Computer Assisted Diagnosis in PET/CT

Machine Learning for Prognosis in Oncological Patients

DISSERTATION

zur Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Mohammadsobhan MOAZEMI GOODARZI

aus Teheran, Iran

Bonn, März 2022

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Thomas SCHULTZ
2. Gutachter: Prof. Dr. Dr. Ralph A. BUNDSCHUH

Tag der mündlichen Prüfung: 24. März 2022
Erscheinungsjahr: 2022

Abstract

Mohammadsobhan MOAZEMI GOODARZI

Computer Assisted Diagnosis in PET/CT
Machine Learning for Prognosis in Oncological Patients

Artificial intelligence (AI) has revolutionised problem solving in a wide range of industrial as well as research domains. Particularly, computer-aided diagnosis (CAD) and clinical decision support systems (CDSSs) as sub-domains of AI, have gained critical importance in many biomedical and clinical domains such as virology, computational neuroscience, and oncology. As making accurate decisions in a timely manner is an inevitable part of daily routines in the medical and clinical domains, machine learning (ML) and deep learning methods are widely applied in CAD and CDSSs to provide diagnostic and prognostic assistance for the researchers and physicians as the domain experts.

Focusing on advanced prostate cancer (PCa) disease as an example, the procedure of disease staging and patient screening using established CAD tools is considered time consuming and attention intensive. In many clinical practices, this procedure includes examining patients' prostate-specific membrane antigen-positron emission tomography/computed tomography (PSMA-PET/CT) scans and analyzing patient-specific clinical factors in a daily routine. Thus, as the main motivation behind this PhD thesis project, AI and ML based methods are utilized to automate the corresponding diagnostic and prognostic pipelines.

Accordingly, providing an automated CDSS which facilitates: 1) visualization and annotation of medical scans, 2) automated segmentation of pathological uptake, 3) prediction of treatment outcome taking advantage of radiomics features extracted from Gallium[68]-(^{68}Ga)-PSMA-PET/CT scans in PCa patients was the main objective of this thesis.

To this end, we introduce AutoPyPetCt, an automated pipeline developed in Python which takes multimodal whole-body baseline ^{68}Ga-PSMA-PET/CT scans and patient-specific clinical parameters as input and applies state-of-the-art statistical, ML, and deep learning techniques to automatically identify and segment pathological uptake all over the body, to anticipate responders to Lutetium[177]-(^{177}Lu)-PSMA therapy, and to predict overall survival of the PCa patients.

To achieve this, on the one hand, multimodal PET/CT scans integrate functional as well as anatomical aids to locate malignancies as volumes and regions of interest (VoIs and and RoIs respectively). On the other hand, a variety of conventional parameters (such as standardized uptake value (SUV)) as well as radiomics features (such as textural heterogeneity features) extracted for the VoIs/RoIs together with patient-specific clinical factors (such as age and prostate-specific antigen (PSA) level) form the basis for statistical and ML-based analyses towards prognostic hypotheses realizing the prediction of patient level outcomes such as treatment response and overall survival.

The main contribution of the methods is to provide automated decision support tools to manage patients with advanced PCa in shorter times and with limited annotation effort. To investigate the relevance and to quantify the performance of the methods, multiple retrospective quantitative as well as qualitative clinical studies have been conducted which resulted in several preliminary conference abstracts, four journal papers, and one conference paper. The studies had been carried out along the whole project's life-cycle, starting by a proof of concept and finalizing with the evaluations of the integrated solution pipeline.

The findings from the clinical studies confirmed the overall relevance of the methods and their potential to replace parts of current clinical routine procedures in the future. Most interestingly, the provided automated segmentation tools achieved high performance in true delineation of pathological uptake which outperformed a standard established thresholding based approach. However, the results of the treatment response prediction studies, regardless of different segmentation methods, identified rooms for improvement.

To conclude, the provided automated decision support system has shown its potential to serve as an assistant for the management of patients diagnosed with advanced prostate cancer disease. However, to further assess the generalizability of the findings and to improve the decision making certainty, studies including multicentric data should be considered as future work.

Keywords: Computer-Aided Diagnosis (CAD), Clinical Decision Support System (CDSS), Machine Learning (ML), Deep Learning (DL), Prostate Cancer (PCa)

Zusammenfassung

Mohammadsobhan MOAZEMI GOODARZI

Computer Assisted Diagnosis in PET/CT
Machine Learning for Prognosis in Oncological Patients

Computergestützte Diagnose in PET/CT
maschinelles Lernen für die Prognose bei onkologischen Patienten

Künstliche Intelligenz (KI) hat die Problemlösung in einer Vielzahl von Industrie- und Forschungsbereichen revolutioniert. Insbesondere computergestützte Diagnose (CAD) und klinische Entscheidungsunterstützungssysteme (CDSSs) als Unterbereiche der KI haben in vielen biomedizinischen und klinischen Bereichen wie Virologie, Computational Neuroscience und Onkologie eine entscheidende Bedeutung erlangt. Da das zeitnahe Treffen genauer Entscheidungen ein unvermeidlicher Bestandteil der täglichen Routine im medizinischen und klinischen Bereich ist, werden maschinelles Lernen (ML) und Deep-Learning-Methoden in CAD und CDSSs weit verbreitet eingesetzt, um Forschern und Ärzten (als den Domänenexperten) diagnostische und prognostische Unterstützung zu bieten.

Am Beispiel des fortgeschrittenen Prostatakarzinoms (PCa) wird das Verfahren des Krankheits-Staging und des Patienten-Screenings mit etablierten CAD-Tools als zeit- und aufmerksamkeitsintensiv angesehen. In vielen klinischen Praxen umfasst dieses Verfahren die Untersuchung der Prostata-spezifischen Membranantigen-Positronen-Emissions-Tomographie/Computertomographie (PSMA-PET/CT) von Patienten und die Analyse patientenspezifischer klinischer Faktoren in einer täglichen Routine. Als Hauptmotivation für dieses Dissertationsprojekt werden daher KI- und ML-basierte Methoden verwendet, um die entsprechenden diagnostischen und prognostischen Pipelines zu automatisieren.

Diese Arbeit stellt ein automatisiertes CDSS bereit, das folgende Funktionen bietet: 1) Visualisierung und Annotation medizinischer Scans, 2) automatisierte Segmentierung des pathologischen Uptakes, 3) Vorhersage des Behandlungsergebnisses der PCa Patienten unter Nutzung der aus Gallium[68]-(^{68}Ga)-PSMA-PET/CT-Scans extrahierten Radiomics Features.

Zu diesem Zweck stellen wir AutoPyPetCt vor, eine in Python entwickelte automatisierte Pipeline, die multimodale Ganzkörper-Baseline-^{68}Ga-PSMA-PET/CT Scans und patientenspezifische klinische Parameter als Input verwendet. Sie nutzt statistische, ML- und Deep-Learning-Techniken, um pathologische Uptakes im ganzen Körper automatisch zu identifizieren und zu segmentieren, um ein Ansprechen auf die Lutetium[177]-(^{177}Lu)-PSMA-Therapie zu antizipieren und das Gesamtüberleben des PCa Patienten vorherzusagen.

Um dies zu erreichen integrieren zunächst multimodale PET/CT-Scans funktionelle sowie anatomische Daten, um pathologische Veränderungen als Volumes und Regions of Interest (VoIs bzw. RoIs) zu lokalisieren. Diese werden mit einer Vielzahl konventioneller Parameter (wie standardized uptake value (SUV)) sowie radiomics

features (wie texturale Heterogenitätsmerkmale) quantifiziert. Zusammen mit patientenspezifischen klinischen Faktoren (wie Alter und Prostata-spezifisches Antigen (PSA)) dienen sie als Grundlage für statistische und ML-basierte Analysen, die eine Vorhersage von Ergebnissen auf Patientenebene wie Wirksamkeit bestimmter Behandlungen und Gesamtüberleben ermöglichen.

Der Hauptbeitrag der Methoden besteht darin, automatisierte Tools zur Entscheidungsunterstützung bereitzustellen, um Patienten mit fortgeschrittenem PCa in kürzerer Zeit und mit begrenztem Annotationsaufwand zu behandeln. Um die Relevanz zu untersuchen und die Leistungsfähigkeit der Methoden zu quantifizieren, wurden mehrere retrospektive quantitative sowie qualitative klinische Studien durchgeführt, die zu mehreren vorläufigen Konferenzabstracts, vier Zeitschriftenbeiträgen und einem Konferenzbeitrag führten. Die Studien wurden entlang des gesamten Projektlebenszyklus durchgeführt, beginnend mit einem Proof of Concept und abschließend mit den Bewertungen der integrierten Lösungspipeline.

Die Erkenntnisse aus den klinischen Studien bestätigten die Gesamtrelevanz der Methoden und ihr Potenzial, zukünftig Teile der aktuellen klinischen Routineverfahren zu ersetzen. Interessanterweise erreichten die bereitgestellten automatisierten Segmentierungstools eine hohe Leistung bei der richtigen Abgrenzung der pathologischen Aufnahme, die einen etablierten, auf Schwellenwerten basierenden Ansatz übertraf. Die Ergebnisse der Studien zur Vorhersage des Behandlungswirksamkeit zeigten jedoch, unabhängig von unterschiedlichen Segmentierungsmethoden, Raum für weitere Verbesserungen.

Zusammenfassend lässt sich sagen, dass das bereitgestellte automatisierte Entscheidungsunterstützungssystem sein Potenzial gezeigt hat, als Assistent für die Behandlung von Patienten mit diagnostizierter fortgeschrittener Prostatakrebserkrankung zu dienen. Um jedoch die Generalisierbarkeit der Ergebnisse weiter zu bewerten und die Entscheidungssicherheit zu verbessern, sollten künftig weitere Studien durchgeführt werden, die insbesondere auch multizentrischen Daten berücksichtigen.

Acknowledgements

In the first place, I would sincerely like to pay tribute to my supervisors Prof. Dr. Dr. Ralph A. Bundschuh and Prof. Dr. Thomas Schultz for their invaluable advice, encouragement, and support during my PhD studies. I should declare that without their gracious mentorship and supervision, the preparation of this dissertation and conducting the corresponding research would have not succeeded. Specially, I appreciate that I was given the opportunity to conduct interdisciplinary research guided by the expertise and qualifications of my supervisors. Moreover, I would be delighted to acknowledge that the Department of Nuclear Medicine at the University Hospital Bonn has provided the funding, resources, and daily supervision for this work. In addition, this work was co-supervised by the Visualization and Medical Image Analysis Group at the Department of Computer Science at the University of Bonn. Also, alongside my PhD supervisors, I would like to express my sincere gratitude to Prof. Dr. Christian Bauckhage and Prof. Dr. rer. nat. Matthias Schmid for their supports as the members of my doctoral examination committee.

Furthermore, I would like to thank my lovely wife, my parents, my sister, and my in-laws for their spiritual support throughout the entire PhD program as well as during writing this thesis. I should sincerely acknowledge that nothing would replace their kindness, sympathy, and support.

Last but not least, I would like to thank the fellow researchers of the Visualization and Medical Image Analysis Group at the Department of Computer Science at the University of Bonn as well as the wonderful staff, collaborators, and supervisors from the Departments of Nuclear Medicine and Neurology at the University Hospital Bonn, in particular, Prof. Dr. med. Markus Essler and Dr. Xenia Kobeleva, for all their collaboration, supervision, and support during my PhD and related side projects.

Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
ANN	**A**rtificial **N**eural **N**etwork
AUC	**A**rea **U**nder the **C**urve
CAD	**C**omputer **A**ided **D**iagnosis
CDSS	**C**linical **D**ecision **S**upport **S**ystem
CNN	**C**onvolutional **N**eural **N**etwork
DNN	**D**eep **N**eural **N**etwork
CT	**C**omputed **T**omography
DL	**D**eep **L**earning
ET	**E**xtra **T**rees
FDG	**F**luoro**D**eoxy**G**lucos
Ga	**Ga**llium
GPU	**G**raphical **P**rocessing **U**nit
Lu	**Lu**tetium
ML	**M**achine **L**earning
MRI	**M**agnetic **R**esonance **I**maging
PCa	**P**rostate **Ca**ncer
PET	**P**ositron **E**mission **T**omography
PSA	**P**rostate **S**pecific **A**ntigen
PSMA	**P**rostate **S**pecific **M**embrane **A**ntigen
RAF	**RA**ndom **F**orest
RBF	**R**adial **B**asis **F**unction
RFE	**R**ecursive **F**eature **E**limination
RFs	**R**adiomics **F**eature**s**
ROC	**R**eceiver **O**perating **C**haracteristic
RS	**R**adiomics **S**ignature
SE	**SE**nsitivity
SP	**SP**ecificity
STD	**ST**andard **D**eviation
SUV	**S**tandardized **U**ptake **V**alue
SVC	**S**upport **V**ector **C**lassifier
SVM	**S**upport **V**ector **M**achine

Chapter 1

Introduction

1.1 Motivation

Along the lines of reshaping many research domains and industries, also the medical practice and clinical research domains have been influenced by the application of Artificial Intelligence (AI) and machine learning (ML). Rather than simplifying time consuming and complicated tasks taking advantage of automatization, AI in general and ML methods in particular provide problem solving approaches which mimic human way of thinking without suffering from experience inconsistency and subjectivity.

According to the fact and figures published by different health organizations, prostate cancer (PCa) is and continues to be a common malignancy which ranks amongst top causes of men's death worldwide [44, 147], imposing increasing demands in terms of costs and resources to healthcare and insurance systems. As a result, there is an increasing need for AI and ML based clinical support solutions to address this still rising issue.

Following the pattern of automatization, computer-aided diagnosis (CAD) has been used extensively to assist physicians and clinical researchers in a variety of fields including oncology and computational neuroscience. In the special case of cancer analysis, CAD-based medical image analysis is commonly applied as it facilitates non-invasive tissue assessment without requiring any biopsies. This becomes even more critical when subjects suffer from numerous metastases spread all over the body. However, the procedure of identification and delineation of the regions or volumes of interest (RoIs or VoIs) using established CAD tools is considered time consuming and attention intensive.

Moreover, inter-observer variability caused by subjective opinions or different experience levels may arise inconsistencies in CAD-based diagnosis and prognosis. Another limitation of uni-modal medical imaging techniques is that they often provide a limited spectrum of characterization to the RoIs or VoIs as they either lack spatial information or functional features. For example, PET imaging provides information about functional uptakes without much spatial context, whereas CT imaging provides high resolution spatial context without providing functional and metabolic characteristics. In contrast to uni-modal scans, multimodal imaging plays an important role as it provides a broader contextual field to locate and characterize malignancies compared to when single mode scans are used.

In this PhD thesis we aim at facilitating an automated pipeline to manage PCa patients which tackles some limitations to the existing CAD-based routines as mentioned above. To assess the disease stage as well as to monitor the progress of the disease, PET/CT scans are commonly used. PET/CTs are multimodal medical imaging techniques which are widely used for different cancer diseases. Multimodal imaging serves nuclear medicine (NM) physicians, oncologists, and neuroscientists for

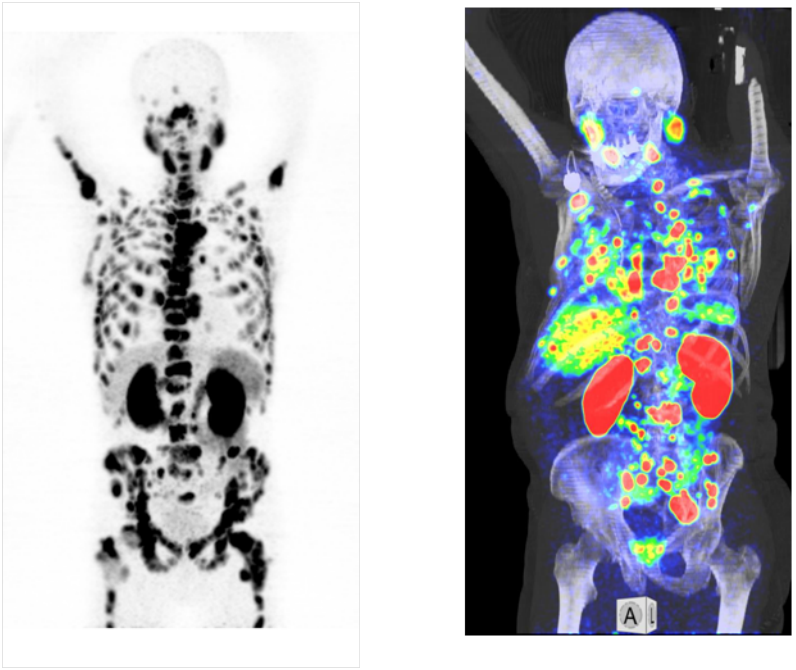


FIGURE 1.1: An example of multimodal imaging for prostate cancer management. Left: the positron emission tomography (PET), right: the overlaid PET/computed tomography (PET/CT). The red uptake in the right panel includes both pathological and physiological uptake. This figure was originally published in our previous work [106].

managing disorders such as advanced prostate carcinoma, lung cancer, and neurodegenerative diseases [21, 76, 107, 139, 142]. A wide variety of imaging modalities including PET, CT, and magnetic resonance imaging (MRI) are used as the essential means for diagnosis and therapy outcome assessment. For the examination of cancer patients, PET imaging is widely used to investigate the degree of metabolic uptake both in early stages of the disease and in metastatic tissues. The anatomical information on the other hand often comes from, e.g., CT or MRI scans. While the sensitivity of PET images is usually higher than that of CT or MRI, lack of anatomical information often disqualifies alone PET images for certain diagnostic purposes [46]. Figure 1.1 shows an example of multimodal PET/CT images.

As shown in Figure 1.1, apart from the primary uptake in the prostate itself and the metastatic uptake spread in different organs (most commonly in bone and lymph nodes), some organs such as the liver, kidneys, and glands feature high uptakes in PET/CT imaging. Therefore, true discrimination of pathological (i. e., malignant) from physiological (i. e., normal) uptake in the absence of domain expert annotation is considered a challenging task for automated CAD-based solutions [42, 107].

Even if taken with the same machine, PET and CT often are attributed with different resolutions and coordination systems. Thus, another important aspect of multimodal medical imaging is the resampling and co-registration of original modalities. If the two modalities only differ in scale or rotation, appropriate affine transformations could bring one modality to the common space as for the other modality. However, often more sophisticated mappings such as deformations and body motions due to respiratory movements should be taken care of [2, 71].

The next critical aspect of CAD systems which assist oncologists is the segmentation. There are many tumor segmentation tools available which can be categorized

from totally manual to fully automated, with regards to the amount of the user interaction they need. For instance, InterView FUSION [98] is commonly used in clinical routines as a standard tool for manual segmentation of tumors in whole-body PET/CT scans and FreeSurfer [45] facilitates automated segmentation of brain MRI scans. Furthermore, thresholding techniques are widely applied to identify pathological uptake in PET scans [73, 168] and to segment different body organs based on Hounsfield scale [40] attributed to CT scans [113, 136].

Most of the manual and semi-automated segmentation techniques require intensive and time-consuming human interaction. Therefore, providing fast and accurate automated segmentation tools is one of the most important objectives of CAD and CDSSs. As ML based approaches, artificial neural networks (ANNs) and deep learning architectures have been used for many years for automated image segmentation and classification. In recent years, ANNs have become more popular in medical image analysis and segmentation as well. In particular, the U-Net model [132] is widely used for segmentation purposes in many tools and studies facilitating diagnosis and treatment outcome assessment pipelines [47, 55]. As part of the developed automated pipeline (AutoPyPetCt), we trained and fit a multi-channel U-Net based model, namely PET-CT-U-Net, for the segmentation of pathological uptake in PSMA-PET/CT scans. Results of a retrospective study [106] that revealed the quantitative performance of the model in terms of precision, recall and Dice coefficient will be presented in the next chapters. Furthermore, qualitative assessments of predicted labels from the developed PET-CT-U-Net are provided by a highly experienced domain expert.

Conventionally, parameters such as standardized uptake value (SUV), metabolic tumor volume (MTV), and total lesion glycolysis (TLG) had been in focus for characterization of pathological uptake in PET scans. Taking advantage of linear regression methods with limited number of independent variables, many studies analyzed these conventional metrics or other metrics such as textural heterogeneity parameters (e. g., homogeneity, entropy, and kurtosis) individually to classify pathological uptake, to predict response to treatment, or to analyze overall survival [73, 78, 97, 120]. More recently, as state-of-the-art ML based methods became more applicable in clinical research, analyses of combinations of such conventional variables with the so-called radiomics features and patient-specific clinical parameters gets more publicity in clinical studies [14, 21].

Radiomics denotes the procedure of extracting numerical quantities out of medical imaging data in terms of two or three dimensional (2D or 3D) intensity-, shape-, or texture based features which characterize tumors or physiological hotspots. Considering diagnosis, therapy response prediction, and survival analysis as the ultimate goals of clinical decision support tools, supervised ML methods are widely used in combination with radiomics features in clinical research. As part of our methods, we took advantage of supervised ML to classify manually delineated tumors as pathological vs physiological based on radiomics features calculated by InterView Fusion with nuclear medicine expert accuracy [42, 107]. In another study [103], we further showed the potential of ML classifiers as applied to radiomics features from manually segmented tumors and patient-specific clinical data to predict responders to ^{177}Lu-PSMA treatment. Recently, we analyzed the potential of radiomics signature calculated by the least absolute shrinkage and selection operator (LASSO) [81] method from Cox proportional hazard model as multivariate survival analysis method [105].

To conduct diagnostic and prognostic analyses at different levels and in different

steps from malignancy detection to treatment outcome prediction, different statistical and ML based techniques should be applied. In general, the utilization of different methods depends on various data- and task oriented factors such as sample size, input modalities, and analysis outcome (e. g., binary classification, segmentation, overall survival prediction etc.). Accordingly, there is an immense set of AI based algorithms and techniques such as support vector machines (SVMs), decision trees, convolutional neural networks (CNNs), and uni- and multivariate survival analysis techniques which can be beneficial for different use-cases. Thus, selecting the most suitable set of methods among the diverse set of available choices has been one of the challenges during this project.

As another technical challenge which needs to be addressed in applied data science is the generalizability, particularly in the absence of sufficiently big and diverse cohorts of data. To account for this, as will be discussed in the Methodology chapter (4), appropriate cross validation as well as splitting the available data cohorts into train and test groups have been included in the methods pipelines. As will be further discussed in 5 and 6, our findings reveals the relevance of the implemented and integrated methods. However, for the task of treatment response prediction using deep learning methods, as they typically need bigger cohorts to converge, our findings identified rooms for improvement which can be facilitated by using alternative CNN models and providing multicentric data.

Also, as findings from related work suggest [6], leveraging deep features can be considered as an alternative approach to radiomics analysis for the assessment of diagnostic and prognostic hypotheses. Deep features refer to numerous hierarchical features extracted from deep neural networks as applied for, e. g., medical image analysis and treatment outcome prediction. This topic is of great importance and should be considered as a possible future work track.

1.2 Contributions

To conclude, the main objective of this thesis is to provide a comprehensive automated pipeline which serves as a CDSS for the management of patients with advanced prostate carcinoma. As will be discussed in the Background (2) and Related Work (3) chapters, there are tons of studies and tools which address parts of such a comprehensive pipeline. Although there have been studies which presented tools which address parts of these building blocks as consecutive steps [51, 134, 152, 166], our research to find comprehensive automated solutions for PCa patients who underwent ^{177}Lu-PSMA therapy resulted in in-concrete findings. Thus, to the best of our knowledge, the lack of such an automated pipeline which addresses the enrollment of ^{68}Ga-PSMA-PET/CT from visualization to treatment outcome prediction is still persistent.

Apart from replacing previous time- and attention intensive manual routines for tumor delineation and disease staging, which itself is an added value to the existing clinical practice, the developed methods combine state-of-the-art statistical and machine learning methods to provide comprehensive analyses on the patients' states in terms of predictions of treatment response and overall survival. Correspondingly, the AutoPyPetCt pipeline takes raw Dicom PET/CT images together with patient-specific clinical parameters as input and utilizes the multi-channel PET-CT-U-Net to identify and segment pathological uptake. Then, for each patient, radiomics features are calculated based on the predicted pathological masks. Finally, supervised

ML classifiers and uni- and multivariate survival estimators are facilitated to come up with the ultimate diagnosis and prognostice outlines.

1.3 Publications and Outline

1.3.1 List of the Publications

To evaluate the methods and to assess the generalizability of the findings, several retrospective clinical studies have been conducted. Based on the clinical studies and findings in accordance with this thesis project, so far, the following manuscripts have been published (note that the shorter version of the author's name, Sobhan Moazemi, was used for all of the publications):

- S. Moazemi, M. Essler, T. Schultz, R. A. Bundschuh. Predicting Treatment Response in Prostate Cancer Patients Based on Multimodal PET/CT For Clinical Decision Support. In: Syeda-Mahmood T. et al. (eds) Multimodal Learning for Clinical Decision Support. ML-CDS 2021. Lecture Notes in Computer Science, vol 13050. Springer, Cham. doi: 10.1007/978-3-030-89847-2_3 [106]
- A. Erle, S. Moazemi, S. Lütje, M. Essler, T. Schultz, R. A. Bundschuh. Evaluating a Machine Learning Tool for the Classification of Pathological Uptake in Whole-Body PSMA-PET-CT Scans. Tomography. 2021; 7(3):301-312. doi: 10.3390/tomography7030027 [42]
- S. Moazemi, A. Erle, Z. Khurshid, S. Lütje, M. Muders, M. Essler, T. Schultz, R. A. Bundschuh. Decision support for treatment with 177Lu-PSMA: machine learning predicts response with high accuracy based on PSMA-PET/CT and clinical parameters. Ann Transl Med 2021. doi: 10.21037/ atm-20-6446 [103]
- S. Moazemi, A. Erle, S. Lütje, M. Essler, R. A. Bundschuh. Estimating the Potential of Radiomics Features and Radiomics Signature from Pretherapeutic PSMA-PET-CT Scans and Clinical Data for Prediction of Overall Survival When Treated with 177Lu-PSMA. Diagnostics 2021, 11, 186. doi: 10.3390/diagnostics11020186 [105]
- S. Moazemi, Z. Khurshid, A. Erle, S. Lütje, M. Essler, T. Schultz, R. A. Bundschuh. Machine Learning Facilitates Hotspot Classification in PSMA-PET/CT with Nuclear Medicine Specialist Accuracy. Diagnostics 2020, 10, 622. doi: 10.3390/diagnostics10090622 [107]

1.3.2 Summary of the Studies

To summarize the methods and findings from the above-mentioned studies, we took advantage of supervised machine learning methods including support vector machines (SVM) [64] and decision trees [129] to classify pathological uptake from pretherapeutnc ^{68}Ga-PSMA-PET/CT scans leveraging RFs from both PET and CT modalities. Our findings illustrated the significant role of combination of PET and CT RFs and the ML methods for the discrimination of pathological from physiological uptake in patients with advanced prostate carcinoma [107]. We measured the performance of our methods in terms of area under the curve (AUC), sensitivity, and specificity. In a next study [42], we evaluated a stable ML based tool based on findings from [107] focusing on the effect of increasing the training cohort size on the performance of the algorithms. We also investigated in more detail in which

body parts it was more challenging for the algorithms to achieve higher specificity. As the prediction of the patients' response to ^{177}Lu-PSMA therapy was another goal of ours, in another study [103], we used difference in prostate specific antigen (PSA) levels in pre- and post-therapeutic scans (called ΔPSA) as treatment response indicator to investigate performance of PSMA-PET/CT RFs and patient-specific clinical parameters for the prediction of the responders to the therapy. To this end, we leveraged linear regression and a similar set of ML classifiers as was used in [107] to train the model and validate our findings. Moreover in a recent study [105], we applied both uni- and multivariate analysis methods such as Cox proportional hazards (CPH) model [34] and Kaplan-Meier (KM) estimator [74] to predict overall survival (OS) of an extended cohort with similar clinical factors as for the previous studies. This study revealed the potential of radiomics signature extracted from pre-therapeutic ^{68}Ga-PSMA-PET/CT, using the least absolute shrinkage and selection operator (LASSO) [81], to predict OS. Finally, we implemented a mutli-channel U-Net based model to automatically segment pathological uptake based on PSMA-PET/CT scans. This was followed by calculating radiomics features for predicted binary masks resulting in an analysis of ML classifiers for predicting treatment response. The results of the last study are recently published in [106].

In the next chapters of this thesis, we present background (2), related literature (3), methods and solutions (4), results and evaluations (5) which mostly address in-house developed software tools as building blocks of AutoPyPetCt which serve as a CDSS for the examination, diagnosis, and treatment response prediction for patients with advanced prostate carcinoma, focusing on findings from pre-therapeutic ^{68}Ga-PSMA-PET/CT scans as well as conventional clinical parameters. In the end, the final chapter (6) presents discussions about the contributions, achievements, and drawbacks of the integrated methods as well as possible future work tracks.

Chapter 2

Background

In this chapter, the basic terminology as well as the fundamental concepts which will be required to understand the methodologies applied in the projects and studies which are realized as part of this thesis are elaborated and discussed briefly. The next chapter, related work (3), will provide the most relevant literature and research material to this thesis with regards to the application domains and implemented methods.

2.1 Medical Imaging Modalities

2.1.1 Positron Emission Tomography (PET)

Positron emission tomography (PET) [9] is a kind of nuclear imaging which is widely used in medical practice, especially in oncology. PET is a functional imaging technique which makes the use of radiotracers to visualize and measure physiological and metabolic activities in various body organs. As different types of tissues react differently to radiotracer intake in terms of absorption and regional concentration, PET imaging can facilitate discrimination of malignant versus healthy tissues. In the 1950s, researchers at the University of Pennsylvania introduced the concept of emission and transmission tomography for the first time. Later on, in 1975, the methods for tomographic imaging were further developed at the Washington University School of Medicine [127, 150].

As described by A.M.J. Paans [118], state-of-the-art PET scanners consist of a radiation detector unit and a computation unit with a relatively high power to perform data acquisition and image reconstruction. The radiation detector is composed of several crystal subdetectors and several photomultiplier tubes (PMTs) which together capture an axial field of view (FOV). Consecutively, the whole-body PET image is then formed as the patient is moved along the scanner containing a source substance (a positron emitter such as ^{68}Ga) in its gantry. The functional differentiation of different tissues is measured in terms of radiation intensity, I:

$$I = I_0 \cdot exp(-\mu.d), \tag{2.1}$$

where I_0 is the radiation intensity before the attenuation, d is the distance at which the object's attenuation appears, and μ is the attenuation coefficient which is defined by the international standard organization (ISO) [119] as:

$$\mu = -\frac{1}{\Phi_e}\frac{d\Phi_e}{dz}, \tag{2.2}$$

where Φ_e is the radiant flux (defined as the radiant energy that is transmitted or received per unit of time) and z is the path length of the beam.

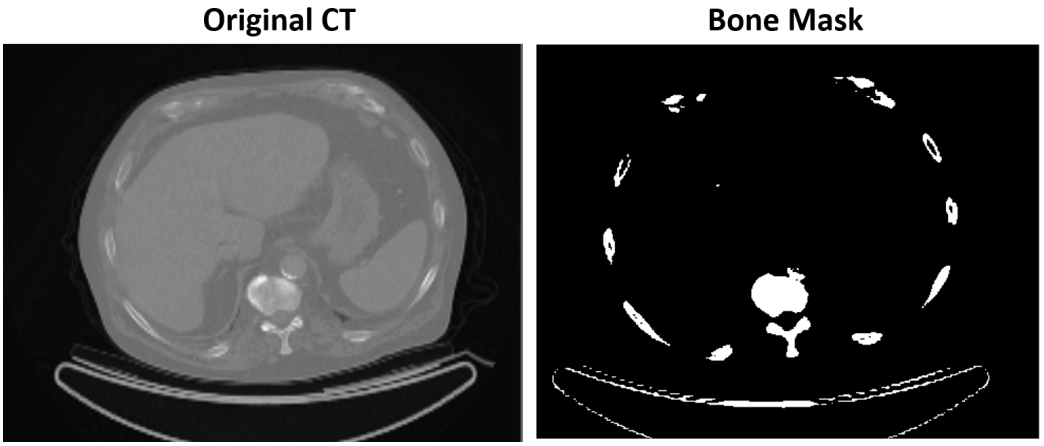


FIGURE 2.1: An example of thresholding based method to segment bone in computed tomography (CT) images using Hounsfield scales. Left: the original CT image. Right: the resulting bone mask.

After applying attenuation- and scatter correction, the activity per pixel is calculated in absolute terms in $[Bq/pixel]$ units. Accordingly, when regions of interest (RoIs) are defined, the concentration of the radioactivity within the RoI (C_{RoI}) in Mega Bq (MBq/ml) at time t is compared to the injected dose (ID) (MBq/g) normalized by the body weight (BW) of the subject to end up with the so-called standardized uptake value (SUV) as defined as:

$$SUV(t) = \frac{C_{RoI}(t)}{ID/BW}, \tag{2.3}$$

and is commonly used in oncological research.

From the application point of view, PET is a well-known medical imaging modality for diagnosis and disease staging in daily clinical routines as well as research projects [36, 125]. PET is often used in combination with anatomical imaging modalities such as computed tomography (CT) or magnetic resonance imaging (MRI).

2.1.2 Computed Tomography (CT)

Computed tomography (CT) is a computerized imaging technique which generates slice-based images using an x-ray tube which rotates around the subject's body, forming 3D volumes as stacked together. Godfrey Hounsfield, who invented the first commercial CT scan machine in 1972 [131] is also known for the Hounsfield scale which is named after him. The Hounsfield scale is a unitless measure for which water and air are arbitrarily set to zero and -1000 units respectively and for the rest of the materials and tissues it is calculated based on the extent of their x-ray absorption [40]. The Hounsfield unit (HU) of a voxel is defined as:

$$HU = 1000 \times \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}}, \tag{2.4}$$

where μ, μ_{water}, and μ_{air} are the attenuation coefficients of the substance (averaged), water, and air respectively.

Compared to PET imaging, CT scans facilitate better anatomical and spatial precision, making localization of RoIs possible. Figure 2.1 presents a sample usage of the Hounsfield scales for bone segmentation in CT scans.

2.1.3 PET/CT

For most of the computer-assisted diagnosis and prognosis tasks in oncology and related fields, examination of both of the functional and anatomical characteristics of the tissues is inevitable, therefore, multimodal imaging has been and continues to be a vital asset in clinical practice and research. As an example of multimodal imaging, PET/CT scans are widely applied [36, 125]. Nevertheless, differences in scan timestamps, quality, resolution, and slice spacing of the joint modalities may arise challenges which need to be addressed by accurate co-registration and resampling of the images [46].

Although CT images of the same patient from other scanners could enhance diagnostic power from a PET only scanner by providing an extra channel to perform co-registration of the two modalities, utilizing CT into the same machine as the PET scanner makes it possible to apply enhanced attenuation correction based on attenuation coefficient (μ) of a reference substance such as bone. The reason behind is that different tissues (e. g., muscles and bones) feature different attenuation coefficients as captured by CT scanners, but when captured by PET scanners the attenuation coefficients of different tissues quantify in similar ranges. Thus, when integrated into the same machine, the attenuation correction can be applied directly as the scan is being performed and the attenuation values are calculated per PET pixels (see equation 2.1) [118].

2.2 Machine Learning Algorithms

Machine learning is a field of artificial intelligence (AI) which takes advantage of statistical analysis methods to teach the computers to model human thinking in order to solve real life problems [101]. The term machine learning (ML) was coined by Arthur Samulen in 1959 and refers to the process of solving problems by computers without being explicitly programmed [80]. Indeed, ML methods apply real world or simulated data to fit and train models which predict the desired outputs. Some common applications of ML are pattern recognition, computer vision, and medicine. Also, ML can be used in a variety of prediction tasks, including classification, regression, object detection, and image processing.

In general, ML methods are subdivided into supervised and unsupervised categories, depending on the presence of true annotations for the training datasets, also known as ground truth (GT) labels. In supervised learning, the machine tries to learn from the training data by fitting a model which can map the input data to the pre-existing ground truth labels. In contrast, unsupervised learning methods try to fit models to training data without pre-existing GT labels. Most common supervised methods are logistic regression (LR) [33], support vector machines (SVMs) [64], and decision trees [129] of which we will focus on random forests (RAF) [18] and Extra-Trees [52]. Clustering methods such as k-means [72] and mixture models [12], and anomaly detection algorithms are examples of unsupervised ML methods. In this thesis, we focus on supervised ML methods.

As another group of methods, artificial neural networks (ANNs), are a category of AI algorithms which aim to mimic biological brain structure to make decisions. ANNs consist of networks of so-called neurons which are connected through so-called edges. When they consist of multiple stacked layers, ANNs are called deep neural networks [87].

To apply supervised ML techniques in practice, it is quite common to separate original datasets into training, validation, and test subgroups and then fit the models

on training subgroup and tune the corresponding hyperparameters by including validation set. In the next step, the tuned and fit model is applied to held-out test cohort to estimate its performance on unseen data. In the sections 2.2.1 to 2.2.5, the basic motivations as well as fundamental mathematical definitions of some common supervised learning approaches are presented.

2.2.1 Logistic Regression

Logistic regression (LR) [33] is a predictive analysis algorithm based on the concept of probability used for solving binary classification problems. Fundamentally, logistic regression is a kind of linear regression models with a special type of activation function, the so-called sigmoid function which is also known as logistic function and defined as:

$$\sigma(x) = \frac{1}{1 + e^{-(x)}}. \tag{2.5}$$

Accordingly, for the feature space X, the LR activation function is defined as:

$$h_\theta(X) = \frac{1}{1 + e^{-(\theta^T \cdot X)}}, \tag{2.6}$$

where θ is the vector of coefficients or parameters representing the correlation between the input variables X and the target variable. The computed $h_\theta(X)$ is limited to the range between 0 and 1 and can be interpreted as the probability that the output variable equals 1 for a given input vector X. Thus, considering a decision boundary d, any value of $h_\theta(X)$ higher than d is mapped to 1 and any value lower than d is mapped to 0. For the special case of a single input variable, the decision boundary would be a flat line with offset d from the origin. Figure 2.2 illustrates how logistic regression works for this special case.

2.2.2 Support Vector Machines

The main idea behind support vector machines (SVMs) as a supervised ML algorithm is to find the optimum hyperplane which distinguishes between data points in one, two, or multi-dimensional space, depending on the complexity of the feature space. Figure 2.3 illustrates how SVM works for a simplified binary classification task based on two input independent variables. In order to maximize the probability of accurate classification of unseen data points, the chosen hyperplane should expose the maximum possible distance, i. e., margin, between the data points of different classes, emphasising the impact of the data points residing nearest to the hyperplane (also known as support vectors).

As defined by Cortes et al. [30], defining the input variable space as vector X, the problem of finding the hyperplane with maximum margin can be formulated as the decision function $h(X)$ which is defined as:

$$h(X) = W_0.X + b_0, \tag{2.7}$$

where W_0 term is the normal vector to the hyperplane and the b_0 term is the offset of the hyperplane from the origin in the direction of the normal vector to the hyperplane. Accordingly, the W_0 term is defined as sum of the linearly combined support vectors z_i in the following formula:

$$W_0 = \sum \alpha_i z_i, \tag{2.8}$$

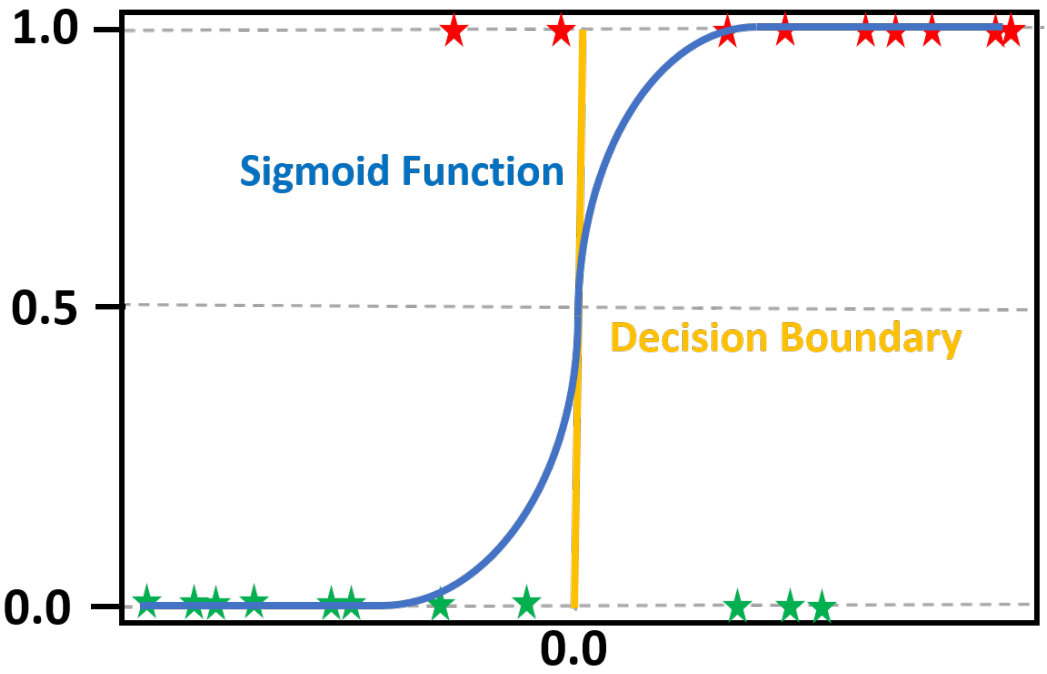


FIGURE 2.2: The illustration of how the logistic regression (LR) classifier works on the simplified case of single input variable.

where α_i are the parameters vector.

