Deep Learning Applications in the Prostate Cancer Diagnostic Pathway

Pritesh Mehta

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Medical Physics and Biomedical Engineering

University College London

October 30, 2022

I, Pritesh Mehta, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Prostate cancer (PCa) is the second most frequently diagnosed cancer in men worldwide and the fifth leading cause of cancer death in men, with an estimated 1.4 million new cases in 2020 and 375,000 deaths. The risk factors most strongly associated to PCa are advancing age, family history, race, and mutations of the BRCA genes. Since the aforementioned risk factors are not preventable, early and accurate diagnoses are a key objective of the PCa diagnostic pathway.

In the UK, clinical guidelines recommend multiparametric magnetic resonance imaging (mpMRI) of the prostate for use by radiologists to detect, score, and stage lesions that may correspond to clinically significant PCa (CSPCa), prior to confirmatory biopsy and histopathological grading. Computer-aided diagnosis (CAD) of PCa using artificial intelligence algorithms holds a currently unrealized potential to improve upon the diagnostic accuracy achievable by radiologist assessment of mpMRI, improve the reporting consistency between radiologists, and reduce reporting time.

In this thesis, we build and evaluate deep learning-based CAD systems for the PCa diagnostic pathway, which address gaps identified in the literature. First, we introduce a novel patient-level classification framework, PCF, which uses a stacked ensemble of convolutional neural networks (CNNs) and support vector machines (SVMs) to assign a probability of having CSPCa to patients, using mpMRI and clinical features. Second, we introduce AutoProstate, a deep-learning powered framework for automated PCa assessment and reporting; AutoProstate utilizes biparametric MRI and clinical data to populate an automatic diagnostic report containing segmentations of the whole prostate, prostatic zones, and candidate CSPCa lesions,

as well as several derived characteristics that are clinically valuable. Finally, as automatic segmentation algorithms have not yet reached the desired robustness for clinical use, we introduce interactive click-based segmentation applications for the whole prostate and prostatic lesions, with potential uses in diagnosis, active surveillance progression monitoring, and treatment planning.

Impact statement

The work presented in this thesis has the potential to encourage further research and development within academia and provide benefits outside of academia.

Within the academic environment, an immediate next step will be to conduct multi-center external validation studies to evaluate the computer-aided diagnosis (CAD) systems presented in Chapters 4, 5, and 6 of this thesis. Multi-centre external validation is important for revealing the extent to which CAD systems can generalise beyond the data used to train them. Pleasingly, planning has already begun on a multi-center study to validate the patient classification framework, PCF, presented in Chapter 4, using the PROMIS trial dataset and to validate AutoProstate, presented in Chapter 5, in an international study comparing CAD systems for prostate cancer (PCa) tumour detection.

Another benefit to academia will be to encourage research groups to build upon the topics explored in this thesis that feature scarcely in the current PCa CAD literature, in particular, patient classification, automatic diagnostic report generation, and interactive segmentation. Furthermore, the introduction of technical enhancements to the CAD systems presented in Chapters 4, 5, and 6 can form the basis of academic scholarships; several possible technical enhancements are outlined in Chapter 7. Beyond enhancements and extensions of the work presented in this thesis, effort can be allocated to the creation of a unified platform composed of the CAD systems presented in this thesis, and extensions thereof; active learning should feature in the platform to facilitate continuous improvement over time.

Outside of academia, deployments to the PCa diagnostic pathway can be envisioned: PCF, presented in Chapter 4, can be deployed to triage patients by their

probability of having clinically significant PCa (CSPCa) or to rule out patients with the lowest risk of CSPCa; AutoProstate, presented in Chapter 5, can be deployed as a companion system for radiologists to improve their diagnostic accuracy and reporting quality; and the click-based interactive segmentation applications for whole prostate and prostatic lesion segmentation presented in Chapter 6 have various applications in diagnosis, active surveillance, and treatment. The work presented in the thesis was a collaboration between University College London (UCL) and King's College London (KCL). At KCL in particular, there is a considerable opportunity to deploy tools for clinical use through the London Medical Imaging & AI Centre for Value Based Healthcare; a key aim of the centre is to provide the infrastructure required for building, deploying, and evaluating AI algorithms in healthcare. Prior to deployment into the PCa diagnostic pathway, prospective validation studies should be conducted by a multi-disciplinary team of algorithm developers and clinicians. Prospective validation studies are critical for understanding how CAD would fit into the clinical workflow and factors outside of the research setting which may adversely impact performance. During prospective validation, substantive efforts must be placed on understanding which cases fail and why, by both algorithm developers and clinical experts, to guide further algorithmic development and training data collection.

Publications list

Peer-reviewed journal papers

1. (PUBLISHED) Mehta, P., Antonelli, M., Ahmed, H. U., Emberton, M., Punwani, S., Ourselin, S. Computer-aided diagnosis of prostate cancer using multiparametric MRI and clinical features: A patient-level classification framework. Medical Image Analysis, 2021.

2. (PUBLISHED) Syer, T.*, Mehta, P.*, Antonelli, M., Mallett, S., Atkinson, D., Ourselin, S., Punwani, S. Artificial Intelligence Compared to Radiologists for the Initial Diagnosis of Prostate Cancer on Magnetic Resonance Imaging: A Systematic Review and Recommendations for Future Studies. Cancers, 2021. (*indicates joint first author)

3. (PUBLISHED) Mehta, P., Antonelli, M., Singh, S., Grondecka, N., Johnston, E. W., Ahmed, H. U., Emberton, M., Punwani, S., Ourselin, S. AutoProstate: Towards Automated Reporting of Prostate MRI for Prostate Cancer Assessment using Deep Learning. Cancers, 2021.

4. (UNDER REVIEW, PREPRINT PUBLISHED) Diaz-Pinto, A., Alle, S., Ihsani, A., Nath, V., Asad, M., Pérez-García, F., Mehta, P., Li, W., Roth, H. R., Vercauteren, T., Ourselin, S., Feng, A., Cardoso, M. J. MONAI Label: A framework for AI-assisted Interactive Labeling of 3D Medical Images. Medical Image Analysis, 2022. Preprint published arXiv:2203.12362.

Peer-reviewed conference papers

1. (PUBLISHED) Mehta, P., Antonelli, M., Ahmed, H. U., Emberton, M., Punwani, S., Ourselin, S. Decision fusion of 3D convolutional neural networks to triage patients with suspected prostate cancer using volumetric biparametric MRI. SPIE Medical Imaging, 2020, Houston, Texas, United States.

2. (PUBLISHED) Diaz-Pinto, A., Mehta, P., Alle, S., Asad, M., Brown, R., Nath, V., Ihsani, A., Antonelli, M., Palkovics, D., Pinter, C., Alkalay, R., Pieper, S., Roth, H. R., Xu, D., Dogra, P., Vercauteren, T., Feng, A., Quiraini, A., Ourselin, S., Cardoso, M. J. DeepEdit: Deep Editable Learning for Interactive Segmentation of 3D Medical Images. MICCAI Workshop on Data Augmentation, Labelling, and Imperfections, 2022, Singapore.

Peer-reviewed conference abstracts

1. (ACCEPTED) Mehta, P., Antonelli, M., Punwani, S., Ourselin, S. A 3D convolutional neural network for diagnosing prostate cancer using volumetric T2-weighted MRI. International Society for Magnetic Resonance in Medicine (ISMRM), 2019, Montreal, Canada.

2. (ACCEPTED) Williams, H., Cattani, L., Li, W., Tabassian, M., Mehta, P., Vercauteren, T., Deprest, J., Dhooge, J. 3D convolutional neural network for segmentation of the urethra in volumetric ultrasound of the pelvic floor. IEEE International Ultrasonics Symposium (IUS), 2019, Glasgow, UK.

3. (ACCEPTED) Gu, R., Antonelli, M., Mehta, P., Amlani, A., Green, A., Neji, R., Ourselin, S., Dregely, I., Goh, V. Using uncertainty estimation to increase the robustness of bone marrow segmentation in T1-weighted Dixon MRI for multiple myeloma. International Society for Magnetic Resonance in Medicine (ISMRM), 2021, online due to COVID-19.

4. (ACCEPTED) Gu, R., Antonelli, M., Mehta, P., Amlani, A., Gervaise-Andre, L., Green, A., Neji, R., Ourselin, S., Dregely, I., Goh, V. Uncertainty-aware CNN improves prediction robustness for bone marrow segmentation with

noisy labels and T1-weighted Dixon MR images. King's College London School of Biomedical Engineering & Imaging Sciences Postgraduate Research Symposium, 2021, online due to COVID-19.

5. (ACCEPTED) Withey, S. J., Azam, A., Sharkey, A., Mehta, P., Rookyard, C., O'Shea, R., Cook, G. J., Goh, V. Radiogenomics in Cancers of the Genitourinary System. RSNA educational exhibit on Radiogenomics in Cancers of the Genitourinary Tract, 2021, Chicago, USA.

6. (ACCEPTED) Gu, R., Antonelli, M., Mehta, P., Amlani, A., Green, A., Neji, R., Ourselin, S., Dregely, I., Goh, V. Automatic segmentation of whole-body MRI using UnnU-Net: Feasibility of whole-skeleton ADC evaluation in plasma cell disorders. International Society for Magnetic Resonance in Medicine (ISMRM), 2022, London, UK.

Acknowledgements

I would like to thank my primary supervisor, Professor Sébastien Ourselin, for his guidance and words of motivation throughout my PhD, for providing the resources that allowed me to succeed, and for the award of my scholarship all those years ago. I would also like to thank my secondary/clinical supervisor, Professor Shonit Punwani, for helping me understand the clinical context of my work and for encouraging collaborations with students in his research group.

I am extremely grateful to my tertiary supervisor, Dr Michela Antonelli. Her knowledge, experience, and dedication to providing quality supervision through ample support and availability, proved invaluable in ensuring several high quality outputs of the PhD. I feel privileged to have been the first PhD student she supervised.

Last, but certainly not least, I would like to thank my parents, sister, and grandmother for their love, support, and encouragement. They do everything in their power to help me succeed and I am forever grateful.

Contents

Abstract **3**

Impact statement **5**

Publications list **7**

Acknowledgements **10**

Contents **11**

List of figures **16**

List of tables **20**

1 Introduction **24**

1.1 Prostate cancer 25

1.1.1 Prostate anatomy and function 25

1.1.2 What is prostate cancer? 26

1.1.3 Causes and risk factors 27

1.1.4 Incidence, mortality, and survival 27

1.2 Prostate cancer diagnostic pathway 28

1.2.1 Prostate-specific antigen blood test and/or digital rectal exam 29

1.2.2 Multiparametric magnetic resonance imaging 29

1.2.3 Prostate biopsy and Gleason scoring 35

1.2.4 Multidisciplinary team meeting 37

1.2.5 Radical treatment, active surveillance, or watchful waiting . 37

1.3 Motivation of this work and thesis overview 38
1.3.1 Rationale 39
1.3.2 Thesis overview 41

2 Literature review **43**
2.1 Machine learning for medical image analysis 43
2.1.1 Classification 45
2.1.2 Segmentation 45
2.2 Machine learning for prostate MRI analysis 48
2.2.1 Whole prostate segmentation 48
2.2.2 Prostatic zones segmentation 49
2.2.3 Computer-aided diagnosis of prostate cancer 50
2.3 Gaps in the literature addressed by this thesis 57

3 Patient datasets for training and evaluation **60**
3.1 PROSTATEx dataset 60
3.1.1 Multiparametric MRI protocol 61
3.1.2 Multiparametric MRI review 61
3.1.3 Histopathological reference standard 61
3.1.4 Contours 62
3.2 PICTURE dataset 62
3.2.1 Multiparametric MRI protocol 63
3.2.2 Multiparametric MRI review 64
3.2.3 Histopathological reference standard 65
3.2.4 Contours 65
3.3 Bias in prostate datasets 66

4 Patient-level classification framework for triage **69**
4.1 Introduction 69
4.2 Method 71
4.2.1 Pre-processing 71
4.2.2 Convolutional neural network feature extraction 76

4.2.3 Forward feature selection . . . 77
4.2.4 Support vector machine classification . . . 78
4.3 Experimental setup . . . 78
4.3.1 Patient data . . . 78
4.3.2 Experiments . . . 80
4.3.3 Experimental settings . . . 81
4.4 Results . . . 83
4.4.1 Intra-dataset model evaluation . . . 83
4.4.2 Inter-dataset model evaluation . . . 87
4.4.3 Clinical evaluation . . . 89
4.5 Discussion . . . 91

5 Towards automated reporting: AutoProstate 95
5.1 Introduction . . . 95
5.2 Method . . . 97
5.2.1 Zone-Segmenter module . . . 99
5.2.2 CSPCa-Segmenter module . . . 100
5.2.3 Report-Generator module . . . 102
5.3 Experimental setup . . . 104
5.3.1 Patient datasets . . . 105
5.3.2 Methodological settings . . . 105
5.3.3 External validation evaluation measures . . . 107
5.4 Results . . . 108
5.4.1 Zone-U-Net and CSPCa-U-Net ten-fold cross-validation . . . 108
5.4.2 AutoProstate external validation analysis: whole prostate and zonal segmentation, prostate size measurements, and PSA density . . . 109
5.4.3 AutoProstate external validation analysis: clinically significant prostate cancer lesion detection and segmentation . . . 112
5.5 Discussion . . . 121

6 Click-based interactive segmentation of the whole prostate and prostatic lesions: a preliminary study 126

6.1 Introduction . . . 126

6.2 Methods . . . 128

6.2.1 DeepGrow . . . 128

6.2.2 DeepEdit . . . 134

6.3 Experimental setup . . . 134

6.3.1 Experiments . . . 135

6.3.2 Experimental pipeline and settings . . . 135

6.4 Results . . . 138

6.4.1 Whole prostate segmentation task . . . 138

6.4.2 Prostatic lesion segmentation task . . . 141

6.5 Discussion . . . 147

7 Summary and future work 154

A Background theory: magnetic resonance imaging 162

A.1 Fundamental nuclear magnetic resonance theory . . . 163

A.1.1 Spin, magnetic moment, and Larmor precession . . . 163

A.1.2 Magnetisation . . . 164

A.1.3 Resonance . . . 166

A.2 The Bloch equation and relaxation . . . 168

A.2.1 Spin-lattice relaxation . . . 168

A.2.2 Spin-spin interaction and transverse decay . . . 169

A.2.3 Magnetic field inhomogeneity . . . 170

A.2.4 The Bloch Equation . . . 171

A.3 Signal acquisition . . . 171

A.3.1 Free induction decay . . . 171

A.3.2 Spin-echo sequence . . . 172

A.3.3 Demodulation . . . 173

A.4 Imaging (spatial localisation) . . . 173

A.4.1 K-space and Fourier imaging . . . 174
A.4.2 Gradient-echo with spatial encoding . . . 174
A.4.3 Spin-echo with spatial encoding . . . 177
A.5 Image contrast . . . 177
A.5.1 Conventional structural contrasts . . . 178
A.5.2 Diffusion-weighted MRI . . . 178
A.5.3 Dynamic contrast-enhanced MRI . . . 180

B Background material: machine learning and deep learning basics 181
B.1 Machine learning theory . . . 181
B.1.1 Supervised and unsupervised learning . . . 181
B.1.2 Evaluating a learning algorithm . . . 182
B.1.3 K-fold cross-validation . . . 182
B.1.4 Bias and variance . . . 183
B.2 Deep learning theory . . . 183
B.2.1 Neural networks representation . . . 184
B.2.2 Training a neural network . . . 186
B.2.3 Optimisation . . . 190
B.2.4 Batch normalisation . . . 193
B.2.5 Hyperparameter selection in neural networks . . . 194
B.2.6 Regularisation . . . 195
B.2.7 Convolutional neural networks . . . 196
B.2.8 Convolutional neural network architectures . . . 200

Bibliography 205

List of figures

1.1 Regional and zonal division of the prostate. 26

1.2 Schematic representation of the PCa diagnostic pathway, with section numbers shown in brackets. 28

1.3 Transverse T2WI slice through the prostate showing a Gleason 3+4 tumour in the PZ. 31

1.4 Transverse ADC map slice through the prostate showing Gleason 3+4 tumour in the PZ. 32

1.5 Transverse ADC map slice through the prostate showing a severe magnetic susceptibility artifact. 33

2.1 Taxonomy of CAD systems for PCa diagnosis. CAD systems are categorized as Patient Classification (PAT-C), ROI Classification (ROI-C), or Lesion Localization and Classification (LL&C). Blue indicates mpMRI/bpMRI input, yellow indicates manual processes, white indicates automated processes, and green indicates intermediate or final outputs. ROI = region of interest. CNN = convolutional neural network. ML = machine learning. ML* here refers to ML algorithms exclusive of CNNs, such as support vector machines, random forest, logistic regression, and artificial neural networks. . . 52

4.1 Workflow of the proposed patient-level classification framework, PCF. Green rectangles indicate original scans or clinical features or the probabilistic classification outcome; yellow rectangles indicate processes applied to the data. In experiments, we refer to PCF with forward feature selection disabled during training as PCF-ALL and PCF with forward feature selection enabled during training as PCF-SEL. 72

4.2 Normalised signal intensity-versus-time curve corresponding to voxel at the center of a Gleason score 3+4 lesion of a patient from the PROSTATEx dataset. 75

4.3 (a) Proposed ResNet3D CNN used to extract features from volumetric images. (b) A bottleneck block, where k = #kernels. 77

5.1 AutoProstate framework diagram. AutoProstate consists of three modules: Zone-Segmenter (green), CSPCa-Segmenter (blue), and Report-Generator (purple); solid boxes correspond to module computations, while dashed boxes correspond to module outputs. Yellow boxes indicate AutoProstate inputs from external sources. . . . 98

5.2 AutoProstate Report template, where xx denotes an automatically populated field. 103

5.8 AutoProstate report for a 64-year-old man with PSA equal to 10.53 ng/ml who participated in the PICTURE study. LESION 1 (probability of CSPCa equal to 95%) corresponds to a biopsy-proven Gleason score 3+4 lesion, while LESION 2 and LESION 3 (probabilities of CSPCa equal to 46% and 7%, respectively) are false-positives. 120

6.1 Schematic representation of interactive segmentation using DeepGrow. The user, who possesses domain knowledge, observes the current segmentation which may contain false-positive and/or false-negative regions. Subsequently, foreground clicks and background clicks are applied as necessary to produce an updated segmentation. 131

6.2 3D Slicer interface for the whole prostate segmentation MONAI Label DeepEdit application. 139

6.3 3D Slicer interface for the prostatic lesion segmentation MONAI Label DeepEdit application. 140

6.4 Whole prostate segmentation Dice score box plots for the 193 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow, DeepEdit-0.25, and DeepEdit-0.5. Dice loss and Click-Loc-Method-I were fixed for comparison. 145

6.5 Prostatic lesion segmentation Dice score distributions for the 200 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow, DeepEdit-0.25, and DeepEdit-0.5. Dice loss and Click-Loc-Method-I were fixed for comparison. 150

A.1 T1-weighted, T2-weighted, and Flair MRI contrasts for a transverse slice through the brain [1]. 162

A.2 Clockwise precession of a magnetic moment about an external magnetic field $\vec{B}_0$ [2][Fig 1.1]. 164

A.3 Net magnetisation vector $\vec{M}$. There is a spin excess in the parallel alignment, giving rise to $\vec{M}$. $\vec{M}$ has no transverse component as the spins lack phase coherence. 166

A.4 Block diagram of the key hardware components found in a MRI scanner. 167

A.5 90° flip of the net magnetisation vector into the transverse plane [2][Fig. 4.2] 168

A.6 Longitudinal magnetisation recovery to equilibrium value M_0 [2][Fig. 4.1a]. 169

A.7 Transverse magnetisation decay from initial value following 90° flip [2][Fig. 4.1b]. 170
A.8 Sequence diagram for a repeated FID experiment [2][Fig. 8.2]. . . . 172
A.9 Sequence diagram and signal representation for a spin-echo sequence [2][Fig. 8.3]. 173
A.10 2D Fourier transform. 174
A.11 2D gradient-echo sequence [2][Fig. 10.14]. 175
A.12 Traversal of k-space for a typical gradient-echo experiment [2][Fig. 10.15b]. 176
A.13 2D spin-echo sequence [2][Fig. 10.17]. 177
A.14 Spin-echo sequence with diffusion sensitising gradients G_{diff}. 179

B.1 Diagram of a simple MLP with two hidden layers [3]. 184
B.2 Operations performed by a neuron with two inputs x_1 and x_2, two weight parameters w_1 and w_2, a bias parameter b, and an activation function $\sigma(\cdot)$. 185
B.3 An illustration of the trajectory of gradient descent towards the minimum of an error function. red: batch, green: mini-batch, purple: online. 191
B.4 An illustration of a 2D convolution operation using a single 2×2 kernel [4]. 198
B.5 An illustration of a 2D max pooling operation using a 2×2 neighbourhood and stride 2. 200
B.6 Residual learning building block [5]. 201
B.7 U-Net architecture [6]. 202
B.8 3D U-Net architecture [7]. 203
B.9 nnU-Net framework [8]. 204

List of tables

1.1 Risk stratification for patients with localised prostate cancer as described by UK NICE guidelines [9]. 37

3.1 PROSTATEx Challenges dataset characteristics. 63
3.2 PROSTATEx training dataset annotations. 64
3.3 PICTURE dataset characteristics. 68

4.1 Characteristics of the PICTURE dataset patients used to evaluate PCF. Interquartile range shown in brackets for age, PSA, total prostate volume (TPV), and PSA density (PSAd). 79
4.2 Characteristics of the PROSTATEx dataset patients used to evaluate PCF. 80
4.3 Intra-dataset evaluation of ResNet3D, SVM, and PCF classifiers. Mean ROC AUC and PR AUC $\pm$ one standard deviation averaged over five-fold cross validation, for the PICTURE and PROSTATEx datasets, shown. Highest value in each column shown in bold. . . . 84
4.4 Inter-dataset evaluation. Mean ROC AUC and PR AUC $\pm$ one standard deviation for ResNet3D and PCF classifiers obtained from the intra-dataset evaluation and subsequently applied to the dataset that was not used to train the classifier, shown. Highest value in each column shown in bold. 88
4.5 Clinical comparison of the patient-level diagnostic performance of radiologist Likert scoring and PCF-SEL, on temporally separated training and test cohorts from the PICTURE dataset. 90

5.1 AutoProstate external validation analysis of whole prostate and zonal segmentations, prostate size measurements, and PSAd, using 80 patients from the PICTURE dataset for which ground-truth segmentations were available. 110

5.2 PICTURE dataset CSPCa lesion detection metrics for the radiologist and AutoProstate. Mean and standard deviation of false-positives per-patient were calculated using the 93 PICTURE dataset patients who were biopsy-negative for CSPCa, rather than over all patients, due to limitations in the ground-truth. All other metrics shown are calculated at the lesion-level for the 147 CSPCa lesions and 63 nCSPCa lesions. 113

5.3 Agreement and disagreement between radiologist Likert scoring and AutoProstate, on PICTURE dataset annotated lesions; grey shading indicates concordance, gold shading indicates superior performance by the radiologist, and blue shading indicates superior performance by AutoProstate. 116

6.1 Kernel size and stride settings configured by nnU-Net for 3D U-Net. 136

6.2 Whole prostate segmentation mean Dice scores $\pm$ one standard deviation, calculated over the 193 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow trained using Dice loss and a hybrid loss composed of Dice loss and Focal loss. Click-Loc-Method-I was fixed for comparison. The highest mean Dice in each column is shown in bold. In addition, P-values, calculated using the Wilcoxon signed-rank test that are less than 0.05 are indicated with an asterisk. 142

6.3 Whole prostate segmentation mean Dice scores ± one standard deviation, calculated over the 193 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow trained using Click-Loc-Method-I and Click-Loc-Method-II. Dice loss was fixed for comparison. The highest mean Dice in each column is shown in bold. In addition, P-values, calculated using the Wilcoxon signed-rank test that are less than 0.05 are indicated with an asterisk. 143

6.4 Whole prostate segmentation mean Dice scores ± one standard deviation for the 193 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow, DeepEdit-0.25, and DeepEdit-0.5. Dice loss and Click-Loc-Method-I were fixed for comparison. The highest mean Dice in each column is shown in bold. The Friedman test for multiple comparisons was used to assess differences between groups, followed by Dunn's pairwise test with Bonferroni correction. P-values less than 0.05 are indicated with an asterisk. . . 144

6.5 Prostatic lesion segmentation mean Dice scores ± one standard deviation, calculated over the 200 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow trained using Dice loss and a hybrid loss composed of Dice loss and Focal loss. Click-Loc-Method-I was fixed for comparison. The highest mean Dice in each column is shown in bold. In addition, P-values, calculated using the Wilcoxon signed-rank test, less than 0.05 are indicated with an asterisk. 146

6.6 Prostatic lesion segmentation mean Dice scores ± one standard deviation, calculated over the 200 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow trained using Click-Loc-Method-I and Click-Loc-Method-II. Dice loss was fixed for comparison. The highest mean Dice in each column is shown in bold. In addition, p-values, calculated using the Wilcoxon signed-rank test, less than 0.05 are indicated with an asterisk. 148

6.7 Prostatic lesion segmentation mean Dice scores ± one standard deviation for the 200 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow, DeepEdit-0.25, and DeepEdit-0.5. Dice loss and Click-Loc-Method-I were fixed for comparison. The highest mean Dice in each column is shown in bold. The Friedman test for multiple comparisons was used to assess differences between groups, followed by Dunn's pairwise test with Bonferroni correction. P-values less than 0.05 are indicated with an asterisk. . . 149

A.1 A brief description of conventional structural MRI contrasts. 178

B.1 Common activation functions used in neural networks. 188

Chapter 1

Introduction

Cancer is among the leading causes of death in every country of the world; in 2020 there were an estimated 19.3 million new cancer cases and 10 million cancer deaths [10]. Unfortunately, the global cancer burden is projected to rise quite substantially over the next two decades, reaching an estimated 28.4 million cases by 2040 [10]. In order to cope with the increased demand on healthcare services posed by the rising case incidence, healthcare systems across the world are working to build more sustainable cancer management infrastructures based foremost on prevention and early diagnosis [10].

Artificial intelligence (AI) has been earmarked as a disruptive technology that will cause a major transformation in the way healthcare is delivered [11]. In the UK, the National Health Service's (NHS) digital transformation unit, NHSX, has setup the NHS Artificial Intelligence Laboratory which is bringing together the government, healthcare providers, academics, and technology companies to build an infrastructure for the safe and ethical deployment of AI-driven technologies at scale [12]. A priority area for NHSX is diagnostic support in radiology due to pressures caused by an increased demand for radiology services and a shortage of radiologists to meet the demand [13]. According to the Royal College of Radiologists (RCR), there was a 33% shortfall in clinical radiology consultants in the UK in 2020, which is projected to grow to 44% by 2025 [14]. Through providing triage as a service, supporting diagnoses, or additional insights, AI deployed into radiology workflows may improve radiologist productivity and decision-making, leading to quicker di-

agnoses and improved patient outcomes [11].

The prostate cancer (PCa) diagnostic pathway has been identified by research groups globally as a cancer pathway that would benefit from the deployment of AI-driven technologies, due to pressures caused by rising case incidence, the increased use of multiparametric magnetic resonance imaging (mpMRI) for diagnosis, and a shortage of specialist radiologists to review mpMRI [15]. Motivated by the aforementioned pressures, we present novel works on the development and evaluation of deep learning-based computer-aided diagnosis (CAD) systems for the PCa diagnostic pathway, in this thesis. The remainder of this chapter is structured as follows: first, we introduce PCa; next, we describe the PCa diagnostic pathway; we conclude by describing the specific motivations of the work presented in this thesis and by providing a thesis overview.

1.1 Prostate cancer

1.1.1 Prostate anatomy and function

The prostate is an accessory gland belonging to the male reproductive system [16]. The prostate is located inferiorly to the neck of the bladder and superiorly to the external urethral sphincter, and lies adjacent to the rectum. The primary function of the prostate is to secrete proteolytic enzymes into the semen, which act to break down clotting factors in the ejaculate. In addition, the muscles of the prostate contract during ejaculation to push seminal fluid into the urethra.

The structure of the prostate is described using a regional division and a zonal division, as shown in Figure 1.1, which in combination with anatomical directions, can be used to accurately describe the location of prostate pathologies. The regional division describes the base, midgland, and apex. The base is the upper portion of the prostate closest to the bladder, the apex is the lower portion of the prostate closest to the external urethral sphincter, and the midgland is the remaining middle portion. Alternatively, the zonal structure of the prostate describes the peripheral zone (PZ), anterior fibromuscular stroma (AFMS), central zone (CZ), transition zone (TZ), and the thin layer of surrounding connective tissue and muscle fibres called the capsule.

