FACULDADE DE MEDICINA DA UNIVERSIDADE DO PORTO

Bi-parametric MRI radiomics in the spectrum of prostate cancer treatment: Assessing the trustworthiness of AI models

Ana Carolina Rodrigues

DOUTORAMENTO EM CIÊNCIA DE DADOS DE SAÚDE



Doutoramento em Ciência de dados de saúde

Supervisor: Nickolas Papanikolaou

Second Supervisor: Inês Domingues

Third Supervisor: Pedro Rodrigues

June 9, 2025

Abstract

Prostate cancer (PCa) diagnosis, prognosis, and treatment decisions are complex processes that may benefit from artificial intelligence (AI) and radiomics. This dissertation explores the use of bi-parametric MRI (bpMRI) radiomics to develop robust AI models for PCa management across diverse clinical scenarios. The focus is on improving the generalizability of AI models, addressing methodological challenges, and assessing their ability to deliver actionable insights in clinical practice.

A central challenge in radiomics is variability in radiologist segmentations, which can reduce model robustness. Several strategies to mitigate this issue were evaluated with the ProstateX dataset, demonstrating that heterogeneous data, leveraging radiomic features extracted from multiple radiologists' segmentations, led to the most robust classifiers.

The following work used the ProstateNet dataset created under the scope of the ProCAncer-I project. Given its unprecedented size, this dissertation also developed a fully automated pipeline for generating whole-gland and index lesion segmentation masks for DWI and ADC MRI sequences, based on the previously generated masks on T2W.

This allowed the use of biparametric MRI radiomics to tackle six different clinical use cases (UC) along the prostate cancer disease continuum, namely: disease aggressiveness (UC 2), development of metastasis (UC 3), biochemical recurrence (UCs 5 and 6), quality of life after radical prostatectomy (UC 7b) and early withdrawal from an active surveillance program due to disease progression (UC 8). Furthermore, different techniques were evaluated and compared, namely: whole-gland vs lesion radiomics; radiomic features alone, with the inclusion of clinical information and/or deep features; all scanner vendors vs vendor-specific training; and bpMRI vs individual sequences.

Data availability differed highly for the different use cases. The smaller use cases (namely UC3, 6, 7b and 8) had no more than 250 patients available, and, thus, were used more as proof-of-concept studies. Whereas the largest use cases, namely UC5 and 2, had over 700 and 10000 patients, respectively, including a subset for prospective validation. The fully automated bpMRI radiomics pipeline to predict clinically significant PCa (UC2), trained on a large, multi-institutional dataset, achieved a high ROC AUC of 0.912 in retrospective testing and 0.908 in prospective validation, demonstrating consistent performance across MRI vendors, clinical parameters, and patient subgroups. This highlights the promise of bpMRI radiomics in clinical practice, with whole-gland radiomics showing superior predictive power over lesion-based approaches. In the prediction of biochemical recurrence after radical prostatectomy (UC5), whole-gland radiomics again outperformed index-lesion-based models, providing a more accurate risk assessment for BCR and surpassing the standard CAPRA-S score. The model's robustness across MRI vendors and the inclusion of clinical variables further enhance its clinical relevance.

Further research explored voxel-based radiomics for PCa lesion detection and segmentation. A logistic regression model trained on voxel-wise radiomic features demonstrated effective lesion detection and segmentation, particularly for aggressive PCa subtypes. This voxel-wise approach

offers an alternative to deep learning methods, achieving high specificity and precision while addressing the limitations of deep learning-based lesion detection, such as the need for large datasets, and the tendency for overfitting.

To support personalized treatment decisions, uplift modeling was employed to develop a treatment recommendation system based on radiomic features and clinical characteristics. The model successfully stratified patients into groups likely to benefit from different treatment modalities, such as radical prostatectomy or radiation therapy. The use of whole-gland radiomic signatures in this context emphasizes their utility in guiding treatment decisions.

This work emphasizes the importance of large, heterogeneous datasets for model training and validation, demonstrating that whole-gland radiomics can yield reliable, generalizable results in PCa management, when treated with high quality, large, and heterogeneous datasets. Moreover, by integrating fairness assessments and prospective validation, this dissertation addresses the trustworthiness and clinical applicability of AI models in PCa diagnosis, prognosis, and treatment.

Keywords: Prostate Cancer, Radiomics, Machine Learning, bpMRI

Resumo

O diagnóstico, o prognóstico e as decisões de tratamento do cancro da próstata são processos complexos que podem beneficiar da inteligência artificial (IA) e análise radiómica. Esta dissertação explora o uso de variáveis radiómicas extraidas de exames biparamétricos de ressonância magnética para desenvolver modelos robustos de IA para a gestão do cancro da próstata em diversos cenários clínicos. O foco está em melhorar a generalização dos modelos de IA, abordando desafios metodológicos e avaliando a sua capacidade de fornecer insights acionáveis na prática clínica.

Um desafio central na análise radiómica é a variabilidade nas segmentações feitas pelos radiologistas, o que pode reduzir a robustez do modelo. Várias estratégias para mitigar esse problema foram avaliadas com o conjunto de dados ProstateX, demonstrando que dados heterogéneos, com variáveis radiómicas extraídas de segmentações de múltiplos radiologistas, levaram a classificadores mais robustos.

O trabalho que se seguiu, utilizou o conjunto de dados ProstateNet criado no âmbito do projeto ProCAncer-I. Dado o seu tamanho, esta dissertação desenvolveu também uma pipeline totalmente automatizada para gerar máscaras de segmentação de lesões e glândulas inteiras para sequências de ressonância magnética do tipo DWI e ADC, com base em segmentações previamente geradas para sequências do tipo T2W.

Isto permitiu o uso da análise radiómica biparamétrica de ressonância magnética para abordar seis casos clínicos (UC) ao longo do continuum da doença do cancro da próstata: agressividade da doença (UC 2), desenvolvimento de metástases (UC 3), recorrência bioquímica (UCs 5 e 6), qualidade de vida após prostatectomia radical (UC 7b) e retirada precoce de um programa de vigilância ativa devido à progressão da doença (UC 8). Além disso, diferentes técnicas foram avaliadas e comparadas: radiómica da glândula total vs lesão; apenas variáveis radiómicas, com inclusão de informações clínicas e/ou características de *deep learning*; todos os fornecedores de scanners versus treino específico para cada fornecedor; e bpMRI vs sequências individuais.

A disponibilidade de dados variou muito para os diferentes casos clínicos. Os casos clínicos mais pequenos (nomeadamente UC3, 6, 7b e 8) tinham um máximo de 250 pacientes disponíveis e, portanto, foram utilizados principalmente como estudos de prova de conceito. Enquanto que os maiores casos clínicos, nomeadamente UC5 e 2, tiveram mais de 700 e 10.000 pacientes, respectivamente, incluindo um subconjunto para validação prospectiva. A pipeline de análise radiómica bpMRI para prever a presença de cancro clinicamente significativo (UC2), treinado com um grande conjunto de dados multiinstitucional, alcançou uma ROC AUC de 0.912 em testes retrospectivos e 0.908 em validação prospectiva, demonstrando desempenho consistente entre fornecedores de ressonância magnética, parâmetros clínicos e subgrupos de pacientes. Isso destaca quão promissor é o uso de análise radiómica na prática clínica, com a análise radiómica de toda a glândula mostrando poder preditivo superior em relação às abordagens baseadas em lesões. Na previsão de recorrência bioquímica (BCR) após prostatectomia radical (UC5), a radiómica de toda a glândula novamente superou os modelos baseados em lesões, fornecendo uma avaliação de risco mais precisa para BCR e superando o CAPRA-S, padrão para avaliação do risco de recorrência. A ro-

bustez do modelo entre os fornecedores de ressonância magnética e a inclusão de variáveis clínicas aumentam ainda mais a sua relevância clínica.

O trabalho seguinte explorou a análise radiómica baseada em voxeis para detecção e segmentação de lesões de PCa. Um modelo de regressão logística treinado com variáveis radiómicas baseadas em voxel demonstrou capacidade de detecção e segmentação eficaz de lesões, particularmente para subtipos agressivos de cancro da próstata. Esta abordagem baseada em voxeis oferece uma alternativa aos métodos de *deep learning*, alcançando alta especificidade e precisão, ao mesmo tempo que aborda as limitações da detecção de lesões baseada em *deep learning*, como a necessidade de grandes conjuntos de dados e a tendência para overfitting.

Para apoiar decisões de tratamento personalizadas, técnicas de uplift foram empregues para desenvolver um sistema de recomendação de tratamento baseado em características radiómicas e clínicas. O modelo estratificou com sucesso os pacientes em grupos que provavelmente beneficiariam de diferentes modalidades de tratamento, como prostatectomia radical ou radioterapia. O uso de variáveis radiómicas de toda a glândula neste contexto enfatiza a sua utilidade na orientação de decisões de tratamento.

Este trabalho enfatiza a importância de conjuntos de dados grandes e heterogéneos para o treino e validação de modelos, demonstrando que a análise radiómica de toda a glândula pode produzir resultados confiáveis e generalizáveis no contexto de cancro da próstata, quando tratada com conjuntos de dados grandes, heterogéneos e de alta qualidade. Além disso, ao integrar avaliações em subgrupos de pacientes e validações prospectivas, esta dissertação aumenta a confiabilidade e a aplicabilidade clínica dos modelos de IA no diagnóstico, prognóstico e tratamento do cancro da próstata.

Keywords: Cancro da Próstata, Radiómica, Aprendizagem automática, bpMRI

Acknowledgements

I would like to thank my supervisors Nikos and Inês, the CCIG and nuclear medicine teams, especially my fellow PhDs Nuno, Ana Sofia and Cláudia. My family, friends and my partner who supported me throughout the whole process and my dogs Tommy and Ellie for much needed emotional support.

Ana Carolina Rodrigues

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Objectives and Research Questions . . . 2
- 1.3 Contributions . . . 3
- 1.4 Document Structure . . . 4

2 Theoretical Background **7**
- 2.1 Clinical Background . . . 7
  - 2.1.1 The Prostate Gland: Function and Anatomy . . . 7
  - 2.1.2 Tumour development . . . 8
  - 2.1.3 Incidence and mortality . . . 8
  - 2.1.4 Diagnosis and screening . . . 8
  - 2.1.5 Multiparametric MRI . . . 9
  - 2.1.6 Disease aggressiveness . . . 10
  - 2.1.7 Treatment planning / disease management . . . 11
  - 2.1.8 Biochemical Recurrence . . . 12
  - 2.1.9 Quality of Life . . . 12
- 2.2 Technical Background . . . 13
  - 2.2.1 Machine Learning . . . 13
  - 2.2.2 Medical Imaging . . . 13
  - 2.2.3 Radiomics Analysis . . . 14
  - 2.2.4 Causal inference and treatment effect estimation . . . 18

3 State of the Art **27**
- 3.1 Methodology . . . 27
- 3.2 Results . . . 27
  - 3.2.1 Detection of csPCa . . . 27
  - 3.2.2 Metastatic development . . . 28
  - 3.2.3 Treatment failure - Disease relapse or progression . . . 29
  - 3.2.4 Other applications . . . 30

4 Handcrafted and deep radiomic features towards training robust classifiers **35**
- 4.1 Introduction . . . 35
- 4.2 Methodology . . . 36
  - 4.2.1 Research Questions . . . 36
  - 4.2.2 Data description . . . 37
  - 4.2.3 Methods specific to RQ1 . . . 37
  - 4.2.4 Methods specific to RQ2 . . . 38

4.2.5 Train/Test split . . . 39
4.2.6 Dataset pre-processing . . . 39
4.2.7 Model Development . . . 40
4.2.8 Model Evaluation . . . 40
4.2.9 Statistical comparison of Classifiers . . . 40
4.2.10 SHAP importance of predictive variables . . . 40
4.3 Results . . . 41
4.3.1 Research Question 1 . . . 41
4.3.2 Research Question 2 . . . 43
4.4 Discussion . . . 43

5 The ProstateNet Dataset **47**
5.1 Introduction . . . 47
5.2 Methodology . . . 48
5.2.1 Inclusion criteria . . . 48
5.2.2 Data curation . . . 49
5.2.3 Included Clinical variables . . . 50
5.2.4 Initial analysis . . . 50
5.3 Results . . . 50
5.3.1 Missing data . . . 50
5.3.2 Exploratory Data Analysis . . . 54
5.4 Discussion . . . 59

6 Generation of Segmentation Masks and Quality Assessment **61**
6.1 Introduction . . . 61
6.2 Methodology . . . 61
6.2.1 Segmentation of T2W volumes . . . 61
6.2.2 Segmentation of DWI and ADC volumes . . . 62
6.2.3 Quality assessment . . . 62
6.3 Results . . . 63
6.3.1 Quality of T2W segmentations . . . 63
6.3.2 Quality of DWI and ADC segmentations . . . 63
6.4 Discussion . . . 64

7 Whole-gland radiomics models **65**
7.1 Introduction . . . 65
7.2 Methodology . . . 66
7.2.1 Radiomic features extraction . . . 66
7.2.2 Deep features . . . 67
7.2.3 Clinical features . . . 67
7.2.4 Dataset construction . . . 68
7.2.5 Model Development . . . 69
7.3 Results . . . 71
7.3.1 Use Case 2 - ISUP {0,1} vs {2,3,4,5} . . . 71
7.3.2 Use Case 2 - ISUP {0,1,2} vs {3,4,5} . . . 76
7.3.3 Use Case 3 - Development of Metastasis . . . 81
7.3.4 Use Case 5 - BCR pre-RP . . . 83
7.3.5 Use Case 5 - BCR post-RP . . . 86
7.3.6 Use Case 6 - BCR post-RT . . . 89

7.3.7 Use Case 7b . . . 92
7.3.8 Use Case 8 - Disease proggression . . . 94
7.4 Discussion . . . 97

8 Lesion Radiomics models **99**
8.1 Introduction . . . 99
8.2 Methodology . . . 100
8.3 Results . . . 103
8.3.1 Use Case 2 - ISUP {0,1} vs {2345} . . . 103
8.3.2 Use Case 2 - ISUP {0,1,2} vs {3,4,5} . . . 107
8.3.3 Use Case 3 - Development of metastasis . . . 112
8.3.4 Use Case 5 - BCR pre-RP . . . 115
8.3.5 Use Case 5 - BCR post-RP . . . 118
8.3.6 Use Case 6 - BCR post-RT . . . 121
8.3.7 Use Case 7b . . . 124
8.3.8 Use Case 8 - Disease proggression . . . 126
8.4 Discussion . . . 129

9 Radiomics-based lesion detection and aggressiveness mapping on mpMRI **131**
9.1 Introduction . . . 131
9.2 Methodology . . . 132
9.2.1 Patient Cohort . . . 132
9.2.2 Voxel-based radiomic features extraction . . . 132
9.2.3 Model development . . . 132
9.2.4 Model Validation . . . 134
9.3 Results . . . 134
9.4 Discussion . . . 136

10 Uplift modeling for treatment effect estimation **139**
10.1 Introduction . . . 139
10.2 Methodology . . . 140
10.2.1 Data description . . . 140
10.2.2 Modelling . . . 141
10.2.3 Model analysis . . . 141
10.3 Results . . . 142
10.4 Discussion . . . 143

11 Discussion and Main Conclusions **149**
11.1 Comments on the research questions . . . 149
11.1.1 RQ 1 . . . 149
11.1.2 RQ 2 . . . 149
11.1.3 RQ 3 . . . 150
11.1.4 RQ 4 . . . 150
11.1.5 RQ 5 . . . 150
11.1.6 RQ 6 . . . 151
11.1.7 RQ 7 . . . 151
11.1.8 RQ 8 . . . 152
11.1.9 RQ 9 . . . 152
11.1.10 RQ 10 . . . 152

11.2 Comments on the methodology 152
11.2.1 Effect of segmentation mask on model performance 152
11.2.2 Effect of data type on model performance 153
11.2.3 Effect of MRI vendor on model performance 154
11.2.4 Effect of MRI sequence on model performance 154
11.3 Comments on the data and data collection process 155
11.4 Limitations and Future directions 156
11.5 Conclusions 156

A Appendix for Chapter 4 **159**

B Appendix for Chapter 6 **160**

C Appendix for Chapter 7 **162**

Bibliography **163**

List of Figures

2.1 MRI sequences included in a bi-parametric MRI examination. 10
2.2 Diagram depicting anatomical distances information. 14
2.3 Anatomical planes: a) axial; b) sagital; c) coronal. 14
2.4 Radiomics analysis flowchart. 15
2.5 Causal structure. X is a confounder of W and Y. 19

4.1 An example of segmentation variability between radiologists. 37
4.2 Representation of the deep feature extraction procedure. 39
4.3 Overall model development pipeline followed in this study. 41
4.4 Receiver Operator Characteristics Curve for the resampledRad classifier. 44
4.5 Five features with the highest impact on model output (SHAP analysis). 45

5.1 Missing values for each clinical variable included in each use case. 53
5.2 Distribution of clinical variables included in use case 2. 54
5.3 Distribution of clinical variables included in use case 3. 55
5.4 Distribution of clinical variables included in use case 5. 56
5.5 Distribution of clinical variables included in use case 6. 57
5.6 Distribution of clinical variables included in use case 7b. 58
5.7 Distribution of clinical variables included in use case 8. 59

6.1 Pipeline for the automatic generation of segmentation masks. 62

7.1 Radiomics model development pipeline. 70
7.2 WG Radiomics - UC2 (ISUP 0, 1 vs 2, 3, 4, 5) - ROC-AUC model performance. . 72
7.3 WG Radiomics - UC2 (ISUP 0, 1 vs 2, 3, 4, 5) - ROC curve of selected model. . 73
7.4 WG Radiomics - UC2 (ISUP 0, 1 vs 2, 3, 4, 5) - confusion matrix, learning curve, precision-recall curve and calibration plot for the selected model. 73
7.5 WG Radiomics - UC2 (ISUP 0, 1 vs 2, 3, 4, 5) - SHAP analysis. 74
7.6 WG Radiomics - UC2 (ISUP 0, 1 vs 2, 3, 4, 5) - Categorical SHAP analysis. . . . 74
7.7 WG Radiomics - UC2 (ISUP 0, 1 vs 2, 3, 4, 5) - Fairness analysis. 75
7.8 WG Radiomics - UC2 (ISUP 0, 1, 2 vs 3, 4, 5) - ROC-AUC model performance. . 77
7.9 WG Radiomics - UC2 (ISUP 0, 1, 2 vs 3, 4, 5) - ROC curve of selected model. . 77
7.10 WG Radiomics - UC2 (ISUP 0, 1, 2 vs 3, 4, 5) - confusion matrix, learning curve, precision-recall curve and calibration plot for the selected model. 78
7.11 WG Radiomics - UC2 (ISUP 0, 1, 2 vs 3, 4, 5) - SHAP analysis. 79
7.12 WG Radiomics - UC2 (ISUP 0, 1, 2 vs 3, 4, 5) - Categorical SHAP analysis. . . . 79
7.13 WG Radiomics - UC2 (ISUP 0, 1, 2 vs 3, 4, 5) - Fairness analysis. 80
7.14 WG Radiomics - UC3 - ROC-AUC model performance. 81
7.15 WG Radiomics - UC3 - ROC curve of selected model. 82

7.16 WG Radiomics - UC3 - confusion matrix, learning curve, precision-recall curve and calibration plot for the selected model. 82
7.17 WG Radiomics - UC3 - SHAP analysis. 82
7.18 WG Radiomics - UC5 (pre-surgery) - ROC-AUC model performance. 83
7.19 WG Radiomics - UC5 (pre-surgery) - ROC curve of selected model. 84
7.20 WG Radiomics - UC5 (pre-surgery) - confusion matrix, learning curve, precision-recall curve and calibration plot for the selected model. 84
7.21 WG Radiomics - UC5 (pre-surgery) - SHAP analysis. 85
7.22 WG Radiomics - UC5 (post-surgery) - ROC-AUC model performance. 86
7.23 WG Radiomics - UC5 (post-surgery) - ROC curve of selected model. 87
7.24 WG Radiomics - UC5 (post-surgery) - confusion matrix, learning curve, precision-recall curve and calibration plot for the selected model. 87
7.25 WG Radiomics - UC5 (post-surgery) - SHAP analysis. 88
7.26 WG Radiomics - UC5 (post-surgery) - Categorical SHAP analysis. 88
7.27 WG Radiomics - UC6 - ROC-AUC model performance. 89
7.28 WG Radiomics - UC6 - ROC curve of selected model. 90
7.29 WG Radiomics - UC6 - confusion matrix, learning curve, precision-recall curve and calibration plot for the selected model. 90
7.30 WG Radiomics - UC6 - SHAP analysis of selected model. 91
7.31 WG Radiomics - UC7b - ROC-AUC model performance. 92
7.32 WG Radiomics - UC7b - Confusion matrix of selected model. 93
7.33 WG Radiomics - UC8 - ROC-AUC model performance. 94
7.34 WG Radiomics - UC8 - ROC curve of selected model. 95
7.35 WG Radiomics - UC8 - confusion matrix, learning curve, precision-recall curve and calibration plot for the selected model. 95
7.36 WG Radiomics - UC8 - SHAP analysis. 96

8.1 IL Radiomics - UC2 (ISUP 0, 1 vs 2, 3, 4, 5) - ROC-AUC model performance. . 103
8.2 IL Radiomics - UC2 (ISUP 0, 1 vs 2, 3, 4, 5) - ROC curve of selected model. . . 104
8.3 IL Radiomics - UC2 (ISUP 0, 1 vs 2, 3, 4, 5) - confusion matrix, learning curve, precision-recall curve and calibration plot for the selected model. 104
8.4 IL Radiomics - UC2 (ISUP 0, 1 vs 2, 3, 4, 5) - SHAP analysis. 105
8.5 IL Radiomics - UC2 (ISUP 0, 1 vs 2, 3, 4, 5) - Categorical SHAP analysis. 105
8.6 IL Radiomics - UC2 (ISUP 0, 1 vs 2, 3, 4, 5) - Fairness analysis. 106
8.7 IL Radiomics - UC2 (ISUP 0, 1, 2 vs 3, 4, 5) - ROC-AUC model performance. . 108
8.8 IL Radiomics - UC2 (ISUP 0, 1, 2 vs 3, 4, 5) - ROC curve of selected model. . . 108
8.9 IL Radiomics - UC2 (ISUP 0, 1, 2 vs 3, 4, 5) - confusion matrix, learning curve, precision-recall curve and calibration plot for the selected model. 109
8.10 IL Radiomics - UC2 (ISUP 0, 1, 2 vs 3, 4, 5) - SHAP analysis. 110
8.11 IL Radiomics - UC2 (ISUP 0, 1, 2 vs 3, 4, 5) - Categorical SHAP analysis. 110
8.12 IL Radiomics - UC2 (ISUP 0, 1, 2 vs 3, 4, 5) - Fairness analysis. 111
8.13 IL Radiomics - UC3 - ROC-AUC model performance. 112
8.14 IL Radiomics - UC3 - ROC curve of selected model. 113
8.15 IL Radiomics - UC3 - confusion matrix, learning curve, precision-recall curve and calibration plot for the selected model. 113
8.16 IL Radiomics - UC3 - SHAP analysis. 114
8.17 IL Radiomics - UC5 (pre-surgery) - ROC-AUC model performance. 115
8.18 IL Radiomics - UC5 (pre-surgery) - ROC curve of selected model. 116

8.19 IL Radiomics - UC5 (pre-surgery) - confusion matrix, learning curve, precision-recall curve and calibration plot for the selected model. 116
8.20 IL Radiomics - UC5 (pre-surgery) - SHAP analysis of selected model. 117
8.21 IL Radiomics - UC5 (post-surgery) - ROC-AUC model performance. 118
8.22 IL Radiomics - UC5 (post-surgery) - ROC curve of selected model. 119
8.23 IL Radiomics - UC5 (post-surgery) - confusion matrix, learning curve, precision-recall curve and calibration plot for the selected model. 119
8.24 IL Radiomics - UC5 (post-surgery) - SHAP analysis. 120
8.25 IL Radiomics - UC5 (post-surgery) - Categorical SHAP analysis. 120
8.26 IL Radiomics - UC6 - ROC-AUC model performance. 121
8.27 IL Radiomics - UC6 - ROC curve of selected model. 122
8.28 IL Radiomics - UC6 - confusion matrix, learning curve, precision-recall curve and calibration plot for the selected model. 122
8.29 IL Radiomics - UC6 - SHAP analysis. 123
8.30 IL Radiomics - UC7b - ROC-AUC model performance. 124
8.31 IL Radiomics - UC7b - Confusion matrix of selected model. 125
8.32 IL Radiomics - UC8 - ROC-AUC model performance. 126
8.33 IL Radiomics - UC8 - ROC curve of selected model. 127
8.34 IL Radiomics - UC8 - confusion matrix, learning curve, precision-recall curve and calibration plot for the selected model. 127
8.35 IL Radiomics - UC8 - SHAP analysis. 128

9.1 Diagram depicting the construction of the tabular dataset. 133
9.2 Predictions on six example cases from the held out test set. 135
9.3 Fitted logistic regression coefficients of the pre-fusion model. 136

10.1 DAG depicting the confounders considered and controlled for, in this study. . . . 140
10.2 Area Under the Uplift Curve (AUUC) performance of all uplift models trained. . 142
10.3 Uplift Curve of the model that reached the highest AUUC. 143
10.4 Nomogram for starting an aggressive treatment or active surveillance. 145
10.5 Nomogram for choosing radiation therapy (RT) or radical prostatectomy (RP). . . 147

11.1 Summary of AUC performances for the different clinical use cases. 153
11.2 Highest AUC performance for each data type in each clinical use case. 154
11.3 Highest AUC performance for each scanner manufacturer in each clinical use case. 155
11.4 Highest AUC performance for each MRI sequence in each clinical use case. . . . 155

List of Tables

2.1 Divergence metrics used to split uplift decision trees. 26

3.1 Summary of the literature search. 31

4.1 Classification performance on three different hold-out test sets. 42
4.2 Statistical comparison of classifier performance on the resampledRad test set. . . 43
4.3 Performance of classifiers trained with radiomic, deep and hybrid datasets. 46

5.1 Description of MRI acquisition parameters included in the ProstateNet. 48
5.2 Variables description in the processed dataset. 51
5.3 Clinical variables included for each use case. 52

6.1 LR coefficients regarding the low quality of the segmentation of T2W volumes. . 63
6.2 LR coefficients regarding the low quality of the segmentation of DWI volumes. . 64

7.1 Whole-gland radiomics extraction parameters. 66
7.2 Clinical variables included for each use case. 67
7.3 Label distribution in the train and test sets for each binary classification problem. 68
7.4 Label distribution in the train and test sets for the multiclass classification problem. 68
7.5 WG Radiomics - Train set sizes for UC2. 69
7.6 WG Radiomics - Train set sizes for UC3. 69
7.7 WG Radiomics - Train set sizes for UC5. 69
7.8 WG Radiomics - Train set sizes for UC6. 69
7.9 WG Radiomics - Train set sizes for UC7. 69
7.10 WG Radiomics - Train set sizes for UC8. 69
7.11 WG Radiomics - UC2 (ISUP 0, 1 vs 2, 3, 4, 5) - Multi-metric performance. . . . 73
7.12 WG Radiomics - UC2 (ISUP 0, 1, 2 vs 3, 4, 5) - Multi-metric performance. . . . 78
7.13 WG Radiomics - UC3 - Multi-metric performance of selected model. 82
7.14 WG Radiomics - UC5 (pre-surgery) - Multi-metric performance of selected model. 84
7.15 WG Radiomics - UC5 (post-surgery) - Multi-metric performance of selected model. 87
7.16 WG Radiomics - UC6 - Multi-metric performance of selected model. 90
7.17 WG Radiomics - UC7b - Multi-metric performance of selected model. 93
7.18 WG Radiomics - UC8 - Multi-metric performance of selected model. 95

8.1 Label distribution in the train and test sets for each binary classification problem. 100
8.2 Label distribution in the train and test sets for each multiclass classification problem. 100
8.3 Index lesion radiomics extraction parameters. 101
8.4 IL Radiomics - Train set sizes for UC2. 102
8.5 IL Radiomics - Train set sizes for UC3. 102

8.6 IL Radiomics - Train set sizes for UC5. 102
8.7 IL Radiomics - Train set sizes for UC6. 102
8.8 IL Radiomics - Train set sizes for UC7. 102
8.9 IL Radiomics - Train set sizes for UC8. 102
8.10 IL Radiomics - UC2 (ISUP 0, 1 vs 2, 3, 4, 5) - Multi-metric performance. 104
8.11 IL Radiomics - UC2 (ISUP 0, 1, 2 vs 3, 4, 5) - Multi-metric performance. 109
8.12 IL Radiomics - UC3 - Multi-metric performance of selected model. 113
8.13 IL Radiomics - UC5 (pre-surgery) - Multi-metric performance of selected model. 116
8.14 IL Radiomics - UC5 (post-surgery) - Multi-metric performance of selected model. 119
8.15 IL Radiomics - UC6 - Multi-metric performance of selected model. 122
8.16 UC7 - Multi-metric performance of selected model. 125
8.17 IL Radiomics - UC8 - Multi-metric performance of selected model. 127

9.1 Voxel-based radiomics extraction parameters. 132
9.2 Label distribution in the train and test sets. 133
9.3 Segmentation performance of the Pre-fusion model. 135
9.4 Segmentation performance of the Post-fusion model. 136

10.1 Distribution of the outcome variable for each treatment in the train set. 141
10.2 Distribution of the outcome variable for each treatment in the test set. 141
10.3 Hyperparameter space searched during optimization. 142
10.4 Statistical comparison of patients who benefit or not from aggressive treatment. . 144
10.5 Statistical comparison of patients who benefit from radical prostatectomy (RP) or radiation therapy (RT). 146

A.1 List of excluded patients. 159
A.2 Model characteristics. 159

B.1 Co-registration parameter file. 160

C.1 Hyperparameter space used for optimization. 162

Abbreviations

ADC	Apparent diffusion coefficient
ADT	Androgen Deprivation Therapy
AI	Artificial Intelligence
AR	Androgen Receptor
AS	Anterior stroma
ATE	Average Treatment Effect
AUC-ROC	Area Under the Receiver-operating Characteristics curve
BCR	Biochemical Recurrence
bpMRI	Bi-parametric Magnetic Resonance Imaging
CATE	Conditional Average Treatment Effects
CRPC	Castration Resistant Prostate Cancer
CSPC	Castration Sensitive Prostate Cancer
CZ	Central zone
DCE	Dynamic Contrast Enhanced imaging
DL	Deep Learning
DRE	Digital Rectal Examination
DWI	Diffusion-weighted imaging
ERC	Endorectal coil
FOV	Field Of View
GLCM	Gray Level Co-occurrence Matrix
GLDM	Gray Level Dependence Matrix
GLRLM	Gray Level Run Length Matrix
GLSZM	Gray Level Size Zones Matrix
GS	Gleason Score
ICC	Intraclass correlation coefficient
ISUP	International Society of Urologic Pathology
ITE	Individual Treatment Effect
IL	Index Lesion
LGBM	Light Gradient Boosting Machine
LoG	Laplacian of Gaussian
ML	Machine Learning
mpMRI	Multi-parametric Magnetic Resonance Imaging
MRI	Magnetic Resonance Imaging
NGTDM	Neighbouring Gray Tone Difference Matrix
PCa	Prostate Cancer
PI-RADS	Prostate Imaging Reporting And Data System
PPV	Positive Predictive Value
PSA	Prostate-specific antigen
PSAd	Prostate-specific antigen density

PZ	Peripheral zone
QoL	Quality of Life
RCTs	Randomized Controlled Trials
ROC	Receiver-operating Characteristics curve
RP	Radical Prostatectomy
RQ	Research Question
RT	Radiation Therapy
SGD	Stochastic Gradient Descent
T1W	T1-weighted imaging
T2W	T2-weighted imaging
TNR	True Negative Rate
TPR	True Positive Rate
TZ	Transitional zone
UC	Use Case
VOI	Volume of Interest
WG	Whole-gland

Chapter 1

Introduction

This Chapter presents the motivation, objectives, research questions, overall methodology, main contributions and document structure of this doctoral thesis.

1.1 Motivation

In 2020, prostate cancer was the second most incident cancer in men worldwide and ranked 5th in terms of mortality (1). The current diagnostic methods for prostate cancer are not ideal. A statistically significant number of patients are unnecessarily submitted to get a biopsy due to increased false positive diagnosis from the radiologists. When the biopsy is performed, it has been estimated that up to 46% of clinically significant cancers might be missed (2). The consequent upstaging or downstaging of the disease is frequent post radical prostatectomy, where the full gland is removed and examined. Playing a part in this lack of concordance is PCa tissue heterogeneity (3), spatial alignment of lesions (4), among other factors (5). Thus, it is essencial to improve patient selection for biopsy, avoiding unnecessary procedures, with the repercussions associated with it.

Post definitive treatment, biochemical recurrence (BCR) is unfortunately common, happening in up to one third of patients (6). Early detection of BCR would go a long way to prevent disease progression and distant metastases (7), which occur in 8% of cases and have a 5-year survival of only 34.1% (8). Several tools for the prediction of BCR have been proposed in the past, including the Cancer of the Prostate Risk Assessment score (CAPRA/CAPRA-S) (9; 10) or the Decipher test (11). However, non-invasive and more accurate alternatives are necessary. Even in the absence of BCR, prostate cancer patients' lives are not free of strain. Their quality of life is heavily affected by the choice of treatment (12; 13; 14), so an accurate prediction of quality of life decline after specific treatments would be surely helpful in the treatment selection state.

During the full prostate cancer disease continuum, MRI is standard of care. It is recommended both as a first-line diagnostic modality and as a surveillance tool at follow-up (15). Its relevance in patient care makes it the most adequate radiological modality for radiomics analysis. The mathematical nature of feature calculation makes radiomics less subjective than common MRI evaluation by a human reader. MRI-derived radiomic features have been shown to be predictive

of Gleason score, PIRADS, biochemical recurrence, and extracapsular extension, among other events or characteristics of a patient's tumour (Chapter 3).

1.2 Objectives and Research Questions

In this work, several research questions are addressed:

RQ 1 *Which is the best approach to train robust classifiers to inter- and intra-reader variability in segmentation margins?*

RQ 2 *Is the combination of deep and handcrafted radiomic features (I) capable of outperforming radiomic features alone (O) in different machine learning tasks (C) for patients in the prostate cancer disease continuum (P)?*

RQ 3 *Is the combination of clinical information and handcrafted radiomic features (I) capable of outperforming radiomic features alone (O) in different machine learning tasks (C) for patients in the prostate cancer disease continuum (P)?*

RQ 4 *Are radiomic signatures extracted from bi-parametric prostate MRI examinations (I) capable of accurately classifying prostate cancer disease aggressiveness (clinical significance defined by ISUP grade) (O) when compared with biopsy (C) for patients with a suspicion of clinically significant prostate cancer (P)?*

RQ 5 *Are signatures of radiomic and/or clinical features extracted from pre-operative bi-parametric prostate MRI examinations (I) more accurate in predicting metastatic development (within six months follow-up) (O) compared with contemporary predictive models (C) for patients with localized prostate cancer (P)?*

RQ 6 *Are signatures of radiomic and/or clinical features extracted from pre-operative bi-parametric prostate MRI examinations (I) more accurate in predicting biochemical relapse (elevated PSA values after follow-up period) (O) compared with contemporary predictive models (C) for patients that underwent some kind of treatment for prostate cancer (P)?*

RQ 7 *Are signatures of radiomic and/or clinical features extracted from pre-operative bi-parametric prostate MRI examinations (I) more accurate in predicting quality-of-life after prostatectomy (given by the EPIC26 score) (O) compared with contemporary prediction techniques (C) for patients that underwent prostatectomy (P)?*

RQ 8 *Are signatures of radiomic and/or clinical features extracted from pre-operative bi-parametric prostate MRI examinations (I) more accurate in predicting early withdrawal from an active surveillance programme (O) compared with contemporary predictive models (C) for patients enrolled in an active surveillance programme (P)?*

RQ 9 *Are voxel-wise radiomic maps extracted from pre-operative bi-parametric prostate MRI examinations (I) capable of performing clinically significant lesion detection and aggressiveness mapping tasks (O) when compared with contemporary deep learning methods (C) for patients with a suspicion of prostate cancer (P)?*

RQ 10 *Are bi-parametric prostate MRI radiomic variables together with clinical information (I) capable of building an accurate causal model to predict optimal treatment choice (O) when compared with contemporary methods (C) for patients with a suspicion of clinically significant prostate cancer (P)?*

1.3 Contributions

The main contributions of this work are the following:

- Comparison of different methodologies aimed at minimizing the influence of segmentation variability and classifier performance in bpMRI radiomics to discriminate prostate cancer aggressiveness (Chapter 4).
- Development and validation by a clinical expert of a fully automatic pipeline for the generation of whole-gland segmentation masks for DWI and ADC volumes, through co-registration methods (Chapter 6).
- Development and prospective validation of a fully automated tool for tumor aggressiveness assessment from MRI examinations. Followed by deployment of the described tool in an in-house research imaging platform (Chapters 7 and 8).
- Development of an aggressive lesion segmentation pipeline based on voxel-wise radiomic variables. Followed by the deployment of the described tool in an in-house research imaging platform (Chapter 9).
- Training of a causality-based treatment effect estimation model for determining the optimal choice of treatment for an individual patient, given their clinical characteristics and MRI (Chapter 10).
- From the work of this thesis, resulted several publications, submissions and conference acceptances:
  - Ana Rodrigues, Nuno Rodrigues, João Santinha, Maria V Lisitskaya, Aycan Uysal, Celso Matos, Inês Domingues, and Nickolas Papanikolaou. Value of handcrafted and deep radiomic features towards training robust machine learning classifiers for prediction of prostate cancer disease aggressiveness. Scientific Reports, 13(1):6206, 2023. (16)

- Nuno M Rodrigues, José Guilherme de Almeida, Ana Rodrigues, Leonardo Vanneschi, Celso Matos, Maria V Lisitskaya, Aycan Uysal, Sara Silva, and Nickolas Papanikolaou. Deep learning features can improve radiomics-based prostate cancer aggressiveness prediction. JCO Clinical Cancer Informatics, 8:e2300180, 2024. (17)
- Giulio Del Corso, Eva Pachetti, Rossana Buongiorno, Ana Carolina Rodrigues, Danila Germanese, M Antonietta Pascali, José Almeida, Nuno Rodrigues, Manolis Tsiknakis, Nickolas Papanikolaou, et al. Radiomics-based reliable predictions of side effects after radiotherapy for prostate cancer. In IEEE International Symposium on Biomedical Imaging (ISBI), pages 1–4. IEEE, 2024. (18)
- Ana Rodrigues, Nuno Rodrigues, José Almeida, Inês Domingues, and Nickolas Papanikolaou. Predicting prostate cancer biochemical recurrence after radical prostatectomy with biparametric MRI radiomics: A multicentric study. In EuSoMII Annual Meeting 2024. (19)
- Ana Rodrigues, Nuno Rodrigues, José Almeida, Inês Domingues, and Nickolas Papanikolaou. Prostate cancer and radiomics for aggressiveness classification – influence of segmentation mask, MRI sequence, and MRI vendor. In RSNA 110th Scientific Assembly and Annual Meeting 2024. (20)
- Ana Rodrigues, Nuno Rodrigues, José Almeida, Ana Gaivão, Carlos Bilreiro, Inês Santiago, Joana Ip, Sara Belião, Inês Domingues, Manolis Tsiknakis, Kostas Marias, Daniele Regge, and Nickolas Papanikolaou. Improving clinically significant Prostate Cancer Detection with a Multi-Modal Machine Learning Approach: A Large-Scale Multicenter Study. In Radiology: Imaging Cancer [Manuscript submitted for publication]. (21)
- Ana Rodrigues, Nuno Rodrigues, José Almeida, Ana Gaivão, Carlos Bilreiro, Inês Santiago, Joana Ip, Sara Belião, Inês Domingues, Manolis Tsiknakis, Kostas Marias, Daniele Regge, and Nickolas Papanikolaou. Radiomics-based lesion detection and aggressiveness mapping on bpMRI. [Manuscript submitted for publication]. (22)
- Ana Rodrigues, Nuno Rodrigues, José Almeida, Ana Gaivão, Carlos Bilreiro, Inês Santiago, Joana Ip, Sara Belião, Inês Domingues, Manolis Tsiknakis, Kostas Marias, Daniele Regge, and Nickolas Papanikolaou. Uplift modeling for treatment effect estimation in the prostate cancer treatment landscape. In IbPRIA 2025: 12th Iberian Conference on Pattern Recognition and Image Analysis. (23)

1.4 Document Structure

The art of applying data science and artificial intelligence to the healthcare domain is known to be challenging from multiple angles. Whether we look at it from an ethical or technical point of view, it presents obstacles.

This thesis presents work on a highly multidisciplinary subject, which is why I've chosen to give the reader not one, but two theoretical introductions (Chapter 2), one from the clinical perspective and another from the technical perspective. At the former, the reader will find information on the whole disease continuum of prostate cancer. At the latter, information on the technical details will be provided, including statistics, informatics and machine learning, among other subjects.

Following the establishment of a theoretical backbone, the reader can find in Chapter 3 the state of the art for the use of MRI-derived radiomics and similar methodologies along the clinical workflow of a patient with prostate cancer.

Chapter 4 describes the work done to answer the first two research questions. This work was performed with the public ProstateX dataset and resulted in a publication (16).

Chapter 5 describes the ProstateNet dataset's curation and some exploratory data analysis.

Chapter 6 describes the development of a fully automated pipeline for generation of whole-gland segmentation masks for DWI and ADC volumes, as well as a quality assessment of said masks by a clinical expert.

Chapters 7 and 8 describe the model development for each clinical use case this dissertation concerns, using respectively whole-gland- and lesion-based radiomics. From the work of Chapters 6 and 7, resulted a manuscript submission (21) and two conference acceptances (19; 20).

Chapter 9 describes the development of a pipeline for the segmentation of clinically significant lesions, based on voxel-wise radiomic maps. This work resulted in a manuscript submission (22).

Chapter 10 describes the efforts taken to construct a nomogram for the prediction of the optimal treatment choice for a prostate cancer patient. This was based on causal inference and treatment effect estimation with uplift models and resulted in a conference acceptance (23).

The main conclusions of this work are summarized and discussed in Chapter 11, as well as some limitations and future work.

Finally, collaboratory work resulted in two publications (17; 18).

Chapter 2

Theoretical Background

In this chapter, we provide the reader with a theoretical introduction to the clinical and technical aspects of this thesis.

2.1 Clinical Background

In this section, we provide the reader with an introduction to the clinical scenario surrounding the prostate cancer disease continuum.

2.1.1 The Prostate Gland: Function and Anatomy

The prostate gland is an accessory organ in the male reproductive system. Located underneath the bladder, it is traversed by the urethra. This strategic location is essential for proper prostate function. The prostate gland plays several roles in a male's reproductive ability and development, such as closing the urethra during ejaculation and participating in hormone metabolism by converting testosterone into dihydrotestosterone, but its primary function relies upon the secretion of prostatic fluid to the semen (24).

The prostatic fluid makes up between 20% and 30% of semen and it is essential for sperm health. It is composed of citric acid, zinc and several proteolytic enzymes. Among the latter is prostate-specific antigen (PSA), an androgen-regulated glycoprotein serine protease whose purpose is to aid the liquefication of the ejaculate by cleaving semenogelin, the main component of human semen (25).

Distinct anatomic zones can be observed within the adult prostate gland: the peripheral zone (PZ), the central zone (CZ), the transitional zone (TZ) and the anterior fibromuscular stroma (AS) (26). The PZ has the most substantial volume and fosters the most tumours, with almost 70% of prostate tumours developing here. This is followed by the transitional zone, where 25% of prostate tumours arise (27).

2.1.2 Tumour development

The risk of prostate cancer development is known to be higher for genetically predisposed men. This might be in the form of impacted first-degree relatives, in which case the risk may be twice as high (28). The risk is further amplified in the presence of germline mutations, such as those in BRCA1 and BRCA2 (29; 30).

Most genetic modifications associated with the development of prostate cancer are copy number alterations (31) or chromosomal structural rearrangements (32). These are particularly dangerous when occurring in oncogenes or tumour suppressor genes. Since they regulate the cell division cycle, a mutation in these genes might make the cell unresponsive to cell cycle controls or apoptosis, leading to uncontrolled cell proliferation, the primary driver of the disease.

Prostate tumours are thought to originate from the basal (33) or luminal (34; 35) prostate epithelial cells. The accelerated mitosis rates may lead primarily to non-malignant lesions such as prostatic intraepithelial neoplasia (36). These will later be classified as cancerous if the epithelium invades the basement membrane. In a more advanced stage, prostate cancer metastasis might arise. The malignant cells tend to downregulate the expression of proteins associated with their attachment to neighbouring cells, allowing some mobility (37). These will travel to locoregional lymph nodes from where they may enter the bloodstream and bind to a secondary site, where they can proliferate and form a new tumour. Prostate adenocarcinoma is mostly known to metastasise to locoregional lymph nodes and to bone (37; 38).

2.1.3 Incidence and mortality

According to the World Health Organization, in 2020, prostate cancer was one of the leading carcinomas in terms of worldwide incidence, surpassed only by lung cancer, and it ranked fifth regarding mortality (1; 39).

Prostate cancer incidence has been reported to be higher in developed countries (40; 41). Here, the life expectancy is higher and screening protocols have been heavily implemented, increasing the rate of early detection, which is also reflected in the high survival rates of prostate cancer patients: 97% (42).

In addition to demographic differences, incidence disparities are evident in race and ethnicity. Black men are by far the most afflicted, with incidence rates being 70% higher than in Caucasian males (42).

2.1.4 Diagnosis and screening

As stated in section 2.1.2, prostate cancer originates from epithelial cells. These are largely responsible for the production of PSA that will later be part of the ejaculate (section 2.1.1). Thus, the exaggerated proliferation of prostate epithelial tissue tends to result in elevated PSA blood levels (43). For this reason, this biomarker is often used when screening for prostate cancer (44; 45; 46).

Additionally, clinical suspicion of prostate cancer can also arise from an abnormality in digital rectal examination (DRE). This consists of a physical palpation of part of the prostate's peripheral zone, in order to evaluate gland enlargement, texture and firmness (47).

Despite this, none of these approaches is specific to prostate cancer. Abnormalities in DRE or elevated PSA levels can, similarly, arise in benign conditions such as prostatitis (prostate inflammation) (48) or benign prostatic hyperplasia (BPH) (49), resulting in false positives. It has been reported that up to 80% of patients with elevated PSA do not have PCa (50). Furthermore, both PSA levels and DRE might be normal even in the presence of prostate cancer, resulting in false negatives (51; 52). As a result, these screening approaches often result in the over- or underdiagnosis of prostate cancer (53), which come with unfortunate physical and psychological burdens such as anxiety and depression (54), and possible uncontrolled disease progression.

The suspicion of prostate cancer is further confirmed with magnetic resonance imaging examination (MRI). This is a crucial tool for prostate cancer management and the current guidelines recommend its use as a first-line diagnostic modality in case of clinical suspicion (15). Once the most suspicious areas within the prostate are identified, cancer is confirmed through biopsy (55; 56), where an aggressiveness level is assigned according to visual assessment of the histopathology of the tissue.

2.1.5 Multiparametric MRI

Multiparametric MRI (mpMRI) is a combination of multiple image contrasts, from anatomic to functional sequences. Its ability to detect prostate cancer has led to further use for biopsy guidance, replacing the traditional method - transrectal ultrasound guided systematic biopsy, while improving the detection of clinically significant cancers and reducing detection of indolent disease (57; 58).

The anatomic MRI sequences include T1- and T2-weighted imaging. T2W imaging has a high spatial resolution, depicting clearly the anatomic boundaries of prostate zones (59), as described in section 2.1.1. Characteristic of this sequence is the hyperintensity of the peripheral zone, given the high water content in the tissue. For this reason, prostate cancer lesions are easily distinguishable here since they appear, in contrast, as a dark hypointense region. Contrary to T2W imaging, T1W does not allow accurate determination of prostate morphology, however, it is useful in the definition of prostate margins or identification of hemorrhage. Since hemorrhage can appear very similar to cancer on T2W, a T1W exam can be used to rule out this possibility and confirm prostate cancer. Likewise, other benign abnormalities such as prostatitis, fibrosis or scar tissue can also appear hypointense on T2W. For this reason, the combination of multiple imaging sequences in mpMRI is extremely powerful, overcoming these limitations (60).

The functional sequences in mpMRI are diffusion-weighted imaging (DWI), including the calculation of the apparent diffusion coefficient (ADC) map, and dynamic contrast-enhanced maps (DCE). DWI assesses the diffusion of water molecules in the tissue, with areas of restricted diffusion, such as cancerous lesions, appearing hyperintense (61). The sensitivity to diffusion is regulated by the b-value parameter. The higher the b-value, the easier the distinction between

healthy and cancerous tissue (62). The ADC map is calculated on a voxel basis, traditionally with the monoexponential model (63). Here, we plot the DWI signal intensity on a logarithmic y-axis against the b-value on a linear x-axis and the ADC value corresponds to the slope (64). Thus, the ADC map represents, for each voxel, the rate of DWI signal loss as the b-value increases. It is known that a low slope is characteristic of prostate cancer lesions, appearing hypointense in the ADC map (65; 66). Finally, DCE-MRI involves the intravenous injection of a contrast agent and evaluates the velocities of its uptake and washout by the tissue (67). This allows radiologists to evaluate, for instance, the degree of neovascularity.

Recently, bi-parametric MRI (bpMRI) has been suggested as a replacement to mpMRI, with a shorter acquisition time, lower cost and similar performance (68; 69). bpMRI includes only one anatomic sequence, T2W, and one functional sequence, DWI, with the respective ADC map (Figure 2.1).

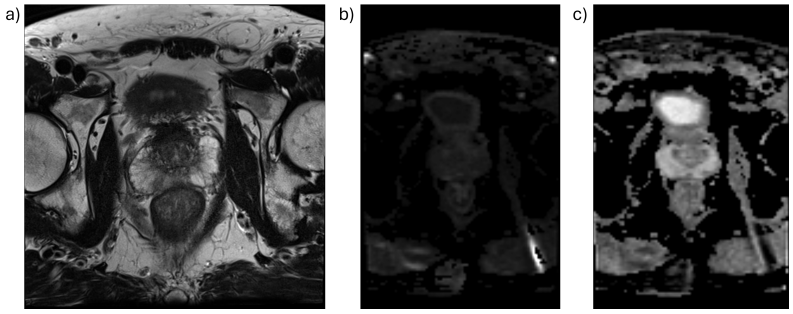


Figure 2.1: MRI sequences included in a bi-parametric MRI examination. a) T2W; b) High b-value DWI; c) ADC.

The qualitative nature of MRI interpretation is prone to some inter-reader variability and lacks objectiveness. In an effort to overcome this, the Prostate Imaging Reporting and Data System (PIRADS 2.1) was introduced in the clinic (70). This is a set of guidelines that aims to objectively give a probability estimate for the presence of clinically significant prostate cancer on MRI (71). However, still, variability remains a problem in the field (72; 73; 74), especially when comparing different healthcare institutions (75), and centres outside specialized clinics (76). Nevertheless, in 2022, Yin Lei et al reported the performance of PIRADS 2.1 for the detection of clinically significant PCa: sensitivity of 0.973, specificity of 0.794, positive predictive value of 0.723 and negative predictive value of 0.982 (77). Though, interestingly, there is some heterogeneity in the reported performance for PI-RADS between radiological and urologic scientific journals, where, in the latter, the performance tends to be lower (76).

2.1.6 Disease aggressiveness

The metric that is clinically used to define prostate cancer disease aggressiveness is the Gleason Score. It ranges from 1 to 5, with a higher value representing a higher degree of aggressiveness. Following biopsy, a patient's two most prominent histological patterns are graded through visual inspection by a pathologist, and the final Gleason score results from their sum. Recently, the

Gleason system has been reorganized into the ISUP grade (78), which more clearly distinguishes gleason scores of 3+4 and 4+3, corresponding to an ISUP grade of 2 and 3, respectively. These were previously encompassed within the Gleason Score of 7, failing to recognize the extensive differences in the clinical significance of the two cases (79; 80).

2.1.7 Treatment planning / disease management

Active surveillance is the option most suited to low-risk prostate cancer patients (Gleason score = 6). Here, patients undergo no treatment, but are monitored for signs of disease progression with scheduled DRE, PSA measurements, MRI examinations and potential biopsies, at which point, a new treatment strategy may be decided if disease progression is found. Active surveillance has several advantages such as preserving sexual, urinary and bowel function while avoiding unnecessary treatment and sustaining quality of life. However, there's a greater likelihood of cancer progression to a more aggressive stage with a need for equally aggressive treatment, failing to benefit from a simpler remedy (81; 82).

Further examinations should be carried out for patients in the intermediate (Gleason score = 7) to high-risk (Gleason score $\geq$ 8) section to identify possible metastasis (83; 15). If the disease is confirmed to be organ-confined, two major treatment options arise: Radical prostatectomy and Radiation therapy.

Radical Prostatectomy (RP) is the surgical removal of the prostate gland, with or without preservation of the surrounding nerve structures. The nerve-sparing option is the least conservative one, ensuring minimal damage to the patient's sexual function (84), but presenting a higher risk of recurrence (85). If lymph node invasion has been previously detected, these can be resected together with the prostate (86). Different surgical strategies can be used which will impact patient outcome, such as open (retropubic or perineal), laparoscopic or robot-assisted laparoscopic, but there is little evidence to favour one strategy over the others (87).

Radiation Therapy (RT) uses high doses of radiation to eliminate prostate cancer cells. There are several techniques used for RT: external beam radiation therapy (EBRT), where the radiation is emitted through the skin and into the prostate, stereotactic body radiation therapy (SBRT), with a more precise and intense dose of radiation, or brachytherapy (BT), where radioactive sources, or seeds, are placed within the prostate (87; 88).

For higher-risk patients or patients with evidence of metastasis, the treatment options include hormone therapy alone or combined with radiotherapy. Hormone therapy targets the most studied oncogene in prostate cancer - the androgen receptor (AR). Prostate epithelial cells express high levels of AR which, when binding of androgen occurs, initiates a transcriptional response that leads to PSA production and increases cell proliferation (89; 90). For this reason, prostate tumour development displays a hormone dependency that can be exploited. Androgen deprivation therapy (ADT) attempts to control prostate cancer cell proliferation by limiting the tumour's androgen supply. The castration levels of androgen lead to a decrease in PSA and tumour size (91). However, ADT is frequently associated with the development of resistance to therapy (92)

through the overexpression of AR, by amplification of the gene and/or alterations to the expression pathway (93), or by AR mutations that reduce its specificity, enabling activation by different ligands (94). Thus, metastatic disease can be divided into castration-sensitive prostate cancer (mCSPC) or castration-resistant prostate cancer (mCRPC), in which case the standard of care would be chemotherapy (95).

2.1.8 Biochemical Recurrence

Biochemical recurrence (BCR) is defined as a rise in PSA to 0.2 ng/ml or greater after radical prostatectomy or a rise of at least 2 ng/ml above the level of PSA after radiation therapy (15). A BCR diagnosis is indicative of disease progression and/or distant metastases (7) and it has been reported that up to 50% of patients experience BCR within the first 10 years after treatment (96; 97). Patients experiencing BCR may be offered local salvage radiotherapy, which has been shown to reduce the risk of progression (98; 99).

2.1.9 Quality of Life

The prostate cancer diagnostic process and subsequent treatments are enormous sources of physical and psychological strain. However, once treated, prostate cancer still leaves its mark in the form of long-term complications. The quality of life of prostate cancer survivors highly depends on the treatment strategy selected.

Radical prostatectomy is mostly associated with problems in erectile (12) and urinary functions (13), while radiotherapy negatively affects urinary and bowel functions (14). Active surveillance presents with the least adverse effects, which is why it is recommended for low-risk patients. However, even patients in an active surveillance programme can experience a decline in sexual and urinary functions due to tumour progression (100).

Regarding more advanced disease states, ADT negatively impacts, sexual, urinary and bowel functions, leading to adverse effects such as decreased libido and/or impotence, loss of muscle mass, fatigue, and bone pain, among others (101; 102). While, for chemotherapy, the list of adverse effects includes, but is not limited to, fever, diarrhoea, nausea, fatigue and sensory polyneuropathy (103).

The physical complications are accompanied by psychological problems such as anxiety and depression in all stages of the disease and across all treatment strategies including active surveillance (104), both in patients and in their partners (105). Toxic masculinity and the societal pressure for men to exhibit the stereotypical “masculine” qualities of physical toughness and emotional control are known to be tremendously harmful to their mental health, especially in the context of cancer where they are experiencing an extreme sense of physical vulnerability. This is aggravated when complications such as erectile dysfunction develop. Experiencing sexual difficulties in a culture where constant pressure is put on a man’s ability to perform sexually can have an enormous impact on a man’s mental and social well-being. As a consequence, long-lasting psychological disorders may develop (106).

The world health organization defines health as not only the absence of disease or infirmity, but as a state of physical, emotional, and social well-being. Thus, these long-term complications, whether physical, psychological or social, should not be disregarded.

2.2 Technical Background

In this section, we provide the reader with an introduction to the technical aspects of this thesis.

2.2.1 Machine Learning

The field of machine learning (ML) can be divided into three main categories: supervised, unsupervised and reinforcement learning. Supervised learning algorithms are trained with labelled data to make predictions for new observations where the outcome label is unknown. On the other hand, unsupervised learning algorithms have no knowledge of the outcome label and attempt to group the observations according to some distance metric. Finally, in reinforcement learning, the system improves its performance according to the feedback it receives from the environment after a certain action.

Within supervised learning, the type of ML algorithm applied is chosen according to the type of outcome variable. For continuous outcome variables, regression algorithms are used, while for categorical outcome variables, the task is of classification. The choice of machine learning algorithm should also take into consideration where they fall on the bias-variance curve.

The discrepancy between a model's ground truth and its prediction for a given instance is known as bias. To facilitate learning, models make assumptions about the data, which might make it more difficult to discern the underlying patterns, resulting in underfitting. The logistic regression is an example of a machine learning algorithm with high bias. The variance indicates the degree to which the target function varies when training data changes. If the data comes from the same distribution, then the algorithm should have minimal variance. A high variance model captures noise in addition to the data's underlying pattern, resulting in overfitting. Tree-based models are examples of high variance algorithms. One cannot avoid the fact that a reduction in bias inevitably results in a rise in variance and vice versa, so the choice of algorithm should be carefully considered for each application.

2.2.2 Medical Imaging

While regular images can be thought of as 2-dimensional matrices of intensity values, medical images will differ in two ways: 1) they can be 2- or 3-dimensional intensity matrices, in which case pixels are renamed as voxels; and 2) they contain anatomical information in the form of the biological location of each voxel. This information results from the combination of three characteristics (Figure 2.2):

- Voxel spacing - x-, y- and z-spacing refer to the inter-voxel distance in each direction.

- Origin - coordinates of the origin voxel, which indicate the anatomical location of voxel (0,0,0). Also known as iso-center.
- Direction matrix - indicates the exact anatomical plane of the image.

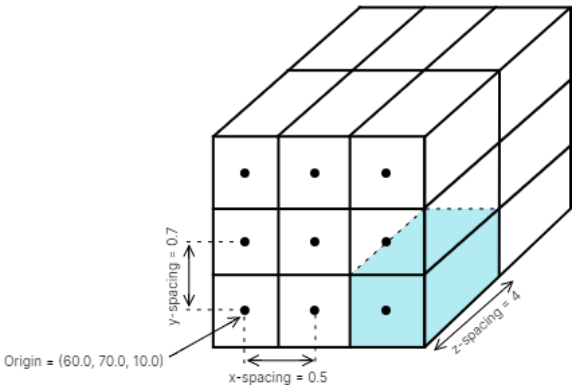


Figure 2.2: Diagram depicting anatomical distances information found in the metadata of medical images.

Anatomical planes (Figure 2.3) refer to the axis chosen for the image, and there are three: the sagital plane, where patients are observed from the side; the coronal plane, where the patient is observed from the front or back; and the axial or transversal plane, where patients are observed from above or below. This dissertation closely focuses on the axial plane as it corresponds to the standard acquisition plane for prostate cancer patients.

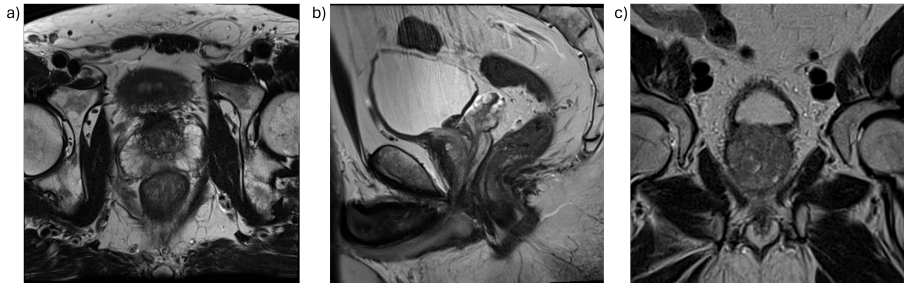


Figure 2.3: Anatomical planes: a) axial; b) sagital; c) coronal.

