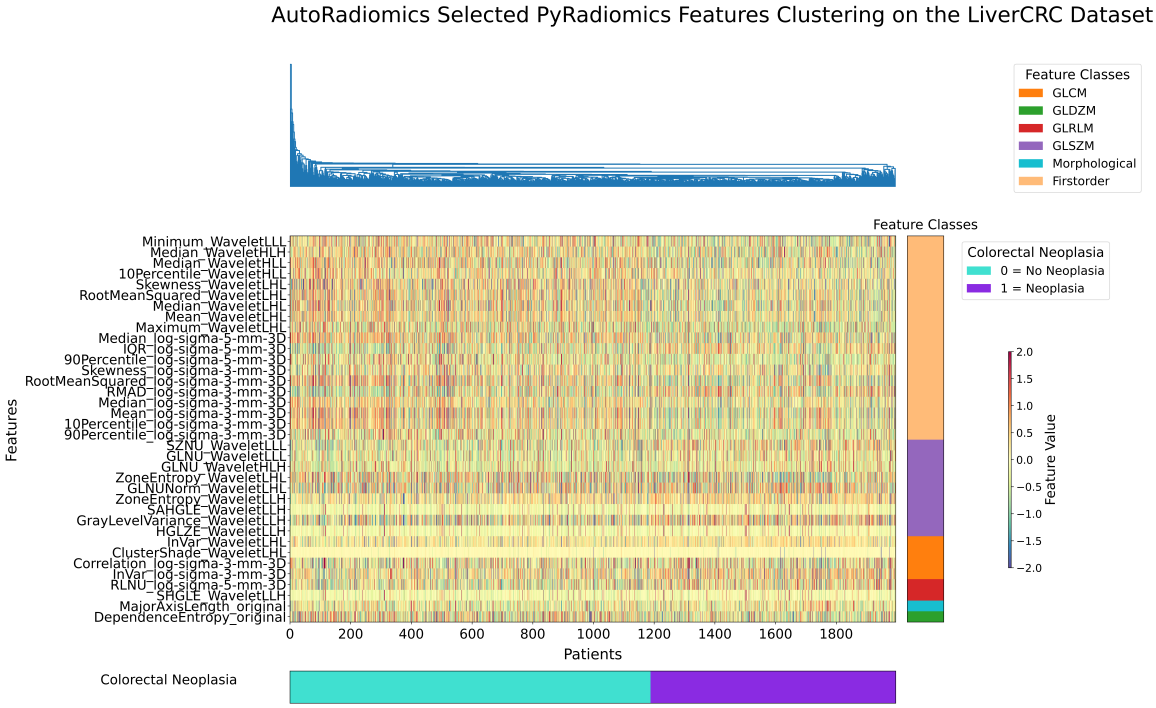
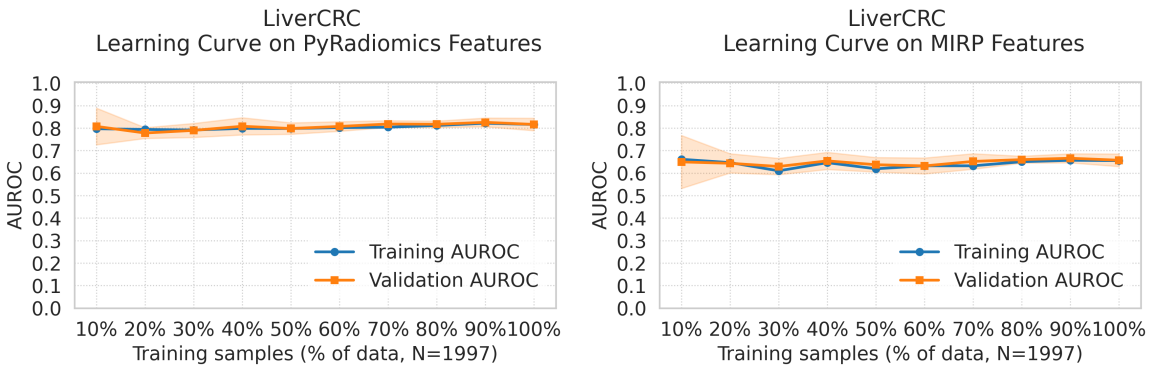


Figure 8.24. AutoRadiomics selected radiomics features. Columns are patients ordered by the colorectal neoplasia label. Cell colors show z-score–normalized feature values. This plot should give an impression on the feature clusters from the selected feature space from AutoRadiomics which show separation of the colorectal neoplasia label. These features were selected from the AutoRadiomics framework based on radiomics features.



(a) Learning curve on selected PyRadiomics Features.

(b) Learning curve on selected MIRP Features.

Figure 8.25. Learning curve on selected PyRadiomics Features (a) and selected MIRP Features (b) from RPTK run. I trained a random forest classifier with 5 fold cv on the LiverCRC data. I generated this figure by using the LearningCurveDisplay function from *scikit-learn* (v 1.5.0) [191].

Table 8.8. Performance metrics of best performing AutoRadiomics models for the LiverCRC dataset, selected based on the best mean validation AUROC (with std). Test AUROC, F1, Sensitivity and Specificity are displayed with 95% CIs after 1000x bootstrapping as described in Section 4.1.8. Included is also the selected feature selection method.

Dataset	Best Model	Feature Selection	Val AUROC	Test AUROC	Test F1	Test Sensitivity	Test Specificity
LiverCRC	Random Forest	ANOVA	0.648 (± 0.029)	0.652 [0.596, 0.706]	0.424 [0.341, 0.495]	0.347 [0.268, 0.421]	0.804 [0.753, 0.851]

Table 8.9. Performance metrics of best performing deep learning models for the LiverCRC dataset, selected based on the best mean validation AUROC (with std). Test AUROC, F1, Sensitivity and Specificity are displayed with 95% CIs after 1000x bootstrapping as described in Section 4.1.8.

Dataset	Best Model	Val AUROC	Test AUROC	Youden	Test F1	Test Sensitivity	Test Specificity
LiverCRC	ResNet18	0.887 (± 0.006)	0.594 [0.533, 0.652]	0.833	0.466 [0.389, 0.538]	0.408 [0.337, 0.482]	0.766 [0.710, 0.817]

8.6 Own Publications

Here I present published papers and papers which are currently in submission or under review.

8.6.1 First Authorships

Bohn J. R.*, Vlachavas, E. I.*, Ückert, F., Nürnberg, S. A Detailed Catalogue of Multi-Omics Methodologies for Identification of Putative Biomarkers and Causal Molecular Networks in Translational Cancer Research. Int. J. Mol. Sci. 2021, 22, 2822. DOI: 10.3390/ijms22062822

Bohn J. R., Heidt, C. M., D. Almeida, S., Kausch, L., Götz, M., Nolden, M., Christopoulos, P., Rheinheimer, S., Peters, A. A., von Stackelberg, O., Kauczor, H. U., Maier-Hein, K. H., Heußel, C. P. & Norajitra, T. RPTK: The Role of Feature Computation on Prediction Performance. In: Woo, J., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops. MICCAI 2023. Lecture Notes in Computer Science, vol 14394. Springer, Cham. (2023) DOI: 10.1007/978-3-031-47425-5_11

Bohn J. R.*, Hinterberger A.*, Trofimova D., Knabe N., Dettling J., Norajitra T., Isensee F., Betge J., O. Schönberg S., Nörenberg D., Grosu S., Loges S., Floca R., Kather J. N., Maier-Hein K.*, and Grawe F.*, (2025) Gut decisions based on the liver: A radiomics approach to boost colorectal cancer screening, Manuscript in submission.

Bohn J. R., Heidt, C. M., D. Almeida, S., Kausch, L., Götz, M., Nolden, M., Christopoulos, P., Rheinheimer, S., Peters, A. A., von Stackelberg, O., Kauczor, H. U., Maier-Hein, K. H., Heußel, C. P. & Norajitra, T. RPTK: Enabling Optimal Radiomics via Comprehensive Feature Space Engineering and Multi-Model Optimization, Manuscript in preperation.

8.6.2 Co-Authorships

Floca, R., **Bohn J. R.**, Haux, C., Wiestler, B. Zollner, F. G. Reinke, A. Weiss, J. Nolden, M. Albert, S. Persigehl, T. Norajitra, T. Baessler, B. Dewey, M. Braren, R. Buchert, M. Fallenberg, E. M. Galldiks, N. Gerken, A. Gotz, M. Hahn, H. K. Haubold, J. Haueise, T. Grosse Hokamp, N. Ingrisch, M. Iuga, A. I. Janoschke, M. Jung, M. Kiefer, L. S. Lohmann, P. Machann, J. Moltz, J. H. Nattenmuller, J. Nonnenmacher, T. Oerther, B. Othman, A. E. Peisen, F. Schick, F. Umutlu, L. Wichtmann, B. D. Zhao, W. Caspers, S. Schlemmer, H. P. Schlett, C. L. Maier-Hein, K. Bamberg, F. Radiomics workflow definition & challenges - German priority program 2177 consensus statement on clinically applied radiomics. Insights Imaging 15, 124 (2024). DOI: 10.1186/s13244-024-01704-w

Peretzke, R., Maier-Hein, K., **Bohn J.**, Kirchhoff Y., Roy, S., Oberli-Palma, S, Becker D., Lenga P., Neher P. atTRACTive: Semi-automatic White Matter Tract Segmen-

*equal contribution

tation Using Active Learning. In: Greenspan, H., et al. Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. MICCAI 2023. Lecture Notes in Computer Science, vol 14227. Springer, Cham. (2023) DOI: 10.1007/978-3-031-43993-3_23

Heidt, C.M., **Bohn J. R.**, Stollmayer, R., von Stackelberg, O., Rheinheimer, S., Bozorgmehr, F. , Senghas, K., Schlamp, K., Weinheimer,O. , Giesel, F. L., Kauczor H. U., Heußel C. P. & Heußel G. Delta-radiomics features of ADC maps as early predictors of treatment response in lung cancer. Insights Imaging 15, 218 (2024). DOI: 10.1186/s13244-024-01787-5

Peretzke, R., Neher, P.F., Brandt, G.A., Fritze, S., Volkmer, S., Daub, J., Northoff, G., **Bohn J.**, Kirchhoff, Y., Roy, S., Maier-Hein, K.H., Meyer-Lindenberg, A., Hirjak, D., Deciphering white matter microstructural alterations in catatonia according to ICD-11: replication and machine learning analysis. Mol Psychiatry 30, 2095–2107 (2025). DOI: 10.1038/s41380-024-02821-0

Holzschuh J. C. , **Bohn J. R.**, Norajitra T., Maier-Hein K., Schlemmer H. P., Johnston O., Bachanek S., Uhlig J., Uhlig A. Radiomics for the Prediction of Postoperative Chronic Kidney Disease in Renal Tumor Patients undergoing Surgical Resection. (2025) Manuscript in preperation.

Bibliography

[1] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler. Radiomics in medical imaging-"how-to" guide and critical reflection. *Insights into Imaging*, 11(1), 2020. URL: <GotoISI>://WOS:000563915600001, doi:ARTN9110.1186/s13244-020-00887-2.

[2] E. Stamoulou, C. Spanakis, G. C. Manikis, G. Karanasiou, G. Grigoriadis, T. Foukakis, M. Tsiknakis, D. I. Fotiadis, and K. Marias. Harmonization strategies in multicenter mri-based radiomics. *J Imaging*, 8(11), 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/36354876, doi:10.3390/jimaging8110303.

[3] H. Horng, A. Singh, B. Yousefi, E. A. Cohen, B. Haghighi, S. Katz, P. B. Noel, R. T. Shinohara, and D. Kontos. Generalized combat harmonization methods for radiomic features with multi-modal distributions and multiple batch effects. *Sci Rep*, 12(1):4493, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35296726, doi:10.1038/s41598-022-08412-9.

[4] B. Tafuri, A. Lombardi, S. Nigro, D. Urso, A. Monaco, E. Pantaleo, D. Diacono, R. De Blasi, R. Bellotti, S. Tangaro, and G. Logroscino. The impact of harmonization on radiomic features in parkinson's disease and healthy controls: A multicenter study. *Front Neurosci*, 16:1012287, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/36300169, doi:10.3389/fnins.2022.1012287.

[5] N. Horvat, N. Papanikolaou, and D. M. Koh. Radiomics beyond the hype: A critical evaluation toward oncologic clinical use. *Radiol Artif Intell*, 6(4):e230437, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38717290, doi:10.1148/ryai.230437.

[6] T. Akinci D'Antonoli, R. Cuocolo, B. Baessler, and D. Pinto Dos Santos. Towards reproducible radiomics research: introduction of a database for radiomics studies. *Eur Radiol*, 34(1):436–443, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/37572188, doi:10.1007/s00330-023-10095-3.

[7] Martijn P.A. Starmans, Milea J.M. Timbergen, Melissa Vos, Guillaume A. Padmos, Dirk J. Grünhagen, Cornelis Verhoef, Stefan Sleijfer, Geert J.L.H. van Leenders, Florian E. Buisman, Francois E.J.A. Willemssen, Bas Groot Koerkamp, Lindsay Angus, Astrid A.M. van der Veldt, Ana Rajicic, Arlette E. Odink, Michel Renckens, Michail Doukas, Rob A. de Man, Jan N.M. IJzermans, Razvan L. Miclea, Peter B. Vermeulen, Maarten G. Thomeer, Jacob J. Visser, Wiro J. Niessen, and Stefan Klein. The worc database: Mri and ct scans, segmentations, and clinical labels for 930 patients from six radiomics studies. *medRxiv*, page 2021.08.19.21262238, 2021. URL: `https://www.medrxiv.org/content/medrxiv/early/2021/08/25/2021.08.19.21262238.full.pdf`, `doi:10.1101/2021.08.19.21262238`.