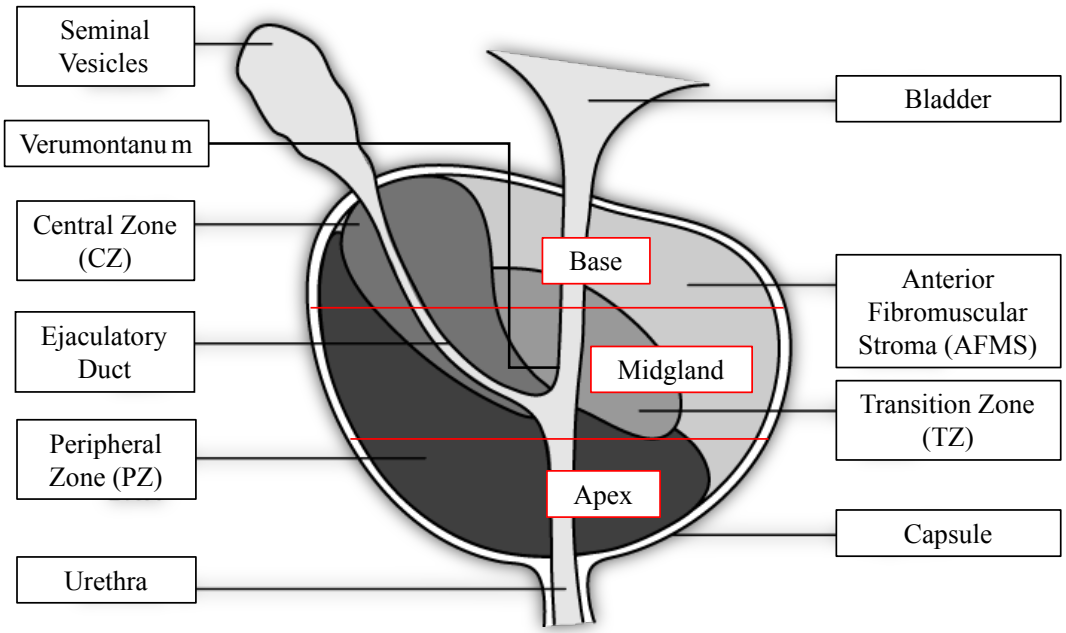


Figure 1.1: Regional and zonal division of the prostate.

The PZ is located at the posterior of the prostate, the AFMS is located at the anterior of the prostate, the TZ surrounds the prostatic urethra, and the CZ sits posterior to the TZ and surrounds the ejaculatory ducts. The AFMS, CZ, and TZ are often grouped into the central gland (CG).

1.1.2 What is prostate cancer?

Cancers occur due to genetic mutations which cause uncontrollable and uninhibited cell growth, forming tumours. A designation of PCa is made if the cancer originates in the prostate. Most cancers that develop in the prostate are adenocarcinomas [17]. These are cancers that develop in glandular tissue, found predominantly in the PZ and TZ of the prostate. Early stage PCa may be asymptomatic [18]. Should tumours grow large enough to put pressure on the urethra or bladder, symptoms may include trouble urinating due to constriction of the urethra or the need to urinate more often due to pressure on the bladder, as well as blood in the urine and/or blood in the semen [18].

Cancer confined within the prostate is referred to as localised PCa. However, PCa can spread to other parts of the body in a process called metastasis. Typically,

PCa spreads to the bones, lymph nodes, liver and/or lungs [17]. Following metastasis, the designation of localised PCa will be replaced by advanced PCa; at this stage, the cancer can be significantly more difficult to control and may lead to a shortening of life [17].

1.1.3 Causes and risk factors

The foremost risk factor for PCa is advancing age [19]. The risk of developing PCa rises dramatically with age due to deoxyribonucleic acid (DNA) damage that accumulates over time. According to Cancer Research UK, incidence rates for PCa are highest in males aged 75-79 [19].

Ethnicity is also a known risk factor for PCa. Black men in the United States and the Caribbean have the highest incidence rates globally [10], and in the UK, PCa risk is higher in Black men compared to White men and Asian men [19].

Inherited factors account for 5-9% of PCa cases [19]. PCa risk is 2.1-2.4 times higher in men whose father has/had the disease, 2.9-3.3 times higher in men whose brother has/had the disease, and 1.9 times higher in men with a second-degree relative (grandfather, uncle, nephew, or half-sibling) who has/had the disease [19]. In addition, studies have shown that PCa risk is 19-24% higher in men whose mother has/had breast cancer, though PCa risk is not associated with breast cancer in a sister [19].

A smaller number of PCa cases can be explained by mutations to the BRCA1 and BRCA2 tumour suppressor genes, inherited Lynch syndrome, and higher than normal levels of insulin-like growth-factor-1 (IGF-1) [19].

1.1.4 Incidence, mortality, and survival

PCa is the second most frequently diagnosed cancer in men worldwide and the fifth leading cause of cancer death; there were an estimated 1.4 million new diagnoses of PCa in 2020 and 375,000 deaths [10]. In 112 of 185 countries, PCa is the most frequently diagnosed cancer in men, including in the Americas, Northern/Western Europe, Australia/New Zealand, and much of Sub-Saharan Africa [10]. In 48 of 185 countries, PCa is the leading cause of cancer death in men, particularly in the

Caribbean, sub-Saharan Africa, and Micronesia/Polynesia [10]. Survival rates vary globally, likely reflecting differences in diagnostic and treatment practices. The CONCORD-2 study [20], which aims to inform global policy on cancer control, compiled survival statistics for PCa from 61 countries; five-year survival rates varied between less than 40% and greater than 95%. In the UK, Cancer Research UK has reported a 86.6% five-year survival rate overall [19]. Furthermore, they report that men diagnosed with early-stage PCa have a five-year survival rate of 100%, compared to a five-year survival rate of 49% for men diagnosed with late-stage PCa [19].

1.2 Prostate cancer diagnostic pathway

In the UK, National Institute of Health and Care Excellence (NICE) guidelines recommend a pathway composed of the following core tests: first, a prostate-specific antigen (PSA) blood test and/or a digital rectal exam (DRE); if suspicion persists, mpMRI; and if suspicion remains, MR-guided targeted biopsy and/or systematic biopsy [9] with histopathological analysis of biopsy samples. Test findings will be discussed at a multidisciplinary team (MDT) meeting to determine whether patients can be discharged, require radical treatment, or require an alternative disease management approach such as active surveillance (AS) or watchful waiting. A schematic representation of the pathway is shown in Figure 1.2.

Figure 1.2: Schematic representation of the PCa diagnostic pathway, with section numbers shown in brackets.

1.2.1 Prostate-specific antigen blood test and/or digital rectal exam

A PSA blood test is typically carried out in men presenting with symptoms concordant with PCa. The PSA blood test is a well-established test for PCa, first introduced in the late 1980s and early 1990s in the United States, Canada, and Australia [10]. PSA is a protein produced by both cancerous and non-cancerous prostate cells, however, elevated PSA is a marker for PCa [9]. According to NICE guidelines, a PSA less than 10 ng/ml may be indicative of low risk disease, a PSA of 10-20 ng/ml may be indicative of intermediate risk disease, while a PSA greater than 20 ng/ml may be indicative of high risk disease. However, several benign conditions can also cause a rise in PSA, such as prostatitis and benign prostatic hyperplasia (BPH) [21]. According to the NHS, about 3 in 4 men with a raised PSA level will not have PCa and PSA can miss about 15% of PCa cases [21].

An alternative test which may be used to detect PCa is a digital rectal exam (DRE). A DRE involves a doctor inserting a lubricated finger into the rectum to feel for hard, lumpy, or abnormal areas in the posterior prostate. However, since DRE cannot achieve complete gland coverage, findings from DRE may be inconclusive. DRE has been shown to be a less effective test for detecting PCa; a meta-analysis found a sensitivity of 51% and a specificity of 59% for PCa detection using DRE [22].

1.2.2 Multiparametric magnetic resonance imaging

MRI is a highly flexible medical imaging technique that can be used to produce detailed anatomical or functional images of parts of the body. The inclusion of MRI in the PCa diagnostic pathway for PCa localisation, grading, and staging is becoming increasingly widespread [23]. In the UK, NICE guidelines recommend mpMRI for men with suspected clinically significant localised PCa [9]. MpMRI of the prostate is a combination of T2-weighted imaging (T2WI), diffusion-weighted imaging (DWI), and dynamic contrast-enhanced imaging (DCEI) [23]. Radiologists interpret the information presented in mpMRI to determine the likelihood of

clinically significant PCa (CSPCa), where CSPCa refers to cancers which carry a heightened mortality risk. While there is no general agreement on the clinical definition of CSPCa, the two most widely used mpMRI reporting guidelines [24, 25] suggest CSPCa should be defined as histopathological Gleason score ≥ 7 (including $3+4$ with prominent but not predominant Gleason 4 component), and/or volume $>$ 0.5 cc, and/or extraprostatic extension (EPE).

The inclusion of mpMRI in the PCa diagnostic pathway has increased CSPCa detection sensitivity. On a cohort of 576 men, the "Prostate MRI Imaging Study" (PROMIS) [26] reported a sensitivity of 93% and a specificity of 41% for the detection of CSPCa by an experienced radiologist reading mpMRI, compared to a sensitivity of 48% and a specificity of 96% for transrectal ultrasound-guided (TRUS) biopsy. However, the low specificity of CSPCa detection on mpMRI currently leads to a large number of unnecessary biopsies [26].

1.2.2.1 T2-weighted imaging

In T2WI, contrast is achieved through tissue T2 relaxation time differentials; T2 is a specific physical constant unique to tissues that can be exploited by certain MRI pulse sequences to produce an image. The contrast produced allows the typically higher signal intensity PZ to be differentiated from the typically lower signal intensity TZ and CZ. In addition, there is a clear differentiation between the prostate and background tissues on T2WI. In the PZ, PCa tumours can appear as areas of low signal intensity against the higher signal intensity background PZ [27]. An example T2WI for a patient with a tumour in the PZ is shown in Figure 1.3.

However, there are limitations in using T2WI alone for PCa diagnosis. In particular, areas of low signal intensity in the PZ do not always represent cancer. Benign abnormalities such as prostatitis, atrophy, scars, post-irradiation or hormonal treatment effects, hyperplasia, and post-biopsy hemorrhage can mimic PCa [27]. In addition, PCa in the TZ and CZ can be difficult to discern from benign prostatic hyperplasia (BPH); BPH occurs naturally with age and may have a similar low signal intensity to PCa [27].

Figure 1.3: Transverse T2WI slice through the prostate showing a Gleason 3+4 tumour in the PZ.

1.2.2.2 Diffusion-weighted imaging

In DWI, contrast is based on the local motion of water molecules. Water diffusion in tissue can be intracellular, within the interstitial space, and between cellular and interstitial spaces. Diffusion sensitising gradients can be applied to produce images that reflect the water diffusion in a voxel. The strength and timing of the diffusion sensitising gradients is determined using the "b-value"; increasing the b-value gives greater sensitivity to water motion, but decreases signal-to-noise ratio (SNR) [27].

By varying the amount of diffusion weighting, a quantitative measure that reflects tissue microstructure can be obtained: the apparent diffusion coefficient (ADC). By solving for the ADC at every voxel, an ADC map is produced. ADC values in healthy prostate tissue can be relatively high due to the presence of gland tubules in the glandular tissue of the prostate [27]. However, PCa can destroy the normal glandular structure of the prostate, which will reduce the motion of water [27]. In addition, tumours have a higher cellular density than healthy tissue, which again restricts water motion. As a result, lower than typical ADC values may be indicative of PCa [27]. Radiologist reporting guidelines recommend a threshold of 750-900 mm^2/sec below which ADC values may correspond to CSPCa [24]. An

example ADC map for a patient with a tumour in the PZ is shown in Figure 1.4.

Figure 1.4: Transverse ADC map slice through the prostate showing Gleason 3+4 tumour in the PZ.

In addition to ADC maps, "high b-value" (b $\geq$ 1400 [24]) DWI is typically collected. Tumours on high b-value DWI appear as hyperintensities against the low signal intensity background due to preservation of signal in areas of restricted diffusion. Recent evidence has suggested that high b-value DWI is advantageous for highlighting index tumours, tumours adjacent to or invading the AFMS, and tumours at the base and apex of the prostate [24, 28].

While high b-value DWI and ADC maps are a valuable component of mpMRI, there are limitations in their use. Due to the use of fast imaging echo-planar sequences for acquisition, DWI has a lower spatial resolution than T2WI. In addition, fast imaging echo-planar sequences are very sensitive to magnetic field inhomogeneities, which can cause magnetic susceptibility artifacts at tissue-air interfaces [27]. In particular, air in the rectum can cause a dome-like distortion of the PZ, which can obscure PZ tumours. An ADC map impacted by a severe magnetic susceptibility artifact is shown in Figure 1.5.

An issue specific to high b-value DWI is diminishing signal-to-noise ratio, which can cause distortion and ghosting artifacts [28]. A method of circumventing the

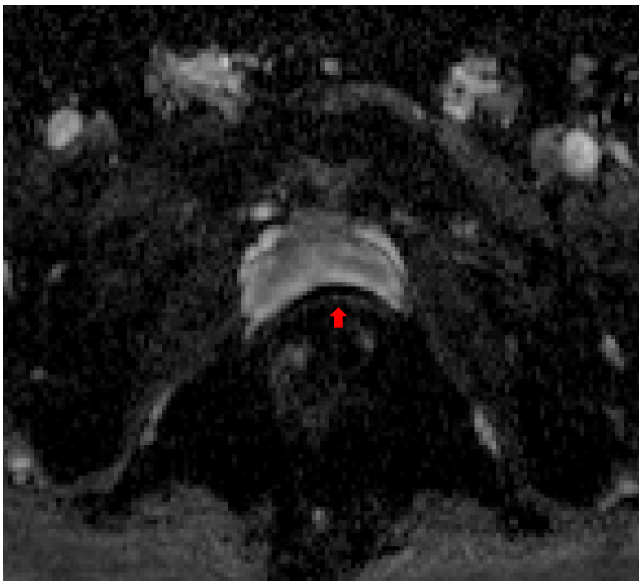


Figure 1.5: Transverse ADC map slice through the prostate showing a severe magnetic susceptibility artifact.

image quality limitations of acquired high b-value imaging is to use lower b-value images to extrapolate a high b-value image [24], which has been shown to achieve higher signal-to-noise ratios [29]. A final limitation of using DWI alone for PCa diagnosis is the presence of benign abnormalities that can mimic PCa on DWI, in particular prostatitis in the PZ [24].

1.2.2.3 Dynamic contrast-enhanced imaging

DCE images are acquired following rapid injection of a bolus of low molecular-weight gadolinium chelate [27]. Gadolinium is paramagnetic, therefore it causes local signal changes as it traverses the vascular system. In particular, the gadolinium chelate contrast agent causes a shortening of the T1 relaxation time, where T1, like T2, is a specific physical constant unique to tissues. Importantly, a T1-weighted pulse sequence will show increased signal where the Gadolinium contrast agent accumulates.

A lack of oxygen and/or nutrients at a PCa tumour site will promote the release of growth factors that induce the formation of new blood vessels [27]. The new vessels are thin, highly permeable, and irregular in shape, structure, and organisation. As a result, the contrast agent washes in and washes out of the tumour more

rapidly than in healthy prostate tissue. This can be observed through a series of fast T1-weighted pulse sequences with a 1-4 second time interval [27].

Through post-processing using compartmental mathematical models, parameters like "washout", "integral area under gadolinium-concentration-time curve", "wash-in gradient", "maximum signal intensity", "time to peak enhancement", and "start of enhancement" can be calculated. A k-trans curve can also be estimated as a measure of capillary permeability by estimating how much contrast agent has accumulated in the extravascular-extracellular space [27].

The main limitation of diagnosing PCa using DCEI is specificity. It is difficult to discriminate PCa from prostatitis in the PZ and BPH nodules in the TZ, both of which can be highly vascularised and can show increased and early enhancement on DCEI [27]. In addition, DCEI has become a subject of clinical debate due to the costs and risks associated to gadolinium injection, and unclear evidence that DCEI improves diagnostic accuracy [30]. A recent systematic review and meta-analysis by Woo et al. [30] sought to investigate whether a performance benefit is obtained from the use of DCEI. Strikingly, their study showed that the performance of radiologists with biparametric MRI (bpMRI) was similar to the performance of radiologists with mpMRI, for the diagnosis of PCa. However, a case may be made for the use of DCEI to support T2WI if DWI contains an obscuring magnetic susceptibility artifact.

1.2.2.4 Radiologist review: scoring

Radiologists score prostate mpMRI for the likelihood of CSPCa. Globally, the most popular scoring system used by radiologists is the Prostate Imaging-Reporting and Data System (PI-RADS). The first version of PI-RADS was an attempt to create a standardised mpMRI reporting structure, following recommendations from a European consensus meeting [31]. Since the release of PI-RADS v1, several modifications have been made in the subsequently released PI-RADS v2 [23] and PI-RADS v2.1 [24]. However, an alternative scoring system is advocated in the UK by NICE. NICE guidelines recommend the Likert scoring system for mpMRI reporting based on multicentre studies that have demonstrated its effectiveness within the NHS [9].

There are a number of similarities and differences between the two scoring systems. Both PI-RADS and Likert use a five-point scoring scale and utilise similar features on mpMRI. The main differences between the two systems are that PI-RADS advocates a sequential read, lesion-level scoring only, and an assignment of scores based on imaging features only, while Likert does not mandate a sequential read, allows the use of clinical data in addition to imaging, and allows scoring of lesions and the whole prostate.

1.2.2.5 Radiologist review: staging

In addition to scoring mpMRI for the likelihood of CSPCa, radiologists use mpMRI for staging PCa. TNM staging is a globally recognised staging methodology used to describe the extent of cancer spread, where T stands for tumour, N stands for node, and M stands for metastasis [32].

The T component of the stage describes the area of the cancer [32]. There are four subdivisions of the T component: T1 means the cancer is too small to be seen on mpMRI or felt during DRE; T2 indicates the cancer can be seen on mpMRI and is completely confined within the prostate; T3 indicates the cancer has broken through the prostate's capsule; and T4 means the cancer has spread to nearby organs, such as the back passage, bladder, or the pelvic wall.

The N component of the stage describes whether the cancer has spread to the lymph nodes [32]. There are two subdivisions of the N component: N0 indicates that the cancer has not spread to the lymph nodes, while N1 indicates that the cancer has spread to the lymph nodes.

The M component describes whether the cancer has metastasised to other parts of the body. There are two subdivisions of the M component: M0 means the cancer has not metastasised to other parts of the body, while M1 means the cancer has spread to other parts of the body outside of the pelvic region.

1.2.3 Prostate biopsy and Gleason scoring

Biopsy remains the only non-surgical method for confirming a PCa diagnosis, though due to sampling error, PCa cannot be ruled out if cancer tissue is not present

in the biopsy sample. NICE guidelines in the UK recommend MRI-guided biopsy of lesions scored greater than or equal to Likert 3 and biopsy omission for patients with mpMRI scored Likert 1 or 2 [9], while PI-RADS v2.1 recommends MRI-guided biopsy of lesions scored PI-RADS 4 or 5, biopsy omission for lesions scored PI-RADS 1 or 2, and the use of other information sources to determine whether a PI-RADS 3 lesion should be biopsied [24].

MRI-guided biopsies are typically performed under local anaesthetic. An alternative biopsy method is transperineal template prostate-mapping (TTPM) biopsy. TTPM biopsy is typically performed under general anaesthetic, and involves taking multiple cores from multiple sites using a grid system. Approximately 20 sites may be systematically sampled, with two or three cores per site. TTPM biopsy is accurate and avoids the image bias associated to MRI-guided biopsy and reduces the sampling error associated to random or systematic transrectal ultrasound-guided (TRUS) biopsy [33]. However, TTPM biopsy does carry a greater potential for side-effects due to the need for general anaesthesia [33]. Therefore, TTPM biopsy is not used routinely in clinic, but has featured in clinical trials of mpMRI to provide a robust reference standard [33, 26].

A Gleason score is given to tissue samples collected from biopsy to establish PCa aggressiveness through microscopic analysis [34]. Gleason grades range from 1-5, with grade 5 indicating the most aggressive tumour cells. By convention, grades are assigned for the two most common patterns seen under the microscope across all the biopsy samples taken from a tumour. The term Gleason score refers to the summation of the two grades. For example, a Gleason score written $3+4$, indicates that most of the cancer cells observed are grade 3 and a smaller proportion are grade 4. They are added together for a Gleason score of 7. Gleason scores will range from 6 (3+3) to 10 (5+5) for PCa. A further categorisation into Gleason Grade Groups with associated histopathological definitions has been outlined by Epstein et al. [35]:

- Group 1 (Gleason score ≤ 6): Only individual discrete well-formed glands.
- Group 2 (Gleason score $3+4=7$): Predominantly well-formed glands with

lesser component of poorly-formed/fused/cribriform glands.

- Group 3 (Gleason score 4+3 = 7): Predominantly poorly-formed/fused/cribriform glands with lesser component of well-formed glands.
- Group 4 (Gleason score $4+4=8$, $3+5=8$, $5+3=8$): Only poorly-formed/fused/cribriform glands or predominantly well-formed glands and lesser component lacking glands or predominantly lacking glands and lesser component of well-formed glands.
- Group 5 (Gleason scores 9-10): Lacks gland formation (or with necrosis) with or without poorly-formed/fused/cribriform glands.

1.2.4 Multidisciplinary team meeting

Following diagnostic tests, biopsy, and histopathological analysis, the multidisciplinary team (MDT) involved in patient care will meet to discuss clinical, imaging, and histopathological findings. NICE guidelines in the UK recommend that each patient should be assigned a risk rating by the MDT [9]; risk categories are shown in Table 1.1.

Level of risk	PSA	Gleason score	Clinical stage
Low	Less than 10 ng/ml and	6 or below and	T1 to T2a
Medium	10 to 20 ng/ml or	7 or	T2b
High	Higher than 20 ng/ml or	8 to 10 or	T2c or higher

Table 1.1: Risk stratification for patients with localised prostate cancer as described by UK NICE guidelines [9].

The main aim of the MDT meeting is to determine whether patients should be treated, and if so, the most appropriate treatment option, or whether patients should be placed on active surveillance or watchful waiting. Patient preferences will be taken into account prior to the MDT meeting and following the MDT meeting [9].

1.2.5 Radical treatment, active surveillance, or watchful waiting

UK NICE guidelines recommend radical treatment for patients with high risk localised or advanced PCa, while patients with low or intermediate risk localised PCa

may be offered a choice between active surveillance, watchful waiting, and radical treatment [9].

The most common radical treatments for PCa are prostatectomy and radiotherapy [36]. Prostatectomy is a surgical procedure to remove the prostate gland and tissues surrounding it - this usually includes the seminal vesicles and some nearby lymph nodes. Radical prostatectomy can cure prostate cancer in men whose cancer is limited to the prostate, though a 20-40% rate of biochemical recurrence has been reported [37]. In addition, common side-effects of prostatectomy are the inability to get an erection and urinary incontinence [36]. Alternatively, radiotherapy uses high energy radiation to kill cancer cells rather than surgical removal of the prostate. External beam radiotherapy uses X-rays administered from outside the body, while brachytherapy is an alternative that uses small radioactive seeds placed within the prostate [36]. For external beam radiotherapy and brachytherapy, biochemical recurrence rates of 30-50% have been reported [38]. Radiotherapy also carries side-effects including diarrhoea, bleeding, discomfort, cystitis, and inability to get an erection [36].

Active surveillance refers to monitoring a patient's PCa over time using PSA testing, DRE, mpMRI, and biopsy. Should there be a marked change in a patient's risk classification, patients on active surveillance may transition to radical treatment. As radical treatment carries risks and potential side-effects, active surveillance presents a viable option for those patients with slow-growing PCa. Studies have shown a 98% ten-year survival rate for patients offered active surveillance and that only 21% of patients offered active surveillance show signs of disease progression [9]. Alternatively, watchful waiting is a less intensive monitoring regime that involves managing symptoms as they arise, whose aim is to not transition patients to radical treatment.

1.3 Motivation of this work and thesis overview

In this work, we build and evaluate deep learning-based CAD systems for the PCa diagnostic pathway. The work in this thesis is motivated by challenges faced in

uroradiology, the growth of deep learning for medical image analysis tasks, and current gaps identified in the PCa CAD literature.

1.3.1 Rationale

Earlier in this chapter, we described the growing global cancer burden and a shortfall of radiologists to meet the projected demand on diagnostic radiology services. In the UK, PCa incidence is projected to rise by 12% between 2014 and 2035 to 233 cases per 100,000 [19]. In addition to the growth in PCa incidence, there is a growing use of mpMRI in the PCa diagnostic pathway. Studies have shown a considerable increase in sensitivity from a pathway that includes mpMRI over a pathway containing TRUS biopsy only [26, 33]. The diagnostic accuracy of radiologists interpreting mpMRI was reported by the PROMIS study [26]; on the task of identifying CSPCa (Gleason score $\geq 3+4$), radiologist Likert scoring (threshold: Likert ≥ 3) had a sensitivity of 88% and specificity of 45% compared to a sensitivity of 48% and a specificity of 99% for TRUS biopsy. While the increase in sensitivity made possible by the introduction of mpMRI is paramount to early diagnosis, improvements are needed to reduce the small proportion of men with CSPCa who are missed by mpMRI, to reduce the large number of men who undergo unnecessary biopsies, and to increase the inter-observer agreement between readers of varying experience and expertise [39]. In summary, the projected growth in PCa incidence, the projected shortfall in the radiology workforce, the increasing use of mpMRI, and the need for greater diagnostic accuracy and inter-observer agreement between radiologists has created the necessary motivation for the research and development of CAD systems for PCa diagnosis.

Deep learning has permeated the entire field of medical image analysis [40]. A review by Litjens et al. [40] summarizing over 300 contributions of deep learning in medical image analysis found successful applications in abdomen, brain, breast, heart, lung, prostate, and retina; tasks included patient classification, lesion classification, organ detection, lesion detection, organ segmentation, lesion segmentation, image registration, image generation, and image enhancement.

Deep learning-based CAD systems can be introduced into the PCa diagnostic

pathway to interpret mpMRI for varied applications. Prior to radiologist assessment of mpMRI, CAD systems can be deployed to perform patient triage using image and clinical data. CAD systems for patient triage could rank patients by disease severity or likelihood of having CSPCa, and potentially identify the lowest risk patients who do not require a clinical read, reducing radiologist workload. Alternatively, CAD systems can be used to provide an independent diagnosis in a second reader setting, which may be used to meet the double reporting recommendation outlined by the Likert assessment guidelines [25]. Rather than provide an independent diagnosis, CAD systems can also be designed to be companion systems for radiologists to help them identify and/or score lesions; several CAD system works have been published to evaluate the companion system paradigm where CAD-generated voxel-level probability maps were used by radiologists to identify and score lesions [15].

In addition to lesion assessment, radiologists use prostate mpMRI to estimate prostate volume and lesion volume using the ellipsoid formula [41]. Primarily, prostate volume is required for calculating the PSA density (PSAd), while lesion volume provides additional information for initial diagnosis, progression monitoring during active surveillance, and radical treatment planning. However, the ellipsoid formula is an approximation that ignores exact prostate and lesion morphology [41], therefore more accurate volume estimation methods are sought. As deep learning algorithms have achieved state-of-the-art performance for medical image segmentation [42], they may be used for prostate and lesion segmentation, from which volume estimates can be derived.

A further use of CAD systems deployed into the clinical workflow can be to enhance reporting quality. In accordance with PI-RADS and Likert guidelines, radiologists typically produce a text-based report of mpMRI findings [43]. Using CAD tools, there is potential for enhanced reporting with pictorial elements, automatic extraction of prostate and lesion characteristics, and automatic generation of report text using natural language processing (NLP).

1.3.2 Thesis overview

In this chapter, we introduced PCa and the PCa diagnostic pathway, which are the pathology and diagnostic pathway that motivate the development of the deep learning-based CAD systems presented in this thesis. Chapter 2 presents a literature review. The literature review explores various automatic and semi-automatic methods for medical image analysis, followed by an in-depth exploration of methods developed for and applied to PCa tasks, namely whole prostate and prostatic zones segmentation, PCa lesion detection, classification, and segmentation, and patient classification. Chapter 3 introduces the prostate datasets used to train and evaluate the CAD systems presented in this thesis, namely the publicly available PROSTATEx dataset [44] and the "Prostate Imaging Compared to Transperineal Ultrasound-guided biopsy for significant prostate cancer Risk Evaluation" (PICTURE) trial dataset [33]. Chapter 4 presents a novel patient classification framework, PCF, that assigns a probability of having CSPCa to patients based on mpMRI and clinical features, with applications in patient triage or as a second reader. In PCF, features are extracted from three-dimensional mpMRI and derived parameter maps using convolutional neural networks (CNNs) and subsequently combined with clinical features by a multi-classifier support vector machine (SVM) scheme. Chapter 5 presents AutoProstate, a deep-learning powered framework for automated MRI-based PCa assessment. AutoProstate uses bpMRI and clinical data to populate an automatic report, which can provide radiologists with useful information at the point of diagnosis, with the aim of increasing diagnostic accuracy and improving reporting quality. Chapter 6 explores interactive click-based segmentation of the whole prostate and prostatic lesions. Fully-automatic segmentation methods can fail under certain circumstances e.g., due to domain shifts between training and test data, image artifacts, and unseen pathologies [45], therefore, clinicians should have the ability to edit segmentations output by deep learning algorithms if clinical deployment is desired. Interactive segmentation has applications in prostate volume and PSA density calculation, diagnosis, active surveillance progression tracking, and radical treatment planning. Finally, Chapter 7 summarises the work presented

in this thesis and presents avenues of future research.