Hard vs Soft Margin

In the case of linearly separable training data, the optimal hyperplane as defined in 2.7 denotes the so-called hard margin. However, in case of not linearly separable training data for a binary classification task with output variable $y = \pm 1$, one could alternatively use hinge loss function which is defined as:

$$l(h_X) = max(0, 1 - y \cdot h(X)), \tag{2.9}$$

where $h(X)$ is the value calculated by the decision function for the input vector X and y is the ground truth label for the input vector X. For data points residing on the correct side of the margin of their corresponding class, this function returns 0, while for the data points on the wrong side of the margin, the loss is proportional to their distance from the margin of their corresponding class. Accordingly, to minimize the loss, the following formula should be minimized:

$$\frac{1}{n}\sum_{1}^{n} l(h_X) + \lambda \|W\|^2, \tag{2.10}$$

where λ is used to increase the margin while ensuring that training data points X lie on the correct side of the plane. To conclude, the main objective of a soft margin is to find the optimal hyperplane which minimizes the number of errors in classifying training data points which are linearly non-separable. In general, this problem is NP-complete. However, in the simplified case in which the cost function is the hinge loss, the SVM's soft margin is a convex problem for which a unique solution would exist [30].

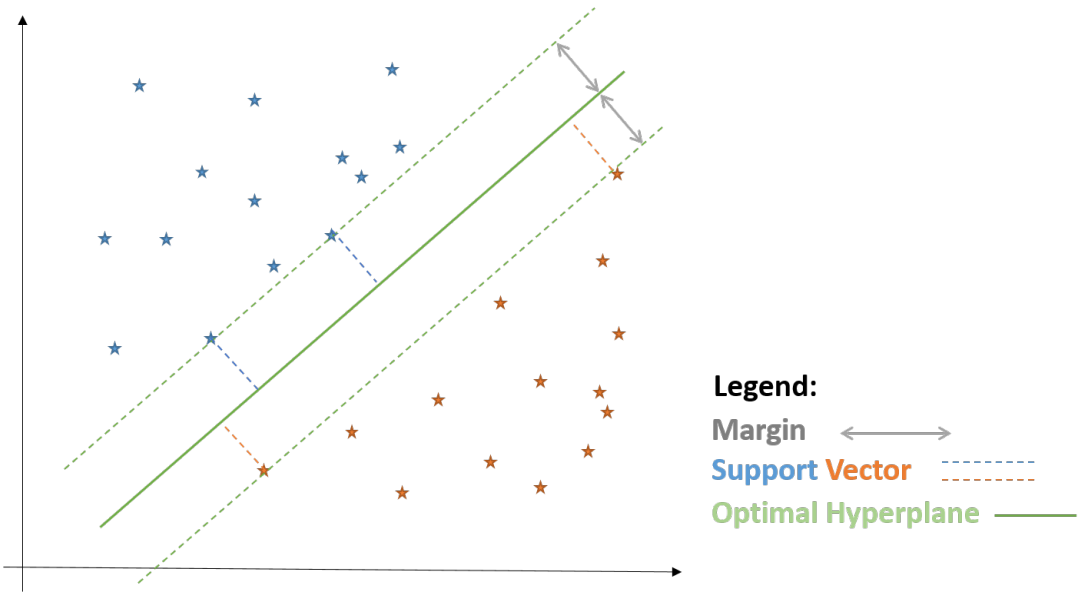


FIGURE 2.3: The simplified illustration of how the support vector machine (SVM) classifier finds out to separate between two groups based on two independent input variables.

2.2.3 Kernel Functions

If the original feature space is not linearly separable, one could map the data to a space with higher dimensionality in which linear algebra and geometry can be used to separate the classes. However, if the dimensionality of the space is very high, two problems might occur [35]: First, the algorithm is subject to overfitting as it would be highly biased with the training samples. Moreover, the high computational power needed in a very high dimensional space limits the size of the problem to be solved by the classifier. To deal with this problem, one could apply kernel functions.

As defined in [35], kernels are inner products in the destination space (i. e., the embedding space) which can often be computed efficiently. A kernel function can be written as:

$$K(X, Z) = (\Phi(X), \Phi(Z)), \tag{2.11}$$

where Φ is the embedding function which maps the data points in the embedding space.

In practice, several kernel functions can be linearly combined and applied to the training points. In the special case of SVM classifiers, the decision function can be described as:

$$f(X) = sign(\sum_i \alpha_i y_i K(X_i, X) + b), \tag{2.12}$$

where X_i are the training points, y_i are the ground truth labels, K is the kernel function, and X is an unseen test data point. Here, the task of finding the hyperplane with maximum margin is mapped to a quadratic programming problem with a convex objective function which can be efficiently maximized under certain constraints. This can be marked as a benefit of the algorithms using kernel functions compared to the algorithms such as decision trees and neural networks which use cost functions with local minima [35].

Two common nonlinear SVM kernels which are used in our methods are polynomial and radial basis function (RBF) as discussed as follows:

Polynomial Kernel SVM

For two sample points X, X' in the feature space, a polynomial kernel with degree of d which maps X to X' is defined as:

$$K(X, X') = (X \cdot X' + c)^d, \tag{2.13}$$

where $c \geq 0$ is a free constant parameter. The special case of $c = 0$ corresponds to the so-called homogeneous polynomial kernel [35].

Radial Basis Function Kernel SVM

For two sample points X, X', the RBF kernel is defined as [157]:

$$k(X, X') = \exp(-\gamma \|X - X'\|^2), \tag{2.14}$$

where γ is equal to $\frac{1}{2\sigma^2}$ for a free parameter σ and $\|X - X'\|^2$ is the squared Euclidean distance d^2 between the two feature vectors as defined here:

$$d^2(p, q) = (p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2, \tag{2.15}$$

where p, q are two points in the feature space with corresponding coordinates of $p_{1:n}$ and $q_{1:n}$ in an n-dimensional space [143].

2.2.4 Decision Trees and Ensemble Methods

Decision trees are kinds of ML algorithms which map the classification problem into tree-structured flowcharts of decisions based on the values of the input features [129]. At each node of such trees, the classifier decides based on a single feature whether to make the final prediction or make another decision based on another feature. The leaves of the decision tree are the final classes or labels. Figure 2.4 shows an example of a decision tree.

Ensemble methods take advantage of a pool of possible decision trees and make the final decision based on aggregations of outcomes of different trees. An example of ensemble methods to cope with a classification problem is shown in Figure 2.5 Here, the mechanism of aggregating the result from different trees varies between different ensemble algorithms such as random forests (RAFs) and extra trees. For instance, RAF applies bagging which denotes facilitation of bootstrapped resampling of the input samples and aggregating different decision trees, while extra trees method uses aggregation of trees and reuses the whole data set without bootstrapping.

Random Forests (RAFs)

As first introduced in 2001 by Leo Breimann [18], random forest (RAF) is a kind of ensemble methods that consists of a large number of decision trees, grown based on random subsets of features, also known as estimators which produce independent predictions. Unlike most standard decision tree classifiers, RAFs are less subject to overfitting, due to the use of ensemble design. As in medical diagnosis often there is

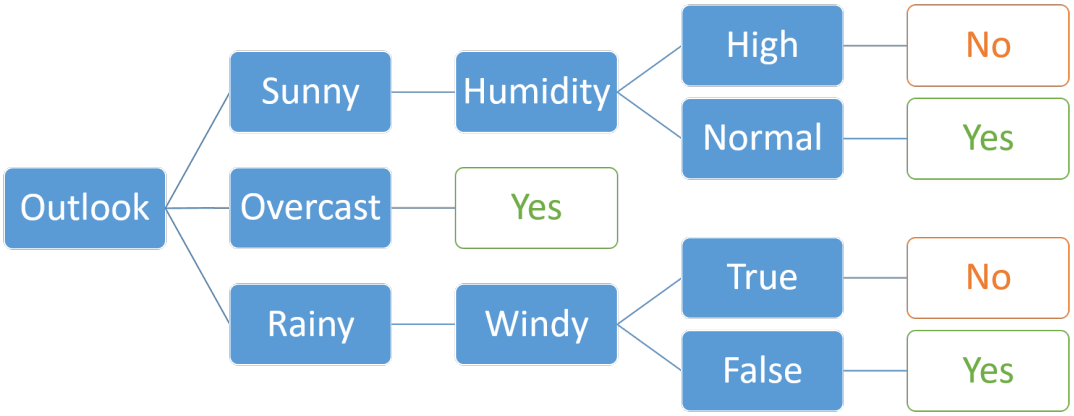


FIGURE 2.4: Represents a decision tree inspired by the original example given by Quinlan, J. R. [129]. This decision tree helps to decide whether to play outside based on different weather conditions.

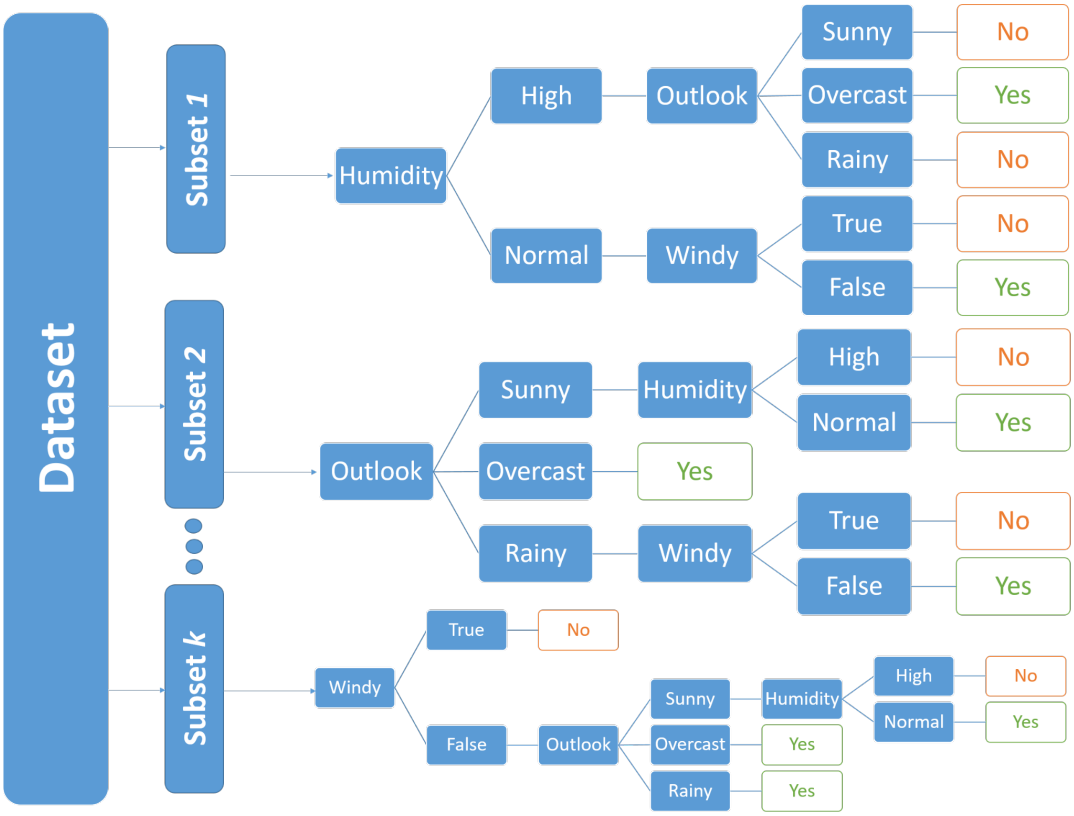


FIGURE 2.5: An example of a forest of trees. First, bootstrapped random samples of the dataset are chosen. Then for each subset, a random decision tree is generated based on a random order of variables.

a large number of input features to a classification or regression problem with each one containing a small fraction of information, taking advantage of a combination of trees grown using random features and letting them vote for the output class or value might result in higher accuracies than those of a single tree classifier [18].

As defined by Breiman, L. [18], a random forest is a classifier which consists of a set of ensemble classifiers $h(X, \theta_k)$ which are defined as:

$$h_k(X) = h(X, \theta_k), k = 1, \cdots, \tag{2.16}$$

where the θ_k are independent random vectors that are distributed identically and each tree has a unit vote on the ultimate decision for input X. Given an ensemble of classifiers $h_k(X)$ and the training set of the feature vectors X, target labels y, and the indicator function $I(.)$, the margin function is defined as:

$$mg(X, y) = avg_k I(h_k(X) = y) - max_{j \neq y} avg_k I(h_k(X) = j), \tag{2.17}$$

and calculates to which extent the average (avg) number of votes for the right class at X, y exceeds the average vote for any other class. Thus to achieve higher confidence in classification, the goal is to maximize the margin and reduce the generalization error PE^* which is defined as:

$$PE^* = P_{X,y}(mg(X, y) < 0). \tag{2.18}$$

As the number of trees grows, for almost all independent random vectors θ_k as described in 2.16, the PE^* would converge to:

$$P_{X,y}(P_\theta(h(X, \theta) = y) - max_{j \neq y} P_\theta(h(X, \theta) = j) < 0), \tag{2.19}$$

which has been proved in [18] and describes why RAF would not overfit as the number of trees grows, although a limiting value of the generalization error PE^* is produced.

Extra Trees (ETs)

Similar to RAF, extremely randomized trees or extra trees are a kind of ensemble algorithms which combine many decision trees to end up with a decision. As first proposed in 2006 by Geurts et al. [52], for a given input feature, the extra trees algorithm selects its cut-point fully independently of the target variable, hence totally randomly. This property of ET makes it more computationally efficient than random forest, as it does not calculate the optimal cut-point at each split. Another difference of the extra trees with other tree-based ensemble methods is that it uses the full learning sample rather than a bootstrapped subsample to grow the ensemble trees. In case of a regression problem, extra trees' predictions are made based on averaged predictions of decision trees. For classification problems, the predictions are made by majority voting from decision trees [52].

More interestingly, slightly modified versions of ensembles of trees (e. g., RAFs and ETs) can be thought of as kernel-based models which makes them more interpretable and easier to analyze [137]. Accordingly, the kernel-based model for an ensemble $\tau = \{t_i : i = 1, \cdots, M\}$ of M trees t_i, is defined as:

$$K_\tau(x, x') = \frac{1}{M} \sum_{i=1}^{M} K_{t_i}(x, x'). \tag{2.20}$$

In case of extremely randomized trees, the kernel $K_\tau(x, x')$ is independent of input values of target variables. For finite M, the kernel is piece-wise constant. For ETs with an infinite number of ensembles ($M \to \infty$), K is continuous and piece-wise multi-linear [52].

As Breiman [18] showed, if the assumptions of uniform prior distribution $P(x)$, infinite sample size, and infinite number of ETs of fixed number of leaves l hold, the kernel function can be approximated as follows:

$$K_\tau(x, x') \approx exp\{-\lambda|x - x'|_1\}, \tag{2.21}$$

where $|x - x'|_1$ is the so-called city-block or Manhattan distance and λ is the sharpness of the kernel which is defined as:

$$\lambda = \frac{log(l)}{n}, \tag{2.22}$$

where n is the dimension of the input space. Here, for balanced trees with a finite sample size N, the number of leaves l is on the order of $\frac{N}{n_{min}}$. This would suggest that higher values of n (i. e., higher dimension) have stronger smoothing effect than that of higher values of n_{min} (the minimum number of the samples in the leaves). Also, Lin et al. [89] show that if the number of samples $n_{t_{i,j}}$ at each terminal node of all trees equals to a constant k, as the sample size $N \to \infty$, then the sharpness λ would be on the order of $k(logN)^{n-1}$. This means that the effect of dimensionality (i. e., number of features) is much stronger than the effect of samples k kept in the leaves. This would explain why in problems with high-dimensional space, increasing the n_{min} has a negligible positive impact on the prediction accuracies [52].

2.2.5 Deep Neural Networks

Inspired by the architectural depth of the human brain, deep neural networks (DNNs) are multi-layered ANNs which consist of at least one input, one output and several hidden layers in between, with each hidden layer applying a different level of non-linear operations. Thus, different layers of such a network represent features at different abstraction levels, each composed of lower-level features [13]. DNNs are successfully applied for a variety of tasks from object detection to image classification. Highlighting DNNs that are best fitted for medical image segmentation, U-Net [132] based models are widely used for clinical decision support [47, 55, 60].

DNN Architecture

Apart from input(s) and output(s), DNNs consist of several hidden layers in between. Each layer of a DNN consists of several neurons with each neuron connected to all or some of the neurons from the previous or the next layer. The input data is fed to the neurons of the first layer and the outputs of each layer are passed to the next layer as its input. This process is repeated until the final layer generates the final output. The simplified architecture of a DNN is shown in figure 2.6. In case of a classification task, the output is the probability of each class and in case of binary image classification, the output is a binary masked image. The neurons of each layer associate with an activation function which based on a threshold, decides either to ignore or to pass to the next neuron through a connection which itself is associated with a weight. The weight quantifies the influence of its input neuron to its output one. At the beginning, all the weights of the network are set at random. At the training, the weights are iteratively updated based on the desired output, i. e., the

ground truth labels. This process is facilitated using a learning mechanism known as the optimizer.

Backpropagation

The term backpropagation or backward propagation denotes the procedure in which the weights of the connections of a neural network are iteratively updated based on the values of an error function quantified as the level of difference between the networks output and the desired output. This is facilitated by calculating the gradient of the error function with regards to the weights of the connections. As the procedure starts from the output layer and goes back until the first layer, the backward term is used. For a feedforward neural network (a neural network in which no circular connection is made between the neurons and units) with parameters vector θ composed of weights w and biases b, the backpropagation is formulated using delta learning rule for the dataset $X = (x_i, y_i)$ (consisting of pairs of inputs x_i and desired outputs y_i, where i takes values between 1 and N which is the sample size) and the error function $E(X, \theta)$. The error function which measures the level of agreement between predicted labels and actual labels of a dataset as fed to a network can be formulated in many ways depending on the type of predictions.

Mean Squared Error (MSE)

One of the most common error functions used for training neural networks is the mean squared error (MSE) defined for the dataset $X = (x_i, y_i)$ as:

$$E(X) = \frac{1}{2N} \sum_{i=1}^{N} (g(w.x_i + b) - y_i)^2, \tag{2.23}$$

where g is the activation function and y_i are the desired output. Therefore, $E = 0$ means $g(.) = y_i$ for all pairs of x_i and y_i. Thus, the goal is to minimize E with respect to w and b.

Cross-Entropy

Another error function, which is used for different classification tasks, is cross-entropy loss which is a measure used in the field of information theory and is defined based on entropy as the dissimilarity between two probability distributions. From the field of information theory [31], the term *information* is defined as the number of bits required to encode and transmit an event. Events with lower probability contain more information and vice versa. Hence, the *information*, $h(x)$, can be measured for an event x based on the probability of the event $P(x)$ as:

$$h(x) = -log(P(x)), \tag{2.24}$$

and the entropy $H(X)$ is defined as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) log(P(x_i)), \tag{2.25}$$

where X is a discrete random variable with possible outcomes x_i with probabilities $P(X_i)$.

Accordingly, the cross-entropy between discrete distributions p, q is defined as:

$$H(p,q) = -\sum_{i=1}^{n} p(x_i) log(q(x_i)). \tag{2.26}$$

For the special case of binary classification, the cross-entropy loss function is calculated as follows:

$$E = -(y_o log(p_o) + (1 - y_o) log(1 - p_o)), \tag{2.27}$$

where y_o is the true binary label and p_o is the predicted probability of the observation o.

Delta Rule

The error function $E(X)$ is usually minimized using gradient descent approach which is mostly applied for differentiable error functions. The gradient descent is an iterative approach aiming to find a local minimum with respect to the parameters w and b. First, the values are set to w_0, b_0 at random. Then at each iteration, the gradient descent updates the values as described in:

$$\begin{aligned} w_{i+1} &= w_i - \alpha \frac{\partial E(X)}{\partial w_i}, \\ b_{i+1} &= b_i - \alpha \frac{\partial E(X)}{\partial b_i}, \end{aligned} \tag{2.28}$$

where w_i, b_i are the corresponding values of w, b at iteration i respectively, α is the learning rate, and $\frac{\partial f}{\partial x}$ is the partial derivative of f with respect to x. The learning rate α, which controls the step size of the gradient descent at each iteration should be set to a relatively small value in order to let the algorithm converge. Choosing very small values for α would mean taking very small steps towards local minimum (hence, a long convergence time), while setting it to a very large value would make the minimization fail to converge. Accordingly, the calculation of the $\Delta w = w_{i+1} - w_i$ and $\Delta b = b_{i+1} - b_i$ is denoted as the delta rule as the special case of backpropagation for single layer perceptrons [160].

Convolutional Neural Networks

Convolutional neural networks (CNNs) are special kinds of deep learning models implemented for datasets attributing grid patterns such as images [162]. CNNs mimic the organization of animals' visual cortex and are used to capture automatically and adaptively the spatial and hierarchical features from low- to high-level patterns in data. CNNs usually consist of three kinds of building blocks: convolution, pooling, and fully connected layers. The convolution and pooling layers contribute to feature extraction and the fully connected layer corresponds to the mapping of the extracted features into final outputs of the model. In the special case of U-shaped networks, no fully connected layer is included. For 2D images, the convolution layer applies to each image position some linear mathematical operations on small grids of parameters called kernels, which are optimizable feature extractors. Each layer feeds its output to the next layer, as a result, a hierarchical set of complex features are extracted. Thus to train a CNN, the network parameters such as kernels should be optimized using backpropagation and gradient descents as described before for DNNs in 2.2.5.

In the following lines, the main components of a CNN are described.

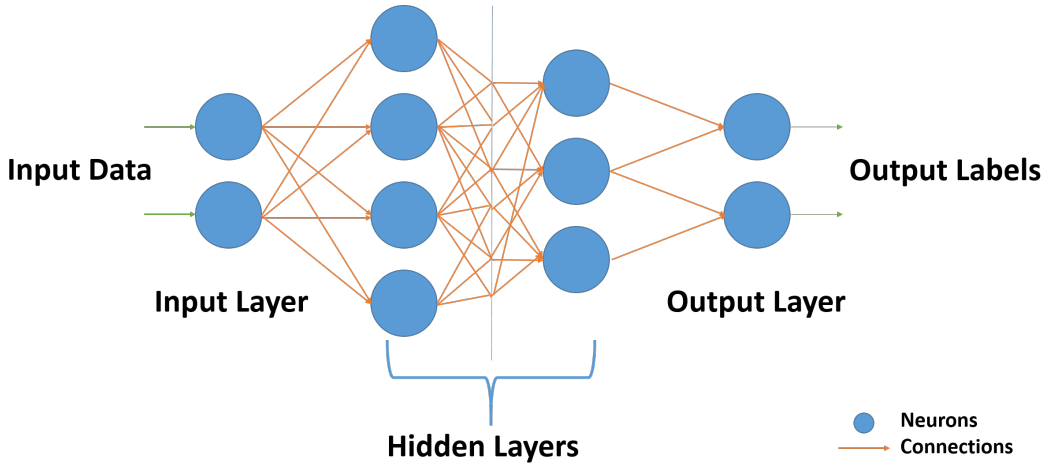


FIGURE 2.6: A simplified example of a deep neural network. The hidden layers can contain several layers of fully or sparsely connected neurons. Each connection has a weight which is updated as the model is trained.

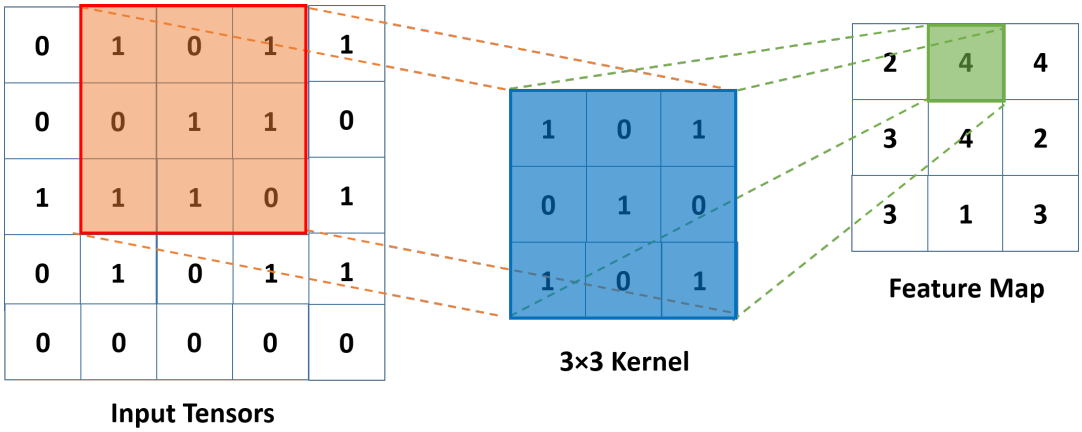


FIGURE 2.7: An example of a 2D convolution operation with a 3×3 sized kernel, stride 1, and no padding.

Convolution Layer: As a fundamental building block of a CNN, the convolution layer combines a convolution operation with an activation function as linear and non-linear components respectively. Convolution is used to extract features leveraging linear operations. It applies a small array of numbers (i. e., kernels) to arrays of numbers formed across the input (i. e., tensors). To end up with the feature map, element-wise products between elements of the kernel and and the input tensors are calculated at all tensor positions and summed together. By applying multiple kernels, an arbitrary number of feature maps representing various characteristics of the input tensors are provided. Therefore, the term kernel can be referred to as a feature extractor. Convolution operations are typically parameterized by their size, denoting the window size of the kernel, and number of kernels they use, determining the depth of output feature maps. Figure 2.7 illustrates a convolution operation including a 3×3 kernel with stride 1.

The convolution operation may also include parameters *padding* and *stride*. The padding is used to be able to further move the center of the kernel window to the boundaries of the input tensor, making it possible to have feature maps in equal

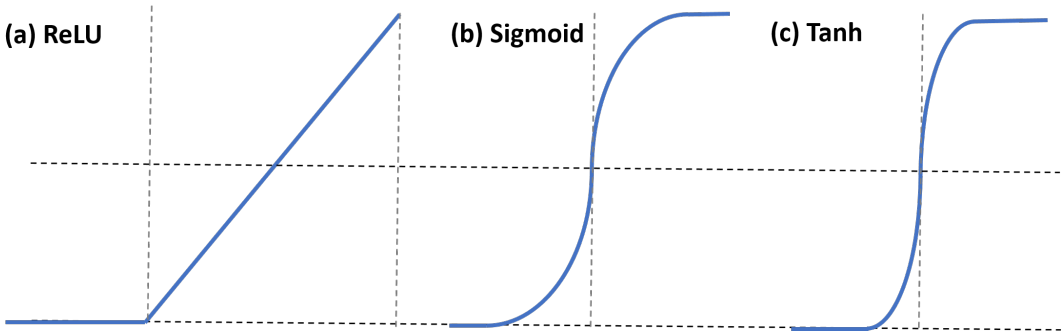


FIGURE 2.8: The most common activation functions applied in neural networks: (a) rectified linear unit (ReLU), (b) sigmoid, and (c) hyperbolic tangent (tanh).

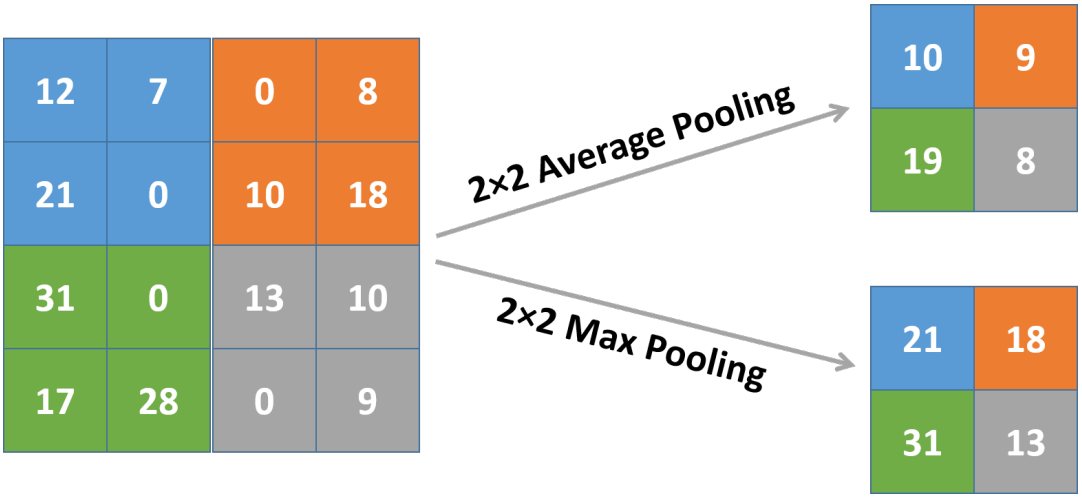


FIGURE 2.9: An example illustration of average pooling and max pooling methods with 2×2 pool size and stride 2.

size as the input tensor. This can facilitate applying successive convolution layers without ending up with feature maps with too small sizes. The stride controls the jumping step size (i. e., the distance between two successive kernel positions) as the kernel window moves across the input window. One impact of strides is to achieve downsampling of feature maps.

The non-linear component of a convolution layer is called the activation function. Conventionally, sigmoid and hyperbolic tangent (*tanh*) have been used as activation functions. Most recently, rectified linear unit (ReLU) is the most common activation function used in modern CNN models (see figure 2.8). For input x, ReLU is simply defined as:

$$ReLU = max(0, x). \tag{2.29}$$

Pooling Layer: Pooling layers apply downsampling operations in order to reduce the in-place dimensionality of the feature maps and therefore introduce a translation invariance to tiny distortions and shifts. Pooling layers do not attribute any learnable parameter; however as for convolution layers, they often attribute parameters such as filter size, padding, and strides. Average and max pooling are common pooling operation used in deep neural networks. Figure 2.9 compared the two types of pooling methods.

Fully Connected Layer: Layers in which all the input nodes are connected to all the output nodes via learnable weights are called dense or fully connected layers. In CNNs used for classification problems, an activation function follows each fully connected layer. In this case, the final fully connected layer of each CNN has the same number of output nodes as the number of the ground truth classes and is usually followed by a different activation function as used for other layers. For instance, sigmoid activation is commonly applied for binary classification tasks. In case of a regression problem, no activation function is used in the output layer of the CNN.

2.2.6 Performance Metrics

To quantify the performance of machine learning classifiers and rank different methods, several approaches and measures can be determined. Confusion matrix (also known as contingency matrix) and receiver operating characteristic (ROC) curves are most common approaches to assess classification outcomes. Moreover, a variety of metrics such as accuracy, sensitivity (or recall), specificity, precision, and area under the curve (AUC) are widely used to quantify classification performance. Last but not least, in case of assessment of binary classification or segmentation methods, Dice similarity coefficient (DSC) is commonly applied.

The measures:

- True positives (TP): the number of the positive cases which were correctly classified as positive by the classifier
- True negatives (TN): the number of the negative cases which were correctly classified as negative by the classifier
- False positives (FP): the number of the negative cases which were incorrectly classified as positive by the classifier
- False negatives (FN): the number of the positive cases which were incorrectly classified as negative by the classifier

form the basis for the upcoming formulations which briefly describe classification performance metrics as follows:

Confusion Matrix

The confusion matrix is represented as a table formed based on the actual (i. e., ground truth) labels and predicted labels (see figure 2.10). Confusion matrices give an overall outline of how the classifier performed and can make it easier to calculate measures such as sensitivity and specificity.

ROC Curve and Area Under the Curve (AUC)

ROC is a curve which connects the points created by plotting the true positive rate against the false positive rate at various thresholds [43]. The bigger the area under the ROC (AUC), the higher classification accuracy. Figure 2.11 gives and illustration of three different ROCs representing three classifiers with high, good, and poor predictive performance.

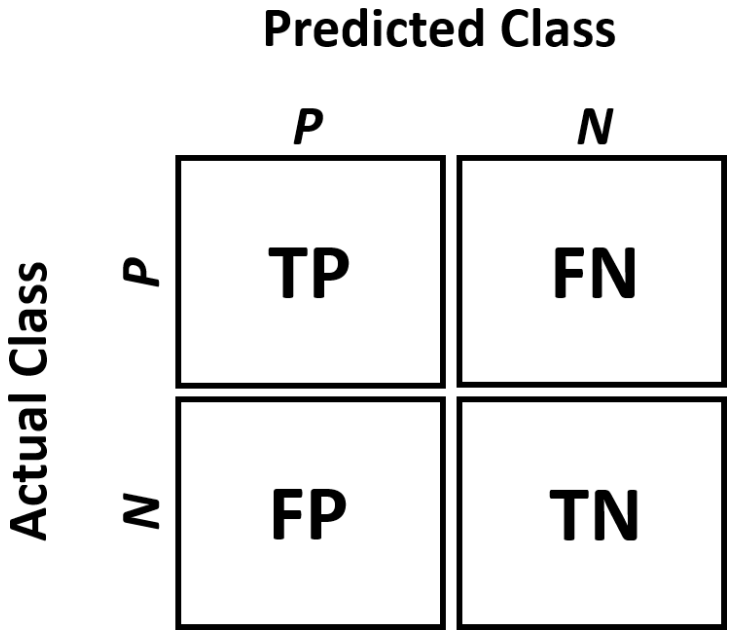


FIGURE 2.10: An example of confusion matrices for a binary classification task (TP: true positives, TN: true negatives, FP: false positives, FN: false negatives).

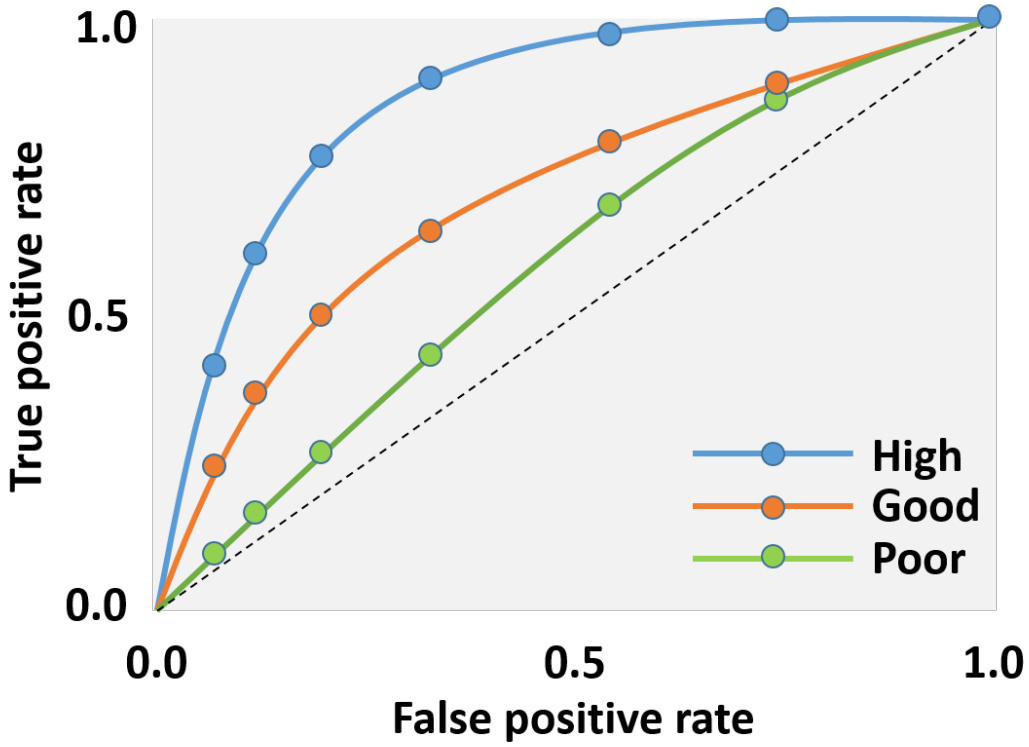


FIGURE 2.11: Three ROC curves. The blue, orange, and green curves represent three classifiers with high, good, and poor predictive performance respectively.

Accuracy

Accuracy is defined as the proportion of the true predictions to the total size of the training samples:

$$acc = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.30}$$

Sensitivity or Recall

Sensitivity or recall which denotes the true positive rate is defined as:

$$Sensitivity = Recall = \frac{TP}{TP + FN}. \tag{2.31}$$

Precision

Precision is defined as the proportion of the true positives to the sum of the true positives and the false positives:

$$Precision = \frac{TP}{TP + FP}. \tag{2.32}$$

Specificity

Specificity is defined as the number of the true negatives divided by the sum of the true negatives and the false positives:

$$Specificity = \frac{TN}{TN + FP}. \tag{2.33}$$

Dice Similarity Coefficient (DSC)

The Dice similarity coefficient, also known as the Sørensen–Dice index [111], of two sets of data is defined as the level of similarity between the two datasets. For the two sets of data X, Y the DSC is formulated as:

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}, \tag{2.34}$$

in which twice the number of common elements of the two sets is divided by the total number of elements in the two sets. Alternatively, in case of boolean classification, DSC can be calculated based on TP, FP, and FN metrics as follows:

$$DSC = \frac{2TP}{2TP + FP + FN}. \tag{2.35}$$

2.2.7 Train, Validate, and Test

Similar to any other field, when experimenting AI methods in medical domain, in order to draw conclusions that are most likely generalizable to wider ranges of unseen cohorts of data, it is critical to subdivide available cohorts into separate train, validate, and test sub-cohorts. The size of the train, validate, and test cohorts should be set appropriately to avoid ending up with classifiers which are highly biased with the training cohort, hence the trained model is less subject to overfitting. For the supervised ML case, different hyperparameters of classifiers need to be tuned

with cross validation and grid search given only the train and validate cohorts samples. If the data cohort is relatively small, the original dataset can be subdivided into train and test cohorts and the cross validation and hyperparameter tuning can be conducted by applying iterative runs of train and validate within the training set.

Cross validation (CV) uses different sub-samples of the training cohort and applies interim train and validate attempts in which each subset of the data is used for both training and validation at least once. As a result, classifier performances are quantified. Common CV methods are KFold and leave-one-out CV which are applied depending on the classification task and attributes of the subject or sample cohort.

Grid search is the procedure in which a classifier is trained with all the possible ranges of standard values for its hyperparameters and as a result, the best set of its hyperparameters which achieve best performance are chosen. For instance, for the SVM classifier with polynomial kernel, the degree of the polynomial is the most important hyperparameter which needs to be tuned. Grid search is usually applied alongside CV in the training step.

In deep learning scenarios, apart from the architectural parameters, e. g., number of encoding and decoding layers, convolution sizes, activation functions, and pooling layers, entities such as number of epochs, learning rate (as described in 2.2.1), and batch size are some of the most important hyperparameters to tune. The number of **epochs** limits the maximum number of times that the entire training dataset is passed through the network to train and fit the model. The epochs number is usually set to a relatively high number $(50, 100, 1000, \cdots)$, aiming at minimizing the error function values as the model gets trained and tested against a separate validation set. The **batch size** is a hyperparameter of gradient descent which defines the number of training samples to work through before the internal model parameters are updated. Most common values for the batch size are 16, 32, 64, and 128, depending on the training cohort size (which is the maximum possible value for batch size). In state-of-the-art libraries for modeling deep and convolutional neural networks such as TensorFlow [1], some early stopping criteria such as monitoring the validation loss and patience can be considered to further customize the training step.

2.2.8 Comparing ML Methods Applied for Diagnosis and Prognosis

Machine learning methods have been extensively used in the medical domain. In particular, diagnostic and prognostic hypotheses of clinical studies can be mapped to classical ML problems. For example, discriminating malignant uptake from normal tissue or prediction of treatment outcome for individual subjects, given hotspot- or patient-level radiomics feature vectors (as will be introduced in 2.4.1) respectively, can be considered as supervised ML problems. Depending on the size of the experiment cohort, one could justify application of deep learning methods as they normally require bigger training cohorts to converge. Furthermore, as radiomics pipelines usually provide hundreds of variables as input feature vectors to the ML problem space, it becomes necessary to apply feature selection techniques to avoid overfitting, especially when dealing with a limited number of subjects or samples compared to the number of radiomics features.