2.2.3 Radiomics Analysis

The qualitative evaluation of medical images by human readers can now be aided by a more objective quantitative assessment thanks to the field of radiomics (107). Radiomics is the analysis of medical images through the extraction of quantitative information, which can then be used for the creation of clinical decision support systems. While the qualitative analysis reflects the expertise of the radiologist, measures such as tumor size, prostate volumes or radiomic features calculated directly from the images, could be considered almost reader-independent, and consequently reduce the discrepancy between readers with different levels of experience or specialized and non-specialized centres (108). The interest in radiomics has grown over the last decade achieving over

2000 publications in 2023 alone. Figure 2.4 describes the main steps in a radiomics analysis. These will be described in detail in the following sections.

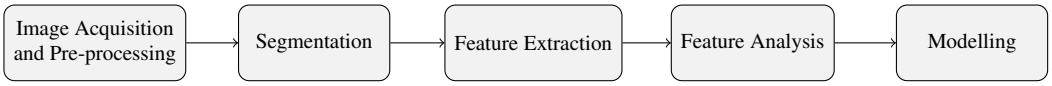


Figure 2.4: Radiomics analysis flowchart.

2.2.3.1 Image acquisition and pre-processing

The first stage of a radiomics analysis is the acquisition of medical images. These are formed by the interaction between the radiation, ultrasound magnetic fields, or radio waves and the biological tissue, reflecting some of its pathophysiological properties. However, clinical images should not be interpreted as the truth, but as a representation of the real object under study as observed by the detection system (109). In fact, the acquisition of biomedical images is conditional on several parameters set according to image modality (x-ray, ultrasound, CT, MRI, etc.), equipment, tissue and aim of the exam. This variability in acquisition protocols is reflected in the images and subsequently in the radiomic features (110; 111; 112; 113; 6), which might give rise to some inaccurate associations with the underlying biological effects. This is why minimizing the influence of acquisition protocols is a key part of the radiomics pipeline. With this goal in mind, some pre-processing techniques have been proposed such as resampling the voxel size, intensity normalization to harmonize the range of gray level values, removing bias or scaling factors, or bias-field correction to adjust low-frequency intensity non-uniformity commonly present in MRI (114).

2.2.3.2 Segmentation

Having acquired the images, the next step in the radiomics pipeline is the delineation of the volume of interest (VOI). Here, a binary mask is generated, where voxels are given a value of 1 if they belong to the region of interest or 0 otherwise. This can be achieved manually, when a radiologist delineates the boundary of the VOI in the image, automatically, by an algorithm trained to segment a certain VOI, or semi-automatically, where the mask is drawn in an automatic fashion and later verified and corrected by a radiologist. This step introduces a source of human bias into the process. Being completely or partially human-dependent, segmentation is a time-consuming and subjective process, that naturally suffers from inter- and intra-reader variability, being dependent on the expert's perception and level of experience (115; 116). This inevitably affects the radiomic features and, subsequently, the final model's external validity. The automated segmentation algorithms overcome this issue to some extent (117), as the technique is human-independent.

2.2.3.3 Feature Extraction

Following segmentation, thousands of pre-defined mathematical transformations are applied to extract the quantitative features which describe VOI shape, tissue texture, and distribution of intensities, among other characteristics with unknown biological correlation. The hypothesis behind radiomics is that these quantitative features are related to underlying pathological or physiological characteristics.

There are four main classes of radiomic features: shape, first-order or histogram-based, second-order or texture, and higher-order features.

Shape-based radiomic features describe the VOI's 2D or 3D size and shape. These are independent from the voxel intensity, relying purely on the margins of the segmentation mask.

First-order or histogram-based radiomic features consider the distribution of intensity values within the VOI. This class of features are highly dependent on the number of bins or bin width selected. If the number of bins is too small (or the bin width too large), the intensity distribution will not be observed and if it is too large (or the bin width too small) the random noise will not allow an accurate representation of the underlying distribution. Optimal binning is thus essential. Loic Duron et al (118) argue why it is imperative to use a fixed bin width and not a fixed number of bins for discretization. They propose that since this method is independent of the VOI's intensity range, it may be less sensitive to inter- and intra-observer segmentation variability. Finally, Tixier F, et al. (119) argue that the ideal number of bins for the calculation of texture features should be between 8 and 128.

Second-order or texture features (120), assess the spatial relationships between voxel intensities, mostly undetectable by the naked eye. To extract this class of features, a matrix is first constructed describing a certain type of spatial relationship, and then some metrics are calculated on it. Different types of spatial relationships can be evaluated, such as the number of occurrences of a certain pair of signal intensities by the Gray Level Co-occurrence Matrix (GLCM), the length of consecutive voxels of the same intensity by the Gray Level Run Length Matrix (GLRLM), the number of connected voxels of the same intensity by the Gray Level Size Zones Matrix (GLSZM), the difference between a voxel's intensity and the average intensity of its neighbours by the Neighbouring Gray Tone Difference Matrix (NGTDM) and the number of connected voxels within a certain distance that are dependent on the central voxel by the Gray Level Dependence Matrix (GLDM).

The final class of radiomic features, higher-order features, can be calculated by applying the same methods described in the previous paragraphs after applying filters or mathematical transformations on the images. Some examples of these filters are Wavelet (121), Laplacian of Gaussian (LoG) (122), Square, SquareRoot, Logarithm, Exponential, Gradient (123), Local Binary Pattern (124), among others.

2.2.3.4 Feature Analysis

Given the unknown biological meaning of the extracted features, it is likely that some are correlated to each other. This issue, commonly known as multicollinearity, is a problem since the lack of feature independence affects coefficient estimates, especially in regression scenarios (125). Even though this is expected not to affect the model's predictive power, it will make it less robust. In this scenario, small changes in one variable will be amplified to all the variables with which it is correlated, having an extreme impact on model coefficients and preventing good external generalizability. Removing these redundant variables is essential.

Because the features are automatically extracted, some might not be applicable in specific use-cases and return with zero or near-zero variance across the dataset. This is somewhat common, and those features can and should be excluded. Still, the number of features can easily reach the thousands, giving rise to the common predicament of the *curse of dimensionality*. This "curse" has two main causes: data sparsity and distance concentration. While the number of predictors increases, so do, exponentially, the number of combinations of predictor values that can be found in the data. When a dataset is unable to cover all these possibilities, we are in the presence of a data sparsity problem, also known as a "blind spot", given that the model will in all likelihood be unable to correctly predict the outcome for feature combinations with which it did not come into contact during the training stage (126). Similarly, distance concentration is the concept that, as dimensionality increases, distances between observations will converge to the same value and become indistinguishable. Since these distances are essential to extract knowledge and patterns from the data, the apparent equidistance will mask meaningful information (127).

This high-dimensional feature space, in addition to being nearly impossible to analyse by the human brain, will lead to the *Curse of Dimensionality*, so, to reduce the dimensionality of the problem, scientists can introduce feature selection techniques. Feature selection is an essential step, not only to minimize the issues that arise with the curse of dimensionality, but also to reduce training time and increase model interpretability. Feature selection techniques can be classified into three types: filter, wrapper, and embedded methods (128). Filter methods select features based on their intrinsic properties in a supervised or unsupervised way, respectively with and without regard to the outcome variable. Examples of unsupervised filter methods are the exclusion of highly correlated predictors or the removal of predictors with zero variance. Supervised filters evaluate how powerful each variable is when predicting the outcome variable. Some examples of supervised filter methods are a univariate analysis with a chi-square test (129) or the calculation of the *information gain* (130) for each feature. Wrapper methods do not evaluate each feature independently, so much as subsets of features. These methods aim to find the feature set that optimizes classifier performance. While they are the most suited to deliver the most useful classifier, they are also extremely computationally expensive when compared to filter methods. Some examples of wrapper methods are forward selection and recursive feature elimination (131), a form of backwards selection. Embedded methods are those that are an intrinsic part of certain classifiers, applied automatically during training. Some examples of classifiers with embedded feature selection are

logistic regressions with LASSO regularization (132) and decision trees (133).

2.2.3.5 Modelling

Finally, machine learning (ML) algorithms can be employed to uncover patterns in the underlying radiomic features. These can be exploited to locate either clusters of similar patients (unsupervised learning) or useful predictors of a certain clinical outcome of interest (supervised learning).

2.2.4 Causal inference and treatment effect estimation

Uplift modeling traditionally refers to the techniques applied by marketing companies to estimate the effects of certain actions (or treatments) on groups of customers. The customers were traditionally divided into four groups: "The Persuadables" (buy if treated, do not buy if not treated), "The Lost Causes" (never buy, independent of treatment), "The Sure Things" (always buy, independent of treatment), and "The Do Not Disturb" or "Sleeping dogs" (buy if not treated, but do not buy if treated). The final purpose of the analysis being to target "The Persuadables" populations and avoid the remaining, especially "The Sleeping dogs".

This type of analysis can be adapted to clinical decision making. For instance, in the case of prostate cancer treatment decisions, correctly identifying "The Lost Causes" (patients that will recur, independent of treatment) and "The Sure Things" (patients with no disease progression, independent of treatment) would prevent the patient from undergoing treatment and suffering through the side-effects of toxicity or low quality-of-life associated. The "Sleeping Dogs", however, (patients that recur, if treated) are somewhat more difficult to translate to the clinical practice, but we could hypothesize that, rarely, the treatment could lead to contamination with tumor cells of areas that were previously free of cancer and, thus, lead to recurrence.

In fact, treatment effect estimation in the form of ATE (average treatment effect) is already commonly used in RCTs (randomized controlled trials) to estimate the overall effect of a drug or procedure on the general population. In contrast, uplift models focus on ITE (individual treatment effect) estimation, allowing for personalized treatment effect estimation, fine-tuned to each patient's characteristics.

2.2.4.1 ITE and the Fundamental Problem of Causal Inference

ITE is defined as the difference in outcomes (Y) between the scenarios where the patient either receives the treatment (Y_1) or not (Y_0):

$$ITE = Y_1 - Y_0 \tag{2.1}$$

This introduces the fundamental problem of causal inference: Having a patient gone through a certain treatment, it is impossible to know what would have happened had the patient not had the treatment. That is where machine learning can help bridge the gap, by estimating the counterfactual scenario.

2.2.4.2 Assumptions

Typically, researchers are interested not in ITE, but in CATE (conditional average treatment effect), that estimates the treatment effect for subgroups of the population:

$$CATE\ or\ uplift = E[Y_i|X_i = x, W_i = 1] - E[Y_i|X_i = x, W_i = 0] \tag{2.2}$$

where Y refers to the outcome, X is a vector of variables describing the patient, and W is the treatment. This formula, however, is only able to estimate CATE under the unconfoundedness assumption, which states that treatment (W_i) and outcome (Y_i) are independent when we condition on X_i. Consider Figure 2.5. There is non-causal association between W and Y, because X is a common cause of both. The non-causal association flows through the path W <- X -> Y. By conditioning on X, we block this alternative path, leaving only the causal relationship between W and Y. In other words, the *unconfoundedness* assumption, means that we assume that X contains all possible confounders of the relationship between W and Y, so conditioning on X, eliminates all possible confounding.

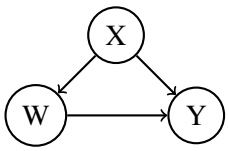


Figure 2.5: Causal structure. X is a confounder of W and Y.

Another important assumption is *Positivity*, stating that: $0 < P(W = 1|X = x) < 1$. By applying Bayes theorem on equation 2.2:

$$E[Y_i|X_i = x, W_i = 1] - E[Y_i|X_i = x, W_i = 0]$$

$$= \sum_y y \times P(Y = y|W = 1, X) - \sum_y y \times P(Y = y|W = 0, X) \tag{2.3}$$

$$= \sum_y y \times \frac{P(Y = y, W = 1, X)}{P(W = 1, X)} - \sum_y y \times \frac{P(Y = y, W = 0, X)}{P(W = 0, X)} \tag{2.4}$$

$$= \sum_y y \times \frac{P(Y = y, W = 1, X)}{P(W = 1|X = x)P(X = x)} - \sum_y y \times \frac{P(Y = y, W = 0, X)}{P(W = 0|X = x)P(X = x)} \tag{2.5}$$

We can see that if $P(W = 1|X = x)$ were to equal 0 in equation 2.5, we would have division by 0, and CATE would be undefined. In other words, the positivity assumption states that, in order to calculate CATE, all subgroups of the population must have a non-zero probability of receiving treatment.

Including more variables in X might be advantageous since we have a higher chance of covering the unconfoundedness assumption. However, as dimensionality increases, subgroups get smaller and there is a higher chance of having groups that either only receive treatment or only

receive the control, violating the positivity assumption. This concept is called *The Positivity-Unconfoundedness Trade-off*.

There are two additional assumptions that we must consider:

- *No interference* - states that one patient's outcome is not influenced by another patient's treatment, only by their own.
- *Consistency (or "no multiple versions of treatment")* - states that if the treatment is T, then the observed outcome Y is the potential outcome under treatment T.

2.2.4.3 Uplift Modeling

The literature describes three main methods for estimating uplift:

- Meta-learners - are estimators that take any classifier and apply it to estimate the missing counterfactual outcome and subsequently uplift. We will consider four meta-learner algorithms: S, T, X, and R.
- Class transformation - these algorithms estimate uplift directly, by applying a transformation to the outcome variable.
- Uplift trees - use causal measures to split the trees.

In the following sections, let n be the total number of patients, p the number of patients that underwent treatment, and m the number of patients that underwent control.

S-Learner

The S-learner uses one single machine-learning classifier. The classifier is fitted with X and W as features, to predict outcome Y (2.6). Inference is done by assigning the treatment variable to 0 or 1, and using that together with X to predict the potential outcome Y in each scenario. The uplift is then defined by the difference between the potential outcome with treatment and the potential outcome with control (2.7).

$$clf.fit\left(\begin{bmatrix} x_{11} & \cdots & x_{1k} & w_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nk} & w_n \end{bmatrix}, \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}\right) \tag{2.6}$$

$$uplift = clf.predict_proba\left(\begin{bmatrix} x_{11} & \cdots & x_{1k} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nk} & 1 \end{bmatrix}\right) - clf.predict_proba\left(\begin{bmatrix} x_{11} & \cdots & x_{1k} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n1} & \cdots & x_{nk} & 0 \end{bmatrix}\right) \tag{2.7}$$

T-Learner

The T-learner uses two machine learning classifiers. One is trained with the subgroup of X that was treated, and the other is trained with the subgroup of X that was not treated, both to predict

Y. Inference is done by getting each classifier's prediction on X. The classifier trained with treated cases returns the potential outcome under treatment, whereas the classifier trained with untreated cases returns the potential outcome under the control. The uplift is the difference in potential outcomes.

$$clf^T.fit\left(\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pk} \end{bmatrix}, \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix}\right) clf^C.fit\left(\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mk} \end{bmatrix}, \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}\right) \tag{2.8}$$

$$uplift = clf^T.predict_proba\left(\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}\right) - clf^C.predict_proba\left(\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}\right) \tag{2.9}$$

X-Learner

The X-learner training is divided into three stages. Stage 1 is equal to the T-learner, so two classifiers are fitted on subgroups of X. Then the "Imputed ITE" is calculated: if the patient was treated the imputed ITE is equal to the actual Y minus the predicted Y, by the classifier trained with control cases; whereas if the patient was in the control group, the imputed ITE is equal to the predicted Y, by the classifier trained with treatment cases, minus the actual Y. In stage 2, two regressors are fitted to predict the imputed ITE, from subgroups of X (treated and control). Finally, in stage 3, a classifier is trained to predict treatment from X, the resulting probability estimate referring to the propensity score (PS). Inference is done by applying input X to predict ITE using the classifiers fitted in stage 2, and to predict the propensity score using the classifier in stage 3. The final uplift is given by:

Stage 1

$$clf_1^T.fit\left(\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pk} \end{bmatrix}, \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix}\right) clf_1^C.fit\left(\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mk} \end{bmatrix}, \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}\right) \tag{2.10}$$

$$imputed\ ITE^T\ (\mu) = Y(T) - clf_1^C.predict_proba\left(\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pk} \end{bmatrix}\right) \tag{2.11}$$

$$imputed\ ITE^C\ (\mu) = clf_1^T.predict_proba\left(\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mk} \end{bmatrix}\right) - Y(C) \tag{2.12}$$

Stage 2

$$reg_2^T.fit\left(\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{p1} & \cdots & x_{pk} \end{bmatrix}, \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}\right) reg_2^C.fit\left(\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mk} \end{bmatrix}, \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_m \end{bmatrix}\right) \tag{2.13}$$

Stage 3

$$clf_3.fit\left(\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix}, \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}\right) \tag{2.14}$$

Inference

$$PS = clf_3.predict_proba(X) \tag{2.15}$$

$$Uplift = PS \times reg_2^C.predict(X) + (1 - PS) \times reg_2^T.predict(X) \tag{2.16}$$

R-Learner

The R-learner uses out-of-fold cross-validation estimates of the conditional mean outcome (m^*) and the treatment propensity (e^*). The authors propose equation 2.17, indicating that $E[Y_0|X_i] + w.\tau^*(X_i)$ approximates $Y(w)$, when conditioned on X and W. Here, the uplift is represented by τ.

$$E[Y(w) - \{E[Y_0|X_i] + w.\tau^*(X_i)\}|X,W] = 0 \tag{2.17}$$

Replacing the actual treatment (w) with the propensity score (e^*), we get the conditional mean outcome m^* (2.18).

$$m^* = E[Y|X_i] = E[Y_0|X_i] + e^*(X_i).\tau^*(X_i) \tag{2.18}$$

Robinson's decomposition states that:

$$Y(w) - m^* = (w_i - e^*(X_i)).\tau^*(X_i) + Y(w) - (E[Y_0|X_i] + w.\tau^*(X_i)) \tag{2.19}$$

Proof:

$$(w_i - e^*(X_i)).\tau^*(X_i) + Y(w) - (E[Y_0|X_i] + w.\tau^*(X_i)) =$$

$$= w_i.\tau^*(X_i) - e^*(X_i).\tau^*(X_i) + Y(w) - E[Y_0|X_i] - w_i.\tau^*(X_i)) =$$

$$= -e^*(X_i).\tau^*(X_i) + Y(w) - E[Y_0|X_i] =$$

$$= Y_i(w) - (E[Y_0|X_i] + e^*(X_i).\tau^*(X_i)) = Y(w) - m^*$$

The authors use Robinson's decomposition to arrive at a loss function for the uplift (τ).

$$= \hat{L}_n(\tau(x)) = \frac{1}{n}\sum_{i=1}^{n}[(Y_i - m^*(X_i)) - (W_i - e^*(X_i)).\tau^*(X_i)]^2$$

DR-Learner

The DR-learner uses cross-fitting with three data partitions $\{Y^i, X^i, W^i\}$, $i = \{1,2,3\}$. In stage 1, the first data partition is used to train a propensity score model, whereas the second data partition is used to fit two outcome regression models, with the treated cases and the control cases, respectively.

Stage 1

$$clf_1.fit(X^1, W^1) \tag{2.20}$$

$$reg_1^T.fit(X_T^2, Y_T^2) \tag{2.21}$$

$$reg_1^C.fit(X_C^2, Y_C^2) \tag{2.22}$$

Stage 2

$$\hat{p} = clf_1.predict_proba(X^3) \tag{2.23}$$

$$\hat{m}_1 = reg_1^T.predict(X^3) \tag{2.24}$$

$$\hat{m}_0 = reg_1^C.predict(X^3) \tag{2.25}$$

$$\phi = \frac{W^3 - \hat{p}}{\hat{p}(1-\hat{p})}.(Y^3 - \hat{m}_w(X^3)) + \hat{m}_1(X^3) - \hat{m}_0(X^3) \tag{2.26}$$

$$reg_2.fit(X^3, \phi) \tag{2.27}$$

Stage 3

Repeat with all data partitions.

Inference

$$uplift = \frac{1}{3}\sum_{i=1}^{3}\phi_i(X) \tag{2.28}$$

Class transformation

In the class transformation approach, a classifier is trained to model a redefined version of the outcome variable:

$$Z = y \times w + (1-y)(1-w) \tag{2.29}$$

This new variable Z will be equal to 1 if W and Y are both 1 or if they are both 0. It can be demonstrated that the probability of Z being 1 when conditioned on X, estimates uplift.

$$P(Z=1|X=x) = \frac{P(Z=1,X=x)}{P(X=x)} = \tag{2.30}$$

$$= \frac{P(Z=1,X=x,W=1)}{P(X=x)} + \frac{P(Z=1,X=x,W=0)}{P(X=x)} \tag{2.31}$$

$$= \frac{P(Z=1,X=x,W=1)}{P(X=x,W=1)} \cdot \frac{P(X=x,W=1)}{P(X=x)} + \frac{P(Z=1,X=x,W=0)}{P(X=x,W=0)} \cdot \frac{P(X=x,W=0)}{P(X=x)} \tag{2.32}$$

By equation 2.29, $Z=1 \cap W=1$ implies $Y=1$, whereas $Z=1 \cap W=0$ implies $Y=0$:

$$= \frac{P(Y=1,X=x,W=1)}{P(X=x,W=1)} \cdot \frac{P(X=x,W=1)}{P(X=x)} + \frac{P(Y=0,X=x,W=0)}{P(X=x,W=0)} \cdot \frac{P(X=x,W=0)}{P(X=x)} \tag{2.33}$$

$$= P(Y=1|X=x,W=1).P(W=1|X=x) + P(Y=0|X=x,W=0).P(W=0|X=x) \tag{2.34}$$

Under the assumption that W is independent from X:

$$= P(Y=1|X=x,W=1).P(W=1) + P(Y=0|X=x,W=0).P(W=0) \tag{2.35}$$

Under the assumption that $P(W=1) = P(W=0) = \frac{1}{2}$:

$$2.P(Z=1|X=x) = 2\left(P(Y=1|X=x,W=1).\frac{1}{2} + P(Y=0|X=x,W=0).\frac{1}{2}\right) \tag{2.36}$$

$$2.P(Z=1|X=x) = P(Y=1|X=x,W=1) + P(Y=0|X=x,W=0) \tag{2.37}$$

$$2.P(Z=1|X=x) = P(Y=1|X=x,W=1) + 1 - P(Y=1|X=x,W=0) \tag{2.38}$$

$$2.P(Z=1|X=x) - 1 = P(Y=1|X=x,W=1) - P(Y=1|X=x,W=0) = uplift \tag{2.39}$$

Transformed outcome

In the transformed outcome approach, a classifier is trained to model a redefined outcome variable:

$$Y' = Y \times \frac{W-p}{p(1-p)} \tag{2.40}$$

The expected value of Y' conditioned on X has been shown to approximate uplift. The proof, as given by (134), follows:

$$E[Y'|X] = E\left[Y \times \frac{W-p}{p(1-p)}\middle|X\right] = \tag{2.41}$$

$$= E\left[W.Y \times \frac{W-p}{p(1-p)} + (1-W).Y \times \frac{W-p}{p(1-p)}\middle|X\right] = \tag{2.42}$$

$$= E\left[W.Y_1 \times \frac{W-p}{p(1-p)} + (1-W).Y_0 \times \frac{W-p}{p(1-p)}\middle|X\right] = \tag{2.43}$$

$$= E\left[Y_1 \times \frac{W(1-p)}{p(1-p)} - Y_0 \times \frac{(1-W)p}{p(1-p)}\middle|X\right] = \tag{2.44}$$

$$= E\left[Y_1 \times \frac{W}{p}\middle|X\right] - E\left[Y_0 \times \frac{(1-W)}{(1-p)}\middle|X\right] = \tag{2.45}$$

$$= E\left[Y_1.W|X\right].\frac{1}{p} - E\left[Y_0.(1-W)|X\right]\frac{1}{(1-p)} = \tag{2.46}$$

$$= E\left[Y_1|X\right].E\left[W|X\right]\frac{1}{p} - E\left[Y_0|X\right].E\left[1-W|X\right]\frac{1}{(1-p)} = \tag{2.47}$$

$$= E\left[Y_1|X\right].p.\frac{1}{p} - E\left[Y_0|X\right].(1-p).\frac{1}{(1-p)} = \tag{2.48}$$

$$= E\left[Y_1|X\right] - E\left[Y_0|X\right] = uplift \tag{2.49}$$

Uplift trees

Uplift trees are tree-based algorithms where the splitting criterion is based on differences in uplift, with the goal of maximizing the differences between the outcome distributions in the treatment and control sets.

$$D_{gain} = D_{after\ split}(Y^T, Y^C) - D_{before\ split}(Y^T, Y^C) \tag{2.50}$$

The difference between the two distributions can be calculated in several ways. Some are described in Table 2.1.

Finally, the uplift in each leaf is given by:

$$uplift = \frac{N_{Y=1}^{W=T}+1}{N^{W=T}+2} - \frac{N_{Y=1}^{W=C}+1}{N^{W=C}+2} \tag{2.51}$$

Table 2.1: Divergence metrics used to split uplift decision trees. k indicates the leaf, p and q are the sample means in the treatment and control groups respectively.

Divergence Metric	
Kullback-Leibler (KL)	$\sum_{k=1}^{2} p_k.log(\frac{p_k}{q_k})$
Euclidean Distance (ED)	$\sum_{k=1}^{2} (p_k - q_k)^2$
χ^2-divergence (Chi)	$\sum_{k=1}^{2} \frac{(p_k - q_k)^2}{q_k}$

2.2.4.4 Uplift Model Evaluation

Because the true uplift is unknown, these models cannot be evaluated in the traditional machine learning sense. One proposed way to evaluate the models is with Uplift curves.

Uplift curves are calculated in two stages: firstly, data points are sorted by predicted uplift (highest to lowest); and secondly, a local ATE is calculated for each fraction of data points. In uplift curves, the x-axis represents the proportion of data points, whereas the y-axis shows the local ATE.

If a model is able to correctly predict uplift, it would be expected that a group of patients with a high predicted uplift would also have a high local ATE. Thus, the uplift curve of a successful model should be left-skewed, having a peak for the first fraction of data points and then gradually descending until it reaches the actual ATE for that population.

The uplift curve is evaluated qualitatively through its shape and quantitatively through the area under the uplift curve (AUUC), which are compared with the shape and AUUC of the theoretical maximum uplift curve. The theoretical maximum uplift curve is obtained by sorting the patients according to treatment response, starting with patients where the treatment effect was positive (high local ATE), such as those who received treatment and had a positive outcome, and patients who did not receive treatment and had a negative outcome, this is followed by patients where the treatment effect was negative (low local ATE), such as those who received treatment and had a negative outcome, or those who did not receive treatment and had a positive outcome.

Chapter 3

State of the Art

In this chapter, the state-of-the-art for different radiomics applications is evaluated. It aims to provide a broad understanding of the existing approaches and current advancements in the field while identifying potential research gaps that this dissertation could address.

3.1 Methodology

On January 2024, a literature search of the online database PubMed was performed to identify recent studies related to the use of prostate bi-parametric MRI radiomics applied to several use cases in the prostate cancer treatment continuum. The search strategy involved the following keywords: radiomics, prostate cancer, magnetic resonance imaging/MRI, artificial intelligence, deep learning, and machine learning. Inclusion criteria were: open access studies, and available AUC-ROC performance. Exclusion criteria: publication year older than 2019.

3.2 Results

A summary of the full results is presented in Table 3.1.

3.2.1 Detection of csPCa

Literature on the use of radiomics for prostate cancer disease characterization is extensive, and there are numerous reviews summarizing and describing recent findings and techniques (135; 136; 137; 138; 108; 139; 140). Due to the enormous quantity of literature on the subject, a further exclusion criteria was taken into consideration: studies where the number of included patients $<$ 200 were excluded. The following paragraphs describe some recent studies that present models for predicting clinical significance in prostate cancer.

Min et al. (141) gathered 280 patients with bpMRI examinations and pathological diagnosis of PCa to train a logistic regression with index-lesion radiomics, reaching an AUC of 0.82. Li et al. (142) collected 381 patients and joined index-lesion radiomics with clinical variables (age, PSA, PSAd, and prostate volume) to train a logistic regression classifier, reaching an AUC of

0.98 on the held-out cohort. Blecker et al. (143) used the ProstateX Grand Challenge dataset to extract index-lesion radiomics, and trained an XGBoost algorithm with variables selected based on domain knowledge, achieving an AUC of 0.816.

Even though the literature on prostate cancer radiomics is vastly filled with lesion-based radiomics analysis, two recent studies stand out using the whole-gland as a region of interest for the radiomics extraction. Gong et al. (144) collected 489 patients and obtained an AUC of 0.8, in a logistic regression model trained with bpMRI whole-gland radiomics and clinical variables (prostate volume, age PSA and PSA density). Rodrigues et al. (145) used the ProstateX Grand Challenge dataset to compare the performance and stability of whole-gland versus index lesion-based radiomics models, reaching a maximum of 0.85 AUC, with a random forest trained with whole-gland radiomics.

Two additional studies are worth mentioning. Even though these did not reach the 200 cases threshold defined in the inclusion criteria for this analysis, they are multi-centric studies, which is not common in the prostate radiomics literature. Castillo et al. (146) gathered 107 patients from 3 centers to build a lesion-based radiomics signature using T2W, DWI, and ADC, reaching 0.75 AUC in the held-out test set. This study stands out for being multi-centric, but also because the lesion mask was derived from pathology-MRI correlation. Similarly, Zhang et al. (147) collected mpMRI data from two centres, using one of them for external validation, where the model reached an AUC of 0.81.

3.2.2 Metastatic development

The literature on AI and prostate cancer metastasis is very clearly divided into two approaches: (1) the detection of metastatic lesions already present in the scan; and (2) the prediction of risk of developing future lesions. Since the lesions are already there, the former strategy should presumably be a simpler problem. In contrast, in the latter scenario, the model must identify features of a high risk of metastasis or even metastatic lesions that are already present but are invisible to the radiologist's naked eye.

Additionally, metastatic development of prostate cancer is primarily found in two locations, namely limph nodes and bone, so literature often describes studies focusing on only one type of metastatic development, instead of any kind metastatic development.

Regarding approach 1, Han et al. (148) used 9113 whole-body scans and CNNs to predict the presence of bone metastasis, reaching a training ROC-AUC of 0.93. Liu et al. (149) trained a 3D-Unet with 334 patients to detect and segment bone metastasis on DWI (AUCs 0.85 in training, 0.94 in a held-out test set) and T1W (AUCs 0.8 in training, 0.89 in a held-out test set). Zhang et al. (150) combined T2W, DWI and DCE lesion-radiomics with clinical risk factors to train a logistic regression to determine the presence of bone metastasis, obtaining a ROC-AUC of 0.93 in the training and 0.92 in testing. Wessels et al. (151) used lymph node histopathology slices and a pre-trained xse_ResNext34 CNN to detect lymph node metastasis, obtaining a ROC-AUC of 0.68 in the held-out test set. Hou et al. (152) combined T2W, DWI and ADC lesion-radiomics with deep learning variables and clinicopathological information to train a random forest to determine

the presence of lymph node metastasis, obtaining a ROC-AUC of 0.93 in training, 0.92 in internal testing, and 0.76 in external testing.

Regarding approach 2, Wang et al. (153) combined T2W and DCE lesion-radiomics with clinical information from 132 patients to train a logistic regression to determine the development of bone metastasis, determined at follow-up (1-10 years), obtaining a ROC-AUC of 0.92 (95% CI 0.847–0.946) in training, and 0.89 in internal testing. Liu et al. (154) gathered 332 patients to compare the performance of whole-gland vs lesion ADC radiomics extracted from manual or automatic segmentations, to predict the development of lymph node metastasis at follow-up. The highest performance was reached with the model trained with automated whole-gland mask radiomic features, with a ROC-AUC of 0.93 on internal validation and 0.73 on external validation.

Despite the success demonstrated by these studies, there are some limitations. Firstly, the small datasets that are used to train these models, tend to cause some overfitting, which was observed in the studies with external validation. Secondly, the follow-up time should extend long enough for possible metastasis to develop, however, a model trained on MRI examinations performed almost a decade ago (as reported by (153)) is not guaranteed to generalize to current exams, due to inevitable technological advancements such as hardware and software updates. To develop AI applications that are clinically relevant, they need to be validated on current data, which is impossible since the outcome for current cases is several years in the future. Thus, in order to develop AI applications that are clinically relevant, studies should focus on short-term metastatic development rather than long-term.

3.2.3 Treatment failure - Disease relapse or progression

Patients enrolled in an active surveillance program could suffer from disease progression to a stage where surveillance is no longer enough and more aggressive treatment options should be considered. Despite this, not many studies explore the predictive power of radiomics in this use case. Sushentsev et al. (155) combined T2W and ADC radiomics with clinical variables to achieve an AUC of 0.75. Two years later, this same team used time series modeling with a long short-term memory recurrent neural network (LSTM RNN) to analyze the changes in radiomic features extracted from longitudinal MRI exams from 76 patients (156). With the latter approach they achieved an AUC of 0.85.

For patients requiring more aggressive treatments (radical prostatectomy or radiotherapy), biochemical recurrence, as described in Chapter 2, can still occur. A number of studies explore the prediction of recurrence from MRI radiomics, a few of which are described in the following paragraphs.

Regarding biochemical recurrence after radical prostatectomy, Bourbonne et al. (157) combined T2W and ADC lesion-radiomics with clinical variables to predict recurrence at a minimum of 2 years follow-up time. With a train set of 70 patients, they obtained a ROC-AUC of 0.79 in training and 0.76 in validation (37 patients). One year later, the same team retrained the model with the full dataset (107 patients) and externally validated this new version with 88 patients from a separate institution, obtaining an accuracy of 0.76 (158). An et al. (159) combined T2W, PWI

and ADC whole-gland radiomics with clinical data to predict recurrence at 1 - 7 years follow-up, obtaining a ROC-AUC of 0.91 in training and 0.92 in internal validation.

Regarding biochemical recurrence after radiotherapy, Zhu et al. (160) used T2W radiomics from 67 patients to predict recurrence after 3 years follow-up, obtaining an AUC of 0.9 in the validation set. Zhong et al. (161) used T2W, T1W and DWI radiomics to train an adaboost model, obtaining an AUC of 0.99 in training and 0.73 in validation. In a slightly different approach, Iglesias et al. (162) experimented with region-wise (PZ, TZ, CZ and AS) and whole-gland radiomics, with T2W, DWI and DCE, obtaining the highest performance with PZ radiomics (AUC of 0.877).

Across the different studies, common trends arise. Relatively small data set sizes are gathered, making them vulnerable to selection bias, with the largest training set comprising only of 145 patients. Additionally, label imbalance is a consistent find in the literature on disease relapse. As recurrence only happens in a minority of prostate cancer patients, this imbalance is reflected in the gathered datasets.

3.2.4 Other applications

The utility of AI and radiomics further extends to other domains, reporting successes in the prediction of a number of clinical outcomes such as extra-capsular extension (163; 164; 165) or radiotherapy toxicity (166; 167; 168; 169; 170). Another relevant clinical outcome is the patients' quality of life, however, we were unable to identify studies in the literature using radiomics to give insights into the patients future quality of life given a certain course of treatment.

Table 3.1: Summary of the literature search. GS - Gleason score; IL - Index lesion; WG - Whole gland; FUp - follow-up; BM - Bone metastasis; LNM - lymph node metastasis; BCR - biochemical recurrence; PSA - prostate-specific antigen; PSAd - PSA density

Ref	Ground truth	Train set	Internal test set	External test set	Sequences	ROI	Clinical variables	Algorithm	AUC	Comments
Clinical significant disease characterization										
Min et al. (2019) (141)	GS >6	187	93	-	T2W DWI ADC	IL	-	LR	0.82	
Li et al. (2020) (142)	GS >6	229	152	-	T1W T2W DWI ADC DCE	IL	Age, psa, psad, prostate volume	LR	0.98	
Bleker et al. (2020) (143)	GS >6	130	76	-	T2W DWI ADC	IL	-	XGB	0.816	ProstateX dataset
Gong et al. (2020) (171)	GS >7	326	163	-	T2W DWI ADC	WG	-	LR	0.79	
Rodrigues et al. (2021) (145)	GS >7	137	46	-	T2W DWI ADC	WG vs IL	-	RF	0.85	ProstateX dataset
Castillo et al. (2021) (172)	GS >6	107 (3 centers)	20%	-	T2W DWI ADC	IL	-	WORC	0.75	Lesion mask obtained through pathology -MRI correlation

Zhang et al. (2020) (147)	GS upgrade at RP	116	50	-	T2W DWI ADC DCE	IL	age, time from biopsy to radical prostatectomy, total PSA, free PSA, the ratio of free/ total PSA, PSAD, number and percentage of positive biopsy cores, clinical stage, biopsy GS, and prostatectomy GS	DT	0.91	
Metastatic Development										
Han et al. (2022) (148)	BM+ (2991) BM- (6142)	9113	?	?	Whole body bone scans			CNN	0.93	
Liu et al. (2021) (149)	BM+ (168) BM- (166)	334	63		DWI T1W			3D UNet	0.94	test set with contemporary data
Zhang et al. (2020) (150)	BM+ (55) BM- (61) determined by ECT/PET-CT	81	35	-	T2W DWI DCE	IL	age, PSA level, T stage, lymph node stage, Gleason score and histological type	LR	0.92	
Wessels et al. (2021) (151)	LNM+ (102) LNM- (116)	118	100	-				pre-trained xse_ResNext34 CNN	0.68	Histopathology slices

Hou et al. (2021) (152)	LNM+ (51) LNM- (229) Found at RP	280 (centre 1)	71 (centre 1)	50 (centre 2)	T2W DWI ADC	IL	age, PSA, PSAD, clinical T stage, biopsy Gleason score, percentage of positive cores and perineural invasion prostate volume, zone of lesion origin, shape, margin, maximum diameter, volumetric ADC, PI-RADS score, MRI T-stage, MRI-based ECE, SVI and PLNM	RF	0.93 (training) 0.92 (internal) 0.76 (external)	Included deep transfer learning features of the slice with the largest lesion volume (Inception v3, SqueezeNet, VGG16, VGG19, DeepLoc); In FUp, patients stratified by the PLNM-Risk calculator showed higher BCR-free survival.
Wang et al. (2019) (153)	BM+ (91) BM- (85) at FUp (min 1 year, max 10 years)	132	44	-	T2W DCE		PSA, gleason, age	LR	0.916 (training) 0.891 (internal)	Only texture radiomics
Liu et al. (2022) (154)	LNM+ (236) LNM- (238) at FUp imaging	332	142	128	ADC				Model 3: 0.93 (internal) 0.73 (external)	Manual gland mask (model 1) vs manual lesion mask (model 2) vs automated gland mask (model 3) vs automated lesion mask (model 4)
Disease progression in patients enrolled in active surveillance programs										
Sushentsev et al. (2021) (155)	Proggression (35) Stable disease (38)	73	-	-	-			RF, LR, LDA, KNN, SVM, GLMnet	0.75	LOOCV
Sushentsev et al. (2023) (156)	Proggression (28) Stable disease (48) 297 scans	76	-	-				LSTM-RNN	0.86	LOOCV
Disease relapse after radical prostatectomy										

Bourbonne et al. (2019) (157)	BCR+ (17) BCR- (90) Min 24 months FUp PSA >0.2 ng/mL (two successive blood samples) +D12	70	37		T2W ADC	IL			0.79 (training) 0.76 (internal)	
Bourbonne et al. (2020) (158)	BCR+ (34) BCR- (46)	107 (centre 1)		88 (centre 2)					Accuracy: 0.78 (training) 0.76 (external)	
An et al. (2023) (159)	BCR+ (77) BCR- (129) 1 - 7 years FUp (PSA ≥0.2 ng/mL)	145	61		T2W PWI ADC	WG			0.91 (training) 0.93 (internal)	
Disease relapse after radiation therapy										
Zhu et al. (2023) (160)	BCR+ (19) BCR- (64) 3 years FUp (PSA ≥0.2 ng/mL)	67	16		T2W				0.9 (internal)	
Zhong et al. (2020) (161)	BCR+ (29) BCR- (62) 10-131 months FUp (PSA ≥0.2 ng/mL)	73	18		T2W DWI T1W			Inception-Resnet v2; Adaboost model	0.99 (training) 0.73 (internal)	
Iglesias et al. (2023) (162)	BCR+ (20) BCR- (108) 10 years FUp (PSA ≥0.2 ng/mL)	128			T2W	PZ, CZ, TZ, WG			0.877 (PZ) 0.841 (CZ+TZ) 0.862 (SV) 0.855 (WG)	Automated segmentation Co-registration of DWI and DCE

Chapter 4

Handcrafted and deep radiomic features towards training robust classifiers

This chapter outlines the work that was done in an effort to get more insight and possibly overcome the reader variability in the segmentation step of the radiomics pipeline. In addition to this, the predictive power of "deep features" extracted from segmentation networks was assessed in the classification of prostate cancer disease aggressiveness. This work resulted in a publication (16).

4.1 Introduction

Radiomics has been shown to be highly dependant on the acquisition protocols and image pre-processing steps (112; 113; 6). Unfortunately, the extreme inconsistency in the extraction process, makes it difficult to reproduce, which has consequences in the final model's generalizability to external data. Overcoming this relies on image pre-processing steps that reduce the protocol's influence on the extracted features, such as intensity normalization and bias-field correction.

Another major limitation is the tight link between the computed radiomic features and the volume of interest (VOI) from where they have been extracted. Tissue or lesion segmentation is performed either manually, when a radiologist outlines the boundary of the VOI in the image, automatically, for example by a deep learning algorithm trained to segment a certain VOI, or semi-automatically, where the mask is drawn in an automated fashion and later verified and corrected by a radiologist. When manually traced, this is a time-consuming and subjective process, that naturally suffers from inter- and intra-reader variability, being dependent on the expert's perception and level of experience (115; 116). These slight real-world differences in human-defined segmentation margins may in principle affect the distribution of the calculated radiomic features, which subsequently affect the algorithms trained with them (116), given that low quality segmentations will negatively impact model performance. Building machine learning algorithms that are robust

to this real-world heterogeneity is essential to a future safe application of AI methods in the clinical setting.

Parallel to the growing interest in radiomic features, deep Neural Networks have emerged as a promising technique for prostate cancer detection and segmentation of anatomic zones or tumorous lesions (139). However, it has been shown that deep learning models tend to overfit when attempting to solve prostate cancer classification problems, not generalizing to out of distribution data (172). Despite the large number of studies published, very few compare the performance of handcrafted radiomic features and deep features on the same data and objective (172; 173).

4.2 Methodology

4.2.1 Research Questions

Two research questions (RQ) were addressed in this chapter, namely RQ 1 and RQ 2, described in Chapter 1 as:

RQ 1 *Which is the best approach to train robust classifiers to inter- and intra-reader variability in segmentation margins?*

RQ 2 *Is the combination of deep and handcrafted radiomic features (I) capable of outperforming radiomic features alone (O) in different machine learning tasks (C) for patients in the prostate cancer disease continuum (P)?*

An example of the common differences in segmentation margins can be found in Figure 4.1. To answer RQ 1, different approaches at training were compared:

1st approach (**stableRad1**) Perform a feature stability analysis and train only with stable features.

2nd approach (**Rad1**) Train with features extracted from masks drawn by radiologist 1 (rad1).

3rd approach (**Rad2**) Train with features extracted from masks drawn by radiologist 2 (rad2).

4th approach (**avgRad**) Train with the feature average between both radiologists.

5th approach (**intersectionRad**) Train with features extracted from the intersection of the two masks.

6th approach (**unionRad**) Train with features extracted from the union of the two masks.

7th approach (**resampledRad**) Train with a randomly resampled dataset where for some patients the extraction was performed from the mask drawn by rad1 and for others the extraction was performed from the mask drawn by rad2.

The dataset that resulted in the most robust classifier was selected for further analysis (RQ 2). To answer RQ 2, three approaches were compared for model training. These included a handcrafted radiomics dataset, a dataset with "deep" features, and a hybrid dataset with both.

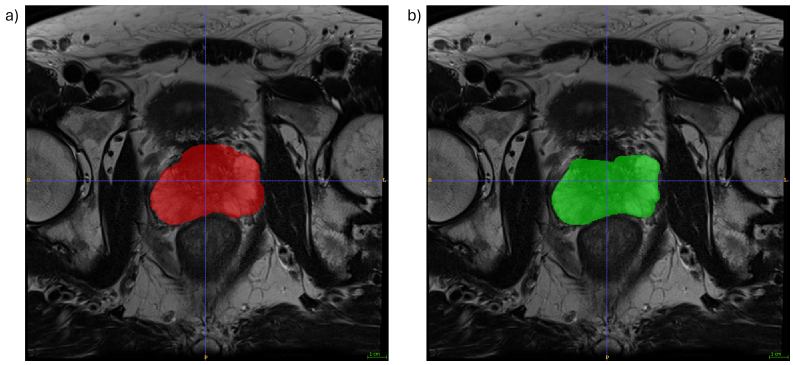


Figure 4.1: An example of the evident segmentation variability in the masks drawn by radiologist 1, (a), and radiologist 2, (b), on patient Prostatex0000.

4.2.2 Data description

The dataset consisted of T2W, DW, and ADC data from the SPIE-AAPM-NCI PROSTATEx challenge (174; 175; 176). The MRI exams were acquired at the Prostate MR Reference Center - Radboud University Medical Centre in the Netherlands. Due to the public nature of the data, ethics committee approval and patient consent were waived for this study. The dataset is composed of 181 patients. Due to complications with some segmentation files, the full extent of the PROSTATEx dataset was not used. The full list of excluded patients can be found in Appendix A (Table A.1). The approximate location of the centroid of each lesion was provided in DICOM coordinates. Cancer was considered to be clinically significant when the biopsy Gleason score was 7 or higher. The lesions were labelled with "TRUE" and "FALSE" for the presence of clinically significant cancer, with a distribution of 67 clinically significant lesions (TRUE) and 214 clinically non-significant lesions (FALSE). A gland was considered to have clinically significant cancer if at least one of its lesions is clinically significant for prostate cancer. This resulted in a label distribution of 122 clinically insignificant glands and 59 clinically significant glands, giving a total of 181 patients.

4.2.3 Methods specific to RQ1

4.2.3.1 Segmentation

Manual segmentations of the whole prostate gland were performed independently by two radiologists (M.L., 10 years of experience, and A.U., radiology resident) on T2W and high b-value DWI images separately.

4.2.3.2 Radiomic Features Extraction

Bias field correction was performed on T2W images using the N4 Bias Field Correction algorithm (177) and the Python package Simple ITK (version 2.0.0) (178). First, each image's x-, y- and z-spacing was checked for discrepancies. Since x- and y-spacings differed from z-spacing, feature extraction was later performed in 2D. Additionally, T2W images' x- and y-spacings differed

within and between patients, so these were resampled to the highest value of 0.5. Non-quantitative images' (T2W and DWI) intensities were normalized. The bin width was selected to produce discretized images with between 30 and 130 bins. This resulted in a bin width of 20 for T2W images, 5 for DWI, and 70 for ADC maps.

Radiomic features were extracted from the whole gland segmentation using the Pyradiomics package (version 3.0) (179) in Python (version 3.7.9) (180). All the pre-processing steps were performed as parameters of the extractor function, except for the bias field correction, which was performed prior to the extraction. The image transforms wavelet, square, squareroot, logarithm, exponential, gradient, and local binary pattern were enabled, resulting in a total of 3111 extracted features, 1037 from each MRI modality (T2W, high b-value DWI and ADC). In the feature extraction of the ADC map, the mask drawn on the DWI was used. The mathematical expressions and semantic meanings of the extracted features can be found at `https://pyradiomics.readthedocs.io/en/latest/`.

4.2.3.3 Spatial Stability of Radiomic Features

For approach 1, spatial stability was assessed by comparing the features extracted from the VOIs created by each radiologist. This analysis was conducted with a two-way, single rater, absolute agreement Intraclass correlation coefficient (ICC) formulation (ICC 2.1) (181). Features with ICC 95% confidence interval lower limit over 0.75 were considered to be robust to segmentation and were kept for further analysis.

4.2.4 Methods specific to RQ2

On top of the radiomic features extracted for RQ1, to answer RQ2, deep features were also extracted.

4.2.4.1 Deep Features Extraction

Deep features were extracted from segmentation models trained to segment the whole gland. A comprehensive description of the training as well as network performance can be found in previous work (182). The ground truth masks used for the gland were obtained as described previously in Section 4.2.3.1, while both the peripheral and transitional zone masks were the publicly available ones from the SPIE-AAPM-NCI PROSTATEx challenge.

The models used to perform the segmentations were all Encoder-Decoder Unet variations, namely: Unet (183); Unet++ (184); Attention Unet (AUnet) (185); Dense Attention Unet (DUnet); Dense-2 Unet (D2Unet) (186); Dense-2 Attention Unet (D2AUnet); Recurrent Residual Unet (R2Unet) (187); Recurrent Residual Attention Unet (R2AUnet).

Taking these segmentation models, we removed the decoder part of the network and devised three different strategies to extract the deep features from the bottleneck (Figure 4.2). The first approach consisted of averaging all the feature maps into a single one before flattening it, producing 200 features. For the second and third approaches, we removed all spatial dependencies

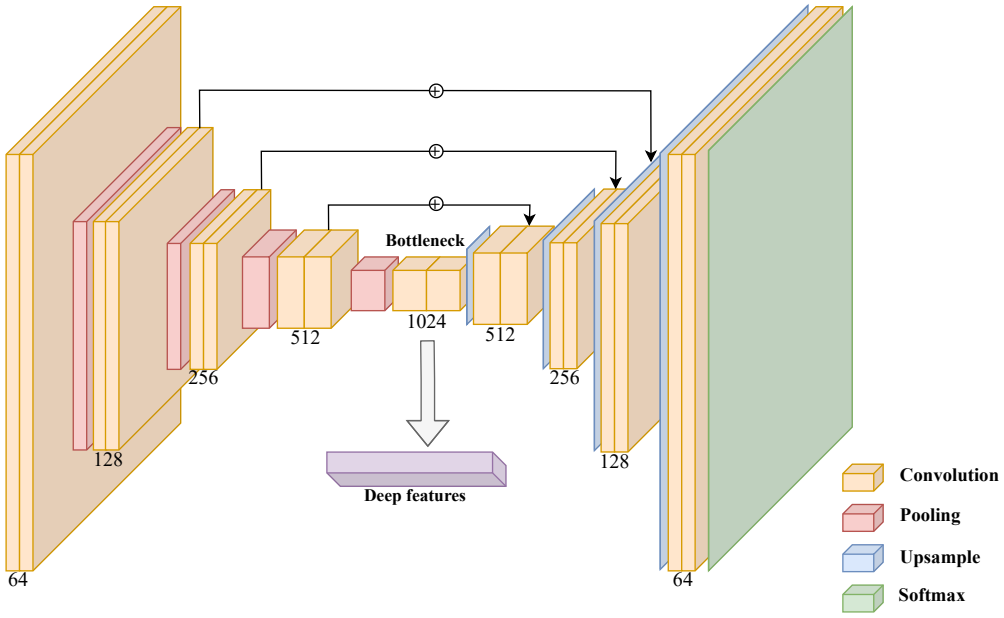


Figure 4.2: Representation of the deep feature extraction procedure from the segmentation models. These features are extracted from the bottleneck of the encoder-decoder model following three different strategies.

of the feature maps by performing pooling operations, one version with Max-pooling and another with Average-pooling, per feature map. This reduced each feature map to a single value, which was then concatenated, producing sets of 1024 features. The 8 different neural networks described were combined with the 3 different strategies for calculating the features, resulting in 24 deep datasets. In addition, the best radiomics dataset, selected in RQ I, was combined with the different whole gland deep features datasets to produce another 24 hybrid datasets.

4.2.5 Train/Test split

The train/test split was performed at a patient level with the Python scikit-learn package (version 0.23.2) (188; 189; 190). The hold-out test set consisted of 33% randomly selected patients and the split was stratified so that both train and test sets have the same proportion of True labels. The patient list of the original split was used to split the remaining datasets, ensuring that the patients belonging to the hold-out test set did not differ between datasets and avoiding data leakage. The hold-out test set comprised 58 patients.

4.2.6 Dataset pre-processing

The preprocessing pipeline described in this section was fitted on the train set and applied on the train and test sets.

Features with zero or near-zero variance were identified and excluded with the nearZeroVar method of the R caret package (version 6.086) (191).

Feature correlation was then assessed. Feature pairs were considered correlated if their Spearman correlation was higher than 0.75. Out of the two, the feature with the highest average correlation across all features was eliminated.

Feature values were scaled to have 0 mean and standard deviation equal to 1.

To cope with the imbalanced nature of the data, the SMOTE (synthetic minority oversampling technique) (192) algorithm was applied.

Finally, due to the high-dimensional feature space, a computationally "light" tree-based feature selection algorithm was also explored, where features with the highest tree importance were selected. The number of features selected was a hyperparameter optimized during model training.

4.2.7 Model Development

The scikit-learn implementation of two machine learning algorithms was used: Random Forest Classifier (193) and Logistic Regression. These were chosen since they are on opposite ends of the bias-variance spectrum. Hyperparameter tuning was performed for each algorithm with an exhaustive grid search and each parameter combination was evaluated through cross-validation. The classifiers' probabilities were calibrated (method = isotonic) using the Python scikit-learn package (version 0.23.2) (188; 189; 190).

4.2.8 Model Evaluation

The probabilistic decision threshold for binary classification was chosen to achieve at least 90% sensitivity on the training set. For that reason, the area under the receiver operating characteristics curve (AUC) was used for hyperparameter tuning and pipeline selection since it is a metric invariable to the decision threshold. The pipelines were validated internally through 3-fold cross-validation, and the one with the highest cross-validation AUC was selected for each approach. For RQ I, the selected classifiers were applied to three hold-out test sets: each independent radiologist's test set (approaches 2 and 3) and the resampled radiologists' test set (approach 7). These hold-out test sets were chosen for being the ones that resemble real-world scenarios the most. For RQ II, the classifiers were applied to the hold-out test set respective to the same dataset as used for training. The overall pipeline followed in this study can be found in Figure 4.3.

4.2.9 Statistical comparison of Classifiers

The classifiers selected for each approach were compared with McNemar's test implemented by the mlextend python package (version 0.21.0) (194). P-values were corrected for multiple comparisons with the FDR method and statistical significance was considered at $\alpha = 0.05$ for all tests.

4.2.10 SHAP importance of predictive variables

A SHapley Additive exPlanations (SHAP) analysis (195) was used to identify the most relevant variables for the prediction. This was completed using the Python package shap (version 0.38.1).

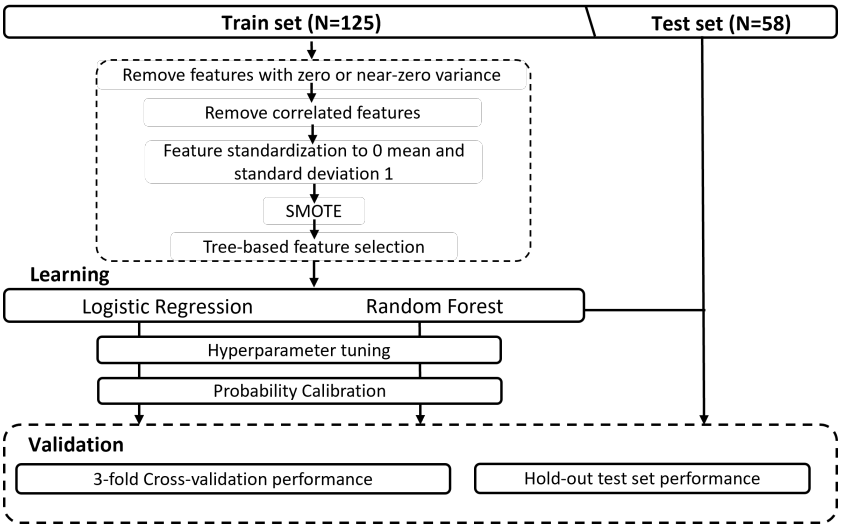


Figure 4.3: Overall model development pipeline followed in this study.

4.3 Results

4.3.1 Research Question 1

4.3.1.1 Hold-out Test Set Performance

For five of the approaches, the selected pipeline included Random Forest, while for the remaining approaches, 1 and 5, it performed better whilst using Logistic Regression. The training hyperparameters, the decision threshold, and the number of features can be found in Appendix A (Table A.2).

Applying the selected classifiers to rad1, rad2, and resampledRad hold-out test sets, Tables 4.1 a), 4.1 b), and 4.1 c), respectively, it is shown that the resampledRad dataset consistently produces the highest performing classifiers, following approach 7. On the contrary, the lowest-performing models were obtained by approach 1, which consisted of training classifiers only with features with high ICC between radiologists.

4.3.1.2 Statistical Comparison of Classifiers

The p-value results for the statistical comparison of the classifiers with FDR correction can be found in Table 4.2. As expected, there is a statistically significant difference between the higher performing classifiers, rad1, rad2, and resampledRad (approaches 2, 3, and 7, respectively), and the lowest performing classifiers stableRad1 and intersectionRad (approaches 1 and 6, respectively). Furthermore, the differences between training classifiers with rad1, rad2, or resampledRad data proved statistically non-significant. The resampledRad dataset was selected for further analysis.

Figure 4.4 shows ROC curve for the resampledRad classifier. The decision threshold chosen for this classifier was 0.32, which ensured a minimum 0.9 sensitivity on the train set. Analysing

Table 4.1: Classification performance on three different hold-out test sets. One from each radiologist (subtables a and b) and one with resampled radiologists (subtable c). The highest value per column is highlighted in **bold**.

a)

Training data	**rad1 Hold-out test-set performance**				
	F2	**CohensKappa**	**AUC**	**Sensitivity**	**Specificity**
stableRad1	0.6696	0.2042	0.7625	0.7895	0.4615
Rad1	0.7353	0.4636	0.7827	0.7895	0.7179
Rad2	0.7143	0.5110	**0.8219**	0.7368	**0.7949**
avgRad	0.6881	0.2758	0.7584	0.7895	0.5385
unionRad	0.7522	0.3073	0.7814	**0.8947**	0.4872
intersectionRad	0.6364	0.1863	0.6802	0.7368	0.4872
resampledRad	**0.7767**	**0.5055**	0.8198	0.8421	0.7179

b)

Training data	**rad2 Hold-out test-set performance**				
	F2	**CohensKappa**	**AUC**	**Sensitivity**	**Specificity**
stableRad1	0.6757	0.2275	0.7605	0.7895	0.4872
Rad1	0.7353	0.4636	0.7794	0.7895	0.7179
Rad2	0.6250	0.4208	0.8151	0.6316	**0.7949**
avgRad	0.7075	0.3525	0.7901	0.7895	0.6154
unionRad	0.7522	0.3073	0.7659	**0.8947**	0.4872
intersectionRad	0.6522	0.1371	0.6356	0.7895	0.3846
resampledRad	**0.7843**	**0.5351**	**0.8381**	0.8421	0.7436

c)

Training data	**resampledRad Hold-out test-set performance**				
	F2	**CohensKappa**	**AUC**	**Sensitivity**	**Specificity**
stableRad1	0.6696	0.2042	0.7537	0.7895	0.4615
Rad1	0.7353	0.4636	0.7841	0.7895	0.7179
Rad2	0.6633	0.4358	**0.8192**	0.6842	**0.7692**
avgRad	0.6944	0.3008	0.7746	0.7895	0.5641
unionRad	0.7522	0.3073	0.7827	**0.8947**	0.4872
intersectionRad	0.6522	0.1371	0.6430	0.7895	0.3846
resampledRad	**0.7767**	**0.5055**	**0.8192**	0.8421	0.7179

Figure 4.4, it is possible to conclude that the decision threshold would need to be lowered to at least 0.25 to reach a sensitivity of 0.9 across the test sets, providing evidence of overfitting.

4.3.1.3 Predictive Variables Selected

The resampledRad classifier uses 40 radiomic features, the majority of which was calculated from the ADC map (24 / 40 features). The T2W and DWI images contributed with 7 and 9 features, respectively. As far as image filters/transforms are concerned, gradient was the most substantial contributor with 12 features, this was followed by the original image with 6 features, and the remaining filters contributed between 1 and 4 features, with the exception of the square filter, which was the only one not represented. Regarding types of radiomic features, the most represented group is texture, with 28 features included. First-order and shape contributed with 9 and 3

Table 4.2: *p*-value results with FDR correction for the statistical comparison of classifier performance on the resampledRad hold-out test set. p-values < 0.05 are highlighted in **bold**. The column and the row corresponding to resampledRad and stableRad1, respectively, have been removed due to not containing information.

	stableRad1	**rad1**	**rad2**	**avgRad**	**unionRad**	**intersectionRad**
rad1	0.0631					
rad2	**0.0466**	0.7784				
avgRad	0.6003	0.1511	0.1799			
unionRad	0.7784	0.1624	0.1802	1.0000		
intersectionRad	0.7784	**0.0097**	**0.0223**	0.1624	0.2958	
resampledRad	**0.0321**	1.0000	1.0000	0.0969	0.0969	**0.0097**

features, respectively.

4.3.1.4 SHAP Importance Analysis

Regarding each feature's contribution to model output, a shap analysis was performed, attempting to explain the predictions of each hold-out test set. The five features with the highest impact were consistent on the three hold-out test sets. Thus, as an example, the shap analysis for the predictions of the resampledRad classifier on the resampledRad test set is displayed in Figure 4.5. The feature DWI_gradient_glcm_lmc1 is inversely associated with a clinically significant output, while the remaining four features are directly associated with it.

4.3.2 Research Question 2

Out of the 24 deep and 24 hybrid datasets generated, the classifiers with the highest performance were selected. The deep features dataset that performed the highest was extracted from a Dense-2 Attention Unet and calculated through the max pooling approach, while the selected hybrid dataset was composed of resampledRad radiomic features and deep features extracted from a Dense-2 Unet and calculated through max pooling.