[8] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*, 18(2):203–211, 2021. URL: `https://www.ncbi.nlm.nih.gov/pubmed/33288961`, `doi:10.1038/s41592-020-01008-z`.

[9] A. Demircioglu. Reproducibility and interpretability in radiomics: a critical assessment. *Diagn Interv Radiol*, 2024. URL: `https://www.ncbi.nlm.nih.gov/pubmed/39463040`, `doi:10.4274/dir.2024.242719`.

[10] P. Whybra, A. Zwanenburg, V. Andrearczyk, R. Schaer, A. P. Apte, A. Ayotte, B. Baheti, S. Bakas, A. Bettinelli, R. Boellaard, L. Boldrini, I. Buvat, G. J. R. Cook, F. Dietsche, N. Dinapoli, H. S. Gabrys, V. Goh, M. Guckenberger, M. Hatt, M. Hosseinzadeh, A. Iyer, J. Lenkowicz, M. A. L. Loutfi, S. Lock, F. Marturano, O. Morin, C. Nioche, F. Orlhac, S. Pati, A. Rahmim, S. M. Rezaeijo, C. G. Rookyard, M. R. Salmanpour, A. Schindele, I. Shiri, E. Spezi, S. Tanadini-Lang, F. Tixier, T. Upadhaya, V. Valentini, J. J. M. van Griethuysen, F. Yousefirizi, H. Zaidi, H. Muller, M. Vallieres, and A. Depeursinge. The image biomarker standardization initiative: Standardized convolutional filters for reproducible radiomics and enhanced clinical insights. *Radiology*, 310(2):e231319, 2024. URL: `https://www.ncbi.nlm.nih.gov/pubmed/38319168`, `doi:10.1148/radiol.231319`.

[11] R. Floca, J. Bohn, C. Haux, B. Wiestler, F. G. Zollner, A. Reinke, J. Weiss, M. Nolden, S. Albert, T. Persigehl, T. Norajitra, B. Baessler, M. Dewey, R. Braren, M. Buchert, E. M. Fallenberg, N. Galldiks, A. Gerken, M. Gotz, H. K. Hahn, J. Haubold, T. Haueise, N. Grosse Hokamp, M. Ingrisch, A. I. Iuga, M. Janoschke, M. Jung, L. S. Kiefer, P. Lohmann, J. Machann, J. H.

Moltz, J. Nattenmuller, T. Nonnenmacher, B. Oerther, A. E. Othman, F. Peisen, F. Schick, L. Umutlu, B. D. Wichtmann, W. Zhao, S. Caspers, H. P. Schlemmer, C. L. Schlett, K. Maier-Hein, and F. Bamberg. Radiomics workflow definition & challenges - german priority program 2177 consensus statement on clinically applied radiomics. *Insights Imaging*, 15(1):124, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38825600, doi:10.1186/s13244-024-01704-w.

[12] B. Kocak, B. Baessler, S. Bakas, R. Cuocolo, A. Fedorov, L. Maier-Hein, N. Mercaldo, H. Muller, F. Orlhac, D. Pinto Dos Santos, A. Stanzione, L. Ugga, and A. Zwanenburg. Checklist for evaluation of radiomics research (clear): a step-by-step reporting guideline for authors and reviewers endorsed by esr and eusomii. *Insights Imaging*, 14(1):75, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/37142815, doi:10.1186/s13244-023-01415-8.

[13] B. Kocak, L. L. Chepelev, L. C. Chu, R. Cuocolo, B. S. Kelly, P. Seebock, Y. L. Thian, R. W. van Hamersvelt, A. Wang, S. Williams, J. Witowski, Z. Zhang, and D. Pinto Dos Santos. Assessment of radiomics research (arise): a brief guide for authors, reviewers, and readers from the scientific editorial board of european radiology. *Eur Radiol*, 33(11):7556–7560, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/37358612, doi:10.1007/s00330-023-09768-w.

[14] P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. de Jong, J. van Timmeren, S. Sanduleanu, R. T. H. M. Larue, A. J. G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F. M. Mottaghy, J. E. Wildberger, and S. Walsh. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*, 14(12):749–762, 2017. URL: https://www.nature.com/articles/nrclinonc.2017.141.pdf, doi:10.1038/nrclinonc.2017.141.

[15] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J. C. Fillion-Robin, S. Pieper, and Hjwl Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*, 77(21):e104–e107, 2017. URL: https://www.ncbi.nlm.nih.gov/pubmed/29092951, doi:10.1158/0008-5472.CAN-17-0339.

[16] Alex Zwanenburg and Steffen Löck. Mirp: A python package for standardised radiomics. *Journal of Open Source Software*, 9(99), 2024. doi:10.21105/joss.06413.

[17] N. Papanikolaou, C. Matos, and D. M. Koh. How to develop a meaningful radiomic signature for clinical use in oncologic patients. *Cancer Imaging*, 20(1):33, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32357923, doi:10.1186/s40644-020-00311-4.

[18] P. Woznicki, F. Laqua, T. Bley, and B. Baessler. Autoradiomics: A framework for reproducible radiomics research. *Front Radiol*, 2:919133, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/37492662, doi:10.3389/fradi.2022.919133.

[19] Martijn P. A. Starmans, Sebastian R. van der Voort, Thomas Phil, Milea J. M. Timbergen, Melissa Vos, Guillaume A. Padmos, Wouter Kessels, David Hanff, Dirk J. Grunhagen, Cornelis Verhoef, Stefan Sleijfer, Martin J. van den Bent, Marion Smits, Roy S. Dwarkasing, Christopher J. Els, Federico Fiduzi, Geert J. L. H. van Leenders, Anela Blazevic, Johannes Hofland, Tessa Brabander, Renza A. H. van Gils, Gaston J. H. Franssen, Richard A. Feelders, Wouter W. de Herder, Florian E. Buism, Francois E. J. A. Willemssen, Bas Groot Koerkamp, Lindsay Angus, Astrid A. M. van der Veldt, Ana Rajicic, Arlette E. Odink, Mitchell Deen, Jose M. Castillo T., Jifke Veenland, Ivo Schoots, Michel Renckens, Michail Doukas, Rob A. de Man, Jan N. M. IJzermans, Razvan L. Miclea, Peter B. Vermeulen, Esther E. Bron, Maarten G. Thomeer, Jacob J. Visser, Wiro J. Niessen, and Stefan Klein. An automated machine learning framework to optimize radiomics model construction validated on twelve clinical applications. *arXiv*, 2025. doi:10.48550/arXiv.2108.08618.

[20] M. F. McNitt-Gray, 3rd Armato, S. G., C. R. Meyer, A. P. Reeves, G. McLennan, R. C. Pais, J. Freymann, M. S. Brown, R. M. Engelmann, P. H. Bland, G. E. Laderach, C. Piker, J. Guo, Z. Towfic, D. P. Qing, D. F. Yankelevitz, D. R. Aberle, E. J. van Beek, H. MacMahon, E. A. Kazerooni, B. Y. Croft, and L. P. Clarke. The lung image database consortium (lidc) data collection process for nodule detection and annotation. *Acad Radiol*, 14(12):1464–74, 2007. URL: https://www.ncbi.nlm.nih.gov/pubmed/18035276, doi:10.1016/j.acra.2007.07.021.

[21] A. Zwanenburg, S. Leger, L. Agolli, K. Pilz, E. G. C. Troost, C. Richter, and S. Lock. Assessing robustness of radiomic features by image perturbation. *Sci Rep*, 9(1):614, 2019. URL: https://www.ncbi.nlm.nih.gov/pubmed/30679599, doi:10.1038/s41598-018-36938-4.

[22] Hinterberger A., Bohn J., Trofimova D., Knabe N., Dettling J., Norajitra T., Isensee F., Betge J., Schönberg S. O., Nörenberg D., Grosu S., Loges S., Floca R., Kather J. N., Maier-Hein K. H., and Grawe F. Gut decisions based on the liver: A radiomics approach to boost colorectal cancer screening. *arXiv*, page 41, 2025. doi:10.48550/arXiv.2510.23687.

[23] Cardoso M. J., Li W., Brown R., Ma N., Kerfoot E., Wang Y., Murrey B., Myronenko A., Zhao C., Yang D., Nath V., He Y., Xu Z., Hatamizadeh A., Zhu W., Liu Y., Zheng M., Tang Y., Yang I., Zephyr M., Hashemian B., Alle S., Darestani M. Z., Budd C., Modat M., Vercauteren T., Wang G., Li Y., Hu Y., Fu Y., Gorman B., Johnson H., Genereaux B., Erdal B. S., Gupta V., Diaz-Pinto A., Dourson A., Maier-Hein L., Jaeger P. F., Baumgartner M., Kalpathy-Cramer J., Flores M., Kirby J., Cooper L. A., Roth H. R., Xu D., Bericat D., Floca R., Zhou S. K., Shuaib H., Farahani K., Maier-Hein K. H., Aylward S., Dogra P., Ourselin S., and Feng A. Monai: An open-source framework for deep learning in healthcare. *arXiv*, 2022. doi:10.48550/arXiv.2211.02701.

[24] J. R. Bohn, C. M. Heidt, S. D. Almeida, L. Kausch, M. Götz, M. Nolden, P. Christopoulos, S. Rheinheimer, A. A. Peters, O. von Stackelberg, H. U. Kauczor, K. H. Maier-Hein, C. P. Heußel, and T. Norajitra. Rptk: The role of feature computation on prediction performance., 2023. doi:10.1007/978-3-031-47425-5_11.

[25] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken, M. Bilello, P. Bilic, P. F. Christ, R. K. G. Do, M. J. Gollub, S. H. Heckers, H. Huisman, W. R. Jarnagin, M. K. McHugo, S. Napel, J. S. G. Pernicka, K. Rhode, C. Tobon-Gomez, E. Vorontsov, J. A. Meakin, S. Ourselin, M. Wiesenfarth, P. Arbelaez, B. Bae, S. Chen, L. Daza, J. Feng, B. He, F. Isensee, Y. Ji, F. Jia, I. Kim, K. Maier-Hein, D. Merhof, A. Pai, B. Park, M. Perslev, R. Rezaiifar, O. Rippel, I. Sarasua, W. Shen, J. Son, C. Wachinger, L. Wang, Y. Wang, Y. Xia, D. Xu, Z. Xu, Y. Zheng, A. L. Simpson, L. Maier-Hein, and M. J. Cardoso. The medical segmentation decathlon. *Nat Commun*, 13(1):4128, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35840566, doi:10.1038/s41467-022-30695-9.

[26] Isensee F., Jäger P. F., Kohl S. A. A., Petersen J., and Maier-Hein K. H. pretrained models for 3d semantic image segmentation with nnu-net

(2.1), 2021. URL: https://doi.org/10.5281/zenodo.4485926, doi:10.5281/zenodo.3734294.

[27] C. Ulrich, F. Isensee, T. Wald, M. Zenk, M. Baumgartner, and K. H. Maier-Hein. Multitalent: A multi-dataset approach to medical image segmentation, 2023. doi:10.1007/978-3-031-43898-1_62.

[28] M. Kachelrie, W. Schlegel, C. P. Karger, and O. Jaekel. *Computertomographie*, pages 153–203. Springer Berlin Heidelberg, 2018. doi:10.1007/978-3-662-54801-1_8.

[29] X. Ou, X. Chen, X. Xu, L. Xie, X. Chen, Z. Hong, H. Bai, X. Liu, Q. Chen, L. Li, and H. Yang. Recent development in x-ray imaging technology: Future and challenges. *Research (Wash D C)*, 2021:9892152, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/35028585, doi:10.34133/2021/9892152.

[30] M. Larobina and L. Murino. Medical image file formats. *J Digit Imaging*, 27(2):200–6, 2014. URL: https://www.ncbi.nlm.nih.gov/pubmed/24338090, doi:10.1007/s10278-013-9657-9.

[31] R. Tuttle and 3rd Kane, J. M. Biopsy techniques for soft tissue and bowel sarcomas. *J Surg Oncol*, 111(5):504–12, 2015. URL: https://www.ncbi.nlm.nih.gov/pubmed/25663366, doi:10.1002/jso.23870.

[32] R. L. van Ineveld, E. J. van Vliet, E. J. Wehrens, M. Alieva, and A. C. Rios. 3d imaging for driving cancer discovery. *EMBO J*, 41(10):e109675, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35403737, doi:10.15252/embj.2021109675.

[33] F. Wang, W. Chen, F. Chen, J. Lu, Y. Xu, M. Fang, and H. Jiang. Risk stratification and overall survival prediction in extensive stage small cell lung cancer after chemotherapy with immunotherapy based on ct radiomics. *Sci Rep*, 14(1):22659, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/39349536, doi:10.1038/s41598-024-73331-w.

[34] H. K. Genant, K. Engelke, and S. Prevrhal. Advanced ct bone imaging in osteoporosis. *Rheumatology (Oxford)*, 47 Suppl 4(Suppl 4):iv9–16, 2008. URL: https://www.ncbi.nlm.nih.gov/pubmed/18556648, doi:10.1093/rheumatology/ken180.

[35] L. Zhang, L. Li, G. Feng, T. Fan, H. Jiang, and Z. Wang. Advances in ct techniques in vascular calcification. *Front Cardiovasc Med*, 8:716822,

2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/34660718, doi:10.3389/fcvm.2021.716822.

[36] W. C. Röntgen. Ueber eine neue art von strahlen. *Annalen der Physik*, 300(1):1–11, 2006. doi:10.1002/andp.18983000102.