According to the report by Uddin et al. [153], while SVM has been the most common supervised learning approach used in disease prediction, random forest has been superior to SVM classifiers in terms of accurate predictions. However, it

is critical to mention that such comparative analyses might be highly biased due to differences in experiment designs and diversity of subject cohorts.

As other common approaches for disease management, deep learning methods can be used in the oncological domain for a variety of use-cases including tumor segmentation and therapy response assessment. For instance, U-Net based models are facilitated for different medical image segmentation tasks [47, 55, 60]. Also, if the study cohort is reasonably big, deep neural networks can be trained and fit to draw subject-specific decisions about disease stages, treatment outcomes or survival risks [83, 158, 167]. As the lack of subjects may lead to overfitting, studies that suffer from this issue mostly conduct solutions such as data augmentation and cross validation within the training cohort.

In the special case of radiomics analysis, unlike other supervised learning methods such as SVMs and random forests which take hand-crafted radiomics features such as textural heterogeneity parameters, CNNs do not necessarily need hand-crafted features. However, CNNs are relatively more data hungry and more computationally expensive, therefore most likely, they require graphical processing units (GPUs) to be trained [162].

2.3 Hotspot Segmentation

As a non-invasive method avoiding multiple biopsies and replacing histopathological analyses, CAD-based tumor delineation has become more justifiable in clinical practice. Therefore, true segmentation of cancerous tissues has been a critical part of clinical decision support tools. Most often, manual segmentation tools such as InterView FUSION [98] are used to identify malignancies and to stage the cancer disease. However, specially in case of patients with multiple metastatic lesions, the procedure of delineating all malignant hotspots becomes more time consuming and attention intensive. To reduce the time and effort required for such diagnostic procedures, thresholding-based and automated segmentation approaches are facilitated. In this part of the thesis, we give a short background of some common techniques applied for tumor segmentation, ranging from fully manual to fully automated algorithms.

2.3.1 Manual Segmentation

Regardless of the imaging modalities, any medical imaging device manufacturer provides scans as 2 or 3 dimensional (2D or 3D) imaging objects. The scanned objects are stored in standard medical imaging formats such as digital imaging and communications in medicine (Dicom) [121] and neuroimaging informatics technology initiative (NifTi) [116]. CAD systems such as InterView FUSION further process these file formats and provide interactive user interfaces (UIs) which facilitate visualization and annotation of the regions or volumes of interest (RoIs or VoIs). The annotated scans then can be used for diagnosis or to provide ground truth (GT) for clinical studies. Although such CAD systems are widely used in clinical routines to manage oncological patients or to conduct research, they require extensive amounts of time and effort by domain experts which can be reduced by taking advantage of automated segmentation techniques.

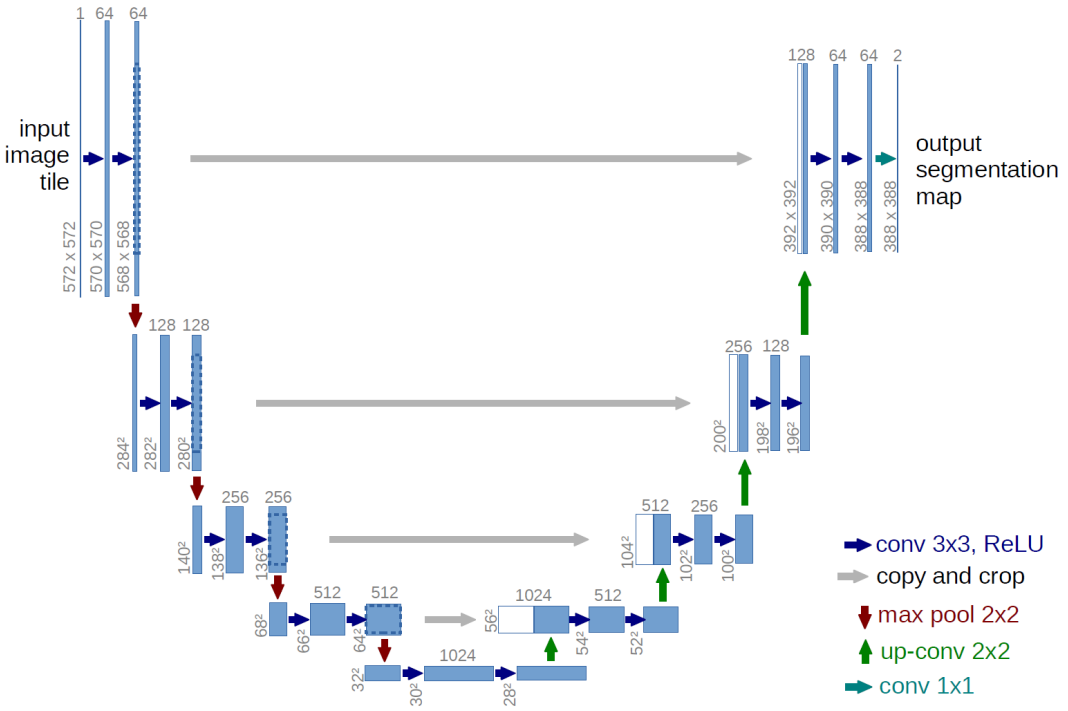


FIGURE 2.12: The original U-Net architecture as proposed by Ronneberger et al. [132].

2.3.2 Thresholding Based Segmentation

As conventional techniques for tumor segmentation, thresholding based methods have been in focus for years. Thresholding based methods can be categorized into fixed and adaptive thresholding, depending on how the thresholds are calculated and applied to the original input images. The choice of threshold in the oncological domain usually follows conventional fixed values based on previous findings and often applies adaptive majors based on cohort- or patient-specific demographic or clinical factors. For instance, for prostate cancer treatment, often 40% of standardized uptake value (40%-SUV_{MAX}) is used to define VoIs. As a limitation of the thresholding based methods' outcomes, diverse characteristics of different medical imaging modalities and scanners such as varieties in scanning resolutions often result in inconsistencies [46].

2.3.3 Automated Segmentation

As discussed in previous sections, automated tumor segmentation tools can support domain experts to conduct decisive actions faster. Most commonly, state-of-the-art deep learning based methods are widely applied in the medical domains such as oncology [25, 166] and computational neuroscience [45, 75]. Since U-Net was proposed in 2015 [132], many convolutional neural networks (CNNs) inspired by U-Net have been introduced and used in clinical research. Figure 2.12 represents the original U-Net architecture. As one of the ultimate goals of this PhD project was to provide an automated decision support tool, we utilized a multi-channel U-Net based model for the segmentation of pathological uptake in PSMA-PET/CT scans. Section 4.3.2 in Methodology chapter will describe the details of the utilized deep segmentation network.

2.4 Diagnosis and Treatment Response Prediction

Timely diagnosis and avoiding unnecessary treatment are among the ultimate objectives of any healthcare system. CADs and CDSSs aim at providing such services in (semi-) automated procedures. The basic fundamentals of different ML techniques for diagnosis and treatment response prediction is already discussed in 2.2. Concerning prostate cancer (PCa) as the main focus of this thesis, an overview of some steps towards automated management of PCa patients is given in this section. First, different types of features extracted from PET/CT scans will be introduced. Furthermore, feature selection techniques which are further used in this thesis are elaborated. Finally, the mapping of diagnostic and prognostic problems to ML-based classification tasks is elaborated.

2.4.1 Features

Different features and parameters can be quantified to characterize RoIs and VoIs in medical imaging. These numerical quantities can be then processed for diagnostic and prognostic purposes such as identification of malignant uptake and prediction of treatment response. To this end, numerous variables ranging from RoI specific pixel-based features to patient-specific parameters are used for different purposes. For simplicity, we categorize them in three groups: conventional parameters, radiomics features, and deep features.

Conventional Parameters

Conventionally, clinical studies in the field of oncology in general and prostate cancer in particular had been focusing on a limited number of parameters based on the degree of malignant uptake in subjects. Most commonly, standardized uptake value (SUV), total lesion glycolysis (TLG), and metabolic tumor volume (MTV) had been in focus of many clinical trials and routines for research or staging purposes. MTV is calculated as the sum of the voxel volumes with SUV exceeding a threshold value in an malignant RoI [15] and TLG is usually defined as MTV multiplied by mean SUV [135]. These parameters mainly aim at quantifying the disease stage for specific subjects rather than characterizing the RoIs or VoIs throughout different organs. Also, due to the limited number of such parameters, mostly, these parameters are analyzed individually using conventional statistical methods such as linear or logistic regression.

Radiomics Features

As high-throughput quantitative image analysis method, the term radiomics denotes the procedure of extracting hundreds of numerical quantities out of medical imaging data in terms of 2D or 3D intensity-, shape-, and texture-based features which characterize RoIs or VoIs regardless of being malignant or normal tissues. These features include but are not limited to first and higher order statistics intensity based features and textural heterogeneity based parameters such as entropy and kurtosis. Recently, as the usage of numerous radiomics features becomes more common in clinical studies, machine learning algorithms are widely applied to hotspot- or patient-level radiomics feature vectors for different purposes.

As will be discussed in Methodology chapter (4), radiomics analysis has been one of the fundamental aspects of this thesis. To mention some of our related findings, tumor identification [42, 107] and therapy response prediction [103, 106] can be

facilitated using supervised ML methods as applied to radiomics features. Depending on the segmentation procedure, we applied different sets of radiomics features: 1) features calculated by InterView FUSION [98] as side product of manual VoI delineation and 2) features calculated using PyRadiomics library [56] based on automatically segmented VoIs. The detailed list and definitions of the most important radiomics features used in our methods pipelines are provided in 4.4.

Deep Features

One important aspect of acquiring deep and convolutional neural networks in the fields of medical image segmentation, diagnosis, and prognosis is the features which are calculated in the layers of these deep networks. As elaborated in 2.2.5, each DNN or CNN consists of several hidden layers in which different filters or kernels with different sizes are applied to input tensors, which themselves are outputs from previous layers, to calculate some feature maps. For instance for U-shaped models, this procedure is normally applied in the encoder parts of the network. As a result, a magnitude of features can be extracted at different levels ranging from primitive edge and contour based features to complex texture, shape, or compositions of features from previous layers. Similar to radiomics features, vectors of deep features can be formed and applied for hotspot- as well as patient-level diagnostic and prognostic tasks such as treatment outcome prediction and survival analyses. For example, Andrearczyk et al. [6] apply and compare radiomics and deep features from a multi-task 3D U-shaped network to predict disease-free survival from FDG-PET/CT scans for a cohort of head and neck cancer patients. Although this thesis is mostly dedicated to conventional parameters and radiomics features, a lack of experiments focusing on deep features and their relevance for treatment response prediction and survival analysis for PCa patients based on multimodal ^{68}Ga-PSMA-PET/CT images has been identified which can be considered as future work.

Feature Selection

In machine learning and statistics domains, feature selection is applied for several purposes such as dimensionality reduction, model simplification, and time and memory optimization. In the special case of radiomics analyses, as the main focus of this thesis, identification of the most relevant radiomics features by removing redundant and irrelevant features is of great importance. On the one hand, identifying the most important radiomics features is highly appreciated in the oncological research domains including disease staging and survival prediction studies [57, 92, 158]. On the other hand, from a technical point of view, ML models trained with few non-redundant feature vectors are less subject to overfitting.

As elaborated in [57], in general, feature selection techniques can be categorized in the following three groups:

Wrapper Methods: this category of variable selection methods take advantage of a predictive model to rank different possible subsets of the feature space. To come up with subset specific scores, first the data cohort is subdivided into train and held-out test subsets. Then, a model is trained on the training subset using only the subset of features which were chosen and the errors in predicting the labels for the held-out set is quantified. As it exposes train and test attempts for each possible subset of features, this approach is considered as very computationally expensive.

Filter Methods: unlike wrapper methods, filter methods apply a proxy measure such as mutual information instead of the error rate to rank variables or subsets of

them. Compared to wrapper methods, filter methods are less computationally expensive and are not bound to predictive models. Moreover, instead of selecting a subset of features, most filter methods provide rankings of all features. Filter methods are often applied as a pre-processing step prior to wrapper methods. Also, recursive feature elimination (RFE) [58] is another filter approach which is widely applied with SVMs. RFE takes an ML classifier and the desired number of features as input and starts from the entire feature space. Then at each recursion step, RFE ranks the features based on importance metrics either using the classifier's importance metric or an independent statistical method and removes least relevant variables. This recursive procedure continues until the desired number of relevant features are selected [58].

Embedded Methods: this group of methods conduct feature selection as part of the model construction procedure. For instance, least absolute shrinkage and selection operator (LASSO) [81] method is applied for constructing a linear model which penalizes the regression coefficients with an L1 penalty which shrinks many of the coefficients to zero. As a result, only the features with non-zero regression coefficients are chosen. LASSO applies regression analyses to facilitate variable selection and regularization, aiming at enhancing the prediction performance of the outcoming statistical model [81].

For a sample with the size of N cases, each consisting of p covariates and a single target variable y, the goal of LASSO method is to solve [151]:

$$\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t, \tag{2.36}$$

where $x_i = (x_1, x_2, \cdots, x_p)_i^T$ and y_i are the covariate vector and the target label for the i^{th} case respectively, β_0 is the constant coefficient, $\beta = (\beta_1, \beta_2, \cdots, \beta_p)$ is the coefficient vector, and t is a free parameter indicating the degree of regularization.

Considering RFE, LASSO, and feature ranking based on feature importance metrics of ML classifiers as common approaches for feature selection, we applied them in our methods for different classification tasks (as will be discussed in 4.4.3). For example, the LASSO method is widely applied in radiomics analysis to calculate the so-called radiomics signature (RS) which is commonly used for the prediction of survival in oncological studies [92, 105].

To elaborate fundamentals of the common feature selection techniques, some terminologies need to be defined as follows:

Well-posed Problem

The term well-posed problem is coined by 20th-century French mathematician Jacques Hadamard and is defined as a problem for which a unique solution exists and the solution's behaviour changes continuously as the initial conditions change [59]. Accordingly, problems which are not well-posed, are called ill-posed problems.

Regularization

The term regularization refers to the procedure of adding information in order to solve an ill-posed problem or to avoid overfitting. In practice, the regularization term (i. e., penalty) imposes a cost on the optimization function to make the optimal solution unique. The solution is comprised of a data term and a regularization term. In machine learning, the data term corresponds to the training dataset and the regularization term corresponds to the chosen model or to the classifier. Hence, the overall goal is to minimize the generalization error [22].

For a classification problem, the regularization term $R(f)$ is added to a loss function as follows:

$$\min_f \sum_{i=1}^{n} V(f(x_i), y_i) + \lambda R(f), \tag{2.37}$$

where V is the loss function describing the cost of predicting $f(x)$ for the ground truth label y and λ is a parameter controlling the impact of the regularization term. $R(f)$ usually imposes a penalty on the complexity of $f(x)$.

As a matter of fact, regularization is a technique to enhance the generalizability of a predictive model by minimizing the so-called generalization error.

Generalization

The main objective of a generalization problem is to find a function which predicts the target label which minimizes the expected error along all possible independent and target variables. Here, the expected error of a function f_n is defined as:

$$I[f_n] = \int_{X \times Y} V(f_n(x), y) \rho(x, y)\, dx\, dy, \tag{2.38}$$

where X, Y are the domains of the independent variables x and their target labels y respectively and $\rho(x, y)$ is the known joint probability distribution for x and y.

In practice, in many learning problems, only subsets of independent input and target labels are available. Thus, measuring the exact value of the expected error becomes impossible. In this case, the best surrogate measure for the expected error for n samples that are available is defined as:

$$I_n[f_n] = \frac{1}{n} \sum_{i=1}^{n} V(f_n(x_i), y_i), \tag{2.39}$$

which is called empirical error or empirical risk as well [109].

Classification

Many clinical diagnostic and prognostic tasks can be easily mapped to well-known statistical and ML-based methods. For example, discriminating pathological (i. e., malignant) from physiological (i. e., normal) tissues based on hotspot-specific radiomics features or identification of responders to certain types of treatment based on patient-specific parameters can be considered as classification problems. Given annotated datasets including input feature vectors and their corresponding ground truth (GT) labels, supervised ML classifiers can be applied to train and fit models which can predict labels for unseen data. As a result, clinical decision support tools can be provided to assist medical domain experts to make decisions in shorter times and with less effort. As described in previous sections, common classification algorithms in the oncological domain include support vector machines (SVMs) [64] and decision trees [129].

2.5 Overall Survival Analyses

Survival analysis is the field in which the probability of an event of interest, e. g., death of the patient in cancer research, is quantified. The main terminology corresponding to survival analysis includes survival time (or event time), survival function, hazard function, and censoring.

2.5.1 Survival Time

The survival time also referred to as event time or failure time denotes the time from the experiment starts (e. g., the patient is diagnosed, or a baseline scan is taken) until the time point at which the event of interest or censoring occurs.

2.5.2 Survival Function

The survival function quantifies the probability P that a subject survives beyond a given time t and is defined as:

$$\hat{S}(t) = P(T > t) = \int_t^{\infty} f(u)\, du = 1 - F(t), \tag{2.40}$$

where $T \geq 0$ is the response variable (a continuous random variable) and $F(t)$ is the cumulative distribution function (CDF) which quantifies the probability that T will take a value less than or equal to t [38]. At the origin time at which $t = 0$, $\hat{S} = 1$ and as $t \to \infty$, $\hat{S} \to 0$. Theoretically, the survival function attributes a continuous and smooth shape. However, in practice, the observations are recorded in discrete periods such as days, weeks, and months which makes $\hat{S}$ more of a descending step function.

2.5.3 Hazard Function

The hazard function $h(t)$ or $\lambda(t)$ quantifies the instantaneous rate at which an event occurs, with no knowledge of any previous events and is defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{\hat{S}(t)}. \tag{2.41}$$

Accordingly, the cumulative hazard $H(t)$ refers to the accumulated risk that event happens until time t and is defined as follows:

$$H(t) = \int_0^t h(u) du, \tag{2.42}$$

and can be interpreted as the cumulative force of mortality, or the number of expected events for each participant by the time t in case the event is a repeatable process [29].

Successively, given any of the three functions $\hat{S}(t)$, $H(t)$, and $h(t)$, the other two can be inferred accordingly as shown here:

$$\begin{aligned} h(t) &= -\frac{d}{dt}[log(\hat{S}(t))], \\ H(t) &= -log(\hat{S}(t)), \\ \hat{S}(t) &= exp(-H(t)). \end{aligned} \tag{2.43}$$

2.5.4 Censoring

In practice, only part of the samples attribute time of the event at the time observation takes place. For the rest of samples, the event of interest is not happened yet. For instance in cancer research, some patients may be still alive at the time the experiment is closed or some patients may have lost the follow-up during study period or some of them may have experienced different events which make further follow-up

impossible. In such cases, the so-called censoring happens for which the accounted survival times underestimate the true, but unknown, time to event. This type of censoring is called **right censoring**. Whereas there is another kind, **left censoring**, which refers to cases in which the actual time of a previous occurrence of a repeatable event of interest (e. g., recurrence after tumor removal surgery) is unknown [29].

To take into account both of the multi- and uni-variate overall survival statistics, Cox proportional hazards model as well as Kaplan-Meier (KM) estimator are applied in this thesis respectively.

2.5.5 Proportional Hazards Model

As a survival analysis model, a proportional hazard model, also referred to as Cox proportional hazard model [32], checks multiple covariates against the chance of occurring a hazardous event by quantifying the effect of unit increases in the covariate as compared to the hazard rate. In the case of treatment response prediction, this can be interpreted as, e. g., taking ^{177}Lu-PSMA could reduce the hazard rate for a patient's death by a half. This holds under the condition that the covariates relate multiplicatively to the hazard rate. Also, the proportional hazards model can cope with both of the binary and continuous types of variables [19]. The proportional hazard model can be considered as a semi-parametric model as it makes a parametric assumption about the effects of the covariates (i. e., the predictors), but no assumptions about the nature of the hazard function $\lambda(t)$ itself [62].

The Cox proportional hazards model, for covariates or parameters vector $X_i = (X_{i1}, \cdots X_{ip})$ (p is the number of parameters) for the i^{th} subject, is defined as:

$$\lambda(t|X_i) = \lambda_0(t) \exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}) = \lambda_0(t) \exp(X_i \cdot \beta), \tag{2.44}$$

which formulates the hazard function at time t for subject i. To solve for β_i this can be thought of as a multivariate regression problem between the predictors X_i and actual event times as target variables.

2.5.6 Kaplan-Meier Estimator

Kaplan-Meier (KM) estimator [53] is a statistical method to estimate the fraction of patients living for a certain amount of time after treatment began. In survival analysis, KM is used to estimate the probability that a patient would survive beyond a certain time point. KM is a univariate survival analysis method which can be applied to continuous as well as categorical variables. In case of a categorical input variable, the KM estimator draws survival curves for each possible value of the variable in question. In case of continuous variables, the variables can be categorized based on fixed or adaptive thresholds based on domain-specific knowledge. For instance, one could categorize the cohort based on the median value of Hemoglobin at the time pre-treatment scan was taken. Figure 2.13 shows an example of KM diagrams.

According to [53], the survival function for a Kaplan-Meier diagram can be defined as:

$$\hat{S}_t = \frac{Num_l(t_0) - Num_d(t)}{Num_l(t_0)}, \tag{2.45}$$

where $Num_l(t_0)$ and $Num_d(t)$ refer to the number of subjects living at start time t_0 and the number of dead subjects at time t respectively.

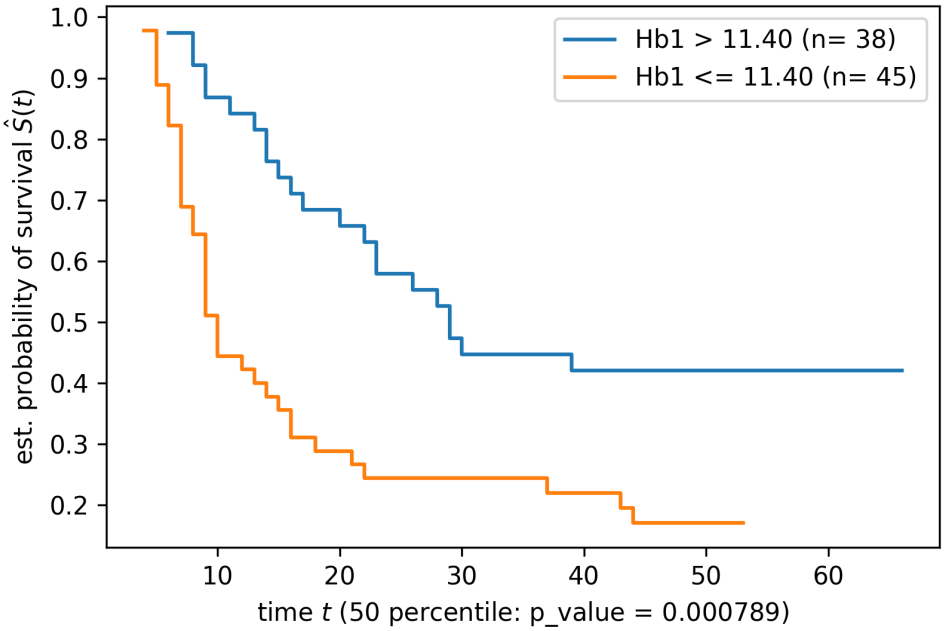


FIGURE 2.13: An example Kaplan-Meier diagram for prediction of overall survival. Here, the cohort is categorized based on the median value of the input variable Hemoglobin on the date that baseline PSMA-PET/CT was taken (Hb1).

As part of our analysis pipeline in [105], we apply Cox proportional hazards model with LASSO method to calculate a radiomics signature for predicting overall survival (OS) of prostate cancer patients. Furthermore, we apply KM estimator to visualize survival functions for some radiomics features and clinical factors. The methodology of this study is further discussed in 4.7.5.

The next chapter, Related Work (3), presents the results of a literature review on the studies and tools related to this thesis subject, highlighting the applications of AI- and ML-based diagnostic and prognostic methods in medical and oncological domains.

Chapter 3

Related Work

Computer-aided diagnosis (CAD) and clinical decision support systems (CDSS) can be considered as general solutions to address the emerging diagnostic and prognostic needs and use-cases in clinical research and routines. In this chapter, results of the literature research based on different aspects of CAD and CDSS will be presented. First of all, different multimodal imaging techniques and their outcoming parameters and radiomics analyses in the literature will be investigated. Moreover, application of machine learning (ML) and statistical methods in diagnosis, treatment planning, and analysis of overall survival (OS) for cancer related diseases will be addressed. Especially, data-driven aspects of such CDSS including processing of image-based as well as clinical factors will be discussed. Section 3.1 presents related work focusing on important aspects and applications of multimodal imaging in clinical studies.

Considering advanced prostate carcinoma as the main focus of this thesis, features and parameters extracted from PET/CT scans are widely used [42, 103, 105, 107]. In general, most of the clinical studies in this category either investigate the impact of conventional parameters such as standardized uptake value (SUV), total lesion glycolysis (TLG), and metabolic tumor volume (MTV) [73, 78, 120] or focus on broader sets of radiomics features (RFs) including textural parameters [76, 107]. Section 3.2 presents research works related to the topic of conventional parameters and RFs respectively.

As for any other cancer examination, the diagnosis of the disease as well as the true identification of malignant uptake is the key for successful treatment for prostate carcinoma at any stage. Thus, focusing on data-driven approaches, most of the studies and clinical decision support tools take advantage of state-of-the-art machine learning (ML) methods. The so-called supervised learning and deep learning methods are commonly used for this purpose [107, 142, 166, 167]. In addition to diagnosis, statistical methods in general and ML methods in particular facilitate treatment planning and analysis of overall survival (OS) by their predictive outcomes based on RFs and patient-specific clinical parameters. Section 3.3 provides a more detailed overview on state-of-the-art statistical and ML methods to address the problems of diagnosis, treatment planning, and analysis of OS.

As will be described in the next sections, CAD and CDSS have several aspects which could be referred to as consecutive steps of a pipeline from the acquisition of medical images to the prediction of overall outcomes of the treatment alternatives. While many studies and software tools focused on the individual steps such as segmentation of tumors [46, 63, 164] or prediction of treatment response [14, 103], some research and development attempts addressed multiple steps of such a decision support pipeline [134, 152, 166].

3.1 Multimodal Imaging

Multimodal imaging has been in focus of nuclear medicine (NM) physicians, oncologists, and neuroscientists for disorders such as advanced prostate carcinoma, lung cancer, and neurodegenerative diseases [21, 76, 107, 139, 142].

In the special case of cancer analysis, multimodal imaging plays an important role. A wide variety of imaging modalities including positron emission tomography (PET), computed tomography (CT), and magnetic resonance imaging (MRI) are used as the essential means for diagnosis and therapy outcome assessment. For the examination of cancer patients, usually, both the metabolic uptake in the tissues as well as the anatomical structure of the lesions and hotspots are of interest. To this end, PET imaging is widely used alongside CT or MRI scan. Thus, the two modalities complement each other by providing both the metabolic as well as the anatomical and spatial context to the reader. Foster et al. [46] reviewed the applications and challenges of integrating multimodal PET/CT scans in clinical practice.

To achieve diagnoses with higher precision, accurate co-registration and resampling of the two modalities is critical. Moreover, the true segmentation of the malignant tissues is of great importance. Thus, it is favorable to support physicians by providing CDSS solutions facilitating precise segmentation as non-invasive aids. Sections 3.1.1 and 3.1.2 present insights about common techniques for resampling and segmentation of multimodal scans.

Another important aspect of medical imaging for oncology is the choice of radiotracers for different cancer disorders. Some radiotracers are specific to a special type of disease (e. g., prostate-specific membrane antigen (PSMA) for prostate cancer (PCa)) [42, 103, 105], while some tracers are more widely used (e. g., fluorodeoxyglucose (FDG) for lung, esophageal, or rectal cancer) [11, 14, 21, 84]. Section 3.1.3 gives an overview of the application of common radiotracers for cancer diagnosis and treatment.

3.1.1 Resampling

One of the most important aspects of multimodal imaging is different resolutions of the scanners. For example, a PET/CT scanner may have different resolutions for PET and CT. Moreover, due to different cameras, different modalities are usually captured in different coordinate spaces. Thus, for further analyses, it is critical to take the data from the different modalities and transform them to a common space. This transformation process in which a raster object (e. g., an image) is re-scaled or projected to a different coordinate aystem is called resampling [123], should be applied prior to any visual assessment or segmentation effort.

Co-registration of multimodal images has been the focus in many medical physics studies [2, 20, 71, 128]. Depending on the degree of freedom between the original spaces, a wide range of simple and sophisticated algorithms are proposed. For instance, if the two modalities only differ in scale or rotation, appropriate affine transformations could bring one modality to the common space as for the other modality. However, often more sophisticated mappings such as deformations and body motions due to respiratory movements should be taken care of. Brock et al. [20] reported common techniques for image registration in radiotherapy. In another study, Jin et al. [71] applied a method based on gradient of mutual information (GMI) to co-register FDG-PET and CT images for esophageal cancer patients. Also, Pépin et al. [128] presented state-of-the-art methods for management of respiratory

motion in PET/CT scans. Furthermore, Abderrahim et al. [2] conducted a comparative study of relevant methods for multimodal image registration, focusing on brain scans.

Most medical imaging formats such as digital imaging and communications in medicine (Dicom) [121] and neuroimaging informatics technology initiative (NIfTI) [116] store affine transformation matrices which facilitate the mapping from voxel coordinates to world coordinates. These affine transformation matrices are used to bring different image modalities (e. g., PET and CT) to the same coordinate space. Many studies [82, 139, 144] focus on affine transformation for resampling and visualization purposes. For example, Lemare et al. [82] apply rigid body transformation to correct respiratory movement artefacts in PET images. Moreover, Shan et al. [139] conducted a retrospective evaluation of registration algorithms for PET/MRI registration, including affine transformations. More interestingly, Steffen et al. [144] analysed the influence of rigid co-registration of PET and CT data and recommended sticking to original PET data to extract parameters such as MTV and SUV_{MAX}. In our methods, we take advantage of affine transformation matrices stored in the Dicom headers of the input images to re-scale the PET/CT data.

3.1.2 Segmentation

Image segmentation is usually done in two related steps: recognition and delineation. Recognition corresponds to *where* the relevant object is located, while delineation denotes how to define the region of the object in the coordination space. For the task of computer-based recognition and delineation of benign or malignant tissues, there are a bunch of software tools and algorithms available, ranging from purely manual to fully automated [45, 96, 98, 148]. These tools and software mostly facilitate delineation of the desired tissues as 2 dimensional (2D) regions of interest (RoIs) or 3 dimensional (3D) volumes of interest (VoIs). Foster et al. [46] present a summary of segmentation techniques for PET/CT scans.

Manual Segmentation

Considering the ethical precautions for clinical human trials which recommend avoiding invasive methods, manually delineated tumor uptake using software tools [96, 98, 148] is commonly used as gold standard or surrogate truth and as ground truth for supervised ML-based methods. Many studies [11, 14, 21, 42, 84, 103, 105, 107] took advantage of available segmentation tools and techniques for manual delineation of the RoIs or VoIs. Reuzé et al. [130] provide a comprehensive review of these studies. For instance, InterView FUSION (Mediso, Budapest, Hungary) [98] is used in clinical routines for the task of segmentation of the pathological hotspots in PET/CT scans. It also calculates a set of RFs including first and higher order statistics as well as textural heterogeneity parameters which can be used for diagnostic and prognostic purposes [21, 42, 84, 103, 105, 107].

As an example of studies which leveraged manual segmentation, Bundschuh et al. [21] assessed the outcome of RFs from baseline fluorodeoxyglucose (FDG)-PET for the prediction of histopathologic response and overall survival (OS) for patients with rectal cancer and compared it to predictive outcome of conventional parameters. Also Lapa et al. [84] have shown that contrast and gray-level non-uniformity (GLNU) extracted from pre-therapeutic ^{68}Ga-PET scans are significantly correlated with progression free survival (PFS) in thyroid cancer patients. In our clinical studies, we analyzed RFs from baseline ^{68}Ga-PSMA-PET/CT scans to detect pathological

uptake [42, 107], to predict responders to ^{177}Lu-PSMA therapy [103], and to analyze overall survival of the patients with advanced prostate carcinoma [105], taking advantage of InterView FUSION to delineate malignant uptake and to provide RF vectors.

Similar to InterView FUSION [98], MaZda [148] and MathWorks [96] provide similar diagnostic tools for physicians. For example, Bang et al. [11] used MaZda to analyze performance of MTV, TLG, and textural parameters from baseline FDG-PET to predict the chemoradiation response for rectal cancer. Also, Beukinga et al. [14] used MathWorks to assess textural features from FDG-PET/CT scans for the task of predicting complete vs incomplete response to neoadjuvant chemoradiotherapy in esophageal cancer.

Thresholding-Based Segmentation

Thresholding is widely applied in both uni- and multimodal medical imaging for segmentation purposes [27, 40, 65, 73, 78, 79, 113, 120, 136]. Regardless of the imaging modality or the ultimate segmentation objectives, thresholding denotes the process of converting gray-scale images to binary format, distinguishing foreground from background in 2D or 3D setups [93, 126, 133]. In general, there are three types of thresholding in image processing: fixed, adaptive, and iterative [46].

In PET/CT imaging, fixed thresholding refers to the process that assigns a predefined threshold (mostly based on SUV levels of PET images or Hounsfield scale of CT images) to distinguish foreground from background. Adaptive thresholding, on the other hand, often takes a second modality (mostly CT) to spatially locate the uptake from PET as a reference and personalized patient information such as body weight and size. As an example of adaptive thresholding methods, Yaremko et al. [164] used a phantom based model to mimic the respiratory and cardiac movements which showed to be beneficial for segmentation of lung tumors.

On the contrary to adaptive thresholding, iterative approaches do not need prior spatial knowledge about volumes of interest (VoI). However, they require the source-to-background ratio (SBR) and the reconstruction algorithm information of the scanner as well as the radiotracer [8]. Baazaoui et al. [8] mentioned several drawbacks (including limited spatial resolution, inability to cope with heterogeneous volumes, and inaccurate registration of the resulting region) for some adaptive and iterative thresholding based segmentation techniques such as adaptive thresholding based on spherical phantoms [16], the iterative threshold method (ITM) [70], and a Monte Carlo (MC) algorithm-based method [114].

PET images are often segmented based on predefined thresholds of conventional parameters such as SUV, TLG, and MTV [73, 78, 120]. For instance, Pak et al. [120] analyzed the prognostic significance of SUV, MTV, and TLG from FDG-PET/CT scans for patients suffering from extranodal nasal type natural killer T (NK/T) cell lymphoma. In another study, Jun et al. [73] used fixed SUV of 2.5 and 40%-SUV_{MAX} as predefined thresholds for the segmentation of FDG-PET tumors in patients with esophageal squamous cell carcinoma. Zaidi et al. [168] compared performance of several segmentation techniques such as adaptive thresholding for the task of segmenting FDG uptake in PET scans in a cohort of pharyngolaryngeal squamous cell carcinoma patients.

CT images on the other hand are segmented in a multilevel approach based on Hounsfield unit (HU) scales for different tissue types. Taking advantage of HU, CT scans are widely used in many clinical studies for tissue segmentation and for diagnostic purposes [113, 136]. In particular, bone segmentation relies significantly

on thresholding techniques leveraging HU from CT images. For example, Scheyerer et al. [136] used HU as a measure of bone density, focusing on its benefits in spine surgery. Also, Narayanan et al. [113] identified the role of lower bone mineral density (BMD) calculated by HU as a significant factor in patients with spontaneous femoral fractures.

Considering the time-consuming nature of manual or semi-automated techniques for bone segmentation, Klein et al. [79] applied convolutional neural networks (CNN) as a fully automated approach for the segmentation of bone in whole-body CT scans. Taghizadeh et al. [149] presented an automated method for CT bone segmentation based on statistical shape modeling and local template matching, focusing on its benefits for personalized surgical instruments and for manufacturing of patient-specific implants. In another study aimed at facilitating treatment of traumatic pelvic injury, Vasilache et al. [156] took advantage of a series of automated region growing algorithms and gradient based segmentation techniques for the segmentation of bone in CT scans.

As another application of HU, Hebb et al. [65] used an extended HU scale to investigate deep brain stimulation (DBS) for the treatment of patients with Parkinson's disease. In another study, Choi et al. [27] used HU scales of CT scans to define and distinguish between benign and malignant RoIs in soft tissue for solitary pulmonary nodules (SPN) in a semi-automated way. More interestingly, Gafita et al. [48] introduced qPSMA, a software package which applies a combination of fixed and adaptive thresholding techniques based on SUV levels from PET and HU values from CT, to calculate whole-body tumor burden based on ^{68}Ga-PSMA-PET/CT for a cohort of patients with prostate cancer (PCa). qPSMA is shown to be a useful semi-automated tool for the quantification of the tumor load in heavily metastasized PCa patients.

Automated Segmentation

Most of the manual and semi-automated segmentation techniques require intensive and time-consuming human interaction. Therefore, providing fast and accurate automated segmentation tools is one of the most important objectives of CAD and CDSSs. Some conventional automated segmentation methods include Gaussian mixture models (GMM), region growing, and graph-based methods [4, 10, 37, 63]. Furthermore, ML-based algorithms, including clustering and artificial neural networks (ANN), have got growing attention in the field of medical image segmentation for different clinical studies [47, 55, 69, 86]. Alternatively, (semi-) automated medical imaging segmentation techniques can be categorized based on the scope of the desired volumes or regions of interest. Hence, some methods and tools focus on desired uptakes or structures in specific body organs [45, 79], while other methods aim at segmenting VoIs or RoIs as spread along the whole-body [10, 69, 86].

As an example of conventional automated segmentation methods, Hatt et al. [63] proposed an approach named fuzzy locally adaptive Bayesian (FLAB), a method based on GMM, as an unsupervised statistical method which performs especially well for delineation of lesions smaller than 2 centimeter and hotspots with inhomogeneous activity distributions. Another conventional method for (semi-) automated segmentation is region growing which was proposed first time by Adams and Bischof [4]. The originally proposed region growing method was dependent on user input to specify some seeds for the algorithm to start. As an example of its application, Day et al. [37] applied a method named confidence connected region growing (CCRG) for tumor volume segmentation on FDG-PET scans for patients

with rectal and anal cancer. CCRG is an iterative method which relies on the mean and standard deviation values of the intensities of the regions, initialized by a sub-region surrounding the pixel with maximum intensity.

As an instance for the application of region growing algorithms for whole-body tumor segmentation, Ballangan et al. [10] proposed a method combining K-means clustering [72] with downhill region growing (DRG) algorithm and decision trees for the purpose of tumor segmentation and hotspot classification in whole-body PET scans for patients suffering from non-small cell lung cancer (NSCLC).

Focusing on brain MRI scans, Fischl et al. [45] introduced FreeSurfer, a set of tools aimed at segmenting the whole brain as well as the subcortical structures including white and gray matter, using a variety of surface based, location based, and isotropic algorithms. For example, to account for within-surface heterogeneity, FreeSurfer applies separately different Gaussian models for different structures and for different points in space. Furthermore for the image prior term, FreeSurfer applies an enhanced version of markov random field (MRF) model which takes the spatial connection between some brain substructures (e. g., amygdala locates always in front of and above hippocampus and never behind or below it) into account. FreeSurfer is widely used in clinical routine as well as computational neuroscience studies for the analyses of oncological or neurodegenerative diseases, mostly as a reference method for brain segmentation [75, 88, 138].

As an ML-based whole-body approach, Jemaa et al. [69] applied a convolutional neural network (CNN) for the segmentation of tumors in whole-body FDG-PET/CT scans in a multicentric setup for patients with non-Hodgkin's lymphoma (NHL) and advanced NSCLC. For the evaluation of their methods, they took advantage of dice coefficients. In another study, Lavdas et al. [86] applied three different automated approaches for segmentation of oncological lesions in whole-body MRI scans. They developed and compared classification forests, a 3D CNN and a multi-atlas based on overlap (dice scores) and surface distance metrics.