Comparing the classifiers trained with radiomic or deep features (Table 4.3), the classifier trained with deep features shows a higher cross-validation (CV) performance, however, the larger difference between CV and hold-out test performance indicates a larger degree of overfitting. These results are concordant with what is currently found in the literature (172). The classifier trained with the hybrid dataset achieves a similar CV performance to the other models. However, even though it is less overfitted, the classifier with the lowest degree of overfitting is still the radiomics one.

4.4 Discussion

In the current study, radiomic and deep features were extracted from the whole prostate gland rather than the lesion. We attempted to answer the well-known issue of inter-reader variability introduced into the radiomics pipeline in the segmentation stage (RQ I). Finally, this study

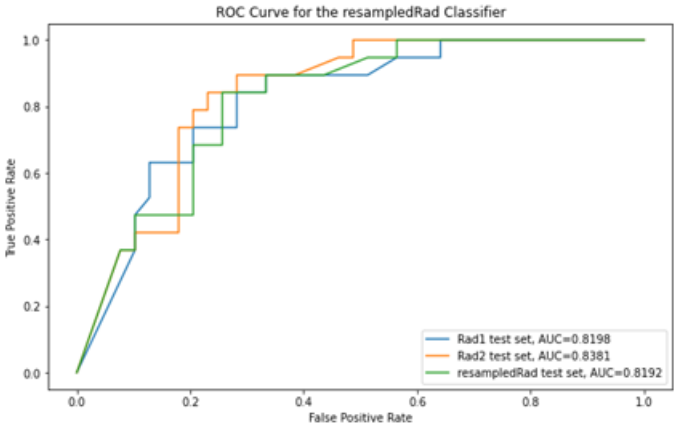


Figure 4.4: Receiver Operator Characteristics Curve for the resampledRad classifier when applied to the rad1, rad2 and resampledRad hold-out test sets, respectively in blue, orange and green. Some of the probability decision thresholds are included as annotations.

demonstrates the promising combination of radiomic and deep features for classifying prostate cancer disease aggressiveness (RQ II).

Publications extracting radiomic features from the whole prostate gland are rare, but there is evidence to support this method in contrast to single-lesion radiomics (196; 145). Previous work (145) has reported that whole-gland radiomic features have higher stability to segmentation, since gland borders are easier to be defined comparing to a lesion's, and that these features produced classifiers with higher performance and less overfitting. This might be justified by the existence of regions in the prostate, outside of lesions, that have relevant information for the classification and would not be easily visually recognized and subsequently segmented by the radiologist.

Radiomic features' stability is essential for developing robust models to use in clinical scenarios. However, it is uncommon to find in the literature studies that assess this issue. From the studies that address it, most use the intraclass correlation coefficient to evaluate the agreement between readers and exclude features where a disagreement is found, training classifiers using purely 'stable' features (197; 198). The issue with this approach is that a disagreement between the segmentation borders does not necessarily mean the resulting features would not be good predictors, it simply means the readers do not agree on what its value should be. Hence, new approaches to combining information from two or more radiologists are of high importance. In this study, this was addressed with research question I.

Regarding this first research question, results revealed that approach 1, corresponding to the evaluation of feature stability through inter-rater absolute agreement and subsequent removal of unstable features, a technique currently recommended by radiomics guidelines and evaluated in the radiomics quality score (199), proved to produce the classifiers with the least ability to generalize to hold-out data. On the other hand, approach 7, corresponding to training classifiers with a radiomics dataset where segmentation masks were randomly chosen from the two available radiologists, proved to be the highest performing across all hold-out test sets. Supporting the hypothesis that the more heterogeneous the training data the more generalizable the classifier may be on

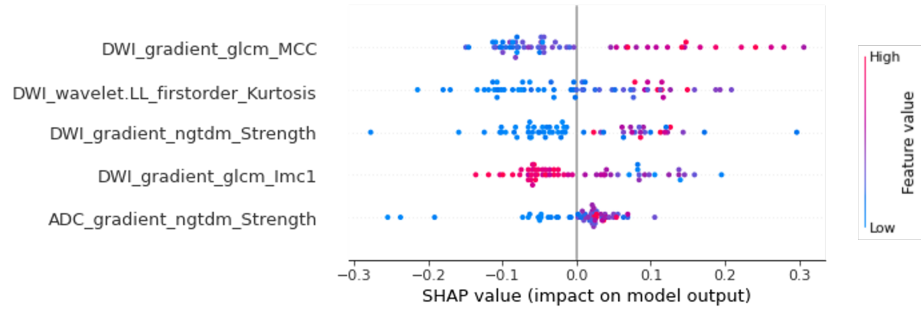


Figure 4.5: Summary beeswarm plot showing the five features with the highest impact on resampledRad model output according to a SHAP analysis explaining the model's predictions for the resampledRad hold-out test set.

unseen data. Additionally, the performance of such classifier was very similar on the different hold-out test sets, indicating its robustness to radiologists with different years of experience. This was further confirmed by the performance on the resampledRad test set, which simulates a real-world clinical environment, where a deployed model would be used by several physicians. Thus, these results are extremely relevant for the clinical translation of AI models. As of right now, but still staying vigilant to further validation studies, the results suggest that gathering segmentations from different radiologists will produce classifiers that are more robust to slight differences in segmentation margins.

Regarding research question II, the few publications comparing deep learning and radiomics-trained classical machine learning on the same classification problem (172; 173) reported higher performance on the train set when using deep learning, but lower performance on the test set when compared to classical machine learning algorithms trained with radiomic features. In this work, deep learning's natural tendency for overfitting was confirmed, both on the deep and hybrid classifiers. Even though the hybrid classifier showed a lot less overfitting than the deep model, it was not enough to outperform the radiomics classifier. Despite this, this hybrid approach is worth exploring further with larger datasets and externally validated.

This study has several limitations. First, this was a retrospective study, so a multi-center prospective analysis should be carried out to validate these results and investigate the impact these predictive models have on patient outcome. Second, only T2W, DWI, and ADC sequences were used. Other sequences, such as dynamic contrast-enhanced MRI, could be worth exploring, however, since these are not consistently part of the MRI examination protocols they were not included in the models. Third, although the overall class imbalance was addressed through SMOTE upsampling of the minority class, the imbalanced nature of the anatomical location of lesions was not, with the large majority of lesions belonging to the peripheral zone. It would be interesting to investigate the model's performance on the different anatomical zones independently. Fourth, using a publicly available dataset increased transparency but limited the access to clinical data, such as PSA levels, patient age, or PI-RADS score, which are fundamental components of a clinician's assessment, but could not be included in the model. Fifth, proper assessment of real-world clinical

Table 4.3: Performance of classifiers trained with radiomic, deep and hybrid datasets. a) 3-fold cross-validation, and b) hold-out test set performance.

a)

Training data	**3-fold cross-validation performance (95% CI)**				
	F2	**CohensKappa**	**AUC**	**Sensitivity**	**Specificity**
Rad	0.8830 (0.8129, 0.9570)	0.7350 (0.5300, 0.8562)	0.8679 (0.7654, 0.9281)	0.8920 (0.8164, 0.9762)	0.8439 (0.6697, 0.9163)
Deep	0.9577 (0.8267, 0.9675)	0.8609 (0.6684, 0.9235)	0.9311 (0.8338, 0.9615)	0.9753 (0.8135, 0.9762)	0.8870 (0.7701, 0.9639)
Hybrid	0.9392 (0.8236, 0.9892)	0.9164 (0.7175, 0.9444)	0.9583 (0.8585, 0.9722)	0.9306 (0.8056, 1)	0.9861 (0.8282, 0.9868)

b)

Training data	**Hold-out test-set performance**				
	F2	**CohensKappa**	**AUC**	**Sensitivity**	**Specificity**
Rad	0.7767	0.5055	0.8192	0.8421	0.7179
Deep	0.3333	0.2648	0.8131	0.2941	0.9355
Hybrid	0.5063	0.4062	0.8710	0.4706	0.9032

performance is only possible through external validation. This important validation step will be addressed in future work. Finally, inherent to the Gleason system is the subjectivity of cancer grading, so note must be taken that the gold standard used in this study is subject to human error and inter or intra-observer variability. In addition, the definition of clinical significance might be based on more than the Gleason score alone, and variables such as tumour volume or tumour category might be relevant.

In conclusion, the results presented in this study are relevant to the clinical translation of AI models. Heterogeneous radiomics datasets where segmentation masks come from more than one radiologist produced classifiers with the highest generalization power. Additionally, the combination of radiomic and deep features in the classification of prostate cancer disease aggressiveness is studied. Here, promising results have been reported with the hybrid approach, which is worth exploring further with larger datasets.

Chapter 5

The ProstateNet Dataset

In Chapter 4, the ProstateX Challenge dataset was used due to the unavailability of the ProstateNet dataset. The latter will be used in the remaining work of this dissertation. Thus, in this Chapter, the contents, organization and curation of the ProstateNet dataset will be described.

5.1 Introduction

The ProCAncer-I is an EU-funded project created for the development of MRI-based AI tools, acting along the full extent of the disease continuum of prostate cancer. Namely, with the purpose of tackling nine specific clinical challenges:

- Use Case 1 - cancer detection on MRI
- Use Case 2 - cancer characterization as low-, intermediate- or high-risk according to its biological aggressiveness (clinical significance).
- Use Case 3 - early identification of present or distant metastatic development.
- Use Case 4 - biological validation of AI models through radiologic - histopathologic correlation.
- Use Case 5 - risk of disease recurrence after radical prostatectomy.
- Use Case 6 - risk of disease recurrence after radiation therapy.
- Use Case 7a - prediction of radiation-induced toxicity side-effects after radiation therapy.
- Use Case 7b - prediction of quality of life side-effects after radical prostatectomy.
- Use Case 8 - prediction of early withdrawal from the active surveillance program, evidence of rapid disease progression.

With this purpose in mind, the ProstateNet dataset was created. The ProstateNet is the largest Prostate Cancer mpMRI dataset reported to date. It is composed of two main objects: the image

repository and the clinical/radiological database. In order to serve the different clinical use cases and to ensure representability, highly diverse data was uploaded by 13 clinical partners, from 9 different countries, 4 scanner manufacturers, and numerous scanner models, as well as annotations from several radiologists with different experience levels.

5.2 Methodology

The work of this dissertation will include use cases 1, 2, 3, 5, 6, 7b, and 8.

5.2.1 Inclusion criteria

MRI studies whose acquisition parameters fall within the ranges described in Table 5.1 were eligible.

Table 5.1: Description of the range of MRI acquisition parameters eligible to be included in ProstateNet.

	T2W	**DWI**
acquisition plane	straight axial or oblique axial matching the long axis of the prostate	same as T2W
slice thickness	$\leq$ 4 mm	$\leq$ 4 mm
FOV	$\leq$ 240 mm	$\leq$ 250 mm
in-plane dimension	$\leq$ 1.2 mm (phase) x $\leq$ 1.2 mm (frequency)	$\leq$ 2.5 mm (phase and frequency)
b-values		$\leq$ 100 s/mm^2 (low) $\leq$ 1000 s/mm^2 (high)

The inclusion criteria for each clinical use case were the following:

- UC1 - Negative cases were of two types: no lesion found on MRI with at least 1 year follow up, or lesion found on MRI with negative biopsy. Positive cases: lesion found on MRI, with biopsy confirmation of PCa.
- UC2 - Lesion found on MRI, and biopsy confirmation of PCa. Ground truth determined by biopsy or radical prostatectomy (if available).
- UC3 - Lesion found on MRI, and biopsy confirmation of PCa. Metastatic status determined at at least six months follow-up.
- UC4 - Only for prospectively collected patients. Out of the scope of this dissertation.
- UC5 - Lesion found on MRI, biopsy confirmation of PCa, and radical prostatectomy performed. Recurrence determined at at least six months follow-up.
- UC6 - Lesion found on MRI, biopsy confirmation of PCa, and radiation therapy performed. Recurrence determined at at least six months follow-up.
- UC7a - Lesion found on MRI, biopsy confirmation of PCa, radiation therapy performed, and available post-treatment toxicity information.

- UC7b - Lesion found on MRI, biopsy confirmation of PCa, radical prostatectomy performed, and available post-treatment quality of life information.
- UC8 - Lesion found on MRI, biopsy confirmation of PCa, and patient enrolled into an active surveillance program. Ground truth determined at six months follow-up.

5.2.2 Data curation

Despite the extent of ProstateNet, many quality issues were identified. The quality issues in the image repository were described and addressed in (200). The curation of the database is described bellow:

1. The initial database file comprised 68 280 rows and 106 columns, corresponding over 12000 patients with an average of 6.8 MRI series uploaded per patient.
2. From these 12000 patients, 10748 had available T2W sequences in the image repository.
3. 21 variables were selected from the 106 available, namely: "patient_id", "use_ca se_form", "study_uid", "series_uid", "manufacturer", "manufacturer_model_name", "receive _coil_name", "provided_by", "prospective", "age_at_baseline", "M_Diagnosis", "epic26", "biochemicalRelapseFUp", "prostatectomy_nerve_sparing", "extra_prostatic_extension", "perineural_invasion", "seminal_vesical_invasion", "resection_margin_status", "prostatectomy _method", "ecrf" and "leftSurveillance".
4. The non-uniform reporting of scanner manufacturer ("SIEMENS", "GE MEDICAL SYSTEMS", "Philips Medical Systems", "Philips Healthcare", "Philips", "Siemens", "Siemens Healthineers", "TOSHIBA", "Siemens HealthCare GmbH" and "TOSHIBA_MEC") was standardized to: "PHILIPS", "SIEMENS", "GE MEDICAL SYSTEMS" and "TOSHIBA".
5. The column "country" was inferred from the data provider (variable "provided_by").
6. The presence of endorectal coil (ERC) in the MRI exam was inferred from the receiver coil used (variable "receive_coil_name"). A new boolean column ("ERC") was created with the value of True whenever the receiver coil was one of the following: "8EIS_eCoil BodyL", "ATDTORSO", "Body E24 AA1", "Body E24 AA2", "Body E24 AA3", "EIS_eCoil Torso". The variable "receive_coil_name" was removed.
7. A column "Treatment" was inferred from the "use_case_form" column in the following manner: patients in UC5 were considered to have undergone radical prostatectomy (RP), patients in UC6 were considered to have undergone radiation therapy (RT) and patients in UC8 were considered to have enrolled in an active surveillance (AS) program.
8. From the "ecrf" variable (electronic case report form) the following information was extracted: whether the patient had had a previous adenomectomy, baseline PSA blood level,

index lesion PI-RADS, index lesion location, Gleason score of the most prominent histological pattern and Gleason score of the second most prominent pattern, respectively the variables: "previousAdenomectomy", "baseline_psa_total", "index_lesion_pirads", "index_lesion_location", "gleason1" and "gleason2". The variable "ecrf" was removed.

9. The column "ISUP grade" was inferred from the two gleason scores in the following manner: GS $\leq$ 6 corresponds to ISUP = 1; GS = 3+4 corresponds to ISUP = 2; GS = 4+3 corresponds to ISUP = 3; GS = 8 corresponds to ISUP = 4; and GS $\geq$ 9 corresponds to ISUP = 5. Gleason scores lower than 1 were considered as absence of cancer.

10. A column "metastasis" was inferred from "M_Diagnosis" (M category of the TNM staging system) in the following manner: "cM0" corresponds to 0; "cM1a", "cM1b" and "cM1c" correspond to 1; and "MX" corresponds to missing data. The variable "M_Diagnosis" was removed.

11. Three levels of quality of life (column "QoL") were defined from the epic26 (Expanded Prostate Cancer Index Composite-26) score so that the 3 groups had the same absolute frequency: 0-70 corresponds to low quality of life; 71-84 corresponds to medium quality of life; and 85-100 corresponds to high quality of life.

12. Finally, all cells with blank spaces were converted to missing data.

The final dataset comprised 10728 rows and 33 columns. The latter are individually described in Table 5.2.

5.2.3 Included Clinical variables

The clinical variables included for each use case can be found in Table 5.3. For UC5, two contexts were considered: pre-surgery and post-surgery. In the latter, information reported during or immediately after the surgery was added.

5.2.4 Initial analysis

The dataset was analysed in terms of missing data and exploratory data analysis per clinical use case.

5.3 Results

5.3.1 Missing data

Figure 5.1 illustrates, for each use case, the missing data in each clinical variable. In UC2, the most impacted variable was baseline PSA (prostate-specific antigen) blood level. This is followed by ISUP grade, whose missing values reflect absent Gleason scores 1 and 2. In UC3, one patient with missing PSA is observed, however, in this instance, the metastatic status, which is also the

Table 5.2: Variables description in the processed dataset. M = mean; SD = standard deviation; RP = radical prostatectomy; RT = radiation therapy; AS = active surveillance; GS = gleason score; PZ = peripheral zone; TZ = transitional zone; CZ = central zone; AS = anterior stroma

Variable	**Description**	**Type**	**Possible values**
patient_id	Patient identifier	str	
use_case_form	Use cases the patient is suitable to join	str	
study_uid	Study identifier	str	
series_uid	Series identifier	str	
manufacturer	Series scanner manufacturer	str	PHILIPS, SIEMENS, GE MEDICAL SYSTEMS, TOSHIBA
manufacturer_model_name	Series scanner model name	str	
ERC	Endorectal coil was used in the MRI exam	bool	1, 0
provided_by	Clinical data provider	str	FPO, HACETTEPE, NCI, RADBOUDUMC, FCHAMPALIMAUD, IPC, QUIRONSALUD, UNIPI, RMH, HULAFE, GAONA, IDIBGI, JCC
country	Country of data provider	str	Italy, Turkey, Lithuania, Netherlands, Portugal, France, Spain, Greece, UK
prospective	Boolean identifier of prospective patients	bool	True, False
age_at_baseline	Patient age at MRI exam	int	14-91 (M = 64.89; SD = 7.87)
metastasis	M category of the TNM staging system	bool	1, 0
epic26	Expanded Prostate Cancer Index Composite-26 (EPIC26) score	int	0-100 (M = 72.62; SD = 21.30)
QoL	Quality of life: low, intermediate and high.	int	1, 2, 3
biochemicalRelapseFUp	Biochemical relapse observed at follow up.	bool	True, False
prostatectomy_nerve_sparing	Neurovascular bundles spared during RP	bool	True, False
extra_prostatic_extension	Extraprostatic extension observed at RP	str	Present, Absent, Not Assessed
perineural_invasion	Perineural invasion observed at RP	str	Present, Absent, Not Assessed
seminal_vesical_invasion	Seminal vesicle invasion observed at RP	str	Present, Absent, Not Assessed
resection_margin_status	Cancerous tissue remained after RP	str	Positive, Negative, Not Assessed
prostatectomy_method	Surgical method	str	RP perineal, RP retropubic, Laparoscopic, Robotic-assisted laparoscopic, Other
previousAdenomectomy	Patient underwent previous adenomectomy.	bool	True, False
baseline_psa_total	PSA blood level at baseline.	float	0.04 - 1968.56 (M = 11.03; SD = 36.79)
index_lesion_pirads	PI-RADS of the index lesion	int	1, 2, 3, 4, 5
gleason1	GS of the most prominent histological pattern	int	1, 2, 3, 4, 5
gleason2	GS of the second most prominent histological pattern	int	1, 2, 3, 4, 5
ISUP grade	International Society of Urological Pathology (ISUP) grade group	int	1, 2, 3, 4, 5
index_lesion_location_PZ	Index lesion located in the PZ	bool	1, 0
index_lesion_location_TZ	Index lesion located in the TZ	bool	1, 0
index_lesion_location_CZ	Index lesion located in the CZ	bool	1, 0
index_lesion_location_AS	Index lesion located in the AS	bool	1, 0
leftSurveillance	Whether a patient was removed from the active surveillance program.	bool	True, False
Treatment	Treatment undergone by the patient.	str	RP, RT, AS

Table 5.3: Clinical variables included for each use case. “num” indicates a numerical variable; “bool” indicates a binary variable; “cat” indicates a categorical variable.

Use Cases	2, 3 and 8	5 (post-surgery) and 7b	5 (pre-surgery) and 6
Clinical variables	age_at_baseline (num) baseline_psa_total (num) index_lesion_pirads (cat) lesion_location_PZ (bool) lesion_location_TZ (bool) lesion_location_CZ (bool) lesion_location_AS (bool)	age_at_baseline (num) baseline_psa_total (num) index_lesion_pirads (cat) lesion_location_PZ (bool) lesion_location_TZ (bool) lesion_location_CZ (bool) lesion_location_AS (bool) prostatectomy_method (cat) resection_margin_status (cat) extraprostatic_extension (cat) perineural_invasion (cat) seminal_vesicle_invasion (cat) gleason1 (num) gleason2 (num) ISUP grade (num) previous_adenomectomy (bool) prostatectomy_nerve_sparing (bool)	age_at_baseline (num) baseline_psa_total (num) index_lesion_pirads (cat) lesion_location_PZ (bool) lesion_location_TZ (bool) lesion_location_CZ (bool) lesion_location_AS (bool) gleason1 (num) gleason2 (num) ISUP grade (num)

target variable in this use case, is the largest contributor to the missing data. This is due to the high prevalence of "MX" cases in the original raw dataset. According to the TNM staging system, "MX" refers to an inability to evaluate the metastatic status, which is why it was treated as missing data. In UC5, we still see a few missing PSA values, but the variable with the highest number of missing values is "prostatectomy_nerve_sparing". In UCs 6 and 7 there are only a few missing vallues in the ISUP grade and baseline PSA, respectively. Finally, in UC8 missing data is only found in the target variable "leftSurveillance".

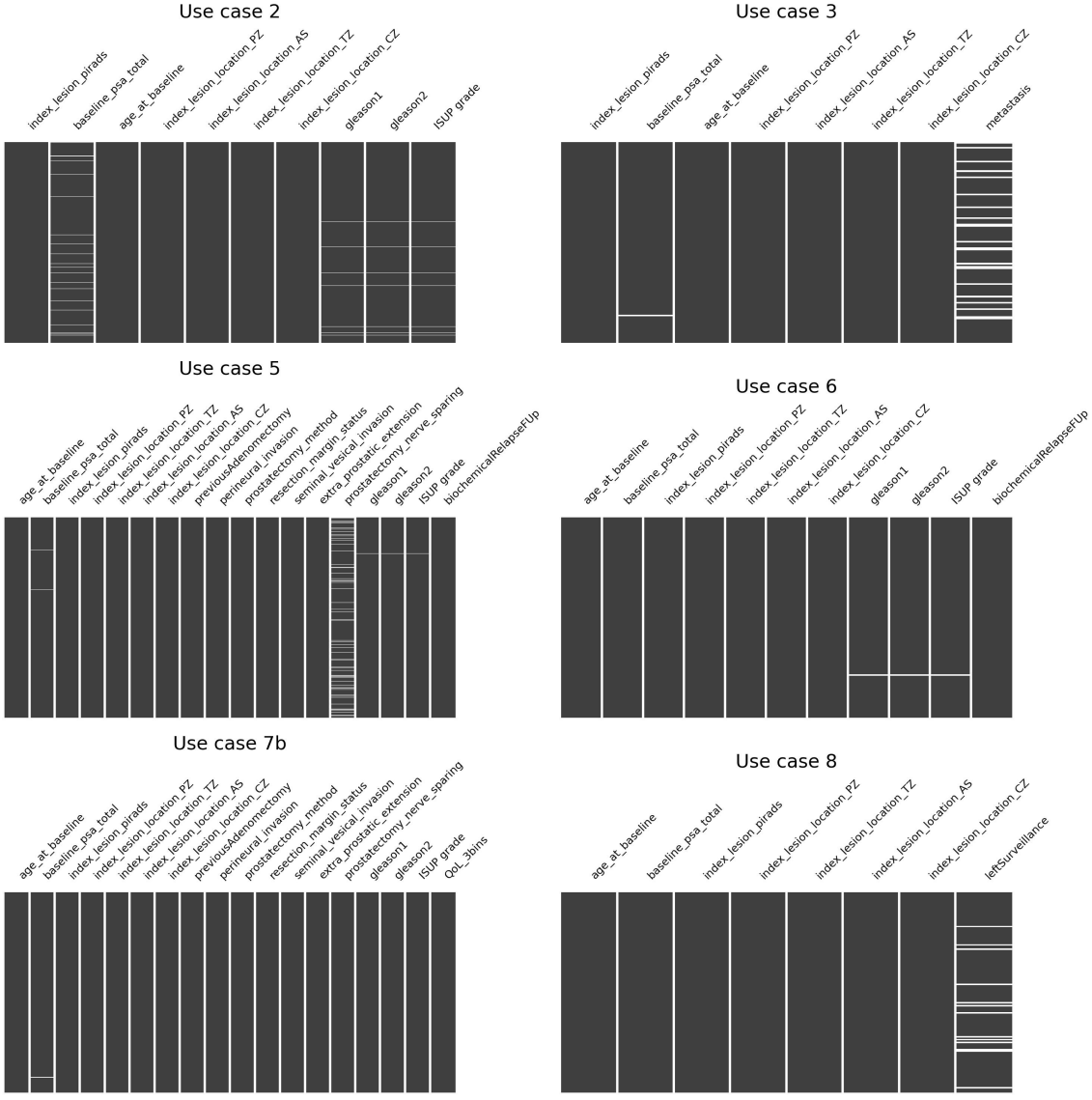


Figure 5.1: Missing values for each clinical variable included in each use case.

5.3.2 Exploratory Data Analysis

Figure 5.2 shows the distribution of each clinical variable used in UC2. There is no variable where the distinction between the different classes is clear, however we can see that the proportion of higher ISUP cases increases as the PIRADS increases.

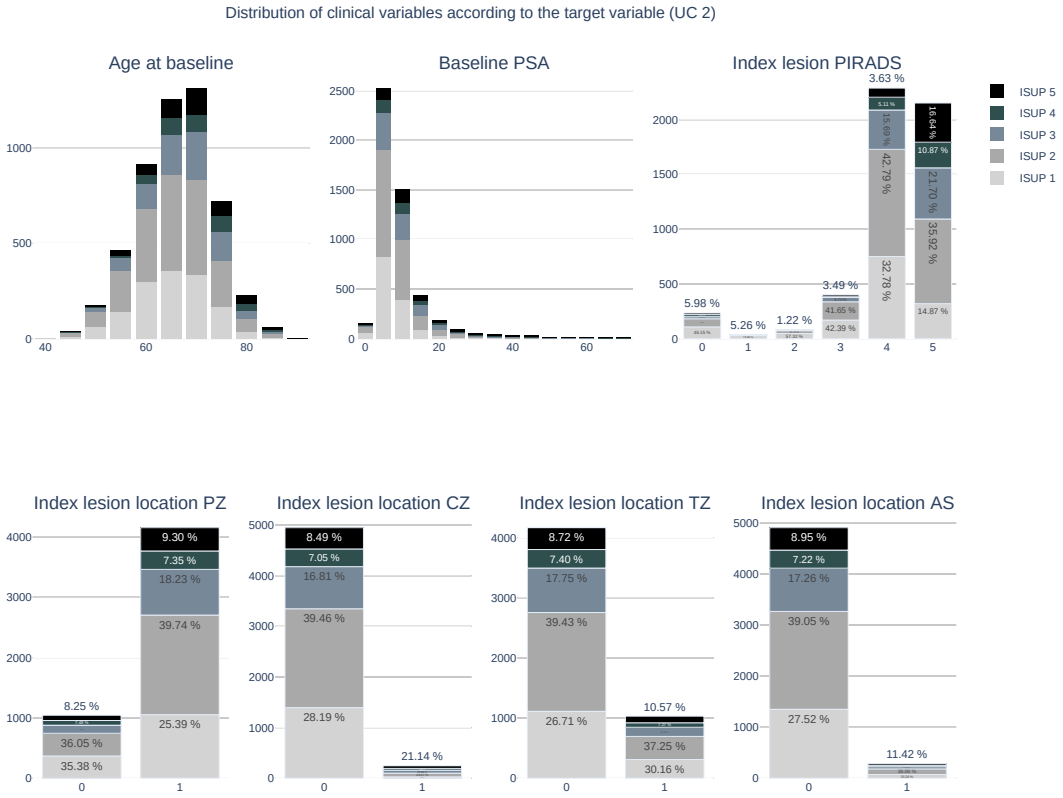


Figure 5.2: Distribution of clinical variables included in use case 2 according to the target variable.

Figure 5.3 shows the distribution of each clinical variable used in UC3. We can see that the proportion of non-metastatic cases is higher in the lower PSA values. Additionally, the proportion of metastatic cases is much higher for PIRADS 5 when compared to PIRADS 4.

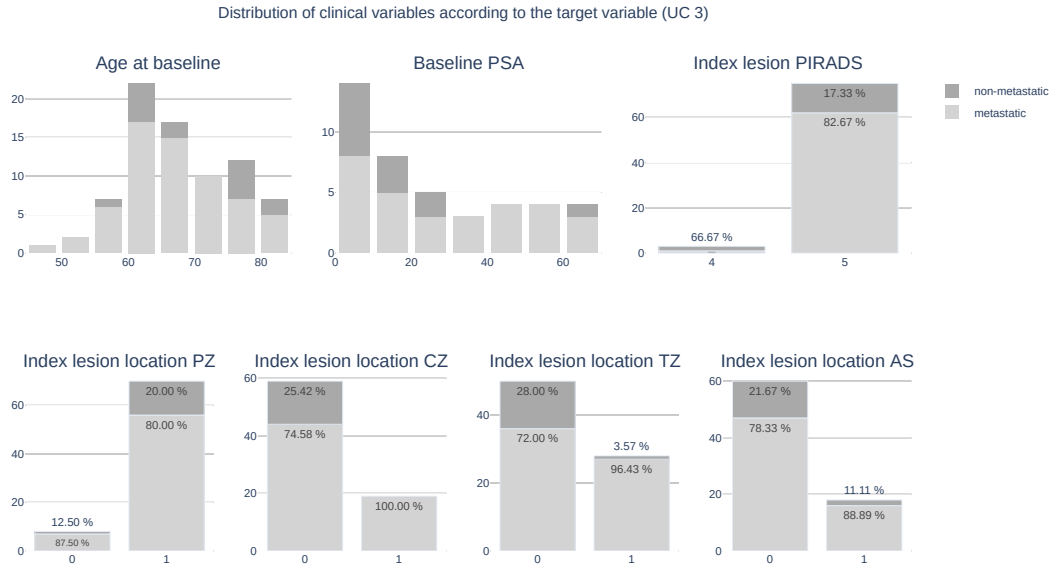


Figure 5.3: Distribution of clinical variables included in use case 3 according to the target variable.

Figure 5.4 shows the distribution of each clinical variable used in UC5. There is no variable where the distinction between the two classes is clear, however we can see that the proportion of biochemical recurrence cases increases as the PIRADS increases. This proportion is also higher for the central ISUP grades, namely ISUP 2 and 3, and, as expected, it is also higher when there is perineural invasion, seminal vesicle invasion, extraprostatic extension or when the resection margins are positive after prostatectomy. A higher proportion of recurrence was also found after laparoscopic prostatectomies (26.9%) compared with robotic-assisted (12.7%) or retro-pubic (3.3%).



Figure 5.4: Distribution of clinical variables included in use case 5 according to the target variable.

Figure 5.5 shows the distribution of each clinical variable used in UC6. There is no variable where the distinction between the two classes is clear, however we can see that the proportion of biochemical recurrence cases increases as the PIRADS increases. Similarly to UC5, the proportion of biochemical recurrence is also higher for the central ISUP grades, namely ISUP 2 and 3.

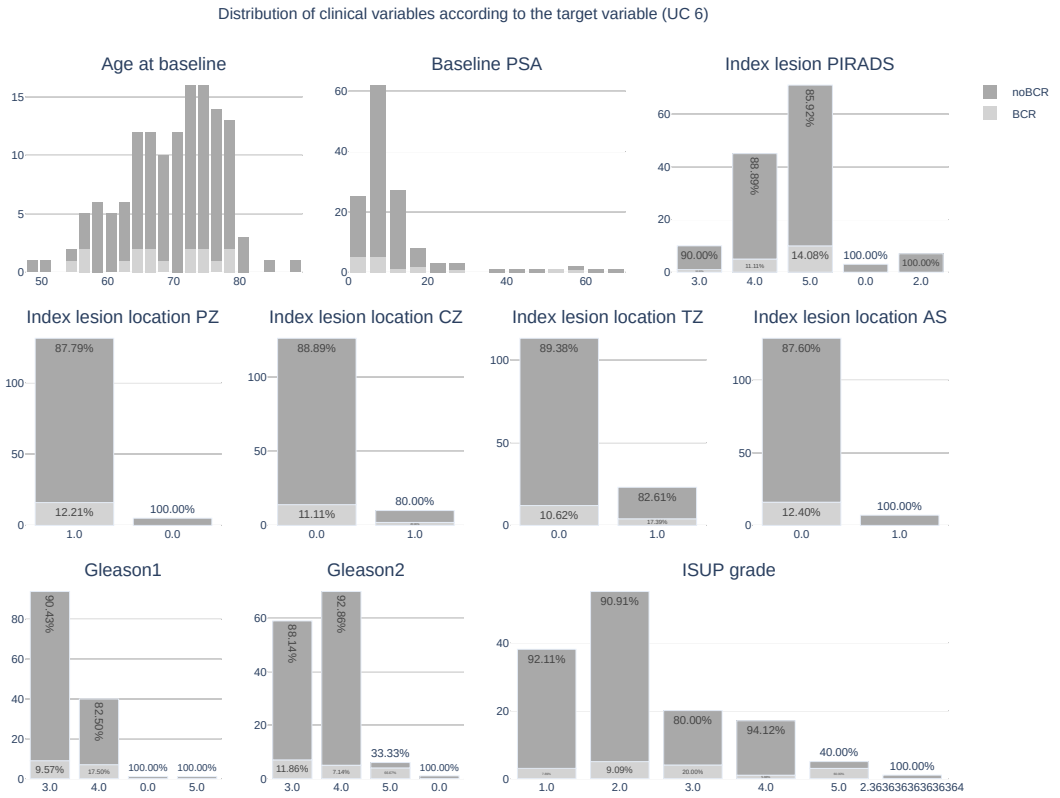


Figure 5.5: Distribution of clinical variables included in use case 6 according to the target variable.

Regarding UC7b (Figure 5.6), it is shown that, interestingly, the higher the PIRADS score, the lower the proportion of low quality of life cases. While, the lower the ISUP grade, the highest the proportion of good quality of life.

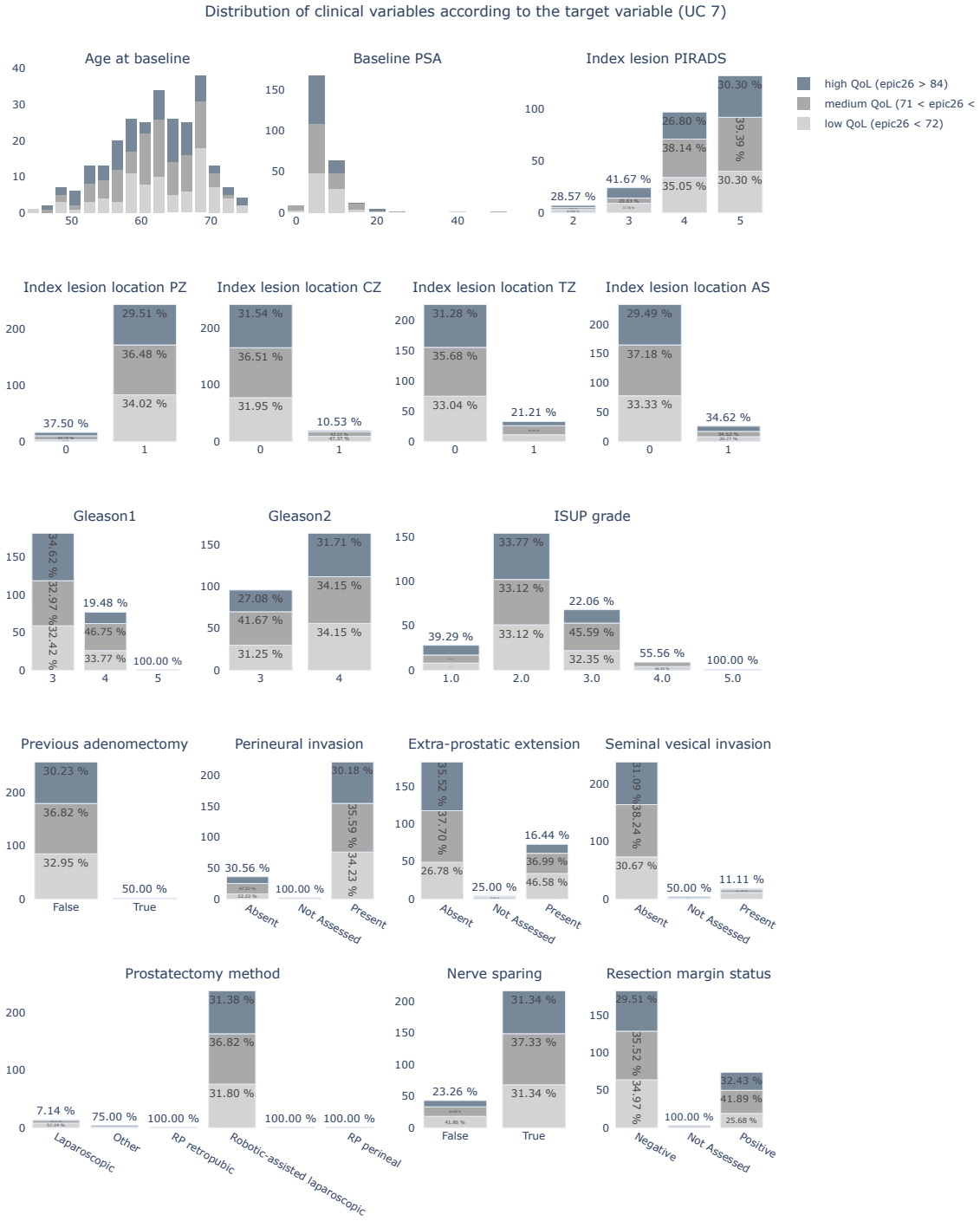


Figure 5.6: Distribution of clinical variables included in use case 7b according to the target variable.

Figure 5.7 shows the distribution of each clinical variable used in UC8. No patient appears to have the index lesion located in the CZ, so this variable is not informative. Regarding the remaining variables, there is none where the distinction between the two classes is clear.

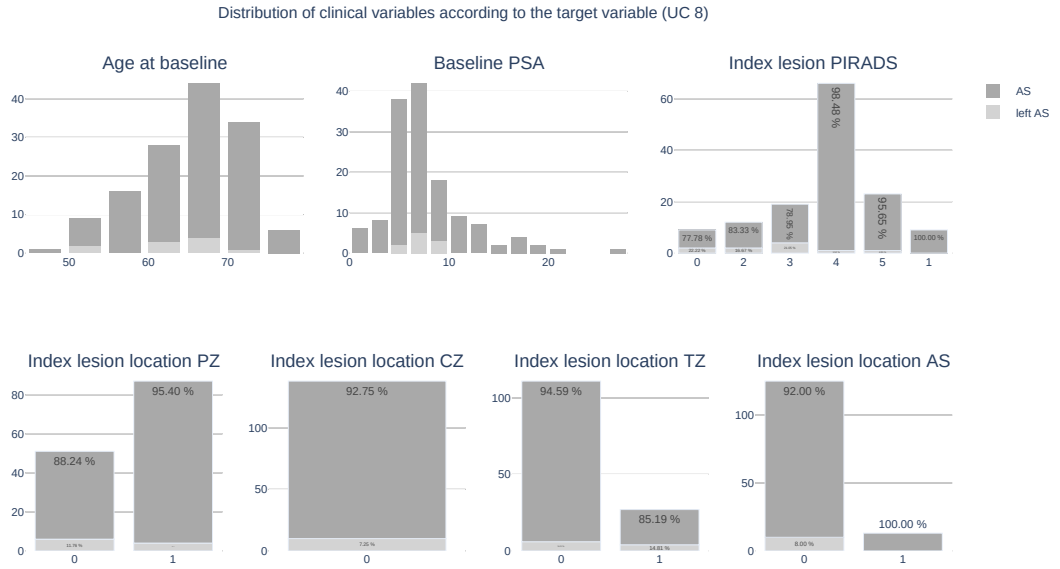


Figure 5.7: Distribution of clinical variables included in use case 8 according to the target variable.

5.4 Discussion

The exploratory data analysis performed was highly informative for the remaining analysis presented in this dissertation.

It was found that the dataset sizes differ greatly across use cases, from over 10000 in UC2 to less than 100 in UC3. Additionally, most use cases are highly imbalanced, with minority classes ranging from 7% to 33% of each use case's dataset. For these reasons, techniques and algorithms will be adapted to suit each use case's characteristics.

Furthermore, there were several variables that appeared problematic in terms of missing values. The items bellow describe how those missing values were dealt with:

- If the missing value belonged to the target variable of the use case, the patient was removed.
- If the missing value did not belong to the target variable of the use case:
  - if the variable is numeric, such as PSA, missing data was considered to be "missing at random" and was imputed with a k-nearest neighbours approach.
  - if the variable is categorical, missing data was set to "Not Assessed".

Chapter 6

Generation of Segmentation Masks and Quality Assessment

This chapter describes the technique developed and adopted for the automatic generation of segmentation masks for the three MRI sequences: T2W, DWI and ADC. The work of this chapter was submitted for publication (21).

6.1 Introduction

A radiomics analysis and, in particular, the extraction of radiomic features is by nature dependent on the existence of segmentation masks, outlining the specific volume of interest (VOI) in a particular study. As has been established in Chapter 4, these segmentations suffer from inter- and intra-reader variability, which inevitably affects the radiomic features and the machine learning algorithm trained to model them. This variability can lead to inconsistent and conflicting results, making it difficult to draw definitive conclusions from the radiomics analysis.

Automated segmentation algorithms are able to overcome this issue to some extent, by being human-independent in the generation process (117; 201). By producing consistent segmentation masks and effectively analysing big datasets, these algorithms lower the possibility of human error and inter-observer variability. The selection of automated techniques for the generation of segmentations is justified by this, as well as the large size of ProstateNet (covered in Chapter 5), which makes it near impossible to manually segment.

6.2 Methodology

6.2.1 Segmentation of T2W volumes

Automatic segmentation of the whole prostate gland was performed on T2W sequences using a previously developed segmentation model. The full details of this model are shown in (182).

The generated masks were post-processed in two stages. Firstly, the largest object was selected. An object was defined as a group of connected voxels. Here, it was assumed that the

largest object would have the highest probability of covering the actual gland. Secondly, so as to smooth mask borders, a Delaunay triangulation was calculated on the convex hull of the selected object.

6.2.2 Segmentation of DWI and ADC volumes

T2W sequences (considered as moving images) were co-registered to the DWI sequences' space (considered as fixed images), and the calculated transformation matrix was then applied to the segmentation mask generated automatically for T2W sequences (Figure 6.1). The co-registration algorithm was a 3-resolution pyramid of rigid registrations. The co-registration parameters file can be found in Appendix B (Table B.1). For wide field-of-view DWI sequences, a center crop was applied to facilitate the co-registration.

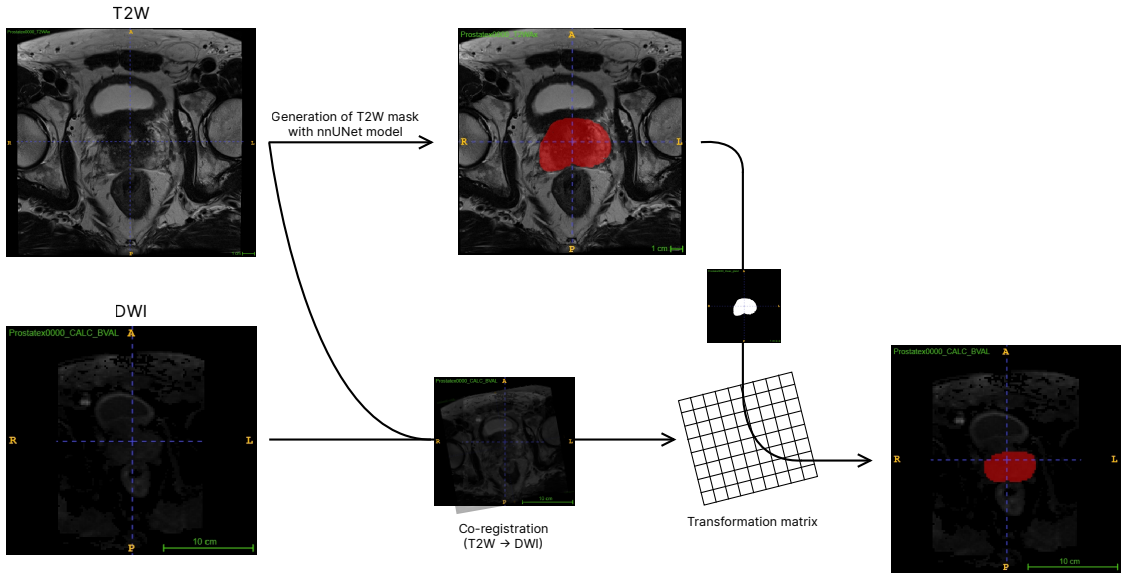


Figure 6.1: Pipeline for the automatic generation of segmentation masks.

6.2.3 Quality assessment

A radiologist was asked to assess the segmentation quality of 125 T2W volumes and respective DWI. A three level grade was assigned in the following manner: 1 was given to good/decent segmentations; 2 to masks where minor corrections were needed; and 3 where major corrections were needed (less than 50% of the gland covered). The radiologist also provided notes/observations about the images and masks that allowed the construction of a dataframe with 17 variables: small_FOV_T2, small_FOV_DWI, big_FOV_DWI, poor_quality_T2, poor_quality_DWI, erc, previous_resection, missed_apex, missed_posterior, mis-sed_anterior, missed_base/superior, missed_PZ, included_seminal_vesicle, lesion_ difficulties, mismatch_apex/base, mismatch_antero /posterior and mismatch_la-tero/lateral, where FOV corresponds to field of view, erc corresponds to presence of endorectal coil, lesion_difficulties indicates whenever a lesion made it more difficult

to segment the gland and the several mismatches correspond to masks dislocated from the actual gland due to co-registration errors. This information was used to train logistic regression models to predict mask quality, and the trained coefficients were analyzed for insights into the causes of low quality.

6.3 Results

6.3.1 Quality of T2W segmentations

The whole gland (WG) segmentations generated for T2W volumes proved to be of high quality (93% with quality level 1), with only 7% (9/125) of masks needing small corrections (quality level 2) and no masks requiring major corrections (quality level 3). Table 6.1 shows that the major contributors to the low quality were the presence of lesions that made the gland margins more difficult to identify or the presence of endorectal coil. This is followed by the absence of parts of the gland (previous resection) and the low quality of the DWI sequence. Additionally, the evidence of the slight drop in quality seems to be in missed apexes and wrongful inclusion of seminal vesicles.

Table 6.1: Coefficients of logistic regression trained to predict low quality of the whole-gland segmentation of T2W volumes.

Features	Coefficient	exp(coefficient)
lesion_difficulties	1.487972	4.428107
erc	0.753808	2.125077
missed_apex	0.703056	2.019917
included_seminal_vesicle	0.589769	1.803572
previous_resection	0.478329	1.613376
poor_quality_DWI	0.374318	1.453999
poor_quality_T2	0	1
mismatch_apex/base	-0.0748	0.927929
missed_PZ	-0.08214	0.921141
small_FOV_T2	-0.09743	0.907166
mismatch_antero/posterior	-0.22899	0.795338
mismatch_latero/lateral	-0.27934	0.756281
missed_anterior	-0.33524	0.715168
big_FOV_DWI	-0.41192	0.662381
small_FOV_DWI	-0.46918	0.625514
missed_base/superior	-0.54922	0.577401
missed_posterior	-0.71678	0.488323

6.3.2 Quality of DWI and ADC segmentations

In contrast, the whole gland segmentations generated for DWI volumes did not show as high quality. Here, 37% (46/125) of masks had quality level 2 and 9% (11/125) showed quality level 3, requiring major corrections. Table 6.2 shows that the major contributors to the low quality were the absence of parts of the gland (previous resection), the low quality of the DWI sequence and the small field of view of the T2 sequence, which affect the co-registration. Regarding the middle

quality (exp(coef_2)), it seems to be mostly due to several mismatches, which indicate imperfect coregistrations.

Table 6.2: Coefficients of logistic regression trained to predict low quality of whole-gland segmentation of DWI volumes.

Features	**coef_1**	**coef_2**	**coef_3**	**exp(coef_1)**	**exp(coef_2)**	**exp(coef_3)**
previous_resection	-0.64316	-0.79734	1.44049	0.52563	0.45053	4.22279
poor_quality_DWI	-0.12803	-0.62484	0.75288	0.87982	0.53535	2.12310
small_FOV_T2	0.15139	-0.39528	0.24389	1.16345	0.67349	1.27621
small_FOV_DWI	0.37782	-0.40817	0.03035	1.45910	0.66487	1.03081
big_FOV_DWI	0.24017	-0.26686	0.02669	1.27146	0.76578	1.02705
poor_quality_T2	0	0	0	1	1	1
lesion_difficulties	-0.17899	0.18172	-0.00273	0.83612	1.19928	0.99727
missed_apex	0.39403	-0.37801	-0.01602	1.48294	0.68523	0.98411
missed_PZ	-0.23069	0.25777	-0.02708	0.79399	1.29403	0.97329
missed_base/superior	-0.31075	0.36941	-0.05866	0.73289	1.44689	0.94302
mismatch_apex/base	-0.13709	0.23201	-0.09492	0.87189	1.26113	0.90944
included_seminal_vesicle	-0.09962	0.20601	-0.10638	0.90518	1.22876	0.89908
missed_anterior	0.96857	-0.78994	-0.17863	2.63418	0.45387	0.83642
mismatch_antero/posterior	-0.91905	1.09925	-0.18019	0.39889	3.00189	0.83511
mismatch_latero/lateral	-0.44828	0.70505	-0.25676	0.63873	2.02394	0.77355
erc	-0.32120	0.80942	-0.48823	0.72528	2.24661	0.61371
missed_posterior	-0.01052	0.51741	-0.50689	0.98953	1.67767	0.60237

6.4 Discussion

Radiomics analysis require segmentations of the target volumes of interest. These can be performed manually, by an expert radiologist; in an automated fashion, such as by deep learning models; or semi-automatically, where the masks are firstly automatically generated and then corrected where needed by an expert radiologist. Due to the large size of the dataset, the whole gland segmentation masks were generated automatically for T2W sequences by a previously trained DL model. For DWI, however, there was no model available so an alternative approach was selected: to coregister the T2W and DWI volumes and apply the transformation matrix to the automatically generated mask. The segmentation quality assessment revealed the T2W segmentations to be of extremely high quality, however, this was not the case for DWI. Further analysis revealed that the lower quality was mostly due to "mismatches", defined by the radiologist as an "apparent translation of the mask from the gland location", which revealed some imperfections in the coregistration stage of the pipeline.

Chapter 7

Whole-gland radiomics models

In this Chapter, the ProstateNet dataset (described in Chapter 5) and the whole-gland segmentation masks (Chapter 6) were used to extract radiomic features and develop classification models pertaining to six separate use cases, defined by the ProCAncer-I project: UC2, UC3, UC5, UC6, UC7b, and UC8. The work of this Chapter resulted in a manuscript submission (21) and two conference acceptances (19; 20).

7.1 Introduction

This chapter presents the application of radiomics analysis in the development of predictive models for prostate cancer. Using the whole-gland segmentation masks generated in Chapter 6, radiomic features were extracted from bpMRI sequences derived from the ProstateNet dataset, as detailed in Chapter 5. Radiomic features, which quantify tissue heterogeneity and microarchitecture, offer valuable insights into tumor characteristics and form the basis of predictive modelling efforts.

This chapter explores six pivotal clinical use cases (UCs) within the prostate cancer continuum. These include predicting disease aggressiveness at diagnosis (UC2), forecasting the development of metastasis within six months (UC3), and estimating the risk of biochemical recurrence following prostatectomy or radiation therapy (UC5 and UC6, respectively). Additionally, the models address quality of life outcomes post-prostatectomy (UC7b) and identify early withdrawal risk from active surveillance programs (UC8). Each of these use cases represents a distinct clinical challenge, spanning from initial diagnosis to long-term disease management.

By developing machine learning models for these specific use cases, we aim to enhance clinical decision-making and improve personalized treatment strategies for prostate cancer patients. The results from this chapter contribute to the broader goal of integrating radiomic features into routine clinical workflows for precision oncology. This work culminated in a manuscript submission (21), reflecting the potential clinical impact of the models developed.

7.2 Methodology

7.2.1 Radiomic features extraction

Bias field correction was performed on T2W sequences using the N4 Bias Field Correction algorithm (177) and the Python package Simple ITK (version 2.0.0) (178). First, each image's x-, y- and z-spacing were assessed for discrepancies. Since x- and y-spacings differed from z-spacing, feature extraction was later performed in 2D. Additionally, images' x- and y-spacings differed within and between patients, so T2W sequences were resampled to the 95th quantile value of 0.6875, and DWI and ADC were resampled to the 95th quantile value of 2.0. Image intensities were also normalized. The bin width was selected for each image filter to produce discretized images with between 30 and 130 bins. The full description of extraction parameters for each modality can be found in Table 7.1.

Table 7.1: Whole-gland radiomics extraction parameters.

T2W extraction parameters	**DWI extraction parameters**	**ADC extraction parameters**
imageType: Original: binWidth: 5 Wavelet: binWidth: 3 Square: binWidth: 3 SquareRoot: binWidth: 8 Logarithm: binWidth: 16 Exponential: binWidth: 0.5 Gradient: binWidth: 5 LBP2D: binWidth: 0.1 LoG: 'sigma' : [1.0, 3.0]	imageType: Original: binWidth: 12 Wavelet: binWidth: 8 Square: binWidth: 8 SquareRoot: binWidth: 16 Logarithm: binWidth: 25 Exponential: binWidth: 3 Gradient: binWidth: 4 LBP2D: binWidth: 0.1 LoG: 'sigma' : [1.0, 3.0]	imageType: Original: binWidth: 5 Wavelet: binWidth: 4 Square: binWidth: 3 SquareRoot: binWidth: 8 Logarithm: binWidth: 12 Exponential: binWidth: 1 Gradient: binWidth: 3 LBP2D: binWidth: 0.1 LoG: 'sigma' : [1.0, 3.0]
featureClass: firstorder: glcm: glrlm: glszm: gldm: ngtdm: shape:	featureClass: firstorder: glcm: glrlm: glszm: gldm: ngtdm: shape:	featureClass: firstorder: glcm: glrlm: glszm: gldm: ngtdm: shape:
setting: binWidth: 5 normalize: True normalizeScale: 100 force2D: True voxelArrayShift: 300 resampledPixelSpacing: [0.6875, 0.6875, 0] geometryTolerance: 0.00001	setting: binWidth: 5 normalize: True normalizeScale: 100 force2D: True voxelArrayShift: 300 resampledPixelSpacing: [2, 2, 0] geometryTolerance: 0.00001	setting: binWidth: 5 normalize: True normalizeScale: 100 force2D: True voxelArrayShift: 300 resampledPixelSpacing: [2, 2, 0] geometryTolerance: 0.00001

Radiomic features were extracted from the whole gland segmentation using the Pyradiomics package (version 3.0) (202) in Python (version 3.7.9) (203). All the pre-processing steps were performed as parameters of the extractor function, except for the bias field correction, which was

performed prior to the extraction. All image filters and feature classes were enabled, resulting in a total of 1223 features per sequence. The mathematical expressions and semantic meanings of the extracted features can be found at https://pyradiomics.readthedocs.io /en/latest/.

7.2.2 Deep features

To generate deep features for each instance, we used the bottleneck of a UNet model pre-trained on segmenting the whole prostate gland using T2W volumes. To calculate a segmentation prediction, the UNet model first encodes the image into a low resolution volume with high semantic information (320 features in our case) and uses this information to obtain a segmentation map for a given object (whole prostate gland in our case). We encode each T2W volume and extract the maximum value of each feature, obtaining a 320-sized vector characterizing each image.

7.2.3 Clinical features

The clinical variables included for each use case can be found in Table 7.2. Missing numerical values were imputed with a KNNImputer. Missing categorical values in the variables perineural invasion, extra prostatic extension, seminal vesical invasion and resection margin status were set to "Not Assessed", while the remaining missing categorical values were imputed to the most frequent category.

For UC5, two contexts were considered: pre-surgery and post-surgery. For the former, all information reported during or immediately after the surgery was removed, namely the variables prostatectomy method, resection margin status, extraprostatic extension, perineural invasion, seminal vesicle invasion, previous adenectomy and prostatectomy nerve sparing were excluded.

Table 7.2: Clinical variables included for each use case. "num" indicates a numerical variable; "bool" indicates a binary variable; "cat" indicates a categorical variable.

Use Cases	2, 3 and 8	5 (post-surgery) and 7b	5 (pre-surgery) and 6
Clinical variables	age_at_baseline (num) baseline_psa_total (num) index_lesion_pirads (cat) lesion_location_PZ (bool) lesion_location_TZ (bool) lesion_location_CZ (bool) lesion_location_AS (bool)	age_at_baseline (num) baseline_psa_total (num) index_lesion_pirads (cat) lesion_location_PZ (bool) lesion_location_TZ (bool) lesion_location_CZ (bool) lesion_location_AS (bool) prostatectomy_method (cat) resection_margin_status (cat) extraprostatic_extension (cat) perineural_invasion (cat) seminal_vesicle_invasion (cat) gleason1 (num) gleason2 (num) ISUP grade (num) previous_adenomectomy (bool) prostatectomy_nerve_sparing (bool)	age_at_baseline (num) baseline_psa_total (num) index_lesion_pirads (cat) lesion_location_PZ (bool) lesion_location_TZ (bool) lesion_location_CZ (bool) lesion_location_AS (bool) gleason1 (num) gleason2 (num) ISUP grade (num)

7.2.4 Dataset construction

The total dataset is composed of 10716 patients (8577 retrospective, 315 retrospective for validation, and 1824 prospective). The train/test split was performed for the larger use cases at patient level with the Python scikit-learn package (version 0.23.2) (204). The hold-out test sets consisted of 200 randomly selected patients for UC2 and 100 for UC5. The split was stratified so that both train and test sets have the same label distribution. The train and test sets label distribution can be found in tables 7.3 and 7.4, for binary and multiclass tasks, respectively. For the smaller use cases, namely 3, 6, 7b and 8, only the cross-validation performance is reported.

Table 7.3: Label distribution in the train and test sets for each binary classification problem. RFV - retrospective for validation.

Use Cases	**Target (binary)**	**Train Set**		**Test Set**		**RFV**		**Prospective**	
		0	1	0	1	0	1	0	1
2	ISUP 01 VS 2345	4725	3652	108	92	220	107	1145	679
	ISUP 012 VS 345	6703	1674	148	52	231	96	1515	309
3	no metastasis in six months VS metastasis developed	15	62	-	-	-	-	-	-
5	no biochemical recurrence after RP at follow-up VS biochemical recurrence	569	94	88	12	9	-	-	-
6	no biochemical recurrence after RT at follow-up VS biochemical recurrence	120	14	-	-	-	-	-	-
8	stayed in active surveillance VS left active surveillance	126	10	-	-	-	-	1	-

Table 7.4: Label distribution in the train and test sets for the multiclass classification problem. RFV - retrospective for validation.

Use Cases	**Target (multiclass)**	**Train Set**			**Test Set**			**RFV**			**Prospective**		
		0	1	2	0	1	2	0	1	2	0	1	2
7b	EPIC26 [0, 71] vs]71, 84] vs]84, 100]	75	71	62	20	15	14	-	-	-	-	-	

Different data subsets were tested for their training ability. Pure radiomics datasets were appended clinical and/or deep features and their performance was compared (radclin - radiomics + clinical variables; raddeep - radiomics + deep features; hybrid - radiomics + clinical + deep features). Training with individual scanner manufacturers was compared to training with all data combined. The exclusion of patients where an endorectal coil had been used was also tested. And, finally, we compared training with the full MRI sequence set to training with each sequence independently. This resulted in 96 training combinations. The discriminated data sizes of the training set are shown in Tables 7.5 - 7.10 for each use case. For UC7b, there were no endorectal coil cases, so this setting was removed.

Table 7.5: Train set sizes for UC2.

UC 2 - Train Sets		sequence			
		T2W	DWI	ADC	T2W& DWI&ADC
scanner	All	8377	7905	7781	7778
	All_noERC	7496	7038	6933	6932
	SIEMENS	3158	3091	3071	3071
	PHILIPS	2965	2882	2858	2857
	GE	2254	1932	1852	1850
	GE_noERC	1373	1065	1004	1004

Table 7.6: Train set sizes for UC3.

UC 3 - Train Sets		sequence			
		T2W	DWI	ADC	T2W& DWI&ADC
scanner	All	77	67	67	67
	All_noERC	73	63	63	63
	SIEMENS	24	24	24	24
	PHILIPS	32	31	31	31
	GE	21	12	12	12
	GE_noERC	17	8	8	8

Table 7.7: Train set sizes for UC5.

UC 5 - Train Sets		sequence			
		T2W	DWI	ADC	T2W& DWI&ADC
scanner	All	663	632	612	611
	All_noERC	631	599	581	581
	SIEMENS	160	153	153	153
	PHILIPS	410	405	401	401
	GE	93	74	58	57
	GE_noERC	61	41	27	27

Table 7.8: Train set sizes for UC6.