[37] J. T. Bushberg, J. A. Seibert, E. M. Leidholdt, and J. M. Boone. *The Essential Physics of Medical Imaging*. Lippincott Williams & Wilkins, 3rd edition, 2011 ISBN: 978-1608312259. URL: https://pubmed.ncbi.nlm.nih.gov/28524933/.

[38] G. N. Hounsfield. Computed medical imaging. nobel lecture, decemberr 8, 1979. *J Comput Assist Tomogr*, 4(5):665–74, 1980. URL: https://www.ncbi.nlm.nih.gov/pubmed/6997341, doi:10.1097/00004728-198010000-00017.

[39] C. Jiang, D. Jin, M. Ni, Y. Zhang, and H. Yuan. Influence of image reconstruction kernel on computed tomography-based finite element analysis in the clinical opportunistic screening of osteoporosis-a preliminary result. *Front Endocrinol (Lausanne)*, 14:1076990, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/36936156, doi:10.3389/fendo.2023.1076990.

[40] I. Vergalasova, M. McKenna, N. J. Yue, and M. Reyhan. Impact of computed tomography (ct) reconstruction kernels on radiotherapy dose calculation. *J Appl Clin Med Phys*, 21(9):178–186, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32889789, doi:10.1002/acm2.12994.

[41] D. K. Jeong, S. S. Lee, J. E. Kim, K. H. Huh, W. J. Yi, M. S. Heo, and S. C. Choi. Effects of energy level, reconstruction kernel, and tube rotation time on hounsfield units of hydroxyapatite in virtual monochromatic images obtained with dual-energy ct. *Imaging Sci Dent*, 49(4):273–279, 2019. URL: https://www.ncbi.nlm.nih.gov/pubmed/31915612, doi:10.5624/isd.2019.49.4.273.

[42] F. Ammon, M. Moshage, S. Smolka, M. Goeller, D. O. Bittner, S. Achenbach, and M. Marwan. Influence of reconstruction kernels on the accuracy of ct-derived fractional flow reserve. *Eur Radiol*, 32(4):2604–2610, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/34735608, doi:10.1007/s00330-021-08348-0.

[43] S. Denzler, D. Vuong, M. Bogowicz, M. Pavic, T. Frauenfelder, S. Thierstein, E. I. Eboulet, B. Maurer, J. Schniering, H. S. Gabrys, I. Schmitt-Opitz, M. Pless, R. Foerster, M. Guckenberger, and S. Tanadini-Lang. Impact of ct convolution

kernel on robustness of radiomic features for different lung diseases and tissue types. *Br J Radiol*, 94(1120):20200947, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/33544646, doi:10.1259/bjr.20200947.

[44] J. Choe, S. M. Lee, K. H. Do, G. Lee, J. G. Lee, S. M. Lee, and J. B. Seo. Deep learning-based image conversion of ct reconstruction kernels improves radiomics reproducibility for pulmonary nodules or masses. *Radiology*, 292(2):365–373, 2019. URL: https://www.ncbi.nlm.nih.gov/pubmed/31210613, doi:10.1148/radiol.2019181960.

[45] R. Da-Ano, I. Masson, F. Lucia, M. Dore, P. Robin, J. Alfieri, C. Rousseau, A. Mervoyer, C. Reinhold, J. Castelli, R. De Crevoisier, J. F. Ramee, O. Pradier, U. Schick, D. Visvikis, and M. Hatt. Performance comparison of modified combat for harmonization of radiomic features for multicenter studies. *Sci Rep*, 10(1):10248, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32581221, doi:10.1038/s41598-020-66110-w.

[46] L. Gallardo-Estrella, D. A. Lynch, M. Prokop, D. Stinson, J. Zach, P. F. Judy, B. van Ginneken, and E. M. van Rikxoort. Normalizing computed tomography data reconstructed with different filter kernels: effect on emphysema quantification. *Eur Radiol*, 26(2):478–86, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/26002132, doi:10.1007/s00330-015-3824-y.

[47] B. E. Matheson and S. K. Boyd. Establishing the effect of computed tomography reconstruction kernels on the measure of bone mineral density in opportunistic osteoporosis screening. *Sci Rep*, 15(1):5449, 2025. URL: https://www.ncbi.nlm.nih.gov/pubmed/39953113, doi:10.1038/s41598-025-88551-x.

[48] N. Ahmed, D. Kassavin, Y. H. Kuo, and R. Biswal. Sensitivity and specificity of ct scan and angiogram for ongoing internal bleeding following torso trauma. *Emerg Med J*, 30(3):e14, 2013. URL: https://www.ncbi.nlm.nih.gov/pubmed/22505301, doi:10.1136/emermed-2011-200376.

[49] K. A. Miles. Functional ct imaging in oncology. *Eur Radiol*, 13 Suppl 5:M134–8, 2003. URL: https://www.ncbi.nlm.nih.gov/pubmed/14989624, doi:10.1007/s00330-003-2108-0.

[50] K. A. Miles, H. Young, S. L. Chica, and P. D. Esser. Quantitative contrast-enhanced computed tomography: is there a need for system calibration? *Eur Radiol*, 17(4):919–26, 2007. URL: https://www.ncbi.nlm.nih.gov/pubmed/17008987, doi:10.1007/s00330-006-0424-x.

[51] S. Bisdas, L. Medov, M. Baghi, G. N. Konstantinou, J. Wagenblast, C. H. Thng, T. J. Vogl, and T. S. Koh. A comparison of tumour perfusion assessed by deconvolution-based analysis of dynamic contrast-enhanced ct and mr imaging in patients with squamous cell carcinoma of the upper aerodigestive tract. *Eur Radiol*, 18(4):843–50, 2008. URL: https://www.ncbi.nlm.nih.gov/pubmed/18175123, doi:10.1007/s00330-007-0827-3.

[52] S. Joyce, O. J. O'Connor, M. M. Maher, and M. F. McEntee. Strategies for dose reduction with specific clinical indications during computed tomography. *Radiography (Lond)*, 26 Suppl 2:S62–S68, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32682731, doi:10.1016/j.radi.2020.06.012.

[53] P. Bornert and D. G. Norris. A half-century of innovation in technology-preparing mri for the 21st century. *Br J Radiol*, 93(1111):20200113, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32496816, doi:10.1259/bjr.20200113.

[54] H. Lu, E. Ayers, P. Patel, and T. K. Mattoo. Body water percentage from childhood to old age. *Kidney Res Clin Pract*, 42(3):340–348, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/37313612, doi:10.23876/j.krcp.22.062.

[55] M. Gaeta, K. Galletta, M. Cavallaro, E. Mormina, M. T. Cannizzaro, L. R. M. Lanzafame, T. D'Angelo, A. Blandino, S. L. Vinci, and F. Granata. T1 relaxation: Chemo-physical fundamentals of magnetic resonance imaging and clinical applications. *Insights Imaging*, 15(1):200, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/39120775, doi:10.1186/s13244-024-01744-2.

[56] A. Heuvelink, P. Saini, O. Tasar, and S. Nauts. Improving pediatric patients' magnetic resonance imaging experience with an in-bore solution: Design and usability study. *JMIR Serious Games*, 13:e55720, 2025. URL: https://www.ncbi.nlm.nih.gov/pubmed/39946688, doi:10.2196/55720.

[57] M. Weiger and K. P. Pruessmann. Short-t(2) mri: Principles and recent advances. *Prog Nucl Magn Reson Spectrosc*, 114-115:237–270, 2019. URL: https://www.ncbi.nlm.nih.gov/pubmed/31779882, doi:10.1016/j.pnmrs.2019.07.001.

[58] R. T. Constable and D. D. Spencer. Repetition time in echo planar functional mri. *Magn Reson Med*, 46(4):748–55, 2001. URL: https://www.ncbi.nlm.nih.gov/pubmed/11590651, doi:10.1002/mrm.1253.

[59] K. Suzuki. Overview of deep learning in medical imaging. *Radiol Phys Technol*, 10(3):257–273, 2017. URL: https://www.ncbi.nlm.nih.gov/pubmed/28689314, doi:10.1007/s12194-017-0406-5.

[60] A. Maier, C. Syben, T. Lasser, and C. Riess. A gentle introduction to deep learning in medical image processing. *Z Med Phys*, 29(2):86–101, 2019. URL: https://www.ncbi.nlm.nih.gov/pubmed/30686613, doi:10.1016/j.zemedi.2018.12.003.

[61] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, Jawm van der Laak, B. van Ginneken, and C. I. Sanchez. A survey on deep learning in medical image analysis. *Med Image Anal*, 42:60–88, 2017. URL: https://www.ncbi.nlm.nih.gov/pubmed/28778026, doi:10.1016/j.media.2017.07.005.

[62] Md Eshmam Rayed, S. M. Sajibul Islam, Sadia Islam Niha, Jamin Rahman Jim, Md Mohsin Kabir, and M. F. Mridha. Deep learning for medical image segmentation: State-of-the-art advancements and challenges. *Informatics in Medicine Unlocked*, 47, 2024. doi:10.1016/j.imu.2024.101504.

[63] B. Sistaninejhad, H. Rasi, and P. Nayeri. A review paper about deep learning for medical image analysis. *Comput Math Methods Med*, 2023:7091301, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/37284172, doi:10.1155/2023/7091301.

[64] M. Li, Y. Jiang, Y. Zhang, and H. Zhu. Medical image analysis using deep learning algorithms. *Front Public Health*, 11:1273253, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/38026291, doi:10.3389/fpubh.2023.1273253.

[65] X. Liu, K. Gao, B. Liu, C. Pan, K. Liang, L. Yan, J. Ma, F. He, S. Zhang, S. Pan, and Y. Yu. Advances in deep learning-based medical image analysis. *Health Data Sci*, 2021:8786793, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/38487506, doi:10.34133/2021/8786793.

[66] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang. Segment anything in medical images. *Nat Commun*, 15(1):654, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38253604, doi:10.1038/s41467-024-44824-z.

[67] I. Wolf, M. Vetter, I. Wegner, T. Bottger, M. Nolden, M. Schobinger, M. Hastenteufel, T. Kunert, and H. P. Meinzer. The medical imaging interaction toolkit. *Med Image Anal*, 9(6):594–604, 2005. URL: https://www.ncbi.nlm.nih.gov/pubmed/15896995, doi:10.1016/j.media.2005.04.005.

[68] Paul Jaccard. The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50, 2006. doi:10.1111/j.1469-8137.1912.tb05611.x.

[69] A. A. Taha and A. Hanbury. An efficient algorithm for calculating the exact hausdorff distance. *IEEE Trans Pattern Anal Mach Intell*, 37(11):2153–63, 2015. URL: https://www.ncbi.nlm.nih.gov/pubmed/26440258, doi:10.1109/TPAMI.2015.2408351.

[70] C. Haarburger, G. Muller-Franzes, L. Weninger, C. Kuhl, D. Truhn, and D. Merhof. Radiomics feature reproducibility under inter-rater variability in segmentations of ct images. *Sci Rep*, 10(1):12688, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32728098, doi:10.1038/s41598-020-69534-6.

[71] H. Kaur, R. Sharma, and J. Kaur. Comparison of deep transfer learning models for classification of cervical cancer from pap smear images. *Sci Rep*, 15(1):3945, 2025. URL: https://www.ncbi.nlm.nih.gov/pubmed/39890842, doi:10.1038/s41598-024-74531-0.

[72] T. Zhou, X. Ye, H. Lu, X. Zheng, S. Qiu, and Y. Liu. Dense convolutional network and its application in medical image analysis. *Biomed Res Int*, 2022:2384830, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35509707, doi:10.1155/2022/2384830.

[73] W. Xu, Y. L. Fu, and D. Zhu. Resnet and its application to medical image processing: Research progress and challenges. *Comput Methods Programs Biomed*, 240:107660, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/37320940, doi:10.1016/j.cmpb.2023.107660.

[74] Y. Yang, L. Zhang, M. Du, J. Bo, H. Liu, L. Ren, X. Li, and M. J. Deen. A comparative analysis of eleven neural networks architectures for small datasets of lung images of covid-19 patients toward improved clinical decisions. *Comput Biol Med*, 139:104887, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/34688974, doi:10.1016/j.compbiomed.2021.104887.

[75] Y. L. Thian, D. W. Ng, Jtpd Hallinan, P. Jagmohan, S. Y. Sia, J. S. A. Mohamed, S. T. Quek, and M. Feng. Effect of training data volume on performance of convolutional neural network pneumothorax classifiers. *J Digit Imaging*, 35(4):881–892, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35239091, doi:10.1007/s10278-022-00594-y.

[76] N. Hasan, Y. Bao, A. Shawon, and Y. Huang. Densenet convolutional neural networks application for predicting covid-19 using ct image. *SN Comput Sci*, 2(5):389, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/34337432, doi:10.1007/s42979-021-00782-7.

[77] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2019. doi:10.1007/s11263-019-01228-7.

[78] Tristan Gomez and Harold Mouchère. Computing and evaluating saliency maps for image classification: a tutorial. *Journal of Electronic Imaging*, 32(02), 2023. doi:10.1117/1.Jei.32.2.020801.

[79] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Comput Biol Med*, 140:105111, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/34891095, doi:10.1016/j.compbiomed.2021.105111.

[80] M. Ennab and H. McHeick. Enhancing interpretability and accuracy of ai models in healthcare: a comprehensive review on challenges and future directions. *Front Robot AI*, 11:1444763, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/39677978, doi:10.3389/frobt.2024.1444763.

[81] A. S. Panayides, A. Amini, N. D. Filipovic, A. Sharma, S. A. Tsaftaris, A. Young, D. Foran, N. Do, S. Golemati, T. Kurc, K. Huang, K. S. Nikita, B. P. Veasey, M. Zervakis, J. H. Saltz, and C. S. Pattichis. Ai in medical imaging informatics: Current challenges and future directions. *IEEE J Biomed Health Inform*, 24(7):1837–1857, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32609615, doi:10.1109/JBHI.2020.2991043.