As ML based approaches, artificial neural networks (ANNs) and deep learning architectures have been used for many years for automated image segmentation and classification. In recent years, ANNs have become more popular in medical image analysis and segmentation as well. In particular, the U-Net model [132] is widely used in many tools and studies facilitating diagnosis and treatment outcome assessment pipelines [47, 55]. The U-Net architecture was first proposed by Ronneberger et al. [132] in 2015 as a CNN with a symmetric structure of encoding and decoding for biomedical image segmentation. U-Net relies on very few images to cope with the lack of ground truth labelled images. To this end, U-Net facilitates the training of the neural network by incorporating data augmentation [132]. Data augmentation in image processing denotes the process of generating augmented image samples based on a few input images by applying a variety of geometric transformations, color space and feature space augmentations, as well as generative adversarial networks (GANs). Liu et al. [90] conducted a survey on U-shaped networks used for medical image segmentation and Shorten et al. [141] compared some of the data augmentation techniques mentioned here.

Gadosey et al. [47] presented SD-UNet, a U-Net based neural network aiming at utilization of biomedical image segmentation on platforms with low computational power. SD-UNet takes advantage of depth-wise separable convolutions in the entire network to build a lightweight deep CNN inspired by U-Net architecture which outperforms the original U-Net in terms of model size and computation time [47]. In another study, Gorgizadeh et al. [55] developed and compared three U-Net-based models for the segmentation of drusen in optical coherence tomography (OCT) scans

to diagnose and track the progression of age-related macular degeneration (AMD) disease. They leveraged CNN-based direct segmentation and combined it with more complex application-specific post processing methods, including a layer-based approach and a method called retinal pigment epithelium (RPE)+drusen complex (RPEDC) [26], for OCT drusen segmentation. Their proposed models outperformed a previous state-of-the-art method for automated drusen segmentation presented by Chen et al. [24] which had used a priori knowledge of normal retinal morphology and anatomical features as well as a method for retinal projection.

3.1.3 Radiotracers and Receptors

Different radiotracers such as fluorodeoxyglucose (FDG) and prostate-specific membrane antigen (PSMA) and receptors such as somatostatin receptors (SSTRs) are widely used for diagnosing and screening of different cancer diseases. For instance, FDG is widely used in PET examinations as a multipurpose radiotracer for different kinds of oncological studies [11, 14, 21]. Bundschuh et al. [21] took the changes in glucose uptake and tumor heterogeneity from FDG-PET/CT findings as a reference of treatment response to investigate the prognostic significance of textural parameters in patients with rectal cancer. In another study focusing on rectal cancer patients, Bang et al. [11] evaluated the power of pre-treatment FDG-PET/CT scans for the prediction of neoadjuvant radiation chemotherapy and 3-year disease free survival (DFS). They took both conventional parameters such as SUV and MTV as well as 3D textural features into account, denoting the association of mean SUV with tumour regression grading (TRG) and that of kurtosis of the absolute gradient (GrKurtosis) with 3-year DFS. Also focusing on neoadjuvant radiation chemotherapy, Beukinga et al. [14] investigated the predictive significance of pre-therapeutic FDG-PET/CT findings in patients with locally advanced esophageal cancer, revealing the outperforming power of clinical parameters such as histologic type and clinical T stage as well as both PET- and CT-based textural parameters in comparison to SUV measures.

As an example of common receptors, some studies took advantage of SSTRs for treatment response assessment [23, 84, 146]. Sun et al. [146] reviewed the application of SSTR for various cancer types, denoting the distinctive receptive characteristics of tumor cells and normal tissues when exposed to somatostatin. These characteristics lead to higher receptive levels in malignant tissues, as a result, various cytotoxic SST conjugates can be developed to target tumors. In a retrospective study, Lapa et al. [84] investigated the role of PET-based heterogeneity parameters to predict the results of peptide receptor radionuclide therapy (PRRT) as a treatment option which requires over-expression of subtype 2 SSTR in patients with thyroid cancer. They identified the power of grey level non uniformity, contrast, and entropy as textural parameters compared to that of conventional PET parameters for the prediction of progression free survival (PFS). However, their findings revealed that all of the investigated parameters would fail to predict overall survival (OS). In a review of lung cancer related literature, Callison et al. [23] assessed the significance of somatostatin receptors for the diagnosis of the disease as well as the application of type 2 SSTRs in targeted chemotherapy and radiotherapy.

In contrast to FDG or SSTR which are applied for examinations in various cancer diseases, prostate-specific membrane antigen (PSMA), as its name suggests, is especially used for detection of different stages of prostate cancer and for screening patients for certain types of treatment. As PSMA-PET/CT usually is attributed with high sensitivity, it facilitates the stratification of patients in primary staging of PCa

for systemic therapy planning or for surgery, by exclusion or detection of metastases [97, 125]. Petersen et al. [125] conducted a systematic review of clinical studies and denoted the high diagnostic accuracy of PSMA-PET for the staging of primary lymph node metastasis in prostate cancer patients with intermediate or high risks. In another multicentric study, Mattiolli et al. [97] illustrated the impact of conventional parameters from ^{68}Ga-PSMA-PET/CT findings on treatment planning of patients with biomedical recurrence.

In the next chapters of this document, we mostly focus on PSMA, as the main objective of the thesis was to provide CDSS to facilitate prostate cancer diagnosis and treatment.

3.2 Feature and Parameter Groups

Conventionally, the impact of intensity related parameters such as mean and maximum SUV as well as metabolic parameters such as MTV and TLG has been in focus of clinical studies aiming at staging and patient screening. More recently, advances in computational power made it feasible to process more computationally complex textural parameters as quantitative image analyses. In the context of medical image processing, radiomics describes the procedure in which quantitative features are extracted from medical imaging modalities such as PET, CT, and MRI [7]. Although some consistency challenges including varying acquisition and reconstruction techniques and lack of standardization exist, radiomics has already shown its potential to be a valuable asset to facilitate personalized treatment leveraging image-based analyses [85]. CAD tools such as InterView FUSION [98] facilitate PET/CT RoI and VoI delineation and calculate SUV metrics as well as radiomics features (RFs) corresponding to each VoI. Furthermore, some open source software libraries such as PyRadimics [56] provide functionalities to calculate hundreds of RFs in development pipelines leveraging Python programming.

In this section, some related work covering the topics of conventional parameters, radiomics, and deep features in clinical research are mentioned.

3.2.1 Conventional Parameters

Standardized uptake value (SUV) is one of the most commonly used metrics in clinical routines and studies for the diagnosis and treatment followup. SUV metrics obtained from both FDG- and PSMA-PET scans facilitate the assessment of patient response to cancer therapy, removing variability artefacts exposed by different patient sizes and various amounts of injection doses [78]. Most often, SUV is analyzed together with other metabolic uptake metrics such as metabolic tumor volume (MTV) and total lesion glycolysis (TLG) [73, 120]. Pak et al. [120] denoted the significant performance of SUV_{Mean}, SUV_{Max}, MTV, and TLG for the prediction of relapse free survival (RFS) in patients diagnosed with extranodal nasal type NK/T cell lymphoma. As discussed earlier, metabolic uptake is also used in thresholding-based segmentation methods [73], treatment planning for recurrent disease [97], and prediction of disease free survival (DFS) [11]. In most of these studies, either the importance of conventional parameters is shown or they are considered as a reference to compare with radiomics and textural parameters [106].

3.2.2 Radiomics

Many oncological studies analyzed the role of RFs in general and textural heterogeneity parameters in particular in detail. While some studies [14, 21, 76, 94] mostly focused on a limited subset of textural parameters such as coefficient of variance (COV), homogeneity, skewness, kurtosis, and entropy to asses their significance on study outcomes individually, other studies [42, 103, 105, 107] focused on biggers sets of RFs and apply machine learning as well as feature selection techniques to identify most discriminative features among input feature vectors. Often, clinical studies applying RFs with ML methods suffer from relatively small subject cohorts. Especially in supervised learning, when the number of subjects in the training cohort is fewer than the size of the input feature vector, there would be a higher chance of overfitting which affects the generalizability of the analyses [122]. Park et al. [122] reviewed the challenges and solutions regarding statistical aspects of reproducibility and generalizability in radiomics modelling. To cope with overfitting, in this thesis we mostly focus on feature selection techniques including feature ranking based on feature importance metrics of classifiers [18, 57], recursive feature elimination (RFE) [58], and least absolute shrinkage and selection operator (LASSO) [81].

As described earlier, Beukinga et al. [14] showed the higher predictive power of RFs from pre-therapeutic FDG-PET/CT scans to that of SUV-based metrics for a cohort of locally advanced esophageal cancer patients. Bundschuh et al. [21], Khurshid et al. [76], and Lv et al. [94] analyzed the performance of textural heterogeneity parameters for prediction of treatment response in patients with rectal, prostate, and nasopharyngeal cancer respectively. Our findings [42, 103, 105, 107] also addressed the role of RFs extracted from pre-therapeutic ^{68}Ga-PSMA-PET/CT images for the classification of pathological uptake [42, 107], for the prediction of treatment response [103], and for the analysis of overall survival (OS) [105]. We took advantage of RFs provided by InterView FUSION and PyRadiomics for different purposes. For instance from InterView FUSION, we used first and higher order statistics features (mean, max, kurtosis, etc.), shape-based features (max diameter and volume), textural features (entropy, contrast, homogeneity, etc.), and volumetric zone and run length statistics (grey-level non-uniformity, short run emphasis, etc.) and applied supervised ML methods to identify pathological uptake in PSMA-PET/CT scans [107] and to predict responders to ^{177}Lu-PSMA therapy [103]. In addition, as recently published [106], we use features calculated by PyRadiomics library to evaluate our in-house developed U-Net-based segmentation which will be discussed in more detail in 4.3.2.

3.2.3 Deep Features

As described in 2.4.1, alongside radiomics features, hundreds of features can be extracted from encoding layers of deep and convolutional neural networks. Some studies investigated the role of deep features for diagnosis and prognosis [6, 83, 115, 124]. Paul et al. [124] compared different compositions of deep features from CT scans fed to a CNN model with a set of radiomics features for the early detection of lung cancer nodules and showed that the combination of deep features and classic radiomics features improved predictive performance of their model. Also, Andrearczyk et al. proposed a multi-task deep model to segment the gross primary tumor volume in FDG-PET/CT scans and predict survival of patients suffering from head and neck cancer. They analyzed the relative importance of deep features and radiomics features for survival prediction and showed that the combined approach

outperformed both the deep radiomics method without segmentation and the standard radiomics model [6]. In another study, Lao et al. [83] investigated and marked the role of radiomics signatures generated from deep features from MRI scans for patients with glioblastoma multiforme (GLM) using LASSO Cox regression model to predict overall survival. Nie et al. proposed a two-staged ML based method to predict OS of patients diagnosed with high-grade gliomas based on multimodal conventional clinical and functional connectivity factors as well as deep features extracted from contrast-enhanced T1 MRI scans, diffusion tensor imaging (DTI), and resting-state functional MRI (rs-fMRI). They further fit an SVM classifier to predict groups of patients with different survival risks with high accuracy [115]. In this thesis, we focused on radiomics features for diagnostic and prognostic tasks [42, 103, 105–107]; however, investigating the role of the deep features for the same analyses is an important topic which should be considered as a track of future work.

In the next sections of this chapter, we consider the role of both conventional parameters and RFs as applied with AI-based methods for (1) the detection of malignant tissues, (2) the prediction of treatment response, and (3) for the analysis of OS.

3.3 Artificial Intelligence Based Solutions

Artificial intelligence (AI) in general, and Machine learning (ML) in particular have been successfully applied in a broad spectrum of domains ranging from image processing, object detection, and outlier detection. Most of the common ML techniques, including supervised and deep learning based methods, have gained critical importance in clinical studies to address problems including but not limited to diagnosis and treatment planning [25, 42, 103, 107, 142]. Thus, considering the wide range of available algorithms and tools, choosing the best ML-based approach for the specific clinical use-case will be vital. On the one hand, many solutions [42, 103, 107] rely on manual segmentation of tumors and make the use of third party CAD tools such as InterView FUSION to calculate (radiomics) feature sets to fit the supervised ML models. On the other hand, deep neural networks support CAD developers and researchers by combining tumor segmentation with (radiomics) feature calculation and often with patient-specific treatment response prediction.

Sollini et al. [142] reviewed AI and ML-based techniques applied with hybrid medical imaging for personalized medicine in oncology, focusing on lung cancer patients. Moreover, Shen et al. [140] focused on deep learning and introduced and reviewed its challenges and potentials for medical image analysis. In another study, Cheng et al. [25] evaluated some deep learning solutions based on the so-called stacked denoising auto-encoder (SDAE) for the classification of malignant tissues from CT scans for patients with lesions in breast area.

The next subsections mention some related work focusing on the application of ML methods together with radiomics image analytics for the topics of diagnosis, treatment response prediction, and analysis of overall survival.

3.3.1 Diagnosis

Supervised ML methods and CNNs are widely used in combination with radiomics in clinical research for diagnosis of cancer disease and delineation of tumors. For instance focusing on bladder cancer management, Ge et al. [51] reviewed the potential of combination of radiomics features from different imaging modalities with ML

for the classification of malignant tissues as well as therapy response prediction for bladder cancer patients. Garapati et al. [50] also compared different ML classifiers such as SVM and random forests (RAF) [18] as applied to RFs extracted from CT scans for the staging of patients diagnosed with urinary bladder cancer. In another study, Gao et al. [49] applied SVM and RAF to classify different tumor grades and pathologic biomarkers of glioma based on RFs extracted from MRI scans.

Given the example of prostate cancer (PCa), Hambarde et al. [60] analyzed radiomics variables from MRI scans to detect and segment pathological uptake associated with PCa, leveraging a 2D U-Net CNN. In another study aiming at detection of PCa from MRI scans, Yoo et al. [166] combined 2D CNNs with random forests to extract first order RFs and to predict patient level diagnoses respectively. Furthermore as a 3D CNN based solution, Liu et al. [91] presented XmasNet for the diagnosis of PCa from multiparametric MRI scans, making the use of data augmentation. As described before, our findings [42, 107] revealed the potential of SVM classifier with different kernels (e. g., linear, polynomial, and radial basis function (RBF)) and decision trees as applied to RFs taken from PSMA-PET/CT for the classification of pathological lesions in patients with advanced prostate carcinoma. We also present a U-Net based method for the segmentation of pathologic uptake (see chapter 4).

3.3.2 Treatment Response Prediction

Combination of ML methods with radiomics also facilitates the prediction of treatment response in oncology. For example, Vallières et al. [154] applied random forests to radiomics features (RFs) from FDG-PET/CT scans for the risk assessment of locoregional recurrence (LR) and distant metastases (DM) in patients diagnosed with head and neck cancer. In another study, Ypsilantis et al. [167] leveraged a convolutional neural network with RFs from FDG-PET scans to predict responders to neoadjuvant chemotherapy in patients with esophageal cancer. As an application of SVMs for prediction of response to carbon ion radiotherapy, Wu et al. [161] analyzed RFs from pre-treatment MRI scans for the screening of patients with PCa. Also aiming at predicting response to intensity-modulated radiation therapy and staging of PCa from MRI findings, Abdollahi et al. [3] took advantage of uni-variate t-tests [77] to identify most significant RFs as well as supervised ML methods for classification purposes. To analyze the objective risk stratification of PCa, Varghese et al. [155] compared seven ML classifiers including linear and Gaussian kernel SVM, logistic regression, and random forests as applied to RFs from multi-parametric MRI scans. In a previous work [103], we applied linear regression to identify most significant radiomics variables from PSMA-PET and CT modalities to correlate with ΔPSA in pre- and post therapy as an indicator of treatment response. Consecutively, we applied supervised ML classifiers (e. g., SVM, RAF, and Extra Trees) to predict responders to ^{177}Lu-PSMA treatment with high AUC.

3.3.3 Analysis of Overall Survival

As conventional metrics of overall survival (OS), variables such as metabolic tumor volume (MTV), total lesion Glycolysis (TLG), and mean/max standardized uptake value (SUV) have been in focus in oncology for many years [110, 112, 120]. Moon et al. [110] reviewed the application of PET-based uptake in analysis of OS for different cancer diseases. For instance, Pak et al. [120] analyzed SUV_{Mean}, SUV_{Max}, MTV, and TLG to predict relapse free survival (RFS) in patients with extranodal nasal type

NK/T cell lymphoma. Moreover, Nappi et al. [112] assessed conventional parameters from fluorodeoxyglucose (FDG)-PET/CT for OS analysis in patients with non-small-cell lung cancer (NSCLC).

As an alternative to conventional metrics, radiomics features (RFs) are also widely used for OS analyses in oncology. To deal with overfitting caused by fewer numbers of subjects than features in the field of radiomics, the so-called term *radiomics signature (RS)* is introduced. The common method to calculate the RS includes a step-wise feature selection followed by least absolute shrinkage and selection operator (LASSO) also known as L1 regularization [81]. Consecutively, Kaplan-Meier (KM) estimator is used to determine whether the calculated RS (or other parameters) can discriminate between high- and low-risk patients in terms of OS. This approach is applied in a bunch of studies for MRI [17, 83, 92, 169], FDG-PET/CT for lung cancer [99], and somatostatin receptor subtype II (SSTR) PET for thyroid cancer [84]. In a review article focusing on supervised and deep learning (DL) techniques, Vial et al. [158] outlined some studies that applied RFs for the analysis of OS of cancer patients.

Lao et al. [83] leveraged deep learning methods to extract RFs and RS from manually segmented lesions from MRI scans and assessed their performance to predict OS in patients with glioblastoma multiforme (GBM) brain tumors. They took into account both handcrafted features (e. g., intensity and textural features) from original images as well as deep features emerged from the CNN. In a study aimed at facilitating post-operative management of high-risk PCa, Bourbonne et al. [17] applied Cox proportional hazards (CPH) model and KM estimator to analyze significantly correlating MRI-derived RFs with biomedical recurrence (BCR) as a survival metric. Furthermore, our findings [105] also revealed the potential of PSMA-PET/CT based RFs and RS for the prediction of OS of PCa patients when treated with ^{177}Lu-PSMA, leveraging Cox and KM as multi- and uni-variate analysis methods respectively.

The next chapter presents details of the methods which have been integrated or implemented to realize the objectives of this thesis.

Chapter 4

Methodology

As described in the previous chapters, in this thesis we aimed at providing a fully automated pipeline for the management of prostate cancer (PCa) patients based on ^{68}Ga-PSMA-PET/CT findings and patient-specific clinical parameters. We focused on ^{68}Ga-PSMA tracer which is commonly used in PET examinations, mostly to locate prostate malignancies. As the automated pipeline is mostly developed using Python programming language and its open source libraries, we named it *AutoPyPetCt*. AutoPyPetCt consists of several modules including multimodal PET/CT resampling and visualization tool box, U-Net based segmentation (called PET-CT-U-Net), radiomics feature extraction and selection, as well as diagnosis, prognosis, and survival analysis tool boxes which together serve as a clinical decision support system (CDSS) for PCa patients management. In this chapter, the methodology of the solutions towards such an automated pipeline, as implemented or applied, will be described. First, an overview of the whole pipeline is presented from different points of view, including a high-level outline of the solutions as well as module-based and process-based outlines. Then, each module will be described in more detail, explaining its corresponding individual sub-processes. Moreover, in section 4.7, the methodology of the published works related to this thesis will be described in detail.

4.1 Methods Overview

Figure 4.1 illustrates the high-level outline of the materials and methods, including in-house developed software tools as well as third-party tools and libraries. The whole pipeline is represented as three consecutive building blocks: 1) segmentation and annotation, 2) processing, and 3) analyses. The segmentation and annotation block consists of various assets facilitating manual and automated segmentation and annotation of multimodal PET/CT images. For instance, InterView FUSION [98] is used for manual delineation of hotspots, while an in-house developed deep segmentation network based on U-Net [132] is used for automated segmentation of pathological uptake. The processing block consists of feature extraction and selection tools, either provided by InterView FUSION or implemented leveraging PyRadiomics library [56], as well as processing of patient-specific clinical data. Finally, the analyses block covers the tools and diagrams aiming at representing the end-user level analyses and insights.

Focusing on the in-house developed tools, figure 4.2 presents the main modules of AutoPyPetCt as a modular automated tool box. This modular representation denotes the use-case based overview of the methods which are mostly implemented in Python. Apart from the Python CORE which represents the basic and third-party

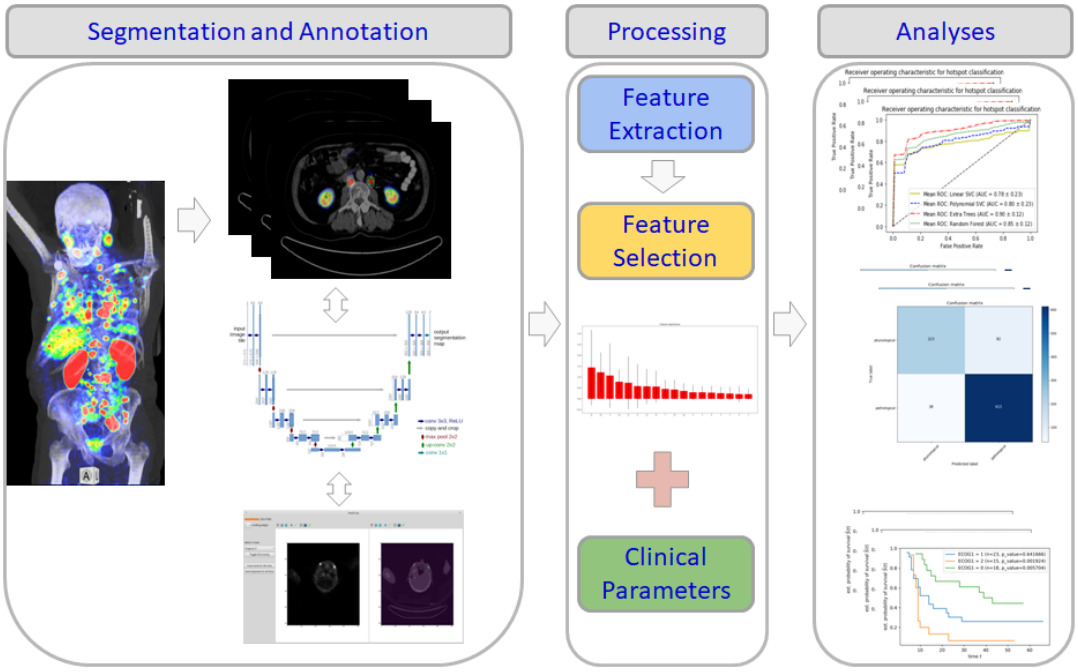


FIGURE 4.1: The high-level outline of the methods.

Python libraries, the main modules are 1) VIEW: the multimodal PET/CT visualization tool box, 2) USEG: the multi-channel U-Net based segmentation unit, also called PET-CT-U-Net, 3) DX: the diagnosis analyses unit, 4) PX: the prognosis analyses unit, and 5) SURV: the survival analyses unit.

Figure 4.3 shows an overview of the processes associated with each of the above-mentioned modules from the implementation point of view. For example, the input data are first processed and structurally organized to make them compatible with visualization and resampling processes. Then, the segmentation tool box applies U-Net based segmentation to identify pathological uptake and calculate radiomics features using PyRadiomics library. Last but not least, diagnosis, prognosis, and survival analyses units process the calculated radiomics features to end up with final intuitions, leveraging supervised machine learning (ML) and feature ranking methods as well as survival statistics.

The details of the integration of the external libraries or the implementation of the in-house developed tools will be described in the next sections of this chapter.

4.2 Visualization

As a predominant visualization and annotation software, InterView FUSION is used in clinical practice at our department. Thus, as a part of our studies, we took advantage of InterView FUSION for labeling PET/CT data as ground truth (GT). Furthermore, to develop an in-house developed software for the visualization of the multimodal PET/CT scans, Python V.2.7 and its corresponding libraries such as Matplotlib [68], PyQt [145] V.4.0, and pydicom [95] have been used.

As described before, resampling of the original imaging modalities and bringing them into the same coordination space is a vital step prior to the visualization of multimodal images. In our case, we used the metadata stored in the Dicom [121] headers of PET and CT images to retrieve the corresponding 3D affine transformation matrices and then to co-align their coordination spaces.

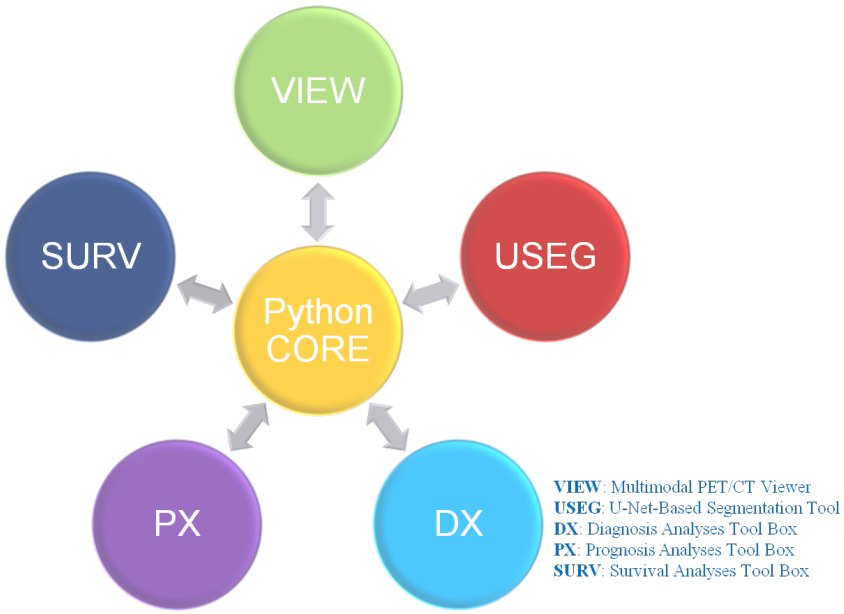


FIGURE 4.2: AutoPyPetCt: Modules overview

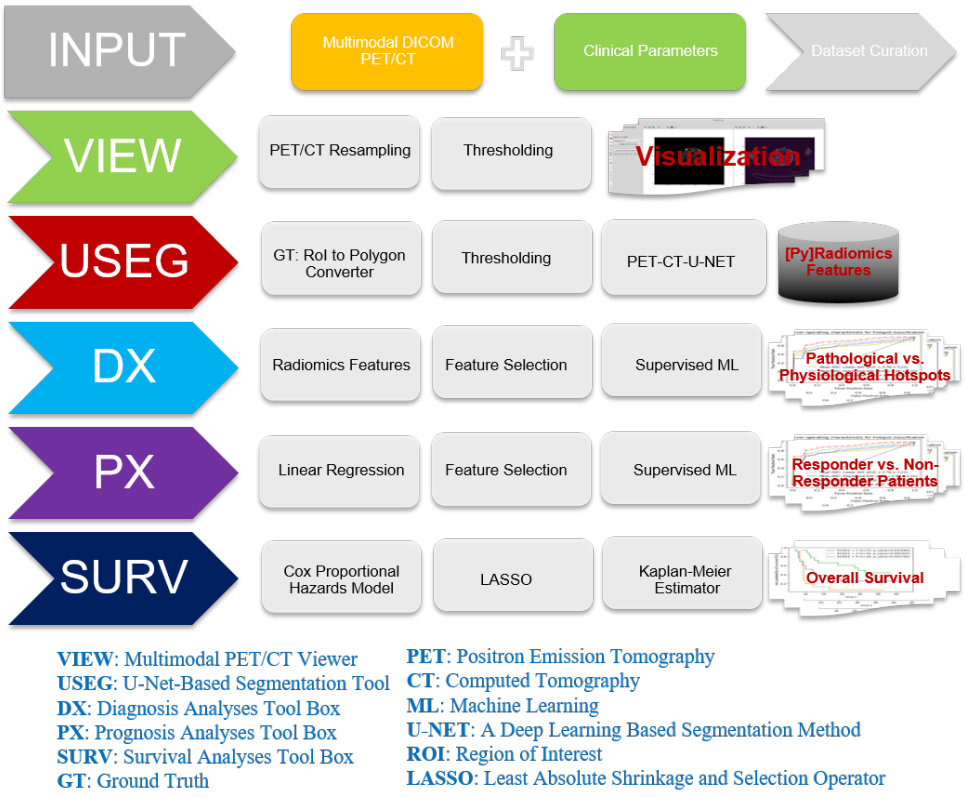


FIGURE 4.3: AutoPyPetCt: Process overview

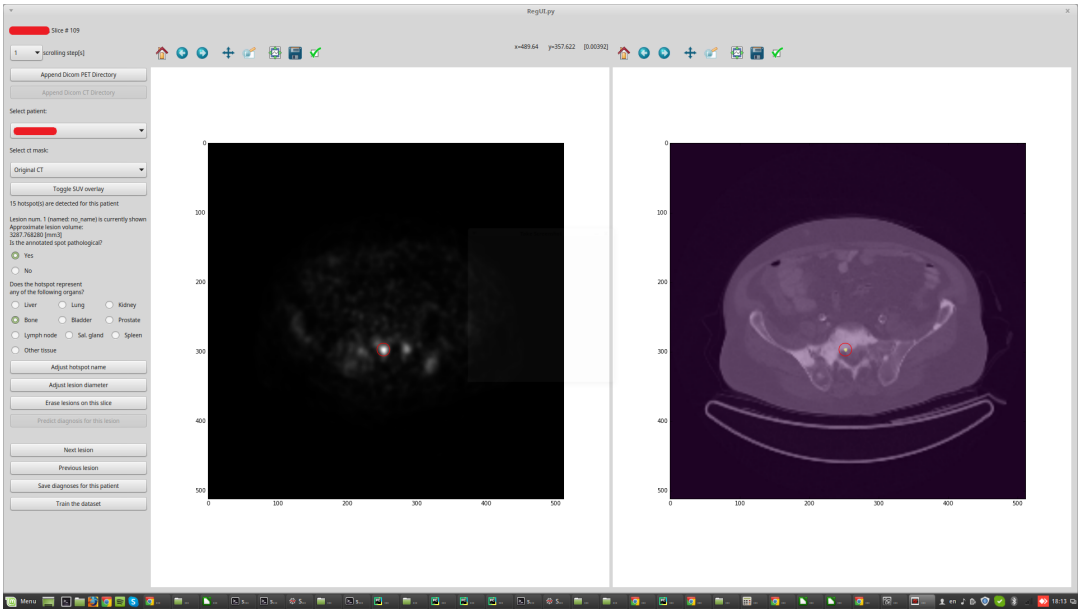


FIGURE 4.4: AutoPyPetCt-VIEW: The multimodal PET/CT visualization tool box.

To provide the interactive visualization toolbox, several modules are developed in Python V.2.7: 1) the Dicom processing module, a script based on pydicom library, to retrieve Dicom metadata information from PET and CT modalities, 2) the resampling module, a script to perform resampling of PET and CT and to organize the backbone of the folder structure for the next processing steps, 3) the module to convert GT labels, a script to convert the GT labeled regions of interest (RoIs) generated by InterView FUSION to 2D polygons and as a result to binary masked images, and 4) the interactive user interface (UI) leveraging Matplotlib and PyQt V.4.0 to visualize the resampled PET/CT images.

The original input data were captured by a Siemens Biograph 2 PET/CT machine (Siemens Healthineers, Erlangen, Germany) (will be referred to as Siemens PET/CT scanner in the next sections for simplicity). Originally, PET data were provided in 128×128 matrices with 5 mm slice thickness and CT data were provided in 512×512 matrices with 5 mm slice thickness. As a result of the resampling process based on the affine transformation matrices stored in the Dicom headers, both of the PET and CT modalities were brought to the same space of 256×256 matrices with 2.5 mm slice thickness. Figure 4.4 shows a screenshot from the PET/CT visualization UI. The UI consists of two main plots illustrating PET and CT (with or without fused PET overlay) slices and a toolbar on the left side which facilitates visualization and annotation options. Some basic interactions such as mouse scroll wheel and circle draw tools as well as keyboard shortcuts are integrated to help users with scrolling and annotation tasks. Moreover, from the user experience (UX) point of view, the mouse and keyboard interactions are consistent to those of InterView FUSION.

4.3 Segmentation

Based on the project requirements, the pipeline provides both manual and automated segmentation tools, leveraging third-party or in-house developed tools respectively.

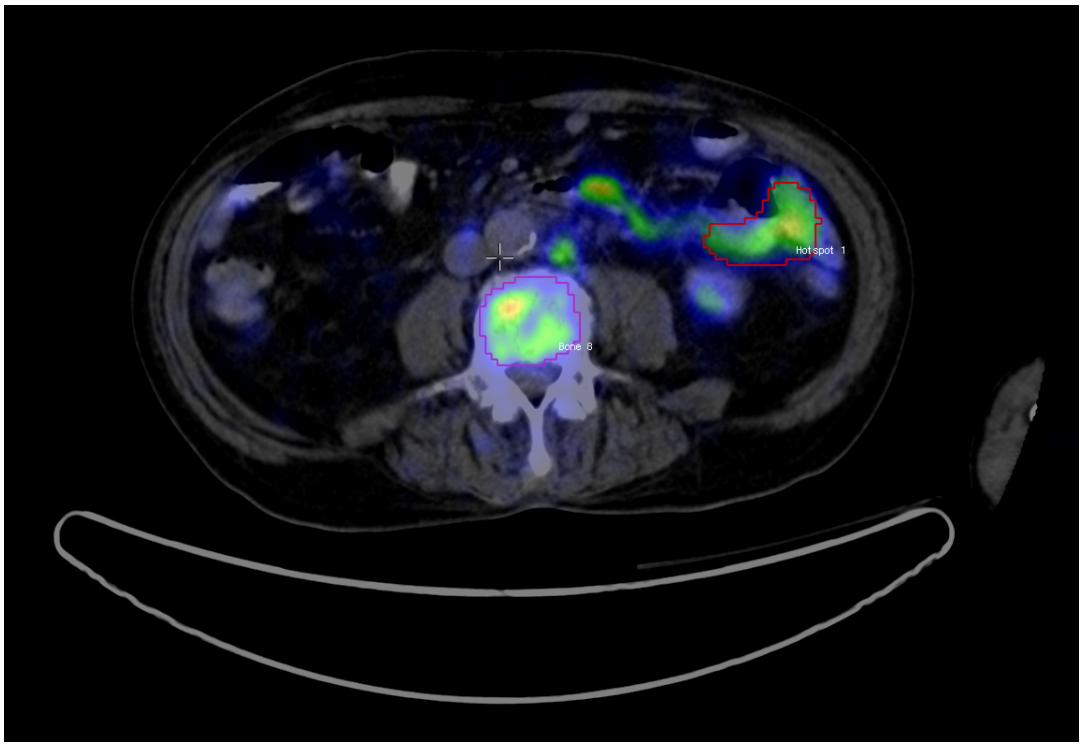


FIGURE 4.5: Manual delineation of pathological and physiological uptake using InterView FUSION (Mediso, Budapest, Hungary). The annotator has delineated two regions of interest in this slice: a bone metastasis (named Bone 8) and a hotspot (named Hotspot 1).

4.3.1 Manual Segmentation

In order to provide ground truth (GT) annotation options such as hotspots definition and delineation, InterView FUSION is used. Using InterView FUSION, the hotspots are defined as volumes of interest (VoIs) in a slice based manner (see Figure 4.5). Thus, each 3D VoI is defined as successively connected 2D RoIs. As a result, for each patient, the hotspots information including name and pixel coordinates of all points belonging to all polygons are stored as extensible markup language (XML) format RoI files. To facilitate further processing, a part of the in-house developed software converts the XML format RoI file to binary format images for both PET and CT modalities as GT masks to train the automated segmentation unit.

4.3.2 Automated Segmentation

For the automated segmentation, in-house developed tools are implemented leveraging thresholding based and U-Net based models. The GT masks converted from predefined RoIs (as described in the previous section) are used to train and fit the automated model as the gold standard for segmentation of hotspots.

Multimodal Thresholding

For the thresholding based method, both PET and CT modalities are taken into account. For PET thresholding, the fixed threshold of 40% of maximum standardized uptake value (40%-SUV_{MAX}) is used to create the slice based binary image masks from the original input PET images. The 40% threshold is chosen as it is widely used

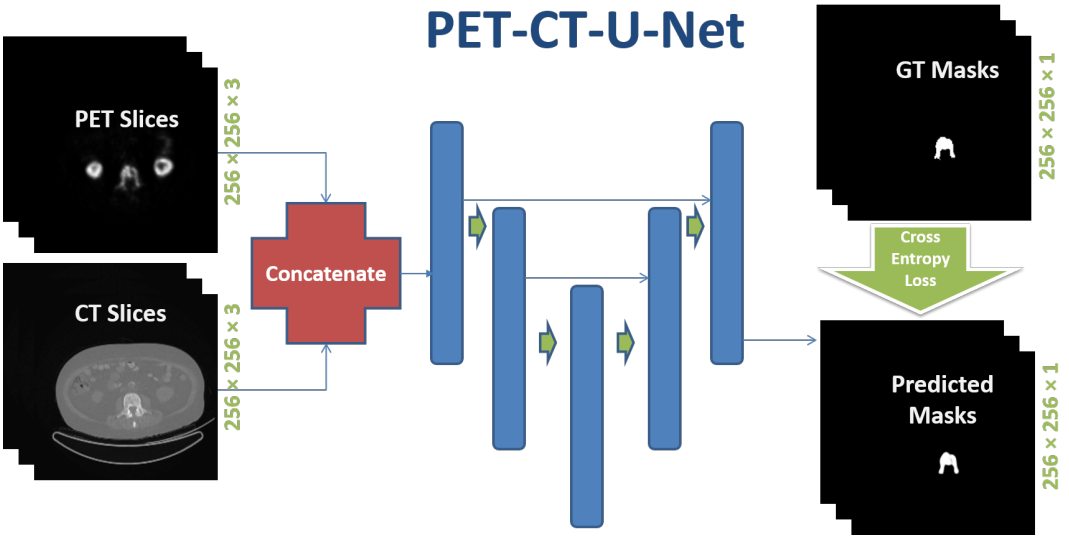


FIGURE 4.6: The simplified schematic of the implemented multi-modal U-Net based segmentation network (PET-CT-U-Net). PET and CT slices are processed as separate channels. The PET-CT-U-Net internally consists of 2 alternative models, one just processing the PET modality, and one processing PET and CT channels simultaneously. In addition, the 40%-SUV$_{MAX}$ mask is internally generated from PET for comparison purposes. Binary cross-entropy serves as the loss function. The figure was originally published in [106].

in oncological studies [78, 120, 170]. Moreover, CT inputs are further processed, taking advantage of thresholding based on Hounsfield scale [40], to create bone masks. As a result, the bone masks are calculated as a co-product of the automated segmentation unit.

U-Net Segmentation

The U-Net based network, named PET-CT-U-Net, takes resampled PET and CT modalities as separate input channels and gets fit and trained by GT masks to predict slice based masks as output. Furthermore, thresholded masks based on PET images are generated to complement the predicted masks for quantitative as well as qualitative analyses. To develop the automated segmentation unit, Python V.3.6, TensorFlow V.2.0 [1] and Keras [28] libraries are utilized to create a convolutional neural network (CNN) inspired by the U-Net architecture. The segmentation network defines two different models based on input channels: single and dual. Figure 4.6 shows the simplified architecture of the U-Net segmentation unit.

To train and fit the two alternative single and dual networks, two different models are defined. Accordingly, the data from PET scans are applied to train and fit the singular model, while the data from both PET and CT modalities are applied to train the dual model. Additionally, to set a baseline for performance analyses of the segmentation networks, the 40%-SUV$_{MAX}$ masks are used. The input image sizes of both PET and CT slices are set to 256×256. Each of the alternative U-Net based models comprised of seven encoding and seven decoding steps connected via a bridge layer. The filter numbers of the subsequent encoding steps are set to: 16, 32, 48, 64, 128, 256, 480, 512 respectively. Correspondingly, the reversed order of these filter sizes are used for the decoding steps. In the encoding round, which loops over the

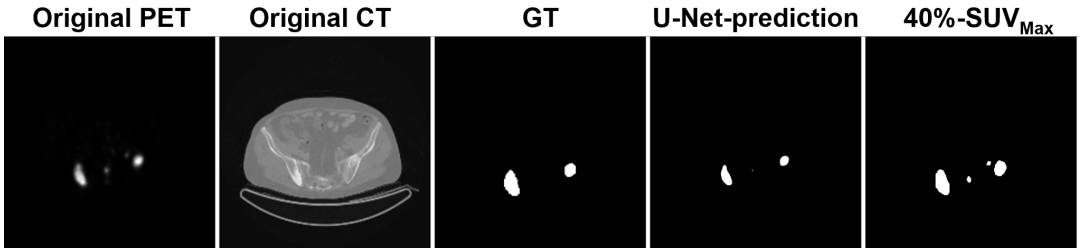


FIGURE 4.7: Sample output of the U-Net based segmentation pipeline. Apart from the PET and CT input channels, the ground truth (GT), U-Net prediction, and 40%-SUV$_{MAX}$ mask are shown.

above-mentioned filter numbers, a 2D convolution block is applied and followed by a 2D max pooling at each iteration. Then a single 2D convolution block is applied at the bridge layer. In the decoding round following the bridge layer, the reverse order of the filter numbers as used for the encoding round is applied. Each iteration of the decoding round includes a 2×2 upsampling followed by a 2D convolution block. As the output layer in the end, a sigmoid activation layer follows a 1×1 2D convolution to end up with the output binary image. The 2D convolution block consists of two 3×3 convolutions, each including a batch normalization and a rectified linear unit (ReLU) activation.