UC 6 - Train Sets		sequence			
		T2W	DWI	ADC	T2W& DWI&ADC
scanner	All	134	133	131	131
	All_noERC	68	67	65	65
	SIEMENS	4	4	3	3
	PHILIPS	60	60	60	60
	GE	70	69	68	68
	GE_noERC	4	3	2	2

Table 7.9: Train set sizes for UC7.

UC 7b - Train Sets		sequence			
		T2W	DWI	ADC	T2W& DWI&ADC
scanner	All	257	254	251	251
	All_noERC	257	254	251	251
	SIEMENS	88	55	55	55
	PHILIPS	197	197	194	194
	GE	2	2	2	2
	GE_noERC	2	2	2	2

Table 7.10: Train set sizes for UC8.

UC 8 - Train Sets		sequence			
		T2W	DWI	ADC	T2W& DWI&ADC
scanner	All	136	131	128	128
	All_noERC	131	127	124	124
	SIEMENS	81	81	81	81
	PHILIPS	13	11	11	11
	GE	42	39	36	36
	GE_noERC	37	35	32	32

7.2.5 Model Development

The preprocessing pipeline described in this section was fitted on the train set and applied on the train and test sets. Features were scaled to have zero mean and standard deviation equal to 1 (Python package scikit-learn version 1.0.2). Features with low variance were identified and excluded. Here, a threshold of 0.01 was considered. Finally, feature correlation was assessed. Feature pairs were considered correlated if their Spearman correlation was higher than 0.8. Out of the two, the feature with the highest average correlation across all features was eliminated.

For models using radiomics or radiomics and deep features, a light gradient boosting machine (LGBM) (205) was trained, while for radclin or hybrid models, which may include categorical data, the CatBoost (206) algorithm was used. Regarding the smaller UCs (UC 3, 6, 7b and 8), a Stochastic gradient descent algorithm (SGD) was selected, as less complexity is prefered. Hyper-parameter tuning was performed for each algorithm and each parameter combination (Appendix C) was evaluated through cross-validation (5 folds for UCs 2, 5 and 7b; 3 folds for UCs 3, 6 and

8). For UCs 3, 6, 7b and 8 a random search approach was selected, as less data is available so a less biased optimization is preferred, while for UCs 2 and 5 tuning was performed with an exhaustive grid search. The overall pipeline can be found in Figure 7.1 and the hyperparameter space used can be found in Appendix B (Table C.1).

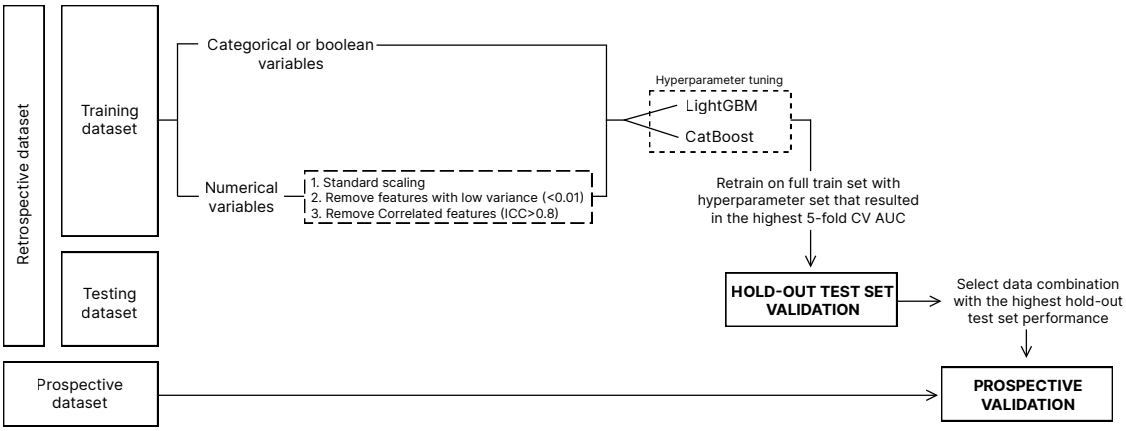


Figure 7.1: Radiomics model development pipeline.

The model names follow the following nomenclature: <*data type*>_<*use case*>_<*sequences*>_<*scanner**>_<*exclusion of ERC cases**>_<*algorithm*>. Where * represents optional fields: the absence of the *scanner* field means all scanner manufacturers were included; the absence of the *exclusion of ERC cases* field means that all cases, taken with or without ERC, were included.

For all models, the ROC curve was analyzed and the probability decision threshold that resulted in the highest youden index was selected for the remaining analysis.

All final models were analyzed in terms of model explainability. A SHapley Additive exPlanations (SHAP) analysis (Python package shap version 0.41.0) (207) was used to identify the most relevant variables for the prediction in the hold-out test set. The 20 most relevant variables for the output of each model were displayed. Each dot in the graph represents a feature's SHAP value for one observation in the hold- out test set. The SHAP value's position on the x-axis expresses whether it is associated with a positive or negative prediction. The red color indicates higher values of a feature and the blue color means lower value.

Model fairness to different population subcohorts was also analysed. However, this was only performed for UC2, as it was considered that the remaining use cases did not have enough data for subcohorts to be representative of the general population. This analysis was performed with the fairlearn python package. ROC-AUC, f2-score, precision and recall are reported for each subgroup, as well as subgroup size on the train and test sets and test set label distribution. For subgroups where only one target label is present the ROC-AUC metric is replaced with Accuracy.

7.3 Results

7.3.1 Use Case 2 - ISUP {0,1} vs {2,3,4,5}

7.3.1.1 Model Selection

Figure 7.2 shows the cross-validation and hold-out test set ROC-AUC model performance for the 96 models trained for UC2 (ISUP 0 & 1 vs 2, 3, 4 & 5). At first glance, we can see that the inclusion of clinical variables (radclin and hybrid) is beneficial for the model performance, both in the cross-validation and hold-out test set. However, the inclusion of deep features or the exclusion of endorectal coil patients does not meaningfully impact model performance or generalizability. Regarding scanner manufacturer, training with only GE data results in the most overfitted classifiers, whereas the least overfitting is seen when models are trained with all data. Overall, the highest hold-out test set performance reached was 0.91 AUC with a model trained with radclin data extracted from the three volumes, all scanners and excluding endorectal coil cases.

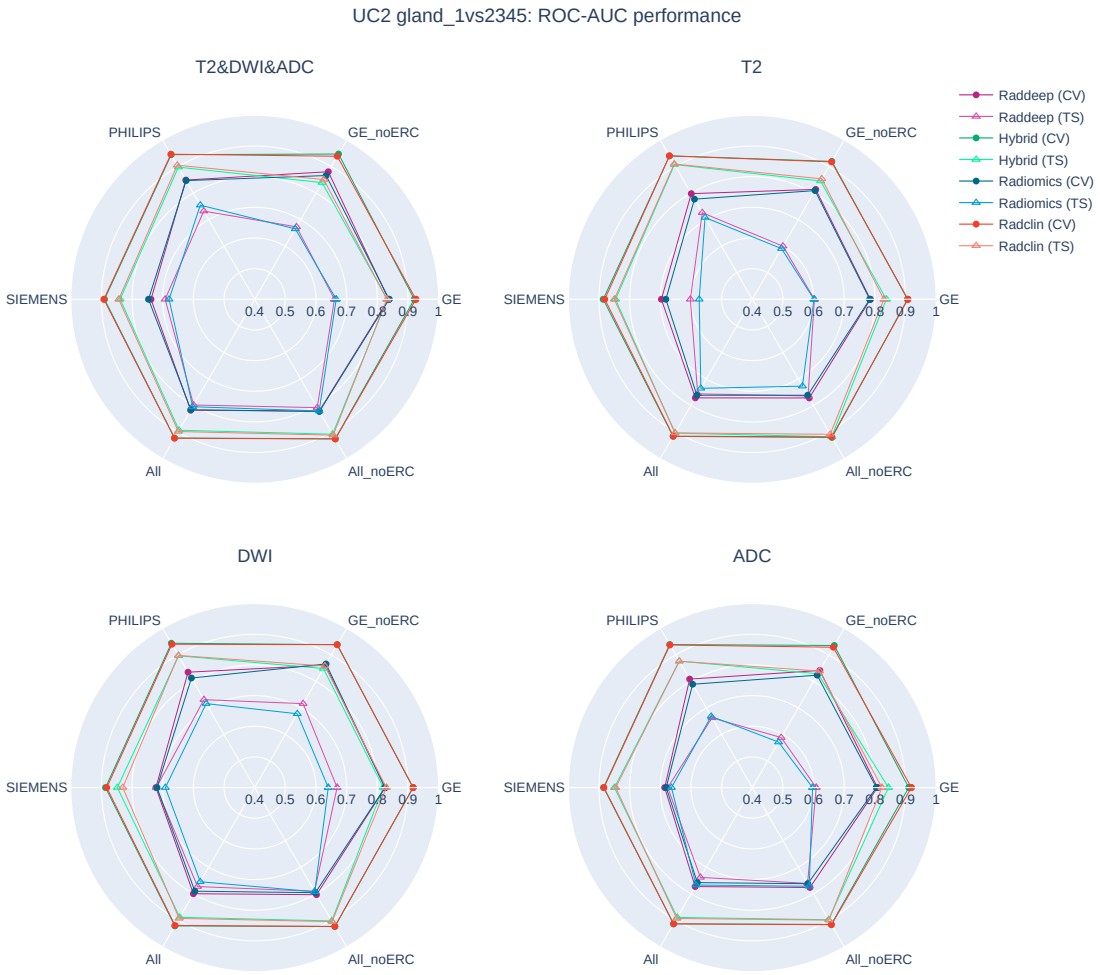


Figure 7.2: Cross-validation and hold-out test set ROC-AUC model performance of 96 models trained to predict disease aggressiveness in UC2 (ISUP 0, 1 vs 2, 3, 4 & 5). The observations are color coded according to the type of data used in training: radiomics in blue, radclin in orange, raddeep in pink and hybrid in green.

7.3.1.2 Model Validation

Given the results of the Section 7.3.1.1, the model radclin_T2W&DWI&ADC_noERC_CatBoost was further analysed. Upon inspection of the ROC curve for the hold-out test set (Figure 7.3), we decided the appropriate probability decision threshold would be 0.5806, above which the model gives an output of clinical significance. The model's performance at this probability decision threshold is shown in Table 7.11 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Figure 7.4.

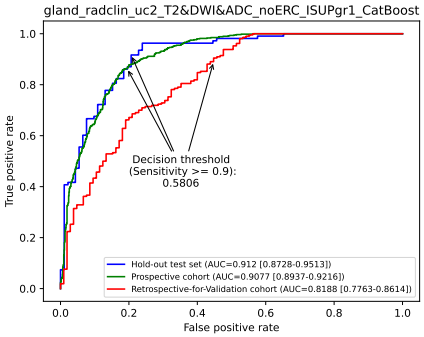


Figure 7.3: Receiver operator characteristics curve of the radclin_uc2_T2&DWI&ADC_no ERC_CatBoost model.

Table 7.11: Multi-metric performance of the radclin_uc2_T2&DWI&ADC_noERC_CatBoost model at 0.5806 probability threshold.

	Hold-out test set	Retrospective for Validation	Prospective cohort
AUC	0.9120 [0.8728-0.9513]	0.8188 [0.7763-0.8614]	0.9077 [0.8937-0.9216]
Sensitivity/Recall/TPR	0.9167 [0.8784-0.9550]	0.8905 [0.8560-0.9250]	0.8629 [0.8463-0.8795]
Specificity/TNR	0.7935 [0.7374-0.8496]	0.5524 [0.4975-0.6073]	0.8051 [0.7860-0.8242]
Precision/PPV	0.8390 [0.7880-0.8899]	0.7991 [0.7549-0.8434]	0.7157 [0.6939-0.7374]
F1	0.8761 [0.8304-0.9218]	0.8423 [0.8021-0.8826]	0.7824 [0.7625-0.8023]
F2	0.9000 [0.8584-0.9416]	0.8706 [0.8335-0.9076]	0.8288 [0.8106-0.8470]
CohensKappa	0.7159 [0.6534-0.9218]	0.4697 [0.4146-0.8826]	0.6396 [0.6165-0.8023]

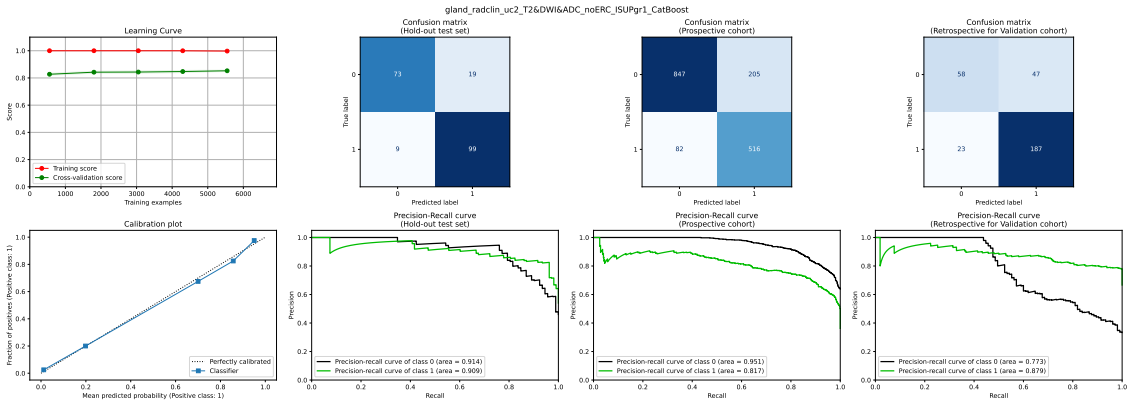


Figure 7.4: Analysis of the radclin_uc2_T2&DWI&ADC_noERC_CatBoost model in terms of confusion matrix, learning curve, precision-recall curve and calibration plot.

7.3.1.3 Explainability analysis

Figure 7.5 shows the SHAP analysis performed on the radclin_uc2_T2&DWI&ADC_noERC_CatBoost model. The plot shows the features ordered by their impact on model output. The model is relying heavily on the clinical variables PI-RADS (Figure 7.6), index lesion in the PZ, PSA blood level and patient age, all positively associated with the aggressive output. The most impactful radiomic feature is related with gland shape, namely LeastAxisLength, which is inversely associated with the aggressive output, indicating that larger glands are typically associated with non-aggressive disease.

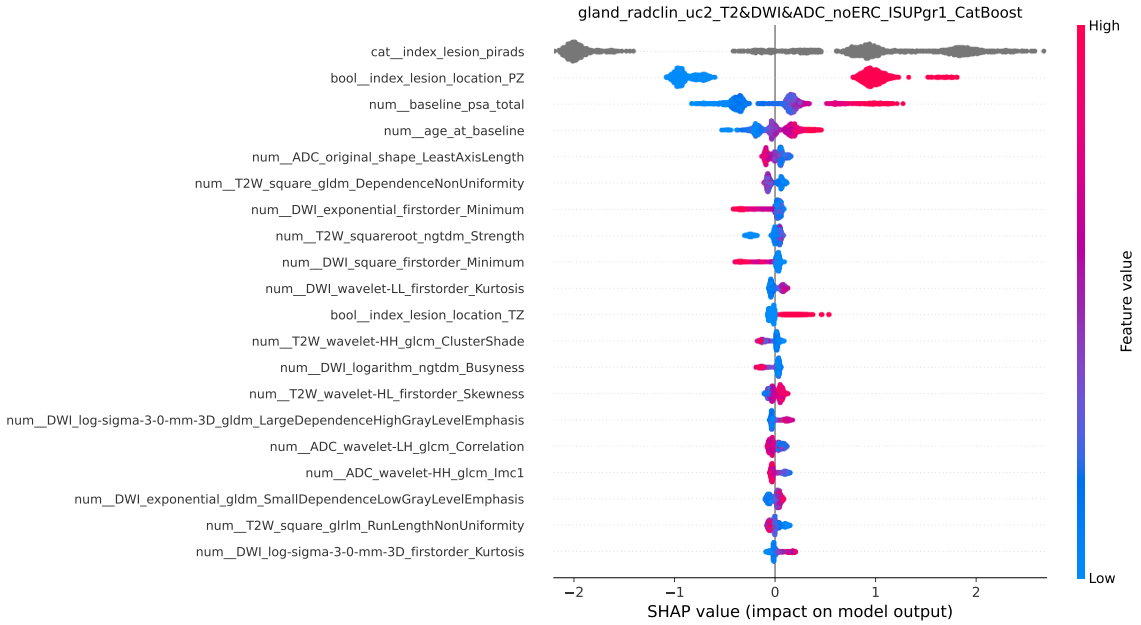


Figure 7.5: SHAP values contribution of features for radclin_uc2_T2&DWI&ADC_noERC_CatBoost model output.

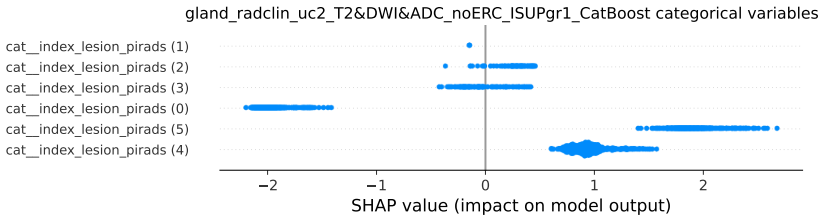


Figure 7.6: SHAP summary plot of the categorical variable index_lesion_pirads within model radclin_uc2_T2&DWI&ADC_noERC_CatBoost

7.3.1.4 Fairness and sub-cohort analysis

Figure 7.7 shows the model's performance on different subsets of the prospective cohort. Regarding scanner manufacturer (top left), there are no extreme differences in performance between

GE and SIEMENS exams, but for PHILIPS there is a decrease in sensitivity and an increase in specificity. In ERC cases there is an increase in sensitivity but an extreme reduction in specificity.

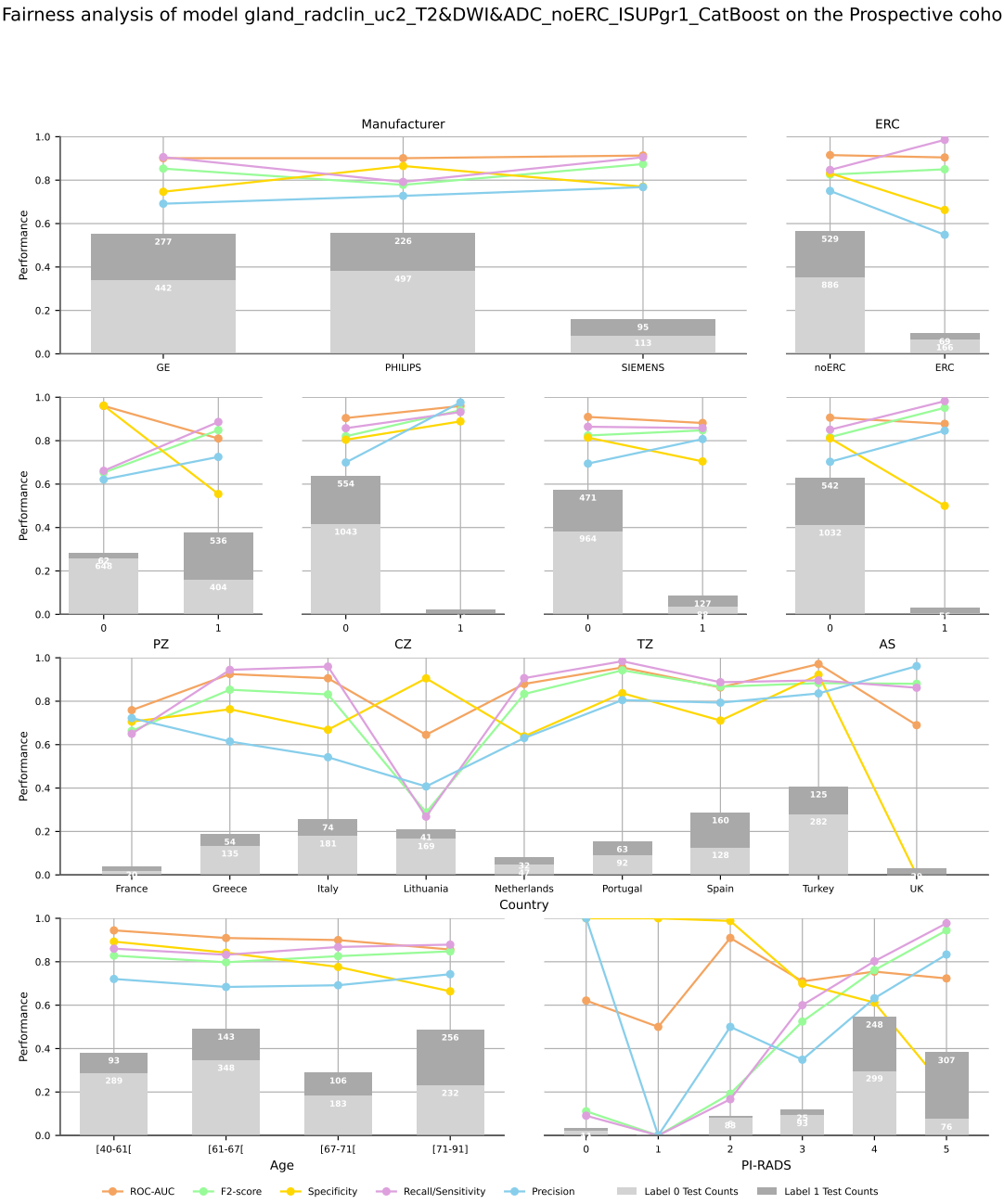


Figure 7.7: Performance of the model radclin_uc2_T2&DWI&ADC_noERC_CatBoost on sub cohorts of the prospective test set. The plots are colour coded according to metric: accuracy in orange, F2-score in green, precision in blue and recall in purple. Each plot also displays stacked bars indicating the subcohort size and label distribution. The four plots in the second row refer to the presence (1) or absence (0) of the index lesion in the different anatomical zones (PZ, TZ, CZ and AS).

Regarding lesion location, a 15% drop in AUC is visible when the index lesion is located within the PZ, although this might be in part due to the differences in target distribution between the two subcohorts. Similarly, there is a slight drop in performance when the index lesion is located in the TZ. Regarding the central zone and anterior stroma, there were not enough samples in the minority cohorts for these to be representative.

Regarding country of origin, both sensitivity and specificity remain consistent across the different countries, with the exception of Lithuania, where there is an inversion: specificity rises to 0.9, whereas sensitivity drops 0.3. This is in all likelihood due to the high prevalence of ISUP 2 cases (37 out of 41 positive cases), which are more challenging for the model to differentiate. Overall, conclusions about country-wise performance should be taken with caution since some subcohorts are relatively small.

In terms of age, sensitivity remains stable, while specificity consistently drops as patients get older. Regarding PI-RADS, there is an increase in sensitivity as PI-RADS increases, but this is accompanied by a reduction in specificity.

7.3.2 Use Case 2 - ISUP {0,1,2} vs {3,4,5}

7.3.2.1 Model Selection

Figure 7.8 shows the cross-validation and hold-out test set ROC-AUC model performance for the 96 models trained for UC2 (ISUP 0, 1 and 2 vs 3, 4 and 5). At first glance, we can see that, the inclusion of clinical variables (radclin and hybrid) leads to a higher overall performance. In terms of generalizability, we can see that training with all sequences leads to the highest test set performance. Out of each scanner manufacturer, the highest performance is achieved with phillips scanners (0.847 AUC on the hold-out test set), though it is still unable to outperform the model trained with all scanners combined. Overall, the highest hold-out test set performance reached was 0.853 AUC with a model trained with DWI sequences, radclin data and excluding enndorectal coil cases.

7.3.2.2 Model Validation

Given the results of the Section 7.3.2.1, we decided to move forward analyzing a radclin model using all sequences and excluding patients with endorectal coil from the training. The probability decision threshold with that results in at least 0.9 sensitivity on the Hold-out test set was 0.1411 (Figure 7.9), above which the model gives an output of clinical significance. The model's performance at this probability decision threshold is shown in Table 7.12 for the hold-out test set and the prospective cohort, and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Figure 7.10. The model was able to generalize to the prospective data, with a slight decrease in specificity of 4.6%, and an increase in sensitivity of 3.6%.

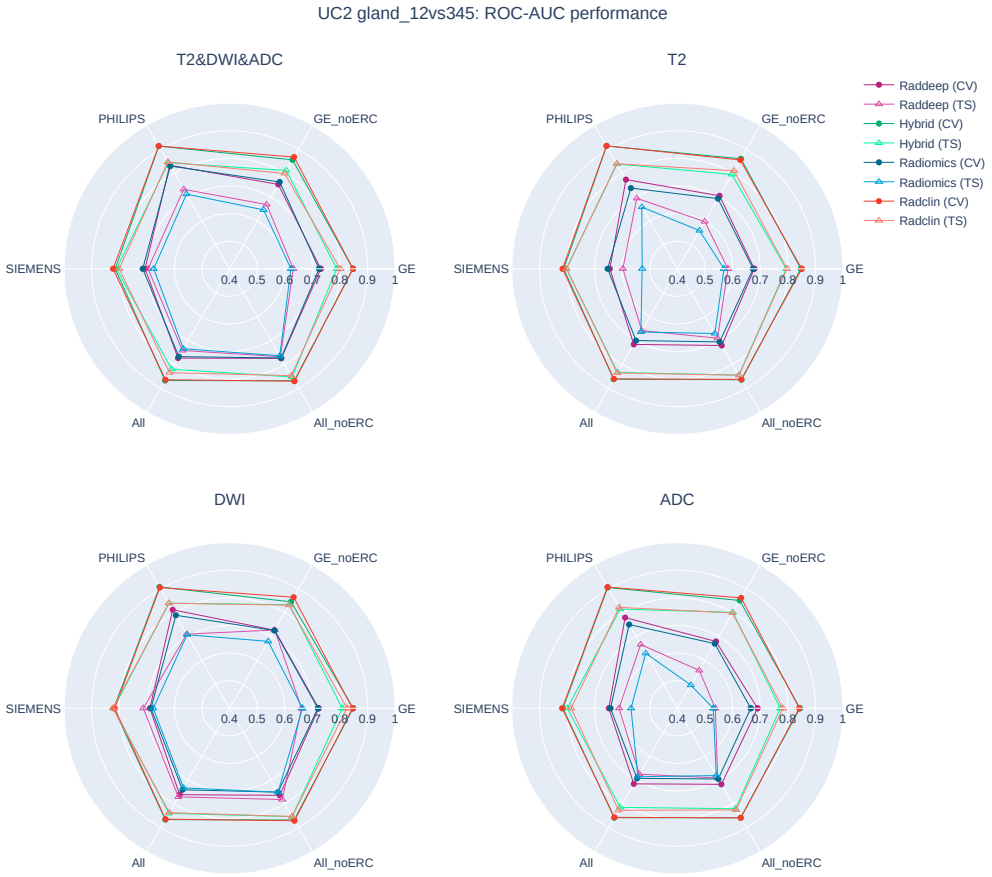


Figure 7.8: Cross-validation and hold-out test set ROC-AUC model performance of 32 models trained to predict disease aggressiveness in UC2 (ISUP 0, 1 & 2 vs 3, 4 & 5). The observations are color coded according to the type of data used in training: radiomics in blue, radclin in orange, raddeep in pink and hybrid in green

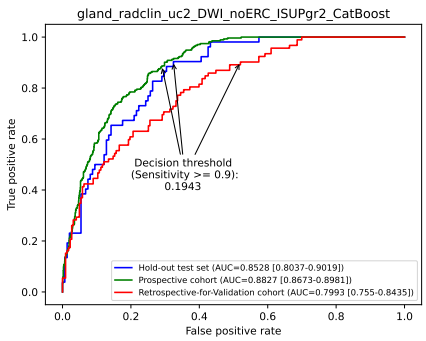


Figure 7.9: Receiver operator characteristics curve of the radclin_uc2_DWI_noERC_CatBoost model.

Table 7.12: Multi-metric performance of the radclin_uc2_DWI_noERC_CatBoost model at 0.1943 probability threshold.

	Hold-out test set	Retrospective for Validation	Prospective cohort
AUC	0.8528 [0.8037-0.9019]	0.7993 [0.755-0.8435]	0.8827 [0.8673-0.8981]
Sensitivity/Recall/TPR	0.9038 [0.863-0.9447]	0.9022 [0.8694-0.935]	0.8869 [0.8717-0.902]
Specificity/TNR	0.6757 [0.6108-0.7406]	0.4664 [0.4113-0.5215]	0.7068 [0.685-0.7285]
Precision/PPV	0.4947 [0.4254-0.564]	0.4109 [0.3566-0.4652]	0.371 [0.3479-0.3941]
F1	0.6395 [0.5729-0.706]	0.5646 [0.5099-0.6194]	0.5231 [0.4993-0.547]
F2	0.7756 [0.7178-0.8334]	0.7281 [0.6789-0.7772]	0.6939 [0.6718-0.7159]
CohensKappa	0.457 [0.3879-0.706]	0.2728 [0.2236-0.6194]	0.3806 [0.3574-0.547]

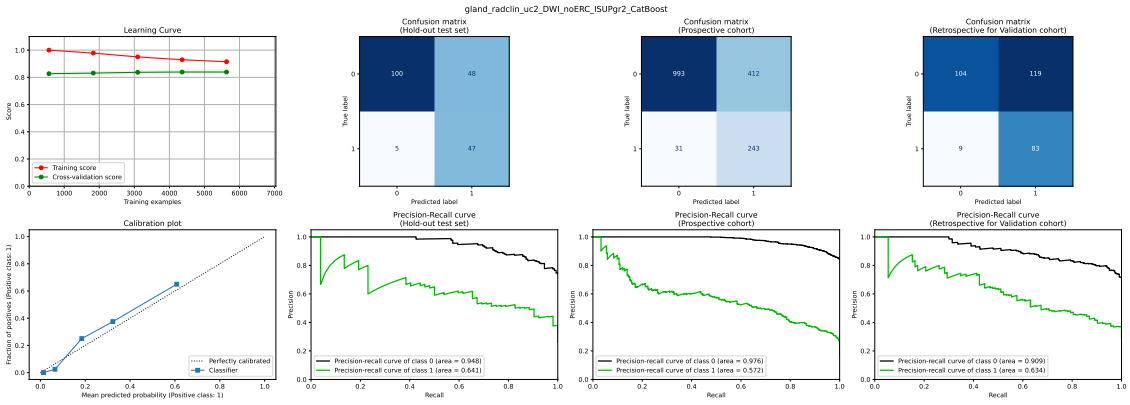


Figure 7.10: Analysis of the radclin_uc2_DWI_noERC_CatBoost model in terms of confusion matrix, learning curve, precision-recall curve and calibration plot.

7.3.2.3 Explainability analysis

Figure 7.11 shows the SHAP analysis performed on the radclin_uc2_DWI_noERC_CatBoost model. The plot shows the features ordered by their impact on model output. The four most relevant features for the model output are clinical variables, namely baseline PSA, patient age, PZ index lesion location and PIRADS (Figure 7.12), which are all positively associated with an aggressive output.

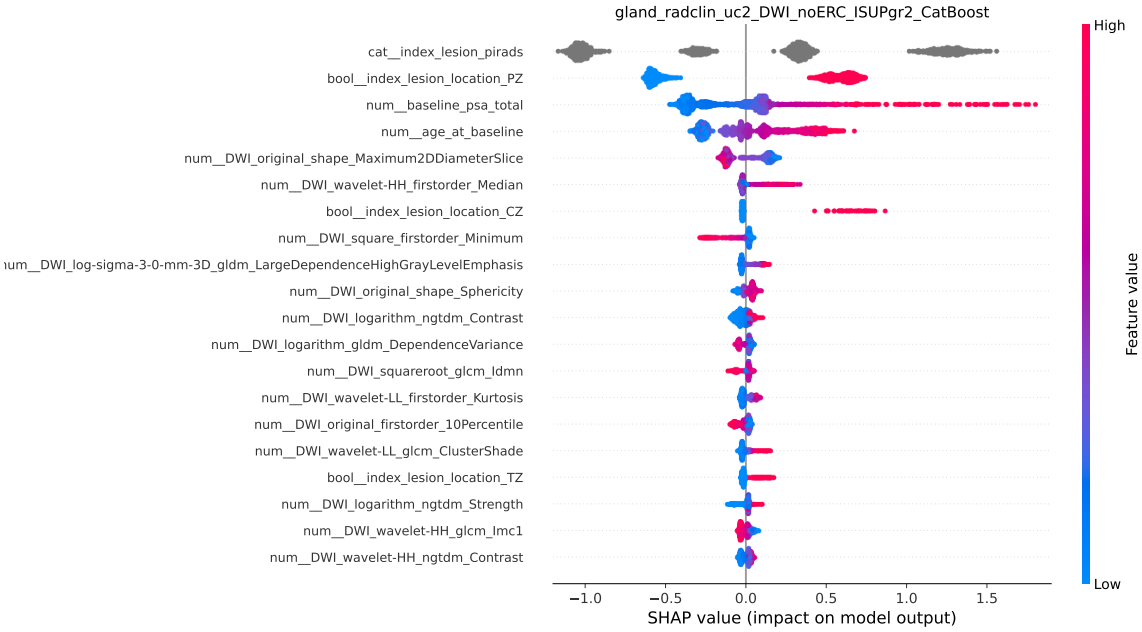


Figure 7.11: SHAP values contribution of features for radclin_uc2_DWI_noERC_CatBoost model.

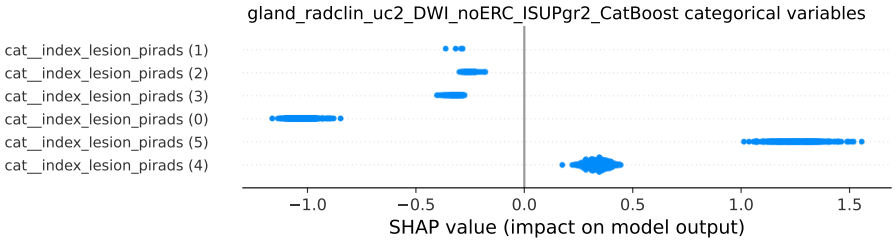


Figure 7.12: SHAP summary plot of the categorical variable index_lesion_pirads within model radclin_uc2_DWI_noERC_CatBoost

7.3.2.4 Fairness and sub-cohort analysis

Figure 7.13 shows the model's performance on different subsets of the prospective cohort. The performance is relatively consistent across scanner manufacturers (top left) and ERC usage (top right).

Regarding lesion location (second row), the model shows robustness to the transitional zone lesion location, but there is a reduction in performance when the index lesion is located in the peripheral zone. Regarding central zone and anterior stroma, there were not enough samples in the minority cohorts for accurate conclusions to be drawn.

Regarding country of origin, there is some variation in performance, but this is likely due to the different target distributions in the subcohorts. In terms of patient age and PI-RADS, there is a consistent reduction in specificity as patients get older or PI-RADS increases.

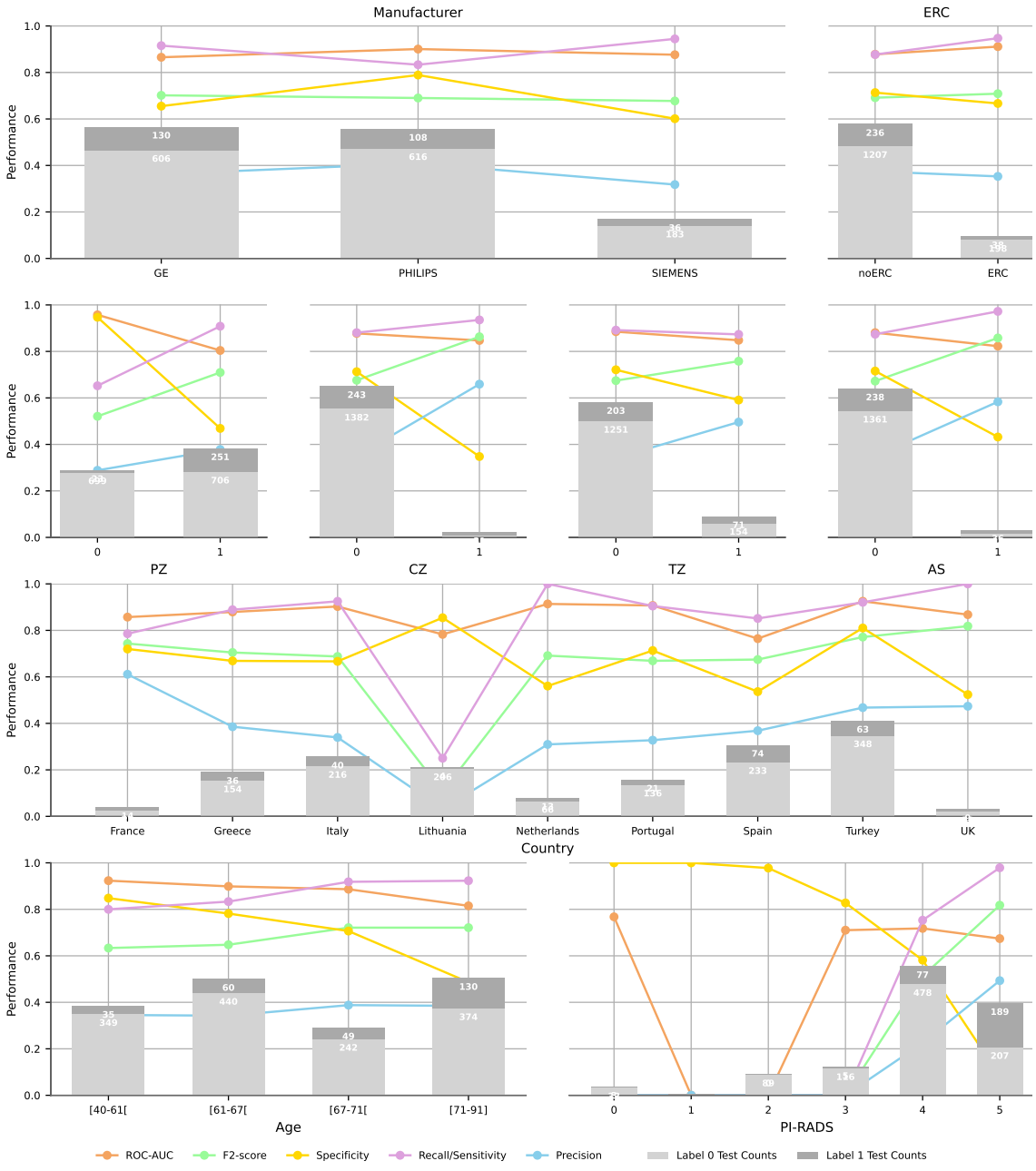


Figure 7.13: Performance of the model radclin_uc2_DWI_noERC_CatBoost on sub cohorts of the prospective cohort. The plots are color-coded according to metric: accuracy in orange, F2-score in green, precision in blue and recall in purple. Each plot also displays stacked bars indicating the subcohort size and label distribution. The four plots in the second row refer to the presence (1) or absence (0) of the index lesion in the different anatomical zones (PZ, TZ, CZ and AS).

7.3.3 Use Case 3 - Development of Metastasis

7.3.3.1 Model Selection

Figure 7.14 shows the cross-validation ROC-AUC model performance for the 32 models trained for UC3. Even though the differences are not extensive, the highest performance is achieved with a raddeep DWI dataset, using uniquely PHILIPS cases.

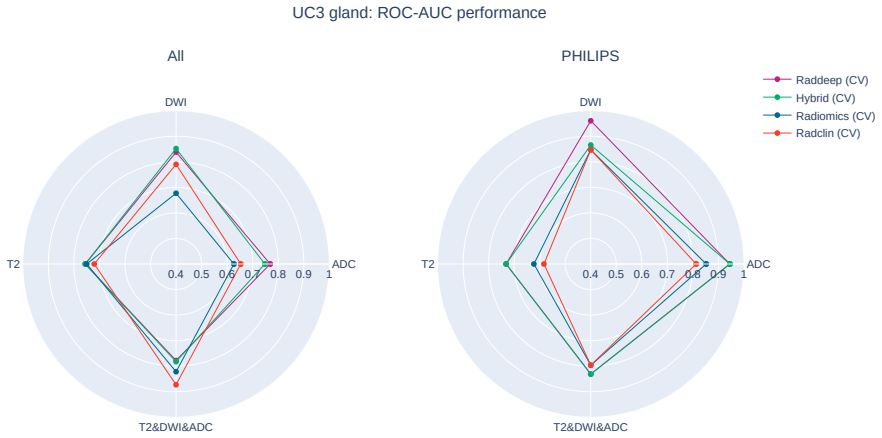


Figure 7.14: Cross-validation ROC-AUC model performance of 32 models trained to predict metastatic development in UC3. The observations are color coded according to the inclusion or exclusion of endorectal-coil exams: all patients in green; excluding ERC cases in pink.

7.3.3.2 Model Validation

Given the results of the Section 7.3.3.1, we decided to move forward analyzing a hybrid model using DWI sequences and excluding ERC patients. The model's performance at a 0.5 (Figure 7.15) probability decision threshold is shown in Table 7.13 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Figure 7.16.

7.3.3.3 Explainability analysis

Figure 7.17 shows the logistic regression coefficients of model raddeep_uc3_DWI_PHILIPS_SGD. Only two variables were selected by the model, both representing texture descriptors.

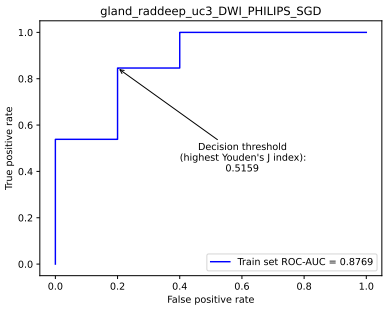


Figure 7.15: Receiver operator characteristics curve of the raddeep_uc3_DWI_PHILIPS_SGD model.

Table 7.13: Multi-metric cross-validation performance of the raddeep_uc3_DWI_PHILIPS_SGD model at 0.5 probability threshold.

	Cross-validation
AUC	0.8769 [0.7613-0.9926]
Sensitivity/Recall/TPR	0.8462 [0.7191-0.9732]
Specificity/TNR	0.8000 [0.6592-0.9408]
Precision/PPV	0.9565 [0.8847-1.0283]
F1	0.8980 [0.7914-1.0045]
F2	0.8661 [0.7463-0.9860]
CohensKappa	0.5201 [0.3443-1.0000]

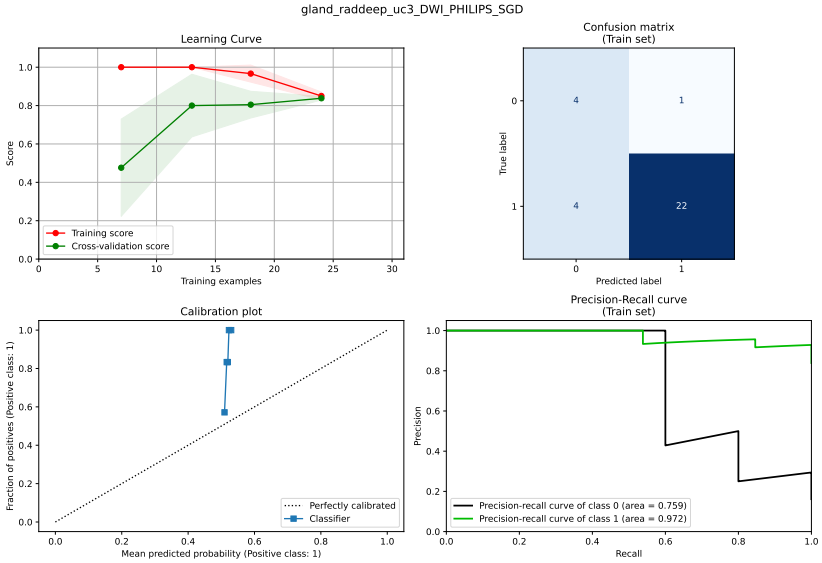


Figure 7.16: Analysis of the raddeep_uc3_DWI_PHILIPS_SGD model in terms of cross-validation confusion matrix, learning curve, precision-recall curve and calibration plot.

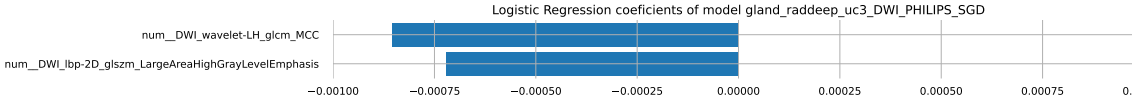


Figure 7.17: Logistic regression coefficients of features for raddeep_uc3_DWI_PHILIPS_SGD model.

7.3.4 Use Case 5 - BCR pre-RP

7.3.4.1 Model Selection

Figure 7.18 shows the cross-validation and hold-out test set ROC-AUC model performance for the 88 models trained for UC5 (pre-surgery context). In terms of cross-validation, it seems that GE models show the highest performance. However, this performance does not generalize to the hold-out test set. The highest hold-out test set performance is found with the radclin_uc5_T2&DWI&ADC_Catboost model.



Figure 7.18: Cross-validation and hold-out test set ROC-AUC model performance of 88 models trained to predict disease aggressiveness in UC5 (pre-surgery context). The observations are color coded according to the type of data used in training: radiomics in blue, radclin in orange, raddeep in pink and hybrid in green.

7.3.4.2 Model Validation

Analyzing the ROC curve (Figure 7.19) for the radclin_uc5_T2&DWI&ADC_CatBoost model, the decision threshold of 0.0328 was chosen, above which the model outputs a high risk of biochemical recurrence. The model's performance at this probability decision threshold is shown in Table 7.14 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Figure 7.20.

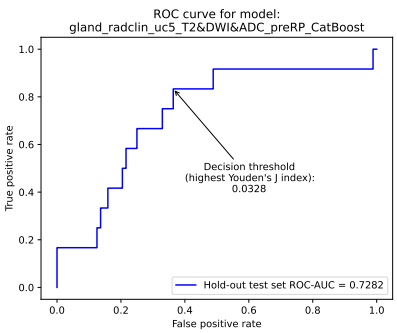


Figure 7.19: Hold-out test set receiver operator characteristics curve of the radclin_uc5_T2&DWI&ADC_CatBoost model.

Table 7.14: Multi-metric performance of the radclin_uc5_T2&DWI&ADC_CatBoost model at 0.0021 probability threshold.

Hold-out test set	
AUC	0.7282 [0.6410-0.8154]
Sensitivity/Recall/TPR	0.8333 [0.7603-0.9064]
Specificity/TNR	0.6364 [0.5421-0.7306]
Precision/PPV	0.2381 [0.1546-0.3216]
F1	0.3704 [0.2757-0.4650]
F2	0.5556 [0.4582-0.6529]
Cohen's Kappa	0.2259 [0.1439-0.4650]

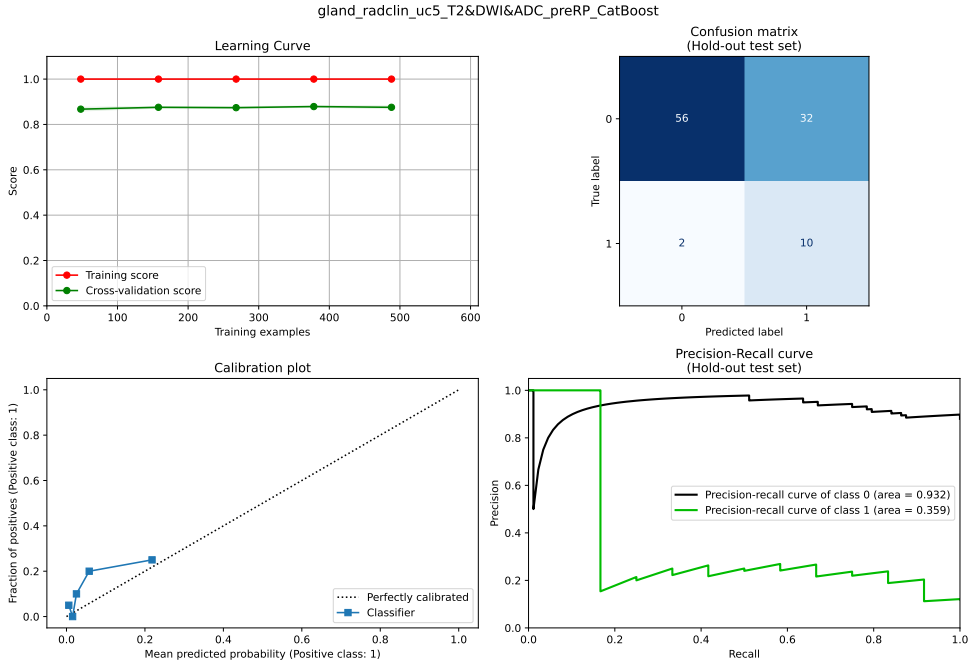


Figure 7.20: Analysis of the radclin_uc5_T2&DWI&ADC_CatBoost model in terms of confusion matrix, learning curve, precision-recall curve and calibration plot.

7.3.4.3 Explainability analysis

Figure 7.21 shows the SHAP analysis performed on the hold-out test set predictions of the radclin_uc5_T2&DWI&ADC_CatBoost model. The most impactful variables are Gleason 1, PSA, DWI texture markers, and PI-RADS.

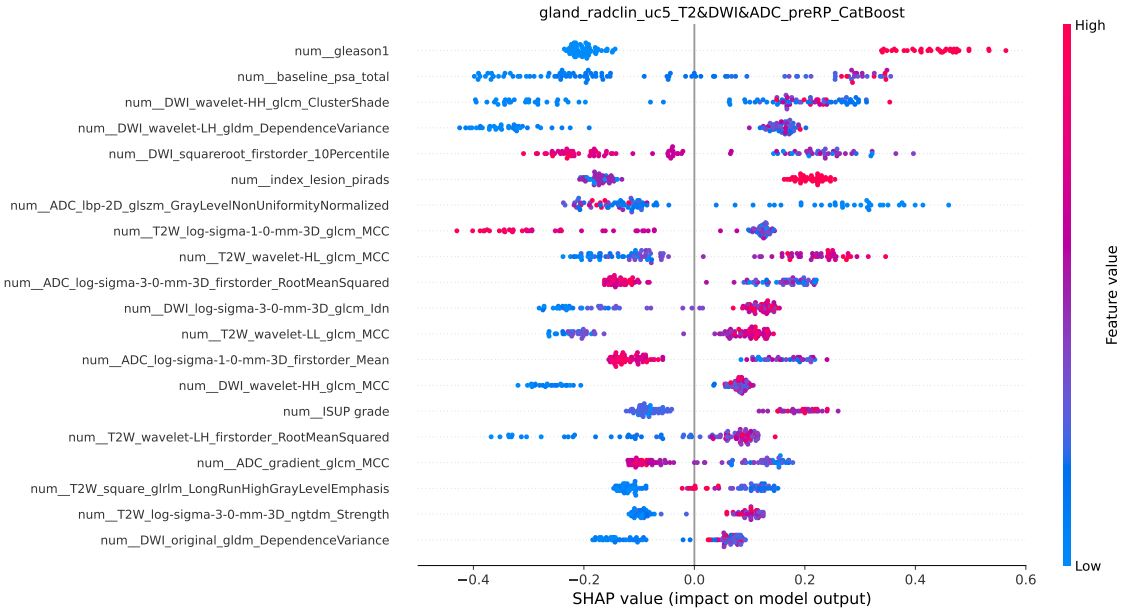


Figure 7.21: SHAP values contribution of features for radclin_uc5_T2&DWI&ADC_CatBoost model.

7.3.5 Use Case 5 - BCR post-RP

7.3.5.1 Model Selection

Figure 7.22 shows the cross-validation and hold-out test set ROC-AUC model performance for the 96 models trained for UC5 (post-surgery context). At first glance, we can see that training with GE data results in very overfitted classifiers, whereas including all scanner manufacturers results in the highest generalizability. Overall, the highest hold-out test set performance was reached with the model radclin_uc5_T2_noERC_CatBoost.



Figure 7.22: Cross-validation and hold-out test set ROC-AUC model performance of 96 models trained to predict biochemical recurrence in UC5. The observations are color coded according to the type of data used in training: radiomics in blue, radclin in orange, raddeep in pink and hybrid in green.

7.3.5.2 Model Validation

Analyzing the ROC curve (Figure 7.23) for the radclin_uc5_T2_noERC_CatBoost model, the decision threshold of 0.0174 was chosen, above which the model outputs a high risk of biochemical recurrence. The model's performance at this probability decision threshold is shown in Table 7.15 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Figure 7.24.

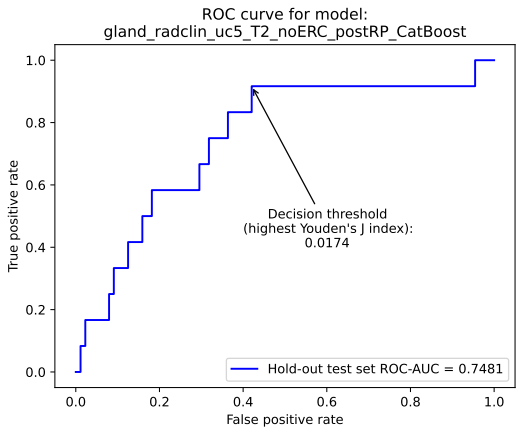


Figure 7.23: Hold-out test set receiver operator characteristics curve of the radclin_uc5_T2_noERC_CatBoost model.

Table 7.15: Multi-metric performance of the radclin_uc5_T2_noERC_CatBoost model at 0.0174 probability threshold.

Hold-out test set	
AUC	0.7481 [0.663-0.8332]
Sensitivity/Recall/TPR	0.9167 [0.8625-0.9708]
Specificity/TNR	0.5795 [0.4828-0.6763]
Precision/PPV	0.2292 [0.1468-0.3115]
F1	0.3667 [0.2722-0.4611]
F2	0.5729 [0.476-0.6699]
CohensKappa	0.2162 [0.1355-0.4611]

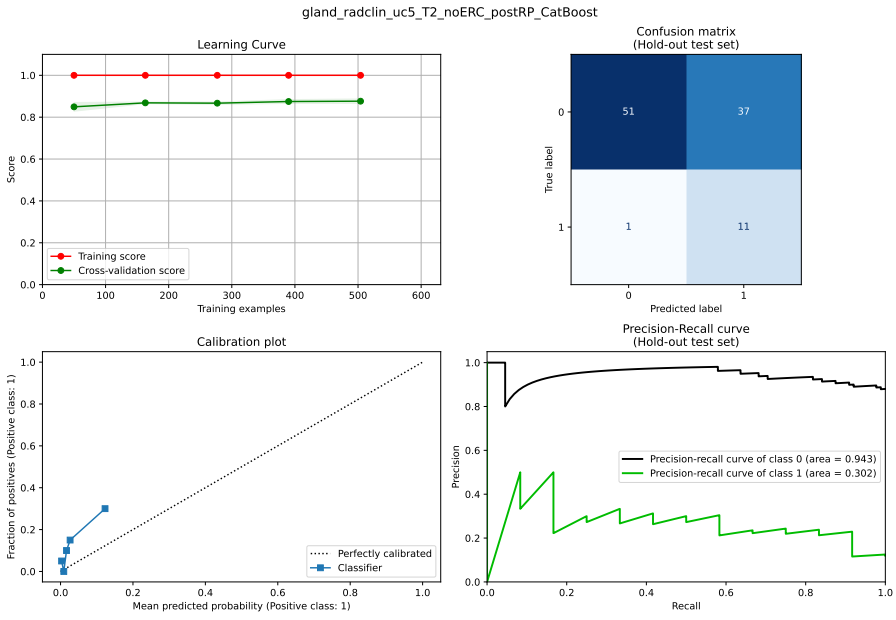


Figure 7.24: Analysis of the radclin_uc5_T2_noERC_CatBoost model in terms of confusion matrix, learning curve, precision-recall curve, and calibration plot.

7.3.5.3 Explainability analysis

Figure 7.25 shows the SHAP analysis performed on the hold-out test set predictions of the radclin_uc5_T2_noERC_CatBoost model. Among the top 5 most influential variables to the model output, there are two clinical variables (namely extraprostatic extension and PSA blood levels) and three texture related radiomic variables. Resection margin status and invasion of perineural regions also play a crucial role in the classification (Figure 7.26), as well as the primary Gleason pattern and the ISUP grade. In concordance with the pre-surgical scenario, radiomic features extracted from the log-sigma or wavelet transformation of the T2W volume show the most predictive power.

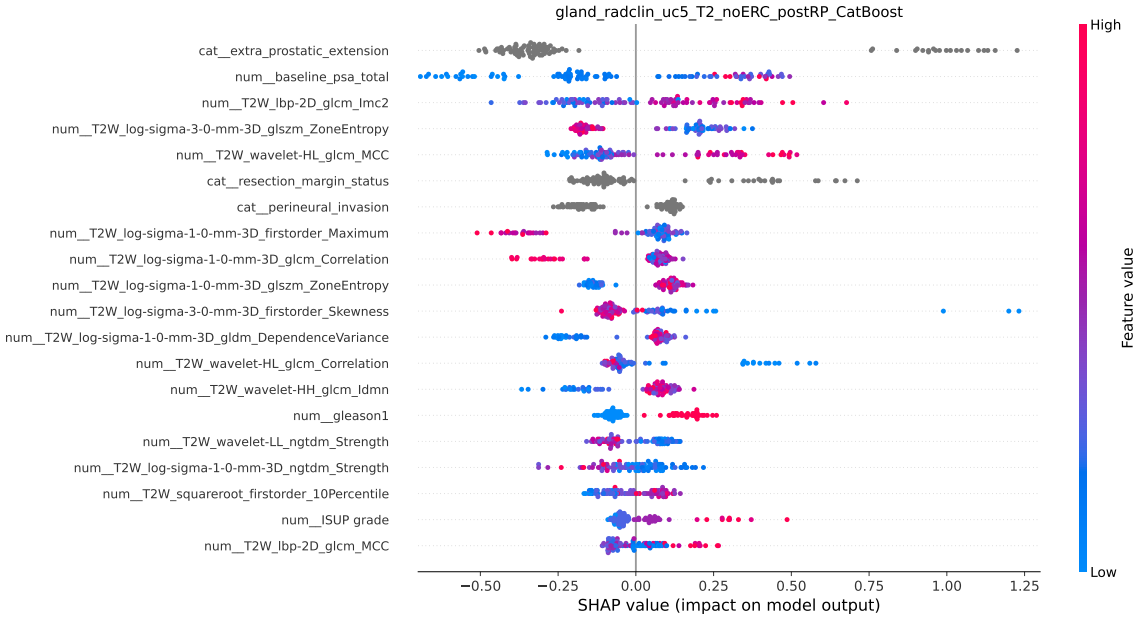


Figure 7.25: SHAP values contribution of features for radclin_uc5_T2_noERC_CatBoost model.

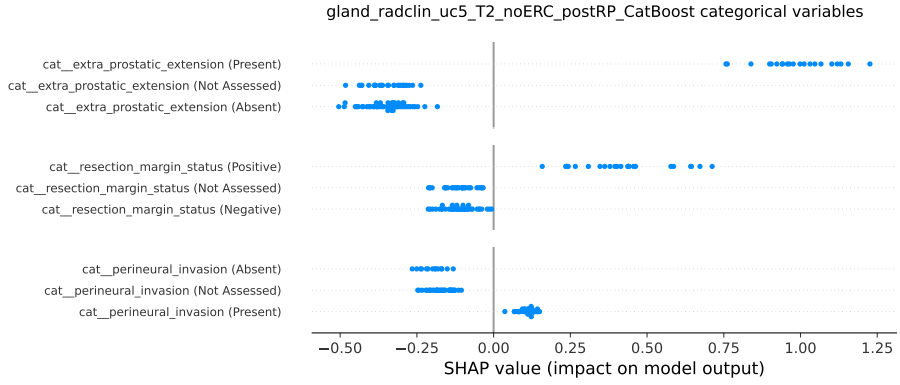


Figure 7.26: SHAP values contribution of categorical features for the radclin_uc5_T2_noERC_CatBoost model output.

7.3.6 Use Case 6 - BCR post-RT

7.3.6.1 Model Selection

Figure 7.27 shows the cross-validation ROC-AUC model performance for the 64 models trained for UC6. The outermost models are consistently those trained on radiomics or raddeep data, with the inclusion of clinical variables underperforming. Overall, the highest performance is achieved with radiomic features from the ADC map (0.85 AUC) and excluding patients with endorectal coil.