[82] N. Ullah, F. Guzman-Aroca, F. Martinez-Alvarez, I. De Falco, and G. Sannino. A novel explainable ai framework for medical image classification integrating statistical, visual, and rule-based methods. *Med Image Anal*, 105:103665, 2025. URL: https://www.ncbi.nlm.nih.gov/pubmed/40505210, doi:10.1016/j.media.2025.103665.

[83] C. Mattiuzzi and G. Lippi. Current cancer epidemiology. *J Epidemiol Glob Health*, 9(4):217–222, 2019. URL: https://www.ncbi.nlm.nih.gov/pubmed/31854162, doi:10.2991/jegh.k.191008.001.

[84] R. W. Ruddon. *Cancer Biology*. Oxford University Press Inc., 2007.

[85] J. Kleeff, M. Korc, M. Apte, C. La Vecchia, C. D. Johnson, A. V. Biankin, R. E. Neale, M. Tempero, D. A. Tuveson, R. H. Hruban, and J. P. Neoptolemos. Pancreatic cancer. *Nat Rev Dis Primers*, 2:16022, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/27158978, doi:10.1038/nrdp.2016.22.

[86] J. Balogh, 3rd Victor, D., E. H. Asham, S. G. Burroughs, M. Boktour, A. Saharia, X. Li, R. M. Ghobrial, and Jr. Monsour, H. P. Hepatocellular carcinoma: a review. *J Hepatocell Carcinoma*, 3:41–53, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/27785449, doi:10.2147/JHC.S61146.

[87] B. Vogelstein and K. W. Kinzler. Cancer genes and the pathways they control. *Nat Med*, 10(8):789–99, 2004. URL: https://www.ncbi.nlm.nih.gov/pubmed/15286780, doi:10.1038/nm1087.

[88] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74, 2011. URL: https://www.ncbi.nlm.nih.gov/pubmed/21376230, doi:10.1016/j.cell.2011.02.013.

[89] W. D. Travis. Pathology of lung cancer. *Clin Chest Med*, 32(4):669–92, 2011. URL: https://www.ncbi.nlm.nih.gov/pubmed/22054879, doi:10.1016/j.ccm.2011.08.005.

[90] F. Siddique, M. Shehata, M. Ghazal, S. Contractor, and A. El-Baz. Lung cancer subtyping: A short review. *Cancers (Basel)*, 16(15), 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/39123371, doi:10.3390/cancers16152643.

[91] A. Marusyk and K. Polyak. Tumor heterogeneity: causes and consequences. *Biochim Biophys Acta*, 1805(1):105–17, 2010. URL: https://www.ncbi.nlm.nih.gov/pubmed/19931353, doi:10.1016/j.bbcan.2009.11.002.

[92] O. Menyhart and B. Gyorffy. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Comput Struct Biotechnol J*, 19:949–960, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/33613862, doi:10.1016/j.csbj.2021.01.009.

[93] E. I. Vlachavas, J. Bohn, F. Uckert, and S. Nurnberg. A detailed catalogue of multi-omics methodologies for identification of putative biomarkers and causal molecular networks in translational cancer research. *Int J Mol Sci*, 22(6), 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/33802234, doi:10.3390/ijms22062822.

[94] D. F. Quail and J. A. Joyce. Microenvironmental regulation of tumor progression and metastasis. *Nat Med*, 19(11):1423–37, 2013. URL: https://www.ncbi.nlm.nih.gov/pubmed/24202395, doi:10.1038/nm.3394.

[95] D. Sonkin, A. Thomas, and B. A. Teicher. Cancer treatments: Past, present, and future. *Cancer Genet*, 286-287:18–24, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38909530, doi:10.1016/j.cancergen.2024.06.002.

[96] V. V. Padma. An overview of targeted cancer therapy. *Biomedicine (Taipei)*, 5(4):19, 2015. URL: https://www.ncbi.nlm.nih.gov/pubmed/26613930, doi:10.7603/s40681-015-0019-4.

[97] M. Reck, D. Rodriguez-Abreu, A. G. Robinson, R. Hui, T. Csoszi, A. Fulop, M. Gottfried, N. Peled, A. Tafreshi, S. Cuffe, M. O'Brien, S. Rao, K. Hotta, M. A. Leiby, G. M. Lubiniecki, Y. Shentu, R. Rangwala, J. R. Brahmer, and Keynote Investigators. Pembrolizumab versus chemotherapy for pd-l1-positive non-small-cell lung cancer. *N Engl J Med*, 375(19):1823–1833, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/27718847, doi:10.1056/NEJMoa1606774.

[98] H. Li, P. A. van der Merwe, and S. Sivakumar. Biomarkers of response to pd-1 pathway blockade. *Br J Cancer*, 126(12):1663–1675, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35228677, doi:10.1038/s41416-022-01743-4.

[99] M. Yi, D. Jiao, H. Xu, Q. Liu, W. Zhao, X. Han, and K. Wu. Biomarkers for predicting efficacy of pd-1/pd-l1 inhibitors. *Mol Cancer*, 17(1):129, 2018. URL: https://www.ncbi.nlm.nih.gov/pubmed/30139382, doi:10.1186/s12943-018-0864-3.

[100] Filippo G. Dall'Olio, Ilaria Maggio, Maria Massucci, Veronica Mollica, Benedetta Fragomeno, and Andrea Ardizzoni. Ecog performance status ≥ 2 as a prognostic factor in patients with advanced non small cell lung cancer treated with immune checkpoint inhibitors—a systematic review and meta-analysis of real world data. *Lung Cancer*, 145:95–104, 2020. doi:10.1016/j.lungcan.2020.04.027.

[101] C. L. Han, G. X. Meng, Z. N. Ding, Z. R. Dong, Z. Q. Chen, J. G. Hong, L. J. Yan, H. Liu, B. W. Tian, L. S. Yang, J. S. Xue, and T. Li. The predictive potential of the baseline c-reactive protein levels for the efficiency of

immune checkpoint inhibitors in cancer patients: A systematic review and meta-analysis. *Front Immunol*, 13:827788, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35211122, doi:10.3389/fimmu.2022.827788.

[102] F. Soria, A. I. Beleni, D. D'Andrea, I. Resch, K. M. Gust, P. Gontero, and S. F. Shariat. Pseudoprogression and hyperprogression during immune checkpoint inhibitor therapy for urothelial and kidney cancer. *World J Urol*, 36(11):1703–1709, 2018. URL: https://www.ncbi.nlm.nih.gov/pubmed/29549485, doi:10.1007/s00345-018-2264-0.

[103] M. J. Duffy, N. Harbeck, M. Nap, R. Molina, A. Nicolini, E. Senkus, and F. Cardoso. Clinical use of biomarkers in breast cancer: Updated guidelines from the european group on tumor markers (egtm). *Eur J Cancer*, 75:284–298, 2017. URL: https://www.ncbi.nlm.nih.gov/pubmed/28259011, doi:10.1016/j.ejca.2017.01.017.

[104] Lesley Seymour, Jan Bogaerts, Andrea Perrone, Robert Ford, Lawrence H. Schwartz, Sumithra Mandrekar, Nancy U. Lin, Saskia Litière, Janet Dancey, Alice Chen, F. Stephen Hodi, Patrick Therasse, Otto S. Hoekstra, Lalitha K. Shankar, Jedd D. Wolchok, Marcus Ballinger, Caroline Caramella, and Elisabeth G. E. de Vries. irecist: guidelines for response criteria for use in trials testing immunotherapeutics. *The Lancet Oncology*, 18(3):e143–e152, 2017. doi:10.1016/S1470-2045(17)30074-8.

[105] J. K. Aronson and R. E. Ferner. Biomarkers-a general review. *Curr Protoc Pharmacol*, 76:9 23 1–9 23 17, 2017. URL: https://www.ncbi.nlm.nih.gov/pubmed/28306150, doi:10.1002/cpph.19.

[106] A. Ahmad, M. Imran, and H. Ahsan. Biomarkers as biomedical bioindicators: Approaches and techniques for the detection, analysis, and validation of novel biomarkers of diseases. *Pharmaceutics*, 15(6), 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/37376078, doi:10.3390/pharmaceutics15061630.

[107] F. Y. Chiu and Y. Yen. Imaging biomarkers for clinical applications in neuro-oncology: current status and future perspectives. *Biomark Res*, 11(1):35, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/36991494, doi:10.1186/s40364-023-00476-7.

[108] J. P. O'Connor, E. O. Aboagye, J. E. Adams, H. J. Aerts, S. F. Barrington, A. J. Beer, R. Boellaard, S. E. Bohndiek, M. Brady, G. Brown, D. L. Buckley, T. L. Chenevert, L. P. Clarke, S. Collette, G. J. Cook, N. M. deSouza, J. C.

Dickson, C. Dive, J. L. Evelhoch, C. Faivre-Finn, F. A. Gallagher, F. J. Gilbert, R. J. Gillies, V. Goh, J. R. Griffiths, A. M. Groves, S. Halligan, A. L. Harris, D. J. Hawkes, O. S. Hoekstra, E. P. Huang, B. F. Hutton, E. F. Jackson, G. C. Jayson, A. Jones, D. M. Koh, D. Lacombe, P. Lambin, N. Lassau, M. O. Leach, T. Y. Lee, E. L. Leen, J. S. Lewis, Y. Liu, M. F. Lythgoe, P. Manoharan, R. J. Maxwell, K. A. Miles, B. Morgan, S. Morris, T. Ng, A. R. Padhani, G. J. Parker, M. Partridge, A. P. Pathak, A. C. Peet, S. Punwani, A. R. Reynolds, S. P. Robinson, L. K. Shankar, R. A. Sharma, D. Soloviev, S. Stroobants, D. C. Sullivan, S. A. Taylor, P. S. Tofts, G. M. Tozer, M. van Herk, S. Walker-Samuel, J. Wason, K. J. Williams, P. Workman, T. E. Yankeelov, K. M. Brindle, L. M. McShane, A. Jackson, and J. C. Waterton. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol*, 14(3):169–186, 2017. URL: `https://www.ncbi.nlm.nih.gov/pubmed/27725679`, `doi:10.1038/nrclinonc.2016.162`.

[109] A. Nejatie, S. S. Yee, A. Jeter, and H. U. Saragovi. The cancer glycocode as a family of diagnostic biomarkers, exemplified by tumor-associated gangliosides. *Front Oncol*, 13:1261090, 2023. URL: `https://www.ncbi.nlm.nih.gov/pubmed/37954075`, `doi:10.3389/fonc.2023.1261090`.

[110] V. Budach and I. Tinhofer. Novel prognostic clinical factors and biomarkers for outcome prediction in head and neck cancer: a systematic review. *Lancet Oncol*, 20(6):e313–e326, 2019. URL: `https://www.ncbi.nlm.nih.gov/pubmed/31162105`, `doi:10.1016/S1470-2045(19)30177-9`.

[111] R. I. Mihaila, A. S. Gheorghe, D. L. Zob, and D. L. Stanculeanu. The importance of predictive biomarkers and their correlation with the response to immunotherapy in solid tumors-impact on clinical practice. *Biomedicines*, 12(9), 2024. URL: `https://www.ncbi.nlm.nih.gov/pubmed/39335659`, `doi:10.3390/biomedicines12092146`.

[112] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher, D. B. Goldgof, L. O. Hall, P. Lambin, Y. Balagurunathan, R. A. Gatenby, and R. J. Gillies. Radiomics: the process and the challenges. *Magn Reson Imaging*, 30(9):1234–48, 2012. URL: `https://www.ncbi.nlm.nih.gov/pubmed/22898692`, `doi:10.1016/j.mri.2012.06.010`.

[113] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies,

and P. Lambin. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*, 5:4006, 2014. URL: https://www.ncbi.nlm.nih.gov/pubmed/24892406, doi:10.1038/ncomms5006.

[114] R. J. Gillies, P. E. Kinahan, and H. Hricak. Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2):563–77, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/26579733, doi:10.1148/radiol.2015151169.

[115] J. Song, Y. Yin, H. Wang, Z. Chang, Z. Liu, and L. Cui. A review of original articles published in the emerging field of radiomics. *Eur J Radiol*, 127:108991, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32334372, doi:10.1016/j.ejrad.2020.108991.

[116] Z. Wang, L. Wang, and Y. Wang. Radiomics in glioma: emerging trends and challenges. *Ann Clin Transl Neurol*, 12(3):460–477, 2025. URL: https://www.ncbi.nlm.nih.gov/pubmed/39901654, doi:10.1002/acn3.52306.

[117] H. Sotoudeh, A. H. Sarrami, G. H. Roberson, O. Shafaat, Z. Sadaatpour, A. Rezaei, G. Choudhary, A. Singhal, E. Sotoudeh, and M. Tanwar. Emerging applications of radiomics in neurological disorders: A review. *Cureus*, 13(12):e20080, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/34987940, doi:10.7759/cureus.20080.

[118] P. Lohmann, K. Bousabarah, M. Hoevels, and H. Treuer. Radiomics in radiation oncology-basics, methods, and limitations. *Strahlenther Onkol*, 196(10):848–855, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32647917, doi:10.1007/s00066-020-01663-3.