Chapter 2

Literature review

2.1 Machine learning for medical image analysis

Medical image analysis describes a field of research concerning automatic and semi-automatic methods for extracting information from medical images, primarily for diagnostic, treatment, and/or monitoring purposes. In a review article by Litjens et al. [40], over 300 machine learning contributions to the field of medical image analysis were described, predominantly for image classification, organ/lesion detection, organ/lesion classification, organ/lesion segmentation, image generation, and image denoising.

Machine learning is rooted in the idea that algorithms can make decisions by "learning" from data rather than being explicitly programmed to do so. Mitchell [46] provides a formal definition of learning: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." Generally speaking, machine learning algorithms can be divided into "supervised" and "unsupervised" learning algorithms [47]. In supervised learning, input-label pairs are available to the algorithm during the training stage during which the algorithm's parameters are optimised, while in unsupervised learning, the algorithm will attempt to find patterns in the input data, during training, without the use of labels [47]. Common supervised learning algorithms include linear regression, logistic regression, support vector machine, random forest, naive Bayes, k-nearest neighbour,

and discriminant analysis, while common unsupervised algorithms include k-means clustering and fuzzy c-means clustering.

Over the last decade, there has been a rapid growth in the use of a particular subclass of machine learning, known as deep learning [40]. Deep learning algorithms typically refer to neural networks with three or more layers [48]. In image analysis, convolutional neural networks (CNNs) are achieving state-of-the-art performance in several tasks [40]. The watershed moment for deep learning was the contribution of Krizhevsky et al. [49] to the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) in 2012. Their CNN, "AlexNet", won the challenge by a large margin, prompting interest and advances in the use of deep learning for computer vision and medical image analysis applications in the years following. The success of AlexNet was attributed to increased network depth compared to previous architectures and the use of novel deep learning components such as rectified linear unit (ReLU) activation [50], local response normalisation [49], overlapping pooling [49], and data augmentation. Critically, training AlexNet would not have been possible without the use of modern GPUs [49]. Since AlexNet, further CNN architectures have been developed that have pushed natural image classification accuracy beyond that of humans [51]. In 2014, "GoogLeNet" [52] won the ILSVRC classification challenge using a CNN architecture featuring "inception units" which allowed feature extraction at multiple scales [52]. In 2015, a "ResNet" CNN with 152 layers won the challenge using "residual units", which solved the degradation problem previously observed with increasing network depth [5]. A further improvement in classification accuracy was made by "SENet"; SENet won the ILSVRC challenge using newly proposed "squeeze-and-excitation" blocks [53], to achieve a a top 5 error rate of just 2.3%. The current state-of-the art for image classification is a family of models known as "EfficientNets" [54]. In their work, neural architecture search (NAS) was used to design a new baseline network, and a new scaling method was proposed to adjust network depth, width, and resolution for different tasks.

We now proceed to discuss classification and segmentation as these are the pertinent medical image analysis tasks explored in this thesis.

2.1.1 Classification

Classical machine learning algorithms have been used to classify patients with suspected mild cognitive impairment (MCI) or Alzheimer's disease (AD) e.g., [55, 56, 57]. In general, discriminant features were extracted from medical images. Extracted features were processed by a machine learning classifier to produce the classification output: Khan et al. [55] used twin SVMs [58] for patient classification using cortical and curvature features extracted from T1-weighted MRI; Jonkreangkrai et al. [56] used a SVM to classify patients using cortical volume and thickness measurements; and Casanova et al. [57] compared several machine learning classifiers for patient classification using T1-weighted MRI voxel values directly.

Several works have investigated the use of deep learning for patient classification e.g., [59, 60, 61, 62]. In Hosseini et al. [59], a 3D CNN was presented to classify patients with suspected MCI/AD using T1-weighted MRI; Shi et al. [62] proposed a deep polynomial network (DPN) to perform the same task as Hosseini et al., but using a combination of MRI and positron emission tomography (PET) images; Antony et al. [60] used a combination of CNN and SVM to classify knee osteoarthritis severity using X-ray images; and Pinaya et al. [61] proposed a deep belief network (DBN) for differentiating healthy controls and patients with schizophrenia. Rather than classifying patients, several works have investigated the use of deep learning for lesion classification e.g., [63, 64, 65]. Setio et al. [63] proposed a multi-stream 2D CNN architecture to classify individual pulmonary nodules on computerised tomography (CT) scans; Harangi [64] presented an ensemble of CNNs with varying architecture to classify skin lesions on dermoscopy images; and Cullell-Dalmau et al. [65] used a pre-trained ResNet-50, subsequently fine-tuned for skin lesions, to perform the same task as Harangi.

2.1.2 Segmentation

Segmentation is a highly relevant task in medical imaging for the delineation of structures, organs, and lesions on medical images [40]. Several methods have been developed for automatic/interactive segmentation of medical images.

2.1.2.1 Automatic segmentation algorithms

Clustering is a type of unsupervised machine learning that can be used to segment medical images, by organising voxels into groups based on image features e.g., intensity, texture, patterns, and shapes [66], and then assigning a class label to each group. A popular and relatively simple clustering algorithm is k-means [67]. The k-means clustering algorithm, in the context of medical image segmentation, assigns voxels to the nearest of k clusters, while keeping the clusters as compact as possible. As an example, in Ng et al. [68] the k-means clustering algorithm was used to segment brain MRI into bone, soft tissue, fat, and background. However, while the k-means algorithm is relatively simple, it does have a tendency to terminate at local optima [66]. On the other hand, the expectation-maximisation clustering algorithm [69] is more robust to local optima [70]. The idea is to assign voxels partially to different clusters instead of assigning them to only one cluster by modelling each cluster using a probability distribution. As an example, in Kwon et al. [71], the EM clustering algorithm was used to segment brain MRI into grey matter, white matter, and cerebrospinal fluid.

Deep learning methods based on CNNs are the current state-of-the-art for automatic 2D and 3D medical image segmentation [8, 72, 73], due in large part to the landmark contributions of Ronneberger et al. [6] (2D U-Net), Çiçek et al. [7] (3D U-Net), and Milletari et al. [74] (V-Net). The Medical Segmentation Decathlon (MSD) [75] is driving current methodological innovations and performance improvements. The MSD challenge asks participants to develop a machine learning algorithm that performs well on 10 different medical image segmentation tasks. At the time of writing, second position on the live leaderboard (`https://decathlon-10.grand-challenge.org/evaluation/challenge/leaderboard/`) is held by nnU-Net by Isensee et al. [8], which is a segmentation pipeline based on U-Net that automatically configures to any new medical image segmentation task, and first place is held by the Differentiable Network Topology Search (DiNTS) scheme, developed by He et al. [72], which allows architectures not confined to pre-defined topologies to be applied

to different medical image segmentation tasks. A recent work by Hatamizadeh et al. [73] introduces UNETR, which utilises a transformer [76] as the encoder, achieving favorable benchmarks on the MSD brain tumour and spleen segmentation tasks.

2.1.2.2 Interactive segmentation algorithms

Despite state-of-the-art performance on several medical image segmentation tasks, automatic segmentation algorithms have not yet reached the desired robustness for clinical use [45]. Interactive segmentation methods accept user-guidance to enable an improved segmentation [45]. Graph-cuts [77], normalised cuts [78], geodesics [79], and random walks [80] have been proposed for interactive segmentation using bounding box or scribble interactions. However, these methods succeed in simple settings with clear structural boundaries, but require extensive user interaction for more complex segmentation tasks [45].

Deep learning-based interactive segmentation methods have been proposed for robust segmentation of digital images [81, 82]. In Xu et al. [81], user foreground and background clicks were converted into euclidean distance maps, and subsequently added as additional input channels to a CNN, while in Agustsson et al. [82], users were expected to provide extreme point clicks and corrective scribbles. Inspired by the aforementioned works and other incremental works, deep learning-based interactive segmentation methods for medical image segmentation have been proposed [83, 45, 84]. In Wang et. al [83], a bounding box and scribble-based CNN segmentation pipeline was proposed, whereby an initial segmentation is obtained within a user-provided bounding box, followed by image-specific fine-tuning using user-provided scribbles. On the tasks of 2D segmentation of multiple organs on fetal MRI and 3D segmentation of brain tumour core and whole tumor using multiple MR sequences, their method proved more robust than state-of-the-art CNNs for segmenting previously unseen objects and more accurate with fewer user-interactions than traditional interactive segmentation methods. In contrast, Sakinis et al. [45] proposed a click-based interactive segmentation method, motivated by the work of Xu et al. [81]. In their work, Gaussian-smoothed user foreground and background clicks were added as input channels to an encoder-decoder CNN. Experi-

ments on multiple organ segmentation on CT showed that their method generated 2D segmentations in a fast and reliable manner, generalised well to unseen structures, and produced accurate results with few clicks. An alternate method that first performs an automatic CNN segmentation, followed by optional refinement through user clicks/scribbles, is proposed by Wang et al. [84]. Their method, DeepIGeoS, achieved substantially improved performance compared to automatic CNNs on 2D placenta and 3D brain tumour segmentation, and higher accuracy with fewer interactions compared to traditional interactive segmentation methods.

2.2 Machine learning for prostate MRI analysis

Machine learning methods are being used for whole prostate segmentation, prostatic zones segmentation, clinically significant prostate cancer (CSPCa) lesion detection/classification/segmentation, and the classification of patients with suspected CSPCa [15].

2.2.1 Whole prostate segmentation

Several automatic whole prostate segmentation methods have been presented in the literature [74, 85, 86, 87, 88, 89, 90]. Above all, the PROMISE12 Challenge has played a key role in ensuring consistent improvements in the performance of whole prostate segmentation algorithms over the past decade [91]. A notable submission to the challenge was made by Milletari et al. [74] in 2016, when they presented a novel encoder-decoder CNN architecture named V-Net for volumetric image segmentation, which was optimised using a novel Dice coefficient loss function. On the PROMISE12 challenge test set, V-Net achieved a mean Dice score of 0.87 for whole prostate segmentation. While the challenge originally took place at the MICCAI conference in 2012, the leaderboard remains active for online entries to the present day; at the time of writing, at least the top seven positions in the leaderboard are held by deep learning algorithms (the individual/team occupying position eight in the leaderboard has not disclosed the details of the algorithm they used), with the top ranking algorithm, named MSD-Net, achieving a mean Dice score of 0.92 on the PROMISE12 challenge test set. Recent works by Aldoj et al. [87] and

Cuocolo et al. [88] have achieved similar mean Dice scores on the publicly available PROSTATEx dataset [44]. Aldoj et al. presented a novel CNN architecture named Dense-2 U-Net, which was inspired by DenseNet [92] and U-Net. Four-fold cross-validation of 188 patients from the PROSTATEx training dataset yielded a mean Dice score of 0.92. In Cuocolo et al. [88], the previously proposed ENet [93] was evaluated on 105 patients from the PROSTATEx training dataset, on which a mean Dice score of 0.91 was achieved.

Segmentation of the whole prostate can be used to obtain an estimate of prostate volume. To the best of our knowledge, only the work by Lee et al. [94] has compared prostate volume estimation using an automatic segmentation method to the clinically utilised ellipsoid formula. On a 70-patient test set, their 3D CNN for whole prostate segmentation achieved a mean Dice coefficient of 0.87 and a mean volume absolute percentage error (Abs%Err) of 11.78%, while the mean volume Abs%Err of the ellipsoid formula was 11.92%.

2.2.2 Prostatic zones segmentation

Several automatic methods for prostatic zones segmentation have been presented in the literature [8, 72, 87, 88, 95]. Makni et al. [95] used evidential C-means clustering to segment prostatic zones on T2WI; on a 31-patient test set, their algorithm achieved mean Dice scores of 0.78 for the peripheral zone (PZ) and 0.88 for the central gland (CG). The zonal segmentation task is included in the Medical Segmentation Decathlon; on a test set of size 16, nnU-Net, introduced by Isensee et al. [8], achieved mean Dice scores of 0.77 and 0.90 for the PZ and CG respectively, while the DiNTS framework, presented by He et al. [72], achieved mean Dice scores of 0.75 and 0.89. Alternatively, on the PROSTATEx dataset, the Dense-2 U-Net CNN presented by Aldoj et al. [87] achieved mean Dice scores of 0.78 and 0.91 for the PZ and CG respectively, for four-fold cross-validation of 188 patients, while E-Net evaluated by Cuocolo et al. [88] achieved mean Dice scores 0.71 and 0.87 for the PZ and CG respectively, on 105 held-out patients.

2.2.3 Computer-aided diagnosis of prostate cancer

A systematic review of CAD systems that use AI for MRI-based PCa diagnosis was published by Syer and Mehta et al. in July 2021 [15]. The key selection criteria for including studies in the systematic review was the need for CAD system performance to be measured against a histopathological reference standard and compared to radiologist interpretation. Since radiologist interpretation of mpMRI is current clinical practice, a comparison to radiologist interpretation provides valuable insight into the utility of the CAD system being evaluated if it were to be deployed into the clinical workflow.

The review of CAD systems for PCa diagnosis presented in this chapter follows from the systematic review by Syer and Mehta et al. The study selection criteria defined by Syer and Mehta et al., is as follows: studies were included if (i) they evaluated CAD for PCa detection or classification on MRI, (ii) CAD performance was compared to radiologist interpretation and against a histopathological reference standard, (iii) the evaluation patient cohort was treatment-naïve, and (iv) a full-text article was available; and studies were excluded if (i) MRI sequences other than T1-weighted imaging (T1WI), T2-weighted imaging (T2WI), diffusion-weighted imaging (DWI), or dynamic contrast-enhanced imaging (DCEI) were used, (ii) the comparator radiologist(s) did not have access to at least axial T2WI and DWI with apparent diffusion coefficient (ADC) map for reporting, and (iii) the patient cohort used for testing was less than thirty patients.

Twenty-seven studies met the selection criteria [96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122]. Two further studies of relevance, published after the systematic review search date (25 March 2021), are included in this chapter [123, 124].

2.2.3.1 Study characteristics

All studies were published between 2013 and 2021 from groups spanning Asia, Europe, and the USA. A retrospective study design was followed by all studies. The size of patient cohorts used for evaluation varied substantially. The smallest patient cohort used for evaluation was 30 patients, used in the work of Niaf et al. [105].

In contrast, the recent work by Saha et al. [123] used institutional scans from 486 patients and external scans from 296 patients for evaluation. Notably, the work by Saha et al. also featured the largest training cohort of size 1950. Histopathological reference standards used in studies also varied substantially. Studies used one or a combination of the following: transperineal template prostate-mapping (TTPM) biopsy, in-bore targeted biopsy, TRUS targeted biopsy, TRUS saturation biopsy, TRUS systematic biopsy, or radical prostatectomy. The majority of studies collected scans using 3T MR scanners, while fewer studies used 1.5T MR scanners [99, 105, 106, 114]. Few studies evaluated CAD systems using multicenter MRI data [96, 113, 117], and few studies used multivendor MRI data for evaluation [96, 100, 113]. Notably, work by Gaur et al. [113] evaluated CAD performance against reader performance on a multicenter external test cohort featuring scans from five institutions based in four countries spread over three continents.

2.2.3.2 Computer-aided diagnosis system taxonomy

CAD system studies, and by extension the CAD systems presented within them, are described as patient classification (PAT-C), region of interest (ROI) Classification (ROI-C), and lesion localisation and classification (LL&C). PAT-C refers to studies where CAD systems classified patients directly, ROI-C refers to studies where CAD systems classified pre-defined ROIs, e.g., manually contoured lesions, and LL&C refers to studies where CAD systems performed simultaneous lesion localisation and classification. The typical workflow of each type of CAD system is shown in Figure 2.1.

2.2.3.3 Patient classification (PAT-C) systems

The work by Deniffel et al. [122] is the only study of type PAT-C among the included studies. Deniffel et al. presented a 3D CNN that classified patients as having clinically significant disease (Gleason score $\geq 3+4$) or not, using T2WI, ADC map, and b1600 DWI. CAD system and radiologist sensitivity and specificity was presented at several probability thresholds. At a probability threshold of ≥ 0.2, their CAD system achieved a per-patient sensitivity and specificity of 100% and 52% respectively, which exceeded the joint performance of two radiologists with 3 and 15

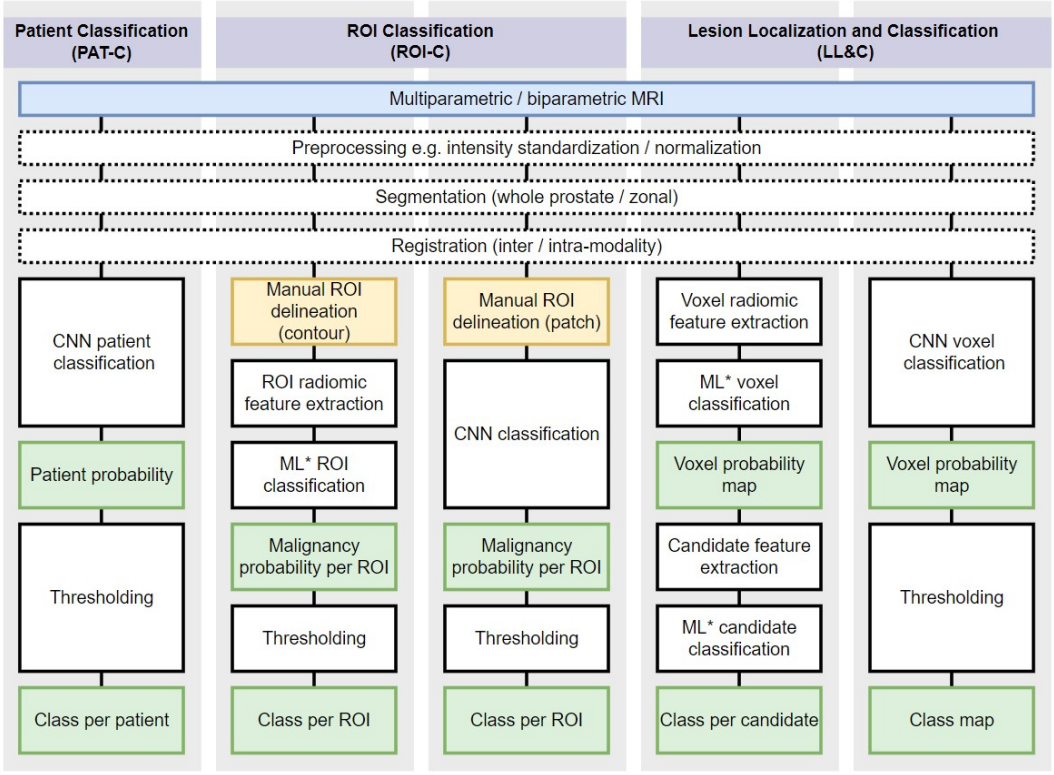


Figure 2.1: Taxonomy of CAD systems for PCa diagnosis. CAD systems are categorized as Patient Classification (PAT-C), ROI Classification (ROI-C), or Lesion Localization and Classification (LL&C). Blue indicates mpMRI/bpMRI input, yellow indicates manual processes, white indicates automated processes, and green indicates intermediate or final outputs. ROI = region of interest. CNN = convolutional neural network. ML = machine learning. ML* here refers to ML algorithms exclusive of CNNs, such as support vector machines, random forest, logistic regression, and artificial neural networks.

years of experience each, who read a subset of cases; the joint radiologist sensitivity and specificity was 95% and 35% respectively. However, since the threshold for the probabilistic output of the CNN was not pre-specified or determined using training data, the performance may not be a true reflection of how the CAD system would perform prospectively.

2.2.3.4 Region of interest classification (ROI-C) systems

Several studies of type ROI-C have been published. Algorithms used for ROI classification include, but are not limited to: logistic regression [99, 102, 103, 104, 106], support vector machine [105, 108], random forest [98, 109], generalised linear mixed model [100, 107], quadratic discriminant analysis [96], linear discriminant

analysis [101], and CNN [111].

Few ROI-C studies have reported superior diagnostic accuracy for CAD compared to radiologist interpretation. In Wang et al. [108], a novel SVM classifier was trained using radiomic features extracted from T2WI, ADC map, b1500 DWI, and DCEI, to classify index lesions using a clinical endpoint of Gleason score $\geq 3+3$ and lesion size $\geq 0.5\text{cm}^3$. Using leave-one-patient-out (LOPO) cross-validation (CV) of 54 patients who had undergone radical prostatectomy, the SVM classifier achieved a sensitivity of 90% and a specificity of 88%, compared to a sensitivity of 76% and a specificity of 91% for the consensus view of two readers with over 10 years' experience in reading prostate MRI; the difference in sensitivity was statistically significant. In Dinh et al. [100], a GLMM classifier was trained using radiomic features extracted from ADC map and DCEI. Using a clinical endpoint of Gleason score $\geq 3+4$, their GLMM classifier achieved a sensitivity of 96% and a specificity of 44% on a temporally separated cohort of 129 patients with combined TRUS and targeted biopsy reference standard, compared to a sensitivity of 100% and a specificity of 14% for nine radiologists of varying experience who read a subset of cases each; only the difference in specificity between the GLMM classifier and radiologists was statistically significant. In Niu et al. [106], zone-specific logistic regression models were trained using T2WI and ADC maps. CAD and reader performance was assessed on 184 patients with combined TRUS biopsy and targeted biopsy reference standard, using a clinical endpoint of Gleason score $\geq 3+4$. For lesions in the PZ, logistic regression achieved a sensitivity of 87% and a specificity of 89%, while radiologist interpretation yielded a sensitivity of 79% and a specificity of 75%, and for lesions in the TZ, logistic regression achieved a sensitivity of 88% and a specificity of 81%, while radiologist interpretation yielded a sensitivity of 73% and a specificity of 77%; in both zones, logistic regression was found to be statistically better than radiologist interpretation in terms of both sensitivity and specificity. In contrast, the work by Transin et al. [107] showed inferior sensitivity for a GLMM classifier compared to a radiologist with 20 years of experience in prostate imaging, with statistical significance; unlike the aforementioned

works, the work by Transin et al. performed an evaluation using externally-obtained test data. In other studies, there were no statistically significant differences between CAD systems and radiologists [96, 97, 98, 99, 109, 111, 103].

CAD systems which accept the radiologist's reporting score as input alongside MRI features have been investigated [108, 110, 104, 103], three of which showed significant improvement upon the radiologist's score alone. Li et al. [103] combined a CAD likelihood score with a PI-RADS v2.1 score and a prostate-specific antigen (PSA) value, using a logistic regression classifier, reporting an increased area under the receiver operating characteristic curve (AUC) compared to radiologist PI-RADS v2.1 assessment alone, with statistical significance. In Litjens et al. [104], a CAD likelihood score was combined with a PI-RADS v1 score, using a logistic regression classifier; they reported an increased specificity over radiologist assessment using PI-RADS v1, with statistical significance. In Wang et al. [108], a support vector machine classifier was used to combine radiomic features and a PI-RADS v2 score; they found an increase in sensitivity over radiologist PI-RADS v2 assessment alone, with statistical significance. A further two studies compared radiologist interpretation with and without knowledge of CAD scores [101, 105], for which no significant differences were demonstrated.

2.2.3.5 Lesion localisation and classification (LL&C) systems

Several CAD systems that perform lesion localisation and classification using classical machine learning algorithms have been published in the literature. A comprehensive work, evaluated on a large patient cohort of size 347, is by Litjens at al. [116]. They presented a two-stage CAD system featuring a radiomic feature extraction and classification stage, to generate a voxel probability map for each patient, followed by candidate selection, candidate feature extraction, and classification of each candidate using a random forest classifier, such that a likelihood of PCa is obtained for each candidate. After excluding biopsy-proven low-grade tumours (Gleason score = 6), they computed a voxel-level AUC of 0.89 and a patient-level AUC of 0.81 for high-grade PCa (Gleason score $\geq$ 7) vs normal/benign, using leave-one-patient-out (LOPO) cross-validation, computed against a targeted biopsy

reference standard. At a high specificity threshold for both CAD system and radiologist ($>$ 20 years' experience in reading prostate MRI), performance did not differ significantly, however, radiologist performance was superior at a high sensitivity threshold. Giannini et al. [114], Greer et al. [115], and Zhu et al. [121] compared CAD system and radiologist performance on independent internal test cohorts. In Giannini et al. [114], quantitative features were extracted from T2WI, ADC map, and DCEI at the voxel level. An SVM classifier was trained to classify voxels as GS $\geq 3+4$ or not. On 89 patients with scans acquired using 1.5T scanners and targeted biopsy or saturation biopsy (at least 28 cores taken from each patient) reference standard, the CAD system achieved a sensitivity of 81%, while the average sensitivity of three radiologists with 2-4 years' experience in reading prostate MRI was 72%; the difference between CAD system and radiologist sensitivity was not statistically significant. In addition, Giannini et al. considered the performance of radiologists using the CAD system output probability map as an aid, where readers were restricted to choose from CAD system highlighted areas only; significant differences in performance were not observed between radiologist performance with and without CAD-assistance. Rather than restrict readers to choose from CAD system highlighted areas only, Zhu et al. [121] compared the unconstrained performance of readers before and after seeing the CAD system's output; they found that CAD-assisted diagnosis increased per-patient sensitivity from 84% to 93%, with statistical significance, compared to readers alone, on an 153-patient independent internal test cohort with combined TRUS and targeted biopsy reference standard. The works by Gaur et al. [113] and Mehralivand et al. [117] must be highlighted for conducting multicenter/multivendor studies; both works evaluated CAD using images acquired from five centers based across multiple countries. Such studies have a large role to play in providing supporting evidence for the clinical translation of CAD systems. In their reader study, Gaur et al., found the use of CAD improved reader specificity with statistical significance and reduced reading time, however a statistically significant drop in reader sensitivity was observed, on a patient cohort of size 216. The study by Mehrilivand reached a somewhat contra-

dictory conclusion, reporting a minimal improvement in reader sensitivity, without statistical significance, a drop in specificity, with statistical significance, and an increase in reading time, for CAD-assistance, on a patient cohort of size 236.

Recent studies have investigated deep learning for simultaneous lesion localisation and classification. Cao et al. [112] presented a novel multi-class CNN named "FocalNet" to jointly detect PCa lesions and predict their aggressiveness. Using five-fold cross-validation of a 417-patient dataset with radical prostatectomy reference standard, FocalNet demonstrated comparable detection sensitivity to a radiologist with over 10 years' experience in reading prostate MRI, for index lesions (89.7%) and clinically significant lesions (87.9%), at one false positive per patient. In the work of Schelb et al. [118], a U-Net CNN was shown [6] to produce similar CSPCa detection performance to PI-RADS v2 scoring by eight radiologists who read a subset of cases each. On the held-out test cohort of 62 men sampled from the same study cohort as the training data, with combined TRUS and targeted biopsy reference standard, their method obtained a patient-level sensitivity of 92% and specificity of 47%, while radiologist assessment yielded a sensitivity of 88% and a specificity of 50%; differences in sensitivity and specificity between the CNN and radiologist assessment were not statistically significant. A recent work by Saha et al. [123] presented a multi-stage 3D CAD system based on CNNs. Deep attention mechanisms in the detection network target structures salient to distinguishing CSPCa from indolent/benign abnormalities, while a residual CNN is used in a novel false-positive reduction step. Notably, they trained the CAD system using a large training cohort of 1950 patients, which was made possible by relaxing the need for biopsy-confirmed ground-truth for training. They tested their CAD system on 486 institutional patient cases without biopsy-confirmation and 296 external patient cases with biopsy-confirmation. On the 486 institutional cases, the CAD system achieved sensitivities of 84% and 93% at 0.50 and 1.46 false-positives per patient respectively. On the external cases, the CAD system achieved a 91% sensitivity at 1.29 false-positives per patient, while radiologists achieved a 91% sensitivity at 0.30 false-positives per patient.