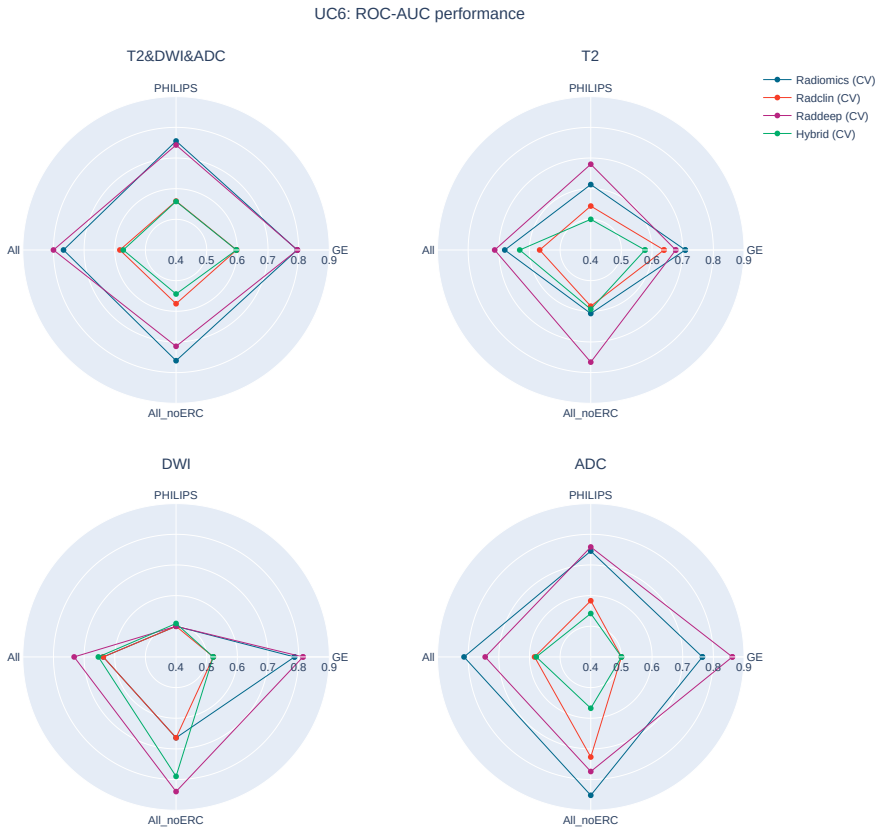


Figure 7.27: Cross-validation ROC-AUC model performance of 64 models trained to predict biochemical recurrence in UC6. The observations are color coded according to the type of data used in training: radiomics in blue, radclin in orange, raddeep in pink and hybrid in green.

7.3.6.2 Model Validation

Upon inspection of the radiomics_uc6_ADC_noERC_SGD model's ROC curve (Figure 7.28), the probability decision threshold of 1.0 was chosen, above which the model outputs a high risk of biochemical recurrence. The model's performance at this probability decision threshold is shown

in Table 7.16 and its confusion matrix, learning curve, precision-recall curve, and calibration plot are displayed in Figure 7.29. There were no prospective cases available for this use case.

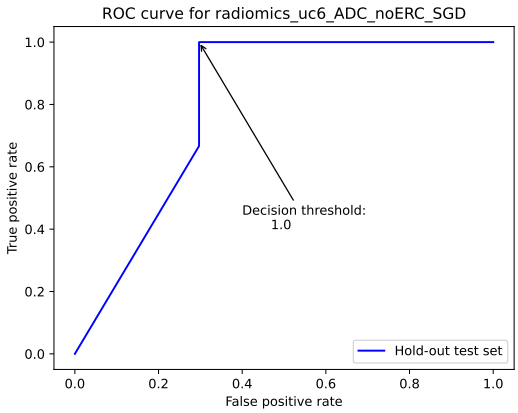


Figure 7.28: Cross-validation receiver operator characteristics curve of the radiomics_uc6_ADC_noERC_SGD model.

Table 7.16: Multi-metric cross-validation performance of the radiomics_uc6_ADC_noERC_SGD model at 1.0 probability threshold.

3-fold cross-validation	
AUC	0.8021 [0.7087-0.8954]
Sensitivity/Recall/TPR	1.0 [1.0-1.0]
Specificity/TNR	0.7031 [0.5961-0.8102]
Precision/PPV	0.24 [0.14-0.34]
F1	0.3871 [0.273-0.5012]
F2	0.6122 [0.4981-0.7264]
Cohen's Kappa	0.2888 [0.1826-0.5012]

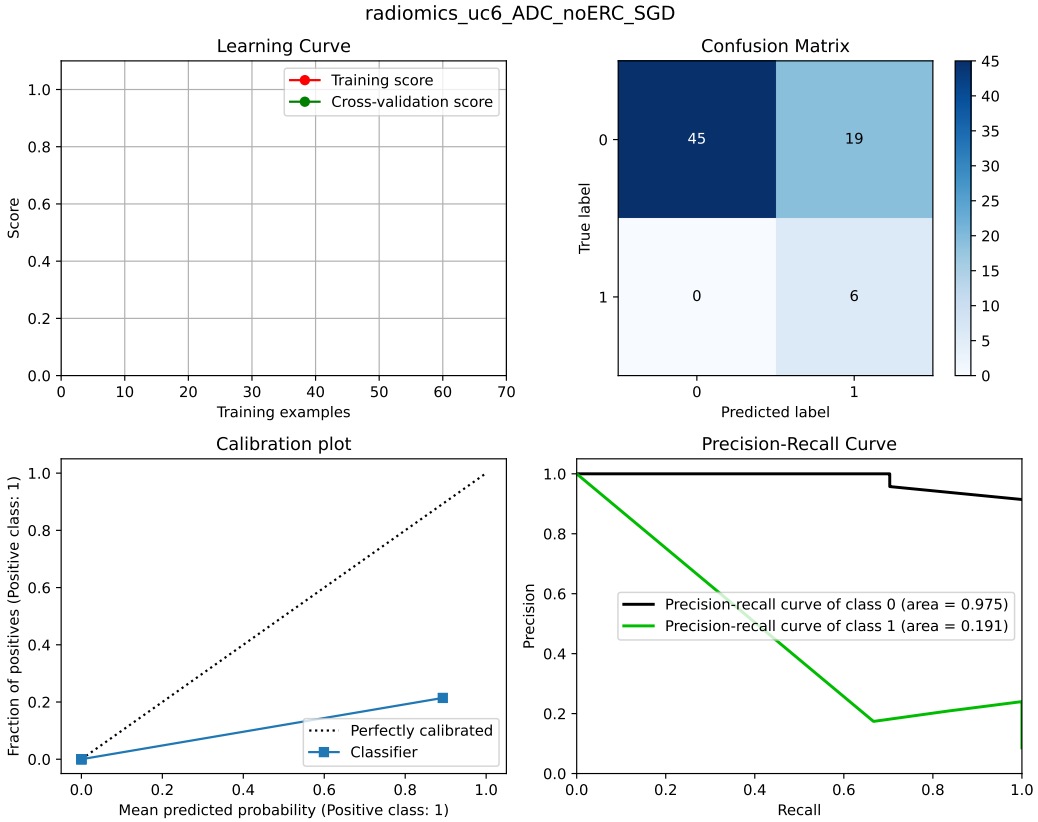


Figure 7.29: Analysis of the radiomics_uc6_ADC_noERC_SGD model in terms of cross-validation confusion matrix, learning curve, precision-recall curve, and calibration plot.

7.3.6.3 Explainability analysis

Figure 7.30 shows the SHAP analysis performed on the radiomics_uc6_ADC_noERC_SGD model. First-order and texture GLCM radiomic features are the most impactful. These are extracted from several transforms of the original volume, namely gradient, squareroot, wavelet and log-sigma. Only one shape feature is present: Maximum 2D diameter row, positively associated with biochemical recurrence.

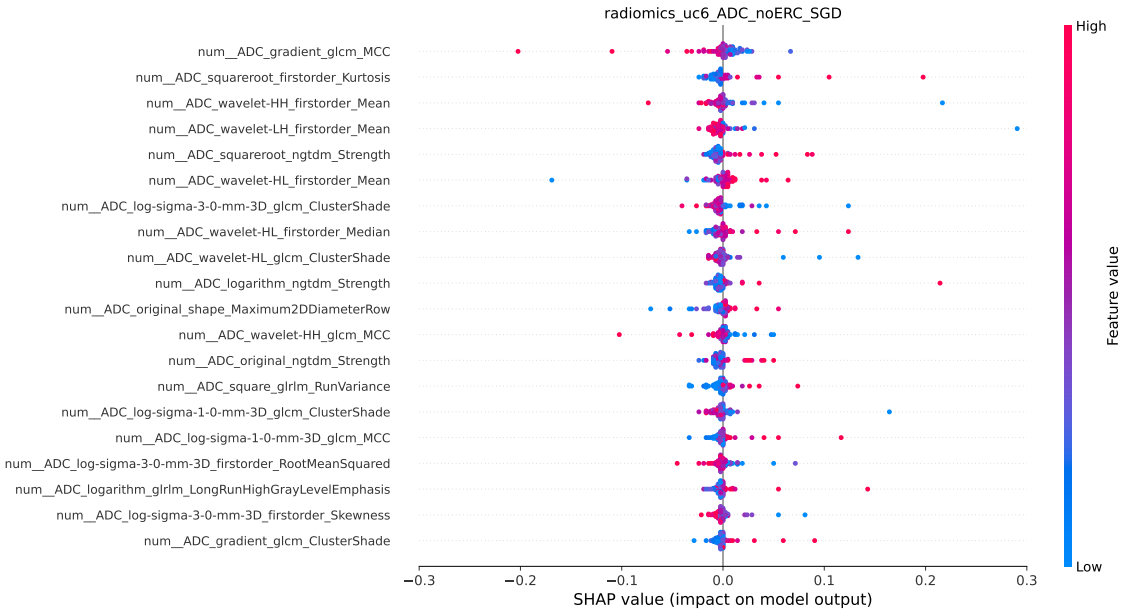


Figure 7.30: SHAP values contribution of features for radiomics_uc6_ADC_noERC_SGD model.

7.3.7 Use Case 7b

7.3.7.1 Model Selection

Figure 7.31 shows the cross-validation ROC-AUC model performance for the 48 models trained for UC7b. Here, we see that the highest cross-validation performances are achieved with SIEMENS data, however these are highly overfitted. Overall, the model that shows the highest performance is a radiomics model trained with DWI sequences all scanner manufacturers (0.68 one-vs-rest AUC).

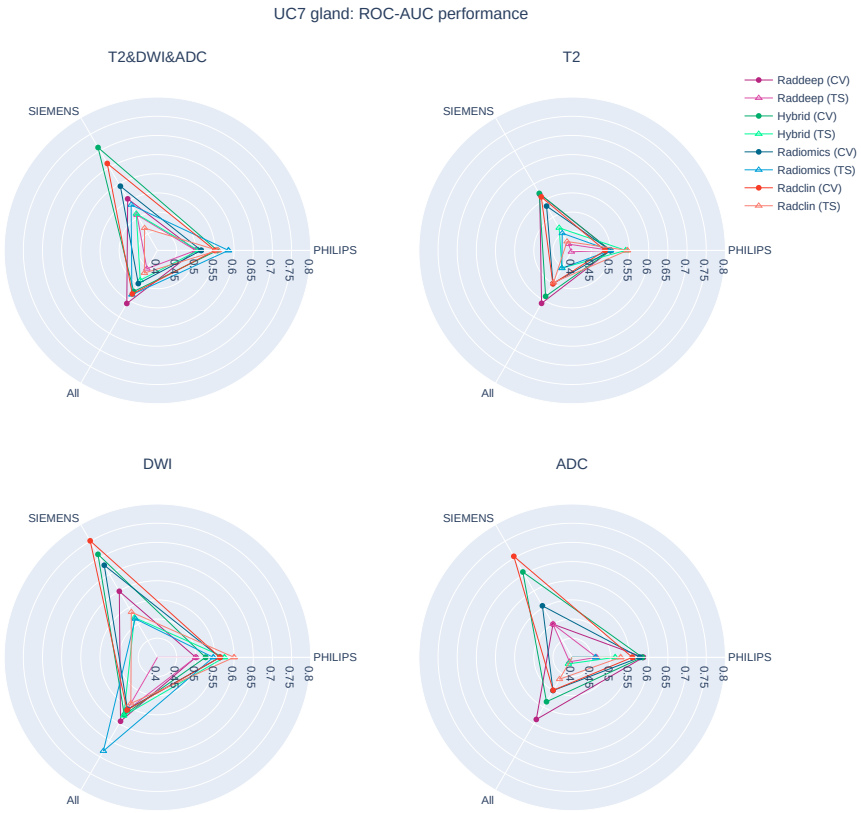


Figure 7.31: Cross-validation ROC-AUC model performance of 48 models trained to predict early quality of life after radical prostatectomy. The observations are color coded according to the type of data used in training: radiomics in blue, radclin in orange, raddeep in pink and hybrid in green.

7.3.7.2 Model Validation

Further analysis of the model selected in Section 7.3.7.1 revealed that its discriminatory power was extremely reduced (Figure 7.32 and Table 7.17). Thus, no further analysis was conducted.

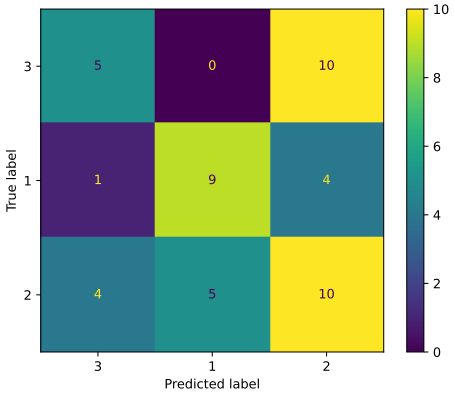


Figure 7.32: Hold-out test set confusion matrix of the radiomics_uc7_DWI_LGBM model.

Table 7.17: Multi-metric hold-out test set performance of the radiomics_uc7_DWI_LGBM model.

Hold-out test set	
AUC	0.6822 [0.5505-0.8139]
Sensitivity/Recall/TPR	0.5008 [0.3594-0.6423]
Precision/PPV	0.5198 [0.3785-0.6612]
F1	0.5027 [0.3612-0.6441]
F2	0.5 [0.3586-0.6414]

7.3.8 Use Case 8 - Disease proggression

7.3.8.1 Model Selection

Figure 7.33 shows the cross-validation ROC-AUC model performance for the 64 models trained for UC8. Here, we see the performance is enhanced in the presence of GE DWI features. The model that shows the highest performance is a radiomics model trained with DWI sequences taken with a GE scanner (0.85 AUC).

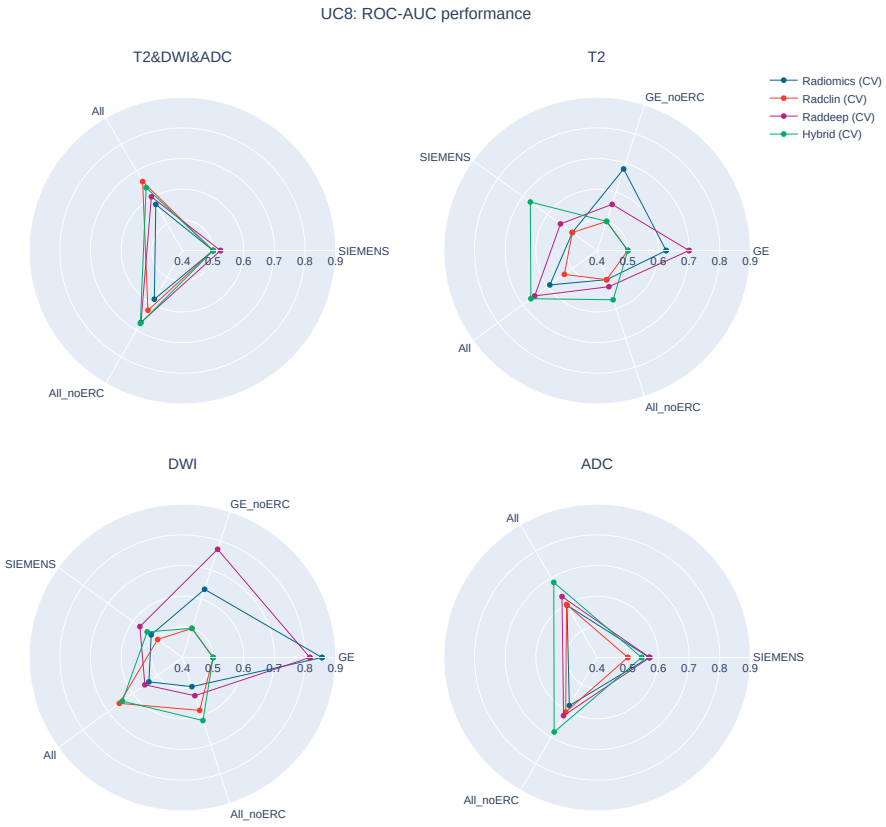


Figure 7.33: Cross-validation ROC-AUC model performance of 64 models trained to predict early withdrawal from the active surveillance program in UC8. The observations are color coded according to the type of data used in training: radiomics in blue, radclin in orange, raddeep in pink and hybrid in green.

7.3.8.2 Model Validation

Given the results of Section 7.3.8.1, we decided to move forward analyzing a radiomics model using DWI sequences from GE scanners. The model's cross-validation performance at a 0.5 (Figure 7.34) probability decision threshold is shown in Table 7.18 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Figure 7.35.

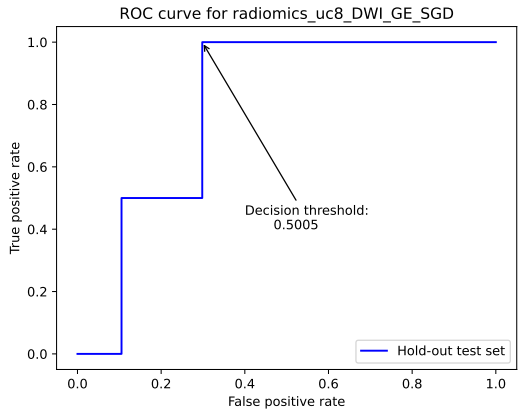


Figure 7.34: Cross-validation receiver operator characteristics curve of the radiomics_uc8_DWI_GE_SGD model.

Table 7.18: Multi-metric cross-validation performance of the radiomics_uc8_DWI_GE_SGD model at 0.5005 probability threshold.

3-fold cross-validation	
AUC	0.7982 [0.6958-0.9006]
Sensitivity/Recall/TPR	1.0 [1.0-1.0]
Specificity/TNR	0.7018 [0.585-0.8185]
Precision/PPV	0.1053 [0.027-0.1836]
F1	0.1905 [0.0903-0.2907]
F2	0.3704 [0.2471-0.4936]
Cohen's Kappa	0.1376 [0.0497-0.2907]

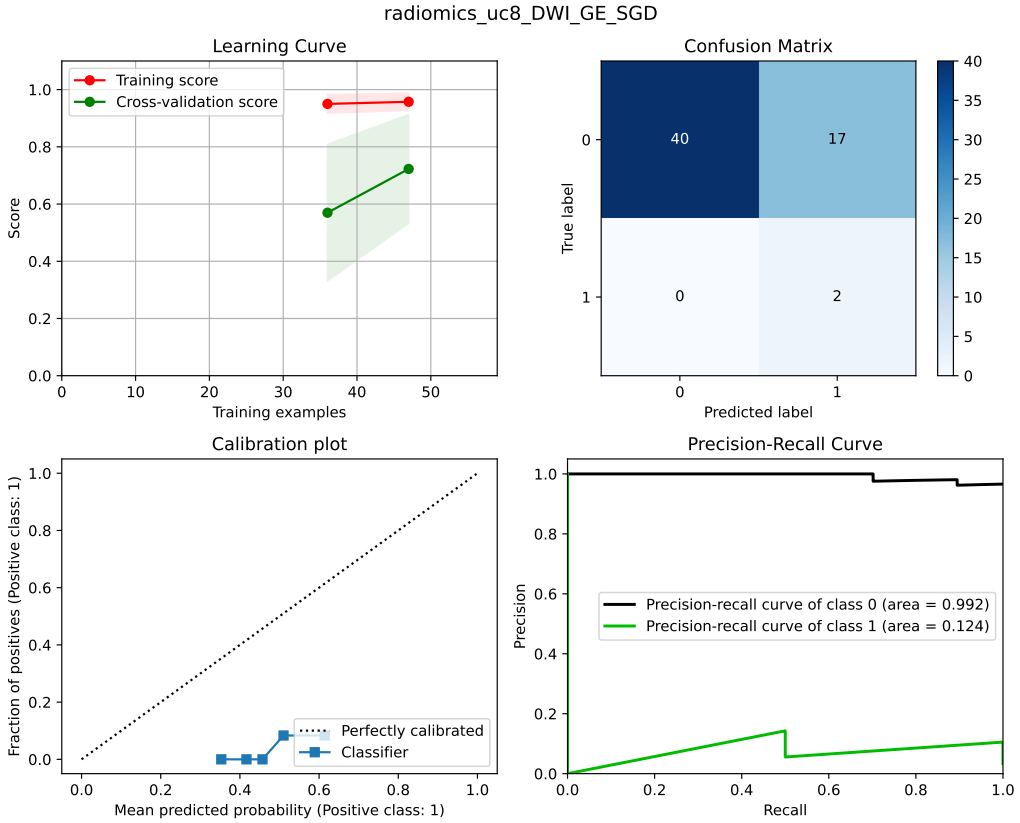


Figure 7.35: Analysis of the radiomics_uc8_DWI_GE_SGD model in terms of cross-validation confusion matrix, learning curve, precision-recall curve and calibration plot.

7.3.8.3 Explainability analysis

Figure 7.36 shows the SHAP analysis performed on the radiomics_uc8_DWI_GE_SGD model.

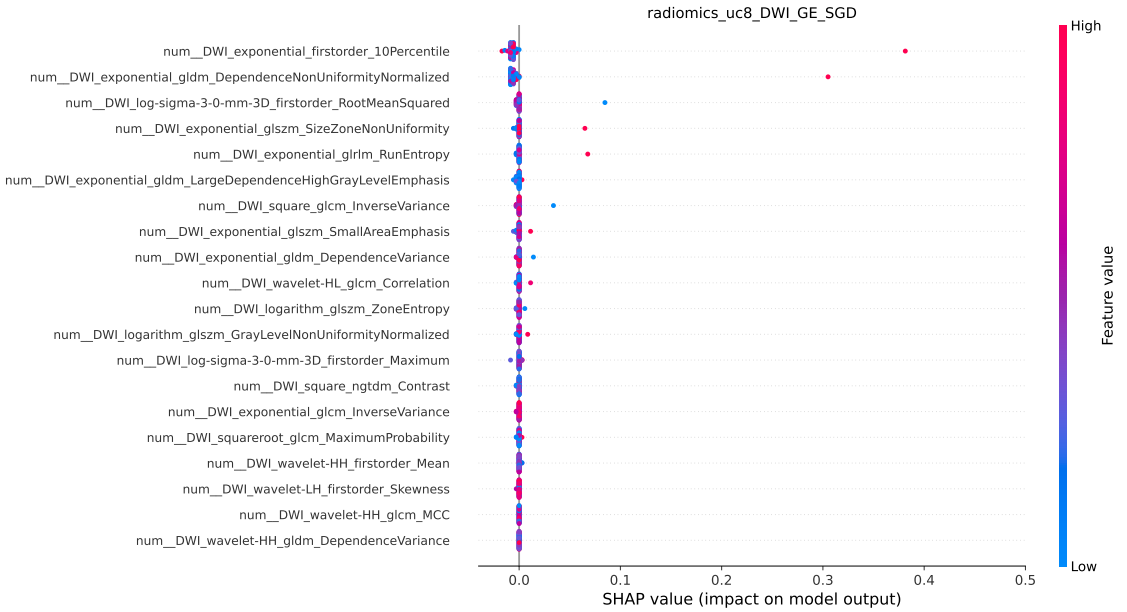


Figure 7.36: SHAP values contribution of features for radiomics_uc8_DWI_GE_SGD model.

7.4 Discussion

In this work, we report the use of bi-parametric MRI whole-prostate-gland radiomics analysis for six separate use cases: the prediction of disease aggressiveness (UC2), development of metastasis within six months (UC3), biochemical recurrence risk after prostatectomy or radiation therapy (UC5 and UC6, respectively), quality of life after prostatectomy (UC7b) and early withdrawal from the active surveillance program (UC8).

Regarding UC2, the ambiguity in cancer grading lead to the training of two binary classification models (ISUP=0,1 vs ISUP=2,3,4,5; and ISUP=0,1,2 vs ISUP=3,4,5). Interestingly, for both ground truth definitions, the highest performing models combined information from the three MRI volumes (T2W, DWI and ADC) to arrive at a decision, and the preferred feature type was the combination of radiomics and clinical data. The latter ground truth showed similar sensitivity but around 10% lower specificity and 40% lower precision, which was reflected in a slightly lower AUC performance. This ground truth definition was expected to be the hardest to model, given the known mixture of clinically significant and insignificant cases in the ISUP classifications 2 and 3, corresponding to GS = 3 + 4 and GS = 4 + 3, respectively. Despite this, both models showed exceptional generalization power to the prospective cases. The explainability analysis revealed that the two models relied heavily on the clinical variables PI-RADS, index lesion location in the PZ, PSA blood level, and age, all showing a positive association with the aggressive output. Both models rely on shape features related to gland size, indicating that larger prostate glands are associated with low aggressiveness. Similarly, both shap analysis revealed a high representation of radiomic features extracted from wavelet filters, which are known for their relative invariability across volumes with big differences in intensity values, such as those found in different scanner manufacturers or in exams taken with or without endorectal coil. This was further supported by the fairness analysis.

Use case 3, prediction of metastatic development within six months, was the smallest use case, with only 77 complete cases. Despite the small dataset, the selected model achieved a 3-fold cross-validation AUC of 0.7904. Surprisingly, the model was very good at identifying the minority class (specificity of 1.0), corresponding to no metastasis after six months, but not so great with the majority class (sensitivity 0.6154), corresponding to metastatic development within six months. Even though the model was trained with a hybrid dataset (radiomics, clinical and deep features), only one clinical variable was found among the 20 most important to the model's output, relying heavily on deep features.

Use case 5 related to the prediction of biochemical recurrence after radical prostatectomy surgery. For this use case, two approaches were taken: pre-surgery prediction and post-surgery prediction. The latter including clinical variables reported during or after the radical prostatectomy surgery. Interestingly, in the post-surgery scenario the selected model was a radclin, while in the pre-surgery scenario a hybrid model was preferred. This indicates that the deep features were capable of compensating for the absence of surgical information such as extra prostatic extension, resection margins, and perineural invasion, that were revealed very impactful for the post-surgery

model output, in the explainability analysis.

Use case 6 aimed to predict biochemical recurrence after radiation therapy. The selected model achieved a 3-fold cross-validation AUC of 0.8021, and sensitivity and specificity of 1.0 and 0.7031, respectively. Interestingly, and in contrast with the remaining clinical use cases, here, the inclusion of clinical variables consistently lead to lower performance. This was the only use case where an ADC-based model was ultimately selected.

Regarding UC7b, the epic26 continuous values were divided into three bins, representing bad, average and good quality of life. With these new labels, a multiclass classification approach was followed. This was not successful, as the results were extremely poor. Following work will explore lesion-based radiomics.

Use case 8 related to the prediction of an early withdrawal from the active surveillance program. Here, we achieved an AUC of 0.7982, and 1.0 and 0.7018 sensitivity and specificity, respectively. The precision, however, was extremely low (10%) indicating a large amount of false positives, which are patients wrongly predicted to show disease progression during the active surveillance program. Even though this is not as unwelcome as false negatives, it still represents a considerable fraction of patients who would unnecessarily be put through aggressive treatment, and suffer through the reduced quality of life associated with it.

Overall, despite the large data size of some use cases, the learning curves demonstrated there still was learning capacity. The discrepancy between training and cross-validation performance was as high as 12% for UC7b, while for use cases 2 and 5 the discrepancy was no larger than 2%.

Several limitations can be identified in the body of work. The first and largest limitation was the very reduced dataset size of some use cases, the most extreme case being UC3. Secondly, the same methodology (or with only slight deviations) was applied to the different use cases. This is not ideal, as different classification problems might thrive with different learning strategies. With more time other methodologies could be explored.

Chapter 8

Lesion Radiomics models

In this Chapter, the index-lesion radiomic features from three MRI volumes (T2W, DWI and ADC) from the ProstateNet dataset (described in Chapter 5) were used to develop classification models pertaining to six separate use cases, within the prostate cancer disease continuum: the prediction of disease aggressiveness (UC2), development of metastasis within six months (UC3), biochemical recurrence risk after prostatectomy or radiation therapy (UC5 and UC6, respectively), quality of life after prostatectomy (UC7b) and early withdrawal from the active surveillance program (UC8).

8.1 Introduction

Building on the analysis from Chapter 7 this chapter focuses on the extraction of radiomic features using index-lesion segmentation masks, rather than whole-gland segmentation, to further refine predictive modelling for prostate cancer. Index lesions, typically representing the largest or most aggressive tumor regions within the prostate, are believed to have a more direct correlation with clinical outcomes.

This chapter develops classification models for the same six critical clinical use cases: predicting disease aggressiveness (UC2), early metastasis within six months (UC3), biochemical recurrence after prostatectomy or radiation therapy (UC5 and UC6, respectively), quality of life post-prostatectomy (UC7b), and withdrawal from active surveillance (UC8). By focusing on the index lesion, the models in this chapter aim to test whether a more localized approach to radiomics can improve predictive accuracy for these clinical outcomes.

The comparison between whole-gland and index-lesion approaches offers valuable insights into which segmentation method yields the most clinically relevant features. This work contributes to the evolving understanding of radiomics in prostate cancer and may help identify the most effective methods for integrating MRI-based feature analysis into clinical practice.

8.2 Methodology

The methodology followed in this chapter is similar to the one followed in Chapter 7, except for the segmentation stage, the dataset size and the radiomics extraction parameters. Automatic index lesion segmentation was performed based on (182), and following the same co-registration framework described in Chapter 6. Unfortunately, it was not possible to perform the quality assessment of the lesion segmentations as had been done for the whole-gland, due to unavailability of the radiologist. The total dataset is composed of 7306 patients (6008 retrospective and 1298 prospective). The label distributions are shown in Tables 8.1 and 8.2, the radiomics extraction parameters can be found in Table 8.3, and the train set sizes are in Tables 8.4 - 8.9.

Table 8.1: Label distribution in the train and test sets for each binary classification problem. RFV - retrospective for validation.

Use Cases	**Target (binary)**	**Train Set**		**Test Set**		**RFV**		**Prospective**	
		0	1	0	1	0	1	0	1
2	ISUP 01 VS 2345	3043	2765	108	92	189	77	703	595
	ISUP 012 VS 345	4354	1454	148	52	177	89	1019	279
3	no metastasis in six months VS metastasis developed	14	60	-	-	-	-	-	-
5	no biochemical recurrence after RP at follow-up VS biochemical recurrence	466	86	84	11	8	-	-	-
6	no biochemical recurrence after RT at follow-up VS biochemical recurrence	89	13	-	-	-	-	-	-
8	stayed in active surveillance VS left active surveillance	92	5	-	-	-	-	1	-

Table 8.2: Label distribution in the train and test sets for each multiclass classification problem. RFV - retrospective for validation.

Use Cases	**Target (multiclass)**	**Train Set**			**Test Set**			**RFV**			**Prospective**		
		0	1	2	0	1	2	0	1	2	0	1	2
7b	EPIC26 [0, 71] vs]71, 84] vs]84, 100]	71	59	53	18	13	12	-	-	-	-	-	-

Table 8.3: Index lesion radiomics extraction parameters.

T2W extraction parameters	**DWI extraction parameters**	**ADC extraction parameters**
imageType: Original: binWidth: 3 Wavelet: binWidth: 3 Square: binWidth: 1 SquareRoot: binWidth: 8 Logarithm: binWidth: 12 Exponential: binWidth: 1 Gradient: binWidth: 4 LBP2D: binWidth: 0.1 LoG: 'sigma' : [1.0, 3.0]	imageType: Original: binWidth: 8 Wavelet: binWidth: 3 Square: binWidth: 4 SquareRoot: binWidth: 8 Logarithm: binWidth: 12 Exponential: binWidth: 3 Gradient: binWidth: 3 LBP2D: binWidth: 0.1 LoG: 'sigma' : [1.0, 3.0]	imageType: Original: binWidth: 3 Wavelet: binWidth: 3 Square: binWidth: 1 SquareRoot: binWidth: 5 Logarithm: binWidth: 5 Exponential: binWidth: 0.5 Gradient: binWidth: 1 LBP2D: binWidth: 0.1 LoG: 'sigma' : [1.0, 3.0]
featureClass: firstorder: glcm: glrlm: glszm: gldm: ngtdm: shape:	featureClass: firstorder: glcm: glrlm: glszm: gldm: ngtdm: shape:	featureClass: firstorder: glcm: glrlm: glszm: gldm: ngtdm: shape:
setting: binWidth: 5 normalize: True normalizeScale: 100 force2D: True voxelArrayShift: 300 resampledPixelSpacing: [0.7031, 0.7031, 0] geometryTolerance: 0.00001	setting: binWidth: 5 normalize: True normalizeScale: 100 force2D: True voxelArrayShift: 300 resampledPixelSpacing: [2, 2, 0] geometryTolerance: 0.00001	setting: binWidth: 5 normalize: True normalizeScale: 100 force2D: True voxelArrayShift: 300 resampledPixelSpacing: [2, 2, 0] geometryTolerance: 0.00001

Table 8.4: Train set sizes for UC2.

UC 2 - Train Sets		sequence			
		T2W	DWI	ADC	T2W& DWI&ADC
scanner	All	5808	5067	5003	4993
	All_noERC	5301	4580	4516	4507
	SIEMENS	2329	2203	2196	2192
	PHILIPS	1934	1593	1582	1578
	GE	1545	1271	1225	1223
	GE_noERC	1038	784	738	737

Table 8.5: Train set sizes for UC3.

UC 3 - Train Sets		sequence			
		T2W	DWI	ADC	T2W& DWI&ADC
scanner	All	74	62	62	62
	All_noERC	71	59	59	59
	SIEMENS	23	23	23	23
	PHILIPS	31	27	27	27
	GE	20	12	12	12
	GE_noERC	17	9	9	9

Table 8.6: Train set sizes for UC5.

UC 5 - Train Sets		sequence			
		T2W	DWI	ADC	T2W& DWI&ADC
scanner	All	552	492	477	475
	All_noERC	531	470	455	454
	SIEMENS	137	126	126	126
	PHILIPS	340	310	308	307
	GE	75	56	43	42
	GE_noERC	54	34	21	21

Table 8.7: Train set sizes for UC6.

UC 6 - Train Sets		sequence			
		T2W	DWI	ADC	T2W& DWI&ADC
scanner	All	102	96	95	95
	All_noERC	62	56	55	55
	SIEMENS	4	4	4	4
	PHILIPS	55	50	50	50
	GE	43	42	41	41
	GE_noERC	3	2	1	1

Table 8.8: Train set sizes for UC7.

UC 7b - Train Sets		sequence			
		T2W	DWI	ADC	T2W& DWI&ADC
scanner	All	226	209	207	207
	All_noERC	226	209	207	207
	SIEMENS	51	45	45	45
	PHILIPS	174	163	161	161
	GE	1	1	1	1
	GE_noERC	1	1	1	1

Table 8.9: Train set sizes for UC8.

UC 8 - Train Sets		sequence			
		T2W	DWI	ADC	T2W& DWI&ADC
scanner	All	97	91	91	91
	All_noERC	93	87	87	87
	SIEMENS	63	63	63	63
	PHILIPS	10	5	5	5
	GE	24	23	23	23
	GE_noERC	20	19	19	19

8.3 Results

8.3.1 Use Case 2 - ISUP {0,1} vs {2345}

8.3.1.1 Model Selection

Figure 8.1 shows the cross-validation and hold-out test set ROC-AUC model performance for the 96 models trained for UC2 (ISUP 0, 1 vs 2, 3, 4 & 5). It can be seen, at first glance, that the inclusion of clinical information is beneficial for the model's overall performance. The exclusion of endorectal coil patients does not seem to have much impact. Overall, the highest hold-out test set performance was 0.9043 AUC with a model trained with radclin data extracted from T2W, DWI and ADC volumes from all scanner manufacturers and excluding endorectal-coil patients. This model was selected for further analysis.

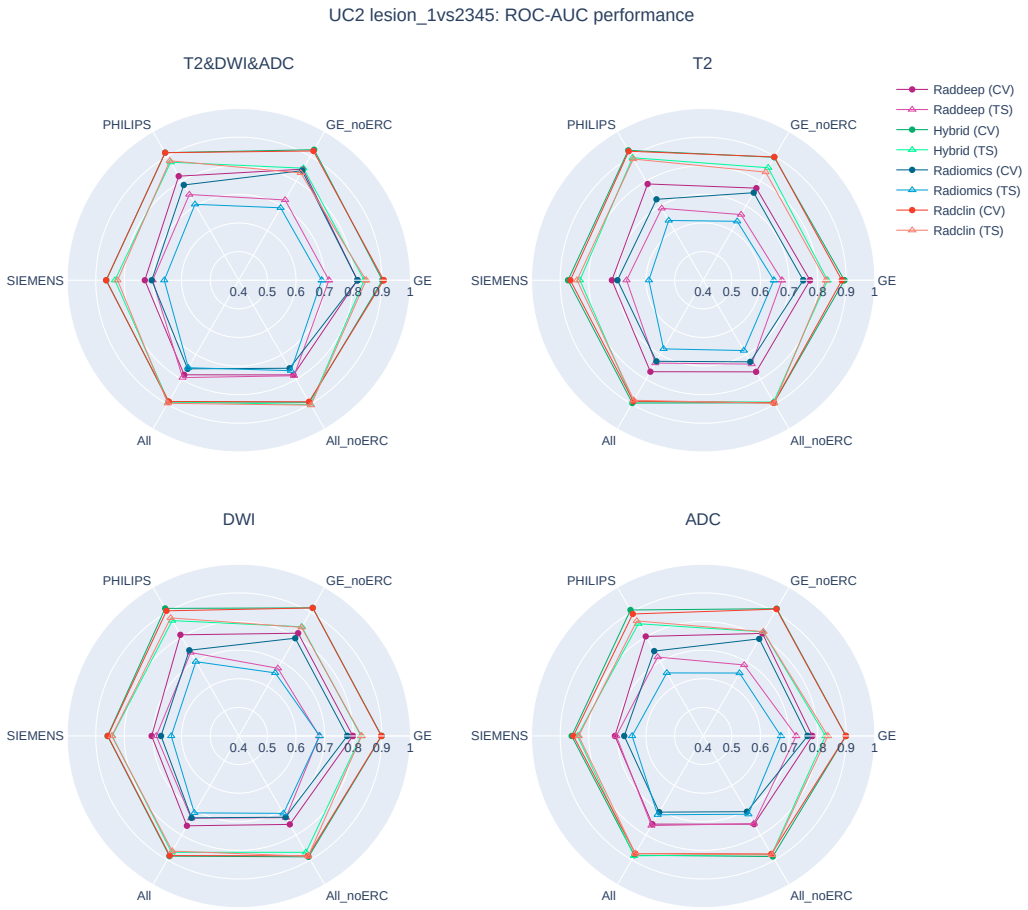


Figure 8.1: Cross-validation and hold-out test set ROC-AUC model performance of 48 models trained to predict lesion aggressiveness in UC2 (ISUP 1 vs 2, 3, 4 & 5). The observations are color coded according to the type of data used in training: radiomics in blue and radclin in orange.

8.3.1.2 Model Validation

Given the results Section 8.3.1.1, the model radclin_T2W&DWI&ADC_noERC_CatBoost was further analysed. Upon inspection of the ROC curve for the hold-out test set (Figure 8.2), it was decided that the appropriate probability decision threshold would be 0.5392, above which the model gives an output of clinical significance. The model's performance at this probability decision threshold is shown in Table 8.10 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Figure 8.3.

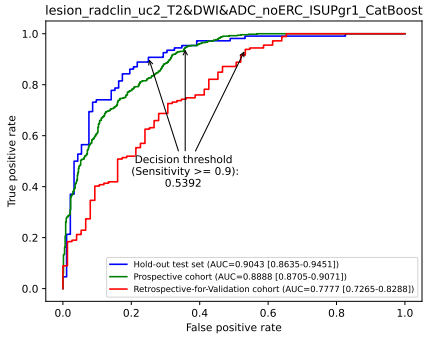


Figure 8.2: Receiver operator characteristics curve of the radclin_uc2_T2&DWI&ADC_no ERC_CatBoost model on the hold-out and prospective cohorts.

Table 8.10: Multi-metric performance of the radclin_uc2_T2&DWI&ADC_noERC_CatBoost model at 0.5392 probability threshold.

	Hold-out test set	Retrospective for Validation	Prospective cohort
AUC	0.9043 [0.8635-0.9451]	0.7777 [0.7265-0.8288]	0.8888 [0.8705-0.9071]
Sensitivity/Recall/TPR	0.9074 [0.8672-0.9476]	0.933 [0.9022-0.9637]	0.9462 [0.933-0.9593]
Specificity/TNR	0.75 [0.69-0.81]	0.4667 [0.4053-0.528]	0.6407 [0.6127-0.6686]
Precision/PPV	0.8099 [0.7555-0.8643]	0.8068 [0.7582-0.8553]	0.69 [0.6631-0.7169]
F1	0.8559 [0.8072-0.9046]	0.8653 [0.8233-0.9073]	0.7981 [0.7747-0.8214]
F2	0.8861 [0.842-0.9301]	0.9047 [0.8685-0.9408]	0.8808 [0.8619-0.8996]
CohensKappa	0.6644 [0.5989-0.9046]	0.4482 [0.3871-0.9073]	0.5705 [0.5417-0.8214]

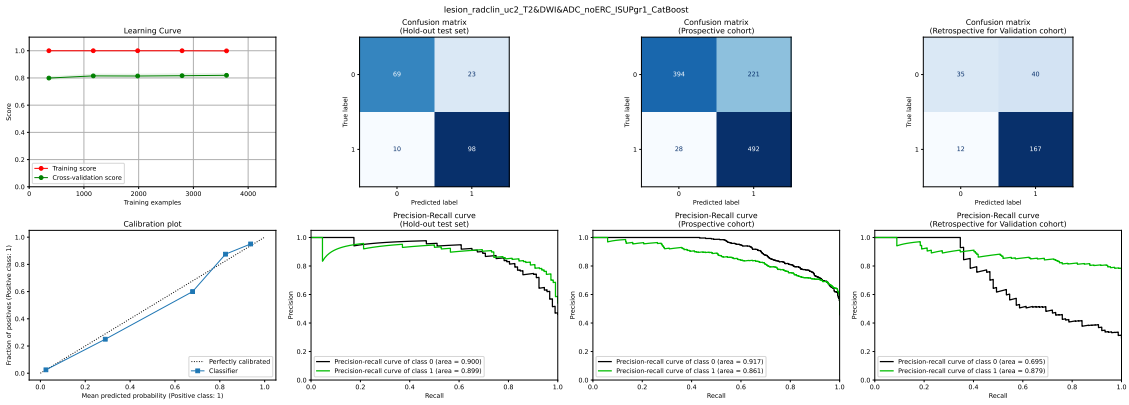


Figure 8.3: Analysis of the radclin_uc2_T2&DWI&ADC_noERC_CatBoost model in terms of confusion matrix, learning curve, precision-recall curve and calibration plot.

8.3.1.3 Explainability analysis

Figure 8.4 shows the SHAP analysis performed on the radclin_uc2_T2&DWI&ADC_noERC_CatBoost model. The most impactful feature is the PI-RADS score (Figure 8.4), reflecting the radiologist's assessment of the sequence. PI-RADS 4 and 5 are positively associated with aggressiveness, while all others are associated with non-aggressiveness (Figure 8.5). Also showing a positive association with the aggressive output is the PSA value, the age and the index lesion location in the peripheral or tranzitional zones.

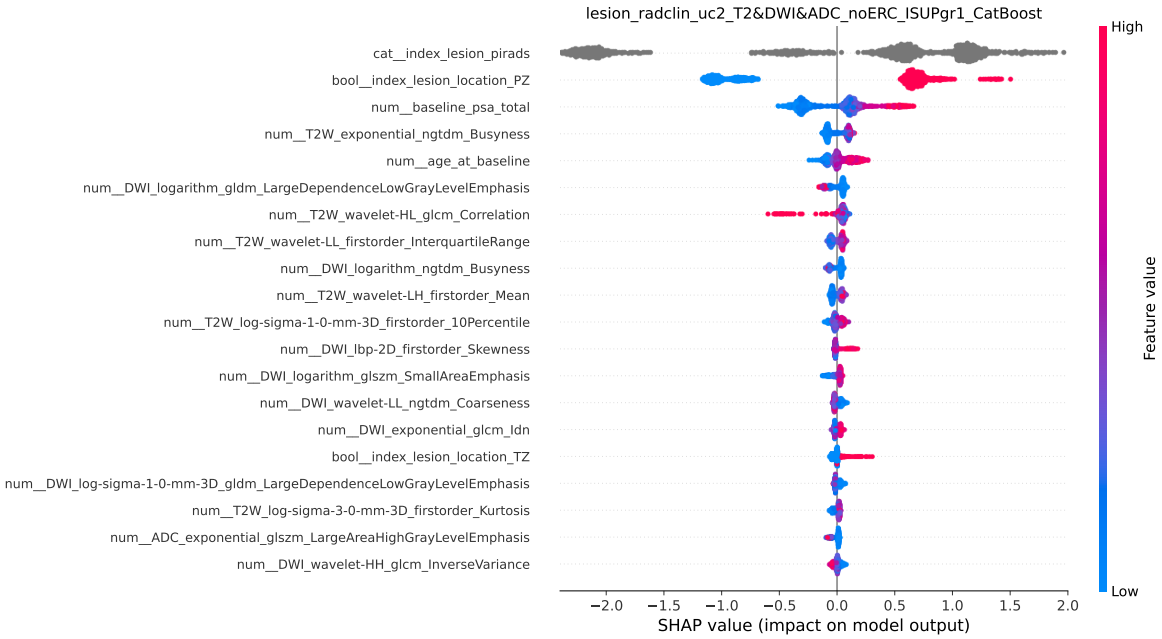


Figure 8.4: SHAP values contribution of features for radclin_uc2_T2&DWI&ADC_noERC_CatBoost model.

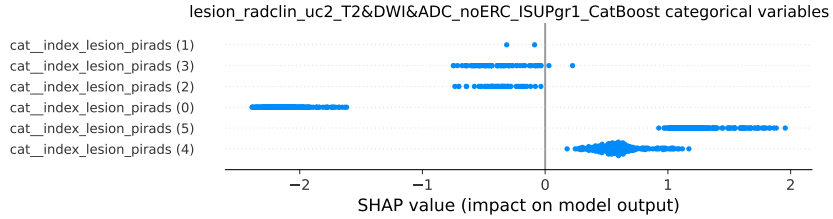


Figure 8.5: SHAP values contribution of the categorical features for radclin_uc2_T2&DWI&ADC_noERC_CatBoost model.

8.3.1.4 Fairness and sub-cohort analysis

Figure 8.6 shows the model's performance on different subsets of the prospective cohort. The model shows robust performance across scanner manufacturers (top left plot) and endorectal coil usage (top right plot).

Fairness analysis of model lesion_radclin_uc2_T2&DWI&ADC_noERC_ISUPgr1_CatBoost on the Prospective cohort

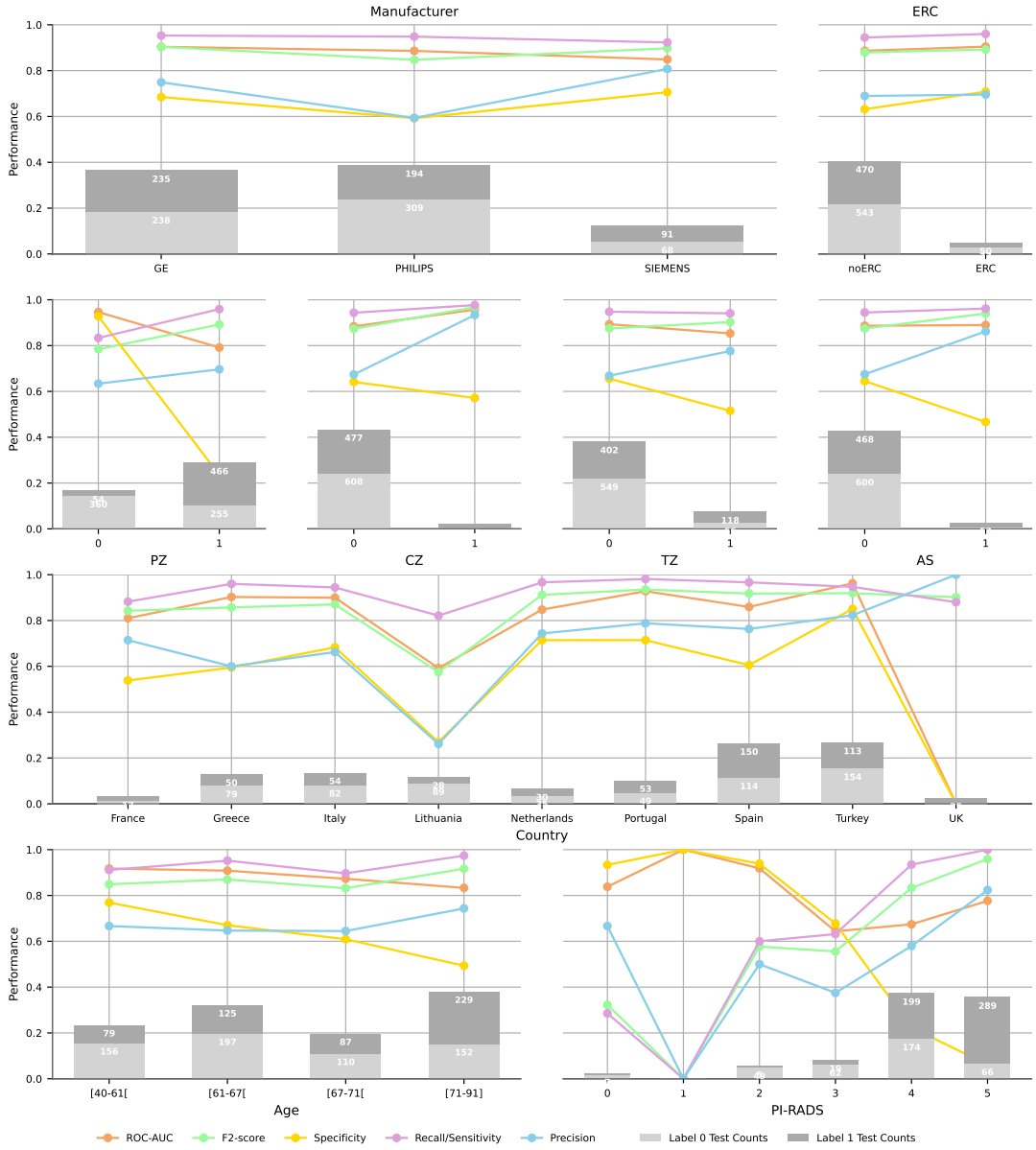


Figure 8.6: Performance of the model radclin_uc2_T2&DWI&ADC_noERC_CatBoost on sub cohorts of the prospective cohort. The plots are colour coded according to metric: accuracy in orange, F2-score in green, precision in blue and recall in purple. Each plot also displays stacked bars indicating the subcohort size and label distribution. The four plots in the second row refer to the presence (1) or absence (0) of index lesion in each anatomical zone (PZ, TZ, CZ and AS).

Regarding lesion location (Figure 8.6, second row), there is a rise in sensitivity when the index lesion is located in the peripheral zone, but this is accompanied by a reduction in specificity. A similar drop in specificity is observed when the index lesion is located in the tranzitional zone. Regarding the central zone and anterior stroma, there were not enough samples in the minority cohorts for accurate conclusions to be drawn.

Regarding country of origin (third row), sensitivity remains consistent while specificity is more variable. The lowest performance is reached with cases from Lithuania and the UK. In terms of patient age, specificity seems to decrease as patients get older. Regarding PI-RADS, the dataset seems unfortunately biased towards high PI-RADS cases, nevertheless, there is a clear trend of increased sensitivity and decreased specificity as PI-RADS increases.

8.3.2 Use Case 2 - ISUP {0,1,2} vs {3,4,5}

8.3.2.1 Model Selection

Figure 8.7 shows the cross-validation and hold-out test set ROC-AUC model performance for the 96 models trained for UC2 (ISUP 0, 1 & 2 vs 3, 4 & 5). In agreement with Section 8.3.1 (ISUP={0,1} vs ISUP={2,3,4,5}), the inclusion of clinical information is beneficial for the model's generalizability to the hold-out test set and the exclusion of endorectal coil patients does not show much impact on the performance. The scanner manufacturer reaching the highest performance is Philips (0.8450 AUC), though it still does not outperform the model trained with all scanners combined (0.8470 AUC). The latter was trained with radclin data using only T2 sequences. This model was selected for further analysis.

8.3.2.2 Model Validation

Given the results of Section 8.3.2.1, the model radclin_T2_noERC_CatBoost was further analysed. Upon inspection of the ROC curve for the hold-out test set (Figure 8.8), it was decided that the appropriate probability decision threshold would be 0.1902, above which the model gives an output of clinical significance. The model's performance at this probability decision threshold is shown in Table 8.11 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Figure 8.9.

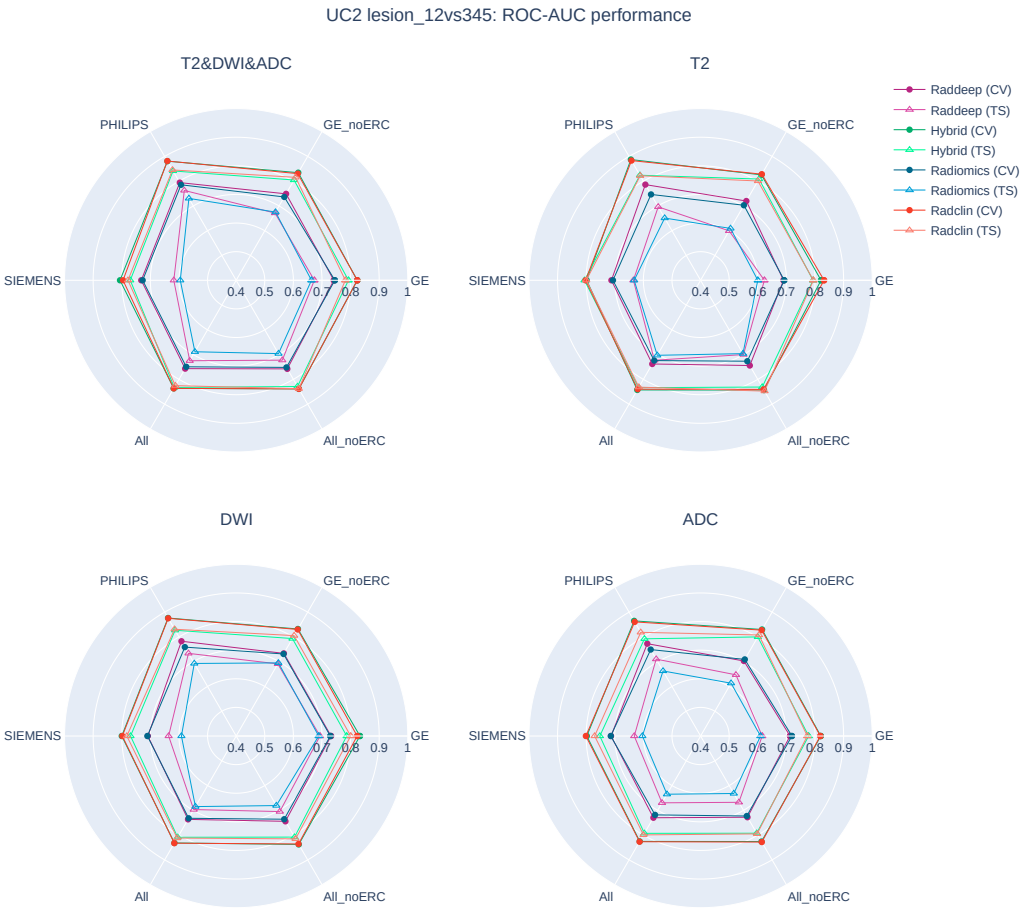


Figure 8.7: Cross-validation and hold-out test set ROC-AUC model performance of 48 models trained to predict lesion aggressiveness in UC2 (ISUP 1 & 2 vs 3, 4 & 5). The observations are color coded according to the type of data used in training: radiomics in blue and radclin in orange.

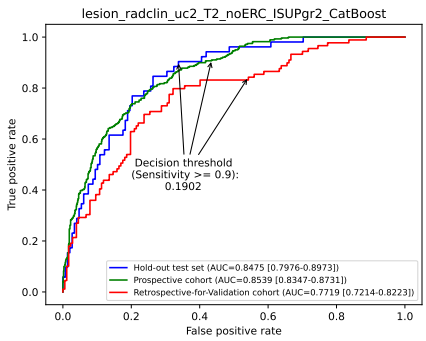


Figure 8.8: Receiver operator characteristics curve of the radclin_uc2_T2_noERC_CatBoost model on the hold-out and prospective cohorts.

Table 8.11: Multi-metric performance of the radclin_uc2_T2_noERC_CatBoost model at 0.1902 probability threshold.

	Hold-out test set	Retrospective for Validation	Prospective cohort
AUC	0.8475 [0.7976-0.8973]	0.7719 [0.7214-0.8223]	0.8539 [0.8347-0.8731]
Sensitivity/Recall/TPR	0.9038 [0.8630-0.9447]	0.8315 [0.7865-0.8764]	0.9104 [0.8949-0.9259]
Specificity/TNR	0.6622 [0.5966-0.7277]	0.4576 [0.3978-0.5175]	0.5633 [0.5363-0.5903]
Precision/PPV	0.4845 [0.4153-0.5538]	0.4353 [0.3757-0.4949]	0.3634 [0.3372-0.3895]
F1	0.6309 [0.5640-0.6978]	0.5714 [0.5120-0.6309]	0.5194 [0.4922-0.5466]
F2	0.7705 [0.7122-0.8288]	0.7034 [0.6485-0.7583]	0.6997 [0.6748-0.7247]
CohensKappa	0.4420 [0.3731-0.6978]	0.2358 [0.1847-0.6309]	0.3063 [0.2812-0.5466]

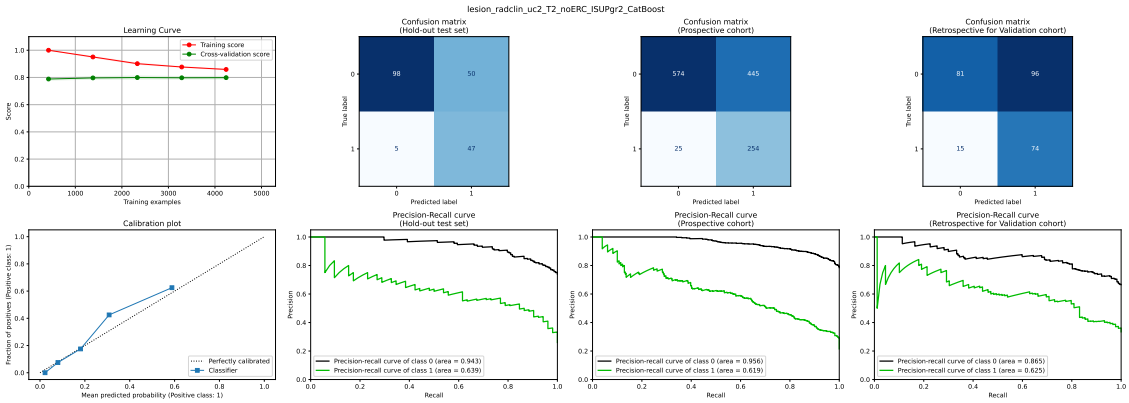


Figure 8.9: Analysis of the radclin_uc2_T2_noERC_CatBoost model in terms of confusion matrix, learning curve, precision-recall curve and calibration plot.

8.3.2.3 Explainability analysis

Figure 8.10 shows the SHAP analysis performed on the radclin_T2_noERC_CatBoost model. The most impactful feature is PI-RADS, followed by whether the index lesion was located in the peripheral zone, the PSA blood level, and patient age, all positively associated with an aggressive output. In concordance with Section 8.3.1 (ISUP=1 VS ISUP=2345), PI-RADS 4 and 5 are positively associated with aggressiveness, while all others are associated with non-aggressiveness (Figure 8.11). Also showing a positive association with the aggressive output is the lesion shape in terms of flatness.

8.3.2.4 Fairness and sub-cohort analysis

Figure 8.12 shows the model's performance on different subsets of the prospective cohort. The model shows a reduction in sensitivity from GE to Philips to Siemens scanners (top left plot), and a reduction in specificity from Philips to Siemens and GE. There is also an increase in overall performance for exams taken with an endorectal coil (top right plot).

Regarding lesion location (Figure 8.12, second row), when compared with Section 8.3.1 (ISUP {0,1} vs {2,3,4,5}) similar trends arise. There is an increased sensitivity and decreased specificity

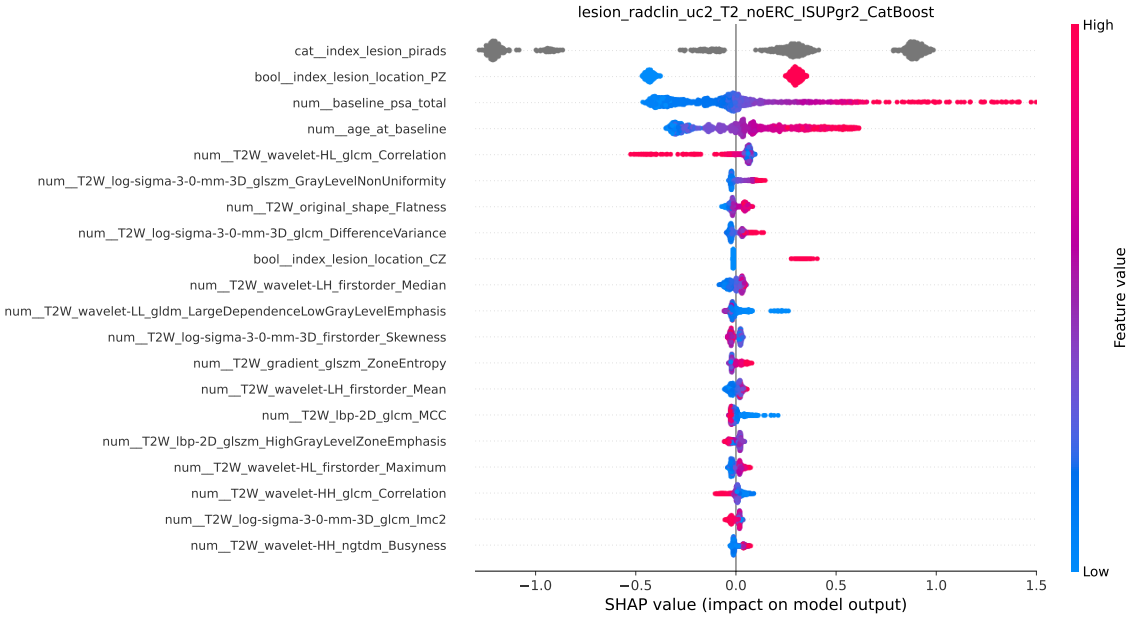


Figure 8.10: SHAP values contribution of features for radclin_uc2_DWI_ISUPgr2_CatBoost model.