[119] A. Zwanenburg, M. Vallieres, M. A. Abdalah, H. J. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G. J. R. Cook, C. Davatzikos, A. Depeursinge, M. C. Desseroit, N. Dinapoli, C. V. Dinh, S. Echegaray, I. El Naqa, A. Y. Fedorov, R. Gatta, R. J. Gillies, V. Goh, M. Gotz, M. Guckenberger, S. M. Ha, M. Hatt, F. Isensee, P. Lambin, S. Leger, R. T. H. Leijenaar, J. Lenkowicz, F. Lippert, A. Losnegard, K. H. Maier-Hein, O. Morin, H. Muller, S. Napel, C. Nioche, F. Orlhac, S. Pati, E. A. G. Pfaehler, A. Rahmim, A. U. K. Rao, J. Scherer, M. M. Siddique, N. M. Sijtsema, J. Socarras Fernandez, E. Spezi, Rjhm Steenbakkers, S. Tanadini-Lang, D. Thorwarth, E. G. C. Troost, T. Upadhaya, V. Valentini, L. V. van Dijk, J. van Griethuysen, F. H. P. van Velden, P. Whybra, C. Richter, and S. Lock.

The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*, 295(2):328–338, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32154773https://escholarship.org/content/qt6316d65s/qt6316d65s.pdf, doi:10.1148/radiol.2020191145.

[120] J. Zhong, J. Lu, G. Zhang, S. Mao, H. Chen, Q. Yin, Y. Hu, Y. Xing, D. Ding, X. Ge, H. Zhang, and W. Yao. An overview of meta-analyses on radiomics: more evidence is needed to support clinical translation. *Insights Imaging*, 14(1):111, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/37336830, doi:10.1186/s13244-023-01437-2.

[121] A. Traverso, L. Wee, A. Dekker, and R. Gillies. Repeatability and reproducibility of radiomic features: A systematic review. *Int J Radiat Oncol Biol Phys*, 102(4):1143–1158, 2018. URL: https://www.ncbi.nlm.nih.gov/pubmed/30170872, doi:10.1016/j.ijrobp.2018.05.053.

[122] V. Kumar, Y. Gu, S. Basu, A. Berglund, S. A. Eschrich, M. B. Schabath, K. Forster, H. J. Aerts, A. Dekker, D. Fenstermacher, D. B. Goldgof, L. O. Hall, P. Lambin, Y. Balagurunathan, R. A. Gatenby, and R. J. Gillies. Radiomics: the process and the challenges. *Magn Reson Imaging*, 30(9):1234–48, 2012. URL: https://www.ncbi.nlm.nih.gov/pubmed/22898692, doi:10.1016/j.mri.2012.06.010.

[123] S. Reuze, A. Schernberg, F. Orlhac, R. Sun, C. Chargari, L. Dercle, E. Deutsch, I. Buvat, and C. Robert. Radiomics in nuclear medicine applied to radiation therapy: Methods, pitfalls, and challenges. *Int J Radiat Oncol Biol Phys*, 102(4):1117–1142, 2018. URL: https://www.ncbi.nlm.nih.gov/pubmed/30064704, doi:10.1016/j.ijrobp.2018.05.022.

[124] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): the tripod statement. *BMC Med*, 13:1, 2015. URL: https://www.ncbi.nlm.nih.gov/pubmed/25563062, doi:10.1186/s12916-014-0241-z.

[125] J. Mongan, L. Moy, and Jr. Kahn, C. E. Checklist for artificial intelligence in medical imaging (claim): A guide for authors and reviewers. *Radiol Artif Intell*, 2(2):e200029, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/33937821, doi:10.1148/ryai.2020200029.

[126] D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang, B. Taylor, E. Rodriguez-Rivera, C. Dodge, A. K. Jones, and L. Court. Measuring computed tomography scanner variability of radiomics features. *Invest Radiol*, 50(11):757–65, 2015. URL: https://www.ncbi.nlm.nih.gov/pubmed/26115366, doi:10.1097/RLI.0000000000000180.

[127] M. Shafiq-Ul-Hassan, G. G. Zhang, K. Latifi, G. Ullah, D. C. Hunt, Y. Balagurunathan, M. A. Abdalah, M. B. Schabath, D. G. Goldgof, D. Mackin, L. E. Court, R. J. Gillies, and E. G. Moros. Intrinsic dependencies of ct radiomic features on voxel size and number of gray levels. *Med Phys*, 44(3):1050–1062, 2017. URL: https://www.ncbi.nlm.nih.gov/pubmed/28112418, doi:10.1002/mp.12123.

[128] Rthm Larue, J. E. van Timmeren, E. E. C. de Jong, G. Feliciani, R. T. H. Leijenaar, W. M. J. Schreurs, M. N. Sosef, Fhpj Raat, F. H. R. van der Zande, M. Das, W. van Elmpt, and P. Lambin. Influence of gray level discretization on radiomic feature stability for different ct scanners, tube currents and slice thicknesses: a comprehensive phantom study. *Acta Oncol*, 56(11):1544–1553, 2017. URL: https://www.ncbi.nlm.nih.gov/pubmed/28885084, doi:10.1080/0284186X.2017.1351624.

[129] F. Orlhac, A. Lecler, J. Savatovski, J. Goya-Outi, C. Nioche, F. Charbonneau, N. Ayache, F. Frouin, L. Duron, and I. Buvat. How can we combat multicenter variability in mr radiomics? validation of a correction procedure. *Eur Radiol*, 31(4):2272–2280, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/32975661, doi:10.1007/s00330-020-07284-9.

[130] S. Volpe, L. J. Isaksson, M. Zaffaroni, M. Pepa, S. Raimondi, F. Botta, G. Lo Presti, M. G. Vincini, C. Rampinelli, M. Cremonesi, F. de Marinis, L. Spaggiari, S. Gandini, M. Guckenberger, R. Orecchia, and B. A. Jereczek-Fossa. Impact of image filtering and assessment of volume-confounding effects on ct radiomic features and derived survival models in non-small cell lung cancer. *Transl Lung Cancer Res*, 11(12):2452–2463, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/36636424, doi:10.21037/tlcr-22-248.

[131] J. Kong, J. Zheng, J. Wu, S. Wu, J. Cai, X. Diao, W. Xie, X. Chen, H. Yu, L. Huang, H. Fang, X. Fan, H. Qin, Y. Li, Z. Wu, J. Huang, and T. Lin. Development of a radiomics model to diagnose pheochromocytoma preoperatively: a multicenter study with prospective validation. *J Transl Med*,

20(1):31, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35033104, doi:10.1186/s12967-022-03233-w.

[132] Fuquan Liu, Zhenyuan Ning, Yanna Liu, Dengxiang Liu, Jie Tian, Hongwu Luo, Weimin An, Yifei Huang, Jialiang Zou, Chuan Liu, Changchun Liu, Lei Wang, Zaiyi Liu, Ruizhao Qi, Changzeng Zuo, Qingge Zhang, Jitao Wang, Dawei Zhao, Yongli Duan, Baogang Peng, Xingshun Qi, Yuening Zhang, Yongping Yang, Jinlin Hou, Jiahong Dong, Zhiwei Li, Huiguo Ding, Yu Zhang, and Xiaolong Qi. Development and validation of a radiomics signature for clinically significant portal hypertension in cirrhosis (chess1701): a prospective multicenter study. *eBioMedicine*, 36:151–158, 2018. doi:10.1016/j.ebiom.2018.09.023.

[133] J. Bleker, C. Roest, D. Yakar, H. Huisman, and T. C. Kwee. The effect of image resampling on the performance of radiomics-based artificial intelligence in multicenter prostate mri. *J Magn Reson Imaging*, 59(5):1800–1806, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/37572098, doi:10.1002/jmri.28935.

[134] D. Marfisi, C. Tessa, C. Marzi, J. Del Meglio, S. Linsalata, R. Borgheresi, A. Lilli, R. Lazzarini, L. Salvatori, C. Vignali, A. Barucci, M. Mascalchi, G. Casolo, S. Diciotti, A. C. Traino, and M. Giannelli. Image resampling and discretization effect on the estimate of myocardial radiomic features from t1 and t2 mapping in hypertrophic cardiomyopathy. *Sci Rep*, 12(1):10186, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35715531, doi:10.1038/s41598-022-13937-0.

[135] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, and Hjwl Aerts. Machine learning methods for quantitative radiomic biomarkers. *Sci Rep*, 5:13087, 2015. URL: https://www.ncbi.nlm.nih.gov/pubmed/26278466, doi:10.1038/srep13087.

[136] W. C. Zhang, Y. Guo, and Q. Y. Jin. Radiomics and its feature selection: A review. *Symmetry-Basel*, 15(10), 2023. URL: <GotoISI>://WOS:001093697400001, doi:ARTN183410.3390/sym15101834.

[137] H. Desaire. How (not) to generate a highly predictive biomarker panel using machine learning. *J Proteome Res*, 21(9):2071–2074, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/36004690, doi:10.1021/acs.jproteome.2c00117.

[138] F. Mariotti, A. Agostini, A. Borgheresi, M. Marchegiani, A. Zannotti, G. Giacomelli, L. Pierpaoli, E. Tola, E. Galiffa, and A. Giovagnoni. Insights into radiomics: a comprehensive review for beginners. *Clin Transl Oncol*, 2025. URL: https://www.ncbi.nlm.nih.gov/pubmed/40355777, doi:10.1007/s12094-025-03939-5.

[139] M. Hatt, C. C. Le Rest, F. Tixier, B. Badic, U. Schick, and D. Visvikis. Radiomics: Data are also images. *J Nucl Med*, 60(Suppl 2):38S–44S, 2019. URL: https://www.ncbi.nlm.nih.gov/pubmed/31481588, doi:10.2967/jnumed.118.220582.

[140] N. S. Mohd Haniff, K. H. Ng, I. Kamal, N. Mohd Zain, and M. K. Abdul Karim. Systematic review and meta-analysis on the classification metrics of machine learning algorithm based radiomics in hepatocellular carcinoma diagnosis. *Heliyon*, 10(16):e36313, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/39253167, doi:10.1016/j.heliyon.2024.e36313.

[141] Jake Lever, Martin Krzywinski, and Naomi Altman. Model selection and overfitting. *Nature Methods*, 13(9):703–704, 2016. doi:10.1038/nmeth.3968.

[142] T. J. Bradshaw, Z. Huemann, J. Hu, and A. Rahmim. A guide to cross-validation for artificial intelligence in medical imaging. *Radiol Artif Intell*, 5(4):e220232, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/37529208, doi:10.1148/ryai.220232.

[143] G. Varoquaux and V. Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit Med*, 5(1):48, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35413988, doi:10.1038/s41746-022-00592-y.

[144] L. Maier-Hein, A. Reinke, P. Godau, M. D. Tizabi, F. Buettner, E. Christodoulou, B. Glocker, F. Isensee, J. Kleesiek, M. Kozubek, M. Reyes, M. A. Riegler, M. Wiesenfarth, A. E. Kavur, C. H. Sudre, M. Baumgartner, M. Eisenmann, D. Heckmann-Notzel, T. Radsch, L. Acion, M. Antonelli, T. Arbel, S. Bakas, A. Benis, M. B. Blaschko, M. J. Cardoso, V. Cheplygina, B. A. Cimini, G. S. Collins, K. Farahani, L. Ferrer, A. Galdran, B. van Ginneken, R. Haase, D. A. Hashimoto, M. M. Hoffman, M. Huisman, P. Jannin, C. E. Kahn, D. Kainmueller, B. Kainz, A. Karargyris, A. Karthikesalingam, F. Kofler, A. Kopp-Schneider, A. Kreshuk, T. Kurc, B. A. Landman, G. Litjens, A. Madani, K. Maier-Hein, A. L. Martel, P. Mattson, E. Meijering, B. Menze,

K. G. M. Moons, H. Muller, B. Nichyporuk, F. Nickel, J. Petersen, N. Rajpoot, N. Rieke, J. Saez-Rodriguez, C. I. Sanchez, S. Shetty, M. van Smeden, R. M. Summers, A. A. Taha, A. Tiulpin, S. A. Tsaftaris, B. Van Calster, G. Varoquaux, and P. F. Jager. Metrics reloaded: recommendations for image analysis validation. *Nat Methods*, 21(2):195–212, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38347141, doi:10.1038/s41592-023-02151-z.

[145] G. Varoquaux. Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage*, 180(Pt A):68–77, 2018. URL: https://www.ncbi.nlm.nih.gov/pubmed/28655633, doi:10.1016/j.neuroimage.2017.06.061.

[146] N. H. Di Cara, N. Zelenka, H. Day, E. D. S. Bennet, V. Hanschke, V. Maggio, O. Michalec, C. Radclyffe, R. Shkunov, E. Tonkin, Z. Turner, and K. Wells. Data ethics club: Creating a collaborative space to discuss data ethics. *Patterns (N Y)*, 3(7):100537, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35845834, doi:10.1016/j.patter.2022.100537.

[147] D. Wilimitis and C. G. Walsh. Practical considerations and applied examples of cross-validation for model development and evaluation in health care: Tutorial. *JMIR AI*, 2:e49023, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/38875530, doi:10.2196/49023.

[148] H. Ghasemzadeh, R. E. Hillman, and D. D. Mehta. Toward generalizable machine learning models in speech, language, and hearing sciences: Estimating sample size and reducing overfitting. *J Speech Lang Hear Res*, 67(3):753–781, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38386017, doi:10.1044/2023_JSLHR-23-00273.

[149] I. Tougui, A. Jilbab, and J. E. Mhamdi. Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. *Healthc Inform Res*, 27(3):189–199, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/34384201, doi:10.4258/hir.2021.27.3.189.

[150] V. Singh, M. Pencina, A. J. Einstein, J. X. Liang, D. S. Berman, and P. Slomka. Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Sci Rep*, 11(1):14490, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/34262098, doi:10.1038/s41598-021-93651-5.