Because of the imbalance in the number of pixels in GT masks and background, weighted binary cross-entropy is used as the loss function between GT and background images. The segmentation quality metrics are measured as precision, recall, and the Dice coefficient of the predicted and GT masks. To tune the networks, different values for the hyperparameters such as batch size (values: 8 and 16), learning rate (values: 0.0001, 0.001, 0.01, and 0.1), and epochs (up to 60) have been considered. To split the dataset into separate train, validate and test groups, the train_test_split function of the model_selection class of Scikit-Learn library is applied. Then, to fit the model using the train and validate subsets, the quality metrics (precision, recall, and Dice) are used to quantify the level of agreement between the predicted masks and GT labels (encoded as 2D binary images) at each epoch. The same procedure is repeated until the maximum epochs are applied or the early stopping criteria are met. To this end, TensorFlow's EarlyStopping function with inputs monitor=validation_loss and patience=10 is applied.

After the training is finished, the fitted model is applied to the held-out test set to predict their corresponding masks. In the final step, both quantitative (i. e., Dice coefficients, accuracy, precision, and recall) as will as qualitative (by an experienced NM expert) measures are applied to assess the performance of the models. As the next methodological step, the input data for the treatment response prediction task should be provided. Therefore, as the subsequent step, the best predicted mask from the U-Net based model is applied to the input images to calculate patient specific radiomics features using PyRadiomics library.

To give a sample outlook, the PET-CT-U-Net applies both thresholding and U-Net based methods on both PET and CT modalities to end up with predicted masks (figure 4.7). For further analysis steps, these predicted masks are applied to calculate radiomics features (RFs) for PET and CT modalities.

TABLE 4.1: List of the radiomics features calculated by InterView FUSION for both PET and CT modalities. Please note that the metabolic tumor volume (MTV) is PET-specific.

First or Higher Order Statistics	Textural
Deviation Mean Max Min Sum PET-MTV Kurtosis Max. Diameter (Size)	Entropy Homogeneity Correlation Contrast Size Variation Intensity Variation Coarseness Busyness Complexity
Volumetric Zone Length Statistics	**Volumetric Run Length Statistics**
Short Zone Emphasis Long Zone Emphasis Low Grey-Level Zone Emphasis High Grey-Level Zone Emphasis Short Zone Low Grey-Level Emphasis Short Zone High Grey-Level Emphasis Long Zone Low Grey-Level Emphasis Long Zone High Grey-Level Emphasis Zone Percentage	Short Run Emphasis Long Run Emphasis Low Grey-Level Run Emphasis High Grey-Level Run Emphasis Short Run Low Grey-Level Emphasis Short Run High Grey-Level Emphasis Long Run Low Grey-Level Emphasis Long Run High Grey-Level Emphasis Grey-Level Non-Uniformity Run Length Non-Uniformity Run Percentage

4.4 Feature Calculation and Feature Selection

In our methods, we used radiomics features either extracted from manually annotated and delineated hotspots or calculated for the masks predicted by the automated U-Net based model (4.3.2). The calculated radiomics features are further used to either classify pathological uptake, predict responders to ^{177}Lu-PSMA therapy, or analyze overall survival. To avoid overfitting, different feature selection methods including recursive feature elimination (RFE) and LASSO as described in 2.4.1 are applied. In this section, different radiomics feature groups are elaborated in more detail.

4.4.1 Radiomics Features From Manual Segmentation

To provide ground truth labels for the clinical studies, InterView FUSION's standard set of radiomics features are calculated for all the volumes of interest (VoIs). These features include first and higher order statistics, diameter, textural heterogeneity parameters, and volumetric run/zone-length statistics. Table 4.1 provides a complete list of the features. For more information on the features, refer to InterView FUSION official documentation provided by Mediso [98].

4.4.2 Radiomics Features From Automated Segmentation

To complement U-Net based pathological uptake predictions, PyRadiomics library [56] is used to calculate radiomics features for the predicted masks from the previous

step. Thus, for each combination of input image and predicted mask, a total of 120 radiomics features including first and higher order statistics are calculated. As a result, the radiomics features are calculated based on the PET-CT-U-Net predicted masks.

As defined in PyRadiomics official documentation, radiomics features are categorized into the following groups: First Order Statistics, 2D/3D Shape-based, Gray Level Co-occurrence Matrix (GLCM), Gray Level Run Length Matrix (GLRLM), Gray Level Size Zone Matrix (GLSZM), Neighbouring Gray Tone Difference Matrix (NGTDM), and Gray Level Dependence Matrix (GLDM) which are elaborated in the following sub-sections. In addition, definitions of some of the features which were identified as significant for diagnostic or prognostic tasks in our retrospective analyses are presented. The feature categories as well as definitions are adapted from the PyRadiomics corresponding documentations [56]. Most of the features calculated by the PyRadiomics library comply with the standard definitions provided by the image biomarker standardisation initiative (IBSI) [170].

First Order Statistics

19 features which describe the distribution of voxel intensities within the RoI defined by the mask through commonly used and basic metrics such as energy, entropy, kurtosis, minimum, mean, maximum, median, 10th and 90th percentile, etc.. Some of the important features of this category are defined as follows:

Energy

$$energy = \sum_{i=1}^{N_p} (\mathbf{X}(i) + c)^2, \tag{4.1}$$

which quantifies the magnitude of voxel values in an image, for which larger values imply greater sum of the squares of these values, where c is an optional value, defined by $voxelArrayShift$ which shifts the intensities to prevent negative values in $\mathbf{X}$, which itself is a set of N_p voxels included in the RoI.

Entropy

$$entropy = -\sum_{i=1}^{N_g} p(i) \log_2 \big(p(i) + \epsilon\big), \tag{4.2}$$

where N_g is the number of discrete intensity levels in the input image. Entropy refers to the uncertainty or randomness in the image values. It quantifies the average amount of information required to encode the image values. In the formula, ϵ is an arbitrarily small positive number ($\approx 2.2 \times 10^{-16}$).

Kurtosis

$$kurtosis = \frac{\mu_4}{\sigma^4} = \frac{\frac{1}{N_p}\sum_{i=1}^{N_p} (X(i) - \bar{X})^4}{\left(\frac{1}{N_p}\sum_{i=1}^{N_p} (X(i) - \bar{X})^2\right)^2}, \tag{4.3}$$

where X is a set of N_p voxels included in the RoI, $\bar{X}$ is the mean value, μ_4 is the 4^{th} central moment, and σ is the standard deviation. Kurtosis is a metric for the so-called *peakedness* of the distribution of values of an RoI. Higher values of kurtosis imply concentration of mass towards the tails rather than towards the mean and vice versa.

Skewness

Skewness refers to the asymmetry of the distribution of values about the Mean value. This value can be positive or negative, depending on where the mass of the distribution is concentrated and the tail is spread:

$$skewness = \frac{\mu_3}{\sigma^3} = \frac{\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i) - \bar{X})^3}{\left(\sqrt{\frac{1}{N_p}\sum_{i=1}^{N_p}(\mathbf{X}(i) - \bar{X})^2}\right)^3}, \tag{4.4}$$

where μ_3 is the 3^{rd} central moment and σ is the standard deviation. As for the kurtosis, X is a set of N_p voxels included in the RoI and $\bar{X}$ is its mean value.

2D/3D Shape-based

26 features which include descriptors of the 2D and 3D size and shape of the RoIs. These features are independent from the gray level intensity distributions of the RoIs. Some of the 2D features are mesh surface, pixel surface, and sphericity. For the 3D case, mesh volume, voxel volume, and spherical disproportion are sample features of this category. In the following lines, some of the important features of this group are defined:

SurfaceVolumeRatio

$$SurfaceVolumeRatio = \frac{A}{V}, \tag{4.5}$$

where A, V represent the surface area and the mesh volume of the RoI respectively and are themselves defined as:

$$\begin{aligned} A_i &= \frac{1}{2}|a_i b_i \times a_i c_i|, \\ A &= \sum_{i=1}^{N_f} A_i, \end{aligned} \tag{4.6}$$

and

$$\begin{aligned} V_i &= \frac{Oa_i \cdot (Ob_i \times Oc_i)}{6}, \\ V &= \sum_{i=1}^{N_f} V_i, \end{aligned} \tag{4.7}$$

where $a_i b_i$ and $a_i c_i$ are the edges of the i^{th} triangle in the mesh formed by the vertices a_i, b_i, c_i and N_f is number of the faces in the triangle mesh. Also for calculating V, for each face i in the mesh defined by a_i, b_i, c_i, the (signed) volume V_f of the tetrahedron defined by the face f and the origin of the image O is calculated. The total surface area A and the total volume V are calculated as the sum of all sub areas A_i and all volumes V_i respectively.

Maximum 3D diameter

Maximum 3D diameter is defined as the largest pairwise Euclidean distance between RoI surface mesh vertices.

Maximum 2D diameter (Slice)

Maximum 2D diameter (Slice) is defined as the largest pairwise Euclidean distance between RoI surface mesh vertices usually in the axial plane.

Gray Level Co-occurrence Matrix (GLCM)

24 features which are defined based on the co-occurrence matrix which itself is defined over an image and quantifies the distribution of co-occurring grayscale pixel values at a given offset and a given angle. As the co-occurrence matrices are typically large and sparse, based on specific use-cases, different metrics (i. e. features) of the matrix are calculated and are often referred to as Haralick features [61].

Let N_g be the number of discrete intensity levels in the input image, then the $(i,j)^{th}$ elements of the GLCM matrix $P(i,j|\delta,\theta)$ represent the number of times the combinations of the levels i,j occur in two pixels of the input image, which are separated by distance δ pixels along angle θ. Two of the most important features belonging to this category are contrast and correlation and are defined as follows:

Contrast

Contrast measures the local intensity variation, promoting values away from the diagonal ($i = j$), where a larger value refers to a greater disparity in intensity values among neighboring voxels:

$$contrast = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g}(i-j)^2 p(i,j). \tag{4.8}$$

Correlation

Correlation holds a value between 0 (uncorrelated) and 1 (perfectly correlated), representing the linear dependency of gray level values to their corresponding voxels in the GLCM:

$$correlation = \frac{\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p(i,j)ij - \mu_x\mu_y}{\sigma_x(i)\sigma_y(j)}, \tag{4.9}$$

where μ_x is the mean gray level intensity of p_x and defined as $\mu_x = \sum_{i=1}^{N_g} p_x(i)i$ and μ_y is the mean gray level intensity of p_y and is defined as $\mu_y = \sum_{j=1}^{N_g} p_y(j)j$, and σ_x, σ_y are the standard deviations of p_x, p_y respectively.

Gray Level Run Length Matrix (GLRLM)

16 features which are defined based on the GLRLM matrix which itself quantifies gray level runs (GLRs) of consecutive pixels that attribute the same gray level value. The $(i,j)^{th}$ element of a GLRLM matrix $P(i,j|\theta)$ describes the number of runs with gray level i and length j as occurred in the image or RoI along angle θ. Some of the important features of this category are described as follows:

RunEntropy (RE)

RE quantifies the uncertainty or randomness in the distribution of run lengths and gray levels. Thus, a higher value refers to more heterogeneity in the texture patterns:

$$RE = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_r} p(i,j|\theta)\log_2(p(i,j|\theta)+\epsilon), \tag{4.10}$$

where, N_r and N_g are the number of discreet run lengths and the number of discreet intensity values in the image respectively and ϵ is an arbitrarily small positive number ($\approx 2.2 \times 10^{-16}$).

Run Length Non-Uniformity (RLNU)

RLNU quantifies the similarity of run lengths throughout the image, where a lower value indicates more homogeneity among run lengths in the image:

$$RLNU = \frac{\sum_{j=1}^{N_r} \left(\sum_{i=1}^{N_g} \mathbf{P}(i,j|\theta) \right)^2}{N_r(\theta)}, \tag{4.11}$$

where, N_r and N_g are the same measures as described in 4.10.

RunLengthNonUniformityNormalized (RLNUN)

RLNUN quantifies the similarity of run lengths throughout the image, where a lower value indicates more homogeneity among run lengths in the image. This is the normalized version of the RLNU formula (4.11) and is defined as follows:

$$RLNUN = \frac{\sum_{j=1}^{N_r} \left(\sum_{i=1}^{N_g} \mathbf{P}(i,j|\theta) \right)^2}{N_r(\theta)^2}, \tag{4.12}$$

here again (and in the next formulas), N_r and N_g are the same measures as described in 4.10.

RunPercentage (RP)

RP quantifies the coarseness of the texture by measuring the ratio of number of runs and number of voxels in the ROI:

$$RP = \frac{N_r(\theta)}{N_p}. \tag{4.13}$$

ShortRunEmphasis (SRE)

SRE measures the distribution of short run lengths, where a greater value describes a shorter run lengths and a more smooth texture:

$$SRE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_r} \frac{\mathbf{P}(i,j|\theta)}{j^2}}{N_r(\theta)}. \tag{4.14}$$

Gray Level Size Zone Matrix (GLSZM)

16 features defined based on the GLSZM matrix which measures gray level zones in an input image and is defined as the number of connected voxels which attribute same gray level intensity. The $(i,j)^{th}$ element of a GLSZM matrix $P(i,j)$ equals the number of zones in the image attributing gray level i and size j. Some of the important features of this group are:

SmallAreaEmphasis (SAE)

SAE quantifies the distribution of small size zones, where a greater value describes more smaller size zones and finer textures:

$$SAE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_s} \frac{\mathbf{P}(i,j)}{j^2}}{N_z}, \tag{4.15}$$

where N_z is the number of dependency zones in the image and equals $\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \mathbf{P}(i,j)$ and $1 \leq N_z \leq N_p$.

SmallAreaLowGrayLevelEmphasis (SALGLE)

SALGLE quantifies the proportion of the joint distribution of smaller size zones with lower gray-level values in the image:

$$SALGLE = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_s} \frac{\mathbf{P}(i,j)}{i^2 j^2}}{N_z}, \tag{4.16}$$

where N_s is the number of discreet zone sizes in the image.

ZonePercentage (ZP)

ZP quantifies the texture's coarseness, computing the ratio of number of zones and number of voxels in the ROI:

$$ZP = \frac{N_z}{N_p}. \tag{4.17}$$

Values of ZP are in the range $[\frac{1}{N_p}, 1]$, where higher values indicate that larger portion of the ROI consists of small zones, hence a finer texture.

Neighbouring Gray Tone Difference Matrix (NGTDM)

5 features which are defined based on the NGTDM matrix which itself measures the difference between a gray value and the average gray value of its neighbours within distance δ and stores the sum of absolute differences for gray level i. Considering $\mathbf{X}_{gl}$ as a set of segmented voxels and $x_{gl}(j_x, j_y, j_z) \in \mathbf{X}_{gl}$ as the gray level of a voxel at position (j_x, j_y, j_z), then the average gray level of the neighbourhood is defined as:

$$\bar{A}_i = \bar{A}(j_x, j_y, j_z) = \frac{1}{W} \sum_{k_x=-\delta}^{\delta} \sum_{k_y=-\delta}^{\delta} \sum_{k_z=-\delta}^{\delta} x_{gl}(j_x + k_x, j_y + k_y, j_z + k_z), \tag{4.18}$$

where $(k_x, k_y, k_z) \neq (0,0,0)$ and $x_{gl}(j_x + k_x, j_y + k_y, j_z + k_z) \in \mathbf{X}_{gl}$, and W represents the number of voxels in the neighbourhood which are also in $\mathbf{X}_{gl}$. Busyness and coarseness are important features of this group and are defined as:

Busyness

Busyness measures the change from a pixel to its neighbour. A busy image which corresponds a high value for busyness attributes rapid changes of intensity between pixels and their neighbourhood:

$$Busyness = \frac{\sum_{i=1}^{N_g} p_i s_i}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} |i p_i - j p_j|}, \text{ where } p_i \neq 0, p_j \neq 0. \tag{4.19}$$

Coarseness

Coarseness measures the average difference between the center voxel and its neighbourhood and indicates the spatial rate of change. A higher value corresponds to a lower spatial change rate and a locally more uniform texture:

$$Coarseness = \frac{1}{\sum_{i=1}^{N_g} p_i s_i}. \tag{4.20}$$

In the formulas 4.19 and 4.20, the p_i are the gray level probabilities and are equal to $\frac{n_i}{N_v}$, n_i is the number of voxels in X_{gl} with gray level i, N_v is the total number of voxels in X_{gl}, and s_i is the sum of absolute differences for gray level i and defined as:

$$s_i = \begin{cases} \sum^{n_i} |i - \bar{A}_i| & \text{for} \quad n_i \neq 0 \\ 0 & \text{for} \quad n_i = 0 \end{cases} \tag{4.21}$$

Gray Level Dependence Matrix (GLDM)

14 features which are defined by the GLDM matrix which itself measures gray level dependencies in an image. A gray level dependency is defined as a the number of connected voxels within distance δ which depend on the center voxel. For $|i - j| \leq \alpha$, a neighbouring voxel with gray level j is considered dependent on center voxel with gray level i. The $(i, j)^{th}$ element of a GLDM matrix $P(i, j)$ refers to the number of times a voxel with gray level i with j dependent voxels in its neighbourhood presents in the image. Gray Level Non-Uniformity (GLNU) is one of the most important features in this category and is defined as follows:

Gray Level Non-Uniformity (GLNU)

GLNU quantifies the similarity of gray-level intensity values in the image. A lower value of the GLNU refers to a greater similarity in intensity values:

$$GLNU = \frac{\sum_{i=1}^{N_g} \left(\sum_{j=1}^{N_d} \mathbf{P}(i, j) \right)^2}{N_z}, \tag{4.22}$$

where N_g is the number of discreet intensity values, N_d is the number of discreet dependency sizes, and N_z is the number of dependency zones in the image and equals $\sum_{i=1}^{N_g} \sum_{j=1}^{N_d} \mathbf{P}(i, j)$.

4.4.3 Feature Selection

Depending on the classification tasks and feature groups provided for each task, different feature selection methods are applied. For instance, as explained in 2.4.1, recursive feature elimination (RFE) and least absolute shrinkage and selection operator (LASSO) are applied. For the hotspot classification task using manually segmented lesions, features were ranked based on the feature importance measure defined for the extra trees classifier provided by the scikit learn library (as will be described in more detail in 4.7.2). Furthermore as will be discussed in 4.7.5, for identifying the features with the most impact on predicting overall survival, the LASSO method is applied. Finally, RFE method is applied to find the most significant features in treatment response prediction based on radiomics features calculate for predicted masks by the automated segmentation pipeline (4.7.6).

4.5 Supervised Machine Learning

In this thesis, supervised ML classifiers are used for two main tasks: 1) pathological hotspot classification and 2) therapy response prediction. Regardless of which group of radiomics features (either calculated by InterView FUSION or measured using PyRadiomics) is used, state-of-the-art surprised ML classifiers are fit in cross validation steps which include grid search for hyperparameter tuning. The classifiers include logistic regression, support vector machine (SVM) with linear, radial basis function (RBF), and polynomial kernels, Extra Trees, and Random Forest. For the task of hotspot classification, VoI-specific radiomics features calculated by InterView FUSION are used [42, 107]. For the task of treatment response prediction,

per patient averaged radiomics features calculated by InterView FUSION or PyRadiomics (based on U-Net predicted masks) are used [103, 106].

To quantify performances of the ML classifiers, area under the receiver operating characteristic (ROC) curve (AUC), sensitivity (SE), and specificity (SP) are measured as each classifier is trained and tested on independent train and test cohorts. In all the experiments, KFold cross validation is conducted along with grid search on the corresponding training cohort to tune the hyperparameters for each classifier.

4.6 Analysis of Overall Survival

For the analysis of overall survival, both of the uni- and multivariate analysis techniques are applied to the per patient averaged radiomics features calculated by InterView FUSION as well as the patient-specific clinical parameters such as age, gleason score, Hemoglobin level, and prostate specific antigen (PSA). On the one hand, Cox proportional hazards model is used to quantify the so-called radiomics signature (RS) out of the radiomics features using the least absolute shrinkage and selection operator (LASSO) method. On the other hand, Kaplan-Meier (KM) diagrams are drawn to identify the most significant variables which better discriminate between groups with high and low survival expectancies.

4.7 Retrospective Clinical Studies

To validate the methods and to evaluate the developed tools, several retrospective clinical experiments have been conducted. In this section, we summarize the materials and methods which correspond to these clinical studies. The content is adapted from the previous work as published in Diagnostics [105, 107], Annals of Translational Medicine [103], and Tomography [42] journals as well as the lecture notes in computer science (LNCS) proceedings of the multimodal learning and fusion across scales for clinical decision support (ML-CDS) workshop of medical image computing and computer assisted intervention (MICCAI) conference in 2021 [106].

4.7.1 Study Cohorts

For the clinical studies which have been conducted to assess the objectives of this thesis, a cohort of 100 patients with advanced prostate carcinoma has been analyzed retrospectively. For the first three studies [42, 103, 107], a subset of 72 patients has been used. For the survival analysis paper [105], a subset of 83 subjects was used. For the analyses of the PET-CT-U-Net [106], all the 100 subjects have been included. The difference in sizes of the cohorts is caused by the availability of annotated data at the beginning times of the experiments. The summary of the clinical information of the patients' cohort is given in table 4.2. All patients gave written and informed consent to the diagnostic procedure. Due to the retrospective character of the data analysis in all of the clinical studies related to this thesis, an ethical statement was waived by the institutional ethical review board according to the professional regulations of the medical board of Nordrheinwestfalen, Germany.

4.7.2 PSMA-PET/CT Radiomics for Hotspot Classification

Given the manually annotated VoIs using the third-party software (InterView FUSION), in the first two studies, we aimed at analyzing the relative performance of

TABLE 4.2: The summary of the clinical information of the patients' cohort (PSA: prostate specific antigen) [106].

	Age [years]	Gleason Score	PSA [ng/ml]
Minimum	48	6	0.25
Maximum	87	10	5910
Average	70.40	8.32	461.57

different supervised ML classifiers for the identification of pathological hotspots in baseline ^{68}Ga-PSMA-PET/CT findings. In the first study [107], we focused on the technical aspects of classifier tuning and performances, while in the other study [42], the main focus was shifted to the analyses of the ultimate relevance of the methods, especially to the cases where the algorithms would fail to truly identify normal, i. e., physiological, uptake. To this end, we quantified the classifiers' performances as applied to different combinations of PET and CT radiomics features.

Patients and Volume of interest (VoI) definition and annotation

The ^{68}Ga-PSMA PET/CT findings from 72 male patients diagnosed with advanced prostate carcinoma has been retrospectively analyzed. The scans had been taken in the period from November 2014 and February 2017 using the Siemens PET/CT scanner. For each of the patients, the scanning procedure started with an intravenous injection of 98 to 159 MBq in-house produced ^{68}Ga-HBED-CC PSMA. After 40 to 80 minutes, a low-dose CT (16mAs, 130 kV) from the base of the skull to mid-thigh was acquired. Then, depending on the body weight of the patient, the PET scan was taken over the same area with 3 or 4 minutes per bed position. In the next step, the PET data was reconstructed in 128×128 resolution, while the CT data was reconstructed in 512×512 matrices. Both of the PET and CT modalities featured the same slice thickness of 5 mm. Furthermore, the manufacturer utilized an attenuation-weighted ordered subsets expectation maximization algorithm for image reconstruction including attenuation and scatter correction. The same scanning as well as reconstruction procedure was applied for all of the cohorts used in the clinical studies conducted in connection with this thesis.

To define and annotate all the pathological as well as the physiological uptake, For each scan, trained nuclear medicine (NM) physicians have identified and manually delineated all the hotspots consecutively using InterView Fusion software (Mediso Medical Imaging, Hungary [98]) (see fig. 4.5). To define each 3D VoI, all its subsequent 2D slices were delineated to form fully-connected 3D volumes. The criteria to choose an uptake was the visible tracer uptake without any predefined threshold. The hotspots included benign and malignant tissues all over the body and any metastatic uptake in bones or lymph nodes as well as any visible physiological uptake in kidneys, livers, glands, etc.. To end up with hotspot level radiomics features, a total of 80 (40 PET-based + 40 CT-based) features were calculated by InterView Fusion software (the standard set of radiomics features provided by the software) for each hotspot. The radiomics features include first and higher order statistics features (SUV_{MEAN}, SUV_{MAX}, kurtosis, etc.), textural heterogeneity features (entropy, contrast, etc.), as well as volumetric zone and run length statistics (grey-level non-uniformity, short run emphasis, etc.). See table 4.1 for the detailed list of the radiomics features. Afterwards, the ground truth labels were merged with the Mediso output using our internal PET/CT scan annotator software (Python V2.7).

Classification

The task of discriminating pathological uptake from normal physiological uptake has been formulated as a supervised machine learning classification problem. Therefore, to set up the annotated ground truth (GT) labels for the study cohort, two experienced NM physicians classified all the hotspots and labeled them as pathological vs physiological. To quantify the relevance of radiomics features from PET and CT modalities, three different feature vectors have been curated for each hotspot: PET only, CT only, and combined PET and CT (PET/CT). Furthermore, five different ML classifiers (linear, radial basis function (RBF), and polynomial kernel SVM, extra trees (ET), and random forest (RF)) have been trained and fit. Consecutively, the performances of all classifiers have been measured as applied to each of the feature groups (e. g., PET with linear SVM or PET/CT with ET).

To quantify the performances of the classifiers, the accuracy measures (area under the receiver operating characteristic (ROC) curve (AUC), standard deviation (STD) of AUCs for the cross validation and the feature ranking steps, sensitivity (SE), and specificity (SP)) were quantified as each of the classifiers applied to each feature group. To establish the hyperparameter values for the ML algorithms, five-fold cross validation (CV) was applied to a training data-set with 48 subjects. Afterwards, the performance of the resulting classifier was evaluated on a test (held-out) cohort consisting of 24 subjects followed by an inter-observer analysis. To obtain insight into which radiomics features contributed most, they were ranked depending on the ET classifier, as it performed the best as applied to the test cohort. The features were sorted with regard to the ET feature importance measure provided by the scikit learn library [159] which quantifies the overall decrease in Gini impurity achieved with a given feature. The means and standard deviations of these importance scores, over the folds of our 5-fold cross validation are reported.

Cross validation (CV)

It is inevitable to subdivide the dataset into train and test cohorts to avoid overfitting and to achieve better generalizable results. Thus, after the dataset is subdivided into independent train and test cohorts randomly, for each classifier, hyperparameters were tuned in an interim cross validation step including grid search. To this end, the first subset, including 48 subjects, was used for training and hyperparameter tuning and the second subset, containing 24 subjects, performed as the test or held-out set. After standardizing the dataset using MinMaxScaler method [100], KFold method with 5 folds was applied to the training set as CV method. In each CV step, a grid search has been performed to identify the best hyperparameters for each ML classifier to predict the true labels for each feature group. For grid search, we took into account the standard ranges of different hyperparameters (e. g., C=[1, 10, 100, 1000, 2^{-5}, 2^{-3}, ..., 2^{15}], gamma=[1e^{-s}, 1e^{-4}, 2^{-15}5, 2^{-13}, 2^{-11}..., 2^{3}], etc.) for all of the five ML classifiers.

As a result of the grid search, based on the selected hyperparameters for each classifier on the training set, the performance of each classifier to predict the labels of the held-out test cohort was quantified. Here as well, the relative importance of each feature group was measured separately. Finally, the accuracy metrics of each combination of classifiers and feature groups as applied to the test (held-out) set are reported as the achieved performances.

Inter-observer Variability

To account for inter-observer variability, both qualitative and quantitative measures have been analyzed. Thus, the whole study cohort has been randomly subdivided into two groups: 1) first subset with 30 patients and 2) second subset with 42 patients. The subsets were manually segmented and annotated by two different experienced NM physicians (referred to Annotator 1 and Annotator 2). Furthermore, the resulting manual segmentations and annotations performed by the two annotators have been reviewed and qualified by a third highly experienced NM physician (referred to as Annotator 3). To quantify the inter-annotator variability of the manual segmentations, additional rounds of CV were conducted. In the first round, the data labelled by Annotator 1 served as training data and the data labeled by Annotator 2 served as the test data. In the second CV round, the train and test cohorts swapped sides. Afterwards, the results of inter-observer CV steps were compared to the main CV results.

Permutation Test

To assess the significance of the results achieved at the classification step, a permutation test was conducted to reject the null hypothesis which stated that the classifiers would feature similar performances as applied to the same data with permuted labels. Thus, a separate five-fold CV has been applied to the training cohort of 48 patients from the first CV step. In total, there were 25000 iterations with the same set of feature groups and ML classifiers as for the first CV. In each CV step of the permutation test, the GT labels were replaced with permuted binary labels. As a result, each AUC that was equal to or higher than the threshold of 0.85 (which was smaller than the lower bound of prediction scores on the test set) was counted. Finally, the resulting count of the AUCs greater than 0.85 was divided by the total number of the iterations (25000) to end up with the p-value of the permutation test:

$$p = \frac{n(AUCs \geq thr)}{N_{iters}}, \tag{4.23}$$

where p is the p-value of the permutation test, $n(.)$ is the number of the $AUCs$ greater than the given threshold (thr), $AUCs$ are the calculated areas under the ROC curves for each classifier as applied to each feature group at each iteration, and N_{iters} is the total number of iterations (equation 4.23).

4.7.3 Follow Up Hotspot Classification Study

To validate the ML methods which had been applied in the first publication [107] on a new set of unseen data and to analyze in more detail the performance of the methods to truly identify normal uptake, a second study has been conducted [42]. As the ML methods had shown promising results in identification of pathological uptake, in the follow up study, the main focus has been to analyze where (i. e., in which organs or body locations) the ML methods would most likely fail to correctly predict physiological uptake as they were applied to the newly tested unseen data.

Patients and Volume of Interest (VoI) Delineation

^{68}Ga-PSMA-PET/CT scans from 87 patients with histologically proven advanced prostate carcinoma were included in this retrospective follow up analysis. The scanning machine and protocols as well as the image reconstruction steps have been

similar to those of the previous study [107]. As for the previous study, the PET/CT images data were analyzed using InterView FUSION Software and the hotspots have been identified based on fused PET and CT data. Also, the volumes of interest (VoIs) were manually delineated slice by slice. In a second step, the hotspots were classified as pathological or physiological, corresponding to the location they were situated in. The hotspots included primary prostate cancer and metastases in the skeletal system, lymph nodes, as well as physiological uptake in kidneys, liver, glands, gastrointestinal tract (gut), etc..

The cohort of 72 patients from the previous study was used as the training dataset. As a result, a total number of 2452 hotspots were delineated and annotated as either pathological (total of 1629) or physiological (total of 823). As the held-out test cohort, 15 remaining patients with similar clinical characteristics as the patients in the training cohort were included. To prepare the test cohort, first, the PET/CT scans were analyzed with Interview FUSION software as explained before. For each patient, 5 to 10 pathological hotspots and the tracer uptake in all glands; 5 physiological hotspots as well as the uptake in the liver were delineated. This analysis resulted in a total of 331 hotspots consisting of 128 pathological and 203 physiological lesions. In the next step, for each hotspot, the same set of radiomics features were calculated by InterView FUSION software as were calculated for the previous study (table 4.1).

Training and Classification

Based on the methods evaluated in the previous work [107], an in-house developed software in Python V.3.5 was developed for training and classification purposes. Initially, a subset of 30 subjects from the training cohort of 72 subjects was used to pre-set, tune, and compare three machine learning classifiers from SciKitLearn library (linear kernel support vector machine (SVM), Extra Trees, and random forest (RF)). In the first round of training, KFold with 3 folds was applied as the cross-validation (CV) method to tune the hyperparameters of the three ML classifiers and to identify the most accurate one. At each CV step, the C and Gamma parameters of the linear SVM as well as the min_sample_leaf and max_depth of the Extra Trees and RF classifiers were tuned using grid search. Here again, standard ranges of the hyperparameters are applied for the hyperparameter tuning. As a result, the best combination of the hyperparameters for each classifier were identified and the best classifier was selected based on the performance metrics (AUC, sensitivity (SE), and specificity (SP)). Afterwards, the performances of the classifiers on the test cohort were further quantified to end up with the best classifier, which was Extra Trees. Consecutively, we used Extra Trees with the tuned hyperparameters (n_estimators = 250, max_depth = 20, min_samples_leaf = 1) to analyze how the algorithm would generalize as the size of the training cohort would increase.

As the next step to investigate the generalizability of the methods, starting with the initial cohort of 30 patients, we added the data from the second training cohort (consisting of 42 patients), one patient after another with a randomized order. Thus, the sizes of the training subsets ranged from 30 to 72. Therefore, in each training step, first, a random subset of the training cohort was selected and then the size of the subset was increased by one patient. In addition, the classification task has been repeated for 100 times with a bootstrapping approach to quantify the accuracy metrics of the classifiers at each training step. As the objective was to assess the performance of our algorithm on unseen data, in each step, the prediction accuracies were calculated as the classifier was applied to the test set (the held-out set with 15

subjects). Finally, the mean and standard deviation (std) of the performance metrics were reported to give an overview of how increasing the size of the training cohort enhanced the classification performance as trained by the training cohorts of 30 to 72 subjects and tested by the held-out test cohort of 15 subjects. The performance metrics were quantified at each training step based on the prediction scores of the ML classifier as trained by the training cohort and tested by the test cohort, then averaged along all the 100 bootstraps. To reduce the risk of overfitting, first, the MinMax standardization method was used to normalize the dataset including feature vectors of training and test cohorts. For this study, we applied cross validation to identify the best performing classifier as well as bootstrapping with replacement and resampling on the training set to better estimate the population statistics.

4.7.4 Treatment Response Prediction based on Manual Segmentation

In another study, first, we took advantage of linear regression to preselect radiomics features and clinical parameters which would show significant correlation with the difference in PSA levels as the treatment response marker. Then, we applied and compared supervised ML classifiers on different groups of features and parameters including PET, CT, clinical, and best correlating variables for the prediction of responders to ^{177}Lu-PSMA therapy. The results of this study have been already published in Annals of Translational Medicine [103].

Patients and Volume of interest (VoI) definition and annotation

In this study, a retrospective cohort of 83 patients with prostate cancer were analyzed. The procedure of the definition and annotation of the lesions using InterView FUSION software has been similar to those of the previous studies [42, 107], except that for this study, only the hotspots representing pathological uptake have been included. Therefore, for each scan, all the pathological hotspots have been identified and manually delineated by a trained nuclear medicine physician (NM) (board certified with 7 years' experience in PET/CT analysis). The lesions included the primary tumor if present as well as metastatic uptake in any body location. The InterView FUSION standard set of radiomics features (table 4.1) were calculated as for the previous studies and averaged on a per patient basis to be further analyzed using linear regression with PSA level difference as treatment response assessment metric.

In addition to the per patient averaged radiomics features, forteen numerical clinical parameters have been analyzed for each individual patient. These variables include clinical parameters such as age, weight, and height as well as therapeutic parameters such as Gleason score, ALP1, Hemoglobin (Hb) and base-line serum PSA level. Table 4.3 provides the detailed list of the clinical parameters, including numerical and categorical variables. The categorical variables are used in the other studies which will be mentioned in the next sections.

In accordance with the previous findings [5] and as surrogate markers for the therapy response, prostate specific antigen (PSA) serum levels had been captured at the time point of the first PET/CT treatment and seven to eight weeks after the beginning of the treatment. The difference in the PSA levels (ΔPSA) between these time points had been calculated for further analyses. As a result, 59 out of the 83 patients have been classified as responders, while the remaining 24 patients were considered as non-responders to the treatment procedure.

Linear Regression

To check whether there was a linear correlation between radiomics features and clinical parameters of individual patients on the one hand and ΔPSA on the other hand, the values of the radiomics features of all of the individual pathological lesions for each patient were averaged to quantify the mean values of the features. The patient-specific clinical parameters were then merged with their corresponding radiomics features to end up with the patient level feature vectors. To correlate individual features and clinical parameters with ΔPSA, linear regression has been applied. The ΔPSA was calculated by subtracting the PSA level at the post therapy scan from the PSA level at the pre-therapy scan. Thus, a negative value of ΔPSA represents a responder to the ^{177}Lu-PSMA treatment and vice versa.

Due to the limited number of non-responders to the ^{177}Lu-PSMA therapy (59 responders vs 24 non-responders) in the original cohort, for the linear regression task, a balanced subset of the cohort with 48 patients including 24 patients in each category of responders or non-responder was randomly selected (the distribution of the demographic and physiological aspects of the cohort was maintained during the sub-sampling). Afterwards, each of the balanced and unbalanced cohorts were subdivided into training and test cohorts to analyze the prediction performance for the supervised ML classifiers. In addition, separate linear regression analyses have been performed on training data-sets of balanced and unbalanced cohorts. Thus, the linear regression analyses have resulted in different sets of radiomics features and clinical parameters which had strong correlations (p-value < 0.05) with ΔPSA for balanced and unbalanced groups. In the next steps of the analysis pipeline, these best correlating features and parameters were used for the task of treatment response prediction using ML classification methods. The procedure of identifying the most relevant variables by taking into account only the data from the training cohort and excluding the test cohort data would minimize the risk of overfitting [117].

Classification

As related works suggest, support vector machines (SVMs) and decision tree based methods are widely applied for clinical treatment outcome prediction (e. g., predicting the optimal cancer drug therapy [66], predicting the outcome of chemotherapy [39], and stratifying the risk in primary prostate cancer [36]). Therefore, for this study, we have applied five ML classifiers from these groups for the treatment response prediction task. The five ML methods (linear, radial basis function (RBF), and polynomial kernel SVM, Extra Trees, and Random Forest) were applied to assess and rank different groups of radiomics features and clinical parameters. Similar performance metrics (AUC, SE, and SP) as for the previous studies were averaged to quantify the precision for each of the classification tasks. Thus, for each classifier as applied to each feature group, AUC, SE, and SP were calculated separately.

Cross-Validation (CV)

In this study, different CV steps are applied for balanced and unbalanced coghorts. Thus, in the first CV step, the unbalanced cohort with 83 patients (consisting of 59 responders and 24 non-responders) is used. In the second CV step, the balanced cohort of 48 subjects (the same cohort which was used for the linear regression task) is used. The idea of conducting a second CV for the balanced cohort would verify whether the classifiers' scores on the unbalanced cohort had been realistic.

Unbalanced Cohort

In the first CV step, the whole data-set of 83 patients was randomly sub-divided into training and held-out test cohorts with 56 and 27 subjects respectively. The demographics and clinical characteristics as well as the ratios of responders to non-responders in the training and test sets were also compatible. Similar to hotspot classification tasks in the previous studies [42, 107], MinMaxScaler method was used to standardize and normalize the data and stratified KFold CV with 3 folds was applied to the training set for hyperparameter tuning. Here as well, a grid search based on standard ranges of the hyperparameters has been conducted at each CV step. Furthermore, the tuned classifiers are ranked as applied to different groups of radiomics features and clinical parameters.

Balanced Cohort

The balanced cohort consisted of 48 patients including 24 responders and 24 non-responders. As for the unbalanced cohort, at the additional CV step for the balanced cohort, the cohort was subdivided into training (32 subjects) and held-out test (16 subjects) groups. Here as well, the demographic parameters and the ratio of responders to non-responders were identical in the training and test subsets. The methodology of the CV and test steps (KFold, grid search, and accuracy measures) for the balanced cohort was quite the same as for the unbalanced cohort.

Permutation Test

To assure that the results were significant, a permutation test was conducted. Similar to the previous work for the classification of pathological uptake [107], the p-value of the permutation test was calculated (equation 4.23). However, as the number of feature groups as well as the minimum achieved test accuracy differ, this time the AUC threshold as well as the total number of iterations would differ.

4.7.5 PSMA-PET/CT Radiomics for Survival Prediction

In a study aiming at estimating the potential of the baseline ^{68}Ga-PSMA-PET/CT radiomics features, we took advantage of Cox proportional hazards model to calculate the radiomics signature. Also, Kaplan-Meier estimator was used to visualize the survival analyses outcomes. The following subsections are adapted from the published work [105] and describe the methodology of this study.