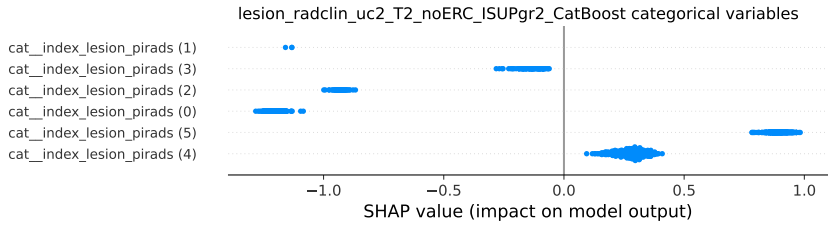


Figure 8.11: SHAP values contribution of the categorical features for radclin_uc2_DWI_ISUPgr2_CatBoost model.

when the index lesion is located in the peripheral or tranzitional zones. However there are not enough samples in the central zone or anterior stroma cohorts for accurate conclusions to be drawn.

Regarding country of origin (third row), the highest performance is achieved with data from the Netherlands, Portugal, Turkey, and Italy. Whereas the lowest performance is reached with cases from Lithuania, Spain and France. In terms of patient age, the model sensitivity seems to increase as patients get older, while specificity tends to decrease. Regarding PI-RADS, the model sensitivity seems to increase as the score increases, while specificity decreases.

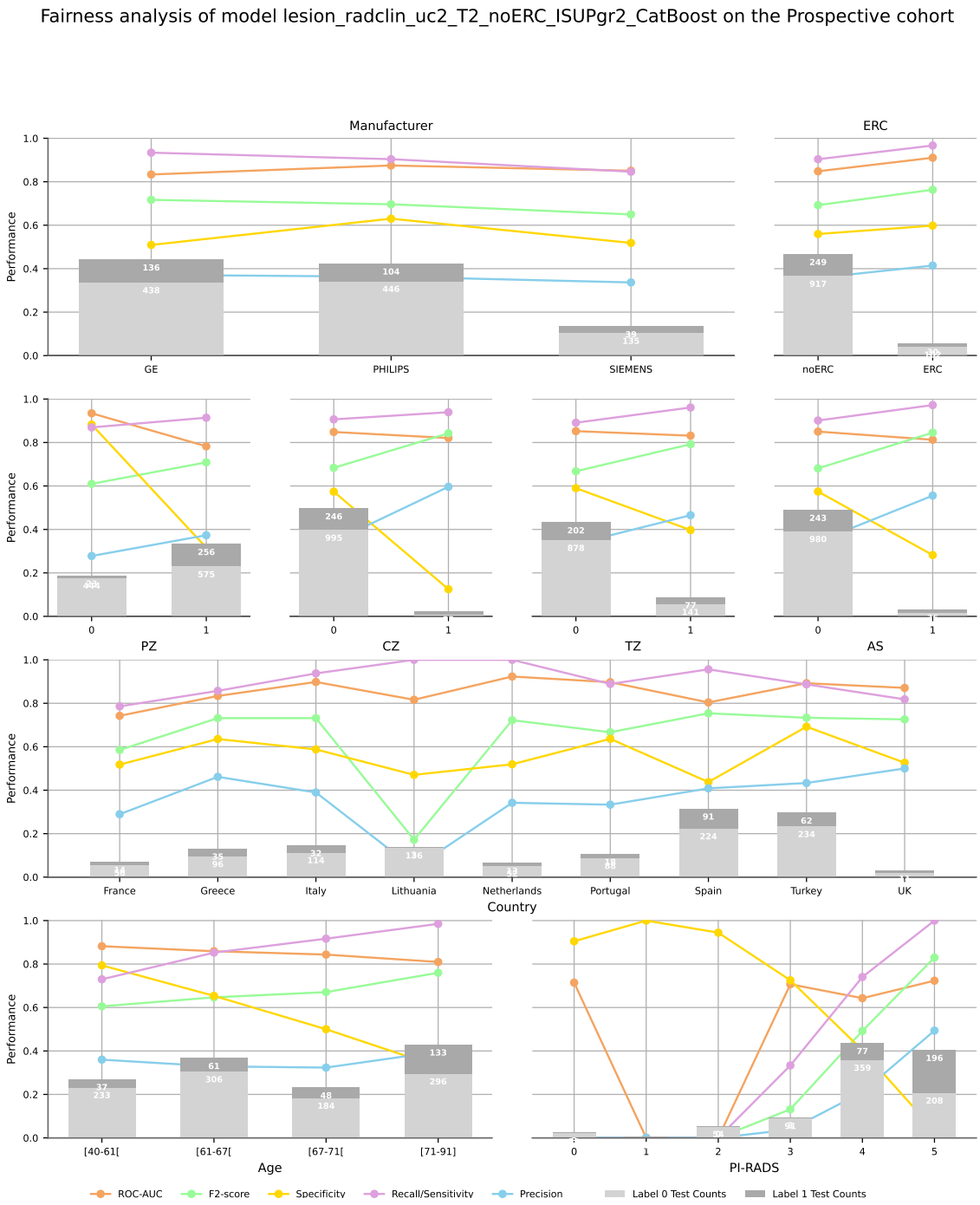


Figure 8.12: Performance of the model radclin_uc2_DWI_ISUPgr2_CatBoost on sub cohorts of the prospective cohort. The plots are colour coded according to metric: accuracy in orange, F2-score in green, precision in blue and recall in purple. Each plot also displays stacked bars indicating the subcohort size and label distribution. The four plots in the second row refer to the presence (1) or absence (0) of index lesion in each anatomical zone (PZ, TZ, CZ and AS).

8.3.3 Use Case 3 - Development of metastasis

8.3.3.1 Model Selection

Figure 8.13 shows the cross-validation ROC-AUC model performance for the 20 models trained for UC3. The highest cross-validation performance was reached with the model radclin_uc3_DWI _SGD (0.8731 AUC).

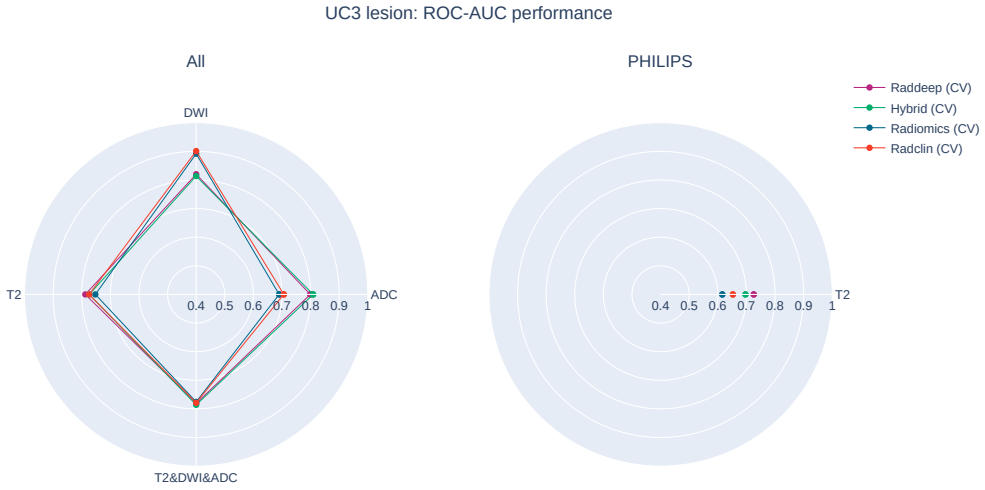


Figure 8.13: Cross-validation ROC-AUC model performance of 20 models trained to predict the development of metastasis in UC3. The observations are color coded according to the type of data used in training: radiomics in blue and radclin in orange.

8.3.3.2 Model Validation

Given the results of the last section, it was decided to move forward analyzing a raddeep model using DWI sequences and cases coming from a PHILIPS scanner. The model's performance at a 0.77 (Figure 8.14) probability decision threshold is shown in Table 8.12 and its confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Figure 8.15.

8.3.3.3 Explainability analysis

Figure 8.16 shows the logistic regression coefficients of the model radclin_uc3_DWI_SGD. The most impactful feature is the PI-RADS score, followed by the PZ lesion location, texture indicators and TZ lesion location.

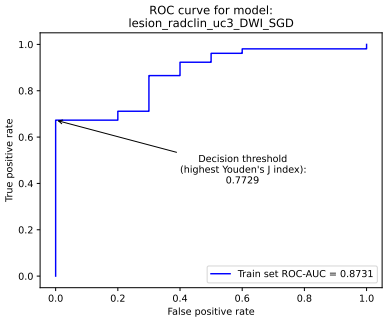


Figure 8.14: Receiver operator characteristics curve of the radclin_uc3_DWI_SGD model on the hold-out and prospective cohorts.

Table 8.12: Multi-metric performance of the radclin_uc3_DWI_SGD model at 0.5099 probability threshold. 95% confidence intervals are shown in brackets.

	Cross-validatiom
AUC	0.8731 [0.7902-0.9559]
Sensitivity/Recall/TPR	0.6731 [0.5563-0.7898]
Specificity/TNR	1.0 [1.0-1.0]
Precision/PPV	1.0 [1.0-1.0]
F1	0.8046 [0.7059-0.9033]
F2	0.7202 [0.6084-0.8319]
CohensKappa	0.3991 [0.2772-0.9033]

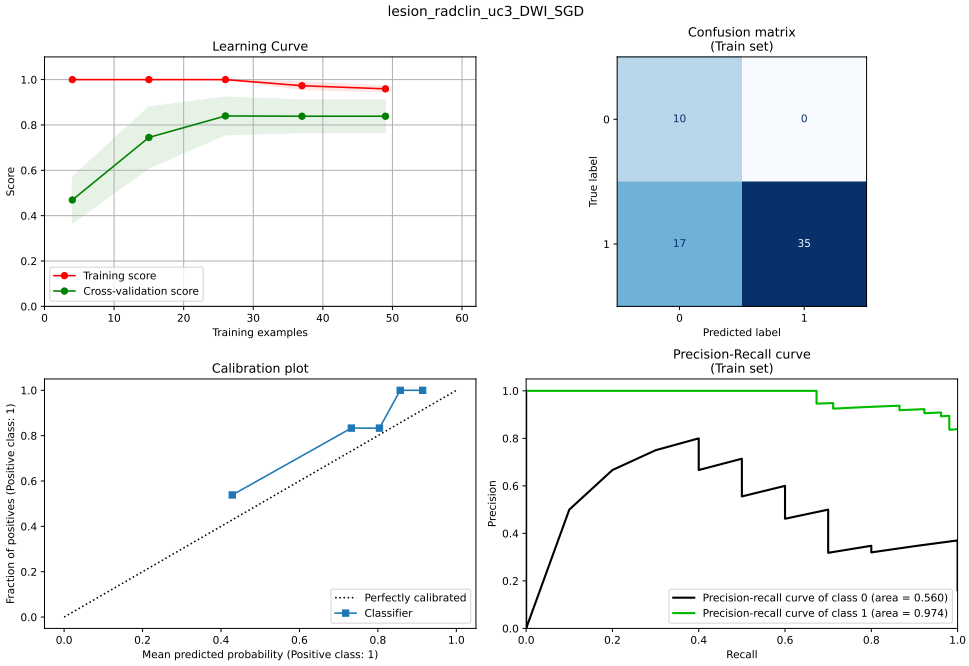


Figure 8.15: Analysis of the radclin_uc3_DWI_SGD model in terms of confusion matrix, learning curve, precision-recall curve and calibration plot.

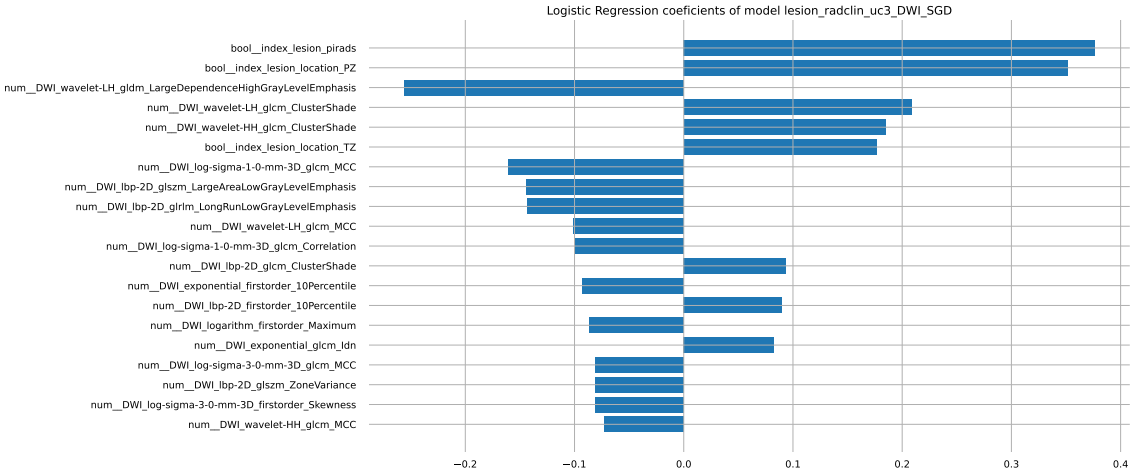


Figure 8.16: Logistic regression coefficients of 20 most impactful features for model radclin_uc3_DWI_SGD.

8.3.4 Use Case 5 - BCR pre-RP

8.3.4.1 Model Selection

Figure 8.17 shows the cross-validation and hold-out test set ROC-AUC model performance for the 88 models trained for UC5 (pre-surgery prediction). Overall, the inclusion of clinical variables outperforms radiomic features alone. The classifier that was selected for further analysis was a radclin model trained with T2W sequences and all scanners (0.7294 AUC on the hold-out test set).



Figure 8.17: Cross-validation and hold-out test set ROC-AUC model performance of 88 models trained to predict biochemical recurrence in UC5 (pre-RP prediction). The observations are color coded according to the type of data used in training: radiomics in blue and radclin in orange.

8.3.4.2 Model Validation

The model's ROC curve (Figure 8.18) was used for the selection of the probability decision threshold of 0.3101, and the respective performance at this threshold is shown in Table 8.13. The confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Figure 8.19.

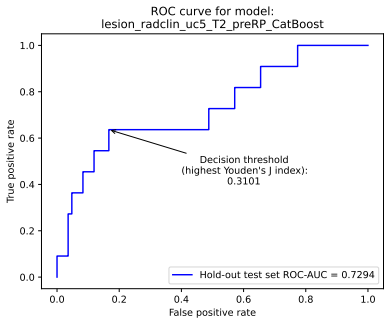


Figure 8.18: Receiver operator characteristics curve of the radclin_uc5_T2_CatBoost model on the hold-out and prospective cohorts.

Table 8.13: Multi-metric performance of the radclin_uc5_T2_CatBoost model at 0.3101 probability threshold. 95% confidence intervals are shown in brackets.

	Hold-out test set
AUC	0.7294 [0.6401-0.8188]
Sensitivity/Recall/TPR	0.6364 [0.5396-0.7331]
Specificity/TNR	0.8333 [0.7584-0.9083]
Precision/PPV	0.3333 [0.2385-0.4281]
F1	0.4375 [0.3377-0.5373]
F2	0.5385 [0.4382-0.6387]
CohensKappa	0.3367 [0.2417-0.5373]

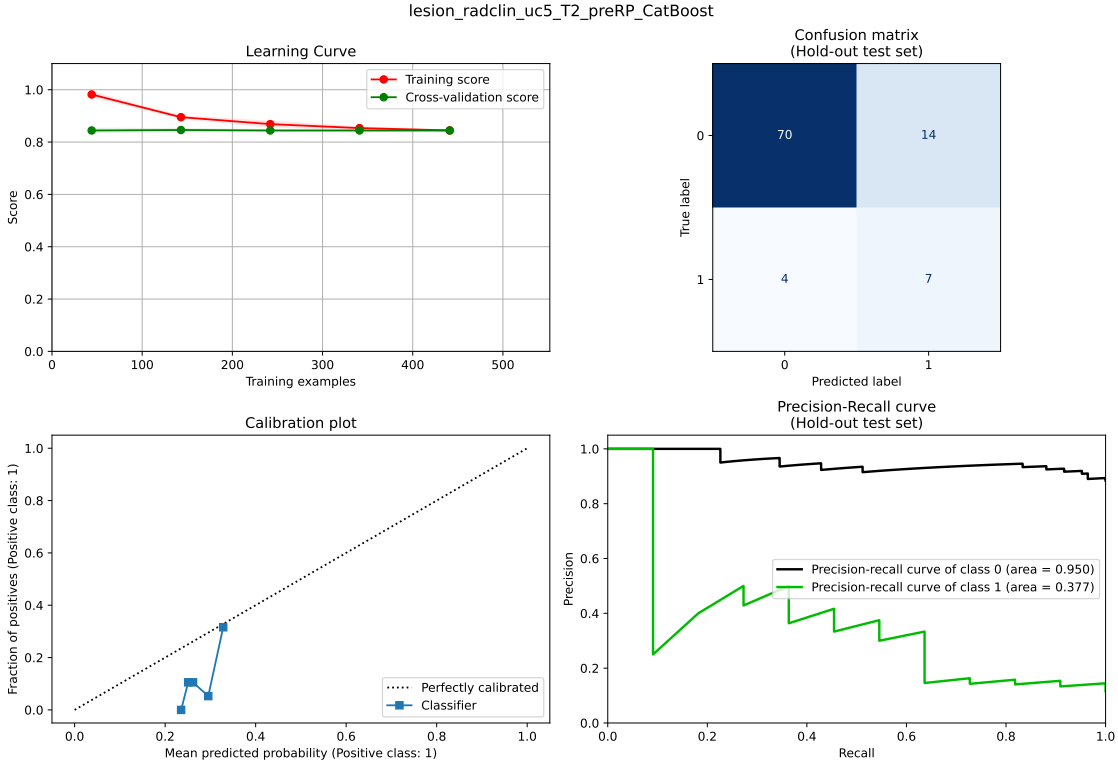


Figure 8.19: Analysis of the radclin_uc5_T2_CatBoost model in terms of confusion matrix, learning curve, precision-recall curve and calibration plot.

8.3.4.3 Explainability analysis

Figure 8.20 shows the SHAP analysis performed on the radclin_uc5_T2_CatBoost model. The model relies heavily on clinical information to predict recurrence, namely the gleason score of the most prominent histological pattern, the ISUP grade the baseline PSA, and PI-RADS, all showing a positive association with the outcome (higher values present with a higher risk of recurrence).

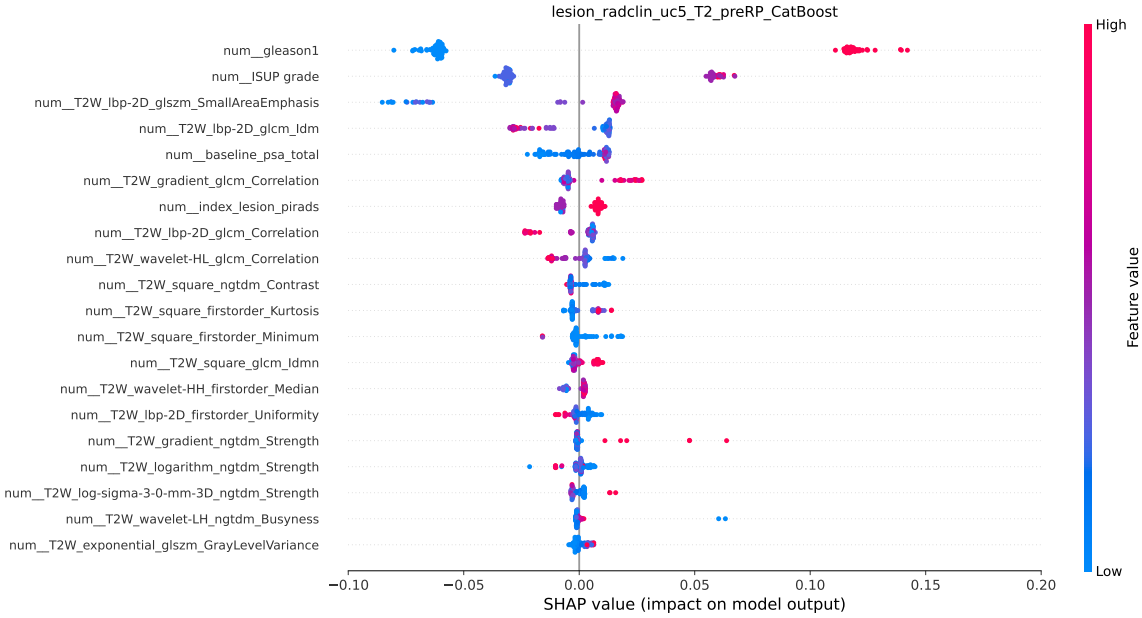


Figure 8.20: SHAP values contribution of features for radclin_uc5_T2_CatBoost model.

8.3.5 Use Case 5 - BCR post-RP

8.3.5.1 Model Selection

Figure 8.21 shows the cross-validation and hold-out test set ROC-AUC model performance for the 88 models trained for UC5 (post-surgery prediction). In concordance with Section 8.3.4, the inclusion of clinical variables seems to outperform radiomic features alone or in combination with deep features. The classifier that was selected for further analysis was a radclin model trained with T2W sequences and all scanners (0.6948 AUC on the hold-out test set).



Figure 8.21: Cross-validation and hold-out test set ROC-AUC model performance of 88 models trained to predict biochemical recurrence in UC5 (pre-RP prediction). The observations are color coded according to the type of data used in training: radiomics in blue and radclin in orange.

8.3.5.2 Model Validation

The model's ROC curve, used for the selection of the probability decision threshold of 0.0749, and the respective performance at this threshold are shown in Figure 8.22 and Table 8.14.

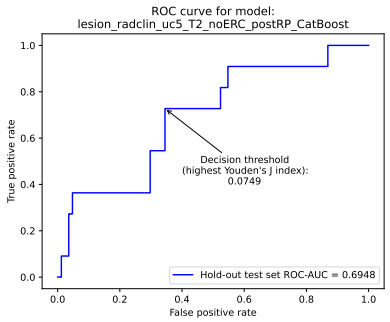


Figure 8.22: Receiver operator characteristics curve of the radclin_uc5_T2_noERC_CatBoost model on the hold-out and prospective cohorts.

Table 8.14: Multi-metric performance of the radclin_uc5_T2_noERC_CatBoost model at 0.0749 probability threshold. 95% confidence intervals are shown in brackets.

	Hold-out test set
AUC	0.6948 [0.6022-0.7874]
Sensitivity/Recall/TPR	0.7273 [0.6377-0.8168]
Specificity/TNR	0.6548 [0.5592-0.7504]
Precision/PPV	0.2162 [0.1334-0.2990]
F1	0.3333 [0.2385-0.4281]
F2	0.4938 [0.3933-0.5944]
CohensKappa	0.1885 [0.1098-0.4281]

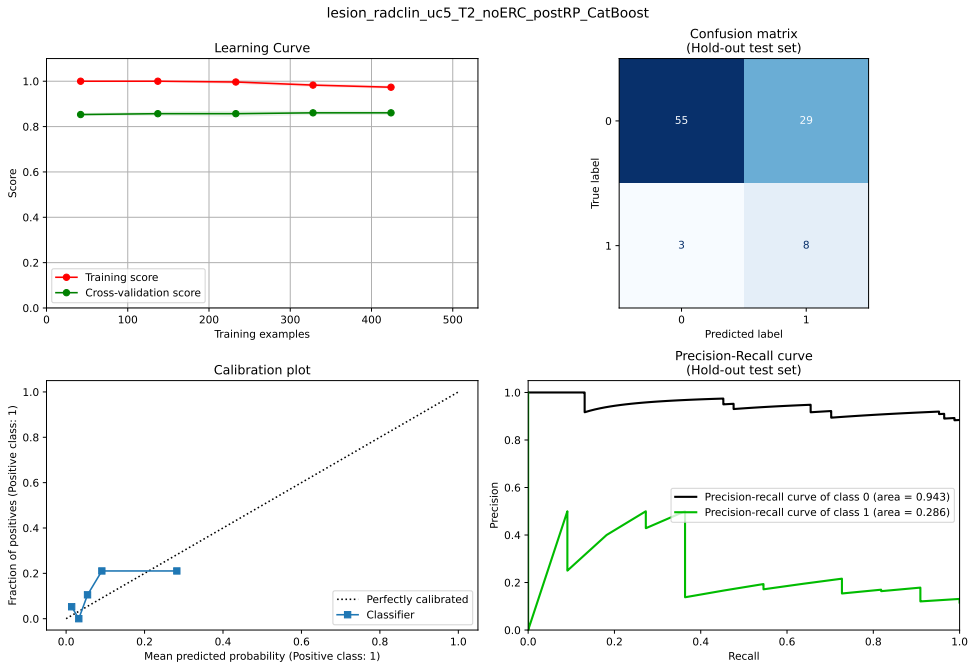


Figure 8.23: Analysis of the radclin_uc5_T2_noERC_CatBoost model in terms of confusion matrix, learning curve, precision-recall curve and calibration plot.

8.3.5.3 Explainability analysis

Figures 8.24 and 8.25 shows the SHAP analysis performed on the radclin_uc5_T2_noERC _CatBoost model. The model relies heavily on clinical information to predict recurrence, namely the presence of extra-prostatic extension, higher values of PSA, presence of perineural invasion, gleason score of the most prominent histological pattern (gleason1) or presence of positive resection margins. Interestingly, in terms of surgical method, the model is associating laparoscopic surgeries with a higher risk of recurrence.

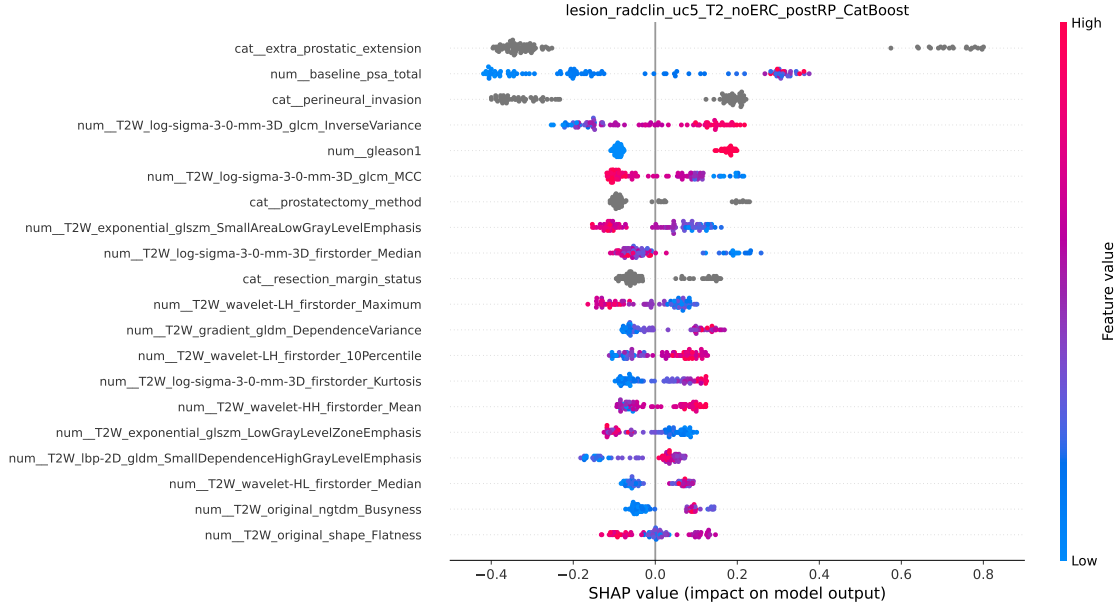


Figure 8.24: SHAP values contribution of features for radclin_uc5_T2_noERC_CatBoost model.

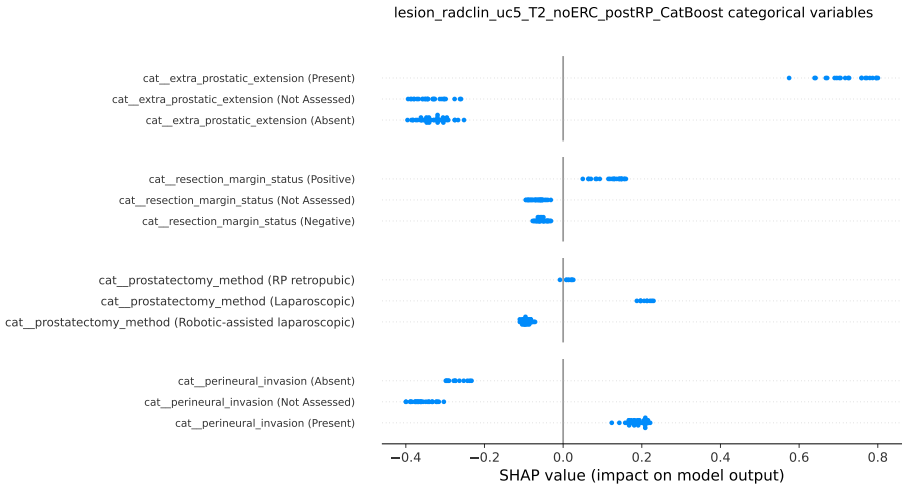


Figure 8.25: SHAP values contribution of the categorical features for radclin_uc5_T2_noERC_CatBoost model.

8.3.6 Use Case 6 - BCR post-RT

8.3.6.1 Model Selection

Figure 8.26 shows the cross-validation ROC-AUC model performance for the 32 models trained for UC6. The highest cross-validation performance was reached with a radiomics model trained with DWI exams coming from GE scanners (0.6806 AUC).

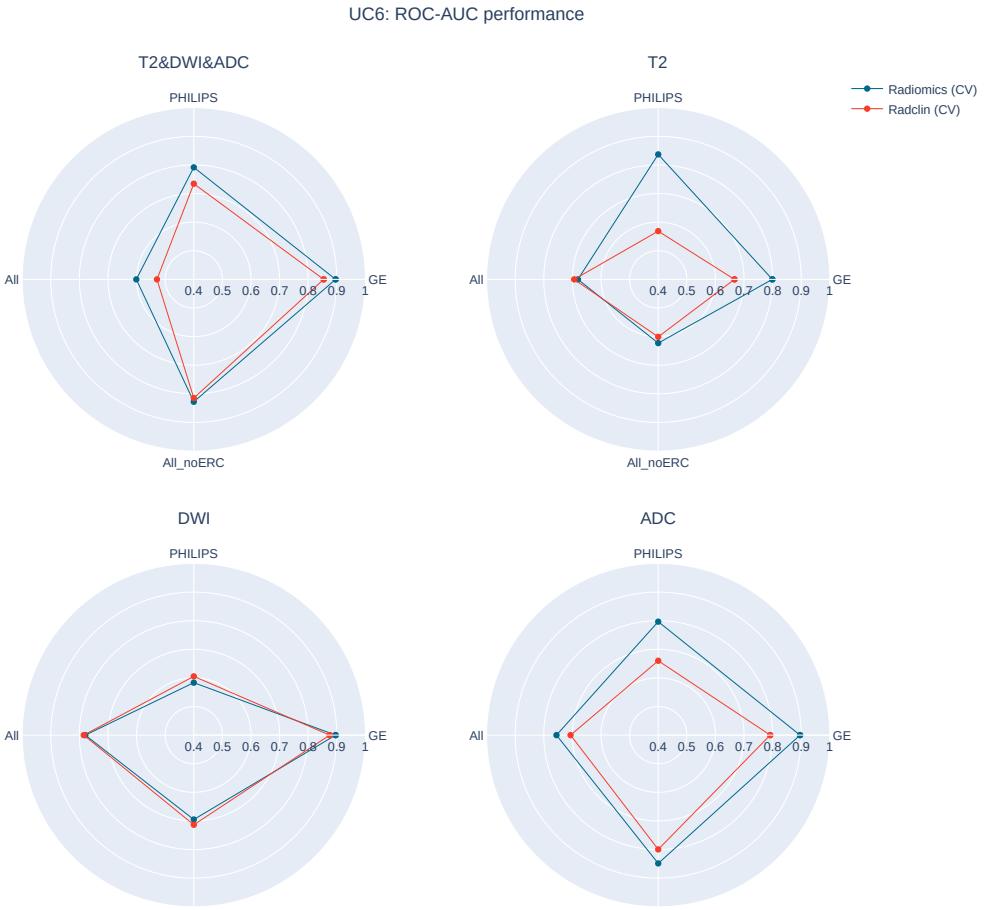


Figure 8.26: Cross-validation ROC-AUC model performance of 32 models trained to predict biochemical recurrence in UC6. The observations are color coded according to the type of data used in training: radiomics in blue and radclin in orange.

8.3.6.2 Model Validation

The model's ROC curve, used for the selection of the probability decision threshold of 0.1048, and the respective performance at this threshold are shown in Figure 8.27 and Table 8.15. No prospective cases were available for this use case. The confusion matrix, learning curve, precision-recall curve and calibration plot are displayed in Figure 8.28.

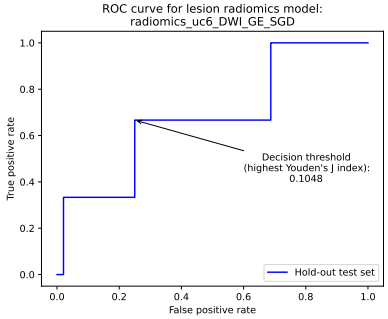


Figure 8.27: Receiver operator characteristics curve of the radiomics_uc6_DWI_GE_SGD model on cross-validation.

Table 8.15: Multi-metric performance of the radiomics_uc6_DWI_GE_SGD model at 0.1048 probability threshold. 95% confidence intervals are shown in brackets.

	Cross-validatiom
AUC	0.6806 [0.5526-0.8085]
Sensitivity/Recall/TPR	0.6667 [0.5373-0.7960]
Specificity/TNR	0.7500 [0.6312-0.8688]
Precision/PPV	0.1429 [0.0468-0.2389]
F1	0.2353 [0.1189-0.3517]
F2	0.3846 [0.2511-0.5181]
CohensKappa	0.1533 [0.0544-0.3517]

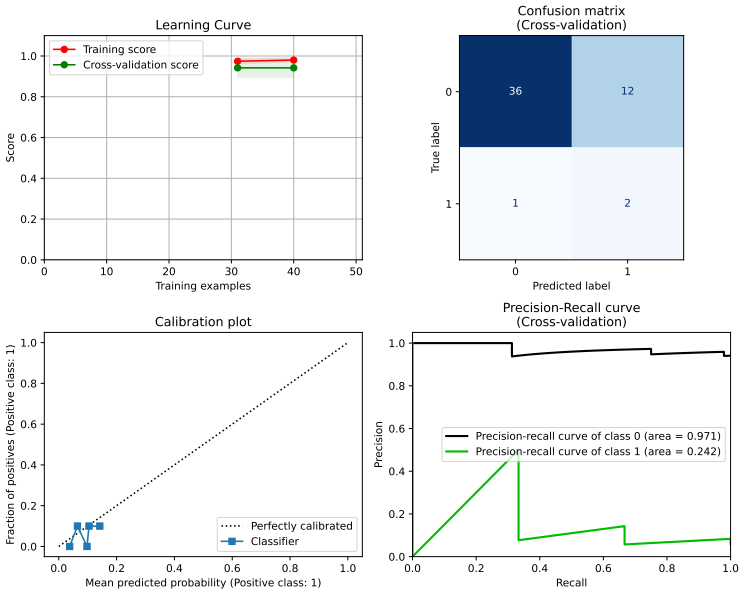


Figure 8.28: Analysis of the radiomics_uc6_DWI_GE_SGD model in terms of confusion matrix, learning curve, precision-recall curve and calibration plot.

8.3.6.3 Explainability analysis

Figure 8.29 shows the logistic regression coefficients of the model radclin_uc6_DWI_GE_SGD. The most impactful features are texture and first-order type radiomic features, extracted from log-sigma and wavelet transforms.

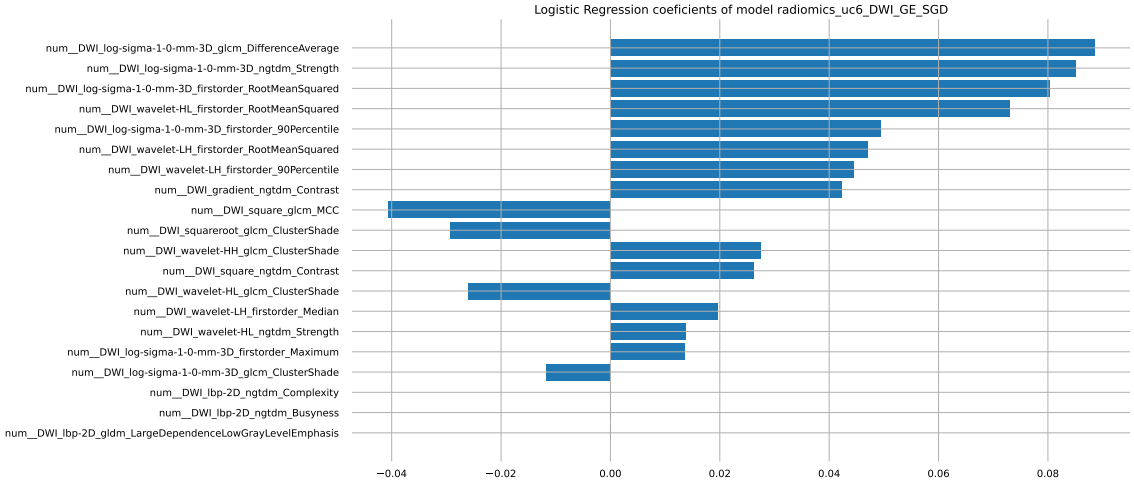


Figure 8.29: Logistic regression coefficients of 20 most impactful features for model radiomics_uc6_DWI_GE_SGD.

Due to the small dataset size, the fairness analysis was not conducted since any subcohort would not be representative of the source population, and it would be difficult to draw conclusions.

8.3.7 Use Case 7b

8.3.7.1 Model Selection

Figure 8.30 shows the cross-validation ROC-AUC model performance for the 48 models trained for UC7b. In concordance with Chapter 7, the highest cross-validation performances are achieved with SIEMENS data, however these are highly overfitted. Overall, the model that shows the highest performance is a radiomics model trained with ADC sequences and all scanner manufacturers (0.60 one-vs-rest AUC).

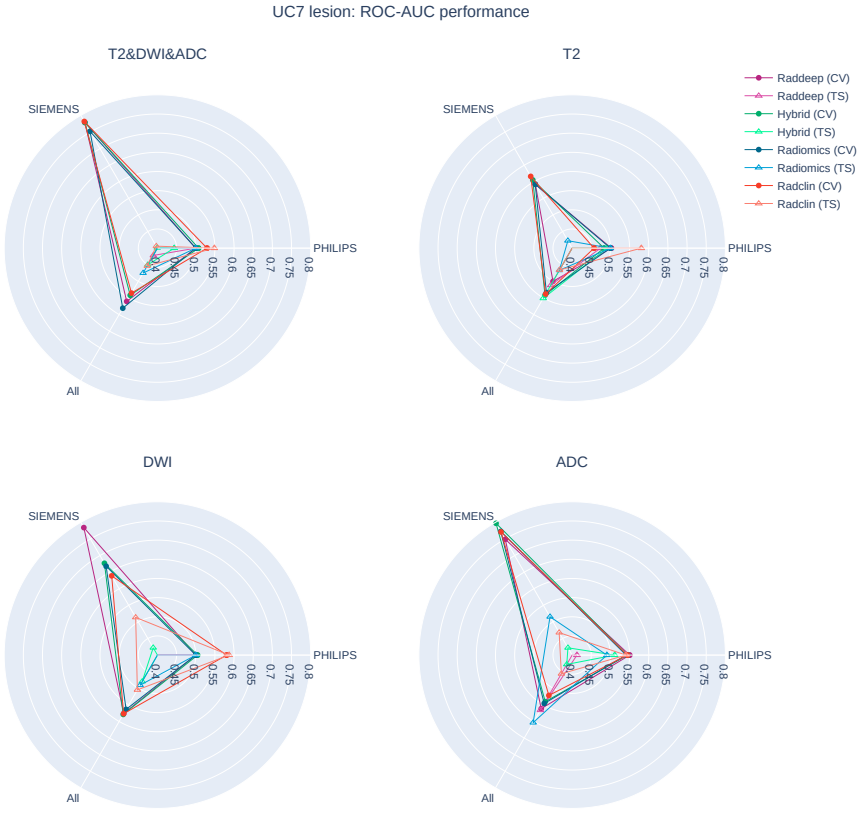


Figure 8.30: Cross-validation ROC-AUC model performance of 48 models trained to predict early quality of life after radical prostatectomy. The observations are color coded according to the type of data used in training: radiomics in blue, radclin in orange, raddeep in pink and hybrid in green.

8.3.7.2 Model Validation

Further analysis of the model selected in Section 8.3.7.1 revealed that the discriminatory power was extremely reduced, In fact, it was lower than the model presented in Chapter 7 (Figure 8.31 and Table 8.16). Thus, no further analysis was conducted.

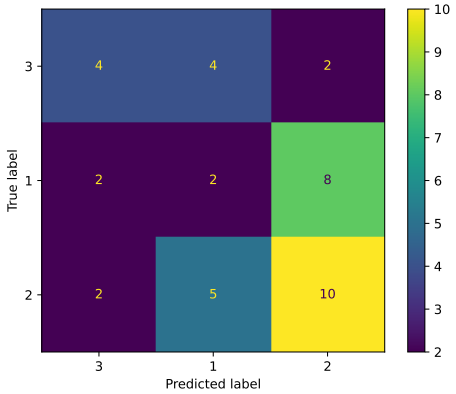


Figure 8.31: Hold-out test set confusion matrix of the radiomics_uc7_ADC_LGBM model.

Table 8.16: Multi-metric hold-out test set performance of the radiomics_uc7_ADC_LGBM model.

Hold-out test set	
AUC	0.6035 [0.4500-0.7570]
Sensitivity/Recall/TPR	0.3850 [0.2323-0.5377]
Precision/PPV	0.3939 [0.2406-0.5473]
F1	0.3863 [0.2335-0.5391]
F2	0.3848 [0.2321-0.5375]

8.3.8 Use Case 8 - Disease proggression

8.3.8.1 Model Selection

Figure 8.32 shows the cross-validation ROC-AUC model performance for the 24 models trained for UC8. The highest cross-validation performance was reached with the model radclin_uc8_ADC _SGD (0.9237 AUC).

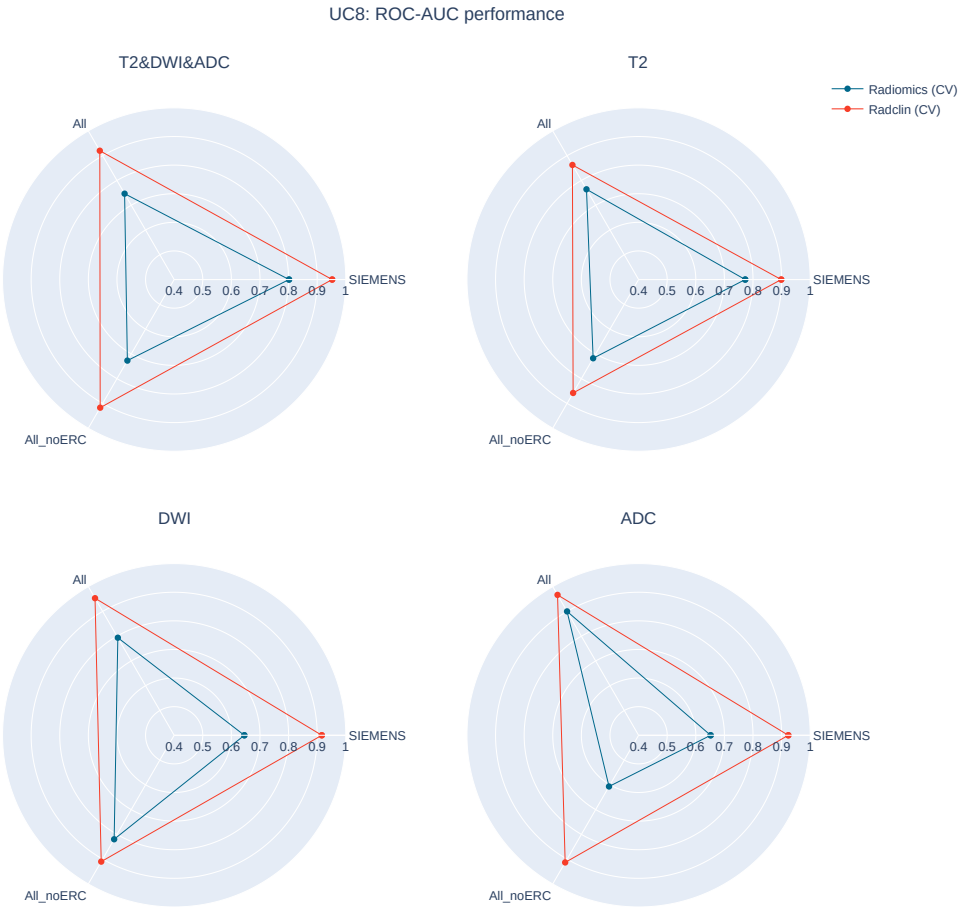


Figure 8.32: Cross-validation ROC-AUC model performance of 24 models trained to predict disease progression in UC8. The observations are color coded according to the type of data used in training: radiomics in blue and radclin in orange.

8.3.8.2 Model Validation

The model's ROC curve, used for the selection of the probability decision threshold of 0.1657, and the respective performance at this threshold are shown in Figure 8.33 and Table 8.17.

8.3.8.3 Explainability analysis

Figure 8.35 shows the logistic regression coefficients of the model radclin_uc8_ADC_SGD. The most impactful features are the PI-RADS score and the PZ index lesion location, both negatively

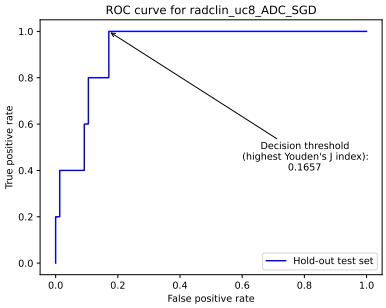


Figure 8.33: Receiver operator characteristics curve of the radclin_uc8_ADC_SGD model on cross-validation.

Table 8.17: Multi-metric performance of the radclin_uc8_ADC_SGD model at 0.1657 probability threshold. 95% confidence intervals are shown in brackets.

	Cross-validatiom
AUC	0.9237 [0.8659-0.9815]
Sensitivity/Recall/TPR	1.0 [1.0-1.0]
Specificity/TNR	0.8289 [0.7469-0.9110]
Precision/PPV	0.2778 [0.1802-0.3753]
F1	0.4348 [0.3268-0.5427]
F2	0.6579 [0.5546-0.7612]
CohensKappa	0.3743 [0.2689-0.5427]

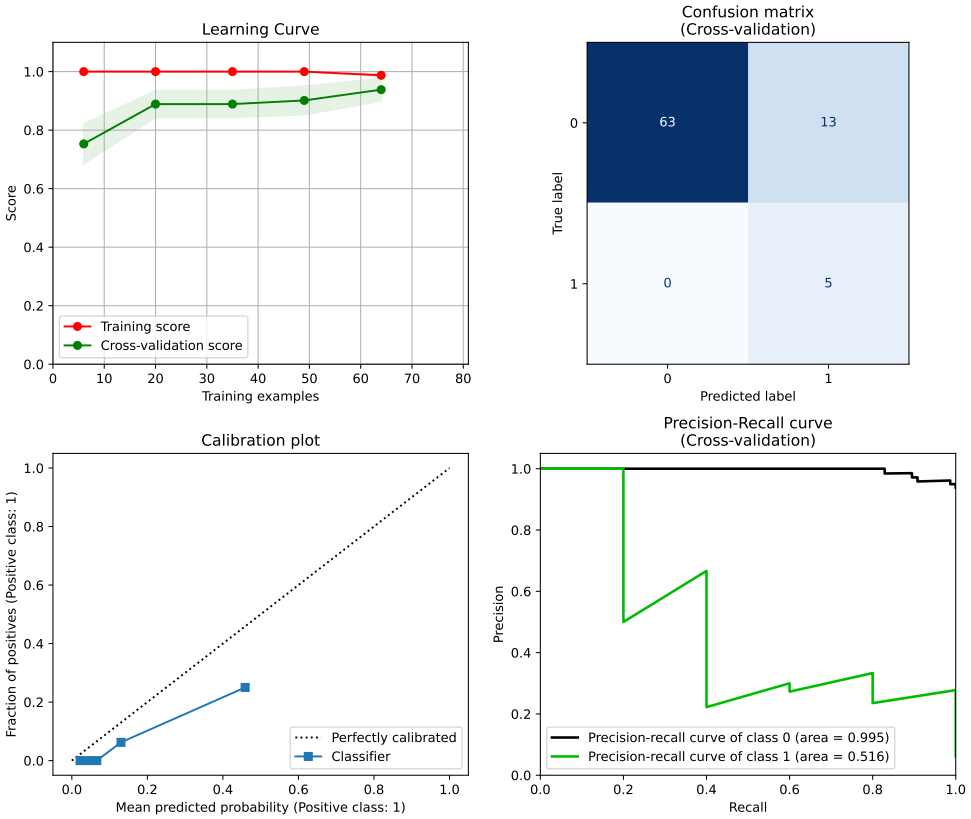


Figure 8.34: Analysis of the radclin_uc8_ADC_SGD model in terms of confusion matrix, learning curve, precision-recall curve and calibration plot.

associated with the disease progression outcome.

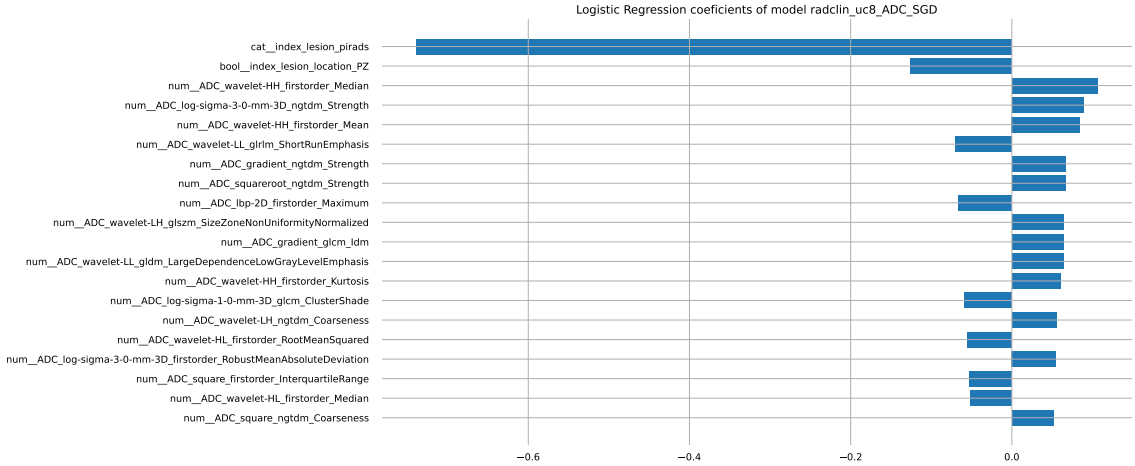


Figure 8.35: Logistic regression coefficients of 20 most impactful features for model radclin_uc8_ADC_SGD.

8.4 Discussion

In this work, the use of bi-parametric MRI index-lesion radiomics analysis is reported for six separate use cases: the prediction of disease aggressiveness (UC2), development of metastasis within six months (UC3), biochemical recurrence risk after prostatectomy or radiation therapy (UC5 and UC6, respectively), quality of life after prostatectomy (UC7b) and early withdrawal from the active surveillance program (UC8).

In agreement with Chapter 7 two aggressiveness definitions were considered (ISUP={0,1} vs ISUP=2,3,4,5; and ISUP=0,1,2 vs ISUP=3,4,5). Interestingly, for both ground truth definitions, the highest performing models combined radiomics and clinical information. However, while the distinction between ISUP=0,1 and ISUP=2,3,4,5 thrived with the three sequences, the distinction between ISUP=0,1, 2 and ISUP=3,4,5 achieved the highest performance with only T2W. The former ground truth (ISUP=0,1 vs ISUP=2,3,4,5) shows consistently 5% higher performance than the latter (ISUP=0,1,2 vs ISUP=3,4,5). This ground truth definition was expected to be the hardest to model, given the known mixture of clinically significant and insignificant cases in the ISUP classifications 2 and 3, corresponding to GS = 3 + 4 and GS = 4 + 3, respectively. Despite this, both models showed exceptional generalization power to the prospective cases. The explainability analysis revealed that the two models relied heavily on the clinical variables PI-RADS, index lesion location in the PZ, PSA blood level, and age, all showing a positive association with the aggressive output. Both shap analysis revealed a high representation of radiomic features extracted from wavelet filters, which are known for their relative invariability across volumes with big differences in intensity values, such as those found in different scanner manufacturers or in exams taken with or without endorectal coil. Robustness that was confirmed in the fairness analysis.

Use case 3, prediction of metastatic development within six months, was the smallest use case, with only 77 complete cases. Despite the small dataset, the selected model achieved a 3-fold cross-validation AUC of 0.8731. Surprisingly, the model was very good at identifying the minority class (specificity of 1.0), corresponding to no metastasis after six months, but not so great with the majority class (sensitivity 0.6731), corresponding to metastatic development within six months. The model was trained with a radclin dataset (radiomics and clinical features), putting a high value on the PI-RADS score.

Use case 5 related to the prediction of biochemical recurrence after radical prostatectomy surgery. For this use case, two approaches were taken: pre-surgery prediction and post-surgery prediction. The latter including clinical variables reported during or after the radical prostatectomy surgery. Interestingly, even though the same model configuration was selected for both scenarios (radclin data from T2W sequences), the pre-surgery classifier was able to achieve a higher performance. This contrasts with the results obtained with the whole-gland radiomics, where the inclusion of post-surgical information displayed predictive advantage.

Use case 6 aimed to predict biochemical recurrence after radiation therapy. The selected model achieved a 3-fold cross-validation AUC of 0.6806, and sensitivity and specificity of 0.67 and 0.75, respectively. Interestingly, and in contrast with the remaining clinical use cases, here, the inclusion

of clinical variables consistently lead to lower performance.

Use case 7 related to the prediction of bad, average, and good quality of life after radical prostatectomy, as defined by the EPIC26 score. The modelling efforts resulted in poor discriminatory performance, with a tendency for predicting the middle class, corresponding to average quality of life.

Use case 8 related to the prediction of an early withdrawal from the active surveillance program. This was one of the few use cases where an ADC model was ultimately selected. Here, an AUC of 0.9237 was achieved, and 1.0 and 0.8289 sensitivity and specificity, respectively. This was a successful proof-of-concept of the methodology for this particular application and should be further validated in future work.

This work has several limitations, all in common with Chapter 7, referring to a similar study with whole-gland radiomics.

Chapter 9

Radiomics-based lesion detection and aggressiveness mapping on mpMRI

This chapter describes the technique developed and adopted for the detection and aggressiveness mapping of Prostate Cancer lesions on multiparametric MRI. The work of this chapter resulted in a manuscript submission (22).

9.1 Introduction

The detection of cancer lesions on medical imaging is of extreme relevance, as early detection can greatly improve patient outcomes and survival rates (208). Using advanced imaging technology and artificial intelligence algorithms can aid in the accurate and timely diagnosis of cancer lesions (209; 210).

AI algorithms have been shown to be capable of not only detecting, but also characterizing lesions from medical images, evaluating features such as size, shape, and texture, to provide more comprehensive information for healthcare professionals (211; 212). AI tasks such as these have in the previous decades been handled by deep learning algorithms, which have shown promising results in the field of medical imaging (213; 214). Neural networks are indubitably powerful and versatile tools, however, they require extraordinarily large and heterogeneous datasets and, even then, have a propensity for overfitting (215).

Radiomics focuses on extracting quantitative data from medical images, providing valuable insights into the characteristics of lesions that may not be visible to the naked eye of the radiologist. This information can then be used to develop algorithms for more accurate diagnosis and efficient treatment planning (160). Recently, voxel-based radiomics has emerged, where features are calculated from a sliding kernel of pre-defined size centered on each voxel. The resulting feature maps can then be used for further analysis and modeling.

This work proposes a radiomics-based lesion detection method, as an alternative to imaging analysis using deep learning algorithms, focusing on prostate cancer lesion detection from pre-diagnostic biparametric MRI exams.

9.2 Methodology

9.2.1 Patient Cohort

The patient cohort used in this study was the group of patients from the ProstateNet dataset (described in Chapter 5) that had manual or semi-automatic index-lesion segmentations available. Two segmentation masks were considered:

- the whole-gland segmentation masks (Chapter 6);
- the index lesion segmentation masks manually or semi-automatically drawn by radiologists in the ProCAncer-I consortium.

9.2.2 Voxel-based radiomic features extraction

Bias field correction was performed on T2W images using the N4ITK algorithm (177; 178). Voxelwise feature extraction was performed in 2D, and T2W and DWI/ADC were resampled on the x- and y-axis to the 95th quantile value of 0.6875 mm and 2 mm, respectively. Image intensities were normalized.

Table 9.1: Voxel-based radiomics extraction parameters.

T2W extraction parameters	DWI extraction parameters	ADC extraction parameters
imageType: Original: setting: normalize: True normalizeScale: 100 force2D: True voxelArrayShift: 300 resampledPixelSpacing: [0.6875, 0.6875, 0] geometryTolerance: 1e-4 voxelSetting: kernelRadius: 1	imageType: Original: setting: normalize: True normalizeScale: 100 force2D: True voxelArrayShift: 300 resampledPixelSpacing: [2, 2, 0] geometryTolerance: 1e-4 voxelSetting: kernelRadius: 1	imageType: Original: setting: normalize: True normalizeScale: 100 force2D: True voxelArrayShift: 300 resampledPixelSpacing: [2, 2, 0] geometryTolerance: 1e-4 voxelSetting: kernelRadius: 1

Radiomic features were extracted from the three sequences for all voxels belonging to the whole gland segmentation. This was done using the Pyradiomics package (version 3.0) (202) in Python (version 3.7.9) (203).

All the pre-processing steps were performed as parameters of the extractor function, except for the bias field correction, which was performed prior to the extraction. No image transforms were included in the analysis, resulting in 93 radiomic maps calculated per sequence, including first-order and texture descriptors. The mathematical expressions and semantic meanings of the extracted features can be found at https://pyradiomics.readthedocs.io/en/latest/.

9.2.3 Model development

The train/test split was performed at patient level and stratified by ISUP grade (280 patients for training, and 100 held-out for validation). Considering ISUP $>$ 1 as aggressive lesions, these

constituted 72% of the dataset (Table 9.2). For the train set, the radiomic maps were flattened and concatenated as described in Figure 9.1. Rows with all zeros were eliminated (corresponding to the edges of the radiomic maps that were not part of the gland segmentation). A preprocessing pipeline was fitted on the train set and applied on the train and test sets: standardization, removal of low variance variables, and removal of highly correlated variables. For each patient, if aggressive disease was present (ISUP $>$ 1), the ground truth constituted of the index-lesion segmentation resampled and cropped to the radiomic maps' space, and subsequently flattened. Whereas, if the patient presented with non-aggressive disease (ISUP = 1) or no cancer (ISUP = 0), the ground truth was a vector of zeros. Random samples of 1 000 000 voxels were selected from each label (1 if the voxel belongs to the index lesion segmentation, and 0 if it does not or the patient had no cancer / non-aggressive disease), forming a final dataset with 2 000 000 cases, that was utilized to train a Logistic Regression classifier.

Table 9.2: Label distribution in the train and test sets.