[151] Bernd Bischl, Martin Binder, Michel Lang, Tobias Pielok, Jakob Richter, Stefan Coors, Janek Thomas, Theresa Ullmann, Marc Becker, Anne-Laure Boulesteix,

Difan Deng, and Marius Lindauer. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, 13(2), 2023. doi:10.1002/widm.1484.

[152] Yasser Ali, Emad Awwad, Muna Al-Razgan, and Ali Maarouf. Hyperparameter search for machine learning algorithms for optimizing the computational complexity. *Processes*, 11(2), 2023. doi:10.3390/pr11020349.

[153] Takuya Akiba Koyama, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori. Optuna: A next-generation hyperparameter optimization framework. *arXiv*, 2019. URL: https://arxiv.org/abs/1907.10902, doi:10.48550/arXiv.1907.10902.

[154] W. Zhang, Q. Guo, Y. Zhu, M. Wang, T. Zhang, G. Cheng, Q. Zhang, and H. Ding. Cross-institutional evaluation of deep learning and radiomics models in predicting microvascular invasion in hepatocellular carcinoma: validity, robustness, and ultrasound modality efficacy comparison. *Cancer Imaging*, 24(1):142, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/39438929, doi:10.1186/s40644-024-00790-9.

[155] E. Bailly, C. Baranton, S. Valot, A. Vincent, H. Begue, C. Beclere, A. Bonnin, D. Costa, P. Poirier, L. Basmaciyan, and F. Dalle. Performance of 30 protocol combinations for the detection of cryptosporidium parvum in stool samples. *J Microbiol Immunol Infect*, 58(3):368–375, 2025. URL: https://www.ncbi.nlm.nih.gov/pubmed/39984420, doi:10.1016/j.jmii.2025.01.003.

[156] W. J. Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–5, 1950. URL: https://www.ncbi.nlm.nih.gov/pubmed/15405679, doi:10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3.

[157] Rocío Aznar-Gimeno, Luis M. Esteban, Gerardo Sanz, and Rafael del Hoyo-Alonso. Comparing the min–max–median/iqr approach with the min–max approach, logistic regression and xgboost, maximising the youden index. *Symmetry*, 15(3), 2023. doi:10.3390/sym15030756.

[158] Y. Wang, Y. Zhang, J. Xiao, X. Geng, L. Han, and J. Luo. Multicenter integration of mr radiomics, deep learning, and clinical indicators for predicting hepatocellular carcinoma recurrence after thermal ablation. *J Hepatocell Carcinoma*, 11:1861–1874, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/39372710, doi:10.2147/JHC.S482760.

[159] S. Majumder, S. Katz, D. Kontos, and L. Roshkovan. State of the art: radiomics and radiomics-related artificial intelligence on the road to clinical translation. *BJR Open*, 6(1):tzad004, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38352179, doi:10.1093/bjro/tzad004.

[160] B. Kocak, A. Borgheresi, A. Ponsiglione, A. E. Andreychenko, A. U. Cavallo, A. Stanzione, F. M. Doniselli, F. Vernuccio, M. Triantafyllou, R. Cannella, R. Trotta, S. Ghezzo, T. Akinci D'Antonoli, and R. Cuocolo. Explanation and elaboration with examples for clear (clear-e3): an eusomii radiomics auditing group initiative. *Eur Radiol Exp*, 8(1):72, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38740707, doi:10.1186/s41747-024-00471-z.

[161] J. Santinha, D. Pinto Dos Santos, F. Laqua, J. J. Visser, K. B. W. Groot Lipman, M. Dietzel, M. E. Klontzas, R. Cuocolo, S. Gitto, and T. Akinci D'Antonoli. Esr essentials: radiomics-practice recommendations by the european society of medical imaging informatics. *Eur Radiol*, 35(3):1122–1132, 2025. URL: https://www.ncbi.nlm.nih.gov/pubmed/39453470, doi:10.1007/s00330-024-11093-9.

[162] Helen Smith. Clinical ai: opacity, accountability, responsibility and liability. *Ai & Society*, 36(2):535–545, 2020. doi:10.1007/s00146-020-01019-6.

[163] R. Fusco, V. Granata, I. Simonetti, S. V. Setola, M. A. D. Iasevoli, F. Tovecci, C. M. P. Lamanna, F. Izzo, B. Pecori, and A. Petrillo. An informative review of radiomics studies on cancer imaging: The main findings, challenges and limitations of the methodologies. *Curr Oncol*, 31(1):403–424, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38248112, doi:10.3390/curroncol31010027.

[164] A. Stanzione, R. Cuocolo, L. Ugga, F. Verde, V. Romeo, A. Brunetti, and S. Maurea. Oncologic imaging and radiomics: A walkthrough review of methodological challenges. *Cancers (Basel)*, 14(19), 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/36230793, doi:10.3390/cancers14194871.

[165] S. P. Singh, L. Wang, S. Gupta, H. Goli, P. Padmanabhan, and B. Gulyas. 3d deep learning on medical images: A review. *Sensors (Basel)*, 20(18), 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32906819, doi:10.3390/s20185097.

[166] J. Zhong, X. Liu, J. Lu, J. Yang, G. Zhang, S. Mao, H. Chen, Q. Yin, Q. Cen, R. Jiang, Y. Song, M. Lu, J. Chu, Y. Xing, Y. Hu, D. Ding, X. Ge, H. Zhang, and

W. Yao. Overlooked and underpowered: a meta-research addressing sample size in radiomics prediction models for binary outcomes. *Eur Radiol*, 35(3):1146–1156, 2025. URL: https://www.ncbi.nlm.nih.gov/pubmed/39789271, doi:10.1007/s00330-024-11331-0.

[167] H. Laci, K. Sevrani, and S. Iqbal. Deep learning approaches for classification tasks in medical x-ray, mri, and ultrasound images: a scoping review. *BMC Med Imaging*, 25(1):156, 2025. URL: https://www.ncbi.nlm.nih.gov/pubmed/40335965, doi:10.1186/s12880-025-01701-5.

[168] B. Varghese, F. Chen, D. Hwang, S. L. Palmer, A. L. De Castro Abreu, O. Ukimura, M. Aron, M. Aron, I. Gill, V. Duddalwar, and G. Pandey. Objective risk stratification of prostate cancer using machine learning and radiomics applied to multiparametric magnetic resonance images. *Sci Rep*, 9(1):1570, 2019. URL: https://www.ncbi.nlm.nih.gov/pubmed/30733585, doi:10.1038/s41598-018-38381-x.

[169] Y. Zhang, A. Oikonomou, A. Wong, M. A. Haider, and F. Khalvati. Radiomics-based prognosis analysis for non-small cell lung cancer. *Sci Rep*, 7:46349, 2017. URL: https://www.ncbi.nlm.nih.gov/pubmed/28418006, doi:10.1038/srep46349.

[170] H. Naseri, S. Skamene, M. Tolba, M. D. Faye, P. Ramia, J. Khriguian, H. Patrick, A. X. Andrade Hernandez, M. David, and J. Kildea. Radiomics-based machine learning models to distinguish between metastatic and healthy bone using lesion-center-based geometric regions of interest. *Sci Rep*, 12(1):9866, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35701461, doi:10.1038/s41598-022-13379-8.

[171] V. S. Parekh and M. A. Jacobs. Deep learning and radiomics in precision medicine. *Expert Rev Precis Med Drug Dev*, 4(2):59–72, 2019. URL: https://www.ncbi.nlm.nih.gov/pubmed/31080889, doi:10.1080/23808993.2019.1585805.

[172] M. Astaraki, G. Yang, Y. Zakko, I. Toma-Dasu, O. Smedby, and C. Wang. A comparative study of radiomics and deep-learning based methods for pulmonary nodule malignancy prediction in low dose ct images. *Front Oncol*, 11:737368, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/34976794, doi:10.3389/fonc.2021.737368.

[173] H. P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou. Deep learning in medical image analysis. *Adv Exp Med Biol*, 1213:3–21, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32030660, doi:10.1007/978-3-030-33128-3_1.

[174] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof. Diffusion models in medical imaging: A comprehensive survey. *Med Image Anal*, 88:102846, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/37295311, doi:10.1016/j.media.2023.102846.

[175] George Webber and Andrew J. Reader. Diffusion models for medical image reconstruction. *BJR—Artificial Intelligence*, 1(1), 2024. doi:10.1093/bjrai/ubae013.

[176] M. A. Bahloul, S. Jabeen, S. Benoumhani, H. A. Alsaleh, Z. Belkhatir, and A. Al-Wabil. Advancements in synthetic ct generation from mri: A review of techniques, and trends in radiation therapy planning. *J Appl Clin Med Phys*, 25(11):e14499, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/39325781, doi:10.1002/acm2.14499.

[177] M. K. Sherwani and S. Gopalakrishnan. A systematic literature review: deep learning techniques for synthetic medical image generation and their applications in radiotherapy. *Front Radiol*, 4:1385742, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38601888, doi:10.3389/fradi.2024.1385742.

[178] K. K. Bressem, L. C. Adams, C. Erxleben, B. Hamm, S. M. Niehues, and J. L. Vahldiek. Comparing different deep learning architectures for classification of chest radiographs. *Sci Rep*, 10(1):13590, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32788602, doi:10.1038/s41598-020-70479-z.

[179] Y. X. Tang, Y. B. Tang, Y. Peng, K. Yan, M. Bagheri, B. A. Redd, C. J. Brandon, Z. Lu, M. Han, J. Xiao, and R. M. Summers. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ Digit Med*, 3:70, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32435698, doi:10.1038/s41746-020-0273-z.

[180] V. K. Raghu, J. Weiss, U. Hoffmann, Hjwl Aerts, and M. T. Lu. Deep learning to estimate biological age from chest radiographs. *JACC Cardiovasc Imaging*, 14(11):2226–2236, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/33744131, doi:10.1016/j.jcmg.2021.01.008.

[181] H. Ieki, K. Ito, M. Saji, R. Kawakami, Y. Nagatomo, K. Takada, T. Kariyasu, H. Machida, S. Koyama, H. Yoshida, R. Kurosawa, H. Matsunaga, K. Miyazawa, K. Ozaki, Y. Onouchi, S. Katsushika, R. Matsuoka, H. Shinohara, T. Yamaguchi, S. Kodera, Y. Higashikuni, K. Fujiu, H. Akazawa, N. Iguchi, M. Isobe, T. Yoshikawa, and I. Komuro. Deep learning-based age estimation from chest x-rays indicates cardiovascular prognosis. *Commun Med (Lond)*, 2(1):159, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/36494479, doi:10.1038/s43856-022-00220-6.

[182] J. Weiss, V. K. Raghu, D. Bontempi, D. C. Christiani, R. H. Mak, M. T. Lu, and Hjwl Aerts. Deep learning to estimate lung disease mortality from chest radiographs. *Nat Commun*, 14(1):2797, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/37193717, doi:10.1038/s41467-023-37758-5.

[183] E. Calli, E. Sogancioglu, B. van Ginneken, K. G. van Leeuwen, and K. Murphy. Deep learning for chest x-ray analysis: A survey. *Med Image Anal*, 72:102125, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/34171622, doi:10.1016/j.media.2021.102125.

[184] A. Diaz-Pinto, S. Alle, V. Nath, Y. Tang, A. Ihsani, M. Asad, F. Perez-Garcia, P. Mehta, W. Li, M. Flores, H. R. Roth, T. Vercauteren, D. Xu, P. Dogra, S. Ourselin, A. Feng, and M. J. Cardoso. Monai label: A framework for ai-assisted interactive labeling of 3d medical images. *Med Image Anal*, 95:103207, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38776843, doi:10.1016/j.media.2024.103207.

[185] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nat Biomed Eng*, 6(12):1399–1406, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/36109605, doi:10.1038/s41551-022-00936-9.

[186] Y. Zhou, M. A. Chia, S. K. Wagner, M. S. Ayhan, D. J. Williamson, R. R. Struyven, T. Liu, M. Xu, M. G. Lozano, P. Woodward-Court, Y. Kihara, U. K. Biobank Eye, Consortium Vision, A. Altmann, A. Y. Lee, E. J. Topol, A. K. Denniston, D. C. Alexander, and P. A. Keane. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/37704728, doi:10.1038/s41586-023-06555-x.

[187] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *Nejm Ai*, 2(1), 2025. doi:10.1056/AIoa2400640.

[188] X. Mei, Z. Liu, P. M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K. E. Link, T. Yang, Y. Wang, H. Greenspan, T. Deyer, Z. A. Fayad, and Y. Yang. Radimagenet: An open radiologic deep learning research dataset for effective transfer learning. *Radiol Artif Intell*, 4(5):e210315, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/36204533, doi:10.1148/ryai.210315.

[189] M. Lindauer, K. Eggensperger, M. Feurer, A. Biedenkapp, D. Deng, C. Benjamins, T. Ruhopf, R. Sass, and Hutter F. Smac3: A versatile bayesian optimization package for hyperparameter optimization. *Journal of Machine Learning Research*, 23:1–9, 2022. URL: http://jmlr.org/papers/v23/21-0888.html, doi:10.48550/arXiv.2109.09831.

[190] Sebastian R. van der Voort and Martijn P. A. Starmans. Predict: A radiomics extensive digital interchangeable classification toolkit (predict), 2023. doi:10.5281/zenodo.8246103.

[191] F. Pedregosa, G. Varoquaux, A Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. doi:10.48550/arXiv.1201.0490.

[192] M. P. A. Starmans, F. E. Buisman, M. Renckens, Feja Willemssen, S. R. van der Voort, B. Groot Koerkamp, D. J. Grunhagen, W. J. Niessen, P. B. Vermeulen, C. Verhoef, J. J. Visser, and S. Klein. Distinguishing pure histopathological growth patterns of colorectal liver metastases on ct using deep learning and radiomics: a pilot study. *Clin Exp Metastasis*, 38(5):483–494, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/34533669, doi:10.1007/s10585-021-10119-6.