Patients and Volume of interest (VoI) definition and annotation

For this study, we retrospectively analyzed ^{68}Ga-PSMA PET/CT scans from 83 male patients who had been histologically diagnosed with advanced prostate carcinoma. The baseline PET/CT scans were carried out between November 2014 and August 2019 and the corresponding ^{177}Lu-PSMA treatments were followed in 5 to 21 days thereafter. The scanning and image reconstruction protocols matched those from the previous studies.

Again, InterView Fusion software was used by a trained nuclear medicine (NM) physician (board certified with 7 years' experience in PET/CT analysis) to define and delineate all the pathological lesions for each patient scan. The hotspots include all the primary tumors if present as well as all the metastatic uptake in all of the

TABLE 4.3: Descriptions of the clinical parameters. The table is adapted from [105]

Parameter	Description
Age	Age at the first PSMA PET
Weight	Weight at the first PSMA PET
Height	Height at the first PSMA PET
Gleason Score	Describes abnormality degree of cancer cells in prostate
ALP1	Serum alkaline phosphatase at the first PSMA PET
PSA1	Serum PSA level at the first PSMA PET
Time Difference	Time between the first diagnosis and the first PSMA PET
Crea1	Serum creatinine at the first PSMA PET
GGT1	Gamma-glutamyltransferase at the first PSMA PET
CRP1	C-reactive protein in serum at the first PSMA PET
Hb1	Hemoglobin at the first PSMA PET
Erys1	Erythrocytes at the first PSMA PET
Thrombose1	Thrombocytes at the first PSMA PET
Leukos1	Leicocytes at the First PSMA PET
ECOG1	Scale of the performance status of the patient
Prostatectomy	whether the patient underwent prostatectomy
Hormonal therapy	whether the patient underwent hormonal therapy
Chemotherapy	whether the patient underwent chemotherapy
Bisphosphonate	whether the patient had taken bisphosphonates
Radiotherapy Prostate	whether the patient underwent radiotherapy of prostate
Radiotherapy Bones	whether the patient underwent radiotherapy of bones
Radiotherapy LN	whether the patient underwent radiotherapy of lymph nodes

organs. The same set of radiomics features as for the previous studies have been calculated for each hotspot (table 4.1). Additionally, for each individual patient, eight categorical therapeutic clinical parameters (such as Gleason score, ECOG1, ALP1, and base-line serum PSA level) as well as forteen numerical (such as age, weight, and height) have been taken into consideration (Table 4.3). All of the numerical variables have been standardized prior to survival analysis steps taking advantage of the MinMax method.

Statistical Analyses

The first step towards the survival analysis is to form the standard structured input for the survival analysis with right-censoring. To this end, the time of death for the patients who had died by the date on which the study began or the fact that the patients were still alive on that date were collected. The standard structured survival data consisted of two variables: 1) a boolean variable representing the status of the patient on the date the experiment began (dead=True or alive=False) and 2) one integer parameter referring to the number of months until the date of patient's death or censoring respectively.

Prior to the survival analysis, to avoid overfitting, feature selection has been performed. The reason behind applying the feature selection was that the number of input variables of the data-set exceeded the number of subjects. For the feature selection, from Cox proportional hazard model, the least absolute shrinkage and selection operator (LASSO) method also known as L1 regularization [81] was applied. The LASSO method is applied to select the most relevant features to predict overall survival (OS). Moreover, it provided coefficients for the selected features which were

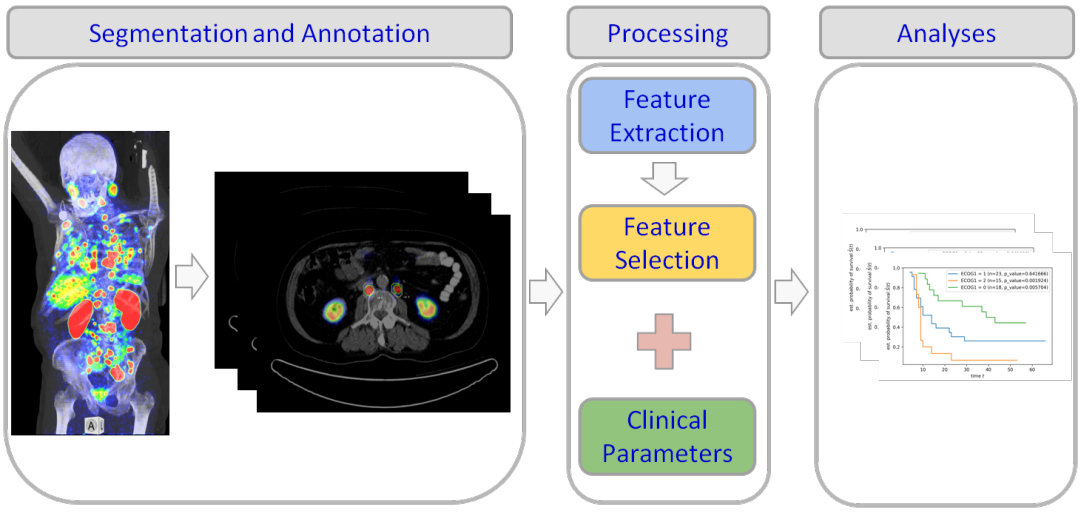


FIGURE 4.8: The overall survival study pipeline. First, the PET/CT images are manually segmented and annotated by an experienced NM physician. Then the radiomics features are extracted and the most relevant features among them are chosen by LASSO method [81] to calculate the radiomics signature. Finally, the Kaplan-Meier estimator [53] is used to analyze and visualize the survival prediction results. This figure was originally published in [105].

then used to calculate the so-called radiomics signature (RS) for each individual patient.

In the next step, to gain more interpretable insight from the survival analyses, Kaplan-Meier (KM) estimator [74] is applied. The KM estimator investigates whether different subject groups separated with regards to predefined cut-off values of given variables have significantly different survival likelihoods. As ending up with too small groups of subjects would affect the generalizability of the analysis results, we used the median value of each numerical variable as well as different possible values of each categorical variable as the cut-off values for the KM estimator. Finally, we compared the predictive performance of some conventional parameters such as MTV and SUV_{Mean}, and SUV_{Max} as well as the clinical parameters to that of the selected variables and calculated RS which was calculated using the LASSO method. An overview of the survival study pipeline is shown in figure 4.8.

4.7.6 Clinical Decision Support Using PET-CT-U-Net

To replace the manual segmentation with automated segmentation pipeline and to perform radiomics analyses directly to the output masks predicted by the automated segmentation network, we trained and fit the multi-channel deep segmentation network (PET-CT-U-Net) and used PyRadiomics library respectively. The automated segmentation and radiomics pipelines are evaluated in a retrospective study [106]. Consecutively, we applied similar methods as in [103] to analyze performance of supervised ML methods for therapy response prediction. In this section, the highlights of the methods used in this study are summarized.

Dataset and Ground Truth Annotation

As described earlier (4.7), the whole study cohort of 100 PCa patients including 67 responders and 33 non-responders have been included in this retrospective study.

The whole cohort was randomly subdivided into train and test cohorts with 61 (40 responders vs 21 non-responders) and 39 (27 responders vs 12 non-responders) subjects respectively. The training cohort was used for fitting the PET-CT-U-Net as well as hyperparameter tuning for ML-based treatment response prediction. The summary of the clinical factors of the study cohort is presented in Table 4.2. Also, the methodology of the definition and annotation of GT labels as used for fitting the PET-CT-U-Net has been quite similar to that of the previous studies.

Automated Segmentation

For the automated delineation of pathological uptake in whole-body PET/CT scans, a multi-channel convolutional neural network (CNN) inspired by U-Net architecture, named PET-CT-U-Net, has been used. Figure 4.6 illustrates the simplified architecture of PET-CT-U-Net. The details of the segmentation network architecture, including encoding and decoding modules as well as convolution and max pooling steps have been described in 4.3.2. To implement the automated segmentation pipeline, TensorFlow and Keras libraries have been used. After training and fitting the network, the predicted pathological masks are used to calculate patient-level radiomics features for PET and CT modalities using PyRadiomics libraries. As a result a total of 120 radiomics features are calculated for each modality. For the complete list of the feature, refer to PyRadiomics official documentation [56]. The performance of the segmentation unit is measured as precision, recall, and the Dice coefficient of the predicted masks as compared to GT masks.

Therapy Response Prediction

Taking the patient-level radiomics features calculated by PyRadiomics, the feature vectors of the whole 100 patients are used for the next analysis step to assess and rank different ML classifiers for prediction of responders to ^{177}Lu-PSMA treatment. To this end, the same training and test cohorts as were used in the automated segmentation step as well as 6 different supervised ML classifiers (logistic regression, support vector machine (SVM) [64] with linear, polynomial and radial basis function (RBF) kernels, extra trees [52] and random forest [18]) have been used. Similar to the previous studies, the training step included 3 fold cross validation and hyperparameter tuning with MinMax standardization for all the classifiers. Furthermore, similar performance metrics (AUC, SE, and SP) are used to rank the classifiers. Finally, recursive feature elimination (RFE) was used to identify most significant features for the classification task.

In the next chapter, the results of the retrospective studies will be presented and discussions about the corresponding findings are given.

Chapter 5

Results and Discussion

In this chapter, the results of the evaluations of the developed methods which were described in the previous chapters are presented and the corresponding discussions about the findings are given. The majority of the findings that are illustrated in this chapter have been published before either as conference abstracts [41, 102, 104, 108] or as journal papers [42, 103, 105, 107] , or as a conference paper [106]. As discussed in the methodology chapter, the whole solution package, which serves as an automated pipeline to assist nuclear medicine (NM) physicians for the management of prostate cancer (PCa) patients, consists of several consecutive modules. Here, we present results which quantify the performance of the modules and illustrate strengths and highlight the limitations of the implemented methods. To this end, both quantitative and qualitative analyses are conducted. The results are quantified either in terms of accuracy metrics (i. e., area under the curve (AUC), sensitivity (SE), specificity (SP), and standard deviation (STD) of those metrics) for the machine learning classifiers or as performance factors such as precision, recall, and Dice coefficient for the automated segmentation pipeline. Moreover, qualitative analyses of the performance of different segmentation methods are further elaborated.

In the coming sections, first, the results achieved in previous work for the classification of pathological uptake in PSMA-PET/CT scans given the radiomics features calculated by third party tools are reported. Afterwards, to predict ^{177}Lu-PSMA treatment responses, the results of ML-based methods as applied to multimodal PET/CT radiomics features calculated either by third-party software or by in-house developed tools are illustrated and compared. Finally, results of uni- and multivariate methods to estimate the potential of PSMA-PET/CT radiomics for the prediction of overall survival are presented. The results which are adapted from previous published work will be cited accordingly.

5.1 Study Cohorts

A retrospective cohort of 100 subjects which had been histologically diagnosed as prostate cancer patients has been analyzed in different clinical experiments. For each patient, baseline ^{68}Ga-PSMA-PET/CT findings as well as the information on the clinical parameters in the period of ^{177}Lu-PSMA treatment had been provided. To provide annotated ground truth (GT) input to machine learning (ML) based pipelines, all the pathological and physiological uptake in the whole PET/CT dataset has been identified and manually delineated by experienced NM physicians. As a result, a total of 3553 hotspots including 2070 pathological vs 1482 physiological lesions have been defined and annotated. Furthermore, the 100-subjects cohort consisted of 67 responders vs 33 non-responders to ^{177}Lu-PSMA therapy. As described in the Methodology chapter, several retrospective clinical studies have been conducted in connection to this PhD thesis project. Due to the different hypotheses and objectives

TABLE 5.1: The distribution of patients and hotspots for the whole subject cohort as well as all the clinical study cohorts. FU: Follow-Up Res.: Responders, N-Res.: Non-Responders., Path.: Pathological, Phys.: Physiological.

Study Cohort	Patients' Distribution			Hotspots' Distribution		
	Res.	N-Res.	Total	Path.	Phys.	Total
Automated Segmentation [106]	67	33	100	2070	1482	3552
Hotspot Classification [107]	59	13	72	1629	790	2419
Hotspot Classification (FU) [42]	59	13	72	1629	790	2419
Therapy Response Prediction [103]	59	24	83	2070	0	2070
Survival Analysis [105]	59	24	83	2070	0	2070

of these clinical studies, also due to availability of annotated data to serve as ground truth, different groups of the subjects and hotspots have been analyzed in different studies. Table 5.1 gives an overview on the patient and hotspot distributions in the clinical study cohorts. Most of the results presented in the remaining sections of this chapter are directly transferred, summarized or adapted from the already published work.

5.2 Hotspot Classification with Manual Segmentation

For the task of discrimination of pathological from physiological uptake in manually delineated hotspots, two clinical studies had been conducted. The first study [107] aimed at providing a proof of concept about the significance of supervised ML methods to classify malignant tissues. The second study [42], which can be considered as a follow up analysis, focused in more detail on the application domain to evaluate the ML methods as applied to hotspots from organs such as glands which normally feature higher physiological uptake compared to other organs. In the subsections 5.2.1 and 5.2.2, we summarize the results from these two studies which have been already published in Diagnostics and Tomography journals respectively.

5.2.1 Hotspot Classification (Proof of Concept)

In this section, the findings from [107] is presented. The corresponding methods have been discussed in 4.7.2. In the first step, a total of 2419 focal tracer accumulations were manually delineated in our complete collective of 72 PCa patients. Out of these hotspots, 1629 and 790 lesions have been classified as pathological and physiological respectively. Based on these data, we applied the five ML classifiers (linear, radial basis function (RBF), and polynomial kernel SVM, extra trees (ET), and random forest (RF)) to the 48 training set patients which were randomly selected from the main cohort. Each ML-algorithm was applied on PET only, CT only, and PET/CT feature groups separately. To tune the hyperparameters for each combination of classifier and feature group, 5 fold cross validation is applied with grid search. For the detailed information on the CV results, refer to the original publication [107].

To avoid overfitting, the tuned classifiers were applied to the remaining 24 patients as the held-out test set. As shown in the figure 5.1 and table 5.2, the performance criteria improved as we compared PET with CT and PET/CT (up to: 98% AUC, 94% SE, 89% SP). Most interestingly, the CT feature group featured surprisingly good results. The results also suggest that the decision tree-based classifiers (RF and ET) outperformed the SVM-based methods, regardless of the subset used.

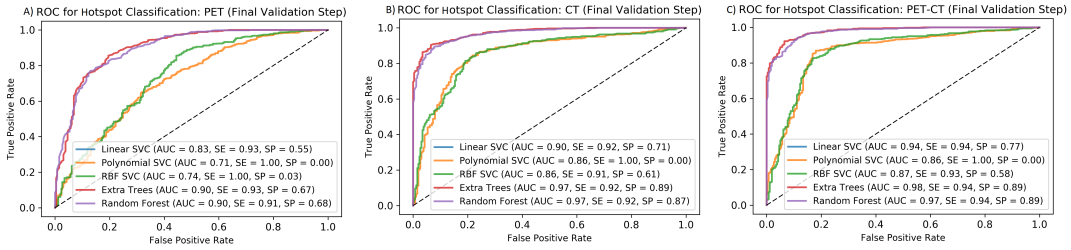


FIGURE 5.1: Results of final test step: ROC curves for five ML methods to predict hotspots labels on the test set using PET (A), CT (B), and all features (C). AUCs, sensitivities (SE), and specificities (SP) are shown for each ML method applied to each feature group. This figure was originally published in [107].

TABLE 5.2: The tuned parameters (from training step) and accuracy measures obtained for ML methods as applied to different feature groups in the final test step. This table is adapted from [107].

Feature Group	**PET**	
Classifier	**Tuned Parameters**	**AUC/SE/SP (%)**
Linear Kernel SVM	C = 0.5	83/93/55
Random Forest	max_depth = 30 min_samples_leaf = 1	90/91/68
Extra Trees	max_depth = 30 min_samples_leaf = 1	90/93/67
RBF Kernel SVM	$C = 2^{13}$ $gamma = 2^{-15}$	74/100/3
Polynomial Kernel SVM	C = 1 degree = 2	71/100/0
Feature Group	**CT**	
Classifier	**Tuned Parameters**	**AUC/SE/SP (%)**
Linear Kernel SVM	C = 1	90/92/71
Random Forest	max_depth = 20 min_samples_leaf = 1	97/92/87
Extra Trees	max_depth = 10 min_samples_leaf = 1	97/92/89
RBF Kernel SVM	$C = 2^{-5}$ $gamma = 2^{-15}$	86/91/61
Polynomial Kernel SVM	C = 1 degree = 2	86/100/0
Feature Group	**All**	
Classifier	**Tuned Parameters**	**AUC/SE/SP (%)**
Linear Kernel SVM	$C = 2^{11}$	94/94/77
Random Forest	max_depth = 20 min_samples_leaf = 1	97/94/89
Extra Trees	max_depth = 10 min_samples_leaf = 1	98/94/89
RBF Kernel SVM	$C = 2^{-3}$ $gamma = 2^{-3}$	87/93/58
Polynomial Kernel SVM	C = 1 degree = 2	86/100/0

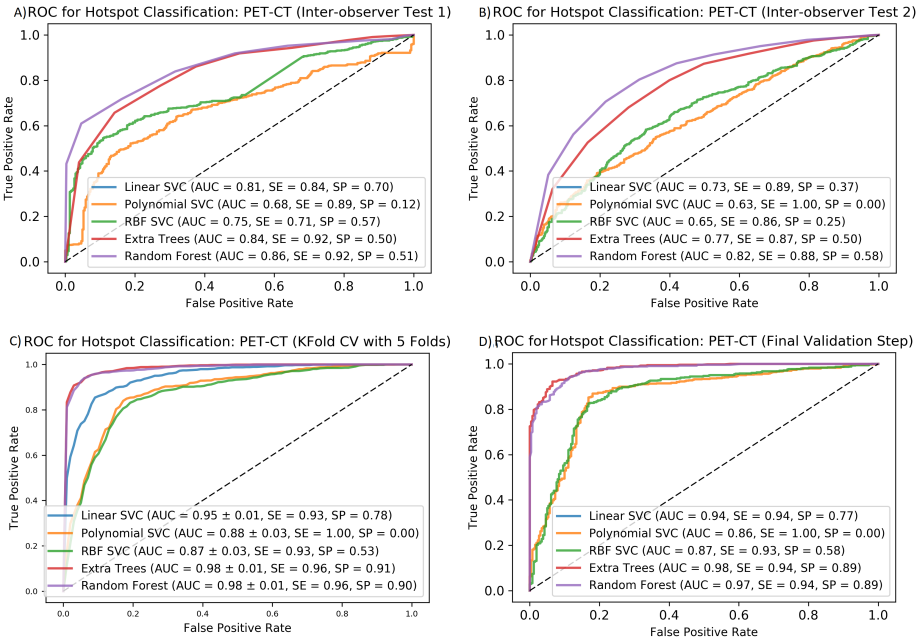


FIGURE 5.2: Mean ROC curves for five ML algorithms to classify hotspots using PET/CT features: results of the inter-observer test 1 (A), the inter-observer test 2 (B), the five-fold cross validation(C), and the final validation step (D). AUCs, sensitivities, and specificities are shown. This figure was originally published in [107].

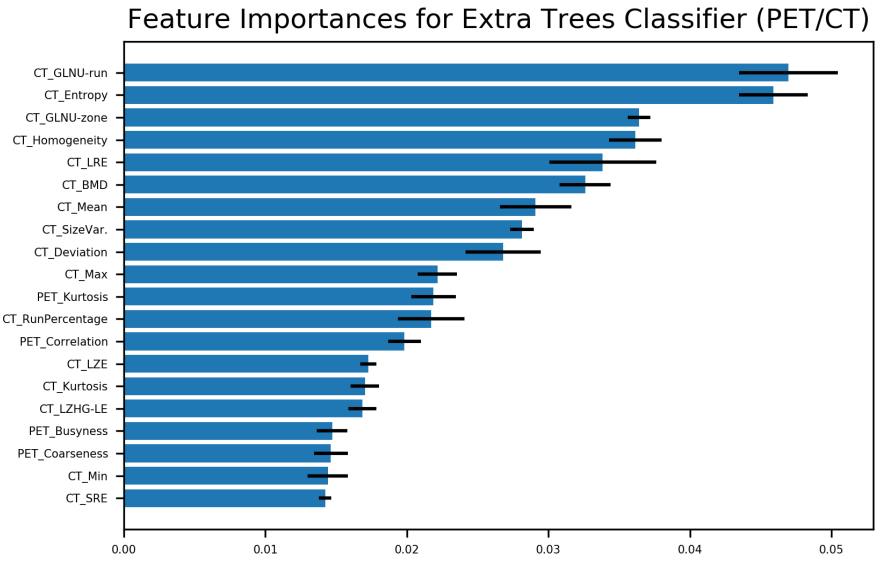


FIGURE 5.3: Best 20 features for hotspot classification based on extra trees classifier and five-fold cross validation. The error bars stand for standard deviation estimated for the CV folds (GLNU: GreyLevelNonUniformity, LRE: LongRunEmphasis, BMD: BoneMineralDensity, LZE: LongZoneEmphasis, LZHG_LE: LongZoneHighGreyLevelEmphasis, SRE: ShortRunEmphasis). This figure was originally published in [107].

The results of the quantitative inter-observer variability analyses are also reported in terms of AUC, SE, and SP obtained by the different ML classifiers. The results suggest that delineation of hotspots by different annotators does not markedly affect the AUCs and sensitivities. Thereby, the RF classifier was the most stable method. However, as we compared the same measures from CV or final validation steps to that of inter-observer analyses (see fig. 5.2), specificity seemed to be affected by the inconsistencies arised by inter-observer variability. To identify the most relevant features to the classification task, we took advantage of the feature importance metric provided by Extra Trees classifier. Figure 5.3 shows the top 20 best ranked features. As the interesting finding from the feature ranking, appearance of CT based features in the highest ranks would suggest the importance of the anatomical and morphological texture for the prediction of malignant tissues. Moreover, as expected, PET based textural heterogeneity based features such as kurtosis, busyness and coarseness possessed high ranks as well. In the final analysis step to assess the significance and generalizability of the findings, the permutation test was conducted which resulted in a p-value of 0.00076 after 25000 iterations.

To conclude, the findings of this study validated the assumption that the combination of the state-of-the-art supervised ML methods with the hotspot level radiomics features could help to identify malignant tissues in whole-body ^{68}Ga-PSMA-PET/CT scans with a sensitivity almost equal to an NM physician. In addition, the analysis of feature importances revealed that apart from PET-based textural heterogeneity parameters (which had been identified as significant in a previous related work [21]), some CT-based features significantly contribute to the classification task. Furthermore, the inter-observer analyses have identified a degree of subjectivity in assigning GT labels. This is a critical issue, specially for multicenter studies, and need to be investigated in the future.

Furthermore, the results have identified a room for improvement of specificity, urging for studies aiming at assessment of the cases in which the classifiers would most likely fail in true labeling of physiological uptake. As a result, a follow-up study has been conducted, the result of which is illustrated in section 5.2.2.

5.2.2 Hotspot Classification (Follow-Up)

In the follow up study [42], we further investigated the performance of ML classifiers on a new set of unseen data and also analyzed in more detail the specificity metric obtained by the algorithms. The corresponding results will be summarized in this section. The methodology of this study has been elaborated in 4.7.3.

To begin, the cross-validation step has identified the Extra Trees classifier as the dominant method compared to linear SVM and random forest classifiers. As shown in figure 5.4, the Extra Trees classifier trained with the data of 30 patients achieved 0.95 AUC, 0.95 sensitivity, and 0.80 specificity. As expected, expanding the size of the training cohort by increasing the sample size, one patient at a time, has improved the performance measures (0.98 AUC, 0.97 sensitivity, and 0.83 specificity) until it reached its overall maximum as the feature vectors from the last subject of the training cohort was added. Table 5.3 lists the mean and STD values of the performance metrics along all the training steps. Based on the training data of 72 subjects and the held-out test cohort of 15 subjects, 125 of the 128 lesions defined by the annotator as pathological were labeled as pathological by the ML method. This corresponds to the sensitivity of 0.97. Moreover, the physiological uptake was identified by the ML method with a high precision. The organ-specific analyses revealed that, while the ML algorithm made accurate predictions of liver, kidneys, gut, etc., the specificity

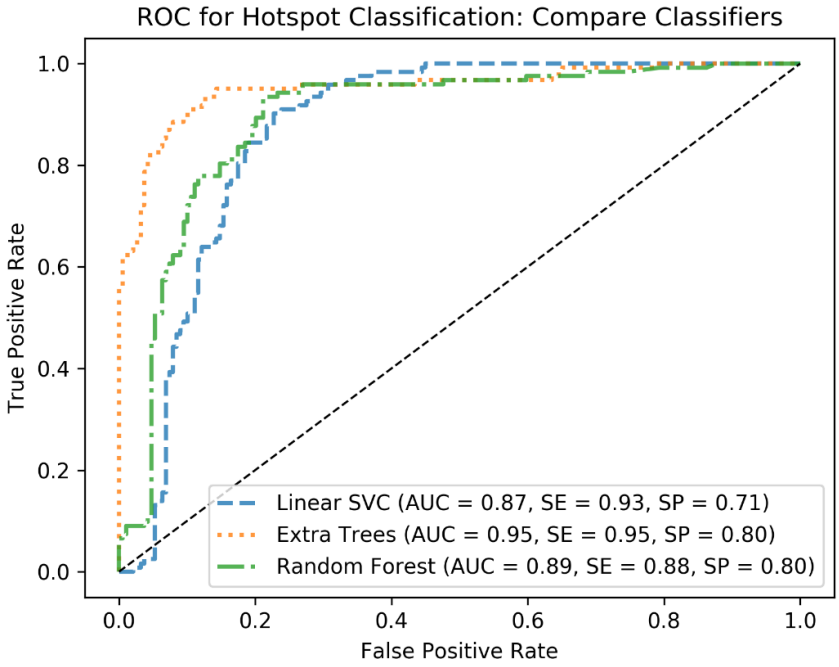


FIGURE 5.4: The receiver operating characteristic (ROC) curves to compare three classifiers. The classifiers are ranked after tuning in the cross-validation step and trained with the first training cohort with 30 subjects and then applied to the test cohort. This figure was originally published in [42].

TABLE 5.3: The mean and standard deviation (std) values of the area under the curves (AUCs), sensitivities, and specificities achieved as the training cohort was extended. This table was originally published in [42].

Accuracy Metric	Mean	Std
AUC	0.98	0.002
Sensitivity	0.97	0.004
Specificity	0.82	0.02

metric of 0.82 suggested that classifying glands correctly as physiological had been more challenging for the algorithm. Especially in sublingual and lacrimal glands, high rates of false positives (9/19 and 7/19) were obtained. The complete outline of the organ-specific analyses is shown in table 5.4. As pathological prostate uptake was only present in 14 subjects from the training cohort, we could not analyze the prediction performance on the test cohort for this category. However, analyzing the 14 prostate hotspots from the training cohort, we achieved a sensitivity of 0.92 (13/14 true positives).

As the results of the follow-up study suggest, the categorization of glands proved to be difficult for the ML classifiers. But, as the head is located far away from typical primary PCa tumor, this limitation of the methods seems to be reasonable for this feasibility study.

In conclusion, the findings from the two studies on hotspot classification based on manually segmented RoIs revealed the high capacity of the AI-based methods for the diagnostic task of malignancy detection and gave us the motivation to replace the

TABLE 5.4: The results of the predictions on the test cohort. (gut: gastrointestinal tract). This table was originally published in [42].

Category	1/Pathological	0/Physiological	Total	Specificity
bladder	0	13	13	1.00
glands	19	92	111	0.82
gut	1	32	33	0.97
liver	0	15	15	1.00
kidney	0	30	30	1.00
ureter	0	1	1	1.00
metastases	125	3	128	0.97

manual segmentation with automated segmentation as described in sections 4.3.2 and 5.5.

5.3 Response Prediction with Manual Segmentation

In this section, the summary of the findings from the clinical study to evaluate performance of supervised ML classifiers for the prediction of responders to ^{177}Lu-PSMA therapy based on patient level radiomics features calculated for manually annotated hotspots are presented. The results, including tables and figures, are adapted from the manuscript [103] published by annals of translational medicine (ATM). The methods of this study have been described in 4.7.4.

5.3.1 Linear Regression-Unbalanced Cohort

The results of the linear regression tests on the training set of the unbalanced cohort have identified 5 radiomics features from both PET (Min and Correlation) and CT (CT_Min, CT_Coarseness, and CT_Busyness) modalities as the best correlating features with PSA level difference as the indicator of the response to the treatment procedure ($p\_values < 0.05$). We named these best correlating features as Best-Radiomics group for further analyses.

5.3.2 Linear Regression-Balanced Cohort

The results of linear regression analyses on the training set of the balanced cohort have identified 3 radiomics features (PET_Min, CT_Busyness, and CT_Coarseness) and 3 clinical parameters (Alp1, Time difference, and Gleason score) as the best correlating features. From these 6 variables, two feature groups are formed: 1) Best-Radiomics including only the 3 best correlating radiomics features and 2) Best-Mixed including all the 6 features or parameters from both of the radiomics and clinical groups.

5.3.3 Classification-Unbalanced Cohort

The results of hyperparameter tuning in the CV step for the unbalanced cohort is shown in table 5.5. The classifiers were further tuned by the given values for the hyperparameters to be validated as applied to the held-out test set. To this end, the cohort of 56 subjects was used as the training data-set while the cohort of 27 subjects served as the test set. Table 5.6 and figure 5.5 illustrate the results of the final validation step for the unbalanced cohort. As the results suggest, the clinical parameters

TABLE 5.5: Results of hyperparameter tuning step, applying 3-Fold cross-validation (CV) for the unbalanced cohort: Tuned hyperparameters of the five ML classifiers on the four different feature or parameter groups on the unbalanced data-set of 56 subjects in the first validation step. This table is adapted from [103].

Feature Group	**Radiomics**	**Clinical**	**Mixed**	**Best-Radiomics**
Classifier	**Tuned Parameters**	**Tuned Parameters**	**Tuned Parameters**	**Tuned Parameters**
Linear Kernel SVM	C=2 gamma=0.001	C=1000 gamma=0.001	C=10 gamma=0.001	C=1 gamma=0.001
Polynomial Kernel SVM	C=1 degree=3	C=1 degree=3	C=1 degree=3	C=32768 degree=3
RBF Kernel SVM	C=1000 gamma=0.5	C=10 gamma=0.5	C=128 gamma=0.5	C=10 gamma=8
Extra Trees	max_depth=20 min_samples_leaf=10	max_depth=20 min_samples_leaf=10	max_depth=10 min_samples_leaf=8	max_depth=10 min_samples_leaf=10
Random Forest	max_depth=15 min_samples_leaf=10	max_depth=5 min_samples_leaf=4	max_depth=20 min_samples_leaf=8	max_depth=1 min_samples_leaf=10

TABLE 5.6: Results of validation step for the unbalanced cohort: Prediction scores of the five ML classifiers on the four different feature or parameter groups on the unbalanced data-set of 56 subjects in the first validation step. This table is adapted from [103].

Feature Group	**Radiomics**	**Clinical**	**Mixed**	**Best-Radiomics**
Classifier	**AUC/SE/SP (%)**	**AUC/SE/SP (%)**	**AUC/SE/SP (%)**	**AUC/SE/SP (%)**
Linear Kernel SVM	88/68/88	46/84/25	95/84/88	99/42/99
Polynomial Kernel SVM	99/58/99	28/63/25	99/84/99	53/58/50
RBF Kernel SVM	81/68/75	37/79/25	76/79/50	96/63/99
Extra Trees	41/11/99	57/79/50	55/16/99	99/21/99
Random Forest	68/26/99	53/95/12	69/32/99	99/53/99

group performed relatively poor, compared to the other feature groups. Furthermore, the results show that the polynomial kernel SVM with hyperparameters C=1 and degree=3 performed the best as applied to the mixture of all radiomics features and clinical parameters with 99% AUC, 84% SE, and 99% SP.

5.3.4 Classification-Balanced Cohort

As for the analyses of the unbalanced cohort, a CV step was conducted followed by a validation step on the balanced training and test cohorts of 32 and 16 subjects respectively. Table 5.7 presents the results of the hyperparameter tuning for the balanced cohort. Table 5.8 and figure 5.6 illustrate the results of the corresponding validation step as the classifiers were trained on the training cohort and tested on the test cohort. As results showed, except for the clinical parameters group which showed very poor performance, the linear, polynomial, and RBF kernel SVM classifiers performed well with 91% AUC, 99% SE, and 62% SP for linear SVM on radiomics group, 88% AUC, 99% SE, and 62% SP for polynomial SVM on radiomics group, and 80% AUC, 75% SE, and 75% SP for RBF SVM on Best-Mixed group.

Finally, the permutation test resulted in a p_value of 0.0043 that confirms the significance of the results.

To summarize the finding from this study, first, we showed that PET-based conventional parameters correlate with the treatment response indicator (ΔPSA) which conform on previous related work by Khurshid et al. [76]. We also identified the significant correlation of some CT-based features as well as some patient-specific clinical parameters with ΔPSA. Furthermore, for the task of response prediction to ^{177}Lu-PSMA treatment, the potential of supervised ML methods as applied to

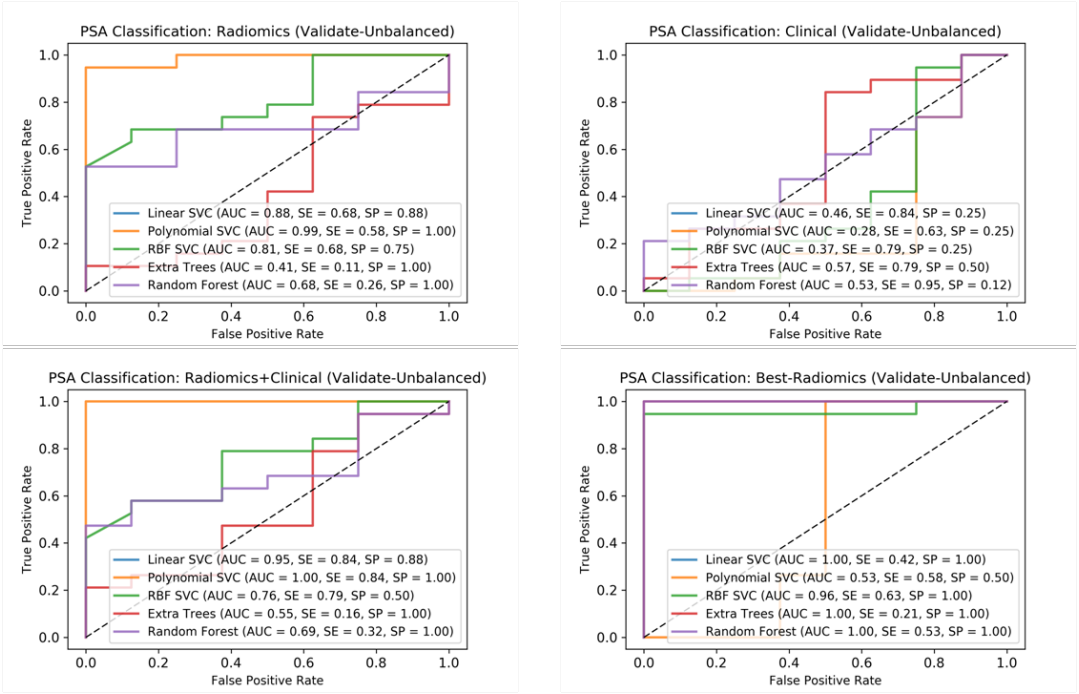


FIGURE 5.5: Receiver operating characteristic (ROC) curves for the final validation step on the unbalanced data-set. The four different diagrams are for the four different feature groups (Radiomics, Clinical, Radiomics and Clinical, and Best Radiomics). This figure was originally published in [103].

TABLE 5.7: Results of hyperparameter tuning step, applying 3-Fold cross-validation (CV) for the balanced cohort: Tuned hyperparameters of the five ML classifiers on the five different feature or parameter groups on the balanced data-set of 32 subjects in the second validation step. This table is adapted from [103].

Feature Group	**Radiomics**	**Clinical**	**Mixed**	**Best-Radiomics**	**Best-Mixed**
Classifier	**Tuned Parameters**	**Tuned Parameters**	**Tuned Parameters**	**Tuned Parameters**	**Tuned Parameters**
Linear Kernel SVM	C=1 gamma=0.001	C=100 gamma=0.001	C=10 gamma=0.001	C=1 gamma=0.001	C=32768 gamma=0.001
Polynomial Kernel SVM	C=1 degree=2	C=10 degree=3	C=10 degree=3	C=32768 degree=3	C=10 degree=3
RBF Kernel SVM	C=1 gamma=2	C=10 gamma=2	C=1 gamma=0.03125	C=100 gamma=0.001	C=100 gamma=8
Extra Trees	max_depth=5 min_samples_leaf=10	max_depth=5 min_samples_leaf=4	max_depth=10 min_samples_leaf=10	max_depth=25 min_samples_leaf=10	max_depth=10 min_samples_leaf=10
Random Forest	max_depth=1 min_samples_leaf=10	max_depth=5 min_samples_leaf=10	max_depth=10 min_samples_leaf=8	max_depth=5 min_samples_leaf=10	max_depth=10 min_samples_leaf=10

TABLE 5.8: Results of validation step for the balanced cohort: Prediction scores of the five ML classifiers on the five different feature or parameter groups on the balanced data-set of 32 subjects in the second validation step. This table is adapted from [103].

Feature Group	**Radiomics**	**Clinical**	**Mixed**	**Best-Radiomics**	**Best-Mixed**
Classifier	**AUC/SE/SP (%)**	**AUC/SE/SP (%)**	**AUC/SE/SP (%)**	**AUC/SE/SP (%)**	**AUC/SE/SP (%)**
Linear Kernel SVM	91/99/62	56/62/50	77/99/62	69/99/38	69/75/50
Polynomial Kernel SVM	88/99/62	58/75/50	75/75/62	80/88/50	75/75/62
RBF Kernel SVM	89/99/50	53/75/50	80/75/62	67/99/38	80/75/75
Extra Trees	86/88/50	45/50/38	80/99/50	68/75/50	61/62/38
Random Forest	80/88/50	42/62/25	81/99/50	71/88/38	75/99/25

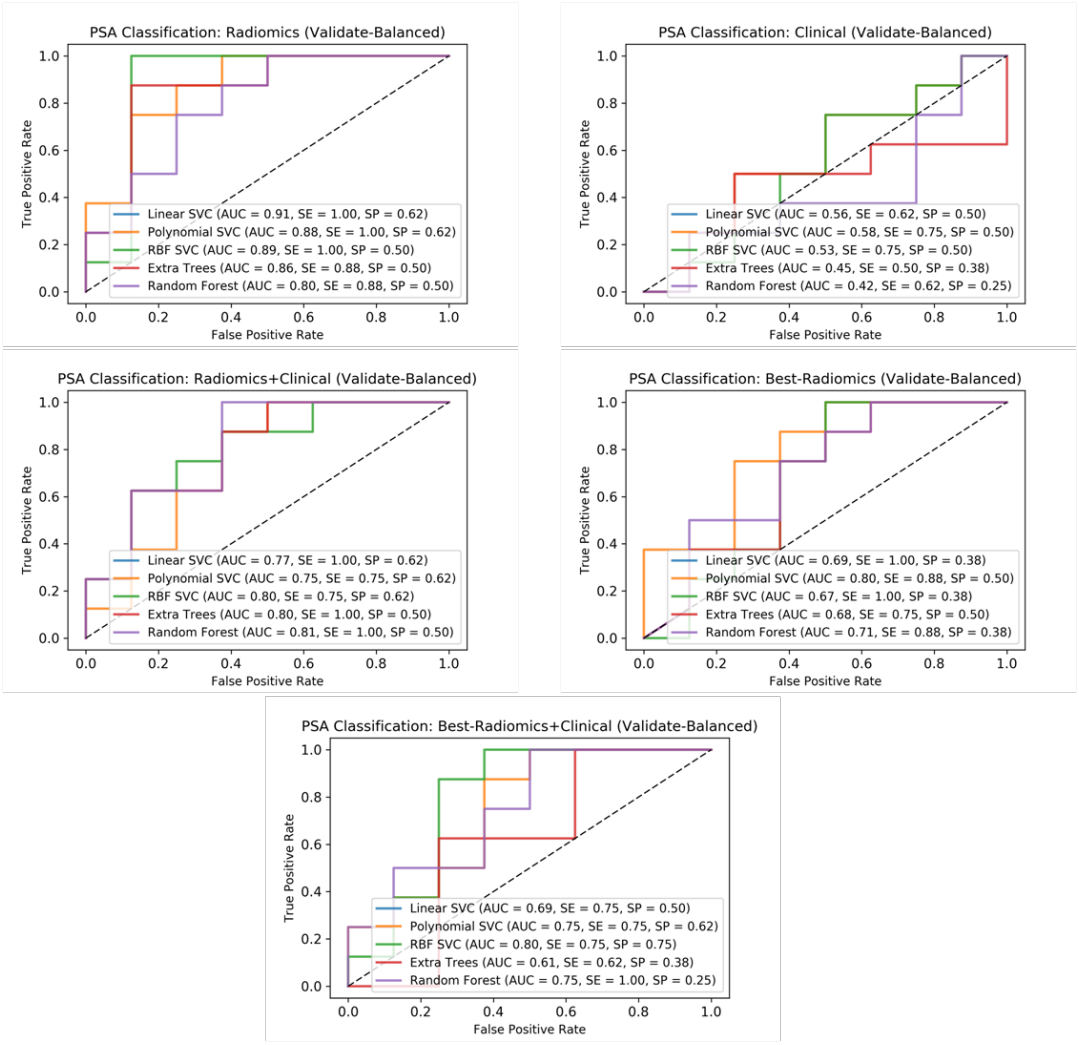


FIGURE 5.6: Receiver operating characteristic (ROC) curves for the final validation step on the balanced data-set. The five different diagrams are for the five different feature groups (Radiomics, Clinical, Radiomics and Clinical, Best Radiomics, and Best Mixed). This figure was originally published in [103].

radiomics features from manually annotated ^{68}Ga-PSMA-PET/CT scans has been shown. These findings have been another motivation to implement automated segmentation pipeline aiming to replace the manual annotation routine or serve as an assistant to the annotator.