	ISUP score					
	0	1	2	3	4	5
Train Set	16	63	107	55	14	25
Test Set	6	23	38	19	5	9

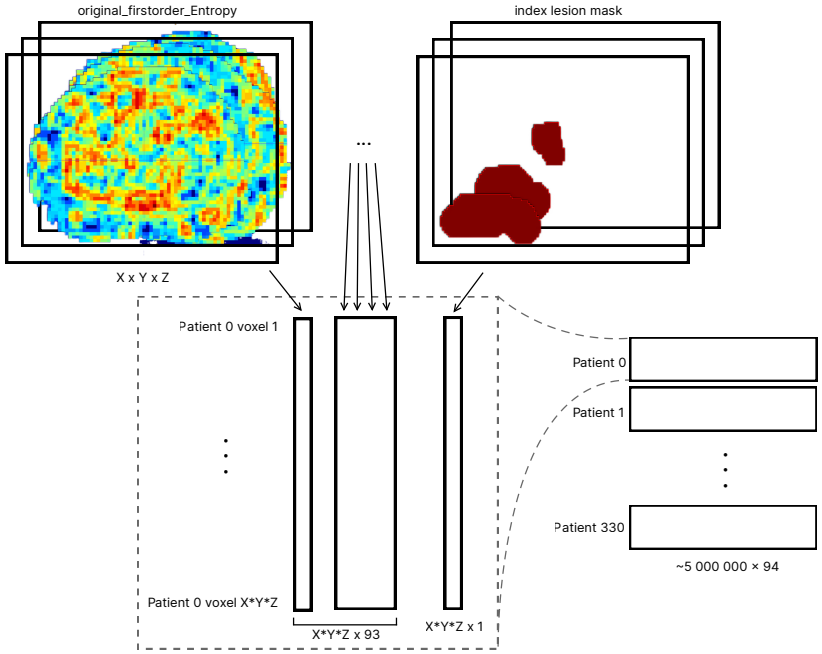


Figure 9.1: Diagram depicting the transformation of the 3-dimensional radiomic maps from each patient into the final tabular dataset. The index lesion masks represent the ground truth.

Two approaches were considered. A pre-fusion Logistic Regression model was trained with the three sequences together, by concatenating the dataframes in the column axis. Whereas the post-fusion approach consisted of training a separate Logistic Regression classifier for each sequence, and the average inferred probability resulted in the final prediction.

9.2.4 Model Validation

The algorithm for model inference for each patient in the test set was the following:

1. Generation of whole-gland segmentation on T2W, DWI and ADC (216).
2. Extraction of radiomic maps from the whole gland, following the parameter files used for the train set.
3. Flattening and concatenation of the radiomic maps.
4. Transform by application of the previously fitted preprocessing pipeline.
5. Voxel-wise model inference.
6. Unflattening of predictions vector to match dimensions of the radiomic maps.
7. Resampling and padding to the initial image's space.

Lesion candidates were extracted from the predicted probability maps using a threshold of 0.55 (217). These were post-processed in two stages: objects smaller than 300 voxels were removed; and holes smaller than 300 voxels were filled. The final predicted segmentations were compared to the original lesion segmentation masks (manual or semi-manual) in terms of Dice coefficient and Intersection over Union, calculated with MedPy (218). All metrics were calculated at the patient level.

9.3 Results

Figure 9.2 shows the probability map predictions of the models on six example cases from the held-out test set. Though both models make very similar predictions, the pre-fusion model appears to make more extreme probability estimates, which makes it easier to discriminate the lesion borders. As expected, the latter model finds no aggressive areas in the ISUP 0 and 1 examples. However, it does find aggressive regions on the other ISUP scores, with increased accuracy as ISUP increases.

A lesion was considered to have been detected if the Dice coefficient was over 10%:

- FN : ISUP $>$ 1 and Dice $<$ 0.1
- FP : ISUP $\leq$ 1 and Dice $\geq$ 0.1
- TN : ISUP $\leq$ 1 and Dice $<$ 0.1
- TP : ISUP $>$ 1 and Dice $\geq$ 0.1

The pre-fusion model presented a sensitivity of 0.6087, specificity of 0.4138, and precision of 0.7119, for aggressive lesion detection. Whereas the post-fusion model presented with a very similar sensitivity of 0.5942, specificity of 0.4483, and precision of 0.7193.

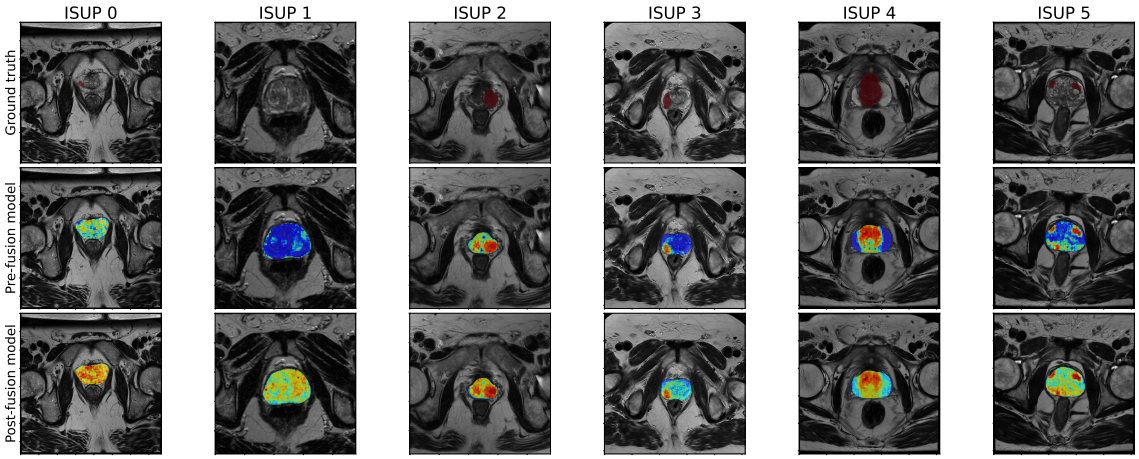


Figure 9.2: Predictions on 6 example cases from the held out test set, from each ISUP classification. The color map goes from dark blue (with probability 0) to dark red (with probability 1).

Tables 9.3 and 9.4 show the average segmentation performance by ISUP grade of the pre- and post-fusion models, respectively. The models perform very similarly both in terms of detection and segmentation. The dice coefficient for ISUP grade 2 is between 32-36%, for ISUP 3 is between 26-28%, for ISUP 5 is between 17-19%, while for ISUP 4 theres's more variability. The majority of false negatives are ISUP 2 cases: 14 or 15 out of 37 ISUP 2 cases being missed by the model. The average Dice was 16% higher for cases where endorectal coil was not used, though the pool of endorectal-coil cases was small (37 cases taken without, and 4 taken with an endorectal coil, among the TP). Regarding scanner manufacturer, the model displayed robust performance: dice coefficients of 0.31, 0.32, and 0.28 for GE, Philips, and Siemens, respectively.

Table 9.3: Segmentation performance of the Pre-fusion model.

	ISUP	Dice	IoU	N
FN	2	0.01 +/- 0.01	0.00 +/- 0.01	15
	3	0.00 +/- 0.01	0.00 +/- 0.00	8
	4	0.00 +/- nan	0.00 +/- nan	1
	5	0.00 +/- 0.01	0.00 +/- 0.00	3
FP	0	0.31 +/- 0.19	0.19 +/- 0.14	3
	1	0.23 +/- 0.15	0.14 +/- 0.11	14
TN	0	0.01 +/- 0.00	0.00 +/- 0.00	3
	1	0.01 +/- 0.02	0.01 +/- 0.01	9
TP	2	0.36 +/- 0.22	0.25 +/- 0.18	22
	3	0.26 +/- 0.18	0.16 +/- 0.13	11
	4	0.45 +/- 0.26	0.31 +/- 0.20	3
	5	0.19 +/- 0.07	0.10 +/- 0.04	6

Figure 9.3 shows the logistic regression coefficients of the pre-fusion model. The model is relying heavily on texture indicators coming essentially from DWI and ADC sequences. Aside from texture markers, the feature *original_firstorder_Minimum* indicates that a low value on DWI contributes to an aggressive prediction, whereas a low value on ADC contributes to a non-aggressive prediction. This is concordant with the MRI interpretation done by radiologists.

Table 9.4: Segmentation performance of the Post-fusion model.

	ISUP	Dice	IoU	N
FN	2	0.01 +/- 0.01	0.00 +/- 0.01	14
	3	0.01 +/- 0.02	0.00 +/- 0.01	9
	4	0.00 +/- 0.00	0.00 +/- 0.00	2
	5	0.00 +/- 0.01	0.00 +/- 0.00	3
FP	0	0.18 +/- 0.09	0.10 +/- 0.06	3
	1	0.25 +/- 0.17	0.15 +/- 0.13	13
TN	0	0.01 +/- 0.01	0.00 +/- 0.00	3
	1	0.01 +/- 0.02	0.01 +/- 0.01	10
TP	2	0.32 +/- 0.22	0.21 +/- 0.18	23
	3	0.28 +/- 0.18	0.18 +/- 0.14	10
	4	0.63 +/- 0.04	0.46 +/- 0.04	2
	5	0.17 +/- 0.07	0.09 +/- 0.04	6

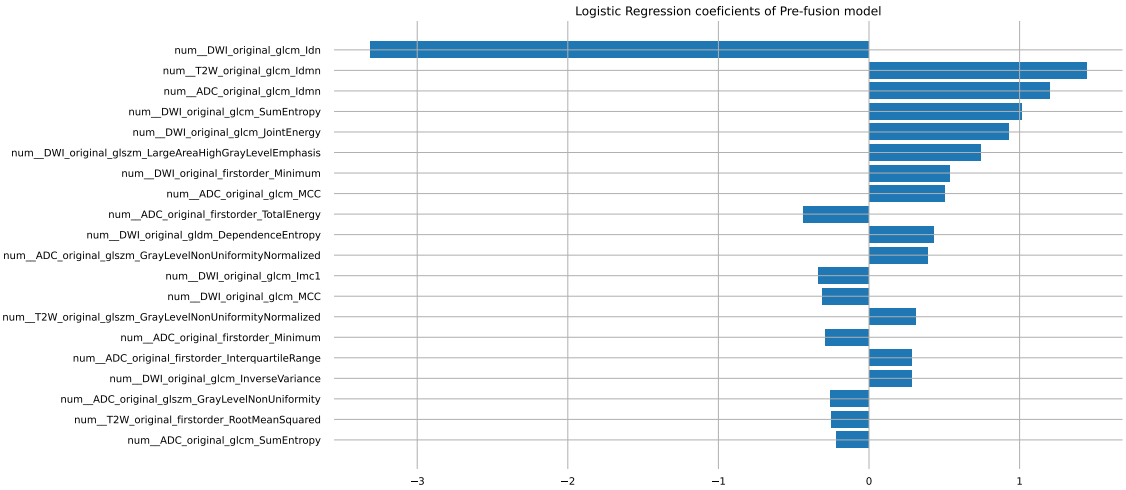


Figure 9.3: Fitted logistic regression coefficients of the pre-fusion model.

9.4 Discussion

This study proposes a simple method to perform aggressive lesion detection and aggressiveness mapping using voxel-wise radiomics and classical machine learning algorithms.

Deep learning algorithms are the current standard for Lesion detection or segmentation problems. On the segmentation side, Kohl et al. (219) employed 152 patients from a single center to train a U-net model and obtained 0.41 dice. Dai et al. (220) trained with 66 patients (2 centers) and reached a dice score of 0.46, for segmentation of aggressive and non-aggressive lesions. Our model was able to achieve comparable performances, relying uniquely on radiomic features.

On the detection side, it is notable to mention that the aim was not to detect all lesions, but only the aggressive ones, which makes it increasingly difficult to find comparable studies in the literature. Nevertheless, Cao et al. (221) achieved 0.81 AUC and 0.88 Sensitivity in aggressive lesion detection with a FocalNet, in a single-center study. Hosseinzadeh et al. (222) reached 0.85 sensitivity, with a dataset of 2734 patients and two convolutional neural networks. Seetharaman et al. (223) achieved between 0.75 and 0.8 AUC with a cohort of 322 patients and an adapted convolutional neural network based on the holistically nested edge detector architecture. The model

proposed by this study did not reach the performances reported in the literature, however, our model was trained and validated with a multi-centric and multi-vendor dataset, showing stability in segmentation performance across scanner manufacturers and endorectal-coil usage.

This study has several limitations. Firstly, each voxel was treated by the model as an independent observation, which is untrue. There are two main dependencies: intra-patient, since groups of voxels belong to the same patient, and regional dependency, since voxels in the same anatomical region are related to each other. Secondly, as the radiologists were instructed to segment only the index lesion, we cannot be certain whether other lesions found by the model, are actual lesions (true positives) or not (false positives). The authors propose that proper validation could be accomplished by inspection of the biopsy or radical prostatectomy specimens by an expert pathologist.

To conclude this Chapter, we have successfully demonstrated the power of voxel-based radiomics, coupled with simple machine learning algorithms, in solving aggressive lesion detection and aggressiveness mapping tasks on prostate biparametric MRI.

Chapter 10

Uplift modeling for treatment effect estimation

This chapter describes the efforts taken to estimate treatment effects or uplift for prostate cancer patients. Section 2.2.4.2 "Assumptions" is based on the description by Brady Neal (224). The work of this chapter resulted in a conference acceptance (23).

10.1 Introduction

Prostate cancer has a complex treatment landscape (15). The management of this disease follows three main courses of treatment (87): radical prostatectomy (RP), the surgical removal of the prostate gland; radiation therapy (RT), cancer cell elimination through radiation; and active surveillance (AS) (82), where lower aggressiveness patients are monitored for signs of disease progression.

Uplift modeling techniques have recently been applied to the health domain (225; 226; 227) with the goal of predicting patient-specific treatment effects (or uplift), and providing a data-driven recommendation for the treatment with the highest chance of a positive outcome. Previous literature predicts uplift by conditioning on relevant clinical variables that represent possible confounders of the relationship between treatment and outcome (225; 226; 227).

In the prostate cancer domain, there are known confounders of the association between treatment and positive or negative outcome. Some of these are patient age (228), PSA blood levels (44; 45; 46) and patient or tumor morphological characteristics, such as lesion location and histopathological aggressiveness (15). Bi-parametric pelvic MRI (bpMRI) is an essential tool in the clinical management of prostate cancer patients (68; 69) and, thus, represents promising insight into these previously unquantified patient and tumor physiological and morphological characteristics. Radiomics (107) has, over the last decade, emerged as a series of quantitative measurements, representing medical images. Thus, it can be used as a representation of these confounding characteristics found on bpMRI.

This study differs from previous research in the sense that it uses, not only clinical variables but also radiomic features. Additionally, instead of binary treatment effects, we present an analysis of two courses of aggressive treatment (radiation therapy and radical prostatectomy) and their impact on patient outcomes when compared to active surveillance. To our knowledge, no other study has attempted to apply uplift techniques to the prostate cancer domain.

10.2 Methodology

10.2.1 Data description

1048 patients were retrieved from the ProstateNet dataset (described in Chapter 5) since they were eligible for clinical use cases 5, 6, and 8, corresponding respectively to the three courses of treatment taken by prostate cancer patients: radical prostatectomy (RP), radiation therapy (RT), and active surveillance (AS).

Radiomic features were extracted from the whole-gland and index lesion segmentations, as per described in chapters 7 and 8, respectively. The clinical variables included were: patient age, baseline PSA, index lesion location, PI-RADS, gleason scores of the first and second most prominent histological patterns, and ISUP grade. These variables were selected since they cover common confounders of the relationship between treatment and outcome (Figure 10.1).

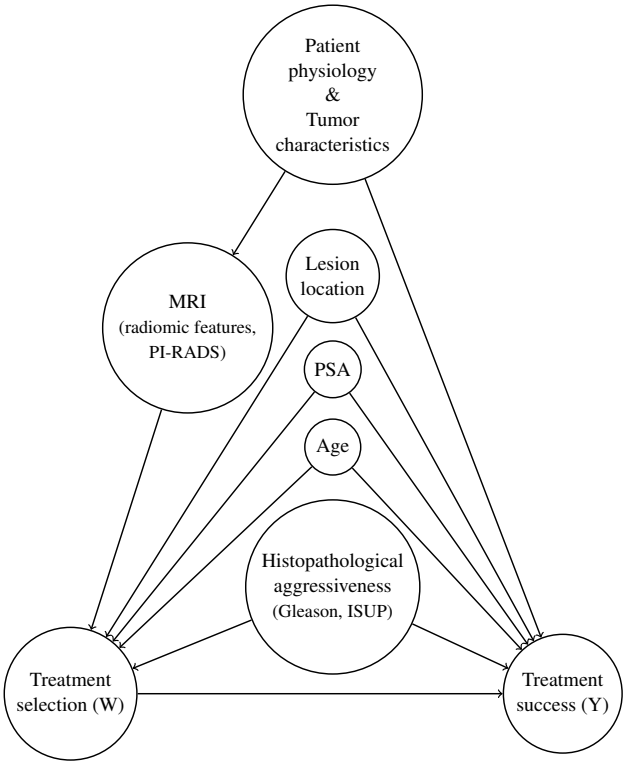


Figure 10.1: DAG depicting the confounders considered and controlled for, in this study.

The outcome variable was defined as "treatment success/failure" (corresponding to 1 or 0, respectively). For use cases 5 and 6, treatment failure refers to biochemical recurrence at six

months follow-up, whereas for use case 8, treatment failure refers to disease progression at six months follow-up. In contrast, treatment success refers to no recurrence or disease progression at follow-up.

Train/test split was done at patient level: 100 patients were randomly assigned to the hold-out test set. The distribution of the outcome variable per treatment is described in Tables 10.1 and 10.2.

Table 10.1: Distribution of the outcome variable for each treatment in the train set.

	Treatment		
	AS	**RP**	**RT**
Treatment failure	10	95	13
Treatment success	128	594	108

Table 10.2: Distribution of the outcome variable for each treatment in the test set.

	Treatment		
	AS	**RP**	**RT**
Treatment failure	0	11	1
Treatment success	13	63	12

10.2.2 Modelling

The preprocessing pipeline described in this section was fitted on the train set and applied on the train and test sets. Features were scaled to have zero mean and standard deviation equal to 1 (Python package scikit-learn version 1.0.2). Features with low variance were identified and excluded. Here, a threshold of 0.01 was considered. Finally, feature correlation was assessed. Feature pairs were considered correlated if their Spearman correlation was higher than 0.8. Out of the two, the feature with the highest average correlation across all features was eliminated.

The Chapter 2 describes several algorithms for treatment effect estimation, all of which considered two treatment options: treatment and control. In this work, three treatment options are considered: active surveillance (control), radical prostatectomy, and radiation therapy. The algorithms are adapted in the following way: subsets of the dataset are considered to estimate the treatment effect of each treatment, after which the treatment with the highest effect is selected. Four meta-learners (S, T, X, and R), one transformed outcome learner, and one uplift Random Forest were trained. Hyperparameter tuning of each learner, and its algorithms, was performed with a custom grid search. Each hyperparameter set was evaluated through 3-fold cross-validation, optimizing the area under the uplift curve (AUUC). The hyperparameter space considered is described in Table 10.3.

10.2.3 Model analysis

The uplift model with the highest AUUC was selected to split the hold-out test set into those who would benefit from treatment (uplift > 0) and those who would not (uplift < 0). A statistical comparison of the two groups was performed with a t-test for numerical variables and Chi-squared test for categorical variables. A logistic regression was then trained to predict these two classes, and based on the fitted coefficients, a nomogram was produced.

Table 10.3: Hyperparameter space searched during optimization.

Learner	**Algorithm**	**Hyperparameter space**
S-learner	Logistic Regression	{'C': [0.01, 0.1, 1, 10, 100], 'penalty': ['l1', 'l2']}
T-Learner	Logistic Regression	{'C': [0.01, 0.1, 1, 10, 100], 'penalty': ['l1', 'l2']}
X-learner	Logistic Regression	{'C': [0.1, 1, 10], 'penalty': ['l1']}
	Support Vector Regressor	{'kernel':('rbf', 'poly'), 'C':[0.1, 1, 10], 'gamma':('auto','scale')}
R-Learner	Logistic Regression	{'C': [0.01, 0.1, 1, 10, 100], 'penalty': ['l1', 'l2']}
	Support Vector Regressor	{'kernel':('linear', 'rbf', 'sigmoid', 'poly'), 'C':[0.1, 1, 10], 'gamma':('auto','scale')}
TO-Learner	Logistic Regression	{'C': [0.01, 0.1, 1, 10, 100], 'penalty': ['l1', 'l2']}
	Support Vector Regressor	{'kernel':('linear', 'rbf'), 'C':[0.1, 1, 10], 'gamma':('auto','scale')}
RF-Learner		{'n_estimators':[10, 50, 100], 'evaluationFunction':['KL', 'ED', 'Chi'], 'max_depth':[5, 10, 15]}

The group of patients predicted to benefit from treatment was then split into those who would most benefit from RP from those where RT is preferred, based on the maximum treatment effect predicted for each treatment. Again, a statistical comparison of the two groups was performed and a second nomogram was constructed through the coefficients of a fitted logistic regression.

10.3 Results

Figure 10.2 shows the hold-out test set area under the uplift curve (AUUC) for all trained models. The area under the theoretical maximum uplift curve for the hold-out test set was 0.7488. The model with the highest performance reached an AUUC of 0.5481. This model was trained with radiomic features extracted from the whole-gland segmentation of T2W exams. Its uplift curve is displayed in Figure 10.3.

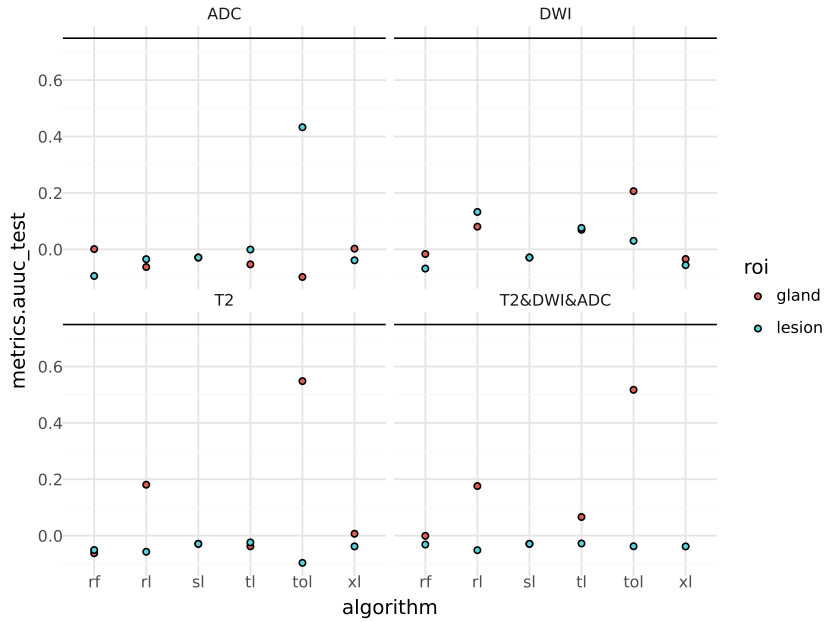


Figure 10.2: Area Under the Uplift Curve (AUUC) performance of all uplift trained models. The maximum AUUC is represented as a horizontal black line.

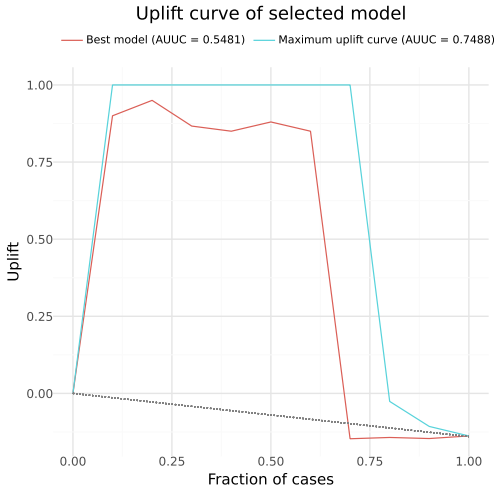


Figure 10.3: Uplift Curve of the model that reached the highest AUUC (red), compared with the maximum uplift curve for the hold-out test set (blue) and the random line (grey).

Based on this model, 95 patients were predicted to benefit from aggressive treatment (uplift > 0) and 5 were predicted to benefit from an active surveillance program (uplift < 0). Even though only the radiomic signature and the PSA levels showed individually significant differences between the two groups (Table 10.4), combinations of the different variables have predictive value. Figure 10.4 shows the nomogram produced to determine whether a patient will benefit from aggressive treatment or if they should be followed in an active surveillance program. It was found that for older patients with high PSA levels, high ISUP, and high radiomic signature, more aggressive treatment is preferred.

Similarly, the model found that, from the 95 patients that would benefit from treatment, 21 would benefit from RT while for 74 RP would be preferred. Considering individual variables, no statistically significant differences were found between the two groups (Table 10.5). Figure 10.5 shows the nomogram produced to determine whether a patient should be sent to RP or RT. It was found that younger patients, patients with high PSA, low PI-RADS, high ISUP, and low Gleason 2 tend to benefit from RT.

10.4 Discussion

This study presents an approach based on uplift modeling to identify prostate cancer patients who would benefit from radical prostatectomy, radiation therapy, or active surveillance.

Uplift modeling has gained popularity in recent years as a technique to predict the effect of interventions or treatments for subgroups of patients (229; 225; 226; 227).

The core idea behind uplift modeling is to identify those individuals who would benefit from treatment, while preventing from receiving treatment those who would not benefit from it or even, who would be harmed by it. This approach has been applied to various use cases in the health domain, including acute kidney injury (226; 227), and persistent atrial fibrillation (225). However,

Table 10.4: Statistical comparison of clinical variables and radiomics signature between the patients predicted to benefit or not from aggressive treatment.

	Do not benefit from treatment (N=5)	**Benefit from Treatment** (N=95)	**P-value**
Age (years)			
Mean (SD)	62.8 (7.66)	64.5 (7.00)	0.652
Median [Min, Max]	62.0 [54.0, 72.0]	65.0 [45.0, 79.0]	
PSA (ug/ml)			
Mean (SD)	5.66 (1.42)	7.79 (5.21)	0.0251
Median [Min, Max]	6.00 [3.21, 6.81]	6.51 [1.03, 43.0]	
PI-RADS			
0	0 (0%)	2 (2.1%)	0.919
2	0 (0%)	1 (1.1%)	
3	0 (0%)	7 (7.4%)	
4	3 (60.0%)	40 (42.1%)	
5	2 (40.0%)	45 (47.4%)	
Index lesion in PZ			
0	0 (0%)	10 (10.5%)	1
1	5 (100%)	85 (89.5%)	
Index lesion in TZ			
0	5 (100%)	79 (83.2%)	0.707
1	0 (0%)	16 (16.8%)	
Index lesion in AS			
0	5 (100%)	91 (95.8%)	1
1	0 (0%)	4 (4.2%)	
Index lesion in CZ			
0	5 (100%)	91 (95.8%)	1
1	0 (0%)	4 (4.2%)	
Gleason 1			
3	5 (100%)	69 (72.6%)	0.397
4	0 (0%)	25 (26.3%)	
5	0 (0%)	1 (1.1%)	
Gleason 2			
3	2 (40.0%)	44 (46.3%)	1
4	3 (60.0%)	51 (53.7%)	
ISUP grade			
1	2 (40.0%)	26 (27.4%)	0.761
2	3 (60.0%)	43 (45.3%)	
3	0 (0%)	18 (18.9%)	
4	0 (0%)	7 (7.4%)	
5	0 (0%)	1 (1.1%)	
Radiomics signature			
Mean (SD)	-8.98 (2.15)	0.415 (1.57)	<0.001
Median [Min, Max]	-9.10 [-11.6, -6.20]	0.500 [-3.52, 4.28]	

to our knowledge, no other study reported this type of analysis applied to the prostate cancer domain.

Li et al. (226) used uplift modeling to identify a subgroup of patients with acute kidney injury who would benefit from renal replacement therapy. He went as far as to identify patient characteristics that were causal predictors of treatment failure, namely: urine output, fluid input, mean blood pressure, body temperature, and lactate levels. Biswas et al. (227) used uplift modeling to identify patients with acute kidney injury who would benefit from an electronic alert system rather than usual care. They concluded that older individuals, women, and patients with lower creati-

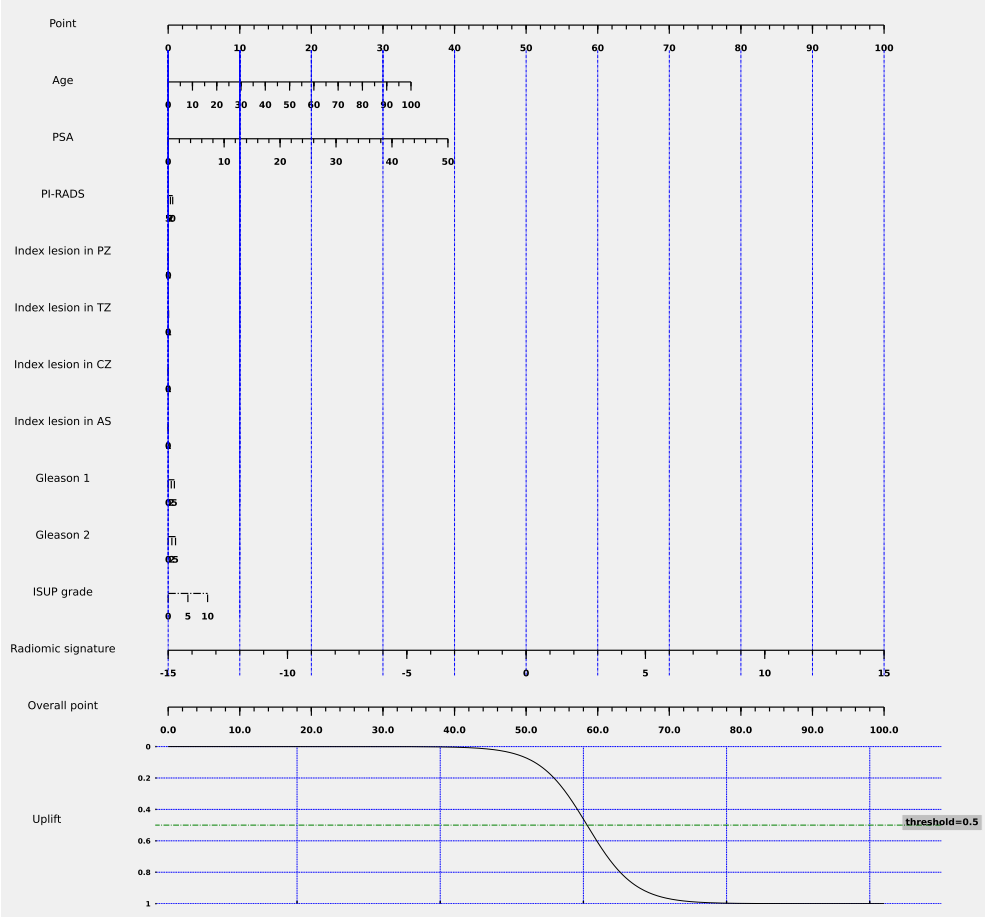


Figure 10.4: Nomogram of the recommendation for starting an aggressive treatment or enrolling in an active surveillance program.

nine levels were the most likely to benefit from electronic alerts. Sato et al. (225) applied uplift techniques to identify patients with persistent atrial fibrillation who would benefit from extensive catheter ablation and pulmonary vein isolation, both controversial treatment options in the area.

This study differs from previous research in the sense that it uses, not only clinical variables, but also radiomic features extracted from biparametric MRI exams, an essential tool in the clinical management of prostate cancer patients. Additionally, instead of binary treatment effects, we present an analysis of two courses of aggressive treatment (radiation therapy and radical prostatectomy) and their impact on patient outcomes when compared to active surveillance. Overall, the findings suggest that a personalised approach to treatment may be more beneficial for prostate cancer patients.

This study has two main limitations: the assumptions made, and the data imbalance. Firstly, regarding the assumptions, including as many relevant variables as possible increases the chance of covering the unconfoundedness assumption. However, as dimensionality increases, there is also a higher chance of having subgroups that only receive treatment or only receive control, which is counterproductive due to the positivity assumption. One example of this is disease aggressiveness. This variable is a known confounder as it impacts both treatment choice and recurrence, so it was

Table 10.5: Statistical comparison of clinical variables and radiomics signature between the patients predicted to benefit from radical prostatectomy (RP) or radiation therapy (RT).

	RP (N=74)	**RT** (N=21)	**P-value**
Age (years)			
Mean (SD)	64.4 (7.50)	64.8 (5.04)	0.774
Median [Min, Max]	65.0 [45.0, 79.0]	65.0 [58.0, 78.0]	
PSA (ug/ml)			
Mean (SD)	7.62 (3.92)	8.39 (8.42)	0.689
Median [Min, Max]	6.58 [1.03, 25.0]	6.50 [1.76, 43.0]	
PI-RADS			
0	2 (2.7%)	0 (0%)	0.756
2	1 (1.4%)	0 (0%)	
3	6 (8.1%)	1 (4.8%)	
4	29 (39.2%)	11 (52.4%)	
5	36 (48.6%)	9 (42.9%)	
Index lesion in PZ			
0	9 (12.2%)	1 (4.8%)	0.567
1	65 (87.8%)	20 (95.2%)	
Index lesion in TZ			
0	59 (79.7%)	20 (95.2%)	0.178
1	15 (20.3%)	1 (4.8%)	
Index lesion in AS			
0	71 (95.9%)	20 (95.2%)	1
1	3 (4.1%)	1 (4.8%)	
Index lesion in CZ			
0	72 (97.3%)	19 (90.5%)	0.448
1	2 (2.7%)	2 (9.5%)	
Gleason 1			
3	53 (71.6%)	16 (76.2%)	0.821
4	20 (27.0%)	5 (23.8%)	
5	1 (1.4%)	0 (0%)	
Gleason 2			
3	34 (45.9%)	10 (47.6%)	1
4	40 (54.1%)	11 (52.4%)	
ISUP grade			
1	19 (25.7%)	7 (33.3%)	0.887
2	34 (45.9%)	9 (42.9%)	
3	15 (20.3%)	3 (14.3%)	
4	5 (6.8%)	2 (9.5%)	
5	1 (1.4%)	0 (0%)	
Radiomics signature			
Mean (SD)	0.418 (1.55)	0.403 (1.69)	0.972
Median [Min, Max]	0.430 [-3.18, 4.28]	0.728 [-3.52, 2.64]	

included in X and, thus, conditioned on. However, it is also known that low aggressiveness patients tend to be put into active surveillance programs, rather than going through aggressive treatment. Therefore, this could break the positivity assumption, as this subpopulation could have zero probability of receiving treatment. The unconfoundedness assumption could similarly be broken by unknown confounders that were not included in X such as the presence of extraprostatic extension. Finally, the consistency assumption could be broken by the application of different radiation doses in radiation therapy, or by different surgeon experience levels in radical prostatectomy. Secondly, regarding the data imbalance, the dataset is heavily skewed towards radical prostatectomy

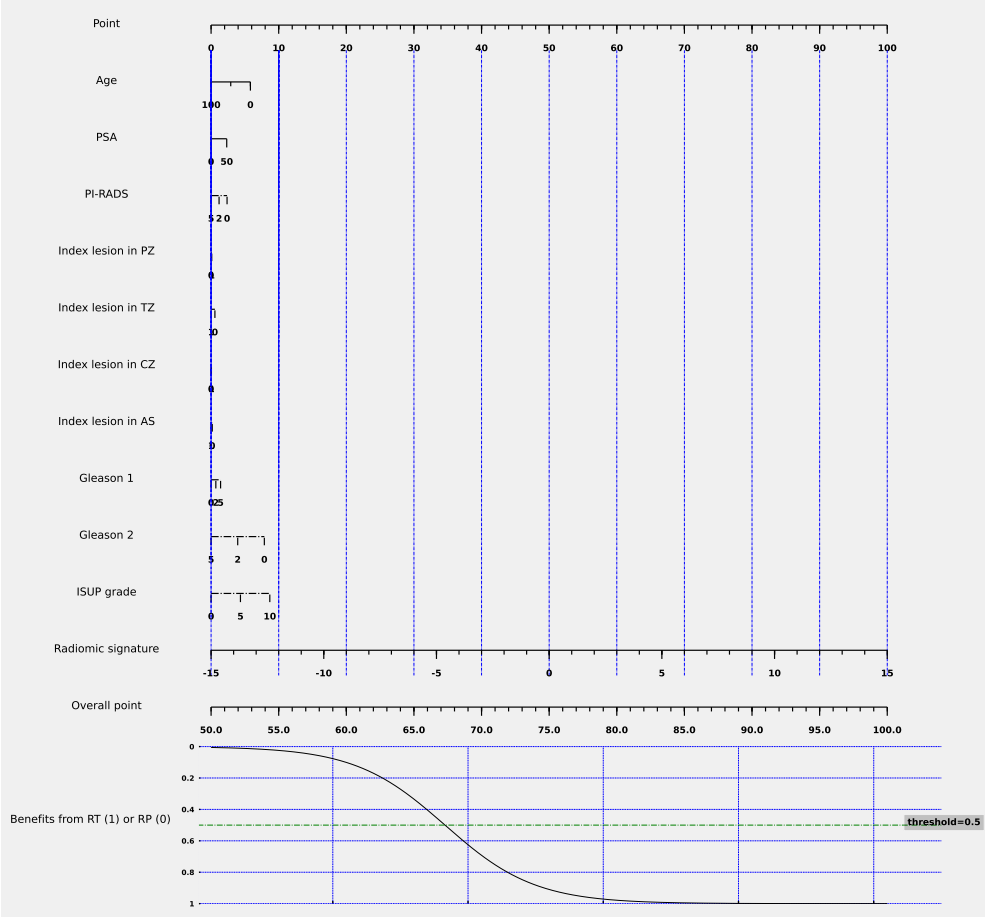


Figure 10.5: Nomogram of the recommendation for starting radiation therapy (RT) or undergoing radical prostatectomy (RP).

cases (73% RP, 13% RT, and 14% AS). The low representation of RT or AS cases could impact model generalizability in these settings, as was evident in inference to the hold out test set, where 74% of patients were predicted to benefit from RP, while only 21% and 5% were predicted to benefit from RT and AS, respectively.

In conclusion, uplift modeling can be a valuable tool for identifying subgroups of patients who may benefit from specific treatments in healthcare. However, it is essential to critically evaluate its use, particularly in the context of cancer care. Prostate cancer, with its complex treatment landscape and varying patient characteristics, presents an exciting opportunity for applying uplift modeling to inform personalized treatment decisions.

Chapter 11

Discussion and Main Conclusions

In this chapter, the overall results and the different techniques tested in this work are discussed. Moreover, this dissertation's main limitations will be communicated and plans for future work will be outlined. Finally, conclusions will be drawn.

11.1 Comments on the research questions

In this section, each research question will be individually answered and the results will be further discussed.

11.1.1 RQ 1

Which is the best approach to train robust classifiers to inter- and intra-reader variability in segmentation margins?

This dissertation demonstrated (in Chapter 4) that the most successful approach to developing classifiers that maintain performance, despite real-world differences in segmentation margins, is through the inclusion in the training set of diverse data labeled by different radiologists with various levels of experience. We hypothesize that training with the radiomic features extracted from the heterogeneous pool of segmentation masks is the only way to prepare the model for the real-world heterogeneity it would find once deployed in a clinical setting.

11.1.2 RQ 2

Is the combination of deep and handcrafted radiomic features (I) capable of outperforming radiomic features alone (O) in different machine learning tasks (C) for patients in the prostate cancer disease continuum (P)?

This research question was explored not only in Chapter 4 but also in Chapters 7 and 8. In Chapter 4, it was found that including deep features improved the cross-validation performance but resulted in classifiers with a higher degree of overfitting. In Chapters 7 and 8, it was found that

the benefit of including deep features was highly dependent on the use case. On top of this, the inclusion of deep features rarely outperformed the best radiomics or radclin approaches.

11.1.3 RQ 3

Is the combination of clinical information and handcrafted radiomic features (I) capable of outperforming radiomic features alone (O) in different machine learning tasks (C) for patients in the prostate cancer disease continuum (P)?

In concordance with Section 11.1.2, Chapters 7 and 8 revealed that this highly depends on the clinical use case. While for most use cases the inclusion of clinical variables was beneficial for model performance, this was not always the case. One example of this was UC6, where the inclusion of clinical variables, consistently lead to lower performance.

11.1.4 RQ 4

Are radiomic signatures extracted from bi-parametric prostate MRI examinations (I) capable of accurately classifying prostate cancer disease aggressiveness (clinical significance defined by ISUP grade) (O) when compared with biopsy (C) for patients with a suspicion of clinically significant prostate cancer (P)?

This research question refers to clinical use case 2 (Chapters 7 and 8), and the answer is "yes". The models presented show over 0.90 AUC performance on an extremely heterogeneous hold-out test set, and on prospective validation with contemporary cases. This model could serve as an aid in patient selection for invasive procedures, potentially helping reduce unnecessary biopsies.

11.1.5 RQ 5

Are signatures of radiomic and/or clinical features extracted from pre-operative bi-parametric prostate MRI examinations (I) more accurate in predicting metastatic development (within six months follow-up) (O) compared with contemporary predictive models (C) for patients with localized prostate cancer (P)?

The work presented was not successful in the prediction of metastatic development. However, given the successes found in the literature for this particular task (Chapter 3 Section 3.2.2) the author proposes two potential justifications: (1) the literature is filled with deep learning approaches, so perhaps radiomics is not the most suitable technique for this particular task; or (2) the dataset size collected within ProstateNet for UC 3 is too small for this particular task. While for most use cases, the largest reduction in the number of patients happened due to unavailability of DWI or ADC volumes, for UC3 the reduced dataset size is mostly attributable to ground truth unavailability. It was found that for around 30% of the original number of patients, the reported metastatic status was "MX", corresponding to an inability to evaluate the metastatic status by the radiologist. These cases were treated as having a missing ground truth and were removed from the analysis, resulting in a dataset that was vastly smaller than those reported in the literature for this type of task.

11.1.6 RQ 6

Are signatures of radiomic and/or clinical features extracted from pre-operative bi-parametric prostate MRI examinations (I) more accurate in predicting biochemical relapse (elevated PSA values after follow-up period) (O) compared with contemporary predictive models (C) for patients that underwent some kind of treatment for prostate cancer (P)?

The current standard for the estimation of biochemical recurrence risk after radical prostatectomy is the CAPRA-S score. An adapted CAPRA-S score, applied on the same hold-out test set used in UC5, achieved an AUC of 0.6766 for biochemical recurrence prediction. Both pre-RP and post-RP scenarios presented for UC5 (Chapters 7 and 8) outperform the current standard, showcasing the potential of radiomics for the task of biochemical recurrence prediction.

Regarding UC6, even though the whole-gland radiomics model (Chapter 7) presents a high AUC performance (0.80), this is not validated in a hold-out test set. Thus, it serves as proof-of-concept for the methodology, but is not ready for further validation in a clinical setting.

11.1.7 RQ 7

Are signatures of radiomic and/or clinical features extracted from pre-operative bi-parametric prostate MRI examinations (I) more accurate in predicting quality-of-life after prostatectomy (given by the EPIC26 score) (O) compared with contemporary prediction techniques (C) for patients that underwent prostatectomy (P)?

This dissertation was not successful in the prediction of quality of life in the form of EPIC26 score in UC7b (Chapters 7 and 8). The author proposes a few justifications:

- Even though this particular use case was not amongst the smallest use cases, 250 patients might not be enough for the task at hand.
- Quality of life is a future event that may depend on other factors and not just anatomical, physiological, radiological or clinical information.
- The ground truth is relatively subjective, even though it is a standardized questionnaire, the answers are dependent on the importance that each patient puts on their urinary/bowel/sexual function and on each patient's willingness to answer honestly about a possible deficiency that, due to societal pressure, may classify them as "less male".
- EPIC26 is a combination of five domain scores: urinary incontinence, urinary irritation, bowel function, sexual function and hormonal function. So, predicting EPIC26 does not give information about each specific domain. the author hypothesises it could be easier to predict a specific domain's score, compared to the five combined in the final EPIC26 score.

11.1.8 RQ 8

Are signatures of radiomic and/or clinical features extracted from pre-operative bi-parametric prostate MRI examinations (I) more accurate in predicting early withdrawal from an active surveillance programme (O) compared with contemporary predictive models (C) for patients enrolled in an active surveillance programme (P)?

The index-lesion radiomics model (Chapter 8) presented a high AUC performance for predicting which patients would be prematurely withdrawn from the active surveillance program due to disease progression. However, in concordance with UCs 3 and 6, this serves uniquely as proof-of-concept of the potential of the radiomics approach for this task and requires further validation before being tested in a clinical environment.

11.1.9 RQ 9

Are voxel-wise radiomic maps extracted from pre-operative bi-parametric prostate MRI examinations (I) capable of performing clinically significant lesion detection and segmentation tasks (O) when compared with contemporary deep learning methods (C) for patients with a suspicion of clinically significant prostate cancer (P)?

The answer is yes. This dissertation presents an innovative methodology (Chapter 9) that is able to achieve a performance similar to current deep learning approaches, using a simpler algorithm that is highly interpretable by nature and needs fewer data than would be necessary to perform a similar task using deep learning.

11.1.10 RQ 10

Are bi-parametric prostate MRI radiomic variables together with clinical information (I) capable of building an accurate causal model to predict optimal treatment choice (O) when compared with contemporary methods (C) for patients with a suspicion of clinically significant prostate cancer (P)?

Overall, the experiment (Chapter 10) was successful as the model achieved high performance on the hold-out test set. However, a lack of prospective validation along with ethical concerns would prevent moving forward with a more thorough validation in a clinical setting.

11.2 Comments on the methodology

11.2.1 Effect of segmentation mask on model performance

Publications extracting radiomic features from the whole prostate gland are rare, but there is evidence to support this method in contrast to the more traditional approach with lesion radiomics (196; 145). Previous work (145) has reported that whole-gland radiomic features have higher stability to segmentation, since gland borders are easier to be defined comparing to delineating a lesion, and that these features produced classifiers with higher performance and less overfitting.

In this work, it was found that different techniques are preferred for different use cases (Figure 11.1). For the aggressiveness prediction in UC2 it was found not only that the whole-gland radiomics outperformed single lesion radiomics but also that models trained with lesion radiomic features were more overfitted than their whole-gland counterparts. This might be justified by the possible presence of regions with predictive power, outside from the index lesion, such as the MRI-invisible lesions reported by van Houdt et al. (230), only discovered at the histopathology analysis. Similar remarks can be made for the prediction of metastasis in UC3.

For both recurrence use cases (UC5 and UC6), whole-gland radiomics outperformed lesion radiomics. Recurrence in prostate cancer patients is typically associated with an incomplete elimination of cancer cells, leading to residual tumors that can continue proliferating. This can occur due to incomplete treatment, undetected micrometastasis or even treatment resistance (231). Either way, a broad look at the entire gland is more informative than focusing on a single lesion since the gland borders should show evidence of characteristics such as extraprostatic extension or lymph node invasion, which are indicative of a high risk of early metastasis and a higher difficulty for obtaining clear resection margins during radical prostatectomy. Similar characteristics will affect the quality of life of each patient post-treatment, which justifies the success of whole-gland radiomics in UC7.

Finally, use case 8, relating to the prediction of disease progression in low aggressiveness patients enrolled in an active surveillance program, was the only one which thrived under lesion radiomics, outperforming its whole-gland counterpart by over 10%.

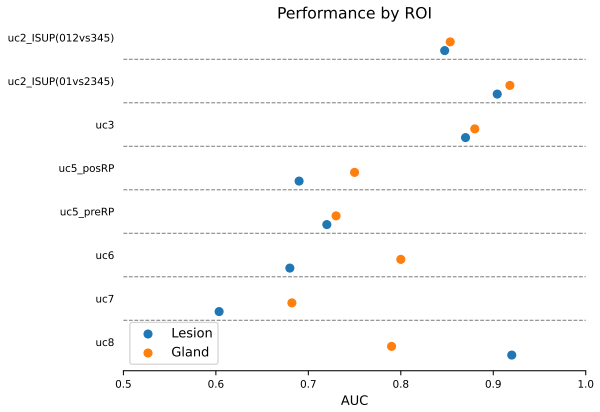


Figure 11.1: Summary of AUC performances for the different clinical use cases that are the subject of this dissertation. The results are colored by type of segmentation mask used in the feature extraction stage (whole-gland in orange and lesion in blue). For use cases 2, 5 and 7, the circle represents hold-out test set performance, while for use cases 3, 6 and 8, it represents cross-validation performance.

11.2.2 Effect of data type on model performance

Figure 11.2 shows the models with the highest ROC-AUC performances for each data type and each use case. The preferred data type for most use cases was radclin (in orange), outperforming

the remaining data types in aggressiveness characterization (UC2), biochemical recurrence after radical prostatectomy (UC5), and disease progression in active surveillance (UC8).

Radiomics alone (in blue) was preferred for the prediction of biochemical recurrence after radiation therapy (UC6) and quality of life after prostatectomy (UC7).

UC3 was the only application where it was clear that the inclusion of deep feature was beneficial to final model performance, as raddeep and hybrid data types outperformed their non-deep counterparts, radiomics and radclin.

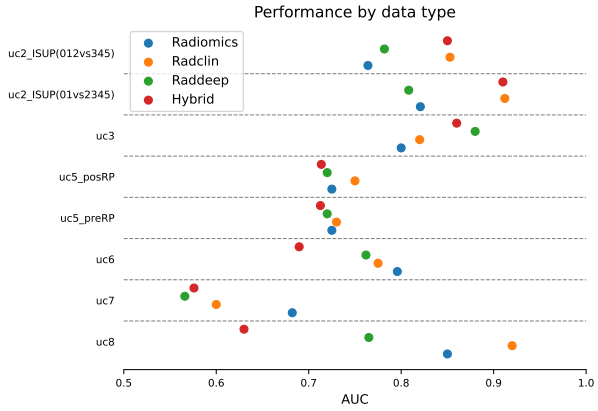


Figure 11.2: AUC performances for the models that reached the highest performance for each data type in each clinical use case. The results are colored by data type. For use cases 2, 5 and 7, the mark represents hold-out test set performance, while for use cases 3, 6 and 8, it represents cross-validation performance.

11.2.3 Effect of MRI vendor on model performance

It was found that training with heterogeneous data, including all scanner manufacturers, consistently outperformed training with each individual vendor (Figure 11.3). This was not the case for the prediction of metastatic development in UC3, where training uniquely with PHILIPS cases outperformed the other settings.

11.2.4 Effect of MRI sequence on model performance

Different MRI sequences assess and display different properties of the tissues. As discussed in Chapter 2, the water content of the tissue is reflected by T2W sequences, the rate of diffusion of water molecules in the tissue is shown by DWI, and the rate at which the DWI signal changes when the b-value parameter is raised is shown by ADC. Therefore, certain MRI sequences may be more or less appropriate to evaluate tissue properties that are predictive of each clinical use case.

It was found that, for the smaller UCs (3, 6, 7, and 8), ADC or DWI alone are the preferred sequences. Whereas for the prediction of disease aggressiveness in UC2, the inclusion of the three sequences outperforms the individual sequences (Figure 11.4).

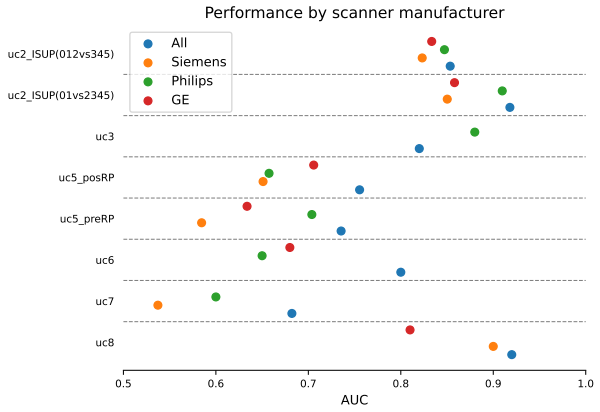


Figure 11.3: AUC performances for the models that reached the highest performance for each scanner manufacturer in each clinical use case. The results are colored by scanner manufacturer. For use cases 2, 5 and 7, the mark represents hold-out test set performance, while for use cases 3, 6 and 8, it represents cross-validation performance.

Regarding UC5, it was found that the pre-surgery prediction thrived with information from the three sequences, while the post-surgery prediction only required T2W to achieve a higher performance. This highlights the importance of the information gathered at the surgery for the model's prediction.

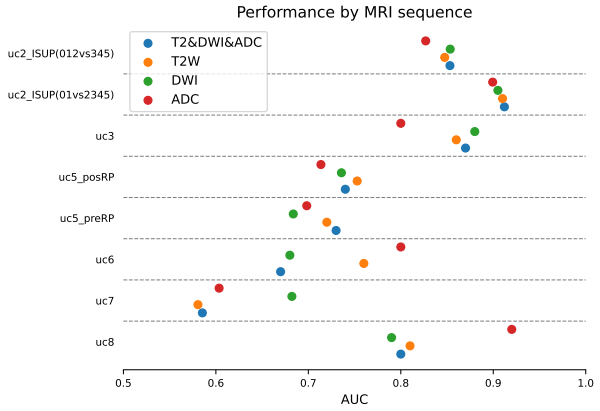


Figure 11.4: AUC performances for the models that reached the highest performance for each MRI sequence in each clinical use case. The results are colored by MRI sequence. For use cases 2 and 5, the mark represents hold-out test set performance, while for use cases 3, 6 and 8, it represents cross-validation performance.

11.3 Comments on the data and data collection process

The ProCAncer-I consortium aimed to make use of data and tools already in use in the clinical setting, to develop AI models. However, as protocols differ between countries and institutions, it was found that, in order to increase the amount of available data, the number of collected variables had to be reduced to a subset that experts in the various institutions collected on their daily patient

assessment. For this reason, information such as family history of cancer or the wheeler ranking was unfortunately not analysed in this dissertation as only very few institutions were able to include it. Similarly, other known risk factors such as patient weight and ethnicity, could have been collected, although the latter might raise some patient privacy concerns.

11.4 Limitations and Future directions

This dissertation has several limitations:

- For the larger use cases (UC2 and UC5), two different machine learning algorithms were employed: LightGBM for radiomics and raddeep data types, and CatBoost for radclin and hybrid data types. As such, for these use cases, it is not known if the differences in performance found between the two data types are in fact due to the differences in the data or if they are due to the different machine learning algorithm.
- The lack of validation of the lesion masks by an expert heavily impacts the comparison of model performance between gland and lesion ROIs, as the lower performance found could be partly due to lower mask quality.
- One of the biggest limitations of this dissertation was the dataset size of the smaller use cases along with the lack of prospective validation. This complicated both the model development and validation, as only cross-validation techniques could be applied to estimate model performance on unseen data.

Future directions involve the deployment of the model developed for UC2 in a clinical platform, so that it can be further tested, evaluating the impact on patients, its cost-effectiveness analysis, and so on.

11.5 Conclusions

This dissertation presents work using radiomics and machine learning techniques for several use cases along the prostate cancer management domain. The author highlights several concluding remarks:

- **Heterogeneous data is beneficial to model generalizability**. Whether it comes from segmentation masks drawn by different radiologists with contrasting years of experience; or including data from different institutions, countries, or scanner manufacturers, heterogeneity is essential for the model to be able to maintain performance upon deployment.
- **An automated pipeline for prostate cancer lesion characterization was successfully developed**. This work demonstrates the power of radiomic features in aggressiveness classification in prostate cancer, using a dataset with unprecedented size and including multiple

centres. The pipeline generalized to the prospective cohort, showing promising results for potentially avoiding unnecessary biopsies.

- **Assessment of model trustworthiness**. Model trustworthiness was a key concern throughout the thesis. Starting with the assessment of how to integrate in the model the real-world variability in manually drawn segmentation masks, to the constant assessment of model explainability and fairness to different population subcohorts throughout the use cases.
- **Development of a radiomics-based lesion detection model**. We have successfully demonstrated the power of voxel-based radiomics in solving aggressive lesion detection task on prostate biparametric MRI. Overcoming challenges such as the tendency for overfitting and the need for large datasets, associated with the use of deep learning methods.
- **Development of an uplift-based model for patient-specific treatment recommendation in prostate cancer**. A causal approach with uplift modeling was taken for the construction of a treatment recommendation system, achieving a high performance.

Appendix A

Appendix for Chapter 4

Table A.1: List of excluded patients.

Prostatex0015	Prostatex0038	Prostatex0074	Prostatex0109	Prostatex0173	Prostatex0201
Prostatex0025	Prostatex0052	Prostatex0101	Prostatex0116	Prostatex0190	Prostatex0202
Prostatex0031	Prostatex0054	Prostatex0102	Prostatex0148	Prostatex0199	Prostatex0203
Prostatex0032	Prostatex0058	Prostatex0107	Prostatex0152	Prostatex0200	

Table A.2: Model characteristics.

Classifier	**Number of features**	**Training Hyperparameters selected during tuning**								**Decision threshold**
		LR penalty	**LR C**	**RF bootstrap**	**RF min_ samples _split**	**RF max_ depth**	**RF min_ samples _leaf**	**RF max_ features**	**RF n_estimators**	
stableRad1	40	L2	1000							0.3739
rad1	30			FALSE	5	10	1	auto	100	0.4543
rad2	30			FALSE	2	10	2	auto	200	0.4446
avgRad	20			FALSE	2	10	1	auto	200	0.2552
unionRad	25	L2	0.01							0.4281
intersectionRad	10			FALSE	2	10	4	auto	200	0.3602
resampledRad	40			FALSE	2	10	1	auto	200	0.3231
Deep	35			TRUE	2	10	1	auto	200	0
Hybrid	30			TRUE	5	10	1	auto	200	0.4906

Appendix B

Appendix for Chapter 6

Table B.1: Co-registration parameter file.

```
// Components
(Registration "MultiResolutionRegistration")
(FixedImagePyramid "FixedRecursiveImagePyramid")
(MovingImagePyramid "MovingRecursiveImagePyramid")
(Interpolator "LinearInterpolator")
(Metric "AdvancedMattesMutualInformation")
(Optimizer "AdaptiveStochasticGradientDescent")
(ResampleInterpolator "FinalBSplineInterpolator")
(Resampler "DefaultResampler")
(Transform "EulerTransform")

// **********Pyramid
(NumberOfResolutions 3)

// **********Transform
(AutomaticTransformInitializationMethod "GeometricCenter")
(AutomaticScalesEstimation "true")

// **********Optimizer
(MaximumNumberOfIterations 300)
(AutomaticParameterEstimation "true")

// **********Several
(WriteTransformParametersEachIteration "false")
(WriteTransformParametersEachResolution "false")
(WriteIterationInfo "false")
(WriteResultImage "true")
(ShowExactMetricValue "false")
(ResultImageFormat "nii")

// **********ImageSampler
(ImageSampler "RandomCoordinate")
(CheckNumberOfSamples "true")
(NewSamplesEveryIteration "true")
(MaximumNumberOfSamplingAttempts 8)
(NumberOfSpatialSamples 2048)
(NumberOfSamplesForExactGradient 4096)
```

```
// **********Interpolator and Resampler
// Order of B-Spline interpolation used for applying the final deformation:
(FinalBSplineInterpolationOrder 3)

// Default pixel value for pixels that come from outside the picture:
(DefaultPixelValue 0)
```

Appendix C

Appendix for Chapter 7

Table C.1: Hyperparameter space used for optimization.

CatBoost Classifier

```
param = {
        'n_estimators': trial.suggest_int('n_estimators', 100, 1000),
        'learning_rate' : trial.suggest_float('learning_rate', 1e-3, 1e-1, log=True),
        'max_depth': trial.suggest_int('max_depth', 2, 10),
        'l2_leaf_reg': trial.suggest_float('l2_leaf_reg', 1e-3, 100, log=True),
        'random_strength': trial.suggest_float('random_strength', 0.2, 10, log=True),
        "objective": trial.suggest_categorical("objective", ["Logloss", "CrossEntropy"]),
        "boosting_type": trial.suggest_categorical("boosting_type", ["Ordered", "Plain"]),
        "bootstrap_type": trial.suggest_categorical("bootstrap_type", ["Bayesian", "Bernoulli", "MVS"]),
        "colsample_bylevel": trial.suggest_float("colsample_bylevel", 0.01, 1),
        'min_data_in_leaf': trial.suggest_int('min_data_in_leaf', 1, 100, log=True),
}
```

LightGBM Classifier

```
param = {
        'n_estimators' : trial.suggest_int('n_estimators', 100, 2000),
        'num_leaves' : trial.suggest_int('num_leaves', 10, 500, log=True),
        'learning_rate' : trial.suggest_float('learning_rate', 1e-3, 1e-1, log=True),
        'max_depth': trial.suggest_int('max_depth', 2, 100),
        'min_child_samples': trial.suggest_int('min_child_samples', 10, 200, log=True),
        'subsample' : trial.suggest_float('subsample', 0, 1),
        'colsample_bytree': trial.suggest_float("colsample_bytree", 0, 1),
        'min_split_gain': trial.suggest_float('min_split_gain', 0, 1),
        'reg_alpha':trial.suggest_float('reg_alpha', 1e-3, 1e1, log=True),
        'reg_lambda':trial.suggest_float('reg_lambda', 1e-3, 1e1, log=True),
}
```

SGD

```
param = {
        'alpha': trial.suggest_float('alpha', 1e-4, 1e3, log=True),
        'max_iter': trial.suggest_int('max_iter', 100, 10000, log=True),
        'learning_rate': trial.suggest_categorical('learning_rate', ['optimal', 'invscaling', 'adaptive']),
        'eta0': trial.suggest_float('eta0', 1e-3, 1e-1, log=True),
        'loss': trial.suggest_categorical('loss', ['log_loss']),
        'penalty': trial.suggest_categorical('penalty', ['l2', 'elasticnet'])
}
```

Bibliography

[1] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[2] Rune Kvåle, Bjørn Møller, Rolf Wahlqvist, Sophie D Fosså, Aasmund Berner, Christer Busch, Anne E Kyrdalen, Aud Svindland, Trond Viset, and Ole J Halvorsen. Concordance between gleason scores of needle biopsies and radical prostatectomy specimens: a population-based study. *BJU international*, 103(12):1647–1654, 2009.