2.3 Gaps in the literature addressed by this thesis

This chapter highlights the extensive efforts of research groups globally who are seeking to address known issues in the PCa diagnostic pathway through the development of AI technologies. However, there are gaps in the literature, which the work presented in the remaining chapters of this thesis seek to address. CAD systems that perform patient classification can be deployed into the PCa diagnostic pathway to perform triage or to provide a second read, following the radiologist's first read. In a triage deployment, a CAD system that performs patient classification can rank patients by likelihood of CSPCa, to ensure the patients with the highest likelihood of harbouring CSPCa are assessed first by a radiologist, providing a smarter alternative to the first-in-first-out approach to assessing patient cases used currently. Ranking patients by likelihood of having CSPCa may also allow ruling-out of the lowest risk patients, alleviating radiologist workload. Further motivation for triage/rule-out is provided by the potential introduction of MRI-based PCa screening programmes for men, which would increase the demand on prostate radiology services quite considerably [125]. As demonstrated, a compelling case can be made for the development of CAD systems for patient classification. As a result, Chapter 4 presents a novel deep learning-based patient classification system, PCF, that uses both mpMRI and clinical features, to output a patient-level probability of CSPCa (Gleason score $\geq 3+4$). The work presented in Chapter 4 departs from the patient classification work by Deniffel et al. [122] in three important ways. First and foremost, PCF considers clinical features such as PSA density (PSAd) alongside MRI features to perform classification, in line with the Likert assessment guidelines for radiologists [25]. Second, our work on PCF includes a methodology for including DCEI as an input, if collected, while the work by Deniffel et al. considers bpMRI only. In PCF, contrast-enhanced images collected over multiple timepoints are converted into parameter maps that reflect characteristics of the enhancement profile of each voxel. Finally, the performance of PCF is evaluated by picking operating thresholds using the training data, which are used to measure performance on the test data used for evaluation. In contrast, Deniffel et al. calculated sensitivity and specificity on the

test data at multiple arbitrary risk thresholds, and therefore we cannot interpret how their system would perform prospectively on unseen test data.

The work in Chapter 5 seeks to address two gaps identified in the literature. Firstly, current studies do not feature CAD systems that can produce an automatic diagnostic report output. Automatic diagnostic reports would provide useful information to radiologists at the time of diagnosis and enhance reporting quality beyond the textual findings recorded by radiologists currently [43]. In Chapter 5, we present AutoProstate, a deep learning-powered framework for PCa assessment and reporting. Importantly, AutoProstate is a complete framework that takes raw image and clinical data as input to output an automatic web-based diagnostic report. The second gap addressed by the work in Chapter 5 is to perform a thorough external validation of our presented system. At present, only the work by Saha et al. [123] has performed an external validation of a deep learning-based CAD system. However, their external dataset does not contain scans acquired from a scanner manufactured by a different vendor to the scans used for training; the need to evaluate CAD systems using multivendor MRI has been outlined in the key considerations for authors, reviewers, and readers of AI Manuscripts in radiology by Bluemke et al. [126]. In our work, AutoProstate is trained using the publicly available PROSTATEx dataset [44] collected using Siemens scanners, and the external validation is performed using the "Prostate Imaging Compared to Transperineal Ultrasound-guided biopsy for significant prostate cancer RISK Evaluation" (PICTURE) dataset [127] which is collected using Phillips scanners. Furthermore, our external validation uses a combined targeted and transperineal template prostate-mapping (TTPM) biopsy reference standard, which avoids biases associated to other reference standards such as prostatectomy, targeted biopsy alone, and TRUS random/systematic biopsy [33]. In addition, our evaluation ensures that the CAD system probability threshold used to segment CSPCa lesions is set without reference to the external validation data, which is a pitfall seen in some studies in the literature [108, 111, 112, 120].

Interactive segmentation methods are vitally important if clinical deployment of CAD systems is desired, to allow robustness to failure cases [45] and to allow for

continuous optimisation of CAD systems through active learning. At present, there are no works in the literature that have investigated interactive segmentation of the whole prostate or lesions within the prostate. Both applications have a diagnostic utility e.g., for computing accurate prostate and lesion volumes, and beyond diagnosis e.g., for monitoring progression in patients placed on active surveillance and for treatment planning. Therefore, in Chapter 6, we build and evaluate applications for click-based interactive whole prostate and prostatic lesion segmentation.

Chapter 3

Patient datasets for training and evaluation

Computer-aided diagnosis (CAD) systems that are based on machine learning algorithms require data for training, validation, and testing. Usually, the training data is used to tune the algorithm's learnable parameters, validation data is used to select hyperparameters, while the testing data (sometimes referred to as "holdout data") is used to obtain an estimate of performance on data that has not been used during parameter/hyperparameter optimisation. Two datasets were used to train, validate, and test the CAD systems presented in this thesis. The first is the publicly available PROSTATEx Challenges data [44] and the second is the "Prostate Imaging Compared to Transperineal Ultrasound-guided biopsy for significant prostate cancer Risk Evaluation" (PICTURE) study dataset [33].

3.1 PROSTATEx dataset

In this thesis, the PROSTATEx dataset refers to the union of the data released for the PROSTATEx Challenge and PROSTATEx-2 Challenge. The PROSTATEx dataset was originally used to train and evaluate the CAD system in a work by Litjens et al. [116], prior to its subsequent release for the two challenges. A total of 346 consecutive patient studies are available for download from the PROSTATEx Challenges database [44], which is hosted by The Cancer Imaging Archive (TCIA). The database features multiparametric MRI (mpMRI) and histopathological findings for

men examined at Radboud University Medical Center between 2011 and 2012. The PROSTATEx dataset is described in detail in the subsections to follow, and a tabular summary is provided in Table 3.1.

3.1.1 Multiparametric MRI protocol

MpMRI was acquired using two 3-Tesla magnetic field scanners (Magnetom Trio and Skyra, Siemens) and a pelvic-phased array coil. All studies included T2-weighted imaging (T2WI), diffusion-weighted imaging (DWI), and dynamic contrast-enhanced imaging (DCEI). T2WI was acquired using a turbo spin echo sequence, at a median in-plane resolution of 0.5 mm and a median slice thickness of 3 mm. DWI was acquired using a single-shot echo planar imaging sequence, at a median in-plane resolution of 2 mm and a median slice thickness of 3 mm. Three b-values were collected (50, 400, and 800); DWI collected at these b-values were used to generate an apparent diffusion coefficient (ADC) map using scanner software. DCEI was acquired using a 3D turbo flash gradient echo sequence, at a median in-plane resolution 1.5 mm, a median slice thickness of 4 mm, and a median temporal resolution of 3.5 seconds. All images were acquired without the use of an endorectal coil, as per Prostate Imaging-Reporting and Data System (PI-RADS) guidelines.

3.1.2 Multiparametric MRI review

All mpMRI studies were reported by an experienced radiologist with over 20 years' experience in reading prostate mpMRI, who highlighted areas of suspicion per modality with a point marker and scored them using PI-RADS v1.

3.1.3 Histopathological reference standard

MR-guided targeted biopsies of marked points with PI-RADS v1 score ≥ 3 were performed, while marked points with PI-RADS v1 score < 3 were assumed clinically insignificant ($< 5\%$ incidence of clinically significant prostate cancer (CSPCa) in PI-RADS v1 < 3 lesions at Radboud University Medical Center) and therefore not biopsied; here, clinically significant refers to Gleason score $\geq 3+4$ disease. Subsequently, biopsy specimens were graded by a histopathologist to determine a

Gleason score for each marked lesion. The marked point coordinate and a ground-truth label (clinically significant equal to true or false) for each marked lesion was released publicly for 204 of the 346 patients (PROSTATEx Challenges training set), while for the remaining 142 patients (PROSTATEx Challenges test set) the marked point coordinate was released publicly, but the ground-truth label was not. In total, 330 lesions were present in the 204 training set patients, while 208 lesions were present in the 142 test set patients; a breakdown of marked lesions by zonal location and Gleason score is shown in Table 3.1.

3.1.4 Contours

Whole prostate, peripheral zone (PZ), central gland (CG), and lesion contours for the 204 PROSTATEx training set patients were performed by an external group at the University of Naples [88]. In summary, contours were produced by a team consisting of two radiology residents ($>$ 2 years' experience in reading prostate mpMRI) and two board-certified radiologists ($>$ 5 years' experience in reading prostate mpMRI). Radiology residents and board-certified radiologists worked in pairs for quality control and annotation of each case. Two hundred and four whole prostate and zonal contours were drawn, while a total of 299 lesion contours were drawn, including 76 CSPCa lesions and 223 low-grade or benign lesions (nCSPCa); a breakdown of annotated lesions by Gleason score and zonal location is shown in Table 3.2.

3.2 PICTURE dataset

Full details of the PICTURE study have previously been reported [127, 33]. The PICTURE study was a paired-cohort validating confirmatory study designed to measure the diagnostic accuracy of mpMRI in men who had undergone an initial standard transrectal ultrasound-guided (TRUS) biopsy, but were advised to have further biopsies as part of standard care. Men were examined at University College London Hospital between 2012 and 2014. Inclusion criteria for the PICTURE study were: (i) men who had undergone an initial standard TRUS biopsy, but concern remained over the accuracy of the subsequent diagnosis; and (ii) men suitable

Variable	PROSTATEx Challenges train set	PROSTATEx Challenges test set
No. of patients	204	142
Total marked lesions	330	208
Per-lesion marked point zone		
Peripheral zone	191	113
Central gland	137	93
Seminal vesicle	2	2
Per-lesion Gleason score		
Not biopsied or benign following biopsy	218	Data not available
Gleason score ≤ 6	36	
Gleason score 3+4	41	
Gleason score 4+3	20	
Gleason score 8	8	
Gleason score 9-10	7	
Per-patient maximum Gleason score		
Not biopsied or only benign findings following biopsy	105	Data not available
Gleason score ≤ 6	29	
Gleason score 3+4	38	
Gleason score 4+3	19	
Gleason score 8	7	
Gleason score 9-10	6	

Table 3.1: PROSTATEx Challenges dataset characteristics.

for further characterisation using transperineal template prostate-mapping (TTPM) biopsy. Exclusion criteria were: (i) previous history of prostate cancer treatment; and (ii) lack of complete gland sampling or inadequate sampling density at TTPM. A breakdown of patient characteristics is shown in Table 3.3.

3.2.1 Multiparametric MRI protocol

MpMRI was acquired using a 3-Tesla magnetic field scanner (Achieva, Philips Healthcare) and a pelvic-phased array coil. All studies included T2WI, DWI, and DCEI. T2WI was acquired using a turbo spin echo sequence, at a median in-plane resolution of 0.4 mm and a median slice thickness of 3 mm. DWI was acquired using a spectral attenuated inversion recovery (SPAIR) sequence, at a median in-plane

Variable	#
No. of patients with annotated lesions	200
No. of patients without annotated lesions	4
Total annotated lesions	299
Per-lesion zone	
Peripheral zone	157
Central gland	122
Both zones	20
Per-lesion Gleason score	
Not biopsied or only benign findings following biopsy	187
Gleason score ≤ 6	36
Gleason score 3+4	41
Gleason score 4+3	20
Gleason score 8	8
Gleason score 9-10	7

Table 3.2: PROSTATEx training dataset annotations.

resolution of 1.25 mm, and a median slice thickness of 5 mm. DWI was acquired at four b-values (0, 150, 500, and 1000); DWI collected at these b-values were used to generate an ADC map using scanner software. In addition, DWI with a higher b-value (2000) was collected using a spectral inversion recovery (SPIR) sequence, with the same resolution and slice thickness as the lower b-value images. DCEI was acquired using a SPAIR sequence, at a median in-plane resolution 1 mm, a median slice thickness of 3 mm, and a median temporal resolution of 13 seconds.

3.2.2 Multiparametric MRI review

All mpMRI studies were reported by an experienced radiologist with over 10 years' experience in reading prostate mpMRI, using a five-point Likert impression scale for the likelihood of CSPCa [31]. Three definitions of clinical significance were used in the PICTURE study. The primary outcome was Gleason score $\geq 4+3$ or cancer core length (CCL) ≥ 6 mm of any Gleason score, the secondary outcome was Gleason score $\geq 3+4$ or CCL ≥ 4 mm of any Gleason score, and the tertiary outcome was Gleason score $\geq 3+4$ of any CCL. Scoring was completed at the lesion, sector, and patient-levels. All definitions of clinical significance were scored

against at the patient-level, while the third definition only was scored against at the sector and lesion levels also. Referral prostate-specific antigen (PSA) was available to the radiologist during scoring to reflect clinical practice. A breakdown of Likert scores at the lesion and patient levels is shown in Table 3.3.

3.2.3 Histopathological reference standard

Men underwent MR-guided targeted biopsy of focal index lesions and TTPM biopsy of the whole gland as the reference standard. For TTPM biopsy, core needles were inserted via a brachytherapy grid with 5 mm spacing, fixed on a stepper. TTPM biopsy was used to overcome the inaccuracies of TRUS biopsy [26] and the selection bias towards men with aggressive disease associated with radical prostatectomy [128]. Altogether, 249 men completed mpMRI and TTPM biopsy. All biopsy samples were reported by one of two expert histopathologists with over 20 years of experience each, who were blinded to mpMRI reports. Furthermore, all negative biopsies were double-reported for quality control. A breakdown of Gleason scores at the patient level is shown in Table 3.3.

3.2.4 Contours

Whole prostate, PZ, and CG contours were drawn by a board-certified radiologist (3 years' experience in the quantitative analysis of prostate mpMRI) for 80 patients. Lesions were delineated by two board-certified radiologists (4 and 5-years' experience in scoring prostate mpMRI using Likert assessment and PI-RADS v2, respectively) who annotated a subset of cases each. First, histopathology reports from MR-guided targeted and TTPM biopsies were reviewed alongside mpMRI to locate the highest Gleason grade focal lesion; if there were multiple focal lesions with the maximum Gleason grade, the highest scoring focal lesion according to Likert or PI-RADS v2 was identified. Next, a single axial T2WI slice was selected corresponding to the center of the identified lesion. Then, all focal lesions on the selected slice were contoured. Additionally, focal benign lesions that were scored Likert or PI-RADS v2 ≥ 4 were contoured in patients that were biopsy-negative for cancer. A total of 210 lesions were delineated (147 CSPCa lesions and 63 nCSPCa

lesions). A breakdown of Likert scores, Gleason scores, and zonal locations for annotated lesions is shown in Table 3.3.

3.3 Bias in prostate datasets

Sources of bias in prostate datasets should be reduced where possible and at minimum, carefully understood in order to comment on the limitations of experiments conducted.

A large source of bias relates to patient selection. Patient selection methods include, but are not limited to, consecutive enrolment, random selection, or selection according to some criteria related to a clinical question to be answered. The PROSTATEx dataset features consecutive patient studies from Radboud University Medical Centre, while the PICTURE dataset features patients who had undergone an initial standard TRUS biopsy, but opted for repeat evaluation due to concerns over the initial diagnosis; as a result, the risk profiles of the PROSTATEx and PICTURE datasets will likely vary. In addition, the patient selection methodology may create class imbalance problem, as can be observed for the PICTURE dataset; of 249 total patients, only 34 patients are without PCa, which is in contrast to the the PROSTATEx dataset, where a larger proportion of patients were found to have benign conditions only following biopsy or were deemed too low-risk to undergo biopsy. Notably, patients can be selected from a single centre or from multiple centres. For the purposes of training and evaluating an AI algorithm, multi-centre datasets, such as those used by Gaur et al. [113] and Mehralivand et al. [117] allow a more comprehensive evaluation of the generalisability of an algorithm, however multi-centre datasets are harder to collect, and the motivation may only arise following strong indications from single-centre evaluations. Both the PROSTATEx and PICTURE datasets are single-centre datasets.

Bias also stems from the acquisition protocol used to collect MRI. A large point of difference between centres concerns the inclusion of DCE imaging in the acquisition protocol. While the PI-RADS and Likert guidelines continue to incorporate DCE into reader guidelines, there is a growing movement towards biparametric

MRI (bpMRI) [30]. Both the PROSTATEx and PICTURE datasets include DCE. Other factors to consider include, but are not limited to, scanner manufacturer, field strength, specific sequence settings, and the type of receiver coil used.

The reference standard can also introduce bias. Reference standards vary substantially across studies and datasets [15]. Prostatectomy achieves complete gland coverage, but includes a selection bias towards men with aggressive disease [33]. Targeted biopsy as a reference standard includes a bias towards the radiologist who scored candidate tumours, relying on their expertise to identify which tumours should and should not undergo biopsy. The PROSTATEx dataset uses a targeted biopsy reference standard, with lesions scored by a radiologist with over 20 years of experience in reading prostate MRI, increasing confidence in the lesions chosen for targeting. Alternatively, the PICTURE study used a reference standard that combines TTPM biopsy and targeted biopsy. TTPM biopsy avoids the biases associated to prostatectomy and targeted biopsy, though is usually reserved for trials [9].

Bias towards a radiologist can also be introduced during lesion contouring if curating a dataset for the task of lesion detection or segmentation. An inter-observer variability has been observed for lesion contouring as reported by Steenbergen et al. [129]. Steps were taken to account for the variability when contouring PROSTATEx dataset lesions by completing lesion contouring in pairs composed of a radiology resident and a board-certified radiologist (see Section 3.1.4). In the PICTURE dataset however, each lesion contour was produced by a single board-certified radiologist only, due to resource constraints (see Section 3.2.4).

Variable	#
No. of patients	249
Median age (years)	62 (IQR: 58 - 67)
Median PSA (ng/ml)	6.85 (IQR: 5.07 - 9.60)
Median PSA density (ng/ml/ml)	0.18 (IQR: 0.13 - 0.28)
Per-patient Likert score (any Gleason score $\geq 3+4$)	
Likert 2	22
Likert 3	83
Likert 4	49
Likert 5	95
Per-patient maximum Gleason score	
No prostate cancer	34
Gleason score 3+3	61
Gleason score 3+4	114
Gleason score 4+3	34
Gleason score 8	5
Gleason score 9-10	1
MRI-detected lesions	
No. of patients without MRI-detected lesions	68
No. of patients with MRI-detected lesions	181
Total annotated lesions	210
Per-annotated-lesion Likert score (any Gleason score $\geq 3+4$)	
Not identified prospectively	33
Likert 2	1
Likert 3	29
Likert 4	47
Likert 5	100
Per-annotated-lesion Gleason score	
Benign	9
Gleason score 3+3	54
Gleason score 3+4	113
Gleason score 4+3	29
Gleason score 8	4
Gleason score 9-10	1
Per-annotated-lesion zone	
Peripheral zone	134
Central gland	57
Both zones	19

Table 3.3: PICTURE dataset characteristics.

Chapter 4

Patient-level classification framework for triage

The work presented in this chapter has been published by Mehta et al. [130] in the paper entitled, "Computer-aided diagnosis of prostate cancer using multiparametric MRI and clinical features: A patient-level classification framework".

4.1 Introduction

Computer-aided diagnosis (CAD) systems that use multiparametric magnetic resonance imaging (mpMRI) for prostate cancer (PCa) diagnosis are actively being investigated [15]. CAD systems that classify patients into those with and without clinically significant PCa (CSPCa) can be deployed into the PCa diagnostic pathway to perform patient triage, prior to radiologist assessment of mpMRI, or to provide a second read following the radiologist's first read. In a triage deployment, a CAD system that performs patient classification can rank patients by likelihood of CSPCa, to ensure the patients with the highest likelihood of having CSPCa are assessed first by the radiologist, providing a smarter alternative to the first-in-first-out approach to caseload management. Ranking patients by likelihood of having CSPCa may also allow ruling-out of the lowest risk patients, alleviating radiologist workload. Further motivation for triage/rule-out is provided by the potential introduction of MRI-based PCa screening programmes for men in the future, which would increase the demand on prostate radiology services quite considerably [125].

As seen in Chapter 2, the majority of PCa CAD systems, including those that are capable of producing a patient-level classification output, require lesion annotations for training. However, producing lesion annotations on mpMRI can be challenging and/or time-consuming for a number of reasons. First and foremost, producing lesions annotations on mpMRI, following prostatectomy or a MRI-blinded biopsy technique such as systematic biopsy, saturation biopsy, or transperineal template prostate-mapping (TTPM) biopsy is not clinical routine and therefore, must be performed retrospectively [112]. Second, cognitive matching of biopsy or prostatectomy findings to mpMRI may be required, which is not trivial. Third, once location on mpMRI is determined, the contour will typically be drawn on T2-weighted MRI (T2WI) (due to its high spatial resolution and superior tissue contrast), in-plane and on all other slices containing the lesion. Should registration issues arise between mpMRI modalities, lesion contours or lesion centroids may be required on the other modalities also. A further challenge is posed by diffuse non-focal tumours and MRI invisible tumours [131]; it is unclear how these tumours should be annotated. Fourth and finally, to account for inter-observer variability [129], lesion annotations should be made by more than one radiologist, which can increase the overall time taken to perform annotations multiplicatively. Due to the annotation difficulties described, CAD systems for PCa are typically trained on small, carefully prepared datasets.

In this chapter, we introduce a novel patient-level classification framework, denoted PCF, that is trained using patient-level labels only, therefore avoiding the need for lesion annotations. In PCF, feature vectors are extracted from three-dimensional T2WI, apparent diffusion coefficient (ADC) map, computed high b-value diffusion-weighted MRI (DWI), and four semi-quantitative parameter maps extracted from dynamic contrast-enhanced MRI (DCEI) using convolutional neural networks (CNNs), where each CNN is a modified 3D ResNet architecture proposed in this work. During the training phase of PCF, feature selection is applied to select the optimal subset of CNN feature vectors and available clinical features for patient classification. Subsequently, selected CNN feature vectors and clinical features are combined for classification using a two-level multi-classifier support

vector machine (SVM) scheme. The output of PCF is a patient probability associated to the presence of CSPCa in the patient's prostate; here, CSPCa is defined as Gleason score $\geq 3+4$ disease. Utilizing features extracted from the full-breadth of mpMRI and available clinical features in combination to enhance classification performance is in line with the guidance provided by the Likert assessment system for radiologists [25]. We envision PCF being applied as a triage tool or as a second reader during routine diagnosis; both applications could help alleviate the workload of radiologists who are an increasingly stretched resource [14].

This chapter is organised as follows: In Section 4.2, we describe the technical details of PCF. In Section 4.3, we describe the subsets of the PROSTATEx and PICTURE datasets, introduced in Chapter 3, that were used to evaluate the performance of PCF, the classification tasks performed, the evaluation measures used, and the experimental settings employed. Section 4.4 presents the results for patient classification. In Section 4.5, we conclude by discussing the implications of our results.

4.2 Method

PCF is visualised schematically in Figure 4.1. First, mpMRI and clinical features are pre-processed. This involves automated prostate region segmentation, calculation of high b-value DWI and semi-quantitative DCEI parameter maps, and finally, normalisation/standardisation of images, parameter maps, and clinical features. Second, CNN encoders are employed to extract feature vectors from three-dimensional MR images and parameter maps. Third, forward feature selection is used to select the CNN feature vectors and clinical features that are most pertinent for classification. Fourth and finally, a two-level SVM scheme is used to output the patient's probability of having CSPCa.

4.2.1 Pre-processing

4.2.1.1 Automated prostate region segmentation

As a first step, the prostate is segmented on T2WI. Segmentation of the prostate creates a simpler classification task, unsullied by excess background information.

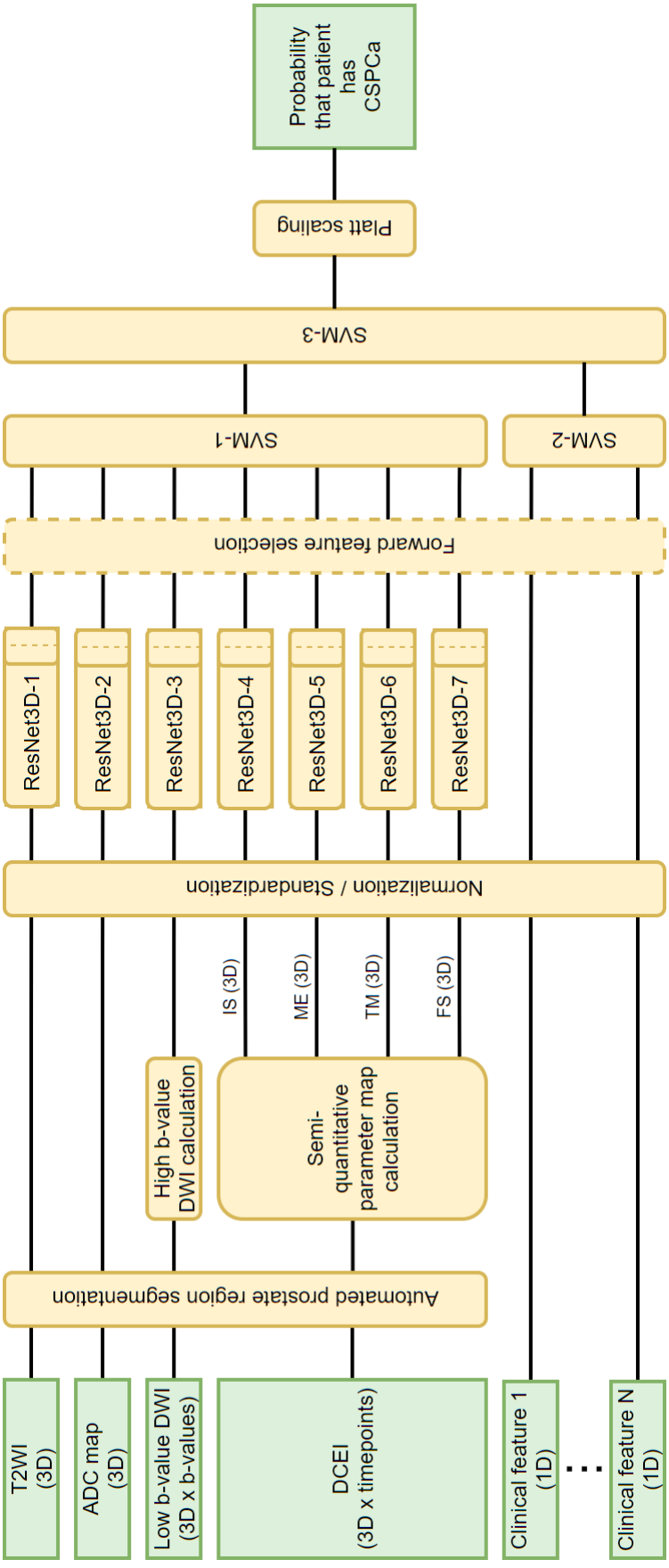


Figure 4.1: Workflow of the proposed patient-level classification framework, PCF. Green rectangles indicate original scans or clinical features or the probabilistic classification outcome; yellow rectangles indicate processes applied to the data. In experiments, we refer to PCF with forward feature selection disabled during training as PCF-ALL and PCF with forward feature selection enabled during training as PCF-SEL.

In PCF, we use HighRes3DNet [132] to segment the prostate on T2WI. HighRes3DNet is a high-resolution compact CNN for volumetric image segmentation. Given a three-dimensional T2WI, I_{T2WI}, HighRes3DNet outputs a prostate mask, S_{T2WI}, with the same spatial dimensions as I_{T2WI}. The prostate mask for DWI, S_{DWI}, is obtained by transforming S_{T2WI} from T2WI space into DWI space using a registration-driven transformation T such that $T : S_{T2WI} \rightarrow S_{DWI}$, which accounts for resolution differences between T2WI and DWI, as well as voluntary/involuntary patient movement between acquisitions, and distortions on DWI caused by air in the rectum [133]. Here, $T = T_{nrd} \circ T_{aff}$, where T_{aff} is the transformation given by the affine registration of I_{T2WI} to the three-dimensional ADC map, I_{ADC}, using the symmetric block-matching algorithm [134] and T_{nrd} is given by the subsequent non-rigid registration using the free-form deformation (FFD) algorithm [135]. The convolution-based fast local normalised correlation coefficient (LNCC) [136] is used as similarity measure for FFD to enable robustness to bias field inhomogeneity. The same approach is used to obtain the prostate mask in DCEI space, S_{DCEI}; in this case driven by the registration of T2WI to the first DCEI timepoint. The prostate masks, S_{T2WI}, S_{DWI}, and S_{DCEI} obtained for each patient are used to crop a sub-volume containing the prostate in corresponding imaging.

4.2.1.2 Computed high b-value DWI

High b-value images with b-value ≥ 1400 are a key component of mpMRI [24]. Computed high b-value DWI has been shown to achieve superior image quality and lesion conspicuity than acquired high b-value DWI [28]. In PCF, we compute high b-value DWI using a monoexponential model [29] for the per-voxel observed signal:

$$s(b) = s(0)\exp(-b \cdot ADC). \tag{4.1}$$

Non-linear least squares is used to fit Equation 4.1 to the observed points given by low b-value DWI intensities, giving per-voxel estimates of s(0) and ADC: $s^*(0)$ and ADC^*. High b-value images are then computed using the equation:

$$s(b_c) = s^*(0)\exp(-b \cdot ADC^*), \tag{4.2}$$

where b_c is the high b-value being extrapolated to.

4.2.1.3 Semi-quantitative DCEI parameter maps

Semi-quantitative analysis of DCEI has been shown to provide good discrimination between benign and malignant lesions [137], while avoiding the challenging estimation of the arterial input function needed for computing pharmacokinetic parameters. In PCF, semi-quantitative analysis of DCEI is used to compute parameter maps in an automated manner.