5.4 Overall Survival Prediction

To estimate the potential of the radiomics features for the prediction of overall survival in patients with advanced prostate carcinoma, we already published a manuscript [105] at the Diagnostics journal. In this section, a summary of the results of this manuscript is provided. The corresponding implemented methods have been clarified in 4.7.5.

5.4.1 Selected Features and Radiomics Signature

As a result of multivariate survival analysis based on Cox proportional hazards model, the least absolute shrinkage and selection operator (LASSO) method has identified SUV_{Min} and kurtosis as the most correlating radiomics features with the overall survival (OS) quantified as number of months that the patient had lived until the censoring time or the patient had died. The LASSO method also calculated the correlation coefficients of 0.984 and -0.118 for SUV_{Min} and kurtosis respectively. These coefficients were then used to form the so-called radiomics signature for each patient:

$$RS_{(i)} = SUV_{Min(i)} \times 0.984 + Kurtosis_{(i)} \times (-0.118), \tag{5.1}$$

where the $RS_{(i)}$ is the radiomics signature for the subject number i and $SUV_{Min(i)}$ and $Kurtosis_{(i)}$ are the values of the selected variables for the subject number i.

5.4.2 Kaplan-Meier Statistics

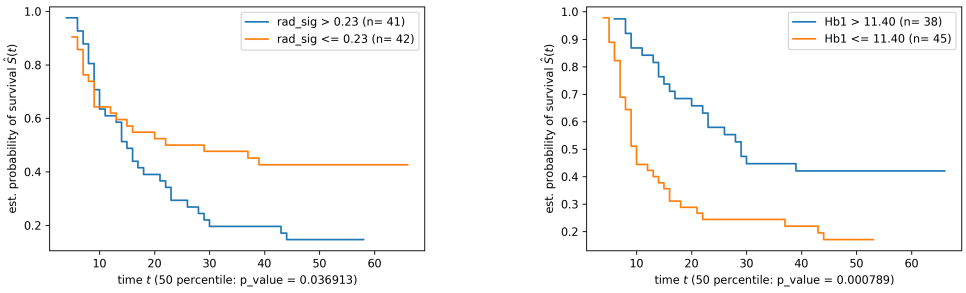
The results of the uni-variate survival analyses performed using Kaplan-Meier estimator (KME) are shown in figure 5.7. As the results revealed, SUV_{Min}, kurtosis, the calculated RS, SUV_{Mean}, as well as three clinical parameters (Hb1, CRP1, and ECOG1) showed higher potential for prediction of OS as they achieved p_value lower than 0.05. The definitions of Hb1, CRP1, and ECOG1 is given in 4.3.

To conclude, this study was aimed to assess the potential of radiomics features from baseline ^{68}Ga-PSMA-PET/CT for the prediction of OS and identified some parameters with high predictive power. To this end, both uni- and multivariate survival analysis techniques were applied. However, the main focus was to set a basis for future analyses, as the retrospective study cohort has been relatively small.

5.5 Response Prediction with Automated Segmentation

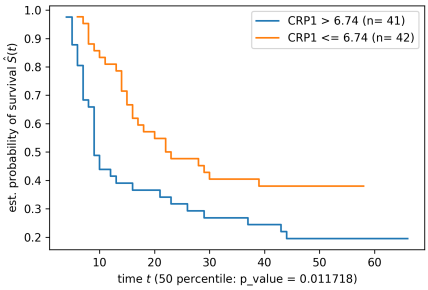
As described in the Methodology chapter (4.3.2, 4.7.6), a multi-channel U-Net based model, named PET-CT-U-Net, has been designed and evaluated for the task of automated segmentation of pathological uptake in whole body ^{68}Ga-PSMA-PET/CT scans. The summary of the results of a previous study which is already published in [106] will be presented in this section.

To provide GT masks for the purpose of training the PET-CT-U-Net, the whole dataset had been manually annotated in a slice-based approach by experienced

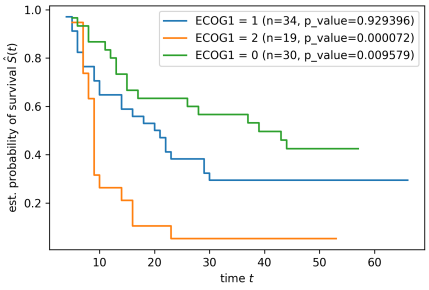


(A)

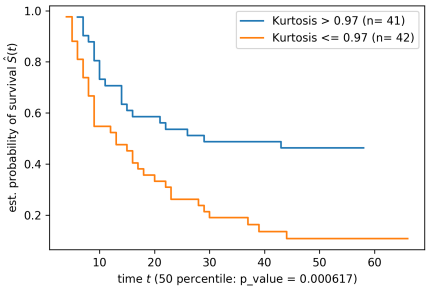
(B)



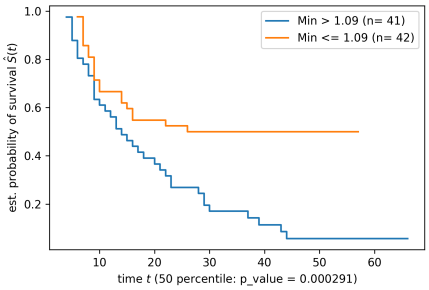
(C)



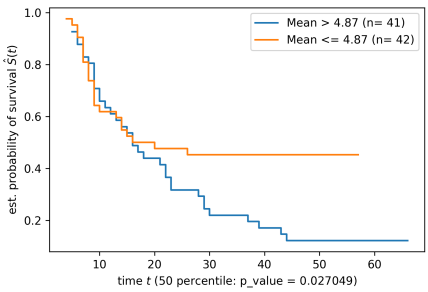
(D)



(E)



(F)



(G)

FIGURE 5.7: The results of Kaplan-Meier Analyses for (A) radiomics signature, (B) Hb1, (C) CRP1, (D) ECOG1, (E) Kurtosis, (F) SUV_{Min}, and (G) SUV_{Mean} (CRP1: C-reactive protein in serum at the first PSMA PET, Hb1: Hemoglobin level at the first PSMA PET, ECOG1: Scale of the performance status of the patient at the first PSMA PET). This figure has been originally published in [105].

TABLE 5.9: The performances of different U-Net based segmentation models as trained and fit with the training cohort and applied to the test cohort (as published in [106]). The performance of 40%-SUV$_{MAX}$ mask has been quantified for comparison. The precision, recall and Dice values are mean and standard deviations over the test subject cohort. (lr: learning rate, acc: accuracy).

Model/Mask	epochs	lr	acc	dice	loss	precision	recall
40%-SUV$_{MAX}$	–	–	99	39.62 ± 16.6	0.01	38.53 ± 21.38	51.48 ± 19.19
PET (Single)	35	0.001	99	71.51 ± 4.9	0.01	83.63 ± 5.3	63.38 ± 4.8
PET/CT (Dual)	32	0.001	99	82.18 ± 4.7	0.01	88.44 ± 4.8	77.09 ± 5.7

NMs. As a result, for each patient scan, all the pathological hotspots were defined as consecutive 2D RoIs in consecutive slices. The PET-CT-U-Net then took resampled PET and CT slices as input and predicted masks as binary images based on weighted cross-entropy loss with regards to the GT masks. As a side product of the automated segmentation network, 40%-SUV$_{MAX}$ masks are produced based on PET images. In a successive step, we took advantage of supervised ML classifiers to predict responders to ^{177}Lu-PSMA therapy using radiomics features calculated based on PET-CT-U-Net predicted masks.

To train and fit our segmentation model, both singular (i. e. just PET) and dual (PET + CT) input channels are used to predict binary masks as the pathological uptakes. As a result, the multi-channel model achieved the best performance results with batch size of 16, 0.99 test accuracy, 0.88 test precision, 0.75 test recall, and 0.81 test Dice. Table 5.9 summarizes the achieved performances from 40%-SUV$_{MAX}$ masks and the alternative U-Net models. Figure 5.8 illustrates a qualitative comparison of the segmentation results. The segmentation results suggest that the U-Net predicted masks perform reasonably well as compared to the GT masks. Furthermore, the U-Net predictions outperformed the 40%-SUV$_{MAX}$ masks, especially, for the identification of physiological uptakes (e. g., in livers and kidneys) which was shown to be a challenging task for ML-based algorithms [42, 107]. Moreover, as the results illustrate, the U-Net based network performs well in identification of bone metastases.

The results of treatment response prediction are summarized in Figure 5.9. Figure 5.9 compares the classifiers' performances as applied to GT and U-Net predicted masks with RFE respectively. Based on GT masks, logistic regression outperformed other classifiers with AUC=0.81, SE=0.70, SP=0.75 as applied to the held-out test set. To assess the performance of radiomics features calculated based on U-Net predicted masks and to identify most relevant features among them, recursive feature elimination technique has been applied for the classification task. Table 5.10 presents the list of 14 features as selected by RFE method. The results conform to our previous findings [103, 107] which denote significance of combination of PET and CT features for the diagnostic and prognostic classification tasks. In conclusion, the overall best performance belonged to the random forest classifier with AUC=0.73, SE=0.81, SP=0.58 as applied to the test cohort.

In summary, the findings of this study [106] suggest that the automated segmentation of pathological hotspots in whole-body ^{68}Ga-PSMA-PET/CT scans using the proposed PET-CT-U-Net model has the potential to replace manual segmentation in future. In doing so, in fact, study of alternative deep neural networks for the automated segmentation could be a track for further improvement of the current method. Moreover, for the task of therapy response prediction, supervised ML classifiers has

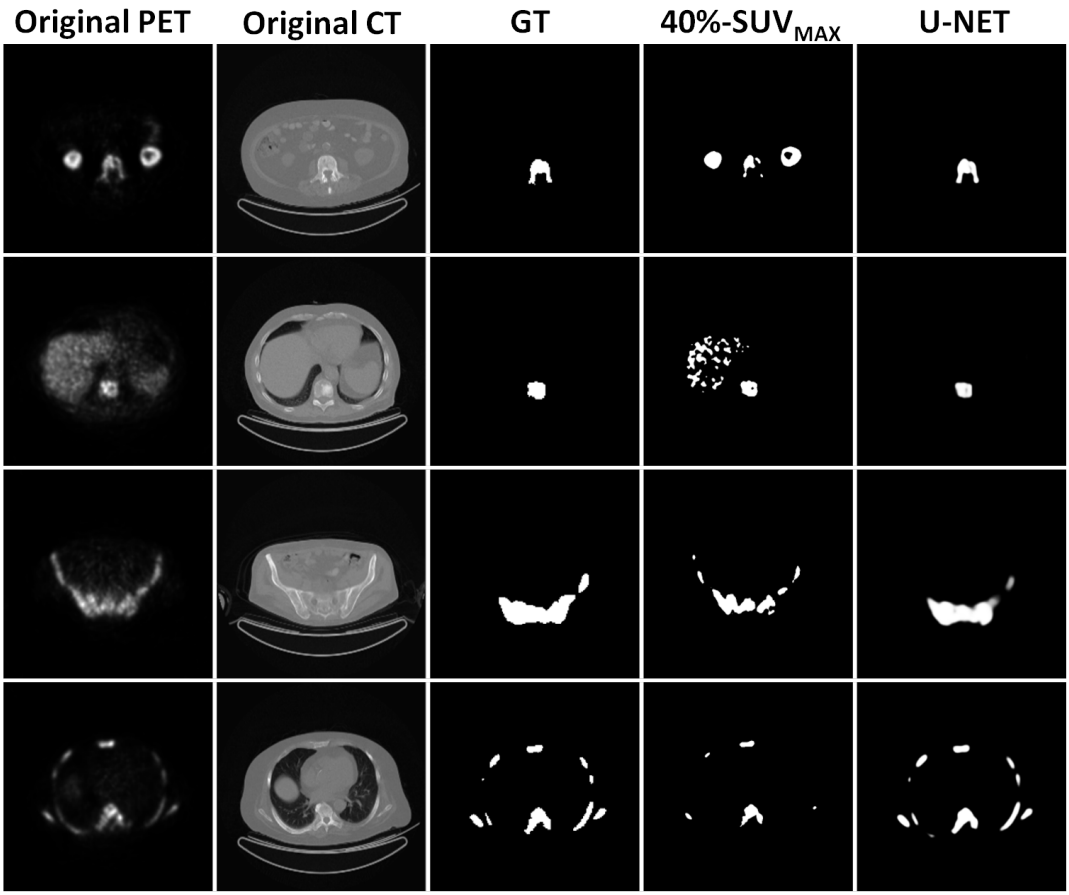


FIGURE 5.8: Example slices of the U-Net based segmentation results. The input PET and CT slices, the ground truth (GT), 40%-SUV$_{MAX}$ PET, and predicted masks are shown. Each row corresponds to an arbitrary 2D slice from an arbitrary subject of the test cohort. This figure has been originally published in [106].

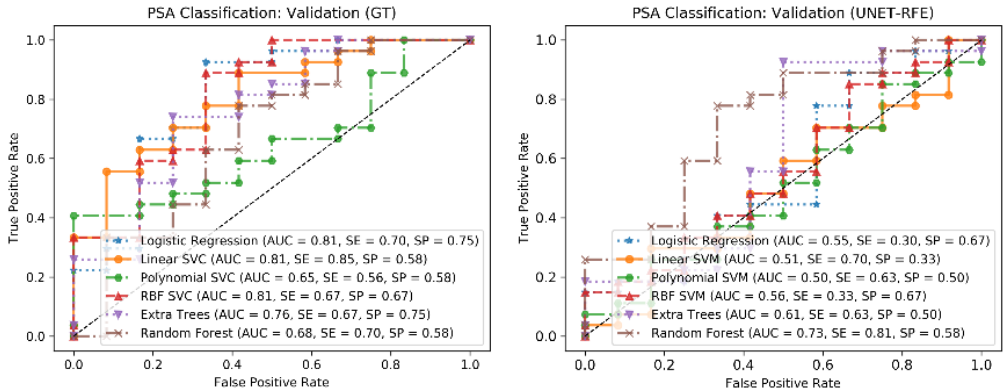


FIGURE 5.9: Receiver operating characteristic (ROC) curves based on GT masks and U-Net predicted masks with feature selection. The 6 classifiers are trained and tuned on the training set and applied to the test set (RBF: radial basis function, RFE: recursive feature elimination, AUC: area under the curve, SE: sensitivity, SP: specificity). This figure was originally published in [106].

TABLE 5.10: The most relevant radiomics features selected by recursive feature elimination (RFE) from both PET and CT modalities. For more information on the radiomics features, refer to 4.4.2 and [56] (glrm: gray level run length matrix, glszm: gray level size zone matrix). This table was originally published in [106].

Feature Group	Feature Subgroup	Feature Name
pet: diagnostics	Image-original	Mean
pet: original	shape	SurfaceVolumeRatio
ct: original	shape	MinorAxisLength
pet: original	firstorder	Energy
pet: original	firstorder	Maximum
pet: original	firstorder	Skewness
pet: original	glrlm	RunEntropy
pet: original	glrlm	RunLengthNonUniformityNormalized
pet: original	glrlm	RunPercentage
ct: original	glrlm	ShortRunEmphasis
ct: original	glszm	SmallAreaEmphasis
ct: original	glszm	SmallAreaLowGrayLevelEmphasis
pet: original	glszm	ZonePercentage
ct: original	glszm	ZonePercentage

shown their potential as applied to radiomics features from U-Net predicted RoIs. These preliminary findings prove the predictive power of the facilitated methods; however, studies with bigger multicentric datasets should be conducted to assess the generalizability of the findings.

5.6 Summary of the Findings

As detailed in previous chapters, several clinical experiments have been conducted to evaluate the methods developed in correspondence to this thesis. The principle idea behind this PhD thesis project was integrating machine learning techniques with radiomics features for diagnostic and prognostic tasks, focusing on the management of prostate cancer patients.

The first two studies [42, 107] aimed at providing a proof of concept for the AI-based methods to identify malignancies in the whole-body ^{68}Ga-PSMA-PET/CT scans. As a result, the predictive importance of ML methods as applied to radiomics features from manually annotated RoIs has been shown, holding the promise for the next coming implementation of a U-Net based segmentation method as described in 4.3.2 and evaluated in [106]. Furthermore, to quantify to which extent would the combination of ML methods with radiomics features be beneficial to predict responders to ^{177}Lu-PSMA treatment based on manual annotations, another study [103] has been conducted. As a matter of fact, the results from this latter study declared more motivation for the study on performance of a fully automated segmentation and response prediction pipeline as proposed in [106]. Last but not least, overall survival prediction based on baseline PET/CT scans has been another side product of our methods. The findings from the corresponding study [105] suggest that the radiomics signature calculated from ^{68}Ga-PSMA-PET/CT holds promise to predict survival likelihood of PCa patients.

To mention some of the important findings from the above-mentioned experiments, first of all, the potential of machine learning for diagnosis and prognosis given the example of prostate cancer disease has been revealed. Moreover, leveraging feature selection methods, our findings implied the significance of variables extracted from both PET and CT modalities as well as some conventional patient-specific clinical parameters which further emphasised relevance of multimodal approaches for clinical decision support. In addition, the automated pipeline for malignant uptake segmentation has shown its superiority to the standard thresholding based method. Here again, the multimodal PET/CT approach outperformed the uni-modal PET only model.

To conclude, although the results of the clinical studies confirm the potential of the integrated and implemented methods, certain improvements remain necessary for the future follow-ups. First, studies with bigger multicentric data cohorts should be conducted to further assess the generalizability of the findings. Additionally, alternative methods for feature extraction assessing the role of deep features should be considered. Furthermore, analyzing alternative convolutional neural networks such as densely connected convolutional networks (dense-Nets) [67] and generative adversarial nets (GANs) [54] could be another direction to proceed. Finally, having provided reasonably big subject cohorts, implementation of end to end deep learning based approaches to clinical decision support, in which the patient level prognosis and survival analyses are directly inferred from the baseline ^{68}Ga-PSMA-PET/CT scans would bring further attention to these findings.

Chapter 6

Conclusion

Artificial intelligence (AI) has reformed problem solving approaches in many scientific as well as industrial sectors in the last decades. On the one hand, compared to humans, computers and machines can be programmed to take over conducting routine tasks without getting tired. On the other hand, taking advantage of AI and machine learning (ML) algorithms, machines can "learn" from the input data by extracting the information and detecting the underlying patterns to predict outcomes, to make decisions, or to assist humans to do so.

Specifically, the medical domain has also benefited from state-of-the-art statistical and ML methods to deal with diagnostic and prognostic problems. For instance in the oncological domain, automated segmentation techniques facilitate tumor delineation leveraging artificial neural networks (ANNs). Furthermore, patient screening and treatment planning can be automatized taking advantage of supervised ML methods.

Machine learning techniques have been in use in disease prediction for a long time. Supervised ML methods in particular have been beneficial in diagnostic and prognostic tasks. Among others, support vector machine (SVM) classifiers have been the most popular methods in the domain [153]. However, specially focusing on studies with higher numbers of features, decision tree based algorithms with randomized kernels (e. g., random forests and extra trees) often outperformed SVMs in similar classification tasks [103, 107]. Deep convolutional neural networks (CNNs) are other common approaches for classification tasks which often outperform SVMs and forests of trees [163, 165]. However, they typically require considerably bigger cohorts to converge [162]. Therefore as elaborated in the Methodology chapter (4), we applied different classifier groups for different purposes. For instance, we compared SVMs with different kernels to decision tree based methods for hotspot and patient classification tasks in [42, 103, 106, 107] and fitted a multi-channel U-Net based model for automated hotspot segmentation in [106].

Nuclear medicine (NM) provides a variety of medical imaging modalities to help physicians from different expertises for diagnosis and prognosis. Specifically, in the oncological domain, multimodal imaging techniques play a critical role to stage the disease and to plan treatment. For instance, positron emission tomography/computed tomography (PET/CT) scans are widely used for prostate cancer (PCa) management. Depending on the disease stage, i. e., degrees of involvement of patients, various biomarkers such as prostate-specific membrane antigen (PSMA) and fluorodeoxyglucose (FDG) might be facilitated. In this thesis, we focused on ^{68}Ga-PSMA-PET/CT scans and used a retrospective dataset from a single NM center using a single PET/CT scanner. However, the findings need to be further compared to those of other scanners as well as other biomarkers such as FDG. To improve these preliminary results, both the U-Net based segmentation as well as the radiomics

analysis pipelines should be enhanced. Furthermore, to implement decision support tools which can take part in clinical routines in near future, we plan to include PET/CT images from different scanners and centers as well as other biomarkers.

Considering common approaches to assess image based metrics in oncological research, previously, quantities such as standardized uptake value (SUV), metabolic tumor volume (MTV), and total lesion glycolysis (TLG) have been in focus in many clinical studies examining PET scans which mostly applied single variable statistical analysis techniques such as linear regression. In contrast, radiomics analysis has brought a broader horizon by introducing hundreds of 2D and 3D texture- and shape-based features for different medical imaging modalities including PET and CT. This justifies application of state-of-the-art ML techniques which facilitate integration of multiple features and parameters for diagnostic and prognostic tasks. It is also worth mentioning that coping with the high numbers of radiomics features might arise new challenges such as overfitting which should be taken care of. To this end, proper cross validation (CV) and feature selection techniques need to be applied. In this thesis, we analyzed radiomics features in combination with different ML based predictive scenarios. For example, we applied CV and feature selection for classification of malignant tissues [42, 107] and prediction of treatment outcome [103, 106].

Apart from conventional parameters and radiomics features, deep features (as sub-products of deep and convolutional neural netwrorks) can be leveraged to conduct prognostic analyses. For instance, Andrearczyk et al. [6] compared radiomics featuress to deep features from a multi-task deep neural network to predict disease-free survival from FDG-PET/CT scans in patients diagnosed with head and neck cancer. As in this thesis we mostly focused on conventional parameters and radiomics features, a possible future work would be to conduct studies assessing the relevance of deep features for therapy response prediction and survival analysis based on multimodal ^{68}Ga-PSMA-PET/CT images from patients with prostate cancer.

Facilitating fast and accurate non-invasive diagnosis and prognosis has been the objective of computer-aided diagnosis (CAD) for years. When it comes to the oncological domain, especially in subjects in advanced metastatic stages, CAD systems take over the histopathological analyses in many clinical practices. This is globally justified as taking multiple biopsies from patients is ethically questionable. Thus, many interactive visualization and annotation tools are provided to assist physicians to locate and delineate regions and volumes of interest manually. However, the procedure of manual delineation of the malignant tissues using established tools such as InterView FUSION is considered time consuming and attention intensive. This limitation motivated us to facilitate automated segmentation tools in practice. Therefore, one goal of this thesis was to develop an automated segmentation tool for multimodal PET/CT scans.

Focusing on PET/CT image segmentation techniques, apart from manual segmentation, thresholding based methods can be applied to both PET and CT scans. On the one hand, SUV based fixed and adaptive thresholding methods such as 40%-SUV$_{MAX}$ are widely applied. On the other hand, CT thresholding based on Hounsfield scale is commonly used to locate malignancies. We also took advantage of 40%-SUV$_{MAX}$ and Hounsfield units in our methods to set a comparison basis for automated segmentation.

As a proof of concept, prior to the development of the automated segmentation tool, we analyzed radiomics features calculated by third party software from manually annotated PET/CT scans and applied supervised machine learning classifiers

to identify pathological uptake in PCa patients. This effort has resulted in preliminary studies [42, 107] which revealed the potential of ML based approach to identify with high accuracy the malignant tissues either as primary uptake in the prostate or metastatic uptake in bone and lymph nodes. Most interestingly, the combination of features from PET and CT modalities outperformed the classifiers' performances as applied to features from single modalities [107]. However, limitations were observed, especially to truly classify physiological uptake in small glands [42]. This is an important topic which need to be further investigated in the future.

Furthermore, in [106], we retrospectively analyzed 2067 pathological hotspots from 100 PCa patients (on average, 20 pathological hotspots per patient). As shown in the Results chapter, our U-Net based multi-channel segmentation network predicts the pathological masks with a high accuracy. Particularly, we showed that including the PET and CT modalities as multiple channels outperforms predictions of the U-Net model as trained only using the PET channel. Also, the qualitative analyses revealed that the multi-channel U-Net prediction is superior in discriminating non-pathological uptake in liver and kidneys compared to 40%-SUV$_{MAX}$ mask as a conventional threshold based method.

Predicting ^{177}Lu-PSMA therapy response has been another goal of this thesis. To analyze the potential of supervised ML classifiers for prediction of treatment response, in a preliminary study [103], we have shown the potential of per patient averaged radiomics features calculated from manually segmented lesions. Radiomics analysis has been successfully used in oncological research for treatment response prediction and analysis of overall survival [103, 105, 154, 167]. Thus, as another contribution of our methods, we combined automated segmentation with radiomics analysis for multimodal ^{68}Ga-PSMA-PET/CT findings. To this end, we calculated radiomics features based on the U-Net predicted masks [106].

Based on the results achieved in [103, 106], the potential of a fully automated approach has been revealed, even though the comparison to predictions based on manual segmentation still implies room for improvement. As a track of future work, we plan to explore an end to end prediction of therapy response using deep neural networks. However, we expect that successfully training such an approach might require a larger cohort. Furthermore, the specificity metric was shown to be a bottleneck in predicting responders to ^{177}Lu-PSMA treatment. A probable cause of this issue could be the in-balanced characteristic of the subject cohort as the most of the patients visiting our NM facility for PCa disease follow-ups are responders to treatment. Therefore, possible future follow-ups should consider solutions to enhance the corresponding prediction performance. One solution could be providing bigger cohorts including more non-responders. Also, integrating alternative CNN models such as generative adversarial nets (GANs) [54] and densely connected convolutional networks (dense-Nets) [67] could enhance the performance of the automated segmentation and consecutively that of the treatment response prediction.

Prediction of overall survival has been another product of the facilitated methods in this thesis. In [105], we applied both uni- and multivariate analysis methods for the prediction of overall survival for patients suffering from advanced prostate carcinoma. For this specific study, we took advantage of Cox proportional hazards model and Kaplan Meier Estimator. We also applied the LASSO method to calculate the so-called radiomics signature (RS). The results revealed the potential of the calculated RS and a couple of clinical parameters for survival prediction. But as a matter of fact, in the future, studies with larger patient cohorts from different tracers and centers need to be conducted to further assess the generalizability of the survival prediction pipeline.

To conclude, the main objective of this PhD thesis was to provide automated clinical decision support solutions for the management of patients with advanced prostate carcinoma to replace pre-existing time consuming and attention intensive diagnostic and prognostic pipelines. Despite some limitations and drawbacks regarding the clinical cohorts used for the assessment of the integrated methods, the results of the retrospective studies suggest the potential of the provided pipeline for future use. The pipelines include several consecutive modules aiming at facilitating manual and automated tools from visualization of volumes of interest (VoIs) to the analysis of ultimate prognostic outlines. As presented in the Related Work chapter (3), there had been so many solutions which address individual modules of such a fully automated pipeline. However, we believe that AutoPyPetCt is the first fully automated pipeline for management of PCa patients from visualization to diagnosis and prognosis, based on ^{68}Ga-PSMA-PET/CT scans. To further empower AutoPyPetCt to serve as a CDSS in clinical routine, studies with multicentric data which address different drawbacks of the integrated methods (as already discussed in this chapter) need to be conducted.

Bibliography

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. "TensorFlow: Large-scale machine learning on heterogeneous systems". URL: http://tensorflow.org.

[2] M. Abderrahim, A. Baâzaoui, and W. Barhoumi. "Comparative Study of Relevant Methods for MRI/X Brain Image Registration". In: *The Impact of Digital Technologies on Public Health in Developed and Developing Countries*. Ed. by M. Jmaiel, M. Mokhtari, B. Abdulrazak, H. Aloulou, and S. Kallel. Cham: Springer International Publishing, 2020, pp. 338–347. ISBN: 978-3-030-51517-1.

[3] H. Abdollahi, B. Mofid, I. Shiri, A. Razzaghdoust, A. Saadipoor, A. Mahdavi, H. M. Galandooz, and S. R. Mahdavi. "Machine learning-based radiomic models to predict intensity-modulated radiation therapy response, Gleason score and stage in prostate cancer". In: *Radiol med* 124 (2019), 555–567. DOI: 10.1007/s11547-018-0966-4. URL: https://doi.org/10.1007/s11547-018-0966-4.

[4] R. Adams and L. Bischof. "Seeded region growing". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16.6 (1994), pp. 641–647. DOI: 10.1109/34.295913.

[5] H. Ahmadzadehfar, S. Wegen, A. Yordanova, R. Fimmers, S. Kürpig, E. Eppard, X. Wei, C. Schlenkhoff, S. Hauser, and M. Essler. "Overall survival and response pattern of castration-resistant metastatic prostate cancer to multiple cycles of radioligand therapy using [177Lu]Lu-PSMA-617". In: *Eur J Nucl Med Mol Imaging* 44 (2017), pp. 1448–54. DOI: 10.1007/s00259-017-3716-2. URL: https://doi.org/10.1007/s00259-017-3716-2.

[6] V. Andrearczyk, P. Fontaine, V. Oreiller, J. Castelli, M. Jreige, J. O. Prior, and A. Depeursinge. "Multi-task Deep Segmentation and Radiomics for Automatic Prognosis in Head and Neck Cancer". In: *Predictive Intelligence in Medicine*. Ed. by I. Rekik, E. Adeli, S. H. Park, and J. Schnabel. Cham: Springer International Publishing, 2021, pp. 147–156. ISBN: 978-3-030-87602-9. DOI: 10.1007/978-3-030-87602-9.

[7] U. I. Attenberger and G. Langs. "How does Radiomics actually work? - Review". In: *Rofo* 2 (2020). DOI: 10.1055/a-1293-8953.

[8] A. Baazaoui, W. Barhoumi, E. Zagrouba, and R. Mabrouk. "A Survey of PET Image Segmentation: Applications in Oncology, Cardiology and Neurology". In: *Current Medical Imaging* 12.13 (2016). DOI: 10.2174/1573405612666151203204003.

[9] D. L. Bailey, D. W. Townsend, P. E. Valk, and M. N. Maisey. *Positron Emission Tomography*. London: Springer-Verlag London Limited, 2005. ISBN: 978-1-84628-007-8. DOI: 10.1007/b136169. URL: https://doi.org/10.1007/b136169.

[10] C. Ballangan, X. Wang, S. Eberl, M. Fulham, and D. Feng. "Automated lung tumor segmentation for whole body PET volume based on novel downhill region growing". In: *Medical Imaging 2010: Image Processing*. Ed. by B. M. Dawant and D. R. Haynor. Vol. 7623. International Society for Optics and Photonics. SPIE, 2010, pp. 1120 –1127. DOI: 10.1117/12.844032. URL: https://doi.org/10.1117/12.844032.

[11] J. I. Bang, S. Ha, Kang S. B., K. W. Lee, H. S. Lee, J. S. Kim, H. K. Oh, H. Y. Lee, and S. E. Kim. "Prediction of neoadjuvant radiation chemotherapy response and survival using pretreatment [18F]FDG PET/CT scans in locally advanced rectal cancer." In: *Eur J Nucl Med Mol Imaging*. 43 (2016), 422–431. DOI: 10.1007/s00259-015-3180-9. URL: https://doi.org/10.1007/s00259-015-3180-9.

[12] R. A. Baxter. "Mixture Model". In: *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G. I. Webb. Boston, MA: Springer US, 2010, pp. 680–682. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_547. URL: https://doi.org/10.1007/978-0-387-30164-8_547.

[13] Y. Bengio. "Learning Deep Architectures for AI". In: *Foundations and Trends in Machine Learning* 2 (2009), p. 1. DOI: 10.1561/2200000006.

[14] R. J. Beukinga, J. B. Hulshoff, L. V. van Dijk, C. T. Muijs, J. G. M. Burgerhof, G. Kats-Ugurlu, R. H. J. A. Slart, C. H. Slump, V. E. M. Mul, and J. T. M. Plukker. "Predicting Response to Neoadjuvant Chemoradiotherapy in Esophageal Cancer with Textural Features Derived from Pretreatment 18F-FDG PET/CT Imaging". In: *Journal of Nuclear Medicine* 58.5 (2017), pp. 723–729. ISSN: 0161-5505. DOI: 10.2967/jnumed.116.180299. eprint: https://jnm.snmjournals.org/content/58/5/723.full.pdf. URL: https://jnm.snmjournals.org/content/58/5/723.

[15] K. J. Biehl, F. M. Kong, F. Dehdashti, J. Y. Jin, S. Mutic, I. El Naqa, B. A. Siegel, and J. D. Bradley. "18F-FDG PET Definition of Gross Tumor Volume for Radiotherapy of Non–Small Cell Lung Cancer: Is a Single Standardized Uptake Value Threshold Approach Appropriate?" In: *Journal of Nuclear Medicine* 47.11 (2006), pp. 1808–1812. ISSN: 0161-5505. eprint: https://jnm.snmjournals.org/content/47/11/1808.full.pdf. URL: https://jnm.snmjournals.org/content/47/11/1808.

[16] Q. C. Black, I. S. Grills, L. L. Kestin, C. Y. Wong, J. W. Wong, A. A. Martinez, and D. Yan. "Defining a radiotherapy target with positron emission tomography." In: *Int J Radiat Oncol Biol Phys.* 60.4 (2004), pp. 1272–82. DOI: 10.1016/j.ijrobp.2004.06.254.

[17] V. Bourbonne, M. Vallières, F. Lucia, L. Doucet, D. Visvikis, V. Tissot, O. Pradier, M. Hatt, and U. Schick. "MRI-Derived Radiomics to Guide Post-operative Management for High-Risk Prostate Cancer". In: *Frontiers in oncology* 9 (2019), p. 807. DOI: 10.3389/fonc.2019.00807. URL: https://doi.org/10.3389/fonc.2019.00807.

[18] L. Breiman. "Random Forests". In: *Machine Learning* 45 (2001), pp. 5–32. URL: https://doi.org/10.1023/A:1010933404324.

[19] N. E. Breslow. "Analysis of Survival Data under the Proportional Hazards Model". In: *International Statistical Review / Revue Internationale de Statistique.* 43 (1975), 45–57. DOI: 10.2307/1402659.

[20] K. K. Brock, S. Mutic, T. R. McNutt, H. Li, and M. L. Kessler. "Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132". In: *Medical Physics* 44.7 (2017), e43–e76. DOI: https://doi.org/10.1002/mp.12256. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1002/mp.12256. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/mp.12256.

[21] R. A. Bundschuh, J. Dinges, L. Neumann, M. Seyfried, N. Zsótér, L. Papp, R. Rosenberg, K. Becker, S. T. Astner, M. Henninger, K. Herrmann, S. I. Ziegler, M. Schwaiger, and M. Essler. "Textural Parameters of Tumor Heterogeneity in 18F-FDG PET/CT for Therapy Response Assessment and Prognosis in Patients with Locally Advanced Rectal Cancer". In: *Journal of Nuclear Medicine* 55.6 (2014), pp. 891–897. ISSN: 0161-5505. DOI: 10.2967/jnumed.113.127340. eprint: https://jnm.snmjournals.org/content/55/6/891.full.pdf. URL: https://jnm.snmjournals.org/content/55/6/891.

[22] P. Bühlmann and S. Van De Geer. *Statistics for High-Dimensional Data*. Springer Series in Statistics, 2011. ISBN: 978-3-642-20191-2. DOI: 10.1007/978-3-642-20192-9.

[23] J. C. Jr. Callison, R. C. Walker, and P. P. Massion. "Somatostatin Receptors in Lung Cancer: From Function to Molecular Imaging and Therapeutics". In: *Journal of lung cancer* 10.2 (2011), 69–76. DOI: 10.6058/jlc.2011.10.2.69. URL: https://doi.org/10.6058/jlc.2011.10.2.69.

[24] Q. Chen, T. Leng, L. Zheng, L. Kutzscher, J. Ma, L. de Sisternes, and Rubin D. L. "Automated drusen segmentation and quantification in SD-OCT images." In: *Med Image Anal.* 17.8 (2013), pp. 1058–72. DOI: 10.1016/j.media.2013.06.003.

[25] J. Z. Cheng, D. Ni, Y. H. Chou, J. Qin, C. M. Tiu, Y. C. Chang, C. S. Huang, D. Shen, and C. M. Chen. "Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans". In: *Scientific reports* 6.24454 (2016). DOI: 10.1038/srep24454. URL: https://doi.org/10.1038/srep24454.

[26] S. J. Chiu, J. A. Izatt, R. V. O'Connell, K. P. Winter, C. A. Toth, and S. Farsiu. "Validated automatic segmentation of AMD pathology including drusen and geographic atrophy in SD-OCT images." In: *Invest Ophthalmol Vis Sci.* 53.1 (2012), pp. 53–61. DOI: 10.1167/iovs.11-7640.

[27] Y. Choi, B. M. Gil, M. H. Chung, W. J. Yoo, N. Y. Jung, Y. H. Kim, S. S. Kwon, and J. Kim. "Comparing attenuations of malignant and benign solitary pulmonary nodule using semi-automated region of interest selection on contrast-enhanced CT." In: *J Thorac Dis.* 11.6 (2019), pp. 2392–2401. DOI: 10.21037/jtd.2019.05.56.

[28] F. Chollet. "Keras. GitHub repository". accessed on 31 August 2021. URL: https://github.com/fchollet/keras.

[29] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman. "Survival analysis part I: basic concepts and first analyses". In: *British journal of cancer* 89.2 (2003), 232–238. DOI: 10.1038/sj.bjc.6601118.

[30] C. Cortes and V. Vapnik. "Support-vector networks". In: *Machine Learning* 20.3 (1995), pp. 273–297. DOI: {10.1007/BF00994018}. URL: https://doi.org/10.1007/BF00994018.

[31] T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. USA: Wiley-Interscience, 2006. ISBN: 0471241954.

[32] D. R. Cox. "Regression Models and Life-Tables". In: *Journal of the Royal Statistical Society, Series B*. 34.2 (1972), 187–220.

[33] D. R. Cox. "The regression analysis of binary sequences". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 20.2 (1958), pp. 215–232. URL: https://doi.org/10.1111/j.2517-6161.1958.tb00292.x.

[34] D. R. Cox and D. Oakes. *Analysis of Survival Data*. New York, NY, USA: Chapman and Hal, 1984.

[35] N. Cristianini and B. Scholkopf. "Support Vector Machines and Kernel Methods: The New Generation of Learning Machines." In: *AI Magazine* 23.3 (), p. 31. DOI: 10.1609/aimag.v23i3.1655.

[36] M. C. F. Cysouw, B. H E. Jansen, T. van de Brug, D. E. Opera-Lager, E. Pfaehler, B. M. de Vries, R. J. van Moorselaar, O. S. Hoekstra, A. N. Vis, and R. Boellaard. "Machine learning-based analysis of [18F]DCFPyL PET radiomics for risk stratification in primary prostate cancer". In: *Eur J Nucl Med Mol Imaging* 48 (2021), pp. 340–9. DOI: 10.1007/s00259-020-04971-z. URL: https://doi.org/10.1007/s00259-020-04971-z.