[193] L. Angus, M. P. A. Starmans, A. Rajicic, A. E. Odink, M. Jalving, W. J. Niessen, J. J. Visser, S. Sleijfer, S. Klein, and A. A. M. van der Veldt. The braf

p.v600e mutation status of melanoma lung metastases cannot be discriminated on computed tomography by lidc criteria nor radiomics using machine learning. *J Pers Med*, 11(4), 2021. doi:10.3390/jpm11040257.

[194] M. P. A. Starmans, M. J. M. Timbergen, M. Vos, M. Renckens, D. J. Grunhagen, Gjlh van Leenders, R. S. Dwarkasing, Feja Willemssen, W. J. Niessen, C. Verhoef, S. Sleijfer, J. J. Visser, and S. Klein. Differential diagnosis and molecular stratification of gastrointestinal stromal tumors on ct images using a radiomics approach. *J Digit Imaging*, 35(2):127–136, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35088185, doi:10.1007/s10278-022-00590-2.

[195] M. J. M. Timbergen, M. P. A. Starmans, G. A. Padmos, D. J. Grunhagen, Gjlh van Leenders, D. F. Hanff, C. Verhoef, W. J. Niessen, S. Sleijfer, S. Klein, and J. J. Visser. Differential diagnosis and mutation stratification of desmoid-type fibromatosis on mri using radiomics. *Eur J Radiol*, 131:109266, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32971431, doi:10.1016/j.ejrad.2020.109266.

[196] M. P. A. Starmans, R. L. Miclea, V. Vilgrain, M. Ronot, Y. Purcell, J. Verbeek, W. J. Niessen, J. N. M. Ijzermans, R. A. de Man, M. Doukas, S. Klein, and M. G. Thomeer. Automated assessment of t2-weighted mri to differentiate malignant and benign primary solid liver lesions in noncirrhotic livers using radiomics. *Acad Radiol*, 31(3):870–879, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/37648580, doi:10.1016/j.acra.2023.07.024.

[197] M. Vos, M. P. A. Starmans, M. J. M. Timbergen, S. R. van der Voort, G. A. Padmos, W. Kessels, W. J. Niessen, Gjlh van Leenders, D. J. Grunhagen, S. Sleijfer, C. Verhoef, S. Klein, and J. J. Visser. Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on mri. *Br J Surg*, 106(13):1800–1809, 2019. URL: https://www.ncbi.nlm.nih.gov/pubmed/31747074, doi:10.1002/bjs.11410.

[198] Samuel G. Armato III, McLennan, Geoffrey, Luc Bidaut, McNitt-Gray, Michael F., Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, Eric A. Hoffman, Ella A. Kazerooni, Heber MacMahon, Edwin J.R. Van Beek, David Yankelevitz, Alberto M. Biancardi, Peyton H. Bland, Matthew S. Brown, Roger M. Engelmann, Gary E. Laderach, Daniel Max, Richard C. Pais, David P.Y. Qing, Rachael Y. Roberts, Amanda R. Smith, Adam Starkey, Poonam Batra, Philip Caligiuri, Ali Farooqi, Gregory W. Gladish, C. Matilda Jude, Reginald F. Munden, Iva Petkovska,

Leslie E. Quint, Lawrence H. Schwartz, Baskaran Sundaram, Lori E. Dodd, Charles Fenimore, David Gur, Nicholas Petrick, John Freymann, Justin Kirby, Brian Hughes, Alessi Vande Casteele, Sangeeta Gupte, Maha Sallam, Michael D. Heath, Michael H. Kuhn, Ekta Dharaiya, Richard Burns, David S. Fryd, Marcos Salganicoff, Vikram Anand, Uri Shreter, Stephen Vastagh, Barbara Y. Croft, and Laurence P. Clarke. Data from lidc-idri, 2015. URL: https://www.cancerimagingarchive.net/collection/lidc-idri/, doi:10.7937/K9/TCIA.2015.LO9QL9SX.

[199] Z. Liao, Y. Xie, S. Hu, and Y. Xia. Learning from ambiguous labels for lung nodule malignancy prediction. *IEEE Trans Med Imaging*, 41(7):1874–1884, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35130152, doi:10.1109/TMI.2022.3149344.

[200] D. Dai, C. Dong, Z. Li, and S. Xu. Ms-net: Learning to assess the malignant status of a lung nodule by a radiologist and her peers. *J Appl Clin Med Phys*, 24(7):e13964, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/36929569, doi:10.1002/acm2.13964.

[201] 3rd Armato, S. G., G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. Van Beeke, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batrah, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. V. Casteele, S. Gupte, M. Sallamm, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, and B. Y. Croft. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Med Phys*, 38(2):915–31, 2011. URL: https://www.ncbi.nlm.nih.gov/pubmed/21452728, doi:10.1118/1.3528204.

[202] Gallée L., Hillenhagen H., Wolf D., Manoj S., and Goetz M. E2mip challenge, 2024 (Accessed May 2024). URL: https://github.com/XRad-Ulm/E2MIP_LIDCI-IDRI_data.

[203] H. Bonab and F. Can. Less is more: A comprehensive framework for the number of components of ensemble classifiers. *IEEE Trans Neural Netw Learn*

Syst, 30(9):2735–2745, 2019. URL: https://www.ncbi.nlm.nih.gov/pubmed/30629518, doi:10.1109/TNNLS.2018.2886341.

[204] Zardad Khan, Asma Gul, Aris Perperoglou, Miftahuddin Miftahuddin, Osama Mahmoud, Werner Adler, and Berthold Lausen. Ensemble of optimal trees, random forest and random projection ensemble classification. *Advances in Data Analysis and Classification*, 14(1):97–116, 2019. doi:10.1007/s11634-019-00364-9.

[205] B. C. Lowekamp, D. T. Chen, L. Ibanez, and D. Blezek. The design of simpleitk. *Front Neuroinform*, 7:45, 2013. URL: https://www.ncbi.nlm.nih.gov/pubmed/24416015, doi:10.3389/fninf.2013.00045.

[206] S. van der Walt, J. L. Schonberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and contributors scikit image. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014. URL: https://www.ncbi.nlm.nih.gov/pubmed/25024921, doi:10.7717/peerj.453.

[207] Y. Zou, Q. Mao, Z. Zhao, X. Zhou, Y. Pan, Z. Zuo, and W. Zhang. Intratumoural and peritumoural ct-based radiomics for diagnosing lepidic-predominant adenocarcinoma in patients with pure ground-glass nodules: a machine learning approach. *Clin Radiol*, 79(2):e211–e218, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38044199, doi:10.1016/j.crad.2023.11.003.

[208] J. Ding, S. Chen, M. Serrano Sosa, R. Cattell, L. Lei, J. Sun, P. Prasanna, C. Liu, and C. Huang. Optimizing the peritumoral region size in radiomics analysis for sentinel lymph node status prediction in breast cancer. *Acad Radiol*, 29 Suppl 1(Suppl 1):S223–S228, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/33160860, doi:10.1016/j.acra.2020.10.015.

[209] Z. Y. Liang, M. L. Yu, H. Yang, H. J. Li, H. Xie, C. Y. Cui, W. J. Zhang, C. Luo, P. Q. Cai, X. F. Lin, K. F. Liu, L. Xiong, L. Z. Liu, and B. Y. Chen. Beyond the tumor region: Peritumoral radiomics enhances prognostic accuracy in locally advanced rectal cancer. *World J Gastroenterol*, 31(8):99036, 2025. URL: https://www.ncbi.nlm.nih.gov/pubmed/40062323, doi:10.3748/wjg.v31.i8.99036.

[210] Q. Qiu, J. Duan, Z. Duan, X. Meng, C. Ma, J. Zhu, J. Lu, T. Liu, and Y. Yin. Reproducibility and non-redundancy of radiomic features extracted from arterial

phase ct scans in hepatocellular carcinoma patients: impact of tumor segmentation variability. *Quant Imaging Med Surg*, 9(3):453–464, 2019. URL: https://www.ncbi.nlm.nih.gov/pubmed/31032192, doi:10.21037/qims.2019.03.02.

[211] M. Ligero, O. Jordi-Ollero, K. Bernatowicz, A. Garcia-Ruiz, E. Delgado-Munoz, D. Leiva, R. Mast, C. Suarez, R. Sala-Llonch, N. Calvo, M. Escobar, A. Navarro-Martin, G. Villacampa, R. Dienstmann, and R. Perez-Lopez. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. *Eur Radiol*, 31(3):1460–1470, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/32909055, doi:10.1007/s00330-020-07174-0.

[212] Z. Xu, L. Zhao, L. Yin, Y. Liu, Y. Ren, G. Yang, J. Wu, F. Gu, X. Sun, H. Yang, T. Peng, J. Hu, X. Wang, M. Pang, Q. Dai, and G. Zhang. Mri-based machine learning model: A potential modality for predicting cognitive dysfunction in patients with type 2 diabetes mellitus. *Front Bioeng Biotechnol*, 10:1082794, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/36483770, doi:10.3389/fbioe.2022.1082794.

[213] J. Yan, B. Zhang, S. Zhang, J. Cheng, X. Liu, W. Wang, Y. Dong, L. Zhang, X. Mo, Q. Chen, J. Fang, F. Wang, J. Tian, S. Zhang, and Z. Zhang. Quantitative mri-based radiomics for noninvasively predicting molecular subtypes and survival in glioma patients. *NPJ Precis Oncol*, 5(1):72, 2021. URL: https://www.ncbi.nlm.nih.gov/pubmed/34312469, doi:10.1038/s41698-021-00205-z.

[214] C. Zhou, L. Hou, X. Tang, C. Liu, Y. Meng, H. Jia, H. Yang, and S. Zhou. Ct-based radiomics nomogram may predict who can benefit from adaptive radiotherapy in patients with local advanced-nsclc patients. *Radiother Oncol*, 183:109637, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/36963440, doi:10.1016/j.radonc.2023.109637.

[215] P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*, 86(2):420–8, 1979. URL: https://www.ncbi.nlm.nih.gov/pubmed/18839484, doi:10.1037//0033-2909.86.2.420.

[216] Kenneth O. McGraw and S. P. Wong. Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46, 1996. doi:10.1037/1082-989x.1.1.30.

[217] Sebastian Raschka. Mlxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *Journal of Open Source Software*, 3(24), 2018. doi:10.21105/joss.00638.

[218] C. Lian, S. Ruan, T. Denoeux, F. Jardin, and P. Vera. Selecting radiomic features from fdg-pet images for cancer treatment outcome prediction. *Med Image Anal*, 32:257–68, 2016. URL: https://www.ncbi.nlm.nih.gov/pubmed/27236221, doi:10.1016/j.media.2016.05.007.

[219] W. Choi, C. J. Liu, S. R. Alam, J. H. Oh, R. Vaghjiani, J. Humm, W. Weber, P. S. Adusumilli, J. O. Deasy, and W. Lu. Preoperative (18)f-fdg pet/ct and ct radiomics for identifying aggressive histopathological subtypes in early stage lung adenocarcinoma. *Comput Struct Biotechnol J*, 21:5601–5608, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/38034400, doi:10.1016/j.csbj.2023.11.008.

[220] A. Demircioglu. The effect of data resampling methods in radiomics. *Sci Rep*, 14(1):2858, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38310165, doi:10.1038/s41598-024-53491-5.

[221] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi:10.1613/jair.953.

[222] C. Li, Z. He, F. Lv, Y. Liu, Y. Hu, J. Zhang, H. Liu, S. Ma, and Z. Xiao. An interpretable mri-based radiomics model predicting the prognosis of high-intensity focused ultrasound ablation of uterine fibroids. *Insights Imaging*, 14(1):129, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/37466728, doi:10.1186/s13244-023-01445-2.

[223] G. Lemaître, F. Nogueira, and C. K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18, 2017. doi:10.48550/arXiv.1609.06570.

[224] C. Tianqi and G. Carlos. *XGBoost: A Scalable Tree Boosting System*, pages 785–794. ACM, San Francisco, California, USA, 2016. doi:10.1145/2939672.2939785.

[225] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: a highly efficient gradient boosting decision tree, 2017. doi:10.5555/3294996.3295074.

[226] Sercan O. Arik Pfister and Tomas. Tabnet: Attentive interpretable tabular learning. *arXiv*, 2020. URL: https://arxiv.org/abs/1908.07442, doi:10.48550/arXiv.1908.07442.

[227] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. URL: https://arxiv.org/abs/1608.03983, doi:10.48550/arXiv.1608.03983.

[228] J.R. Bohn. Rptk (radiomics processing toolkit) – the program v. 1.0, 2025. URL: https://zenodo.org/records/17431545, doi:10.5281/zenodo.17431544.

[229] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–45, 1988. URL: https://www.ncbi.nlm.nih.gov/pubmed/3203132.

[230] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. C. Sanchez, and M. Muller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011. URL: https://www.ncbi.nlm.nih.gov/pubmed/21414208, doi:10.1186/1471-2105-12-77.

[231] E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958. doi:10.1080/01621459.1958.10501452.

[232] Richard Peto and Julian Peto. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society. Series A (General)*, 135(2), 1972. doi:10.2307/2344317.

[233] David P. Harrington and Thomas R. Fleming. A class of rank test procedures for censored survival data. *Biometrika*, 69(3):553–566, 1982. doi:10.1093/biomet/69.3.553.