[3] Michael C Haffner, Wilbert Zwart, Martine P Roudier, Lawrence D True, William G Nelson, Jonathan I Epstein, Angelo M De Marzo, Peter S Nelson, and Srinivasan Yegnasubramanian. Genomic and phenotypic heterogeneity in prostate cancer. *Nature Reviews Urology*, 18(2):79–92, 2021.

[4] Harry C Brastianos, Jure Murgic, Adriana Salcedo, Melvin LK Chua, Alice Meng, Michael Fraser, Michael Brundage, Neil E Fleshner, Theodorus van der Kwast, Robert G Bristow, et al. Determining the impact of spatial heterogeneity on genomic prognostic biomarkers for localized prostate cancer. *European Urology Oncology*, 5(3):362–365, 2022.

[5] Stefano Salciccia, Anna Laura Capriotti, Aldo Laganà, Stefano Fais, Mariantonia Logozzi, Ettore De Berardinis, Gian Maria Busetto, Giovanni Battista Di Pierro, Gian Piero Ricciuti, Francesco Del Giudice, et al. Biomarkers in prostate cancer diagnosis: from current knowledge to the role of metabolomics and exosomes. *International journal of molecular sciences*, 22(9):4367, 2021.

[6] Janita E Van Timmeren, Davide Cester, Stephanie Tanadini-Lang, Hatem Alkadhi, and Bettina Baessler. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into imaging*, 11(1):1–16, 2020.

[7] Sorel Kurbegovic, Kasper Drimer Berg, Frederik Birkebæk Thomsen, Lisa Gruschy, Peter Iversen, Klaus Brasso, and Martin Andreas Røder. The risk of biochemical recurrence for intermediate-risk prostate cancer after radical prostatectomy. *Scandinavian journal of urology*, 51(6):450–456, 2017.

[8] Cancer stat facts: Prostate cancer. national cancer institute. https://seer.cancer.gov/statfacts/html/prost.html, 2024. Accessed: 2024-10-22.

[9] Jonathan S Brajtbord, Michael S Leapman, and Matthew R Cooperberg. The capra score at 10 years: contemporary perspectives and analysis of supporting studies. *European urology*, 71(5):705–709, 2017.

[10] Matthew R Cooperberg, Joan F Hilton, and Peter R Carroll. The capra-s score: a straightforward tool for improved prediction of outcomes after radical prostatectomy. *Cancer*, 117(22):5039–5046, 2011.

[11] Hyung L Kim, Ping Li, Huei-Chung Huang, Samineh Deheshi, Tara Marti, Beatrice Knudsen, Hatem Abou-Ouf, Ridwan Alam, Tamara L Lotan, Lucia LC Lam, et al. Validation of the decipher test for predicting adverse pathology in candidates for prostate cancer active surveillance. *Prostate cancer and prostatic diseases*, 22(3):399–405, 2019.

[12] O Bratu, I Oprea, D Marcu, D Spinu, A Niculae, B Geavlete, and D Mischianu. Erectile dysfunction post-radical prostatectomy–a challenge for both patient and physician. *Journal of medicine and life*, 10(1):13, 2017.

[13] Alicja Popiołek, Bartosz Brzoszczyk, Piotr Jarzemski, Małgorzata Piskunowicz, Marcin Jarzemski, Alina Borkowska, and Maciej Bieliński. Quality of life of prostate cancer patients undergoing prostatectomy and affective temperament. *Cancer Management and Research*, pages 1743–1755, 2022.

[14] Jenny L Donovan, Freddie C Hamdy, J Athene Lane, Malcolm Mason, Chris Metcalfe, Eleanor Walsh, Jane M Blazeby, Tim J Peters, Peter Holding, Susan Bonnington, et al. Patient-reported outcomes after monitoring, surgery, or radiotherapy for prostate cancer. *New England Journal of Medicine*, 375(15):1425–1437, 2016.

[15] Nicolas Mottet, Roderick CN van den Bergh, Erik Briers, Thomas Van den Broeck, Marcus G Cumberbatch, Maria De Santis, Stefano Fanti, Nicola Fossati, Giorgio Gandaglia, Silke Gillessen, et al. EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer—2020 update. part 1: screening, diagnosis, and local treatment with curative intent. *European urology*, 79(2):243–262, 2021.

[16] Ana Rodrigues, Nuno Rodrigues, João Santinha, Maria V Lisitskaya, Aycan Uysal, Celso Matos, Inês Domingues, and Nickolas Papanikolaou. Value of handcrafted and deep radiomic features towards training robust machine learning classifiers for prediction of prostate cancer disease aggressiveness. *Scientific Reports*, 13(1):6206, 2023.

[17] Nuno M Rodrigues, José Guilherme de Almeida, Ana Rodrigues, Leonardo Vanneschi, Celso Matos, Maria V Lisitskaya, Aycan Uysal, Sara Silva, and Nickolas Papanikolaou. Deep learning features can improve radiomics-based prostate cancer aggressiveness prediction. *JCO Clinical Cancer Informatics*, 8:e2300180, 2024.

[18] Giulio Del Corso, Eva Pachetti, Rossana Buongiorno, Ana Carolina Rodrigues, Danila Germanese, M Antonietta Pascali, José Almeida, Nuno Rodrigues, Manolis Tsiknakis, Nickolas Papanikolaou, et al. Radiomics-based reliable predictions of side effects after radiotherapy for prostate cancer. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2024.

[19] Ana Rodrigues, Nuno Rodrigues, José Almeida, Inês Domingues, and Nickolas Papanikolaou. Predicting prostate cancer biochemical recurrence after radical prostatectomy with biparametric MRI radiomics: A multicentric study. In *EuSoMII Annual Meeting 2024*, 2024.

[20] Ana Rodrigues, Nuno Rodrigues, José Almeida, Inês Domingues, and Nickolas Papanikolaou. Prostate cancer and radiomics for aggressiveness classification – influence of segmentation mask, MRI sequence, and MRI vendor. In *RSNA 110th Scientific Assembly and Annual Meeting 2024*, 2024.

[21] Ana Rodrigues, Nuno Rodrigues, José Almeida, Ana Gaivão, Carlos Bilreiro, Inês Santiago, Joana Ip, Sara Belião, Inês Domingues, Manolis Tsiknakis, Kostas Marias, Daniele Regge, and Nickolas Papanikolaou. Improving clinically significant prostate cancer detection with a multi-modal machine learning approach: A large-scale multicenter study. *Radiology: Imaging Cancer*, 2025. [Manuscript submitted for publication].

[22] Ana Rodrigues, Nuno Rodrigues, José Almeida, Ana Gaivão, Carlos Bilreiro, Inês Santiago, Joana Ip, Sara Belião, Inês Domingues, Manolis Tsiknakis, Kostas Marias, Daniele Regge, and Nickolas Papanikolaou. Radiomics-based lesion detection and aggressiveness mapping on bpMRI. *IEEE Transactions on Medical Imaging*, 2024. [Manuscript submitted for publication].

[23] Ana Rodrigues, Nuno Rodrigues, José Almeida, Ana Gaivão, Carlos Bilreiro, Inês Santiago, Joana Ip, Sara Belião, Inês Domingues, Manolis Tsiknakis, Kostas Marias, Daniele Regge, and Nickolas Papanikolaou. Uplift modeling for treatment effect estimation in the prostate cancer treatment landscape. *IbPRIA 2025: 12th Iberian Conference on Pattern Recognition and Image Analysis*, 2025.

[24] VL Kumar and PK Majumder. Prostate gland: structure, functions and regulation. *International urology and nephrology*, 27:231–243, 1995.

[25] Paolo Verze, Tommaso Cai, and Stefano Lorenzetti. The role of the prostate in male fertility, health and disease. *Nature Reviews Urology*, 13(7):379–386, 2016.

[26] John E McNeal. The zonal anatomy of the prostate. *The prostate*, 2(1):35–49, 1981.

[27] John E McNeal, Elise A Redwine, Fuad S Freiha, and Thomas A Stamey. Zonal distribution of prostatic adenocarcinoma: correlation with histologic pattern and direction of spread. *The American journal of surgical pathology*, 12(12):897–906, 1988.

[28] LE Johns and RS Houlston. A systematic review and meta-analysis of familial prostate cancer risk. *BJU international*, 91(9):789–794, 2003.

[29] Colin C Pritchard, Joaquin Mateo, Michael F Walsh, Navonil De Sarkar, Wassim Abida, Himisha Beltran, Andrea Garofalo, Roman Gulati, Suzanne Carreira, Rosalind Eeles, et al. Inherited dna-repair gene mutations in men with metastatic prostate cancer. *N Engl J Med*, 375:443–453, 2016.

[30] Colin C Pritchard, Kenneth Offit, and Peter S Nelson. Dna-repair gene mutations in metastatic prostate cancer. *The New England journal of medicine*, 375(18):1804–1805, 2016.

[31] Haley Hieronymus, Nikolaus Schultz, Anuradha Gopalan, Brett S Carver, Matthew T Chang, Yonghong Xiao, Adriana Heguy, Kety Huberman, Melanie Bernstein, Melissa Assel, et al. Copy number alteration burden predicts prostate cancer relapse. *Proceedings of the National Academy of Sciences*, 111(30):11139–11144, 2014.

[32] Brett S Carver, Jennifer Tran, Zhenbang Chen, Arkaitz Carracedo-Perez, Andrea Alimonti, Caterina Nardella, Anuradha Gopalan, Peter T Scardino, Carlos Cordon-Cardo, William Gerald, et al. Ets rearrangements and prostate cancer initiation. *Nature*, 457(7231):E1–E1, 2009.

[33] Andrew S Goldstein, Jiaoti Huang, Changyong Guo, Isla P Garraway, and Owen N Witte. Identification of a cell of origin for human prostate cancer. *Science*, 329(5991):568–571, 2010.

[34] Chee Wai Chua, Maho Shibata, Ming Lei, Roxanne Toivanen, LaMont J Barlow, Sarah K Bergren, Ketan K Badani, James M McKiernan, Mitchell C Benson, Hanina Hibshoosh, et al. Single luminal epithelial progenitors can generate prostate organoids in culture. *Nature cell biology*, 16(10):951–961, 2014.

[35] Suk Hyung Lee and Michael M Shen. Cell types of origin for prostate cancer. *Current opinion in cell biology*, 37:35–41, 2015.

[36] David G Bostwick and Junqi Qian. High-grade prostatic intraepithelial neoplasia. *Modern pathology*, 17(3):360–379, 2004.

[37] Richard B Berish, Aymon N Ali, Patrick G Telmer, John A Ronald, and Hon S Leong. Translational models of prostate cancer bone metastasis. *Nature Reviews Urology*, 15(7):403–421, 2018.

[38] Noel W Clarke, Claire A Hart, and Mick D Brown. Molecular mechanisms of metastasis in prostate cancer. *Asian journal of andrology*, 11(1):57, 2009.

[39] Jacques Ferlay, Murielle Colombet, Isabelle Soerjomataram, Donald M Parkin, Marion Piñeros, Ariana Znaor, and Freddie Bray. Cancer statistics for the year 2020: An overview. *International journal of cancer*, 149(4):778–789, 2021.

[40] Geolani W Dy, John L Gore, Mohammad H Forouzanfar, Mohsen Naghavi, and Christina Fitzmaurice. Global burden of urologic cancers, 1990–2013. *European urology*, 71(3):437–446, 2017.

[41] Martin CS Wong, William B Goggins, Harry HX Wang, Franklin DH Fung, Colette Leung, Samuel YS Wong, Chi Fai Ng, and Joseph JY Sung. Global incidence and mortality for prostate cancer: analysis of temporal patterns and trends in 36 countries. *European urology*, 70(5):862–874, 2016.

[42] Rebecca L Siegel, Kimberly D Miller, Nikita Sandeep Wagle, and Ahmedin Jemal. Cancer statistics, 2023. *CA: a cancer journal for clinicians*, 73(1):17–48, 2023.

[43] Hans Lilja, David Ulmert, and Andrew J Vickers. Prostate-specific antigen and prostate cancer: prediction, detection and monitoring. *Nature Reviews Cancer*, 8(4):268–278, 2008.

[44] Matthew B Gretzer and Alan W Partin. PSA markers in prostate cancer detection. *Urologic Clinics*, 30(4):677–686, 2003.

[45] Peter C Albertsen. Prostate cancer screening and treatment: where have we come from and where are we going? *BJU international*, 126(2):218–224, 2020.

[46] Dragan Ilic, Mia Djulbegovic, Jae Hung Jung, Eu Chang Hwang, Qi Zhou, Anne Cleves, Thomas Agoritsas, and Philipp Dahm. Prostate cancer screening with prostate-specific antigen (PSA) test: a systematic review and meta-analysis. *bmj*, 362, 2018.

[47] Nicolas Mottet, Joaquim Bellmunt, Michel Bolla, Erik Briers, Marcus G Cumberbatch, Maria De Santis, Nicola Fossati, Tobias Gross, Ann M Henry, Steven Joniau, et al. EAU-ESTRO-SIOG guidelines on prostate cancer. part 1: screening, diagnosis, and local treatment with curative intent. *European urology*, 71(4):618–629, 2017.

[48] Timothy J Coker and Daniel Dierfeldt. Acute bacterial prostatitis: diagnosis and management. *American family physician*, 93(2):114–120, 2016.

[49] Eric H Kim, Jeffrey A Larson, and Gerald L Andriole. Management of benign prostatic hyperplasia. *Annual review of medicine*, 67:137–151, 2016.

[50] Sadhna Verma and Arumugam Rajesh. A clinically relevant approach to imaging prostate cancer. *American Journal of Roentgenology*, 196(3_supplement):S1–S10, 2011.

[51] Ian M Thompson, Donna K Pauler, Phyllis J Goodman, Catherine M Tangen, M Scott Lucia, Howard L Parnes, Lori M Minasian, Leslie G Ford, Scott M Lippman, E David Crawford, et al. Prevalence of prostate cancer among men with a prostate-specific antigen level

lower or equal to 4.0 ng per milliliter. *New England Journal of Medicine*, 350(22):2239–2246, 2004.

[52] Edith Canby-Hagino, Javier Hernandez, Timothy C Brand, Dean A Troyer, Betsy Higgins, Donna Pauler Ankerst, Ian M Thompson, Robin J Leach, and Dipen J Parekh. Prostate cancer risk with positive family history, normal prostate examination findings, and PSA less than 4.0 ng/ml. *Urology*, 70(4):748–752, 2007.

[53] Stacy Loeb, Marc A Bjurlin, Joseph Nicholson, Teuvo L Tammela, David F Penson, H Ballentine Carter, Peter Carroll, and Ruth Etzioni. Overdiagnosis and overtreatment of prostate cancer. *European urology*, 65(6):1046–1055, 2014.

[54] RC Macefield, C Metcalfe, JA Lane, JL Donovan, KNL Avery, JM Blazeby, L Down, DE Neal, FC Hamdy, and K Vedhara. Impact of prostate cancer testing: an evaluation of the emotional consequences of a negative biopsy result. *British journal of cancer*, 102(9):1335–1340, 2010.

[55] Veeru Kasivisvanathan, Antti S Rannikko, Marcelo Borghi, Valeria Panebianco, Lance A Mynderse, Markku H Vaarala, Alberto Briganti, Lars Budäus, Giles Hellawell, Richard G Hindley, et al. MRI-targeted or standard biopsy for prostate-cancer diagnosis. *New England Journal of Medicine*, 378(19):1767–1777, 2018.

[56] Hashim U Ahmed, Ahmed El-Shater Bosaily, Louise C Brown, Rhian Gabe, Richard Kaplan, Mahesh K Parmar, Yolanda Collaco-Moraes, Katie Ward, Richard G Hindley, Alex Freeman, et al. Diagnostic accuracy of multi-parametric MRI and trus biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *The Lancet*, 389(10071):815–822, 2017.

[57] Morgan R Pokorny, Maarten De Rooij, Earl Duncan, Fritz H Schröder, Robert Parkinson, Jelle O Barentsz, and Leslie C Thompson. Prospective study of diagnostic accuracy comparing prostate cancer detection by transrectal ultrasound–guided biopsy versus magnetic resonance (MR) imaging with subsequent MR-guided biopsy in men without previous prostate biopsies. *European urology*, 66(1):22–29, 2014.

[58] Armando Stabile, Francesco Giganti, Mark Emberton, and Caroline M Moore. MRI in prostate cancer diagnosis: do we need to add standard sampling? a review of the last 5 years. *Prostate Cancer and Prostatic Diseases*, 21(4):473–487, 2018.

[59] Les C Thompson and Morgan R Pokorny. Multiparametric MRI in the diagnosis of prostate cancer-a generational change. *Australian Family Physician*, 44(8):597–602, 2015.

[60] Baris Turkbey, Peter A Pinto, Haresh Mani, Marcelino Bernardo, Yuxi Pang, Yolanda L McKinney, Kiranpreet Khurana, Gregory C Ravizzini, Paul S Albert, Maria J Merino, et al. Prostate cancer: value of multiparametric mr imaging at 3 t for detection—histopathologifc correlation. *Radiology*, 255(1):89–99, 2010.

[61] Junqian Xu, Peter A Humphrey, Adam S Kibel, Abraham Z Snyder, Vamsidhar R Narra, Joseph JH Ackerman, and Sheng-Kwei Song. Magnetic resonance diffusion characteristics of histologically defined prostate cancer in humans. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 61(4):842–850, 2009.

[62] Anwar R Padhani, Guoying Liu, Dow Mu-Koh, Thomas L Chenevert, Harriet C Thoeny, Taro Takahara, Andrew Dzik-Jurasz, Brian D Ross, Marc Van Cauteren, David Collins, et al. Diffusion-weighted magnetic resonance imaging as a cancer biomarker: consensus and recommendations. *Neoplasia*, 11(2):102–125, 2009.

[63] Jussi Toivonen, Harri Merisaari, Marko Pesola, Pekka Taimen, Peter J Boström, Tapio Pahikkala, Hannu J Aronen, and Ivan Jambor. Mathematical models for diffusion-weighted imaging of prostate cancer using b values up to 2000 s/mm2: Correlation with gleason score and repeatability of region of interest analysis. *Magnetic resonance in medicine*, 74(4):1116–1124, 2015.

[64] Nicola Schieda, Christopher S Lim, Fatemeh Zabihollahy, Jorge Abreu-Gomez, Satheesh Krishna, Sungmin Woo, Gerd Melkus, Eran Ukwatta, and Baris Turkbey. Quantitative prostate MRI. *Journal of Magnetic Resonance Imaging*, 53(6):1632–1645, 2021.

[65] Baris Turkbey, Vijay P Shah, Yuxi Pang, Marcelino Bernardo, Sheng Xu, Jochen Kruecker, Julia Locklin, Angelo A Baccala Jr, Ardeshir R Rastinehad, Maria J Merino, et al. Is apparent diffusion coefficient associated with clinical risk scores for prostate cancers that are visible on 3-t mr images? *Radiology*, 258(2):488–495, 2011.

[66] Meltem Esen, Mehmet Ruhi Onur, Nusret Akpolat, Irfan Orhan, and Ercan Kocakoc. Utility of ADC measurement on diffusion-weighted MRI in differentiation of prostate cancer, normal prostate and prostatitis. *Quantitative imaging in medicine and surgery*, 3(4):210, 2013.

[67] Nicola Schieda, Satheesh Krishna, and Matthew S Davenport. Update on gadolinium-based contrast agent–enhanced imaging in the genitourinary system. *American Journal of Roentgenology*, 212(6):1223–1233, 2019.

[68] Jiahui Zhang, Lili Xu, Gumuyang Zhang, Xiaoxiao Zhang, Xin Bai, Zhigang Ji, Yu Xiao, Hao Sun, and Zhengyu Jin. Comparison between biparametric and multiparametric MRI diagnosis strategy for prostate cancer in the peripheral zone using PI-RADS version 2.1. *Abdominal Radiology*, 47(8):2905–2916, 2022.

[69] Sungmin Woo, Chong Hyun Suh, Sang Youn Kim, Jeong Yeon Cho, Seung Hyup Kim, and Min Hoan Moon. Head-to-head comparison between biparametric and multiparametric MRI for the diagnosis of prostate cancer: a systematic review and meta-analysis. *American Journal of Roentgenology*, pages W226–W241, 2018.

[70] Baris Turkbey, Andrew B Rosenkrantz, Masoom A Haider, Anwar R Padhani, Geert Villeirs, Katarzyna J Macura, Clare M Tempany, Peter L Choyke, Francois Cornud, Daniel J Margolis, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. *European urology*, 76(3):340–351, 2019.

[71] Jelle O Barentsz, Jonathan Richenberg, Richard Clements, Peter Choyke, Sadhna Verma, Geert Villeirs, Olivier Rouviere, Vibeke Logager, and Jurgen J Fütterer. Esur prostate mr guidelines 2012. *European radiology*, 22:746–757, 2012.

[72] Sungmin Woo, Chong Hyun Suh, Sang Youn Kim, Jeong Yeon Cho, and Seung Hyup Kim. Diagnostic performance of prostate imaging reporting and data system version 2 for detection of prostate cancer: a systematic review and diagnostic meta-analysis. *European urology*, 72(2):177–188, 2017.

[73] Clayton P Smith, Stephanie A Harmon, Tristan Barrett, Leonardo K Bittencourt, Yan Mee Law, Haytham Shebel, Julie Y An, Marcin Czarniecki, Sherif Mehralivand, Mehmet Coskun, et al. Intra-and interreader reproducibility of PI-RADSv2: A multireader study. *Journal of Magnetic Resonance Imaging*, 49(6):1694–1703, 2019.

[74] Antonio C Westphalen, Charles E McCulloch, Jordan M Anaokar, Sandeep Arora, Nimrod S Barashi, Jelle O Barentsz, Tharakeswara K Bathala, Leonardo K Bittencourt, Michael T Booker, Vaughn G Braxton, et al. Variability of the positive predictive value of PI-RADS for prostate MRI across 26 centers: experience of the society of abdominal radiology prostate cancer disease-focused panel. *Radiology*, 296(1):76–84, 2020.

[75] Matthew D Greer, Joanna H Shih, Nathan Lay, Tristan Barrett, Leonardo Bittencourt, Samuel Borofsky, Ismail Kabakus, Yan Mee Law, Jamie Marko, Haytham Shebel, et al. Interreader variability of prostate imaging reporting and data system version 2 in detecting and assessing prostate cancer lesions at prostate MRI. *AJR. American journal of roentgenology*, page 1, 2019.

[76] Kimia Kohestani, Jonas Wallström, Niclas Dehlfors, Ole Martin Sponga, Marianne Månsson, Andreas Josefsson, Sigrid Carlsson, Mikael Hellström, and Jonas Hugosson. Performance and inter-observer variability of prostate MRI (PI-RADS version 2) outside high-volume centres. *Scandinavian journal of urology*, 53(5):304–311, 2019.

[77] Yin Lei, Tian Jie Li, Peng Gu, Yu kun Yang, Lei Zhao, Chao Gao, Juan Hu, and Xiao Dong Liu. Combining prostate-specific antigen density with prostate imaging reporting and data system score version 2.1 to improve detection of clinically significant prostate cancer: A retrospective study. *Frontiers in Oncology*, 12, 2022.

[78] Geert JLH Van Leenders, Theodorus H Van Der Kwast, David J Grignon, Andrew J Evans, Glen Kristiansen, Charlotte F Kweldam, Geert Litjens, Jesse K McKenney, Jonathan

Melamed, Nicholas Mottet, et al. The 2019 international society of urological pathology (ISUP) consensus conference on grading of prostatic carcinoma. *The American journal of surgical pathology*, 44(8):e87, 2020.

[79] Theresa Y Chan, Alan W Partin, Patrick C Walsh, and Jonathan I Epstein. Prognostic significance of gleason score 3+ 4 versus gleason score 4+ 3 tumor at radical prostatectomy. *Urology*, 56(5):823–827, 2000.

[80] Jennifer R Stark, Sven Perner, Meir J Stampfer, Jennifer A Sinnott, Stephen Finn, Anna S Eisenstein, Jing Ma, Michelangelo Fiorentino, Tobias Kurth, Massimo Loda, et al. Gleason score and lethal prostate cancer: does 3+ 4= 4+ 3? *Journal of Clinical Oncology*, 27(21):3459, 2009.

[81] Richard Choo, Laurence Klotz, Cyril Danjoux, Gerard C Morton, Gerrit DeBoer, Ewa Szumacher, Neil Fleshner, Peter Bunting, and George Hruby. Feasibility study: watchful waiting for localized low to intermediate grade prostate carcinoma with selective delayed intervention based on prostate specific antigen, histological and/or clinical progression. *The Journal of urology*, 167(4):1664–1669, 2002.

[82] Colton H Walker, Kathryn A Marchetti, Udit Singhal, and Todd M Morgan. Active surveillance for prostate cancer: selection criteria, guidelines, and outcomes. *World Journal of Urology*, pages 1–8, 2021.

[83] Manuela A Hoffmann, Helmut J Wieler, Christian Baues, Nicholas J Kuntz, Ines Richardsen, and Mathias Schreckenberger. The impact of 68ga-psma PET/CT and PET/MRI on the management of prostate cancer. *Urology*, 130:1–12, 2019.

[84] Antônio Rebello Horta Görgen, Lucas Medeiros Burttet, Eduardo Tosetto Cachoeira, Pedro Glusman Knijnik, Pietro Waltrick Brum, Artur de Oliveira Paludo, Renan Desimon Cabral, Tiago Elias Rosito, Milton Berger, Christian Paul Pavlovich, et al. Association of nerve-sparing grading in robotic radical prostatectomy and trifecta outcome. *World journal of urology*, pages 1–6, 2022.

[85] Hiroki Komori, Leandro Blas, Masaki Shiota, Dai Takamatsu, Takashi Matsumoto, Ken Lee, Keisuke Monji, Eiji Kashiwagi, Junichi Inokuchi, and Masatoshi Eto. Impact of nerve sparing in robot-assisted radical prostatectomy on the risk of positive surgical margin and biochemical recurrence. *International Journal of Urology*, 29(8):824–829, 2022.

[86] Nicola Fossati, Peter-Paul M Willemse, Thomas Van den Broeck, Roderick CN van den Bergh, Cathy Yuhong Yuan, Erik Briers, Joaquim Bellmunt, Michel Bolla, Philip Cornford, Maria De Santis, et al. The benefits and harms of different extents of lymph node dissection during radical prostatectomy for prostate cancer: a systematic review. *European urology*, 72(1):84–109, 2017.

[87] Vérane Achard, Cédric Michael Panje, Daniel Engeler, Thomas Zilli, and Paul Martin Putora. Localized and locally advanced prostate cancer: treatment options. *Oncology*, 99(7):413–421, 2021.

[88] Mamello Sekhoacha, Keamogetswe Riet, Paballo Motloung, Lemohang Gumenku, Ayodeji Adegoke, and Samson Mashele. Prostate cancer review: Genetics, diagnosis, treatment options, and alternative approaches. *Molecules*, 27(17):5730, 2022.

[89] Ian G Mills. Maintaining and reprogramming genomic androgen receptor activity in prostate cancer. *Nature Reviews Cancer*, 14(3):187–198, 2014.

[90] Alastair D Lamb, Charlie E Massie, and David E Neal. The transcriptional programme of the androgen receptor (ar) in prostate cancer. *BJU Int*, 113(3):358–366, 2014.

[91] Philip Cornford, Joaquim Bellmunt, Michel Bolla, Erik Briers, Maria De Santis, Tobias Gross, Ann M Henry, Steven Joniau, Thomas B Lam, Malcolm D Mason, et al. EAU-ESTRO-SIOG guidelines on prostate cancer. part ii: treatment of relapsing, metastatic, and castration-resistant prostate cancer. *European urology*, 71(4):630–642, 2017.

[92] MR Smith, F Kabbinavar, F Saad, A Hussain, Y-J Hei, E Small, R Cook, and C Higano. Natural history of rising serum PSA in men with castrate nonmetastatic prostate cancer. *Journal of Clinical Oncology*, 23(16_suppl):4514–4514, 2005.

[93] Siu Chiu Chan and Scott M Dehm. Constitutive activity of the androgen receptor. *Advances in pharmacology*, 70:327–366, 2014.

[94] Diane M Robins. Androgen receptor gene polymorphisms and alterations in prostate cancer: of humanized mice and men. *Molecular and cellular endocrinology*, 352(1-2):26–33, 2012.

[95] Silke Gillessen, Gerhardt Attard, Tomasz M Beer, Himisha Beltran, Anders Bjartell, Alberto Bossi, Alberto Briganti, Rob G Bristow, Kim N Chi, Noel Clarke, et al. Management of patients with advanced prostate cancer: report of the advanced prostate cancer consensus conference 2019. *European urology*, 77(4):508–547, 2020.

[96] Thomas Van den Broeck, Roderick CN van den Bergh, Erik Briers, Philip Cornford, Marcus Cumberbatch, Derya Tilki, Maria De Santis, Stefano Fanti, Nicola Fossati, Silke Gillessen, et al. Biochemical recurrence in prostate cancer: the european association of urology prostate cancer guidelines panel recommendations. *European urology focus*, 6(2):231–234, 2020.

[97] Gerard Nkengurutse, Feng Tian, Sixiong Jiang, Qi Wang, Ying Wang, and Weibing Sun. Preoperative predictors of biochemical recurrence-free survival in high-risk prostate cancer following radical prostatectomy. *Frontiers in Oncology*, 10:1761, 2020.

[98] Stephen A Boorjian, R Jeffrey Karnes, Paul L Crispen, Laureano J Rangel, Eric J Bergstralh, and Michael L Blute. Radiation therapy after radical prostatectomy: impact on metastasis and survival. *The Journal of urology*, 182(6):2708–2715, 2009.

[99] Andrew Kneebone, Carol Fraser-Browne, Gillian M Duchesne, Richard Fisher, Mark Frydenberg, Alan Herschtal, Scott G Williams, Chris Brown, Warick Delprado, Annette Haworth, et al. Adjuvant radiotherapy versus early salvage radiotherapy following radical prostatectomy (trog 08.03/anzup raves): a randomised, controlled, phase 3, non-inferiority trial. *The Lancet Oncology*, 21(10):1331–1340, 2020.

[100] Jonathan Bergman and Mark S Litwin. Quality of life in men undergoing active surveillance for localized prostate cancer. *Journal of the National Cancer Institute Monographs*, 2012(45):242–249, 2012.

[101] Deborah P Lubeck, Gary D Grossfeld, and Peter R Carroll. The effect of androgen deprivation therapy on health-related quality of life in men with prostate cancer. *Urology*, 58(2):94–99, 2001.

[102] Paul L Nguyen, Shabbir MH Alibhai, Shehzad Basaria, Anthony V D'Amico, Philip W Kantoff, Nancy L Keating, David F Penson, Derek J Rosario, Bertrand Tombal, and Matthew R Smith. Adverse effects of androgen deprivation therapy and strategies to mitigate them. *European urology*, 67(5):825–836, 2015.

[103] Richard J. Rebello, Christoph Oing, Karen E. Knudsen, Stacy Loeb, David C. Johnson, Robert E. Reiter, Silke Gillessen, Theodorus Van der Kwast, and Robert G. Bristow. Prostate cancer. *Nature Reviews Disease Primers*, 7(1), February 2021.

[104] Sam Watts, Geraldine Leydon, Brian Birch, Philip Prescott, Lily Lai, Susan Eardley, and George Lewith. Depression and anxiety in prostate cancer: a systematic review and meta-analysis of prevalence rates. *BMJ open*, 4(3):e003901, 2014.

[105] AC Wootten, JM Abbott, A Farrell, DW Austin, and B Klein. Psychosocial interventions to support partners of men with prostate cancer: a systematic and critical review of the literature. *Journal of Cancer Survivorship*, 8:472–484, 2014.

[106] Marigdalia K. Ramirez-Fort, Paula Suarez, Margely Carrion, Daniel Weiner, Claire Postl, Ricardo Arribas, Mehdi Sayyah, Digna V. Forta, M. Junaid Niaz, Amir Feily, Christopher S. Lange, Zhahedia Zhaythseff Fort, and Migdalia Fort. Prostatic irradiation-induced sexual dysfunction: A review and multidisciplinary guide to management in the radical radiotherapy era (part III on psychosexual therapy and the masculine self-esteem). *Reports of Practical Oncology Radiotherapy*, 25(4):625–631, July 2020.

[107] Clayton P Smith, Marcin Czarniecki, Sherif Mehralivand, Radka Stoyanova, Peter L Choyke, Stephanie Harmon, and Baris Turkbey. Radiomics and radiogenomics of prostate cancer. *Abdominal Radiology*, 44:2021–2029, 2019.

[108] Matteo Ferro, Ottavio de Cobelli, Gennaro Musi, Francesco Del Giudice, Giuseppe Carrieri, Gian Maria Busetto, Ugo Giovanni Falagario, Alessandro Sciarra, Martina Maggi, Felice Crocetto, et al. Radiomics in prostate cancer: An up-to-date review. *Therapeutic Advances in Urology*, 14:17562872221109020, 2022.

[109] Camilla Scapicchio, Michela Gabelloni, Andrea Barucci, Dania Cioni, Luca Saba, and Emanuele Neri. A deep look into radiomics. *La radiologia medica*, 126(10):1296–1311, 2021.

[110] Honglan Mi, Mingyuan Yuan, Shiteng Suo, Jiejun Cheng, Suqin Li, Shaofeng Duan, and Qing Lu. Impact of different scanners and acquisition parameters on robustness of MR radiomics features based on women's cervix. *Scientific reports*, 10(1):20407, 2020.

[111] Fabrizio Urraro, Valerio Nardone, Alfonso Reginelli, Carlo Varelli, Antonio Angrisani, Vittorio Patanè, Luca D'Ambrosio, Pietro Roccatagliata, Gaetano Maria Russo, Luigi Gallo, et al. MRI radiomics in prostate cancer: A reliability study. *Frontiers in Oncology*, 11, 2021.

[112] Clément Bailly, Caroline Bodet-Milin, Solène Couespel, Hatem Necib, Françoise Kraeber-Bodéré, Catherine Ansquer, and Thomas Carlier. Revisiting the robustness of PET-based textural features in the context of multi-centric trials. *PloS one*, 11(7):e0159984, 2016.

[113] Linda Bianchini, João Santinha, Nuno Loução, Mário Figueiredo, Francesca Botta, Daniela Origgi, Marta Cremonesi, Enrico Cassano, Nikolaos Papanikolaou, and Alessandro Lascialfari. A multicenter study on radiomic features from t2-weighted images of a customized MR pelvic phantom setting the basis for robust radiomic models in clinics. *Magnetic resonance in medicine*, 85(3):1713–1726, 2021.

[114] Hajar Moradmand, Seyed Mahmoud Reza Aghamiri, and Reza Ghaderi. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. *Journal of applied clinical medical physics*, 21(1):179–190, 2020.

[115] P Steenbergen, K Haustermans, F Pos, R Oyen, S Heijmink, L De Wever, R Kalisvaart, J Teertstra, L Van den Bergh, and U Van der Heide. Prostate tumor delineation using multiparametric MRI: Inter observer variability and pathology validation. *Radiotherapy and Oncology*, 111:S53–S54, 2014.

[116] Binsheng Zhao. Understanding sources of variation to improve the reproducibility of radiomics. *Frontiers in oncology*, page 826, 2021.

[117] Zia Khan, Norashikin Yahya, Khaled Alsaih, Mohammed Isam Al-Hiyali, and Fabrice Meriaudeau. Recent automatic segmentation algorithms of MRI prostate regions: a review. *IEEE Access*, 9:97878–97905, 2021.

[118] Loïc Duron, Daniel Balvay, Saskia Vande Perre, Afef Bouchouicha, Julien Savatovsky, Jean-Claude Sadik, Isabelle Thomassin-Naggara, Laure Fournier, and Augustin Lecler. Gray-level discretization impacts reproducible MRI radiomics texture features. *PLoS One*, 14(3):e0213459, 2019.

[119] Florent Tixier, Mathieu Hatt, Catherine Cheze Le Rest, Adrien Le Pogam, Laurent Corcos, and Dimitris Visvikis. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18f-fdg PET. *Journal of Nuclear Medicine*, 53(5):693–700, 2012.

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

[121] Robert M Haralick, K Shanmugam, and Its'hak Dinstein. Texture analysis using gray level run lengths. *IEEE Transactions on pattern analysis and machine intelligence*, 9(1):58–63, 1990.

[122] Peter Burt and Edward Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1986.

[123] David Marr and E Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217, 1980.

[124] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In *European Conference on Computer Vision*, pages 469–481. Springer, 2004.

[125] Michael Patrick Allen. The problem of multicollinearity. *Understanding regression analysis*, pages 176–180, 1997.

[126] Visar Berisha, Chelsea Krantsevich, P Richard Hahn, Shira Hahn, Gautam Dasarathy, Pavan Turaga, and Julie Liss. Digital medicine and the curse of dimensionality. *NPJ digital medicine*, 4(1):153, 2021.

[127] Damien François, Vincent Wertz, and Michel Verleysen. The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):873–886, 2007.

[128] Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79, 2018.

[129] Bruce A Draper and Hajime B Hirshman. Feature selection for computer vision tasks. In *Proceedings of the SPIE*, volume 1708, pages 312–325. International Society for Optics and Photonics, 1992.

[130] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1:81–106, 1986.

[131] Isabelle Guyon and Andre Elisseeff. Feature selection through recursive feature elimination. *Journal of Machine Learning Research*, 3:1157–1182, 2002.

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

[133] Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. Classification and regression trees. *Chapman & Hall/CRC*, 1984.

[134] Susan Athey and Guido W Imbens. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050(5):1–26, 2015.

[135] Renato Cuocolo, Maria Brunella Cipullo, Arnaldo Stanzione, Valeria Romeo, Roberta Green, Valeria Cantoni, Andrea Ponsiglione, Lorenzo Ugga, and Massimo Imbriaco. Machine learning for the identification of clinically significant prostate cancer on MRI: a meta-analysis. *European Radiology*, 30:6877–6887, 2020.

[136] Arnaldo Stanzione, Michele Gambardella, Renato Cuocolo, Andrea Ponsiglione, Valeria Romeo, and Massimo Imbriaco. Prostate MRI radiomics: A systematic review and radiomic quality score assessment. *European journal of radiology*, 129:109095, 2020.

[137] Giuseppe Cutaia, Giuseppe La Tona, Albert Comelli, Federica Vernuccio, Francesco Agnello, Cesare Gagliardo, Leonardo Salvaggio, Natale Quartuccio, Letterio Sturiale, Alessandro Stefano, et al. Radiomics and prostate MRI: current role and future applications. *Journal of Imaging*, 7(2):34, 2021.

[138] Federico Midiri, Federica Vernuccio, Pierpaolo Purpura, Pierpaolo Alongi, and Tommaso Vincenzo Bartolotta. Multiparametric MRI and radiomics in prostate cancer: A review of the current literature. *Diagnostics*, 11(10):1829, 2021.

[139] Huanye Li, Chau Hung Lee, David Chia, Zhiping Lin, Weimin Huang, and Cher Heng Tan. Machine learning in prostate MRI for prostate cancer: Current status and future opportunities. *Diagnostics*, 12(2):289, 2022.

[140] Linda My Huynh, Yeagyeong Hwang, Olivia Taylor, and Michael J Baine. The use of MRI-derived radiomic models in prostate cancer risk stratification: A critical review of contemporary literature. *Diagnostics*, 13(6):1128, 2023.

[141] Xiangde Min, Min Li, Di Dong, Zhaoyan Feng, Peipei Zhang, Zan Ke, Huijuan You, Fangfang Han, He Ma, Jie Tian, et al. Multi-parametric MRI-based radiomics signature for discriminating between clinically significant and insignificant prostate cancer: Cross-validation of a machine learning method. *European journal of radiology*, 115:16–21, 2019.

[142] Mengjuan Li, Tong Chen, Wenlu Zhao, Chaogang Wei, Xiaobo Li, Shaofeng Duan, Libiao Ji, Zhihua Lu, and Junkang Shen. Radiomics prediction model for the improved diagnosis of clinically significant prostate cancer on biparametric MRI. *Quantitative Imaging in Medicine and Surgery*, 10(2):368, 2020.

[143] Jeroen Bleker, Thomas C Kwee, Rudi AJO Dierckx, Igle Jan de Jong, Henkjan Huisman, and Derya Yakar. Multiparametric MRI and auto-fixed volume of interest-based radiomics signature for clinically significant peripheral zone prostate cancer. *European radiology*, 30:1313–1324, 2020.

[144] Lixin Gong, Min Xu, Mengjie Fang, Bingxi He, Hailin Li, Xiangming Fang, Di Dong, and Jie Tian. The potential of prostate gland radiomic features in identifying the gleason score. *Computers in biology and medicine*, 144:105318, 2022.

[145] Ana Rodrigues, João Santinha, Bernardo Galvão, Celso Matos, Francisco M Couto, and Nickolas Papanikolaou. Prediction of prostate cancer disease aggressiveness using bi-parametric mri radiomics. *Cancers*, 13(23):6065, 2021.

[146] Jose M Castillo T, Martijn PA Starmans, Muhammad Arif, Wiro J Niessen, Stefan Klein, Chris H Bangma, Ivo G Schoots, and Jifke F Veenland. A multi-center, multi-vendor study to evaluate the generalizability of a radiomics model for classifying prostate cancer: high grade vs. low grade. *Diagnostics*, 11(2):369, 2021.

[147] Yongsheng Zhang, Wen Chen, Xianjie Yue, Jianliang Shen, Chen Gao, Peipei Pang, Feng Cui, and Maosheng Xu. Development of a novel, multi-parametric, MRI-based radiomic nomogram for differentiating between clinically significant and insignificant prostate cancer. *Frontiers in oncology*, 10:888, 2020.

[148] Sangwon Han, Jungsu S Oh, and Jong Jin Lee. Diagnostic performance of deep learning models for detecting bone metastasis on whole-body bone scan in prostate cancer. *European Journal of Nuclear Medicine and Molecular Imaging*, pages 1–11, 2022.

[149] Xiang Liu, Chao Han, Yingpu Cui, Tingting Xie, Xiaodong Zhang, and Xiaoying Wang. Detection and segmentation of pelvic bones metastases in MRI images for patients with prostate cancer based on deep learning. *Frontiers in Oncology*, 11:773299, 2021.

[150] Wenjie Zhang, Ning Mao, Yongsheng Wang, Haizhu Xie, Shaofeng Duan, Xuexi Zhang, and Bin Wang. A radiomics nomogram for predicting bone metastasis in newly diagnosed prostate cancer patients. *European Journal of Radiology*, 128:109020, 2020.

[151] Frederik Wessels, Max Schmitt, Eva Krieghoff-Henning, Tanja Jutzi, Thomas S Worst, Frank Waldbillig, Manuel Neuberger, Roman C Maron, Matthias Steeg, Timo Gaiser, et al. Deep learning approach to predict lymph node metastasis directly from primary tumour histology in prostate cancer. *BJU international*, 128(3):352–360, 2021.

[152] Ying Hou, Jie Bao, Yang Song, Mei-Ling Bao, Ke-Wen Jiang, Jing Zhang, Guang Yang, Chun-Hong Hu, Hai-Bin Shi, Xi-Ming Wang, et al. Integration of clinicopathologic identification and deep transferrable image feature representation improves predictions of lymph node metastasis in prostate cancer. *EBioMedicine*, 68, 2021.

[153] Yueren Wang, Bing Yu, Fei Zhong, Qiyong Guo, Kexin Li, Yang Hou, and Nan Lin. MRI-based texture analysis of the primary tumor for pre-treatment prediction of bone metastases in prostate cancer. *Magnetic Resonance Imaging*, 60:76–84, 2019.

[154] Xiang Liu, Xiangpeng Wang, Yaofeng Zhang, Zhaonan Sun, Xiaodong Zhang, and Xiaoying Wang. Preoperative prediction of pelvic lymph nodes metastasis in prostate cancer using an ADC-based radiomics model: comparison with clinical nomograms and PI-RADS assessment. *Abdominal Radiology*, 47(9):3327–3337, 2022.

[155] Nikita Sushentsev, Leonardo Rundo, Oleg Blyuss, Vincent J Gnanapragasam, Evis Sala, and Tristan Barrett. MRI-derived radiomics model for baseline prediction of prostate cancer progression on active surveillance. *Scientific Reports*, 11(1):12917, 2021.

[156] Nikita Sushentsev, Leonardo Rundo, Luis Abrego, Zonglun Li, Tatiana Nazarenko, Anne Y Warren, Vincent J Gnanapragasam, Evis Sala, Alexey Zaikin, Tristan Barrett, et al. Time series radiomics for the prediction of prostate cancer progression in patients on active surveillance. *European Radiology*, pages 1–9, 2023.

[157] Vincent Bourbonne, Martin Vallières, François Lucia, Laurent Doucet, Dimitris Visvikis, Valentin Tissot, Olivier Pradier, Mathieu Hatt, and Ulrike Schick. MRI-derived radiomics to guide post-operative management for high-risk prostate cancer. *Frontiers in oncology*, 9:807, 2019.

[158] Vincent Bourbonne, Georges Fournier, Martin Vallières, François Lucia, Laurent Doucet, Valentin Tissot, Gilles Cuvelier, Stephane Hue, Henri Le Penn Du, Luc Perdriel, et al. External validation of an MRI-derived radiomics model to predict biochemical recurrence after surgery for high-risk prostate cancer. *Cancers*, 12(4):814, 2020.

[159] Peng An, Yong Lin, Yan Hu, Ping Qin, YingJian Ye, Weiping Gu, Xiumei Li, Ping Song, and Guoyan Feng. Predicting model of biochemical recurrence of prostate carcinoma (PCa-BCR) using MR perfusion-weighted imaging-based radiomics. *Technology in Cancer Research & Treatment*, 22:15330338231166766, 2023.

[160] Xuehua Zhu, Zenan Liu, Jide He, Ziang Li, Yi Huang, and Jian Lu. MRI-derived radiomics model to predict the biochemical recurrence of prostate cancer following seed brachytherapy. *Archivos Espanoles de Urologia*, 76(4):264–269, 2023.

[161] Qiu-Zi Zhong, Liu-Hua Long, An Liu, Chun-Mei Li, Xia Xiu, Xiu-Yu Hou, Qin-Hong Wu, Hong Gao, Yong-Gang Xu, Ting Zhao, et al. Radiomics of multiparametric MRI to predict

biochemical recurrence of localized prostate cancer after radiation therapy. *Frontiers in oncology*, 10:731, 2020.

[162] Ángel Sánchez Iglesias, Virginia Morillo Macías, Alfonso Picó Peris, Almudena Fuster-Matanzo, Anna Nogué Infante, Rodrigo Muelas Soria, Fuensanta Bellvís Bataller, Marcos Domingo Pomar, Carlos Casillas Meléndez, Raúl Yébana Huertas, et al. Prostate region-wise imaging biomarker profiles for risk stratification and biochemical recurrence prediction. *Cancers*, 15(16):4163, 2023.

[163] Are Losnegård, Lars AR Reisæter, Ole J Halvorsen, Jakub Jurek, Jörg Assmus, Jarle B Arnes, Alfred Honoré, Jan A Monssen, Erling Andersen, Ingfrid S Haldorsen, et al. Magnetic resonance radiomics for prediction of extraprostatic extension in non-favorable intermediate-and high-risk prostate cancer patients. *Acta Radiologica*, 61(11):1570–1579, 2020.

[164] Shuai Ma, Huihui Xie, Huihui Wang, Chao Han, Jiejin Yang, Zhiyong Lin, Yifan Li, Qun He, Rui Wang, Yingpu Cui, et al. MRI-based radiomics signature for the preoperative prediction of extracapsular extension of prostate cancer. *Journal of Magnetic Resonance Imaging*, 50(6):1914–1925, 2019.

[165] Shuai Ma, Huihui Xie, Huihui Wang, Jiejin Yang, Chao Han, Xiaoying Wang, and Xiaodong Zhang. Preoperative prediction of extracapsular extension: radiomics signature based on magnetic resonance imaging to stage prostate cancer. *Molecular Imaging and Biology*, 22:711–721, 2020.

[166] Shayan Mostafaei, Hamid Abdollahi, Shiva Kazempour Dehkordi, Isaac Shiri, Abolfazl Razzaghdoust, Seyed Hamid Zoljalali Moghaddam, Afshin Saadipoor, Fereshteh Koosha, Susan Cheraghi, and Seied Rabi Mahdavi. CT imaging markers to improve radiation toxicity prediction in prostate cancer radiotherapy by stacking regression algorithm. *La radiologia medica*, 125:87–97, 2020.

[167] Hamid Abdollahi, Seied Rabi Mahdavi, Bahram Mofid, Mohsen Bakhshandeh, Abolfazl Razzaghdoust, Afshin Saadipoor, and Kiarash Tanha. Rectal wall MRI radiomics in prostate cancer patients: prediction of and correlation with early rectal toxicity. *International journal of radiation biology*, 94(9):829–837, 2018.

[168] Hamid Abdollahi, Seied Rabi Mahdavi, Isaac Shiri, Bahram Mofid, Mohsen Bakhshandeh, and Kazem Rahmani. Magnetic resonance imaging radiomic feature analysis of radiation-induced femoral head changes in prostate cancer radiotherapy. *Journal of cancer research and therapeutics*, 15(8):S11–S19, 2019.

[169] Hamid Abdollahi, Kiarash Tanha, Bahram Mofid, Abolfazl Razzaghdoust, Afshin Saadipoor, Leila Khalafi, Mohsen Bakhshandeh, and Seied Rabi Mahdavi. MRI radiomic analysis of IMRT-induced bladder wall changes in prostate cancer patients: a relationship

with radiation dose and toxicity. *Journal of Medical Imaging and Radiation Sciences*, 50(2):252–260, 2019.

[170] Joshua W Lorenz, Diane Schott, Lisa Rein, Farshad Mostafaei, George Noid, Colleen Lawton, Meena Bedi, XA Li, Christopher J Schultz, Eric Paulson, et al. Serial t2-weighted magnetic resonance images acquired on a 1.5 tesla magnetic resonance linear accelerator reveal radiomic feature variation in organs at risk: an exploratory analysis of novel metrics of tissue response in prostate cancer. *Cureus*, 11(4), 2019.

[171] Lixin Gong, Min Xu, Mengjie Fang, Jian Zou, Shudong Yang, Xinyi Yu, Dandan Xu, Lijuan Zhou, Hailin Li, Bingxi He, et al. Noninvasive prediction of high-grade prostate cancer via biparametric MRI radiomics. *Journal of Magnetic Resonance Imaging*, 52(4):1102–1109, 2020.

[172] Jose M Castillo T, Muhammad Arif, Martijn PA Starmans, Wiro J Niessen, Chris H Bangma, Ivo G Schoots, and Jifke F Veenland. Classification of clinically significant prostate cancer on multi-parametric MRI: A validation study comparing deep learning and radiomics. *Cancers*, 14(1):12, 2021.

[173] Elena Bertelli, Laura Mercatelli, Chiara Marzi, Eva Pachetti, Michela Baccini, Andrea Barucci, Sara Colantonio, Luca Gherardini, Lorenzo Lattavo, Maria Antonietta Pascali, et al. Machine and deep learning prediction of prostate cancer aggressiveness using multiparametric MRI. *Frontiers in oncology*, 11:802964–802964, 2021.

[174] Geert Litjens Oscar Debats Jelle Barentsz Nico Karssemeijer and Henkjan Huisman. "prostatex challenge data", the cancer imaging archive (2017). In *"ProstateX Challenge data", The Cancer Imaging Archive (2017)*, 2017.

[175] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. Computer-aided detection of prostate cancer in MRI. *IEEE Transactions on Medical Imaging*, 33(5):1083–1092, 2014.

[176] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013.

[177] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging*, 29(6):1310–1320, June 2010.

[178] Ziv Yaniv, Bradley C Lowekamp, Hans J Johnson, and Richard Beare. SimpleITK Image-Analysis notebooks: a collaborative environment for education and reproducible research. *J. Digit. Imaging*, 31(3):290–303, June 2018.

[179] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JWL Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.

[180] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[181] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

[182] Nuno M Rodrigues, Sara Silva, Leonardo Vanneschi, and Nickolas Papanikolaou. A comparative study of automated deep learning segmentation models for prostate MRI. *Cancers*, 15(5):1467, 2023.

[183] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015.

[184] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation, 2018.

[185] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, M. J. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *ArXiv*, abs/1804.03999, 2018.

[186] Nader Aldoj, F. Biavati, Florian Michallek, S. Stober, and M. Dewey. Automatic prostate and prostate zones segmentation of magnetic resonance images using densenet-like u-net. *Scientific Reports*, 10, 2020.

[187] Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation, 2018.

[188] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[189] Charles R. Harris, K. Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020.

[190] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[191] Max Kuhn. Building predictive models in r using the caret package. *Journal of Statistical Software, Articles*, 28(5):1–26, 2008.

[192] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[193] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[194] Sebastian Raschka. MLxtend: Providing machine learning and data science utilities and extensions to python's scientific computing stack. *The Journal of Open Source Software*, 3(24), April 2018.

[195] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[196] Chao Han, Shuai Ma, Xiang Liu, Yi Liu, Changxin Li, Yaofeng Zhang, Xiaodong Zhang, and Xiaoying Wang. Radiomics models based on apparent diffusion coefficient maps for the prediction of high-grade prostate cancer at radical prostatectomy: comparison with preoperative biopsy. *Journal of Magnetic Resonance Imaging*, 54(6):1892–1901, 2021.

[197] Xiang-ke Niu, Zhi-fan Chen, Lin Chen, Jun Li, Tao Peng, and Xin Li. Clinical application of biparametric MRI texture analysis for detection and evaluation of high-grade prostate cancer in zone-specific regions. *American Journal of Roentgenology*, 210(3):549–556, 2018.

[198] Min Xu, Mengjie Fang, Jian Zou, Shudong Yang, Dongdong Yu, Lianzhen Zhong, Chaoen Hu, Yali Zang, Di Dong, Jie Tian, et al. Using biparametric MRI radiomics signature to differentiate between benign and malignant prostate lesions. *European journal of radiology*, 114:38–44, 2019.

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

[200] ProCAncer-I consortium. D5.3 deep learning master model and radiomic signatures, 2024. Accessed on March 28, 2024.

[201] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018,*

Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, pages 3–11. Springer, 2018.

[202] Joost J M van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G H Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J W L Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer Res.*, 77(21):e104–e107, November 2017.

[203] Guido Van Rossum and Fred Drake. *The python language reference manual*. Network Theory, Bristol, England, March 2011.

[204] Raul Garreta and Guillermo Moncecchi. *Learning scikit-learn: Machine Learning in Python*. Packt Publishing Ltd, November 2013.

[205] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.*, 30, 2017.

[206] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. CatBoost: unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.*, 31, 2018.

[207] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.*, 30, 2017.

[208] David Crosby, Sangeeta Bhatia, Kevin M Brindle, Lisa M Coussens, Caroline Dive, Mark Emberton, Sadik Esener, Rebecca C Fitzgerald, Sanjiv S Gambhir, Peter Kuhn, et al. Early detection of cancer. *Science*, 375(6586):eaay9040, 2022.

[209] Barbara Kenner, Suresh T Chari, David Kelsen, David S Klimstra, Stephen J Pandol, Michael Rosenthal, Anil K Rustgi, James A Taylor, Adam Yala, Noura Abul-Husn, et al. Artificial intelligence and early detection of pancreatic cancer: 2020 summative review. *Pancreas*, 50(3):251–279, 2021.

[210] Ioannis Sechopoulos, Jonas Teuwen, and Ritse Mann. Artificial intelligence for breast cancer detection in mammography and digital breast tomosynthesis: State of the art. In *Seminars in cancer biology*, volume 72, pages 214–225. Elsevier, 2021.

[211] OT Jones, RN Matin, M Van der Schaar, K Prathivadi Bhayankaram, CKI Ranmuthu, MS Islam, D Behiyat, R Boscott, N Calanzani, Jon Emery, et al. Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: a systematic review. *The Lancet Digital Health*, 4(6):e466–e476, 2022.

[212] Muhammad Shahid Iqbal, Waqas Ahmad, Roohallah Alizadehsani, Sadiq Hussain, and Rizwan Rehman. Breast cancer dataset, classification and detection using deep learning. In *Healthcare*, volume 10, page 2395. MDPI, 2022.

[213] Ella Mahoro and Moulay A Akhloufi. Applying deep learning for breast cancer detection in radiology. *Current Oncology*, 29(11):8767–8793, 2022.

[214] Maryam Naqvi, Syed Qasim Gilani, Tehreem Syed, Oge Marques, and Hee-Cheol Kim. Skin cancer detection using deep learning—a review. *Diagnostics*, 13(11):1911, 2023.

[215] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, 9:611–629, 2018.

[216] Nuno Miguel Rodrigues, José Guilherme de Almeida, Ana Sofia Castro Verde, Ana Mascarenhas Gaivão, Carlos Bilreiro, Inês Santiago, Joana Ip, Sara Belião, Raquel Moreno, Celso Matos, et al. Analysis of domain shift in whole prostate gland, zonal and lesions segmentation and detection, using multicentric retrospective data. *Computers in Biology and Medicine*, page 108216, 2024.

[217] Joeran S Bosma, Anindo Saha, Matin Hosseinzadeh, Ilse Slootweg, Maarten de Rooij, and Henkjan Huisman. Annotation-efficient cancer detection with report-guided lesion annotation for deep learning-based prostate cancer detection in bpMRI. *arXiv preprint arXiv:2112.05151*, 2021.

[218] Oskar M., Pradeep Reddy Raamana, Mamrehn, and Rémi Bèges. loli/medpy: Medpy 0.3.0 python 3, September 2017.

[219] Simon Kohl, David Bonekamp, Heinz-Peter Schlemmer, Kaneschka Yaqubi, Markus Hohenfellner, Boris Hadaschik, Jan-Philipp Radtke, and Klaus Maier-Hein. Adversarial networks for the detection of aggressive prostate cancer. *arXiv preprint arXiv:1702.08014*, 2017.

[220] Zhenzhen Dai, Eric Carver, Chang Liu, Joon Lee, Aharon Feldman, Weiwei Zong, Milan Pantelic, Mohamed Elshaikh, and Ning Wen. Segmentation of the prostatic gland and the intraprostatic lesions on multiparametic magnetic resonance imaging using mask region-based convolutional neural networks. *Advances in radiation oncology*, 5(3):473–481, 2020.

[221] Ruiming Cao, Amirhossein Mohammadian Bajgiran, Sohrab Afshari Mirak, Sepideh Shakeri, Xinran Zhong, Dieter Enzmann, Steven Raman, and Kyunghyun Sung. Joint prostate cancer detection and gleason score prediction in mp-MRI via focalnet. *IEEE transactions on medical imaging*, 38(11):2496–2506, 2019.

[222] Matin Hosseinzadeh, Anindo Saha, Patrick Brand, Ilse Slootweg, Maarten de Rooij, and Henkjan Huisman. Deep learning–assisted prostate cancer detection on bi-parametric MRI: minimum training data size requirements and effect of prior knowledge. *European Radiology*, pages 1–11, 2022.

[223] Arun Seetharaman, Indrani Bhattacharya, Leo C Chen, Christian A Kunder, Wei Shao, Simon JC Soerensen, Jeffrey B Wang, Nikola C Teslovich, Richard E Fan, Pejman Ghanouni, et al. Automated detection of aggressive and indolent prostate cancer on magnetic resonance imaging. *Medical Physics*, 48(6):2960–2972, 2021.

[224] Brady Neal. *Introduction to Causal Inference From a Machine Learning Perspective*. -, 2020.

[225] Taiki Sato, Yohei Sotomi, Shungo Hikoso, Tetsuhisa Kitamura, Daisaku Nakatani, Katsuki Okada, Tomoharu Dohi, Akihiro Sunaga, Hirota Kida, Yuki Matsuoka, et al. Uplift modeling to identify patients who require extensive catheter ablation procedures among patients with persistent atrial fibrillation. *Scientific Reports*, 14(1):2634, 2024.

[226] Guanggang Li, Bo Li, Bin Song, Dandan Liu, Yue Sun, Hongyan Ju, Xiuping Xu, Jingkun Mao, and Feihu Zhou. Uplift modeling to predict individual treatment effects of renal replacement therapy in sepsis-associated acute kidney injury patients. *Scientific Reports*, 14(1):5833, 2024.

[227] Aditya Biswas, Chirag R Parikh, Harold I Feldman, Amit X Garg, Stephen Latham, Haiqun Lin, Paul M Palevsky, Ugochukwu Ugwuowo, and F Perry Wilson. Identification of patients expected to benefit from electronic alerts for acute kidney injury. *Clinical Journal of the American Society of Nephrology*, 13(6):842–849, 2018.

[228] Seth K Bechis, Peter R Carroll, and Matthew R Cooperberg. Impact of age at diagnosis on prostate cancer treatment and survival. *Journal of Clinical Oncology*, 29(2):235–241, 2011.

[229] Yuanda Zhu, Hang Wu, and May D Wang. Feature exploration and causal inference on mortality of epilepsy patients using insurance claims data. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–4. IEEE, 2019.

[230] Petra J van Houdt, Ghazaleh Ghobadi, Ivo G Schoots, Stijn WTPJ Heijmink, Jeroen de Jong, Henk G van der Poel, Floris J Pos, Susanne Rylander, Lise Bentzen, Karin Haustermans, et al. Histopathological features of MRI-invisible regions of prostate cancer lesions. *Journal of Magnetic Resonance Imaging*, 51(4):1235–1246, 2020.

[231] Walter Artibani, Antonio Benito Porcaro, Vincenzo De Marco, Maria A Cerruto, and Salvatore Siracusano. Management of biochemical recurrence after primary curative treatment for prostate cancer: a review. *Urologia internationalis*, 100(3):251–262, 2018.