First, per-voxel signal intensity-versus-time curves are normalised to a standard pre-contrast level using a mean baseline computed from the first three signal values from T timepoints:

$$\hat{s}_t = \frac{s_t}{k}, \quad t = 1, \cdots, T, \quad k = \sum_{t=1}^{3} \frac{s_t}{3}. \tag{4.3}$$

Then, four voxel-wise variables are extracted: initial slope of enhancement (IS), maximum enhancement (ME), time to maximum enhancement (TM), and final slope (FS); originally defined in [137] and [138], but we are the first to use them to construct three-dimensional parameter maps and in addition, to later extract spatial features from the parameter maps using CNNs. An illustration of the variables is given in Figure 4.2.

IS is assumed to be the gradient of the steepest portion of the normalised signal intensity-versus-time curve. First, an averaging window of length l_{IS} is passed over the normalised signal $\{\hat{s}_t\}_{t=1}^{T}$ in steps of one timepoint as in [138]. The gradient of the linear best fit in each window is computed, giving the set of gradients $\{g_t\}_{t=1}^{T-l_{IS}+1}$. Subsequently, IS is computed by taking the maximum of the gradients:

$$IS = \max(\{g_t\}_{t=1}^{T-l_{IS}+1}), \tag{4.4}$$

where l_{IS} is determined empirically based on the temporal resolution of the DCEI.

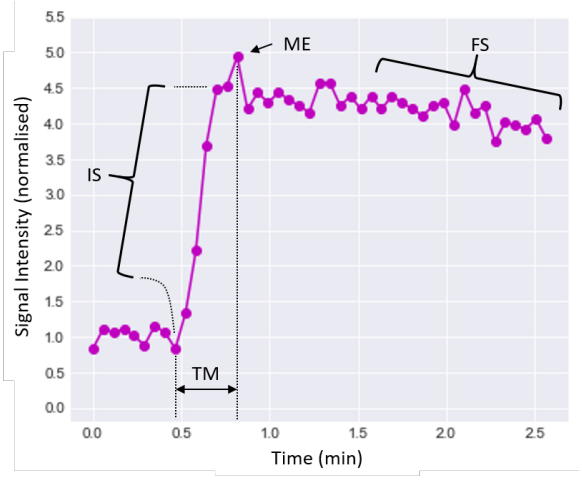


Figure 4.2: Normalised signal intensity-versus-time curve corresponding to voxel at the center of a Gleason score 3+4 lesion of a patient from the PROSTATEx dataset.

ME is calculated as the maximum value of the normalised signal:

$$\mathrm{ME} = \max(\{\hat{s}_t\}_{t=1}^{T}). \tag{4.5}$$

TM is calculated as the difference between onset time, denoted t_{OS}, and the time of ME, denoted t_{ME}, in minutes, where t_{OS} is defined as the first time point in the averaging window to which IS corresponds:

$$\mathrm{TM} = t_{ME} - t_{OS}. \tag{4.6}$$

FS is defined as the gradient of the normalised signal over the wash-out phase of the contrast agent. Here, we compute it as the the gradient g_{FS} of the linear best fit over the final m_{FS} minutes of the normalised signal, where m_{FS} is determined empirically based on the length of the wash-out phase of the contrast agent.

4.2.1.4 Normalisation/standardisation

Histogram-based standardisation is applied to the prostate region in each patient's T2WI to homogenise tissue intensities across patients in line with the work by Toivonen et al. [139]. Images are transformed by matching their intensity histograms to a mean histogram calculated using training data. The algorithm, includ-

ing pseudo-code, is presented in the work by Nyul et al. [140]. A simple per-patient z-score normalisation is then applied to each patient's standardised T2WI, computed high b-value DWI, and DCEI parameter maps in line with the work by Isensee et al. [8], who showed this to be an effective strategy for MRI as CNN input. ADC maps are not normalised since ADC is a quantitative measurement.

Each clinical feature included in PCF is standardised to have zero mean and unit standard deviation, using mean and standard deviation computed from training data.

4.2.2 Convolutional neural network feature extraction

ResNet CNN architectures have demonstrated good performance in several image classification tasks [141, 142]. In PCF, seven identical 3D ResNet CNN architectures, denoted $\{\text{ResNet3D-i}\}_{i=1}^{7}$, are employed to extract features from T2WI, ADC map, computed high b-value DWI, and each of the four DCEI parameter maps. ResNet3D is a modified 3D implementation of the standard 2D ResNet. Our implementation is composed of a convolutional layer C_1, followed by four bottleneck blocks B_1, B_2, B_3, and B_4, and a fully-connected layer FC. A network diagram is shown in Figure 4.3a. Bottleneck blocks reduce the computational load of 3D convolutional layers by performing a channel reduction and restoration operation either side of the core convolution operation as shown in Figure 4.3b. Preactivation [142] (batch normalisation and rectified linear unit activation prior to weight layer computation) is used to ease optimisation and regularise the networks. The last bottleneck block, B_4, outputs a set of feature maps to which global average pooling is applied to transform each feature map f_j into a feature value v_j.

During the training phase of PCF, each ResNet3D-i is trained end-to-end. The feature values v_j are linearly combined in the FC layer, followed by softmax to produce a classification output, followed by loss computation, backpropagation, and weight updates.

During inference, the feature values v_j, corresponding to ResNet3D-i, are grouped into a feature vector $V_i = \{v_j\}_{j=1}^{128}$. Subsequently, each feature vector V_i corresponding to each ResNet3D-i is passed to a two-level SVM scheme where the

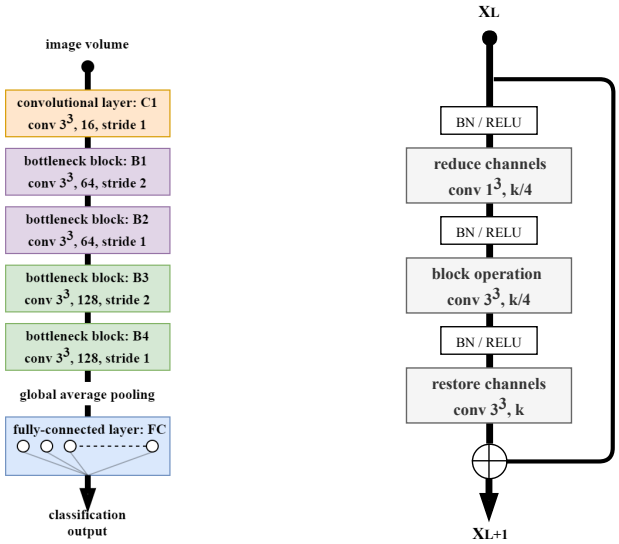


Figure 4.3: (a) Proposed ResNet3D CNN used to extract features from volumetric images. (b) A bottleneck block, where k = #kernels.

final patient classification is made.

4.2.3 Forward feature selection

The optimal subset of ResNet3D feature vectors $V = \{V_i\}_{i=1}^{7}$ and normalised clinical features $F = \{F_j\}_{j=1}^{N}$, for some quantity of clinical features N, is found during the training phase of PCF using forward feature selection (FFS) [143]. FFS is used to remove features that are acting as noise; removing noise is especially important when training classification algorithms using small datasets. In our implementation of FFS, each ResNet3D feature vector V_i is considered a feature. We denote the total feature set $ALL = V \cup F$. We begin by assuming the null set of selected features $SEL = \varnothing$. At each iteration we induct the feature into SEL which maximises an evaluation metric M computed over SEL. The FFS procedure is summarised as follows:

1. Initialise $SEL = \{\varnothing\}$;
2. For each feature $X_k \in ALL$, $k = 1, \ldots, N+7$, compute $M(X_k)$ and select the feature $\hat{X}_k$ that maximises M;
3. Remove $\hat{X}_k$ from the set ALL and add $\hat{X}_k$ to the set SEL; thus ALL :=

ALL$\backslash\{\hat{X}_k\}$ and SEL := $\{\hat{X}_k\}$;

4. Repeat until a decrease in M is observed:

   (a) For each X_k in ALL, compute M(SEL$\cup\{X_k\}$) and select the feature $\hat{X}_k$ that maximises M;

   (b) Remove $\hat{X}_k$ from the set ALL and add $\hat{X}_k$ to the set SEL; thus ALL := ALL$\backslash\{\hat{X}_k\}$ and SEL := SEL$\cup\{\hat{X}_k\}$;

4.2.4 Support vector machine classification

A two-level multi-classifier SVM scheme is used to combine the selected ResNet3D feature vectors and normalised clinical features to produce a final patient classification. Two SVMs, denoted SVM-1 and SVM-2, are included in the first layer and a third SVM, denoted SVM-3, is included in the second layer. First, SVM-1 takes the ResNet3D feature vectors $V_i \in$ SEL as input and outputs a patient classification probability $\hat{y}_1$. Concurrently, SVM-2 takes the normalised clinical features $F_j \in$ SEL as input and outputs a patient classification probability $\hat{y}_2$. Since SVMs do not naturally output a probability, Platt scaling [144] is used to obtain probability estimates. Then, SVM-3 accepts $\hat{y}_1$ and $\hat{y}_2$ as input to output a final classification probability $\hat{y}$ associated to the positive class i.e., probability that the patient has CSPCa. It should be noted that if clinical features are either not available or not selected by FFS, the final classification is made by SVM-1, and SVM-2 and SVM-3 will be omitted.

4.3 Experimental setup

In this section we describe the patient datasets used in this work, the classification tasks completed, the validation measures used to evaluate PCF, and the methodological settings employed for conducting experiments.

4.3.1 Patient data

The performance of PCF was evaluated using the "Prostate Imaging Compared to Transperineal Ultrasound-guided biopsy for significant prostate cancer Risk Eval-

uation" (PICTURE) trial dataset [127] and the the publicly available PROSTATEx dataset [44]. A detailed description of both datasets is given in Chapter 3.

The PICTURE dataset patient-level ground truth used for training and evaluating PCF was established as follows: a patient was allocated to the CSPCa class if any core sampled during transperineal template prostate-mapping (TTPM) biopsy or targeted biopsy was positive for Gleason score $\geq 3+4$. Five patient studies were removed due to one or more missing MRI sequences and 34 patient studies were removed due to severe magnetic susceptibility artifacts on DWI. Characteristics of the included patients are shown in Table 4.1.

Total patients following exclusions	210
Median age (years)	62 (58-67)
Median PSA (ng)	7 (5-10)
Median TPV (ml)	40 (28-51)
Median PSAd (ng/ml)	0.18 (0.13-0.28)
Breakdown by max Gleason score	# of patients
Normal/Benign	30
GS 3+3	50
GS 3+4	96
GS 4+3	30
GS > 8	4

Table 4.1: Characteristics of the PICTURE dataset patients used to evaluate PCF. Interquartile range shown in brackets for age, PSA, total prostate volume (TPV), and PSA density (PSAd).

The PROSTATEx dataset patient-level ground truth used for training and evaluating PCF was established as follows: a patient was allocated to the CSPCa class if the patient's prostate contained any lesion with Gleason score $\geq 3+4$. 64 patient studies were removed due to missing ground-truth labels; of these, two patients belonged to the PROSTATEx Challenges training set and 62 patients belonged to the PROSTATEx Challenges test set. Characteristics of the remaining 282 patient studies are shown in Table 4.2.

Total patients following exclusions	282
Breakdown by max Gleason score	# of patients
No CSPCa*	212
GS 3+4	38
GS 4+3	19
GS $>$ 8	13

*Either GS $\leq$ 6, benign, or PI-RADS = 2. PI-RADS = 2 lesions were not biopsied; assumed not CSPCa, CSPCa occurrence in PI-RADS = 2 lesions at Radboud Medical Center less than 5%.

Table 4.2: Characteristics of the PROSTATEx dataset patients used to evaluate PCF.

4.3.2 Experiments

PCF was trained to classify patients into those with CSPCa and those without CSPCa, where CSPCa refers to the presence of max Gleason score $\geq 3+4$ tissue, as determined through histopathological analysis. In the PICTURE dataset a total of 130 patients with CSPCa and 80 patients without CSPCa were available for analysis, while in the PROSTATEx dataset a total of 70 patients with CSPCa and 212 patients without CSPCa were available for analysis. The following experiments were conducted:

- Intra-dataset evaluation: The following classifiers were trained: (i) ResNet3D with individual MRI modalities or parameter maps (ResNet3D-x, where x $\in$ X = $\{$T2WI, ADC, Cb2000, IS, ME, TM, FS$\}$); (ii) SVM with the individual clinical features prostate-specific antigen (PSA), total prostate volume (TPV), and PSA density (PSAd) (SVM-y, where y $\in$ Y = $\{$PSA, TPV, PSAd$\}$); (iii) PCF with the set of available MRI modalities and parameter maps (PCF-ALL-MR); (iv) PCF with the set of MRI modalities and parameter maps selected by FFS (PCF-SEL-MR); (v) PCF with the set of available MRI modalities, parameter maps, and clinical features (PCF-ALL); and (vi) PCF with the set of MRI modalities, parameter maps. and clinical features selected by FFS (PCF-SEL). The performance of classifiers was evaluated using a five-fold cross-validation on the PICTURE and PROSTATEx datasets separately. The mean receiver operating characteristic (ROC) curve, precision-recall (PR) curve,

and respective areas under the curve (AUC) were calculated in each instance. The Wilcoxon-signed rank test for pairwise comparison [145] was applied to statistically validate the comparison between different classifiers.

- Inter-dataset evaluation: ResNet3D classifiers, PCF-ALL-MR classifiers, and PCF-SEL-MR classifiers, obtained from the PICTURE dataset intra-dataset five-fold cross-validation were used to perform inference on the PROSTATEx dataset and vice versa. The mean and standard deviation of the ROC and PR AUCs of the five cross-validation models is presented.

- Clinical evaluation: The PICTURE dataset alone was used for clinical evaluation as radiologist PI-RADS v1 scores associated to the PROSTATEx dataset have not been released publicly. The PICTURE dataset was divided temporally into 170 patients for training (scan dates: 11/01/2012 to 25/06/2013) and 40 patients for testing (scan dates: 26/06/2013 to 29/01/2014). The test set comprised of 20 patients with CSPCa and 20 patients without CSPCa. PROSTATEx scans were used to augment the training set. PCF-SEL was used in the clinical evaluation. The probabilistic output of PCF-SEL was binarised by selecting a cutoff that matched the sensitivity of the radiologist on the PICTURE training set at two cutoffs: Likert ≥ 3 and Likert ≥ 4. The sensitivity, specificity, precision, and negative predictive value (NPV) were computed; 95% confidence intervals (CI) were calculated using bootstrapping. McNemar's test [146] was used to statistically compare the sensitivity and specificity of the radiologist and PCF-SEL, while the weighted generalised score (WGS) test statistic [147] was used to compare the precision and NPV of the radiologist and PCF-SEL.

4.3.3 Experimental settings

In this section we describe the methodological settings used for conducting experiments with PCF.

4.3.3.1 Pre-processing settings

HighRes3DNet was trained using the T2WI of 82 patients from the PICTURE dataset for which manual contours of the whole prostate were available, and 50 training cases from the publicly available PROMISE12 dataset [91]. All images were whitened and resampled to isotropic 1mm resolution as preprocessing, and resampled to original voxel resolution as post-processing. During training subvolumes of size 64^3 were sampled to maintain a 50:50 ratio of foreground to background voxels. Flip and rotation augmentations were applied on-the-fly. Training was conducted using Dice loss [74], Adam optimisation [148], learning rate equal to 0.001, and batch size 4. The network was trained until we observed a plateau in performance on the validation set. The trained network was used to segment the remainder of the PICTURE dataset and the entirety of the PROSTATEx dataset. A mean Dice score of 0.90 was achieved on a ten-fold cross-validation of the 82 PICTURE dataset patients.

Registration of T2WI to ADC map and first timepoint of DCEI, used to obtain the transformation of prostate masks into DWI and DCEI space, used default parameters for affine registration via symmetric block-matching [134]. The subsequent non-rigid FFD registration used a Gaussian kernel with standard deviation equal to 5mm for LNCC calculation, control point spacing equal to 10mm, and bending energy constraint equal to 0.1.

A high b-value, $b_c = 2000$, was selected for computing high b-value DWI as in Verma et al. [28]. We refer to computed high b-value DWI with a b-value of 2000 as computed b2000 DWI (Cb2000).

The DCEI parameter IS was calculated using an averaging window of length $l_{IS} = 3$ for the PICTURE dataset and $l_{IS} = 5$ for the PROSTATEx dataset. DCEI parameter FS was calculated over the final $m_{FS} = 2$ minutes of the normalised signal for the PICTURE dataset and $m_{FS} = 1$ minutes of the normalised signal for the PROSTATEx dataset.

As recommended in Nyul et al. [140], deciles were used as landmarks for histogram standardisation of T2WI.

4.3.3.2 Training settings

For each experiment, training data was further subdivided 80:20 into training and validation sets. The training set was used for training constituent ResNet3D and SVM classifiers, while the validation set was used for selecting feature vectors and normalised clinical features during FFS.

All images were resized to a common size of $65 \times 65 \times 45$ prior to ResNet3D training. Each ResNet3D in PCF was trained using cross-entropy loss, Adam optimisation, learning rate equal to 0.00001, and batch size 8. In-plane flip and random deformation augmentations were applied to the training set to balance classes and reduce overfitting.

The following metric M is proposed for observation during FFS:

$$\mathrm{M} = \frac{\mathrm{ROC\,AUC} + \mathrm{PR\,AUC}}{2}, \tag{4.7}$$

as it maximises both model evaluation metrics of interest.

A radial basis kernel was used in SVM-1, SVM-2, and SVM-3 as there existed no reason to assume linear separability of data. The misclassification penalty was set to C = 0.1 for SVM-1 and SVM-2, and C = 1 for SVM-3, in all experiments.

4.4 Results

In this section we present the results obtained from the intra-dataset and inter-dataset evaluations of PCF, as well as the clinical evaluation of PCF using a temporally separated patient cohort from the PICTURE dataset.

4.4.1 Intra-dataset model evaluation

The mean ROC and PR AUCs averaged over five-fold cross-validation for ResNet3D, SVM, and PCF classifiers are shown in Table 4.3 for both the PICTURE and PROSTATEx datasets. Figure 4.4a, 4.4b, 4.4c, and 4.4d show the mean ROC and PR curves calculated for PCF-SEL for both datasets. Reliability diagrams for PCF-SEL are shown in Figure 4.4e for both datasets.

For the PICTURE dataset, ResNet3D-ADC had the best performance among

	PICTURE dataset		PROSTATEx dataset	
Classifier	**Mean ROC AUC**	**Mean PR AUC**	**Mean ROC AUC**	**Mean PR AUC**
ResNet3D-T2WI	0.70 ± 0.06	0.79 ± 0.05	0.78 ± 0.07	0.60 ± 0.08
ResNet3D-ADC	0.74 ± 0.09	0.83 ± 0.08	0.80 ± 0.05	0.64 ± 0.04
ResNet3D-Cb2000	0.67 ± 0.10	0.79 ± 0.08	0.82 ± 0.05	0.66 ± 0.07
ResNet3D-IS	0.65 ± 0.03	0.75 ± 0.04	0.70 ± 0.08	0.47 ± 0.07
ResNet3D-ME	0.67 ± 0.06	0.76 ± 0.06	0.79 ± 0.05	0.55 ± 0.07
ResNet3D-TM	0.68 ± 0.13	0.77 ± 0.09	0.70 ± 0.08	0.44 ± 0.11
ResNet3D-FS	0.65 ± 0.08	0.75 ± 0.05	0.72 ± 0.03	0.44 ± 0.03
PCF-ALL-MR	0.72 ± 0.09	0.80 ± 0.07	0.82 ± 0.04	0.63 ± 0.05
PCF-SEL-MR	0.77 ± 0.11	0.84 ± 0.09	$\mathbf{0.86 \pm 0.04}$	$\mathbf{0.72 \pm 0.03}$
SVM-PSA	0.54 ± 0.06	0.64 ± 0.03	n/a	n/a
SVM-TPV	0.70 ± 0.12	0.80 ± 0.10	n/a	n/a
SVM-PSAd	0.73 ± 0.07	0.82 ± 0.06	n/a	n/a
PCF-ALL	0.74 ± 0.10	0.81 ± 0.09	0.82 ± 0.04	0.63 ± 0.05
PCF-SEL	$\mathbf{0.79 \pm 0.09}$	$\mathbf{0.86 \pm 0.07}$	$\mathbf{0.86 \pm 0.04}$	$\mathbf{0.72 \pm 0.03}$

Table 4.3: Intra-dataset evaluation of ResNet3D, SVM, and PCF classifiers. Mean ROC AUC and PR AUC $\pm$ one standard deviation averaged over five-fold cross validation, for the PICTURE and PROSTATEx datasets, shown. Highest value in each column shown in bold.

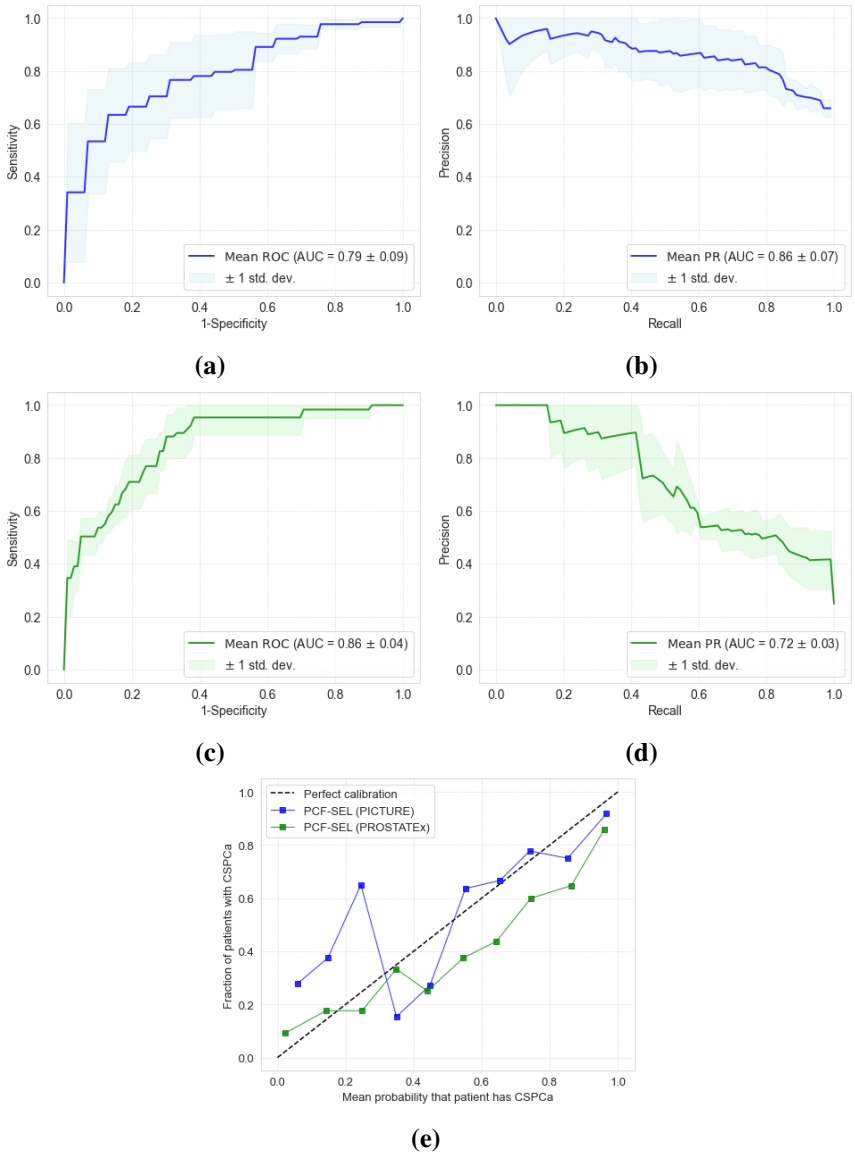


Figure 4.4: Intra-dataset evaluation of PCF-SEL. Graphs (a, b) show the mean ROC and PR curves, averaged over five-fold cross validation, for the PICTURE dataset, while graphs (c, d) correspond to the PROSTATEx dataset. Reliability diagrams for PCF-SEL are shown in (e), for both datasets.

ResNet3D and SVM classifiers that were trained using a single MRI modality, parameter map, or clinical feature. PCF-ALL did not improve the result. However, PCF-SEL did improve upon the result of ResNet3D-ADC, with an increase in ROC AUC from 0.74 to 0.79 ($p < 0.05$) and an increase in PR AUC from 0.83 to 0.86 ($p = 0.08$); during the five-fold cross-validation of PCF-SEL, FFS selected ADC map, PSAd, Cb2000 DWI, and TM map in the majority of fold experiments run.

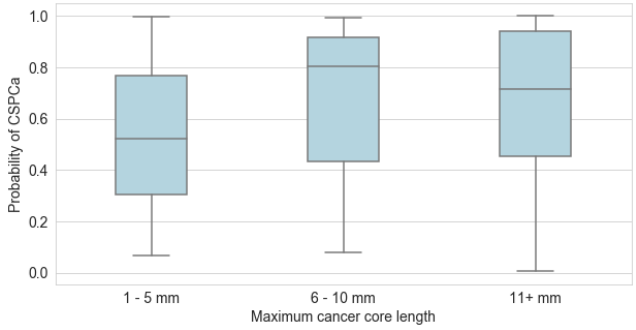
For the PROSTATEx dataset, ResNet3D-Cb2000 had the best performance among ResNet3D classifiers that were trained using a single MRI modality of parameter map. PCF-ALL did not improve the result. However, PCF-SEL did improve upon the result of ResNet3D-Cb2000, with an increase in ROC AUC from 0.82 to 0.86 ($p = 0.07$) and an increase in PR AUC from 0.66 to 0.72 ($p = 0.07$); during the five-fold cross-validation of PCF-SEL, FFS selected Cb2000 DWI, ADC map, and ME map in the majority of fold experiments run.

In addition to the ability to discriminate between classes, it is desirable for models to produce well-calibrated probability estimates. For output probability $\hat{P}$, perfect calibration is defined as:

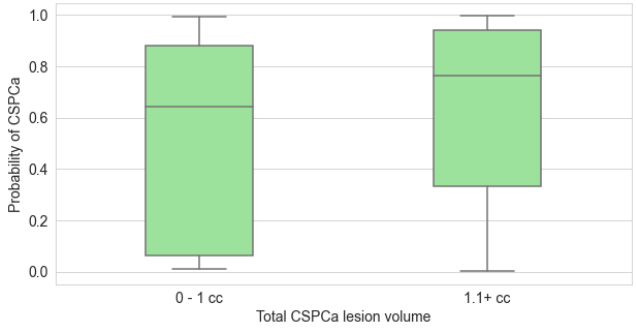
$$P(\text{CSPCa} \mid \hat{P} = p) = p, \quad \forall p \in [0, 1], \tag{4.8}$$

i.e., $\hat{P}$ should represent a true probability [149]. Figure 4.4e shows reliability diagrams for PICTURE and PROSTATEx dataset patient probabilities output by PCF-SEL. Perfect calibration is represented by the identity diagonal. As observed for both datasets, the identity diagonal is broadly tracked indicating reasonable calibration. For the PICTURE dataset we observe better calibration at the higher probability end, while for the PROSTATEx dataset we observe better calibration at the lower probability end. This may be explained by the higher prevalence of patients with CSPCa in the PICTURE dataset and the higher prevalence of patients with benign conditions or low-grade PCa in the PROSTATEx dataset.

An additional analysis was completed to investigate the relationship between the predicted probability of CSPCa and lesion volume. The distribution of patient probabilities output by PCF-SEL by maximum cancer core length (MCCL) (PICTURE dataset) and lesion volume (PROSTATEx dataset) is shown in Figure 4.5; lesion volume could not be computed for the PICTURE dataset, therefore MCCL is used as a surrogate measure. As may be expected, for both datasets, we observe a lower median probability of CSPCa and a higher variability, for patients with a low MCCL/total lesion volume.



(a)



(b)

Figure 4.5: Intra-dataset evaluation of PCF-SEL. Figure (a) shows the probability of CSPCa for PICTURE dataset patients with biopsy-proven CSPCa, separated by maximum cancer core length (MCCL). Figure (b) shows the probability of CSPCa for PROSTATEx dataset patients with biopsy-proven CSPCa, separated by total CSPCa lesion volume. In both boxplots, the limits represent the minimum and maximum values, while the middle line represents the median.

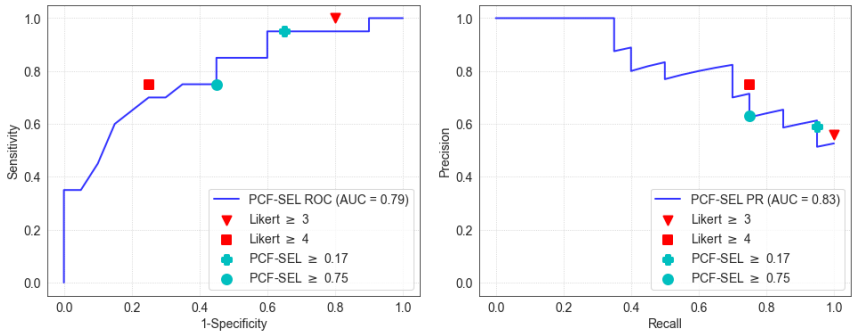
4.4.2 Inter-dataset model evaluation

ResNet3D classifiers, PCF-ALL-MR classifiers, and PCF-SEL-MR classifiers, obtained from the PICTURE dataset intra-dataset five-fold cross-validation, were used to perform inference on the PROSTATEx dataset and vice versa. The mean and standard deviation of the ROC and PR AUCs of the five cross-validation models is presented in Table 4.4. Clinical features were not considered since they were not available for the PROSTATEx dataset.