[234] H. Wang, W. Chen, S. Jiang, T. Li, F. Chen, J. Lei, R. Li, L. Xi, and S. Guo. Intra- and peritumoral radiomics features based on multicenter automatic breast volume scanner for noninvasive and preoperative prediction of her2 status in breast cancer: a model ensemble research. *Sci Rep*, 14(1):5020, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38424285, doi:10.1038/s41598-024-55838-4.

[235] M. D. Holbrook, S. J. Blocker, Y. M. Mowery, A. Badea, Y. Qi, E. S. Xu, D. G. Kirsch, G. A. Johnson, and C. T. Badea. Mri-based deep learning segmentation and radiomics of sarcoma in mice. *Tomography*, 6(1):23–33, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32280747https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7138523/pdf/tom23.pdf, doi:10.18383/j.tom.2019.00021.

[236] A. Hassouneh, B. Bazuin, A. Danna-Dos-Santos, I. Acar, I. Abdel-Qader, and Initiative Alzheimer's Disease Neuroimaging. Feature importance analysis and machine learning for alzheimer's disease early detection: Feature fusion of the hippocampus, entorhinal cortex, and standardized uptake value ratio. *Digit Biomark*, 8(1):59–74, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38650695, doi:10.1159/000538486.

[237] S. K. Corbacioglu and G. Aksel. Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turk J Emerg Med*, 23(4):195–198, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/38024184, doi:10.4103/tjem.tjem_182_23.

[238] S. Chen and S. Wu. Identifying lung cancer risk factors in the elderly using deep neural networks: Quantitative analysis of web-based survey data. *J Med Internet Res*, 22(3):e17695, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32181751, doi:10.2196/17695.

[239] G. Lee, S. H. Bak, H. Y. Lee, J. Y. Choi, H. Park, S. H. Lee, Y. Ohno, M. Nishino, E. J. R. van Beek, and K. S. Lee. Measurement variability in treatment response determination for non-small cell lung cancer: Improvements using radiomics. *J Thorac Imaging*, 34(2):103–115, 2019. URL: https://www.ncbi.nlm.nih.gov/pubmed/30664063, doi:10.1097/RTI.0000000000000390.

[240] Y. Gao, A. Kalbasi, W. Hsu, D. Ruan, J. Fu, J. Shao, M. Cao, C. Wang, F. C. Eilber, N. Bernthal, S. Bukata, S. M. Dry, S. D. Nelson, M. Kamrava, J. Lewis, D. A. Low, M. Steinberg, P. Hu, and Y. Yang. Treatment effect prediction for sarcoma patients treated with preoperative radiotherapy using radiomics features from longitudinal diffusion-weighted mris. *Phys Med Biol*, 65(17):175006, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/32554891, doi:10.1088/1361-6560/ab9e58.

[241] H. Wang, R. Yang, K. Zhou, S. Wang, C. Cheng, D. Liu, and W. Li. Association between pretreatment c-reactive protein level and survival in non-small cell lung

cancer patients treated with immune checkpoint inhibitors: A meta-analysis. *Int Immunopharmacol*, 124(Pt B):110937, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/37757636, doi:10.1016/j.intimp.2023.110937.

[242] Y. Uehara, T. Hakozaki, R. Kitadai, K. Narita, K. Watanabe, K. Hashimoto, S. Kawai, M. Yomota, and Y. Hosomi. Association between the baseline tumor size and outcomes of patients with non-small cell lung cancer treated with first-line immune checkpoint inhibitor monotherapy or in combination with chemotherapy. *Transl Lung Cancer Res*, 11(2):135–149, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35280320, doi:10.21037/tlcr-21-815.

[243] X. Han, Y. Wang, X. Jia, Y. Zheng, C. Ding, X. Zhang, K. Zhang, Y. Cao, Y. Li, L. Xia, C. Zheng, J. Huang, and H. Shi. Predictive value of delta-radiomic features for prognosis of advanced non-small cell lung cancer patients undergoing immune checkpoint inhibitor therapy. *Transl Lung Cancer Res*, 13(6):1247–1263, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38973966, doi:10.21037/tlcr-24-7.

[244] J. Gong, X. Bao, T. Wang, J. Liu, W. Peng, J. Shi, F. Wu, and Y. Gu. A short-term follow-up ct based radiomics approach to predict response to immunotherapy in advanced non-small-cell lung cancer. *Oncoimmunology*, 11(1):2028962, 2022. URL: https://www.ncbi.nlm.nih.gov/pubmed/35096486, doi:10.1080/2162402X.2022.2028962.

[245] F. Cousin, T. Louis, S. Dheur, F. Aboubakar, B. Ghaye, M. Occhipinti, W. Vos, F. Bottari, A. Paulus, A. Sibille, F. Vaillant, B. Duysinx, J. Guiot, and R. Hustinx. Radiomics and delta-radiomics signatures to predict response and survival in patients with non-small-cell lung cancer treated with immune checkpoint inhibitors. *Cancers (Basel)*, 15(7), 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/37046629, doi:10.3390/cancers15071968.

[246] E. R. Fearon and B. Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–67, 1990. URL: https://www.ncbi.nlm.nih.gov/pubmed/2188735, doi:10.1016/0092-8674(90)90186-i.

[247] Z. Yu, B. Li, S. Zhao, J. Du, Y. Zhang, X. Liu, Q. Guo, H. Zhou, and M. He. Uptake and detection rate of colorectal cancer screening with colonoscopy in china: A population-based, prospective cohort study. *Int J Nurs Stud*, 153:104728, 2024. URL: https://www.ncbi.nlm.nih.gov/pubmed/38461798, doi:10.1016/j.ijnurstu.2024.104728.

[248] S. Kim, J. H. Jung, K. Han, S. J. Koh, J. P. Im, B. G. Kim, J. S. Kim, and H. J. Lee. Association between nonalcoholic fatty liver disease and risk of early-onset colorectal cancer. *Clin Gastroenterol Hepatol*, 2025. URL: https://www.ncbi.nlm.nih.gov/pubmed/40349893, doi:10.1016/j.cgh.2025.02.020.

[249] A. I. Valderrama-Trevino, B. Barrera-Mera, J. C. Ceballos-Villalva, and E. E. Montalvo-Jave. Hepatic metastasis from colorectal cancer. *Euroasian J Hepatogastroenterol*, 7(2):166–175, 2017. URL: https://www.ncbi.nlm.nih.gov/pubmed/29201802, doi:10.5005/jp-journals-10018-1241.

[250] Y. Feng, J. Gong, Y. Wang, Y. Cui, and T. Tong. Explainable multi-modal radiomics for early prediction of liver metastasis in rectal cancer: a multicentric study. *Insights Imaging*, 16(1):142, 2025. URL: https://www.ncbi.nlm.nih.gov/pubmed/40579653, doi:10.1186/s13244-025-02010-9.

[251] J. Yan and Q. Zhou. Lncrna foxp4-as1 silencing inhibits metastasis and epithelial-mesenchymal transition in nasopharyngeal carcinoma via mir-136-5p/mapk1. *Anticancer Drugs*, 34(10):1104–1111, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/36961080, doi:10.1097/CAD.0000000000001510.

[252] R. Inchingolo, C. Maino, R. Cannella, F. Vernuccio, F. Cortese, M. Dezio, A. R. Pisani, T. Giandola, M. Gatti, V. Giannini, D. Ippolito, and R. Faletti. Radiomics in colorectal cancer patients. *World J Gastroenterol*, 29(19):2888–2904, 2023. URL: https://www.ncbi.nlm.nih.gov/pubmed/37274803, doi:10.3748/wjg.v29.i19.2888.

[253] J. Scherer, M. Nolden, J. Kleesiek, J. Metzger, K. Kades, V. Schneider, M. Bach, O. Sedlaczek, A. M. Bucher, T. J. Vogl, F. Grunwald, J. P. Kuhn, R. T. Hoffmann, J. Kotzerke, O. Bethge, L. Schimmoller, G. Antoch, H. W. Muller, A. Daul, K. Nikolaou, C. la Fougere, W. G. Kunz, M. Ingrisch, B. Schachtner, J. Ricke, P. Bartenstein, F. Nensa, A. Radbruch, L. Umutlu, M. Forsting, R. Seifert, K. Herrmann, P. Mayer, H. U. Kauczor, T. Penzkofer, B. Hamm, W. Brenner, R. Kloeckner, C. Duber, M. Schreckenberger, R. Braren, G. Kaissis, M. Makowski, M. Eiber, A. Gafita, R. Trager, W. A. Weber, J. Neubauer, M. Reisert, M. Bock, F. Bamberg, J. Hennig, P. T. Meyer, J. Ruf, U. Haberkorn, S. O. Schoenberg, T. Kuder, P. Neher, R. Floca, H. P. Schlemmer, and K. Maier-Hein. Joint imaging platform for federated clinical data analytics. *JCO Clin Cancer Inform*, 4:1027–1038, 2020. URL: https://www.ncbi.nlm.nih.gov/pubmed/33166197, doi:10.1200/CCI.20.00045.

[254] K. Kades, J. Scherer, M. Zenk, M. Kempf, and K. Maier-Hein. Towards real-world federated learning in medical image analysis using kaapana. *Lecture Notes in Computer Science*, 13573, 2022. doi:10.1007/978-3-031-18523-6_13.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to Prof. Dr. Klaus Maier-Hein for his continuous support, patience, and invaluable feedback, both as a mentor and as a researcher. Without his guidance, this thesis and the projects it builds upon would not have been possible. His openness to new ideas, his encouragement, and his efforts to foster collaborations across the DKFZ have greatly shaped my scientific and personal growth throughout this journey.

I would also like to sincerely thank Prof. Dr. Benedikt Brors for his valuable feedback on the projects and analyses presented in this work, as well as for his guidance throughout the submission process. His thoughtful recommendations and scientific advice have been of great importance, helping me to frame my research according to high standards of scientific practice.

A very big thanks goes to Dr. Tobias Norajitra for his continuous support during my PhD, not only as a group leader but also as a friend. His critical questions, constructive feedback, and willingness to help in all situations have been essential to improving my research and making the past years an enjoyable and rewarding experience. I am truly grateful for his guidance and encouragement.

I would also like to thank the group 3 for the nice group meetings, the interesting exchange of information and the honest feedback on my reports as well as for all the good news you shared with me. These meetings were not only scientifically interesting but also in a friendly environment within a great team spirit. I am very glad to know you and I hope we will stay in contact.

I would also like to express my sincere appreciation to Prof. Dr. Frank Ückert for giving me the opportunity to pursue my research interests independently and for welcoming me into his group. His trust, support, and openness allowed me to explore my ideas freely and grow as a researcher. I am looking forward to continuing our collaborations and shared projects in Hamburg.

My thanks also go to Dr. Fabian Isensee for his technical support with the GPU cluster and GitHub repositories, as well as for his honest and practical advice on technical solutions. I would also like to thank Dr. Ralf Floca for including me in

collaborations and giving me the chance to grow further within the radiomics community. Thank you for your valuable advice on the dos and don'ts of research, for being open to my ideas, and for your support in analysis and contributions to the quality of my work. I would further like to thank Amine Yamlahi for the great time we shared in the office and for our many discussions that helped solve deep learning and technical challenges, it was a real pleasure to have you as a desk neighbor.

I am very grateful to my collaboration partners Prof. Dr. Claus Peter Heußel, PD. Dr. Petros Christopoulos, Dr. Alan Arthur Peters, Dr. Stephan Rheinheimer, and Dr. Oyunbileg Stackelberg for their kind support in collecting and annotating the data for immunotherapy response prediction, and for their valuable scientific input and feedback. I am looking forward to continuing our fruitful collaborations in the future. Special thanks also go to Dr. Anna Hinterberger and Dr. Freba Grawe for contributing your clinical expertise and helping to realize projects with a direct clinical application. I am excited to continue our joint efforts to translate research into patient benefit.

I would like to thank all my colleagues from E240 for including me in their social activities and for brightening many days with good humor and camaraderie. My colleagues from E230 deserve special thanks for the wonderful time we shared at retreats, conferences, and celebrations, and for maintaining such an inspiring and positive team spirit.

To my office mates from E220, thank you for the great moments, for our little traditions, and, of course, for the delicious cakes. I will miss our everyday conversations and your warm integration into the group. Many thanks also to Dr. Ina Kurth for her help in improving the office situation amidst all the construction work, and to Dr. Mareike Roscher and Dr. Olga Ximena Giraldo Pasmin for your kind support, patience, and encouragement during my thesis writing period, you truly helped me through the challenging times.

I would also like to thank my friends I met during my time at the DKFZ in Heidelberg, especially Yassin Harrim, Sandro Hoffmann, Nicola Biondi, Christina Blume, Giulia Di Muzio, Paul Schwerd-Kleine, and Daniela Kocher for all the wonderful moments, and celebrations we shared. I am very glad to know you all and to have spent so many memorable times together.

My heartfelt gratitude goes to my family, especially my mother and my brother, for their unconditional support throughout my PhD journey. You have always helped me to focus on what truly matters and encouraged me when things became difficult. You are an essential part of my life, and this achievement would not have been possible without your support.

The biggest thanks go to my partner, Vivien Ionasz, for being my greatest source of emotional support and stability throughout these years in Heidelberg. I cannot express how much your love, patience, and encouragement have meant to me. Thank you for standing by my side, for believing in me, and for sharing both the everyday moments and the most challenging times. I am deeply grateful that our paths crossed in Heidelberg, and I look forward to starting our next chapter together in Hamburg. You make my life complete and fill it with joy, love, and purpose.

I dedicate this thesis to my deceased father, whose memory continues to inspire and guide me every day.

Disclaimer

The text of this dissertation is original and was entirely written by Jonas Bohn.
It has been proofread and edited using ChatGPT.
This thesis was written and typeset in LaTeX.