[37] E. Day, J. Betler, D. Parda, B. Reitz, A. Kirichenko, S. Mohammadi, and M. Miften. "A region growing method for tumor volume segmentation on PET images for rectal and anal cancer patients." In: *Med Phys.* 36.10 (2009), pp. 4349–58. DOI: 10.1118/1.3213099.

[38] M. P. Deisenroth, A. A. Faisal, and C. S. Ong. *Mathematics for Machine Learning*. Cambridge University Press., 2020. ISBN: 9781108455145. URL: https://mml-book.github.io/.

[39] T. M. Deist, Dankers F. J. W. M., G. Valdes, R. wijsman, I. C. Hsu, C. Oberije, T. Lustberg, J. van Soest, F. Hoebers, A. Jochems, I. El Naqa, et al. "Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers." In: *Med Phys* 45 (2018), pp. 3449–59. DOI: 10.1002/mp.12967. URL: https://doi.org/10.1002/mp.12967.

[40] T. D. DenOtter and Schubert J. "Hounsfield Unit". In: *StatPearls [Internet]*. [Updated 2020 May 11]. Treasure Island (FL): StatPearls Publishing, 2021. URL: https://www.ncbi.nlm.nih.gov/books/NBK547721/.

[41] A. Erle, S. Moazemi, M. Essler, T. Schultz, and R. A. Bundschuh. "Evaluating a machine learning based tool for the detection of pathological hotspots in whole-body PSMA-PET-CT scans". In: *Nuklearmedizin* 59.02 (2020), pp. 99–100. DOI: 10.1055/s-0040-1708151.

[42] A. Erle, S. Moazemi, S. Lütje, M. Essler, T. Schultz, and R. A. Bundschuh. "Evaluating a Machine Learning Tool for the Classification of Pathological Uptake in Whole-Body PSMA-PET-CT Scans". In: *Tomography* 7.3 (2021), pp. 301–312. DOI: 10.3390/tomography7030027. URL: https://doi.org/10.3390/tomography7030027.

[43] T. Fawcett. "An introduction to ROC analysis". In: *Pattern Recognition Letters* 27.8 (2006). ROC Analysis in Pattern Recognition, pp. 861–874. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010. URL: https://www.sciencedirect.com/science/article/pii/S016786550500303X.

[44] J. Ferlay, F. Lam, M. Colombet, L. Mery, M. Pineros, A. Znaor, I. Soerjomataram, and F. Bray. "Global cancer observatory: cancer today. Lyon, France: International Agency for Research on Cancer." 2021. URL: https://gco.iarc.fr/today.

[45] B. Fischl. "FreeSurfer". In: *NeuroImage* 62.2 (2012), 774–781. DOI: 10.1016/j.neuroimage.2012.01.021. URL: https://doi.org/10.1016/j.neuroimage.2012.01.021.

[46] B. Foster, U. Bagci, A. Mansoor, Z. Xu, and Mollura D. J. "A review on segmentation of positron emission tomography images." In: *Comput Biol Med.* 50 (2014), pp. 76–96. DOI: 10.1016/j.compbiomed.2014.04.014.

[47] P. K. Gadosey, Y. Li, E. A. Agyekum, T. Zhang, Z. Liu, P. T. Yamak, and F. Essaf. "SD-UNet: Stripping down U-Net for Segmentation of Biomedical Images on Platforms with Low Computational Budgets". In: *Diagnostics* 10.2 (2020). ISSN: 2075-4418. DOI: 10.3390/diagnostics10020110. URL: https://www.mdpi.com/2075-4418/10/2/110.

[48] A. Gafita, M. Bieth, M. Krönke, G. Tetteh, F. Navarro, H. Wang, E. Günther, B. Menze, W. A. Weber, and M. Eiber. "qPSMA: Semiautomatic Software for Whole-Body Tumor Burden Assessment in Prostate Cancer Using 68Ga-PSMA11 PET/CT". In: *Journal of nuclear medicine : official publication, Society of Nuclear Medicine* 60.9 (2019), 1277–1283. DOI: 10.2967/jnumed.118.224055. URL: https://doi.org/10.2967/jnumed.118.224055.

[49] M. Gao, S. Huang, X. Pan, X. Liao, R. Yang, and J. Liu. "Machine Learning-Based Radiomics Predicting Tumor Grades and Expression of Multiple Pathologic Biomarkers in Gliomas". In: *Front. Oncol.* 10.1676 (2020). DOI: 10.3389/fonc.2020.01676.

[50] S. S. Garapati, L. Hadjiiski, K. H. Cha, H. P. Chan, E. M. Caoili, R. H. Cohan, A. Weizer, A. Alva, C. Paramagul, J. Wei, and C. Zhou. "Urinary bladder cancer staging in CT urography using machine learning". In: *Med. Phys* 44 (2017), pp. 5814–5823. DOI: 10.1002/mp.12510. URL: https://doi.org/10.1002/mp.12510.

[51] L. Ge, Y. Chen, C. Yan, P. Zhao, P. Zhang, A R, and J. Liu. "Study Progress of Radiomics With Machine Learning for Precision Medicine in Bladder Cancer Management". In: *Front. Oncol.* 9.1296 (2019). DOI: 10.3389/fonc.2019.01296.

[52] P. Geurts, D. Ernst, and L Wehenkel. "Extremely randomized trees". In: *Mach Learn* 63 (2006), pp. 3–42. URL: https://link.springer.com/content/pdf/10.1007/s10994-006-6226-1.pdf.

[53] M. K. Goel, P. Khanna, and J. Kishore. "Understanding survival analysis: Kaplan-Meier estimate". In: *International journal of Ayurveda research.* 1.4 (2010), 274–278. URL: https://doi.org/10.4103/0974-7788.76794.

[54] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative Adversarial Nets". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 2014, 2672–2680.

[55] S. Gorgi Zadeh, M. Wintergerst, V. Wiens, S. Thiele, F. Holz, R. Finger, and T. Schultz. "CNNs Enable Accurate and Fast Segmentation of Drusen in Optical Coherence Tomography". In: *Deep Learning in Medical Image Analysis (DLMIA)*. LNCS. Springer, 2017.

[56] J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J. C. Fillon-Robin, S. Pieper, and H. J. W. L. Aerts. "Computational Radiomics System to Decode the Radiographic Phenotype". In: *Cancer Research* 77.21 (2017), e104–e107. DOI: 10.1158/0008-5472.CAN-17-0339. URL: https://doi.org/10.1158/0008-5472.CAN-17-0339.

[57] I. Guyon, A. Elisseeff, and L. P. (Ed.) Kaelbling. "An introduction to variable and feature selection". In: *Journal of Machine Learning Research* 3.7-8 (2003), 1157–1182. DOI: 10.1162/153244303322753616. URL: https://doi.org/10.1162/153244303322753616.

[58] I. Guyon, J. Weston, and S. Barnhill. "Gene selection for cancer classification using support vector machines". In: *Mach. Learn* 46 (2002), 389–422. DOI: 10.1023/A:1012487302797. URL: https://doi.org/10.1023/A:1012487302797.

[59] J. Hadamard. "Sur les Problemes Aux Derivees Partielles et Leur Signification Physique". In: *Princeton university bulletin* (1902), pp. 49–52. URL: https://ci.nii.ac.jp/naid/10030321135/en/.

[60] P. Hambarde, S. Talbar, A. Mahajan, S. Chavan, M. Thakur, and N. Sable. "Prostate lesion segmentation in MR images using radiomics based deeply supervised U-Net". In: *Biocybernetics and Biomedical Engineering* 40.4 (2020), pp. 1421–1435. ISSN: 0208-5216. DOI: https://doi.org/10.1016/j.bbe.2020.07.011. URL: https://www.sciencedirect.com/science/article/pii/S0208521620300929.

[61] R. M. Haralick, K. Shanmugam, and I. Dinstein. "Textural Features for Image Classification". In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-3.6 (1973), pp. 610–621. DOI: 10.1109/TSMC.1973.4309314.

[62] F. E. Harrell. "Cox Proportional Hazards Regression Model". In: *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY: Springer New York, 2001, pp. 465–507. ISBN: 978-1-4757-3462-1. DOI: 10.1007/978-1-4757-3462-1_19. URL: https://doi.org/10.1007/978-1-4757-3462-1_19.

[63] M. Hatt, C. Cheze le Rest, A. Turzo, C. Roux, and D. Visvikis. "A fuzzy locally adaptive Bayesian segmentation approach for volume determination in PET." In: *IEEE Trans Med Imaging.* 28.6 (2009), pp. 881–93. DOI: 10.1109/TMI.2008.2012036.

[64] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. "Support vector machines". In: *IEEE Intelligent Systems and their Applications* 13.4 (1998), pp. 18–28. DOI: 10.1109/5254.708428.

[65] A. O. Hebb and A. V. Poliakov. "Imaging of Deep Brain Stimulation Leads Using Extended Hounsfield Unit CT." In: *Stereotact Funct Neurosurg* 87 (2009), pp. 155–160. DOI: 10.1159/000209296.

[66] C. Huang, R. Mezencev, J. F. McDonald, and F. Vannberg. "Open source machine-learning algorithms for the prediction of optimal cancer drug therapies". In: *PLoS One* 12 (2017). DOI: 10.1371/journal.pone.0186906. URL: https://doi.org/10.1371/journal.pone.0186906.

[67] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. *Densely Connected Convolutional Networks*. 2018. arXiv: 1608.06993 [cs.CV].

[68] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[69] S. Jemaa, J. Fredrickson, R. A. D Carano, T. Nielsen, A. de Crespigny, and T. Bengtsson. "Tumor Segmentation and Feature Extraction from Whole-Body FDG-PET/CT Using Cascaded 2D and 3D Convolutional Neural Networks." In: *J Digit Imaging* 33 (2020), 888–894. DOI: 10.1007/s10278-020-00341-1. URL: https://doi.org/10.1007/s10278-020-00341-1.

[70] W. Jentzen, L. Freudenberg, E. G. Eising, M. Heinze, W. Brandau, and A. Bockisch. "Segmentation of PET volumes by iterative image thresholding". In: *J Nucl Med* 48.1 (2007), pp. 108–114. URL: https://jnm.snmjournals.org/content/48/1/108.

[71] S. Jin, D. Li, H. Wang, and Y. Yin. "Registration of PET and CT images based on multiresolution gradient of mutual information demons algorithm for positioning esophageal cancer patients". In: *Journal of Applied Clinical Medical Physics* 14.1 (2013), pp. 50–61. DOI: https://doi.org/10.1120/jacmp.v14i1.3931. eprint: https://aapm.onlinelibrary.wiley.com/doi/pdf/10.1120/jacmp.v14i1.3931. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1120/jacmp.v14i1.3931.

[72] X. Jin and J. Han. "K-Means Clustering". In: *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G. I. Webb. Boston, MA: Springer US, 2010, pp. 563–564. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_425. URL: https://doi.org/10.1007/978-0-387-30164-8_425.

[73] S. Jun, J. G. Park, and Y. Seo. "Accurate FDG PET tumor segmentation using the peritumoral halo layer method: a study in patients with esophageal squamous cell carcinoma". In: *Cancer Imaging* 18.35 (2018). DOI: 10.1186/s40644-018-0169-1. URL: https://doi.org/10.1186/s40644-018-0169-1.

[74] "Kaplan-Meier Estimator: SciKitSurvival Official Website". accessed on 31 August 2021. URL: https://scikit-survival.readthedocs.io/en/latest/api/generated/sksurv.nonparametric.kaplan_meier_estimator.html.

[75] A. R. Khan, L. Wang, and M. F. Beg. "FreeSurfer-initiated fully-automated subcortical brain segmentation in MRI using Large Deformation Diffeomorphic Metric Mapping". In: *NeuroImage* 41.3 (2008), pp. 735–746. ISSN: 1053-8119. DOI: 10.1016/j.neuroimage.2008.03.024. URL: https://www.sciencedirect.com/science/article/pii/S1053811908002528.

[76] Z. Khurshid, H. Ahmadzadehfar, F. C. Gaertner, L. Papp, N. Zsóter, M. Essler, and R. A. Bundschuh. "Role of textural heterogeneity parameters in patient selection for 177Lu-PSMA therapy via response prediction". In: *Oncotarget* 9.70 (2018), 33312–33321. DOI: 10.18632/oncotarget.26051. URL: https://doi.org/10.18632/oncotarget.26051.

[77] T. K. Kim. "T test as a parametric statistic". In: *Korean journal of anesthesiology* 68.6 (2015), 540–546. URL: https://doi.org/10.4097/kjae.2015.68.6.540.

[78] P. E. Kinahan and J. W. Fletcher. "Positron emission tomography-computed tomography standardized uptake values in clinical practice and assessing response to therapy". In: *Seminars in ultrasound, CT, and MR* 31.6 (2010), 496–505. DOI: 10.1053/j.sult.2010.10.001. URL: https://doi.org/10.1053/j.sult.2010.10.001.

[79] A. Klein, J. Warszawski, J. Hillengaß, and K. H. Maier-Hein. "Automatic bone segmentation in whole-body CT images". In: *Int J CARS* 14 (2019), 21–29. DOI: 10.1007/s11548-018-1883-7. URL: https://doi.org/10.1007/s11548-018-1883-7.

[80] J. R. Koza, F. H. Bennett, D. Andre, and M. A. Keane. "Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming". In: *Artificial Intelligence in Design*. Dordrecht: Springer, 1996, 151–170. DOI: 10.1007/978-94-009-0279-4_9.

[81] S. L. Kukreja, J. Löfberg, and M. J. Brenner. "A least absolute shrinkage and selection operator (lasso) for nonlinear system identification". In: *IFAC Proc.* 39 (2006), pp. 814–819. DOI: 10.3182/20060329-3-AU-2901.00128.

[82] F. Lamare, T. Cresson, J. Savean, C. Cheze-Le Rest, A. Turzo, Y. Bizais, A. J. Reader, and D. Visvikis. "Affine transformation of list mode data for respiratory motion correction in PET". In: *IEEE Symposium Conference Record Nuclear Science 2004.* Vol. 5. 2004, 3151–3155 Vol. 5. DOI: 10.1109/NSSMIC.2004.1466349.

[83] J. Lao, Y. Chen, Z. C. Li, Q. Li, J. Zhang, J. Liu, and G. Zhai. "A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme". In: *Scientific reports* 7.1 (2017), p. 10353. DOI: 10.1038/s41598-017-10649-8. URL: https://doi.org/10.1038/s41598-017-10649-8.

[84] C. Lapa, R. A. Werner, J. S. Schmid, L. Papp, N. Zsótér, J. Biko, C. Reiners, K. Herrmann, A. K. Buck, and R. A. Bundschuh. "Prognostic value of positron emission tomography-assessed tumor heterogeneity in patients with thyroid cancer undergoing treatment with radiopeptide therapy". In: *Nuclear Medicine and Biology* 42.4 (2015), pp. 349–354. ISSN: 0969-8051. DOI: https://doi.org/10.1016/j.nucmedbio.2014.12.006. URL: https://www.sciencedirect.com/science/article/pii/S096980511400571X.

[85] R. T. Larue, G. Defraene, D. De Ruysscher, P. Lambin, and W. van Elmpt. "Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures". In: *The British journal of radiology* 90.1070 (2017). DOI: 10.1259/bjr.20160665. URL: https://doi.org/10.1259/bjr.20160665.

[86] I. Lavdas, B. Glocker, K. Kamnitsas, D. Rueckert, H. Mair, A. Sandhu, S. A. Taylor, E. O. Aboagye, and Rockall A. G. "Fully automatic multiorgan segmentation in normal whole body magnetic resonance imaging (MRI), using classification forests (CFs), convolutional neural networks (CNNs), and a multi-atlas (MA) approach." In: *Med Phys. 2017* 44.10 (2017), pp. 5210–5220. DOI: 10.1002/mp.12492.

[87] Y. LeCun, Y. Bengio, and G. Hinton. "Deep Learning". In: *Nature.* 521 (2015), 436–444. DOI: 10.1038/nature14539.

[88] D. k. Lee, U. Yoon, K. Kwak, and J. M. Lee. "Automated Segmentation of Cerebellum Using Brain Mask and Partial Volume Estimation Map". In: *Computational and Mathematical Methods in Medicine* (2015). DOI: 10.1155/2015/167489. URL: https://doi.org/10.1155/2015/167489.

[89] Y. Lin and Y. Jeon. "Random Forests and Adaptive Nearest Neighbors". In: *Journal of the American Statistical Association* 101.474 (2006), pp. 578–590. DOI: 10.1198/016214505000001230. eprint: https://doi.org/10.1198/016214505000001230. URL: https://doi.org/10.1198/016214505000001230.

[90] L. Liu, j. Cheng, Q. Quan, F. X. Wu, Y. P. Wang, and J. Wang. "A survey on U-shaped networks in medical image segmentations". In: *Neurocomputing* 409 (2020), pp. 244–258. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2020.05.070. URL: https://www.sciencedirect.com/science/article/pii/S0925231220309218.

[91] S. Liu, H. Zheng, Y. Feng, and W. Li. "Prostate cancer diagnosis using deep learning with 3D multiparametric MRI". In: *Medical Imaging 2017: Computer-Aided Diagnosis*. Ed. by Samuel G. Armato III and Nicholas A. Petrick. Vol. 10134. International Society for Optics and Photonics. SPIE, 2017, pp. 581 –584. DOI: 10.1117/12.2277121. URL: https://doi.org/10.1117/12.2277121.

[92] X. Liu, Y. Li, Z. Qian, Z. Sun, K. Xu, K. Wang, S. Liu, X. Fan, S. Li, Z. Zhang, T. Jiang, and Y. Wang. "A radiomic signature as a non-invasive predictor of progression-free survival in patients with lower-grade gliomas." In: *Neuroimage Clin.* 20 (2018), pp. 1070–1077. DOI: 10.1016/j.nicl.2018.10.014.

[93] M. Luessi, M. Eichmann, G. M. Schuster, and A. K. Katsaggelos. "Framework for efficient optimal multilevel image thresholding". In: *Journal of Electronic Imaging* 18.1 (2009), pp. 1 –10. DOI: 10.1117/1.3073891. URL: https://doi.org/10.1117/1.3073891.

[94] W. Lv, Q. Yuan, Q. Wang, J. Ma, Q. Feng, W. Chen, A. Rahmim, and L. Lu. "Radiomics Analysis of PET and CT Components of PET/CT Imaging Integrated with Clinical Parameters: Application to Prognosis for Nasopharyngeal Carcinoma." In: *Mol Imaging Biol.* 21.5 (2019), pp. 954–964. DOI: 10.1007/s11307-018-01304-3.

[95] D. Mason. "SU-E-T-33: pydicom: an open source DICOM library". In: *Medical Physics* 38.6Part10 (2011), pp. 3493–3493.

[96] "MathWorks by Matlab, Image Processing Toolbox". Official Company Website for the Software, Available online: URL: https://www.mathworks.com/products/image.html.

[97] A. B. Mattiolli, A. Santos, A. Vicente, M. Queiroz, D. Bastos, D. Herchenhorn, M. Srougi, F. A. Peixoto, L. Morikawa, da Silva Jlf, and E. Etchebehere. "Impact of 68GA-PSMA PET / CT on treatment of patients with recurrent / metastatic high risk prostate cancer - a multicenter study." In: *Int Braz J Urol.* 44.5 (2018), pp. 892–899. DOI: 10.1590/S1677-5538.IBJU.2017.0632.

[98] "Mediso InterView Fusion". Official Company Website for the Software. Available online: URL: https://www.mediso.de/Interview-fusion.html.

[99] X. Meng, B. Zhao, R. Xi, B. Guo, B. Huang, S. Li, Z. Wu, and K. Yun. "The radiomic signature derived from pre-treatment PET and CT images: A predictor of overall survival in non-small cell lung cancer". In: *J. Nucl. Med* 60 (2019), p. 1333.

[100] "MinMaxScaler Normalization Method: Scikitlearn Official Website". accessed on 31 August 2021. URL: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html.

[101] T. Mitchell. *Machine Learning*. New York: McGraw Hill., 1997. ISBN: 07-042807-7.

[102] S. Moazemi, A. Erle, M. Essler, T. Schultz, and R. A. Bundschuh. "Analyzing different combinations of radiomics features and clinical data for treatment response prediction based on whole-body PSMA-PET-CT scans: A machine learning based approach". In: *Nuklearmedizin* 59.02 (2020), pp. 170–171. DOI: 10.1055/s-0040-1708365.

[103] S. Moazemi, A. Erle, Z. Khurshid, S. Lütje, M. Muders, M. Essler, T. Schultz, and R. A. Bundschuh. "Decision-support for treatment with 177Lu-PSMA: machine learning predicts response with high accuracy based on PSMA-PET/CT and clinical parameters". In: *Annals of translational medicine* 9.9 (2021), p. 818. DOI: 10.21037/atm-20-6446.

[104] S. Moazemi, A. Erle, S. Lütje, M. Essler, and R. A. Bundschuh. "Analyzing the potential of radiomics features and radiomics signature from pretherapeutic PSMA-PET-CT scans and clinical data for prediction of overall survival when treated with Lu[177]-PSMA". In: *Nuklearmedizin* 60.02 (2021), pp. 181–182. DOI: 10.1055/s-0041-1726854.

[105] S. Moazemi, A. Erle, S. Lütje, F. C. Gaertner, M. Essler, and R. A. Bundschuh. "Estimating the Potential of Radiomics Features and Radiomics Signature from pre-therapeutic PSMA-PET-CT Scans and Clinical Data for Prediction of Overall Survival When Treated with 177Lu-PSMA". In: *Diagnostics (Basel, Switzerland)* 11.2 (2021), p. 186. DOI: 10.3390/diagnostics11020186. URL: https://doi.org/10.3390/diagnostics11020186.

[106] S. Moazemi, M. Essler, T. Schultz, and R. A. Bundschuh. "Predicting Treatment Response in Prostate Cancer Patients Based on Multimodal PET/CT for Clinical Decision Support". In: *Multimodal Learning for Clinical Decision Support*. Ed. by T. Syeda-Mahmood, X. Li, A. Madabhushi, H. Greenspan, Q. Li, R. Leahy, B. Dong, and H. Wang. Cham: Springer International Publishing, 2021, pp. 22–35. ISBN: 978-3-030-89847-2. DOI: 10.1007/978-3-030-89847-2_3.

[107] S. Moazemi, Z. Khurshid, A. Erle, S. Lütje, M. Essler, T. Schultz, and R. A. Bundschuh. "Machine Learning Facilitates Hotspot Classification in PSMA-PET/CT with Nuclear Medicine Specialist Accuracy". In: *Diagnostics (Basel)Aug* 10.9 (2020), p. 622. DOI: 10.3390/diagnostics10090622.

[108] S. Moazemi, Z. Khurshid, M. Essler, T. Schultz, and R. A. Bundschuh. "Automated detection of pathological lesions in PSMA PET/CT scans in prostate cancer patients: analyzing the relative importance of different groups of features". In: *Nuklearmedizin* 58.02 (2019), p. 107. DOI: 10.1055/s-0039-1683476.

[109] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. 2nd ed. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press, 2018. 504 pp. ISBN: 978-0-262-03940-6.

[110] S. H. Moon, S. H. Hyun, and Choi J. Y. "Prognostic significance of volume-based PET parameters in cancer patients." In: *Korean J Radiol.* 14.1 (2012), pp. 1–12. DOI: 10.3348/kjr.2013.14.1.1.

[111] C. Moore and D. Bell. "Dice similarity coefficient". Reference article, . Available online: URL: https://radiopaedia.org/articles/75056.

[112] A. Nappi, R. Gallicchio, V. Simeon, A. Nardelli, A. Pelagalli, A. Zupa, G. Vita, A. Venetucci, M. Di Cosola, F. Barbato, and G. Storto. "[F-18] FDG-PET/CT parameters as predictors of outcome in inoperable NSCLC patients." In: *Radiol Oncol.* 49.4 (2015), pp. 320–6. DOI: 10.1515/raon-2015-0043.

[113] A. Narayanan, A. Cai, Y. Xi, N. M. Maalouf, C. Rubin, and A. Chhabra. "CT bone density analysis of low-impact proximal femur fractures using Hounsfield units." In: *Clin Imaging.* 57 (2019), pp. 15–20. DOI: 10.1016/j.clinimag.2019.04.009.

[114] S. A. Nehmeh, H. El-Zeftawy, C. Greco, J. Schwartz, Y. E. Erdi, A. Kirov, C. R. Schmidtlein, A. B. Gyau, S. M. Larson, and Humm JL. "An iterative technique to segment PET lesions using a Monte Carlo based mathematical model." In: *Med Phys.* 36.10 (2009), pp. 4803–9. DOI: 10.1118/1.3222732.

[115] D. Nie, J. Lu, H. Zhang, E. Adeli, J. Wang, Z. Yu, L. Liu, Q. Wang, J. Wu, and D. Shen. "Multi-Channel 3D Deep Feature Learning for Survival Time Prediction of Brain Tumor Patients Using Multi-Modal Neuroimages." In: *Sci Rep* 9 (2019), p. 1103. DOI: 10.1038/s41598-018-37387-9. URL: https://doi.org/10.1038/s41598-018-37387-9.

[116] *NIfTI: Neuroimaging Informatics Technology Initiative*. URL: https://nifti.nimh.nih.gov/.

[117] L. O'Donnell and T. Schultz. "Statistical and Machine Learning Methods for Neuroimaging: Examples, Challenges, and Extensions to Diffusion Imaging Data". In: *Visualization and Processing of Higher Order Descriptors for Multi-Valued Data*. Springer, 2015, pp. 299–319.

[118] A. M. J. Paans. "Positron emission tomography". In: (2006). DOI: 10.5170/CERN-2006-012.363. URL: http://cds.cern.ch/record/1005065.

[119] A. M. J. Paans. "Thermal insulation — Heat transfer by radiation — Physical quantities and definitions". In: *ISO 9288:1989* (1989). URL: https://www.iso.org/standard/16943.html.

[120] K. Pak, B. S. Kim, K. Kim, I. J. Kim, S. Jun, Y. J. Jeong, H. K. Shim, S. D. Kim, and Cho K. S. "Prognostic significance of standardized uptake value on F18-FDG PET/CT in patients with extranodal nasal type NK/T cell lymphoma: A multicenter, retrospective analysis." In: *Am J Otolaryngol.* 39.1 (2017), pp. 1–5. DOI: 10.1016/j.amjoto.2017.10.009.

[121] C. Parisot. "The Dicom standard". In: *Int J Cardiac Imag* 11 (1995), pp. 171–177. DOI: 10.1007/BF01143137. URL: https://doi.org/10.1007/BF01143137.

[122] J. E. Park, S. Y. Park, H. J. Kim, and H. S. Kim. "Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives". In: *Korean journal of radiology* 20.7 (2019), 1124–1137. DOI: 10.3348/kjr.2018.0070. URL: https://doi.org/10.3348/kjr.2018.0070.

[123] J. A. Parker, R. V. Kenyon, and D. E. Troxel. "Comparison of Interpolating Methods for Image Resampling". In: *IEEE Transactions on Medical Imaging* 2.1 (1983), pp. 31–39. DOI: 10.1109/TMI.1983.4307610.

[124] R. Paul, S. H. Hawkins, M. B. Schabath, R. J. Gillies, L. O. Hall, and D. B. Goldgof. "Predicting malignant nodules by fusing deep features with classical radiomics features." In: *J Med Imaging (Bellingham)* 5.1 (2018). DOI: 10.1117/1.JMI.5.1.011021..

[125] L. J. Petersen and H. D. Zacho. "Psma Pet for primary lymph node staging of intermediate and high-risk prostate cancer: an expedited systematic review." In: *Cancer Imaging.* 20.1 (2020). DOI: 10.1186/s40644-020-0290-9.

[126] N. A. Pham, A. Morrison, J. Schwock, S. Aviel-Ronen, V. Iakovlev, M. S. Tsao, J. Ho, and D. W. Hedley. "Quantitative image analysis of immunohistochemical stains using a CMYK color model". In: *Diagnostic pathology* 2.8 (2007). DOI: 10.1186/1746-1596-2-8. URL: https://doi.org/10.1186/1746-1596-2-8.

[127] M. E. Phelps, E. J. Hoffman, N. A. Mullani, and M. M. Ter-Pogossian. ""Application of annihilation coincidence detection to transaxial reconstruction tomography"". In: *Journal of Nuclear Medicine* 16.3 (1975).

[128] A. Pépin, J. Daouk, P. Bailly, S. Hapdey, and Meyer M. E. "Management of respiratory motion in PET/computed tomography: the state of the art." In: *Nucl Med Commun.* 35.2 (2014), pp. 113–22. DOI: 10.1097/MNM.0000000000000048.

[129] J. Quinlan. "Induction of Decision Trees". In: *Mach Learn* 1 (1986), pp. 81–106. URL: https://doi.org/10.1023/A:1022643204877.

[130] S. Reuzé, 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." In: *Int J Radiat Oncol Biol Phys.* 102.4 (2018), pp. 1117–1142. DOI: 10.1016/j.ijrobp.2018.05.022.

[131] C. Richmond. "Obituary – Sir Godfrey Hounsfield". In: *BMJ* 329.687 (2004). DOI: 10.1136/bmj.329.7467.687.

[132] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science*. Ed. by N. Navab, J. Hornegger, W. Wells, and A. Frangi. Springer, Cham., 2015. DOI: 10.1007/978-3-319-24574-4_28. URL: https://doi.org/10.1007/978-3-319-24574-4_28.

[133] P. K. Saha and J. K. Udupa. "Optimum image thresholding via class uncertainty and region homogeneity". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.7 (2001), pp. 689–706. DOI: 10.1109/34.935844.

[134] J. Sanyal, I. Banerjee, L. Hahn, and D. Rubin. "An Automated Two-step Pipeline for Aggressive Prostate Lesion Detection from Multi-parametric MR Sequence". In: *AMIA Joint Summits on Translational Science proceedings* (2020), 552–560.

[135] Y. Satoh, A. Nambu, T. Ichikawa, and H. Onishi. "Whole-body total lesion glycolysis measured on fluorodeoxyglucose positron emission tomography/computed tomography as a prognostic variable in metastatic breast cancer". In: *BMC Cancer* 14 (2014). DOI: 10.1186/1471-2407-14-525.

[136] M. J. Scheyerer, B. Ullrich, G. Osterhoff, U. A. Spiegl, and K. J. Schnake. "Arbeitsgruppe Osteoporotische Frakturen der Sektion Wirbelsäule der Deutschen Gesellschaft für Orthopädie und Unfallchirurgie. „Hounsfield units" als Maß für die Knochendichte – Anwendungsmöglichkeiten in der Wirbelsäulenchirurgie." In: *Unfallchirurg.* 122.8 (2019). [Hounsfield units as a measure of bone density-applications in spine surgery], pp. 654–661. DOI: 10.1007/s00113-019-0658-0.

[137] E. Scornet. *Random forests and kernel methods*. 2015. arXiv: 1502.03836 [math.ST].

[138] K. Selvaganesan, E. Whitehead, P. M. DeAlwis, M. K. Schindler, S. Inati, Z. S. Saad, J. E. Ohayon, I. C. M. Cortese, B. Smith, S. Jacobson, A. Nath, D. S. Reich, S. Inati, and G. Nair. "Robust, atlas-free, automatic segmentation of brain MRI in health and disease". In: *Heliyon* 5.2 (2019), e01226. ISSN: 2405-8440. DOI: https://doi.org/10.1016/j.heliyon.2019.e01226. URL: https://www.sciencedirect.com/science/article/pii/S2405844018354860.

[139] Z. Y. Shan, S. J. Mateja, W. E. Reddick, J. O. Glass, and B. L. Shulkin. "Retrospective Evaluation of PET-MRI Registration Algorithms". In: *J Digit Imaging* 24 (2011), pp. 485–493. DOI: 10.1007/s10278-010-9300-y. URL: https://doi.org/10.1007/s10278-010-9300-y.

[140] D. Shen, G. Wu, and H. I. Suk. "Deep Learning in Medical Image Analysis". In: *Annual review of biomedical engineering* 19 (), 221–248. DOI: 10.1146/annurev-bioeng-071516-044442. URL: https://doi.org/10.1146/annurev-bioeng-071516-044442.

[141] C. Shorten and T. M. Khoshgoftaar. "A survey on Image Data Augmentation for Deep Learning". In: *J Big Data* 60 (2019). DOI: 10.1186/s40537-019-0197-0. URL: https://doi.org/10.1186/s40537-019-0197-0.

[142] M. Sollini, F. Bartoli, A. Marciano, R. Zanca, R. H. J. A. Slart, and P. A. Erba. "Artificial intelligence and hybrid imaging: the best match for personalized medicine in oncology". In: *European J Hybrid Imaging* 4.24 (2020). DOI: 10.1186/s41824-020-00094-8. URL: https://doi.org/10.1186/s41824-020-00094-8.

[143] N. H. Spencer. "Essentials of Multivariate Data Analysis". In: CRC Press, 2013, p. 95. ISBN: 978-1-4665-8479-2. URL: https://doi.org/10.1201/b16344.

[144] I. G. Steffen, F. Hofheinz, J. M. Rogasch, C. Furth, H. Amthauer, and J. Ruf. "Influence of rigid coregistration of PET and CT data on metabolic volumetry: a user's perspective." In: *EJNMMI Res* 3.85 (2013). DOI: 10.1186/2191-219X-3-85. URL: https://doi.org/10.1186/2191-219X-3-85.

[145] M. Summerfield. "Rapid GUI Programming with Python and Qt : the Definitive Guide to PyQt Programming." Upper Saddle River, NJ :Prentice Hall, 2008.

[146] L. C. Sun and Coy D. H. "Somatostatin receptor-targeted anti-cancer therapy". In: *Curr Drug Deliv* 8.1 (2011), pp. 2–10. DOI: 10.2174/156720111793663633.

[147] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and Bray F. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries." In: *CA Cancer J Clin* 71 (2021), pp. 209–249. DOI: 10.3322/caac.21660.. URL: https://gco.iarc.fr/today.

[148] P. M. Szczypinski, M. Strzelecki, and A. Materka. "Mazda - a software for texture analysis". In: *2007 International Symposium on Information Technology Convergence (ISITC 2007)* (2007), pp. 245–249. DOI: 10.1109/ISITC.2007.15.

[149] E. Taghizadeh, A. Terrier, F. Becce, A. Farron, and P. Büchler. "Automated CT bone segmentation using statistical shape modelling and local template matching". In: *Computer Methods in Biomechanics and Biomedical Engineering* 22.16 (2019). PMID: 31482715, pp. 1303–1310. DOI: 10.1080/10255842.2019.1661391. eprint: https://doi.org/10.1080/10255842.2019.1661391. URL: https://doi.org/10.1080/10255842.2019.1661391.

[150] M. M. Ter-Pogossian, M. E. Phelps, E. J. Hoffman, and N.A. Mullani. "A positron-emission transaxial tomograph for nuclear imaging (PETT)". In: *Radiology* 114.1 (1975), pp. 89–98. DOI: 10.1148/114.1.89.

[151] R. Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1 (1996), 267–288. URL: http://www.jstor.org/stable/2346178.

[152] J. Toivonen, I. M. Perez, P. Movahedi, H. Merisaari, M. Pesola, P. Taimen, P. J. Boström, J. Pohjankukka, A. Kiviniemi, T. Pahikkala, H. J. Aronen, and I. Jambor. "Radiomics and machine learning of multisequence multiparametric prostate MRI: Towards improved non-invasive prostate cancer characterization." In: *PLoS One.* 14.7 (2019). DOI: 10.1371/journal.pone.0217702.

[153] S. Uddin, A. Khan, M. Hossain, M. E. Hossain, and M. A. Moni. "Comparing different supervised machine learning algorithms for disease prediction". In: *BMC Medical Informatics and Decision Making* 19.1 (2019), p. 281. DOI: 10.1186/s12911-019-1004-8.

[154] M. Vallières, E. Kay-Rivest, L. J. Perrin, X. Liem, C. Furstoss, H. Aerts, N. Khaouam, P. F. Nguyen-Tan, C. S. Wang, K. Sultanem, J. Seuntjens, and I. El Naqa. "Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer". In: *Scientific reports* 7.1 (2017), p. 10117. DOI: 10.1038/s41598-017-10371-5. URL: https://doi.org/10.1038/s41598-017-10371-5.

[155] 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". In: *Sci Rep* 9 (2019), p. 1570. DOI: 10.1038/s41598-018-38381-x. URL: https://doi.org/10.1038/s41598-018-38381-x.

[156] S. Vasilache and K. Najarian. "Automated bone segmentation from Pelvic CT images". In: *2008 IEEE International Conference on Bioinformatics and Biomeidcine Workshops*. 2008, pp. 41–47. DOI: 10.1109/BIBMW.2008.4686207.

[157] J. P. Vert, K. Tsuda, and B. Schölkopf. "1 A primer on kernel methods". In: 2004.

[158] A. Vial, D. Stirling, M. Field, M. Ros, C. Ritz, M. Carolan, L. Holloway, and A. A. Miller. "The role of deep learning and radiomic feature extraction in cancer-specific predictive modelling: a review". In: *Translational Cancer Research* 7.3 (2018). URL: https://tcr.amegroups.com/article/view/21823.

[159] SciKitLearn Official Website. accessed on 31 August 2021. URL: http://scikit-learn.org/stable.

[160] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. "Chapter 10 - Deep learning". In: *Data Mining (Fourth Edition)*. Ed. by I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. Fourth Edition. Morgan Kaufmann, 2017, pp. 417–466. ISBN: 978-0-12-804291-5. DOI: https://doi.org/10.1016/B978-0-12-804291-5.00010-6. URL: https://www.sciencedirect.com/science/article/pii/B9780128042915000106.

[161] S. Wu, Y. Jiao, Y. Zhang, X. Ren, P. Li, Q. Yu, Q. Zhang, Q. Wang, and S. Fu. "Imaging-Based Individualized Response Prediction Of Carbon Ion Radiotherapy For Prostate Cancer Patients". In: *Cancer management and research* 11 (2019), 9121–9131. DOI: 10.2147/CMAR.S214020. URL: https://doi.org/10.2147/CMAR.S214020.

[162] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi. "Convolutional neural networks: an overview and application in radiology". In: *Insights Imaging* 9 (2018), 611–629. DOI: 10.1007/s13244-018-0639-9.

[163] O. Yardimci and B. Ç. Ayyıldız. "Comparison of SVM and CNN classification methods for infrared target recognition". In: *Automatic Target Recognition XXVIII*. Ed. by Firooz A. Sadjadi and Abhijit Mahalanobis. Vol. 10648. International Society for Optics and Photonics. SPIE, 2018, pp. 14 –20. URL: https://doi.org/10.1117/12.2303504.

[164] B. Yaremko, T. Riauka, D. Robinson, B. Murray, A. Alexander, A. McEwan, and W. Roa. "Thresholding in PET images of static and moving targets." In: *Phys Med Biol.* 50.24 (2005), pp. 5969–82. DOI: 10.1088/0031-9155/50/24/014.

[165] C. Yoo, D. Han, J. Im, and B. Bechtel. "Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 157 (2019), pp. 155–170. ISSN: 0924-2716. DOI: https://doi.org/10.1016/j.isprsjprs.2019.09.009. URL: https://www.sciencedirect.com/science/article/pii/S0924271619302205.

[166] S. Yoo, I. Gujrathi, M. A. Haider, and F. Khalvati. "Prostate Cancer Detection using Deep Convolutional Neural Networks". In: *Sci Rep* 9 (2019), p. 19518. DOI: 10.1038/s41598-019-55972-4. URL: https://doi.org/10.1038/s41598-019-55972-4.

[167] P. P. Ypsilantis, M. Siddique, H. M. Sohn, A. Davies, G. Cook, V. Goh, and G. Montana. "Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks." In: *PLoS One.* 10.9 (2015). DOI: 10.1371/journal.pone.0137036.

[168] H. Zaidi, M. Abdoli, C. L. Fuentes, and El Naqa I. M. "Comparative methods for PET image segmentation in pharyngolaryngeal squamous cell carcinoma." In: *Eur J Nucl Med Mol Imaging.* 39.5 (2012), pp. 881–91. DOI: 10.1007/s00259-011-2053-0.

[169] S. Zhao, Y. Su, J. Duan, Q. Qiu, X. Ge, A. Wang, and Y. Yin. "Radiomics signature extracted from diffusion-weighted magnetic resonance imaging predicts outcomes in osteosarcoma." In: *J Bone Oncol.* 19 (2019). DOI: 10.1016/j.jbo.2019.100263.

[170] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. W. L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G. J. R. Cook, et al. "The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping". In: *Radiology* 295.2 (2020), 328–338. ISSN: 1527-1315. DOI: 10.1148/radiol.2020191145. URL: http://dx.doi.org/10.1148/radiol.2020191145.