ResNet3D-ADC trained using the PROSTATEx dataset and applied to the PICTURE dataset maintained a similar performance level to ResNet3D-ADC trained with the PICTURE dataset. Similarly, ResNet3D-Cb2000 trained using the PICTURE dataset and applied to the PROSTATEx dataset maintained a similar performance level to ResNet3D-Cb2000 trained with the PROSTATEx dataset. For both

Classifier	PROSTATEx train, PICTURE test		PICTURE train, PROSTATEx test	
	Mean ROC AUC	**Mean PR AUC**	**Mean ROC AUC**	**Mean PR AUC**
ResNet3D-T2WI	0.70 ± 0.01	0.79 ± 0.01	0.75 ± 0.01	0.51 ± 0.03
ResNet3D-ADC	$\mathbf{0.73 \pm 0.03}$	$\mathbf{0.82 \pm 0.02}$	0.73 ± 0.02	0.47 ± 0.01
ResNet3D-Cb2000	0.68 ± 0.00	0.79 ± 0.00	$\mathbf{0.81 \pm 0.02}$	$\mathbf{0.65 \pm 0.03}$
ResNet3D-IS	0.63 ± 0.01	0.72 ± 0.01	0.51 ± 0.08	0.30 ± 0.06
ResNet3D-ME	0.62 ± 0.02	0.72 ± 0.01	0.59 ± 0.06	0.33 ± 0.05
ResNet3D-TM	0.65 ± 0.03	0.74 ± 0.02	0.58 ± 0.07	0.33 ± 0.06
ResNet3D-FS	0.60 ± 0.04	0.69 ± 0.02	0.62 ± 0.04	0.33 ± 0.03
PCF-ALL-MR	$\mathbf{0.73 \pm 0.01}$	0.80 ± 0.01	0.75 ± 0.03	0.50 ± 0.04
PCF-SEL-MR	0.72 ± 0.03	0.81 ± 0.03	0.77 ± 0.07	0.56 ± 0.11

Table 4.4: Inter-dataset evaluation. Mean ROC AUC and PR AUC $\pm$ one standard deviation for ResNet3D and PCF classifiers obtained from the intra-dataset evaluation and subsequently applied to the dataset that was not used to train the classifier, shown. Highest value in each column shown in bold.



(a) **(b)**

Figure 4.6: Graph (a) shows the ROC curve and graph (b) shows the PR curve for PCF-SEL on the temporally separated PICTURE dataset test cohort. Radiologist performance at Likert cutoffs (≥ 3 and ≥ 4) and PCF-SEL performance at probability cutoffs (≥ 0.17 and ≥ 0.75) are shown, where probability cutoffs for PCF-SEL were selected using the PICTURE dataset training cohort.

datasets we observed a decrease in the performance of ResNet3D classifiers trained using DCEI parameter maps likely due to the differences in temporal resolution of the DCEI between the PICTURE and PROSTATEx datasets (13s vs. 3.5s). Notably, for both datasets we observed a drop in the performance of PCF-SEL-MR as compared to its performance in the intra-dataset evaluation, primarily as the datasets do not share the same optimal modalities and due to the reduction in performance of constituent ResNet3D classifiers trained using DCEI parameter maps.

4.4.3 Clinical evaluation

In this section we present the results of the clinical evaluation of PCF-SEL. To simulate prospective use, we temporally split the PICTURE dataset into 170 patients for training and 40 patients for testing (20 patients with CSPCa and 20 patients without CSPCa). The performance of PCF-SEL was compared to the performance of an experienced radiologist (10 years of experience in reading prostate mpMRI) who assigned a Likert score to each patient. To enable calculation of sensitivity, specificity, precision, and NPV for PCF-SEL, the probabilistic output of PCF-SEL was thresholded to match the sensitivity of the radiologist on the training set. The results of the clinical evaluation are shown in Table 4.5. Figure 4.6 shows the training and test set ROC and PR curves calculated for PCF-SEL, as well as the performance of the radiologist and PCF-SEL at two operating thresholds.

Metric	Value	95% CI	Value	95% CI	P-val	Value	95% CI	Value	95% CI	P-val
	Training set (n = 170)					Test set (n = 40)				
Threshold 1	Likert ≥ 3		PCF-SEL ≥ 0.17			Likert ≥ 3		PCF-SEL ≥ 0.17		
Sensitivity / Recall (%)	95 (104/110)	90-98	95 (104/110)	90-98	1.00	100 (20/20)	100-100	95 (19/20)	83-100	1.00
Specificity (%)	33 (20/60)	22-45	65 (39/60)	53-77	< 0.01	20 (4/20)	5-39	35 (7/20)	14-57	0.51
Precision / PPV (%)	72 (104/144)	65-79	83 (104/125)	76-90	< 0.01	56 (20/36)	39-71	59 (19/32)	42-76	0.47
NPV (%)	77 (20/26)	59-92	87 (39/45)	76-96	0.27	100 (4/4)	100-100	87 (7/8)	60-100	0.46
Threshold 2	Likert ≥ 4		PCF-SEL ≥ 0.75			Likert ≥ 4		PCF-SEL ≥ 0.75		
Sensitivity / Recall (%)	69 (76/110)	60-78	69 (76/110)	60-78	1.00	75 (15/20)	55-94	75 (15/20)	55-94	1.00
Specificity (%)	77 (46/60)	66-87	87 (52/60)	78-95	0.21	75 (15/20)	55-93	55 (11/20)	33-77	0.23
Precision / PPV (%)	84 (76/90)	77-92	90 (76/84)	84-96	0.14	75 (15/20)	55-93	63 (15/24)	43-82	0.34
NPV (%)	58 (46/80)	47-68	60 (52/86)	50-71	0.54	75 (15/20)	55-94	69 (11/16)	44-91	0.57

Table 4.5: Clinical comparison of the patient-level diagnostic performance of radiologist Likert scoring and PCF-SEL, on temporally separated training and test cohorts from the PICTURE dataset.

FFS selected SEL = {T2WI, ADC map, Cb2000 DWI, PSAd}. Using the training cohort a probability threshold equal to 0.17 was selected for PCF-SEL to match the sensitivity of the radiologist at Likert threshold ≥ 3, while a probability threshold equal to 0.75 was selected to match the sensitivity of the radiologist at Likert threshold ≥ 4. On the test cohort, PCF-SEL achieved sensitivities of 95% and 75%, compared to the radiologist who achieved sensitivities of 100% and 75% and PCF-SEL achieved specificities of 35% and 55%, compared to the radiologist who achieved specificities of 20% and 75%.

While differences in specificity can be observed in favour of PCF-SEL at the higher sensitivity setting and in favour of the radiologist at the lower sensitivity setting, McNemar's test did not find statistically significant differences between PCF-SEL and the radiologist on the test cohort.

4.5 Discussion

In this work we proposed a patient-level classification framework, denoted PCF, that uses volumetric mpMRI, derived parameter maps, and clinical features, jointly, to classify patients into those with and without CSPCa. PCF is trained using patient-level labels only, thus avoiding the need for lesion annotations, which can be challenging and time-consuming to obtain. The performance of PCF was evaluated using the PICTURE and PROSTATEx datasets. We performed an intra-dataset five-fold cross-validation, an inter-dataset generalisation experiment, and a clinical evaluation of PCF on a temporally separated patient cohort from the PICTURE dataset.

In the intra-dataset five-fold cross-validation, the performance of PCF with feature selection enabled (PCF-SEL) was compared to the performance of PCF with feature selection disabled (PCF-ALL), to assess whether feature selection in PCF has a performance benefit. Further comparison was made to ResNet3D and SVM classifiers trained using individual MRI modalities, parameter maps, and clinical features. On both the PICTURE and PROSTATEx datasets, PCF-SEL outperformed all other classifiers. On the PICTURE dataset, PCF-SEL achieved a mean ROC AUC of 0.79 and mean PR AUC of 0.86; ADC map, PSAd, Cb2000 DWI and

TM map were selected for inference in at least three out of five folds during the five-fold cross-validation. On the PROSTATEx dataset, PCF-SEL achieved a mean ROC AUC of 0.86 and mean PR AUC of 0.72; for this dataset, Cb2000 DWI, ADC map, and ME map were selected for inference in at least three out of five folds during the five-fold cross-validation. Three observations are made based on the results of the intra-dataset evaluation. First, we observe that the inclusion of feature selection during the training stage of PCF yields a performance benefit, as shown by the superior performance of PCF-SEL as compared to PCF-ALL. The feature selection step improves generalizability to unseen data by removing MRI modalities, parameter maps, and clinical features that are acting as noise; removing sources of noise is especially important when training classification algorithms with small datasets which are common in PCa CAD works primarily due to the need for a consistent and accurate reference standard. Second, we observe that PCF-SEL successfully uses clinical features alongside MRI to improve patient classification performance. Our method uses a stacked ensemble of SVMs, where MRI features and clinical features are processed by separate dedicated SVMs, whose outputs are combined by a third SVM, to produce to final patient classification. Using both clinical features and MRI features for improved classification performance is in line with works by Antonelli et al. [85] and Woznicki et al. [110] who showed the utility of PSAd in lesion classification tasks. Third, we observed a performance benefit from using DCEI parameter maps. The semi-quantitative DCEI parameters calculated in this work avoid the challenging estimation of the arterial input function needed for computing pharmacokinetic parameters [150]. However, prior to clinical adoption it would be important to consider whether the gain in performance from using DCEI justifies the additional costs and risks associated to gadolinium injection. The costs of DCEI include the cost of injection, the administering nurses time, and increased scanner-time, while the risks include minor side effects such as injection site pain, nausea, headaches, and dizziness, and rare side effects include gadolinium toxicity and nephrogenic systemic fibrosis. For the reasons mentioned, and systematic review evidence suggesting a lack of performance improvement from using DCE

[30], there is an increasing sentiment towards the omittance of DCEI from prostate cancer imaging protocols.

In the intra-dataset evaluation we considered the ability of PCF to generalise to unseen patient data from the same distribution as the training patient data. However, it is also of interest to consider the ability of CAD systems to generalise to external patient cohorts, since this type of generalisability, if observed, would allow for wider deployment of a trained system. However, our inter-dataset evaluation revealed a generalisation gap. More precisely, for the PICTURE dataset we observed a drop in the performance of PCF-SEL-MR as compared to its performance in the intra-dataset evaluation, from a ROC AUC of 0.77 to 0.72. As the feature selection step uses validation data from the same distribution as the training data, it does not guarantee selection of the optimal modalities in the external dataset. However, a small increase in ROC AUC was observed for PCF-ALL-MR from 0.72 to 0.73. On the PROSTATEx dataset, both PCF-SEL-MR and PCF-ALL-MR had diminished performance, again as the datasets do not share the same optimal modalities and additionally due to the reduction in the performance of constituent ResNet3D classifiers. Our findings suggest that training CAD systems with data from the institution in which deployment is intended is the optimal strategy and should be sought where possible.

It is important to clinically evaluate prostate CAD systems. Central to this is the need to compare CAD system performance to the performance of radiologists who are the current clinical standard. Moreover, clinical evaluations should consider how CAD systems may perform prospectively which can be simulated using a temporally separated patient cohort or an external patient cohort. Furthermore, an effective clinical evaluation requires the probabilistic output of the CAD system to be thresholded, allowing measures such as sensitivity, specificity, precision, and NPV to be reported as opposed to ROC AUC or PR AUC, which allow model comparison, but are less useful measures clinically. We compared the performance of PCF-SEL to the performance of a radiologist with 10 years of experience in reading prostate mpMRI, who gave a Likert score to each patient's prostate indicating

the likelihood of CSPCa. On a temporally separated cohort of 40 patients from the PICTURE dataset, the radiologist achieved a sensitivity of 100% and a specificity of 20% at Likert threshold ≥ 3, while PCF-SEL achieved a sensitivity of 95% and a specificity of 35% at a probability threshold equal to 0.17. At Likert threshold ≥ 4, the radiologist achieved a sensitivity of 75% and a specificity of 75%, whereas PCF-SEL achieved a sensitivity of 75% and specificity of 55% at a probability threshold equal to 0.75. The differences in performance between the radiologist and PCF-SEL were not found to be statistically significant, providing initial evidence for PCF-SEL to be evaluated in a larger clinical trial.

There were four main limitations in our study. Firstly, our training data was limited. In the temporal validation of PCF, ResNets were trained using 452 patients (170 patients from the PICTURE dataset and 282 patients from the PROSTATEx dataset), while the SVM trained with PSAd input was trained with 170 patients from the PICTURE dataset only, as PSA was not available for the PROSTATEx dataset. Secondly, due to the lack of PSA availability for the PROSTATEx dataset, we were not able to include PSA or PSAd in the inter-dataset model evaluation. Third, cases with severe magnetic susceptibility artifact on DWI were removed. Finally, PCF was not evaluated in any specific clinical setting e.g., as a triage system or as a second reader.

Chapter 5

Towards automated reporting: AutoProstate

The work presented in this chapter has been published by Mehta et al. [151] in the paper entitled, "AutoProstate: Towards Automated Reporting of Prostate MRI for Prostate Cancer Assessment using Deep Learning".

5.1 Introduction

Radiologists use prostate multiparametric magnetic resonance imaging (mpMRI) to detect, score, and stage lesions that may correspond to clinically significant prostate cancer (CSPCa), whose status can later be confirmed using MR-guided targeted biopsy and histopathological grading [26]. However, the current diagnostic approach must be improved to reduce the small proportion of men with CSPCa who are missed by mpMRI, to reduce the large number of men who undergo unnecessary biopsies, and to increase the inter-observer agreement between readers [39]. In addition to lesion assessment, radiologists use prostate mpMRI to estimate prostate volume using the ellipsoid formula [152]. Primarily, prostate volume is needed for calculating prostate-specific antigen (PSA) density (PSAd), which has been shown to be a useful predictor of CSPCa [153]. However, the ellipsoid formula is an approximation which ignores exact prostate morphology [152], therefore more accurate volume estimation methods are sought. Computer-aided diagnosis (CAD) systems that use mpMRI for prostate volume estimation and CSPCa lesion detec-

tion and/or segmentation may provide the desired performance improvements over current clinical practice.

CAD systems for lesion detection and segmentation are actively being investigated, as demonstrated by a vast and growing literature [112, 113, 114, 115, 118, 120, 121]. In addition, several automatic segmentation algorithms have been proposed for whole prostate segmentation, driven largely by the PROMISE12 Challenge [91]; an accurate automatic segmentation of the whole prostate can enable a more accurate prostate volume and PSAd calculation. However, the CAD system studies to-date have not considered prostate volume and PSAd estimation, alongside CSPCa lesion detection and segmentation, in a combined system for PCa assessment. Furthermore, CAD system studies to-date have not considered automatic diagnostic report generation, which can improve reporting quality beyond the text findings currently recorded by radiologists [43], as well as providing timely information at the point of diagnosis to improve the diagnostic accuracy of radiologists. In addition, there are a lack of studies which present a robust external validation of their presented CAD systems [15], particularly using data acquired from scanners that were not used to acquire the training data.

The primary aim of this chapter is to introduce AutoProstate: a deep learning-powered framework for automated MRI-based prostate cancer (PCa) assessment and reporting. In particular, AutoProstate segments the prostatic zones on T2-weighted MRI (T2WI), detects and segments CSPCa lesions using biparametric MRI (bpMRI), and generates a novel automatic web-based report containing four sections: *Patient Details*, *Prostate Size and PSA Density*, *Clinically Significant Lesion Candidates*, and *Findings Summary*, which posits it close to clinical deployment. Notably, AutoProstate uses up-to-date deep learning techniques for training and inference, such as hybrid losses [154], test-time dropout [155], test-time augmentation [156], and model ensembling, to enhance performance. The second aim of the work presented in this chapter is to perform a high-quality single-center external validation of AutoProstate, as a first step towards clinical deployment, ahead of multicenter external validation and prospective validation in a clinical setting.

In our experiment, AutoProstate is trained using the publicly available PROSTATEx dataset [44], and externally validated using the "Prostate Imaging Compared to Transperineal Ultrasound-guided biopsy for significant prostate cancer Risk Evaluation" (PICTURE) trial dataset [33]. The external validation follows the key considerations for authors, reviewers, and readers of AI Manuscripts in radiology by Bluemke et al. [126]. In particular, the external test set contains MRI acquired using scanners manufactured by a different vendor to the scanners used to acquire the training set and is confirmed using transperineal template prostate-mapping (TTPM) biopsy, which avoids the biases associated to MR-guided targeted biopsy and prostatectomy [33]. Furthermore, we compare the performance of AutoProstate to the performance of an experienced radiologist who at the time of the PICTURE trial had 10 years' experience of reading prostate mpMRI; since radiological assessment of prostate mpMRI is current clinical practice, preliminary evidence of CAD system efficacy can be gained through comparing CAD system performance to radiologist performance.

This chapter is organised as follows: in Section 5.2, we describe the technical details of AutoProstate; in Section 5.3, we explain how the PROSTATEx and PICTURE datasets were used to train and evaluate AutoProstate, the experimental settings employed, the experiments performed, and the evaluation measures used; in Section 5.4, we present the results of the external validation; and in Section 5.5, we conclude by discussing the implications of our results.

5.2 Method

AutoProstate, visualised schematically in Figure 5.1, consists of three modules: Zone-Segmenter, CSPCa-Segmenter, and Report-Generator. Methodological aspects of each module are described in detail in the subsections to follow, while specific experimental parameters used to collect results are described in Section 5.3.

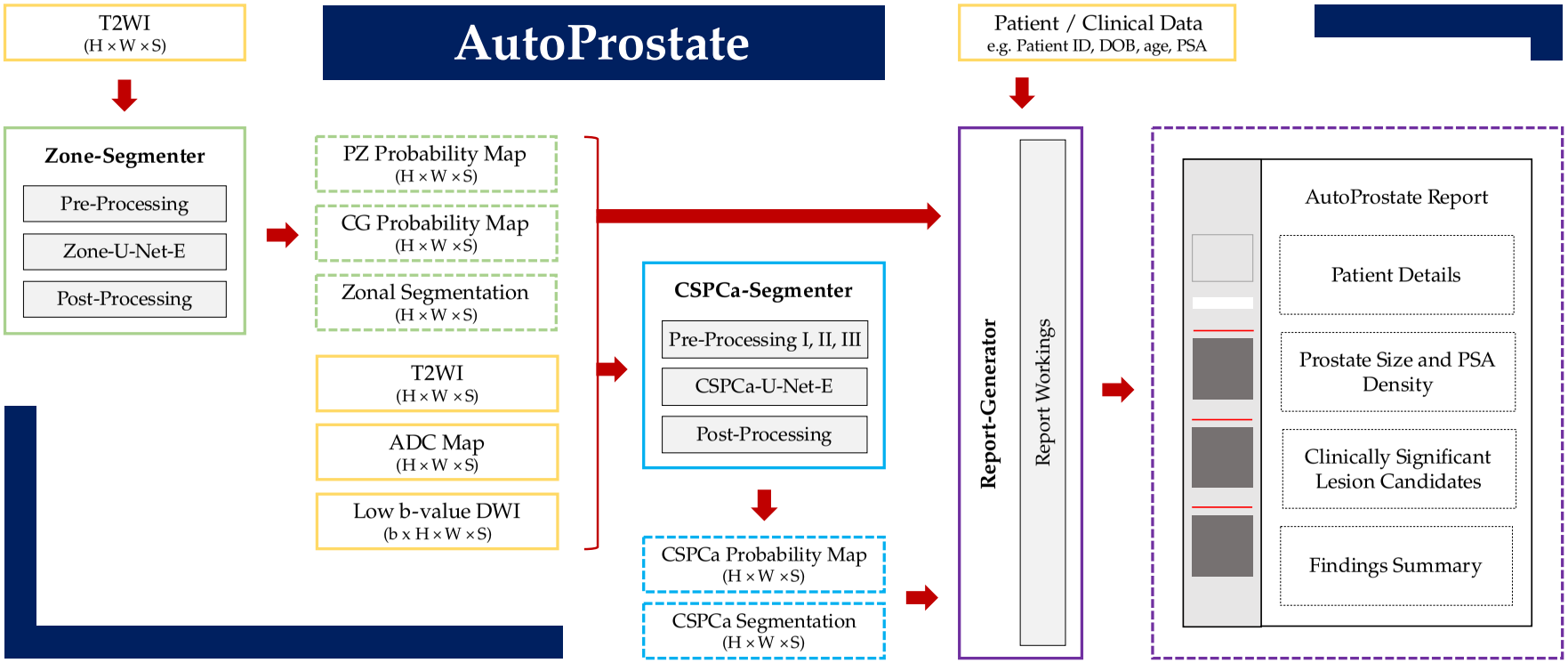


Figure 5.1: AutoProstate framework diagram. AutoProstate consists of three modules: Zone-Segmenter (green), CSPCa-Segmenter (blue), and Report-Generator (purple); solid boxes correspond to module computations, while dashed boxes correspond to module outputs. Yellow boxes indicate AutoProstate inputs from external sources.

5.2.1 Zone-Segmenter module

The Zone-Segmenter module segments peripheral zone (PZ), central gland (CG), and background tissues on T2WI.

5.2.1.1 Pre-processing

T2W images are first resampled to a common in-plane resolution and cropped to a common in-plane shape, and then normalised by whitening of image voxel intensities.

5.2.1.2 Zone-U-Net-E

After pre-processing, each T2WI slice is segmented by an ensemble of 2D U-Nets [6] with hyperparameters taken from the work by Isensee et al. [8] on the nnU-Net framework, modified for the task of zone segmentation where required; we refer to each constituent 2D U-Net as Zone-U-Net and the ensemble of Zone-U-Nets as Zone-U-Net-E. Specifically, each Zone-U-Net features six encoding blocks and five decoding blocks. Each encoding block consists of two convolutional layers with stride one 3×3 convolutions with zero padding, leaky rectified linear unit (LReLU) activation (neg. slope 1e-2) and instance normalisation [157], followed by a stride two 2×2 max pooling operation in the first five encoding blocks. Thirty-two feature maps are output by convolutional layers in the first encoding block, with feature maps doubling in each subsequent encoding block. Upsampling deconvolution operations are used in the decoding blocks, which receive semantic information from the last encoding block and higher resolution feature maps from encoder-to-decoder skip connections. The output of each Zone-U-Net is slice-wise PZ, CG, and background probability maps. Per-voxel averaging is used to combine the probability map outputs of each Zone-U-Net $\in$ Zone-U-Net-E, followed by restacking of slices to form PZ, CG, and background probability map volumes.

5.2.1.3 Post-processing

The PZ, CG, and background probability maps output by Zone-U-Net-E are transformed to the original T2WI size and voxel resolution using padding and resampling operations. As a final step, a zonal segmentation map is obtained from PZ, CG, and

background probability maps using a per-voxel argmax operation.

5.2.2 CSPCa-Segmenter module

The CSPCa-Segmenter module detects and segments CSPCa lesions using each patient's T2WI, apparent diffusion coefficient (ADC) map, low b-value diffusion-weighted MRI (DWI), and PZ and CG probability maps output by Zone-Segmenter.

5.2.2.1 Pre-processing I: computed high b-value DWI

AutoProstate generates computed high b-value DWI from available DWI corresponding to low b-values (typically $b \in [0, 1000]$ s/mm^2 [28]) using a monoexponential model for the per-voxel observed signal [29]:

$$s(b_c) = s^*(0) \cdot \exp(-b \cdot ADC^*), \tag{5.1}$$

where b_c is the high b-value being extrapolated to.

5.2.2.2 Pre-processing II: registration

Image registration is used to align ADC maps and computed high b-value DWI to T2WI to account for voluntary/involuntary patient movement between T2WI and DWI acquisitions and differences in resolution. First, ADC maps are affinely registered to T2WI using the symmetric block matching algorithm [158]. Next, a non-rigid registration is applied to the transformed ADC map using the free-form deformation (FFD) algorithm [135], with the convolution-based fast local normalised correlation coefficient (LNCC) similarity measure to enable robustness to bias field inhomogeneity [136]. Finally, the transformation obtained from the composition of both types of registration is used to register computed high b-value DWI to T2WI.

5.2.2.3 Pre-processing III: resampling, cropping, and normalisation

T2WI, registered ADC map and computed high b-value DWI, and PZ and CG probability maps are resampled to a common in-plane resolution and cropped to a common in-plane shape, centered on the prostate; image cropping is used for memory efficiency. Then, T2WI and computed high b-value DWI are normalised by dividing voxel intensities by the interquartile mean of CG voxel intensities. Our approach is

a modification of the normalisation approach suggested by Bonekamp et al. [98], where voxel intensities were divided by the mean of PZ voxel intensities. We opt for normalisation using CG voxel intensities since CG segmentations are typically more reliable than PZ segmentations [8], and we opt for the interquartile mean of CG voxel intensities as opposed to the mean of all CG voxel intensities, to remove extremes that may correspond to abnormalities unique to a patient. ADC maps were not normalised as they contain a quantitative measurement.

5.2.2.4 CSPCa-U-Net-E

After pre-processing, each slice of a patient's T2WI, ADC map, computed high b-value DWI, and PZ and CG probability maps are input channel-wise to an ensemble of 2D U-Nets with hyperparameters taken from the the nnU-Net framework, modified for the task of CSPCa lesion segmentation where required; the addition of PZ and CG guidance as input to a CNN alongside MRI has been shown to increase CSPCa lesion detection performance as the occurrence and appearance of PCa is dependent on its zonal location [159]. We refer to each constituent 2D U-Net as CSPCa-U-Net and the ensemble of CSPCa-U-Nets as CSPCa-U-Net-E. Each CSPCa-U-Net is further modified to account for prediction uncertainty. CSPCa-U-Net features five encoding blocks and four decoding blocks. Each encoding block features two convolutional layers with stride one 3×3 convolutions with zero padding, LReLU activation (neg. slope 1e-2) and instance normalisation, followed by a stride two 2×2 max pooling operation in the first four encoding blocks; 64 feature maps are output by convolutional layers in the first encoding block, with feature maps doubling in each subsequent encoding block. Upsampling deconvolution operations are used in the decoding blocks, which receive semantic information from the last encoder block and higher resolution feature maps from encoder-to-decoder skip connections. In each CSPCa-U-Net, we model epistemic uncertainty using test-time dropout, following the approach in Kendall et al. [160] i.e., dropout layers are inserted after the central three encoder units and two decoder units, with dropout probability equal to some value P. We model aleatoric uncertainty using test-time augmentation as in Wang et al. [156]. The output of each CSPCa-U-

Net is slice-wise CSPCa probability maps. Per-voxel averaging is used to combine the probability map outputs of each CSPCa-U-Net $\in$ CSPCa-U-Net-E, followed by restacking of slices to form a probability map volume.

5.2.2.5 Post-processing

The CSPCa probability map output by CSPCa-U-Net-E is transformed to the original T2WI size and voxel resolution using padding and resampling operations. Next, probabilities are calibrated using an isotonic regression calibration module [161], to allow more interpretable CSPCa likelihoods. CSPCa lesion segmentations are obtained by thresholding CSPCa probability maps using a cut-off value C; C is chosen during experimentation using training data to match AutoProstate's detection sensitivity and specificity to that of an experienced radiologist. Finally, a false-positive reduction step is applied to remove connected components smaller than MinSize mm^2.

5.2.3 Report-Generator module

The Report-Generator module generates an automatic report using input bpMRI and clinical data, and the outputs of the Zone-Segmenter and CSPCa-Segmenter modules; the report template is shown in Figure 5.2.

The left-hand pane contains interactive report elements including a patient selector and transverse, frontal, and sagittal views of zone and CSPCa lesion segmentation outputs overlaid on T2WI, with associated widgets for slice selection.

The topmost section of the main report interface is named *Patient Details*. This section includes *Patient Name*, *Hospital Number*, *Date of Birth*, *Scan Date*, *Age* (years), and *PSA* (ng/mL).

The second report section is named *Prostate Size and PSA Density*. This section presents calculated prostate lengths and volumes, and the PSAd. The *Transverse*, *Anterior–Posterior*, and *Cranio–Caudal* lengths of the prostate, in cm, are calculated using the maximum extents of the prostate on the whole prostate segmentation, where the whole prostate segmentation is the union of the PZ and CG segmentations. *Prostate Volume*, *Peripheral Zone Volume*, and *Central Gland Vol-*

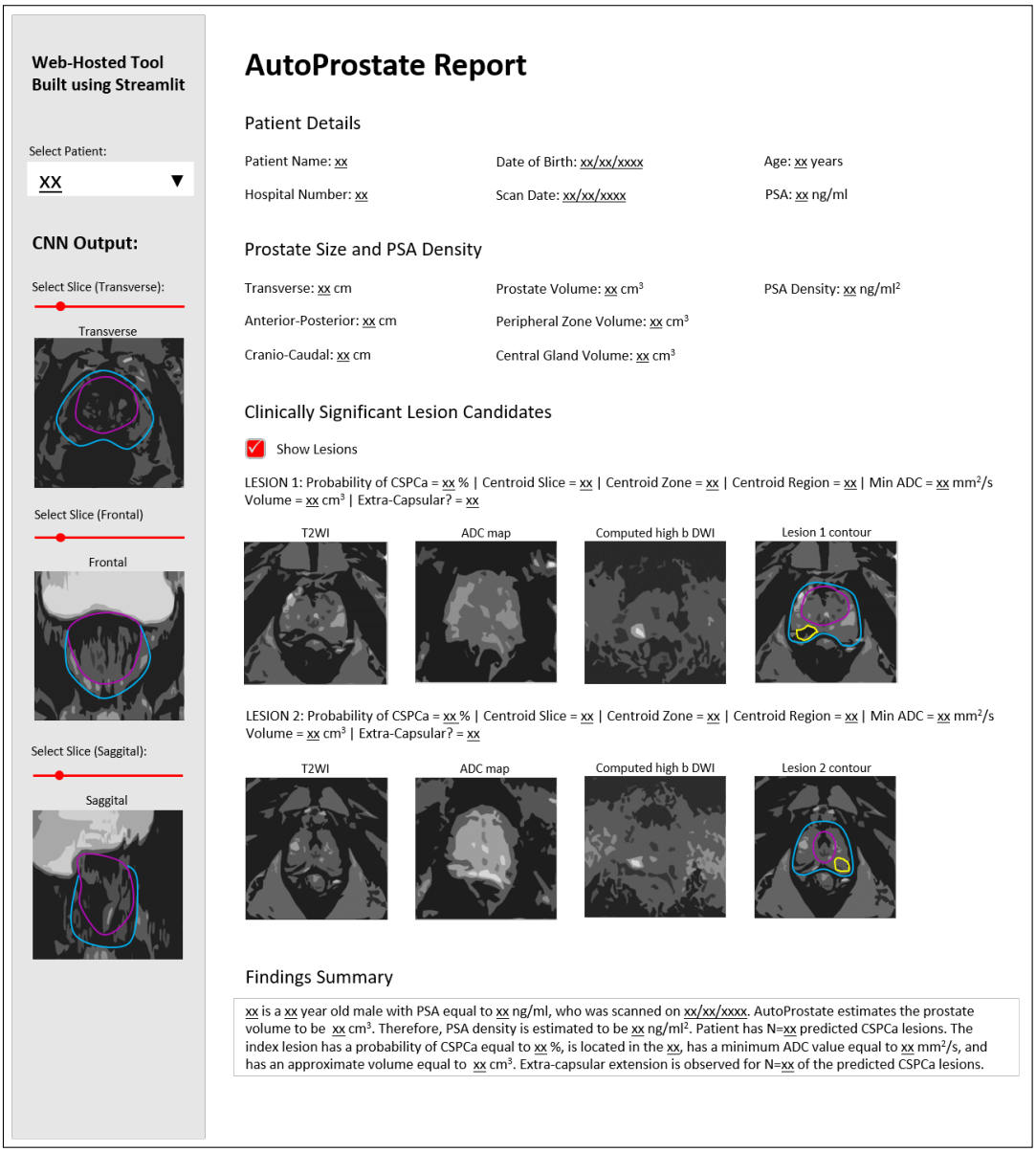


Figure 5.2: AutoProstate Report template, where xx denotes an automatically populated field.

ume, in cm^3, are calculated by multiplying voxel counts by voxel volume. The *PSA Density* (ng/mL^2) is calculated by dividing PSA by the calculated whole prostate volume.

The third report section is named *Clinically Significant Lesion Candidates*. This section presents a listing of all detected CSPCa lesions, sorted in descending order of *Probability of CSPCa*. The *Centroid Slice*, *Centroid Zone* (PZ or CG), and *Centroid Region* (base, midgland, or apex) are determined based on the location of the lesion centroid; our region determination follows the methodology outlined by Litjens et al. [91] for evaluating the PROMISE12 Challenge, where the apex is defined as the caudal-most third of the prostate, the base is the cranio-most third of the prostate, and the midgland is the remaining portion. The *Min ADC* (mm^2/s) is calculated as the minimum ADC value inside the predicted CSPCa lesion contour. As in the *Prostate Size and PSA Density* report section, *Volume* (cm^3) is calculated by multiplying voxel counts by voxel volume. Finally, the flag *Extra-Capsular?* is set to true if the lesion contour protrudes beyond the whole prostate contour, otherwise it is set to false.

The last section of the report is named *Findings Summary*, where key information (denoted xx in Figure 5.2) from other report sections is used to populate a template paragraph.

Following patient selection, the report is built using Streamlit (version 0.75.0; Available online: `https://streamlit.io` (accessed on 21 January 2021). Streamlit is an open-source Python library for creating shareable interactive web applications.

5.3 Experimental setup

In this section, we describe the datasets used for training and testing AutoProstate, the methodological settings employed, and the evaluation measures used to assess performance.

5.3.1 Patient datasets

AutoProstate was trained using the publicly available PROSTATEx challenges training dataset [44], and externally validated using the "Prostate Imaging Compared to Transperineal Ultrasound-guided biopsy for significant prostate cancer Risk Evaluation" (PICTURE) study dataset [33]. A detailed description of both datasets is given in Chapter 3, including details of mpMRI protocol, radiologist scoring, histopathological reference standard, and manually-drawn contours used as ground-truth in this work. In this work, two PICTURE dataset patients were removed due to missing MRI data.

5.3.2 Methodological settings

In this section, we describe the training and inference settings used for conducting experiments with AutoProstate.

5.3.2.1 Zone-Segmenter Module

T2WI were resampled to a common in-plane resolution of 0.4018 mm $\times$ 0.4018 mm and cropped to a common in-plane shape of 320 $\times$ 320.

A ten-fold cross-validation analysis of Zone-U-Net was conducted using the PROSTATEx dataset to optimise training hyperparameters, loss function, and augmentations. Zone-U-Net performed optimally when trained for 50 epochs with learning rate equal to 0.0001, batch size equal to eight, Adam optimisation [148], an equally-weighted hybrid loss composed of Dice loss [74] and Focal loss [162], and horizontal flip (probability 0.5), rotation (–20°, 20°), and scaling (-10%, 20%) augmentations.

Following the ten-fold cross-validation, the ten trained Zone-U-Nets were used to construct Zone-U-Net-E; cross-validation ensembles have been shown to be an effective ensembling strategy [8].

5.3.2.2 CSPCa-Segmenter Module

A high b-value, b_c = 2000, was selected for computing high b-value DWI as in Verma et al. [28]. The registration of ADC maps to T2WI employed default parameters for affine registration via symmetric block-matching. The subsequent non-

rigid FFD registration used a Gaussian kernel with standard deviation equal to 5 mm for LNCC calculation, control point spacing equal to 10 mm, and bending energy constraint equal to 0.1. Registrations were run using NiftyReg (version 1.3; `https://github.com/KCL-BMEIS/niftyreg`). Through visual inspection, we observed that the degree of misregistration was inconsequentially small for all PROSTATEx and PICTURE dataset cases. Therefore, no manual steps were taken to correct any instances of misregistration, and cases with misregistration were not excluded from our analysis.

T2WI, registered ADC maps and computed b2000 (Cb2000) DWI, and PZ and CG probability maps, were resampled to a common in-plane resolution of 0.4018 mm $\times$ 0.4018 mm and cropped to a common in-plane shape of 256 $\times$ 256, centered on the prostate.

Like Zone-U-Net, the training settings for CSPCa-U-Net were determined through ten-fold cross-validation using the PROSTATEx dataset. CSPCa-U-Net performed optimally when trained for 50 epochs with learning rate equal to 0.0001, batch size equal to 12, Adam optimisation, a dropout probability of P = 0.2 for central dropout, a hybrid loss composed of the sum of Dice loss multiplied by 0.5 and Focal loss multiplied by 1.0, and horizontal flip (probability 0.5), rotation (–20°, 20°), and scaling (-10%, 20%) augmentations. The same dropout probability and augmentation settings were used for test-time dropout and test-time augmentation.

CSPCa probability maps output by CSPCa-U-Net for each fold were calibrated using separate isotonic calibration modules for each fold. Following calibration, CSPCa probability maps were thresholded using cut-off values determined for each fold, corresponding to a lesion-level sensitivity of 93% and specificity of 37%, in the fold's training set. The aforementioned sensitivity and specificity correspond to reference radiologist performance at PI-RADS v1 cut-off ≥ 4 on a separate patient cohort from Radboud Medical Center, reported on in Litjens et al. [104], which was used since prospective radiologist performance was not available for the PROSTATEx dataset. As a final post-processing step, connected components smaller than 40 mm^3 were removed. UK National Institute for Health and Care Excellence (NICE)

guidelines recommend a minimum size of 200 mm^3 for CSPCa lesions [9]; we picked a minimum size of 40 mm^3 (20% of 200 mm^3) considering some CSPCa lesions may only be partially segmented.

Following the ten-fold cross-validation, the ten trained CSPCa-U-Nets were used to construct CSPCa-U-Net-E. CSPCa-U-Net-E was calibrated using isotonic calibration. For thresholding, a cut-off value C = 4.5% was determined to match radiologist performance in the training set for CSPCa-U-Net-E i.e., the entire PROSTATEx dataset. For false-positive reduction, connected components smaller than 40 mm^3 were removed, as in the cross-validation analysis.

5.3.3 External validation evaluation measures

Whole prostate and zonal segmentations were evaluated using the Dice coefficient. Prostate size measurements (transverse, anterior-posterior, and cranio-caudal lengths), as well as whole prostate and zonal volumes, were evaluated using absolute percentage error (Abs%Err); the ground-truth lengths and volumes used in the calculation of Abs%Err were derived from the manually-drawn whole prostate and zonal contours. The PSAd estimated by AutoProstate was evaluated using absolute error (AbsErr), since the absolute value of PSAd has a meaning relative to risk definitions [9]; the ground-truth PSAd value used in the calculation of AbsErr was obtained by dividing PSA by the whole prostate volume calculated using the manually-drawn whole prostate contour. The aforementioned evaluation metrics were calculated over 80 patients from the PICTURE dataset for which manually-drawn whole prostate and zonal segmentations were available.

Receiver operating characteristic (ROC) area under the curve (AUC) and precision-recall (PR) AUC were calculated to quantify AutoProstate's ability to differentiate between CSPCa lesions and nCSPCa lesions. After thresholding and false-positive reduction, we calculated sensitivity, specificity, and precision at lesion-level and average false-positives at patient-level. For the PICTURE dataset, the calculation of average false-positives was made using 93 patients who were biopsy-negative for CSPCa, due to limitations in the ground-truth prohibiting false-positive determination in biopsy positive patients. In addition, CSPCa lesion Dice

and Abs%Err of lesion area were calculated on contoured slices only.

Prostate volume, PSAd, and lesion detection metrics computed for AutoProstate were compared to the same metrics calculated for an experienced radiologist (10 years' experience in reading and scoring prostate mpMRI) who prospectively filled out a case report for each patient. Prostate volume was estimated using the ellipsoid formula and lesions were scored using a five-point Likert scale [31]. Statistical tests were used to compare the performances of AutoProstate and the experienced radiologist. The Wilcoxon's signed-rank test [145] was used to statistically compare prostate volume and PSAd estimates, DeLong's test was used to statistically compare lesion ROC AUC, McNemar's test [146] was used to statistically compare sensitivity and specificity, the weighted generalized score (WGS) test statistic [147] was used to statistically compare precision, and Wilcoxon's signed-rank test was used to statistically compare average false-positives.

5.4 Results

AutoProstate, trained using the PROSTATEx dataset, was externally validated using the PICTURE dataset. This section presents the results of the cross-validation of Zone-U-Net and CSPCa-U-Net which are building blocks of AutoProstate, a detailed analysis of the external validation of AutoProstate using the PICTURE dataset, with comparisons made to the performance of an experienced radiologist with 10 years' experience in reading prostate mpMRI, where possible.

5.4.1 Zone-U-Net and CSPCa-U-Net ten-fold cross-validation

Using the settings described in Section 5.3.2.1, Zone-U-Net achieved mean Dice coefficients of 0.78, 0.86, and 0.91 for PZ, CG, and whole prostate segmentation, respectively.

Using the settings described in Section 5.3.2.2, CSPCa-U-Net achieved a mean ROC AUC of 0.85 and a mean PR AUC of 0.70. After thresholding, CSPCa-U-Net achieved a mean sensitivity of 93%, a mean specificity of 37%, a mean precision of 34%, and a mean false-positive count per-patient of 6.9. Following false-positive reduction, mean sensitivity dropped marginally to 92%, mean specificity increased

to 46%, mean precision increased to 37%, and mean false-positives per-patient dropped significantly to 3.3 ($p < 0.01$). Furthermore, CSPCa-U-Net achieved a mean Dice coefficient of 0.39 for CSPCa lesion segmentation.

5.4.2 AutoProstate external validation analysis: whole prostate and zonal segmentation, prostate size measurements, and PSA density

Table 5.1 and Figure 5.3 present summaries of the distribution of Dice coefficients for whole prostate and zonal segmentations, the distribution of Abs%Err for prostate size measurements, and the distribution of AbsErr for PSAd calculation, for 80 patients from the PICTURE dataset for which ground-truth segmentations were available.

Mean Dice coefficients of 0.75, 0.80, and 0.89 were obtained for the PZ, CG, and whole prostate, respectively. AutoProstate's Zone-Segmenter module found PZ segmentation a more difficult task than CG segmentation, while whole prostate segmentation had a higher mean Dice coefficient than both zonal segmentations, suggesting an ease of distinguishing prostate tissue from background tissues, but a difficulty in distinguishing between PZ and CG tissue. As expected, the mean Dice coefficients for the PZ, CG, and whole prostate segmentations were lower than those obtained on the PROSTATEx dataset during the ten-fold cross-validation of Zone-U-Net (0.78, 0.86, and 0.91 for PZ, CG, and whole prostate segmentation, respectively) which may be indicative of a generalization gap due to acquisition/population differences.

The transverse, anterior-posterior, and cranio-caudal lengths of the prostate were estimated using the whole prostate segmentation output by Zone-Segmenter. A mean Abs%Err of 3%, 5%, and 20% were obtained for transverse, anterior-posterior, and cranio-caudal lengths, respectively. In addition to the lowest mean Abs%Err, the transverse length had a smaller standard deviation than anterior-posterior and cranio-caudal lengths. Through visual inspection of segmentation outputs, we attribute the variability in the accuracy of the anterior-posterior mea-

	Mean (SD)	Median (IQR)	Min - Max
Experienced radiologist (ellipsoid formula used to estimate whole prostate volume)			
Whole prostate volume Abs%Err	13 (11)	11 (5 – 20)	0 – 66
PSA density AbsErr	0.031 (0.032)	0.022 (0.008 – 0.043)	0.000 – 0.158
AutoProstate			
Segmentation			
Peripheral zone Dice coefficient	0.75 (0.06)	0.75 (0.70 – 0.79)	0.55 – 0.88
Central gland Dice coefficient	0.80 (0.07)	0.81 (0.77 – 0.85)	0.56 – 0.90
Whole prostate Dice coefficient	0.89 (0.03)	0.90 (0.88 – 0.92)	0.75 – 0.93
Lengths			
Transverse length Abs%Err	3 (2)	2 (1 – 4)	0 – 12
Anterior-posterior length Abs%Err	5 (4)	4 (2 – 7)	0 – 22
Cranio-caudal length Abs%Err	20 (15)	16 (10 – 31)	0 – 100
Volumes and PSA density			
Peripheral zone volume Abs%Err	12 (10)	10 (4 – 18)	0 – 49
Central gland volume Abs%Err	18 (15)	14 (10 – 25)	0 – 112
Whole prostate volume Abs%Err*	9 (7)	8 (5 – 12)	0 – 37
PSA density AbsErr*	0.019 (0.020)	0.014 (0.006 – 0.025)	0.000 – 0.129

AbsErr: absolute error; Abs%Err: absolute percentage error; IQR: interquartile range; Max: maximum; Min: minimum; PSA: prostate-specific antigen; SD: standard deviation. An asterisk indicates a p-value < 0.05 for AutoProstate compared to the experienced radiologist.

Table 5.1: AutoProstate external validation analysis of whole prostate and zonal segmentations, prostate size measurements, and PSAd, using 80 patients from the PICTURE dataset for which ground-truth segmentations were available.

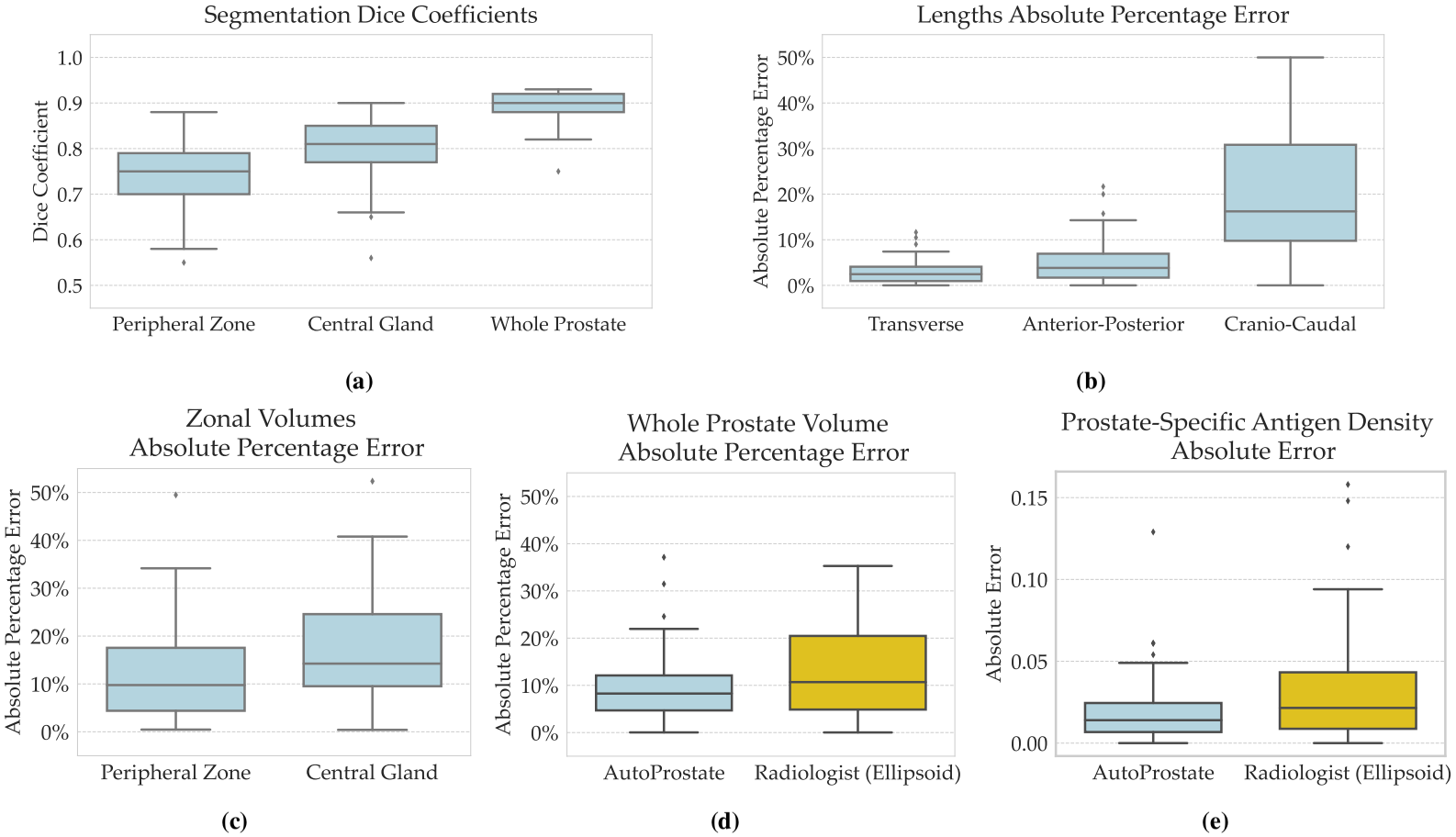


Figure 5.3: AutoProstate external validation analysis of whole prostate and zonal segmentations, prostate size measurements, and PSAd, using 80 patients from the PICTURE dataset for which ground-truth segmentations were available: (a) Distribution of Dice coefficients for PZ, CG, and whole prostate segmentation; (b) Distribution of Abs%Err for transverse, anterior-posterior, and cranio-caudal lengths; (c) Distribution of Abs%Err for PZ and CG volumes; (d) Distribution of Abs%Err for whole prostate volume estimations by AutoProstate and the experienced radiologist; and (e) Distribution of AbsErr for PSAd calculated by AutoProstate and the experienced radiologist; the ground-truth PSAd value used to compute the AbsErr for AutoProstate and the experienced radiologist was calculated by dividing PSA by the whole prostate volume derived from the ground-truth whole prostate segmentation.

surement to the difficulty of determining prostate extent in the anterior fibromuscular stroma, and similarly, we attribute the variability in the accuracy of the cranio-caudal measurement to the difficulty of determining prostate extent at the base and apex regions of the prostate. Strikingly, a large maximum Abs%Err of 100% was observed for the cranio-caudal measurement, which was found to be due to under-segmentation of the base region in the ground-truth.

PZ, CG, and whole prostate volumes were calculated using PZ, CG, and whole prostate segmentations output by Zone-Segmenter. A mean Abs%Err of 12%, 18%, and 9% were obtained for PZ, CG, and whole prostate volumes, respectively. Strikingly, a large maximum Abs%Err of 112% was observed for the CG, which was found to be due to over-segmentation of the CG in the base region. We compare the Abs%Err of the whole prostate volume calculated by AutoProstate to the same calculated by the experienced radiologist who used the ellipsoid formula, which is clinically advocated. AutoProstate had a mean Abs%Err of 9%, while the experienced radiologist's mean Abs%Err was 13%; the difference was statistically significant ($p < 0.05$). Using the whole prostate volumes computed by AutoProstate and the experienced radiologist, PSAd was calculated. AutoProstate achieved a mean AbsErr of 0.019, while the experienced radiologist's mean AbsErr was 0.031; the difference was statistically significant ($p < 0.05$).

5.4.3 AutoProstate external validation analysis: clinically significant prostate cancer lesion detection and segmentation

CSPCa lesion detection performance for AutoProstate and the experienced radiologist are shown in Table 5.2, while Figure 5.4 shows the ROC and PR curves for AutoProstate and the experienced radiologist.

AutoProstate achieved a mean ROC AUC of 0.70 and a mean PR AUC of 0.84, calculated using output CSPCa probability maps prior to thresholding. After thresholding the CSPCa probability maps using a cut-off value equal to 4.5%, the following were obtained: a sensitivity of 78%, a specificity of 49%, a precision of 78%, and a mean false-positive count of 6.1. Following false-positive reduction, mean sensitivity dropped marginally to 76%, mean specificity increased to 57%,

Radiologist (Likert scoring)	
ROC AUC	0.64 (0.56 – 0.72)
PR AUC	0.78 (0.71 – 0.84)
Post-thresholding (cut-off: Likert ≥ 4)	
Sensitivity / recall (%)	78 (71 – 84)
Specificity (%)	48 (35 – 60)
Precision (%)	78 (71 – 84)
Mean false-positives per-patient	0.3 (0.2 – 0.4)
AutoProstate	
ROC AUC	0.70 (0.62 – 0.78)
PR AUC	0.84 (0.77 – 0.90)
Post-thresholding (cut-off: $\geq 4.5\%$)	
Sensitivity / recall (%)	78 (71 – 85)
Specificity (%)	49 (37 – 62)
Precision (%)	78 (71 – 85)
Mean false-positives per-patient*	6.1 (5.5 – 6.8)
Post-thresholding (cut-off: $\geq 4.5\%$) and false-positive reduction (<40 mm^3)	
Sensitivity / recall (%)	76 (68 – 82)
Specificity (%)	57 (45 – 69)
Precision (%)	80 (74 – 87)
Mean false-positives per-patient*	2.5 (2.2 – 2.8)

AUC: area under curve; PR: precision-recall; ROC: receiver operating characteristic.
An asterisk indicates a p-value < 0.001 for AutoProstate compared to the radiologist.

Table 5.2: PICTURE dataset CSPCa lesion detection metrics for the radiologist and AutoProstate. Mean and standard deviation of false-positives per-patient were calculated using the 93 PICTURE dataset patients who were biopsy-negative for CSPCa, rather than over all patients, due to limitations in the ground-truth. All other metrics shown are calculated at the lesion-level for the 147 CSPCa lesions and 63 nCSPCa lesions.

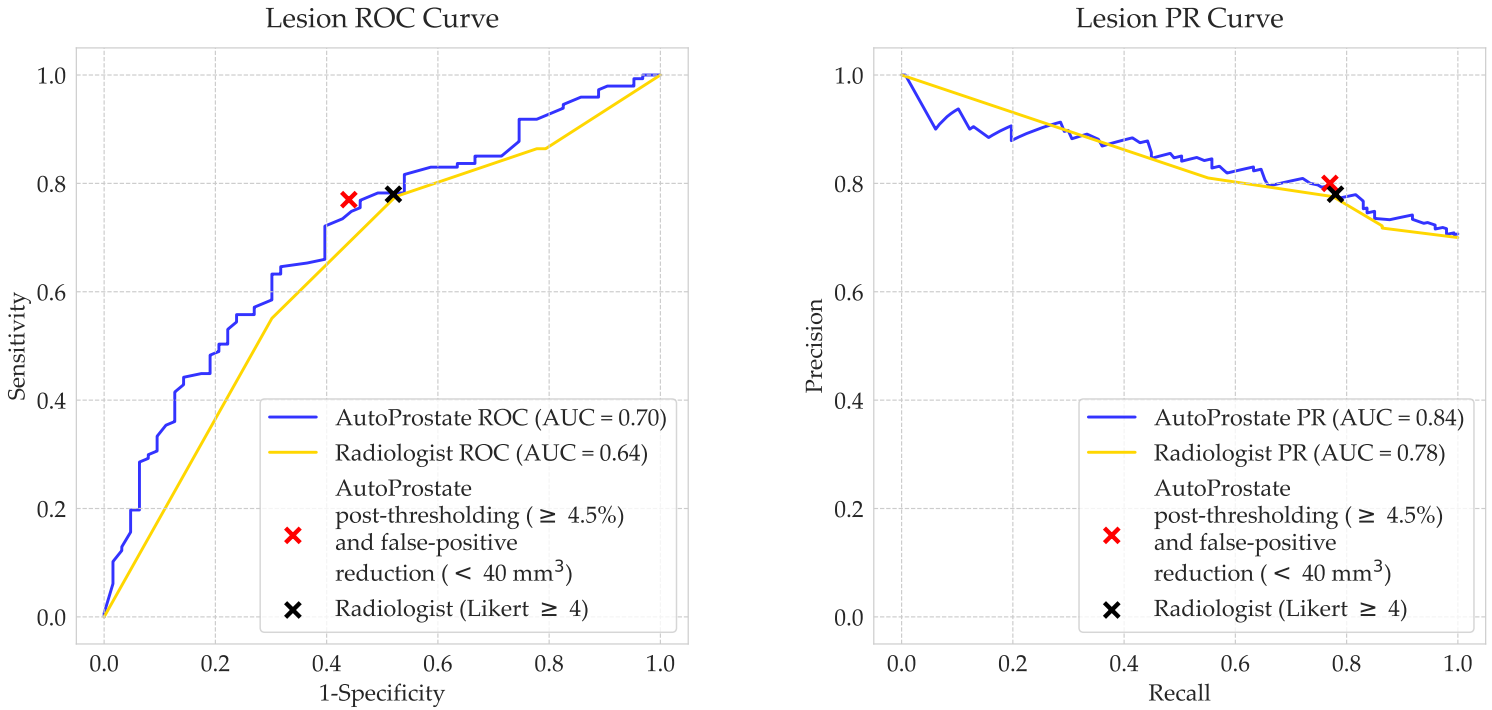


Figure 5.4: PICTURE dataset CSPCa lesion detection (a) ROC curves and (b) PR curves, corresponding to the radiologist (gold) and AutoProstate (blue).

mean precision increased marginally to 80%, and the mean false-positive count per-patient dropped to 2.5. The relationship between detection accuracy and lesion size was investigated. CSPCa lesion detection accuracy varied by lesion area (lesion area is used as a surrogate measure for lesion size since a single slice only was contoured for each identified lesion). We found an overall CSPCa lesion detection accuracy of 79% over the total 147 CSPCa lesions contoured. For CSPCa lesions with an in-slice area of less than 50 mm^3, the detection accuracy was found to be 70%, which increases to 75% for CSPCa lesions with an in-slice area greater than or equal to 50 mm^3 and less than 100 mm^3, and increases further to 94% for CSPCa lesions with an in-slice area greater than or equal to 100 mm^3.

Likert scores assigned to suspicious lesions by the experienced radiologist were used to calculate ROC and PR curves; radiologist Likert scoring gave a ROC AUC of 0.64 and PR AUC 0.78. After thresholding at cut-off score Likert ≥ 4, the following were obtained: a sensitivity of 78%, a specificity of 48%, a precision of 78%, and a mean false-positive count of 0.3. Differences between the ROC AUC, PR AUC, sensitivity, specificity, and precision of AutoProstate and the experienced radiologist were not statistically significant. However, the difference between mean false-positives was statistically significant ($p < 0.001$).

A further analysis was completed to assess the level of agreement between AutoProstate and the radiologist's Likert scores, on annotated lesions, as shown in Table 5.3. For CSPCa lesions, there was a 78% (114/147) concordance between AutoProstate and the radiologist, while for nCSPCa lesions, there was a 62% (39/63) concordance.

CSPCa lesion segmentation accuracy, evaluated using the Dice coefficient, was calculated using slices containing a corresponding ground-truth CSPCa lesion contour. The following Dice coefficient metrics were obtained: a mean of 0.46 (SD: 0.32), a median of 0.58 (IQR: 0.10 - 0.72), and a min-max range of 0.00 – 0.90. Example CSPCa lesion segmentations are shown in Figure 5.5, Figure 5.6, and Figure 5.7. Furthermore, an example automatic report generated by AutoProstate is shown in Figure 5.8.

		CSPCa Lesions (n = 147)		nCSPCa Lesions (n = 63)	
		AutoProstate			
		Predicted nCSPCa	Predicted CSPCa	Predicted nCSPCa	Predicted CSPCa
Radiologist	Not identified prospectively	8% (12/147)	5% (8/147)	16% (10/63)	5% (3/63)
	Likert 2	0% (0/147)	0% (0/147)	2% (1/63)	0% (0/63)
	Likert 3	4% (6/147)	5% (7/147)	16% (10/63)	10% (6/63)
	Likert 4	7% (10/ 147)	16% (23/147)	14% (9/63)	8% (5/63)
	Likert 5	5% (8/ 147)	50% (73/147)	10% (6/63)	21% (13/63)

CSPCa: clinically significant prostate cancer; nCSPCa: not clinically significant prostate cancer.

Table 5.3: Agreement and disagreement between radiologist Likert scoring and AutoProstate, on PICTURE dataset annotated lesions; grey shading indicates concordance, gold shading indicates superior performance by the radiologist, and blue shading indicates superior performance by AutoProstate.

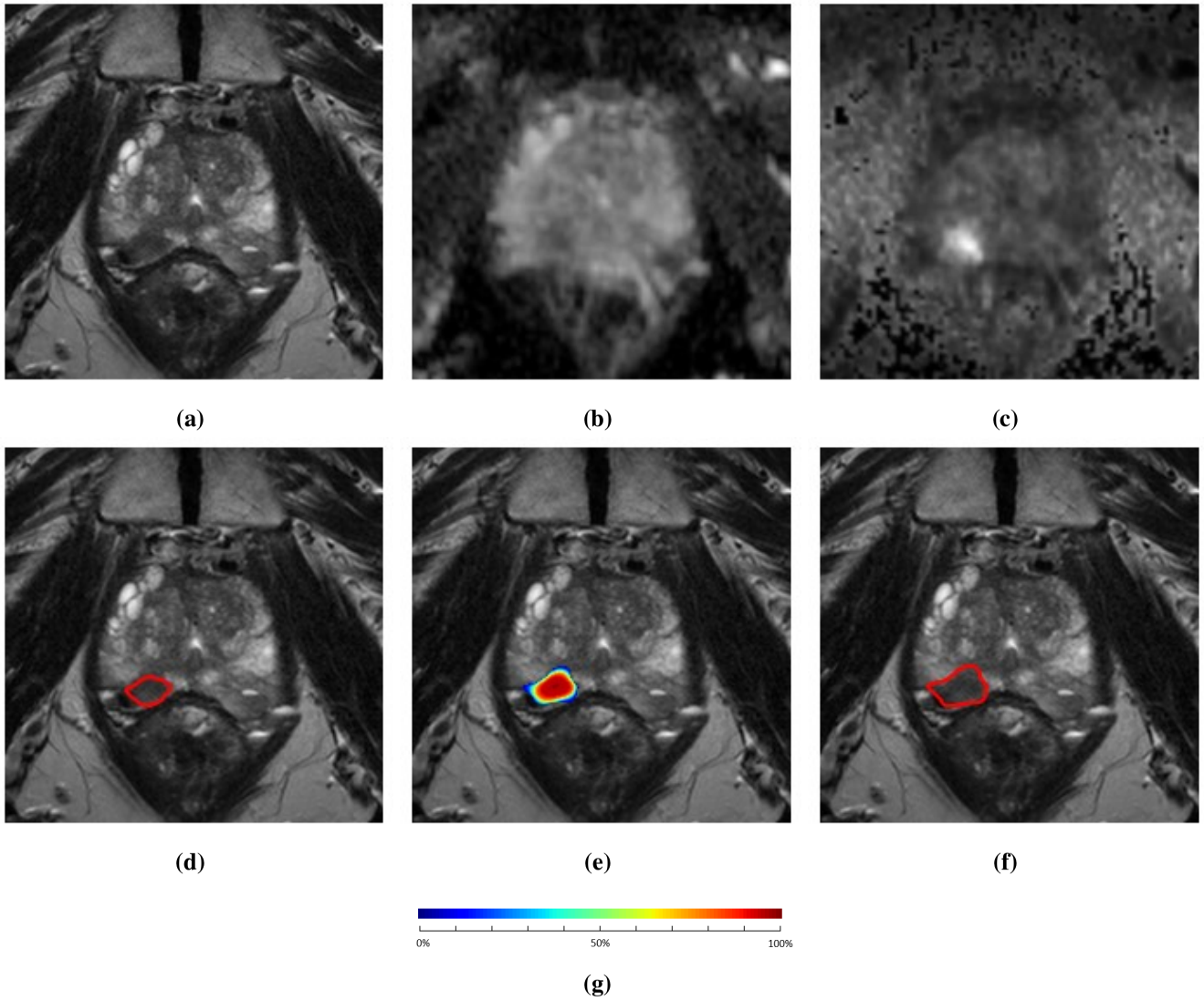


Figure 5.5: 79-year-old man with PSA 12.57 ng/ml, and a Likert 5 midgland PZ Gleason score 4+3 lesion, from the PICTURE dataset. (a) axial T2WI, (b) ADC map, (c) Cb2000 DWI, (d) ground-truth lesion contour overlaid on T2WI, (e) probability map overlaid on T2WI, (f) segmentation overlaid on T2WI, and (g) probability map legend.

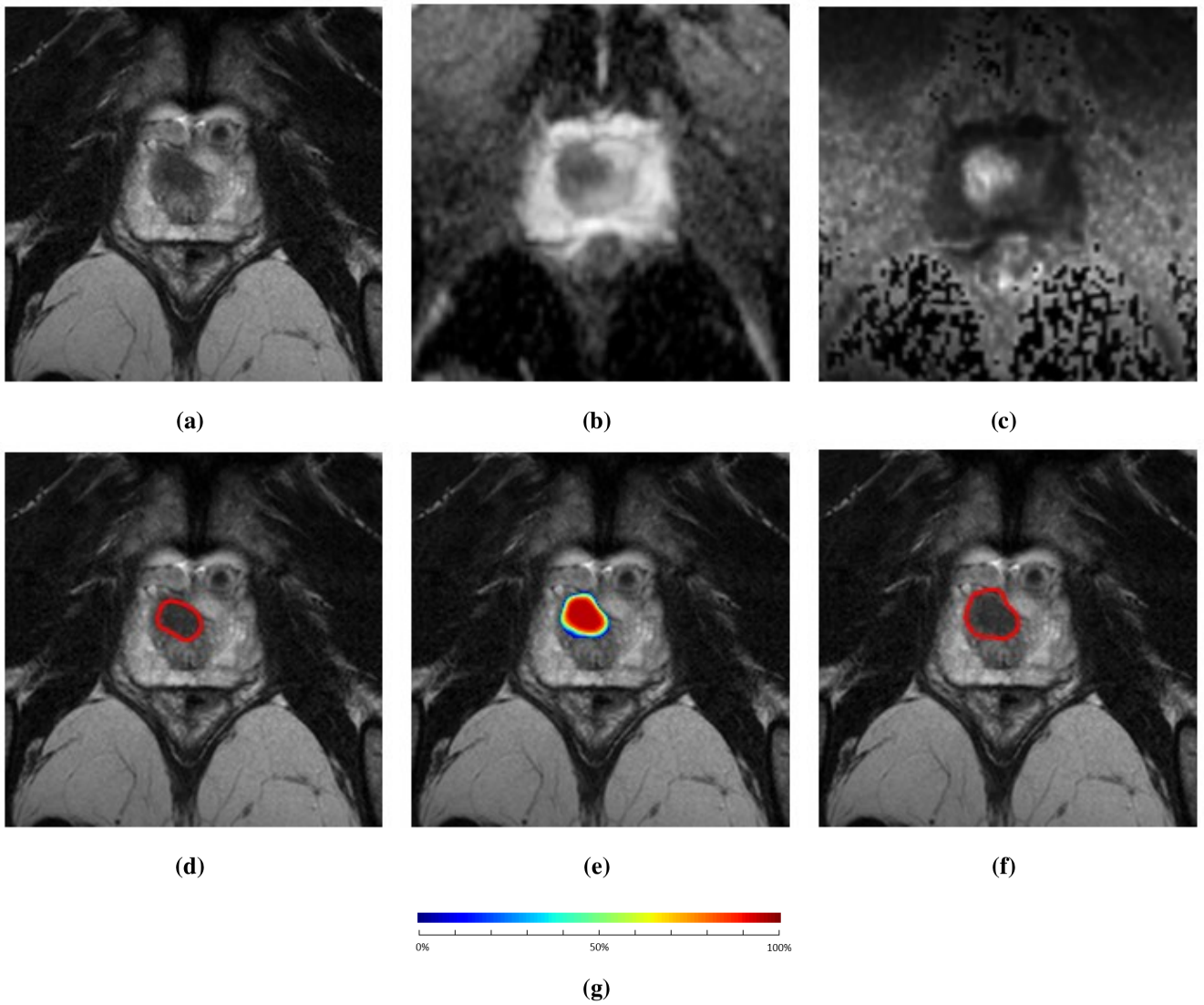


Figure 5.6: 64-year-old man with PSA 10.53 ng/ml, and a Likert 5 apex CG Gleason score 3+4 lesion, from the PICTURE dataset. (a) axial T2WI, (b) ADC map, (c) Cb2000 DWI, (d) ground-truth lesion contour overlaid on T2WI, (e) probability map overlaid on T2WI, (f) segmentation overlaid on T2WI, and (g) probability map legend.

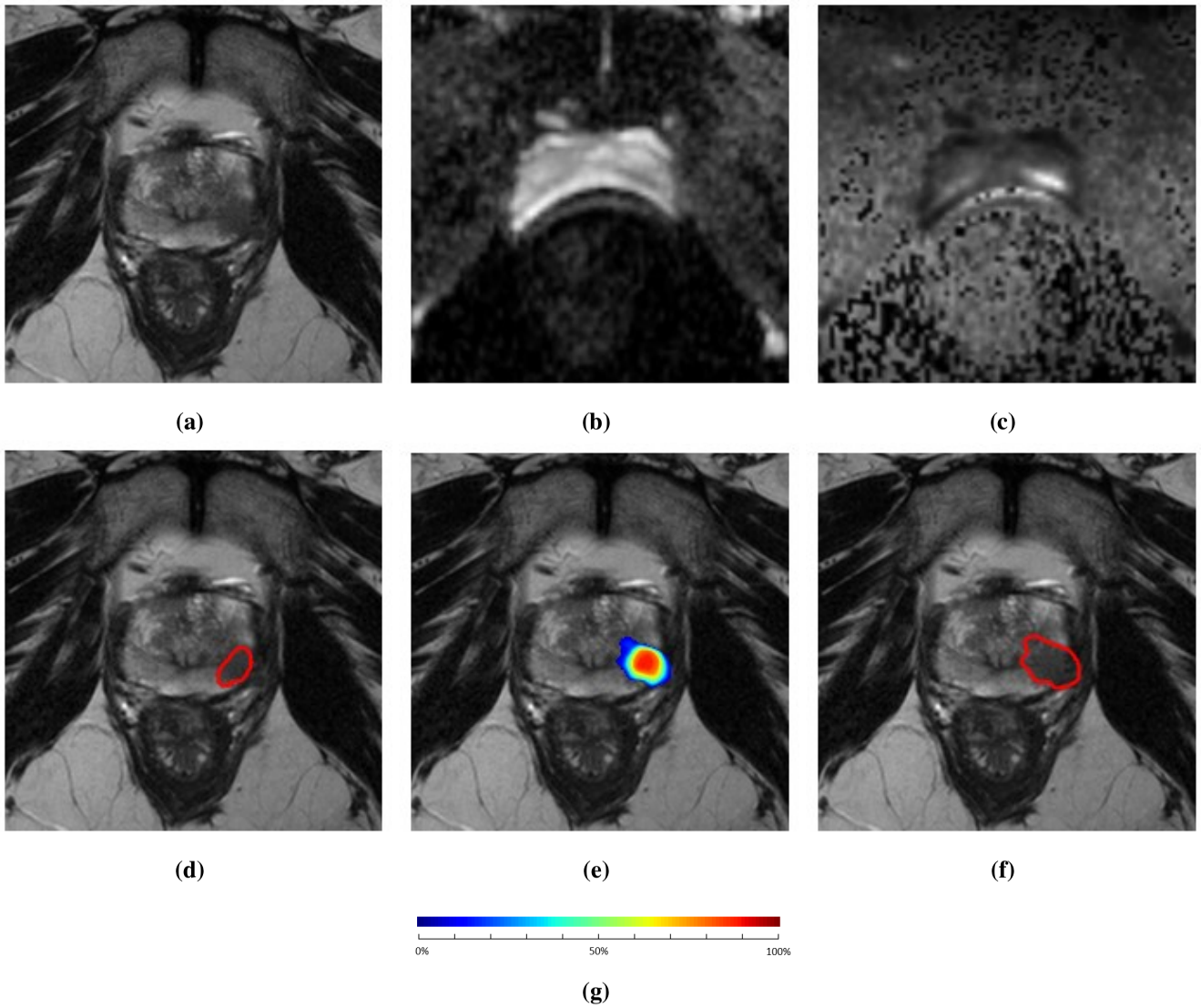


Figure 5.7: 60-year-old man with PSA 6.15 ng/ml, and a Likert 5 midgland PZ Gleason score 3+4 lesion, from the PICTURE dataset. (a) axial T2WI, (b) ADC map, (c) Cb2000 DWI, (d) ground-truth lesion contour overlaid on T2WI, (e) probability map overlaid on T2WI, (f) segmentation overlaid on T2WI, and (g) probability map legend.

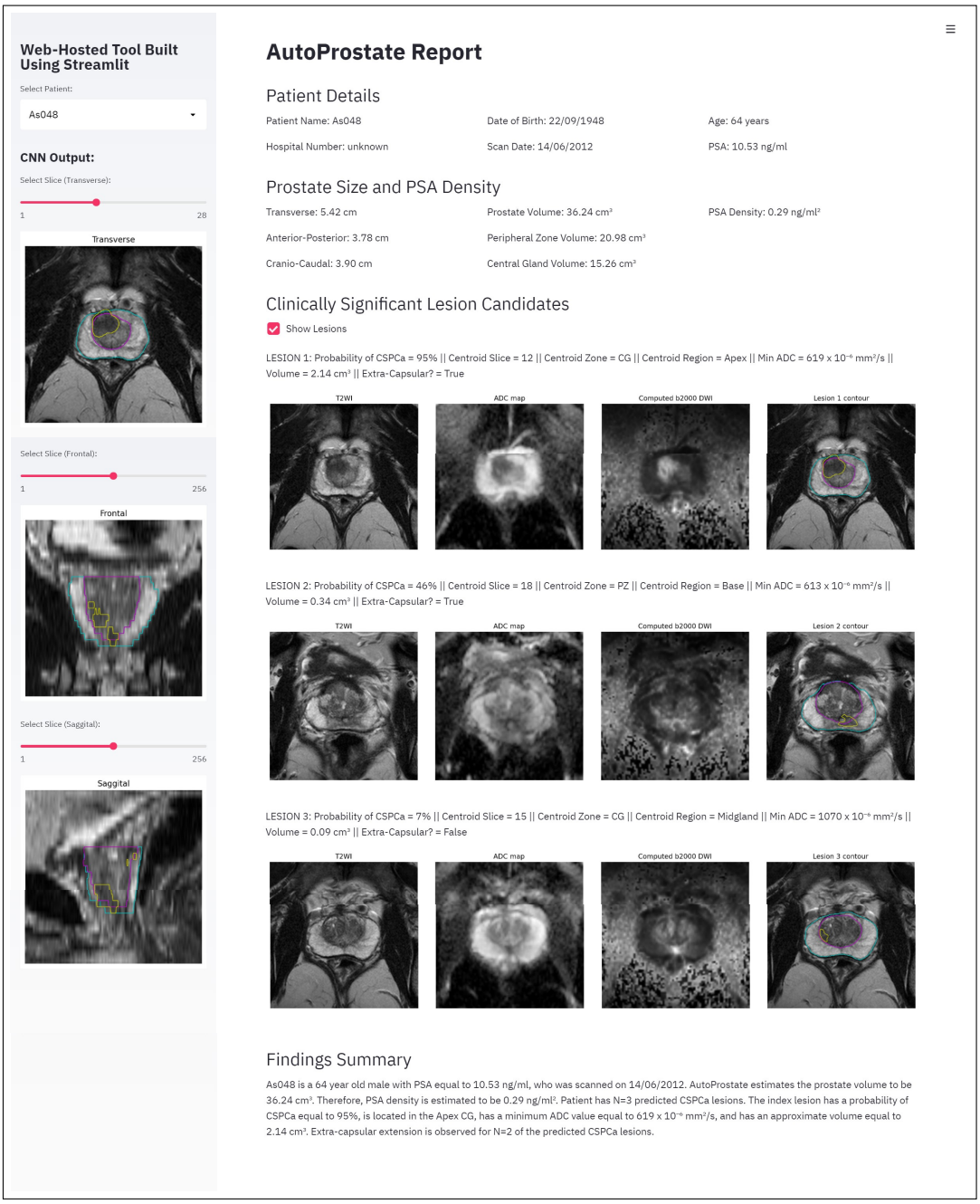


Figure 5.8: AutoProstate report for a 64-year-old man with PSA equal to 10.53 ng/ml who participated in the PICTURE study. LESION 1 (probability of CSPCa equal to 95%) corresponds to a biopsy-proven Gleason score 3+4 lesion, while LESION 2 and LESION 3 (probabilities of CSPCa equal to 46% and 7%, respectively) are false-positives.

5.5 Discussion

In this work, we introduced AutoProstate, a deep learning-powered framework for automated MRI-Based PCa assessment. AutoProstate consists of three modules: Zone-Segmenter, CSPCa-Segmenter, and Report-Generator. The output of AutoProstate is an automatic web-based report that presents patient details, prostate size measurements and PSAd, a listing of candidate CSPCa lesions with derived characteristics, and a findings summary. AutoProstate, trained using the publicly available PROSTATEx dataset, was externally validated using the PICTURE dataset. During the external validation, the performance of AutoProstate was compared to the performance of an experienced radiologist with 10 years' experience in reading prostate mpMRI, who prospectively estimated prostate volume and PSAd using the ellipsoid formula, and scored lesions using a five-point Likert scale.

PZ, CG, and whole prostate segmentations are output by AutoProstate's Zone-Segmenter module. During the experimental setup phase, we tested Zone-U-Net, prior to ensembling of Zone-U-Nets to form Zone-U-Net-E. Zone-U-Net achieved mean Dice coefficients of 0.78, 0.86, and 0.91 for PZ, CG, and whole prostate segmentation, respectively, in ten-fold cross-validation using the PROSTATEx dataset. Our result compares well to recent works by Aldoj et al. [87], where their proposed Dense-2 U-Net CNN was evaluated using four-fold cross-validation of a 188-patient subset from the PROSTATEx dataset, and to a recent work by Cuocolo et al. [88], where the previously proposed ENet CNN [93] was evaluated using a 105-patient test set from the PROSTATEx dataset. Aldoj et al. obtained mean Dice coefficients of 0.78, 0.91, and 0.92, and Cuocolo et al. obtained mean Dice coefficients of 0.71, 0.87, and 0.91, for PZ, CG, and whole prostate segmentation, respectively. However, direct comparisons between our work and the works of Aldoj et al. and Cuocolo et al. is not possible due to the use different subsets of data for testing. During the external validation of AutoProstate using the PICTURE dataset, where Zone-U-Net-E was used for PZ, CG, and whole prostate segmentation, AutoProstate achieved mean Dice coefficients of 0.75, 0.80, and 0.89, respectively, on 80 patients for which ground-truth segmentations were available. Antonelli et al. [85] previ-

ously reported segmentation results for the PICTURE dataset. A multi-atlas segmentation approach featuring a novel genetic atlas selection strategy was proposed; mean Dice coefficients of 0.72 and 0.83 were reported for PZ and CG segmentation, using leave-one-out cross-validation, and a mean Dice coefficient of 0.83 was reported for whole prostate segmentation, using atlases from the PROMISE12 dataset [91].

As Ghavami et al., [163] have shown it is not sufficient to evaluate an automatic whole prostate segmentation algorithm using Dice score alone; the performance gain, or lack thereof, in downstream tasks must also be considered, in this case downstream calculations of prostate volume and PSAd. AutoProstate's estimate of prostate volume was compared to an estimate obtained using the ellipsoid formula, which is clinically advocated [152]. AutoProstate achieved a mean Abs%Err of 9%, while the radiologist computed ellipsoid formula estimate had a mean Abs%Err of 13%. Notably, the difference in mean Abs%Err was statistically significant ($p = 0.0051 < 0.05$). Furthermore, we compared PSAd estimates obtained using the volume estimates; we found a mean AbsErr of 0.019 for AutoProstate and a mean AbsErr of 0.031 for the radiologist; again, the difference was statistically significant ($p = 0.0018 < 0.05$). Since PSAd is used clinically to inform the decision to biopsy or to discharge patients [9] and furthermore, to monitor patients on active surveillance, as recommended by NICE guidelines in the UK [9], we believe a case exists for replacement of the ellipsoid formula with automated methods such as ours.

AutoProstate's foremost purpose is to detect and segment CSPCa lesions. During the experimental setup phase, we tested CSPCa-U-Net, prior to ensembling of CSPCa-U-Nets to form CSPCa-U-Net-E. Markedly, CSPCa-U-Net achieved a lesion-level mean ROC AUC of 0.85 in ten-fold cross-validation using the PROSTATEx dataset, while previous studies have reported a lesion-level mean ROC AUC of 0.81 on the same subset of PROSTATEx data used in this study, using the same input modalities. During the external validation of AutoProstate using the PICTURE dataset, where CSPCa-U-Net-E was used to segment CSPCa lesions,

AutoProstate achieved a lesion-level ROC AUC of 0.70. Notably, we observed a large reduction in ROC AUC on the PICTURE dataset from that seen during the PROSTATEx dataset ten-fold cross-validation. We believe that the main reason for the reduction in ROC AUC was the use of TTPM biopsy in the PICTURE dataset reference standard, which allowed lesions that were not prospectively identified by the radiologist, to be retrospectively contoured using TTPM biopsy findings. Other reasons may include a high occurrence of magnetic susceptibility artifacts on DWI in the PICTURE dataset and a possible generalization gap between training data and external testing data due to population/acquisition differences. On the PICTURE dataset, radiologist Likert assessment achieved a lesion-level ROC AUC of 0.64; the difference in ROC AUC between AutoProstate and the radiologist was not statistically significant. Following thresholding and false-positive reduction, AutoProstate achieved a lesion-level sensitivity of 76%, a lesion-level specificity of 57%, and 2.5 false-positives per-patient (calculated over patients without CSPCa, only). In comparison, radiologist Likert assessment thresholded at Likert ≥ 4, achieved a lesion-level sensitivity of 78%, a lesion-level specificity of 48%, and 0.3 false-positives per-patient (calculated over patients without CSPCa, only); only the difference between the number of false-positive detections by AutoProstate and the radiologist was statistically significant ($p < 0.001$). While AutoProstate has demonstrated an ability to differentiate between CSPCa lesions and low-grade/benign lesions at the level of an experienced radiologist, further work is needed to reduce the number of false-positives produced. Interestingly, AutoProstate achieved a similar sensitivity and improved specificity compared to the radiologist on annotated CSPCa and low-grade/benign (nCSPCa) lesions but had a higher overall false-positive count. Therefore, it's possible that the additional false-positives produced by AutoProstate, that were not prospectively scored by the radiologist, may be easy for radiologists to rule-out.

Several aspects of this study have been guided by the set of nine key considerations for authors, reviewers, and readers of artificial intelligence studies in radiology by Bluemke et al. [126]. As recommended, we maintained a clear sep-

aration between training data and testing data. In particular, we avoided a common pitfall observed in previous studies [120, 112], by determining the probability cut-off value using training data, rather than a biased approach involving the test data itself. In line with further recommendations by Bluemke et al., we were able to externally validate AutoProstate using the PICTURE dataset. Furthermore, the PICTURE dataset was acquired using Phillips' scanners, while the PROSTATEx dataset, used to train AutoProstate, was acquired using Siemens' scanners, meaning a further recommendation on using multivendor data for evaluation was met. Moreover, we compared AutoProstate to an expert radiologist who prospectively reported PICTURE dataset patients, and both AutoProstate and the radiologist were compared to an accepted reference standard which combined TTPM and MR-guided targeted biopsies; TTPM biopsy is highly accurate and avoids biases associated to MR-guided targeted biopsy, transrectal ultrasound-guided (TRUS) biopsy, and prostatectomy [33].

CAD system studies should describe how the CAD system will be deployed clinically, so future prospective trials can be planned accordingly. Our goal in this study was to understand the strengths, weaknesses, and idiosyncrasies of AutoProstate through a comparison against an experienced radiologist. In the clinical workflow, we envision AutoProstate as a radiologist companion system during clinical reads to allow enhanced clinical reporting. It should be acknowledged that current CAD systems for MRI-based PCa diagnosis contain varying degrees of error in terms of producing too many false-positives, false-negatives, or both. Since the automatic report produced by AutoProstate presents visual segmentation outputs, as well as derived measurements, all outputs produced by AutoProstate can be rapidly verified by the radiologist. In particular, automatic report information deemed to be accurate can be used to prepare the patient's clinical report, while erroneous information can be recalculated using current clinical methods or ignored if not required. However, prior to deployment, focus groups must be held with radiologists to refine all aspects of the report, followed by prospective validation in the clinical workflow. Particular attention must be given to the "Probability of

CSPCa" computed by AutoProstate for each CSPCa lesion candidate. Probabilities are subject to human interpretation, which itself is dependent on several factors such as clinical experience, technical understanding, and level of confidence in the system. Furthermore, it has been shown that clinicians can struggle to correctly interpret diagnostic information to make decisions, particularly when this involves probabilistic reasoning [164]. Successful clinical adoption may hinge on the provision of a study-backed guide on how to use data provided from the system to make decisions. In the longer term, as the number of AI-based systems for PCa diagnosis, such as ours, proliferates, there may need to be a careful adaption of radiologist reporting guidelines.

There were three limitations in our study. Firstly, our training data was limited to 76 CSPCa lesions and 223 nCSPCa lesions; we may expect improved detection sensitivity and reduced false-positives if a bigger training dataset with more lesions is available. Secondly, our external validation was limited to a single external site. Thirdly, lesion contours for each PICTURE dataset patient were drawn by a single radiologist only. While the location and Gleason score of lesions was confirmed by a combination of TTPM and MR-guided targeted biopsies, we were not able to overcome the inter-reader variation known to exist in lesion boundary determination [129].

Chapter 6

Click-based interactive segmentation of the whole prostate and prostatic lesions: a preliminary study

The results presented in this chapter have been published by Diaz-Pinto et al. [165] in the conference paper entitled, "DeepEdit: Deep Editable Learning for Interactive Segmentation of 3D Medical Images", in which the author of this thesis is the second author.

6.1 Introduction

Obtaining accurate segmentations of the whole prostate and prostatic lesions, on magnetic resonance imaging (MRI), is of interest clinically for prostate cancer (PCa) diagnosis, active surveillance, and treatment. In particular, accurate segmentations can enable more precise calculations of prostate and lesion volume. Prostate volume is needed for calculating prostate-specific antigen (PSA) density (PSAd), which is a predictor of clinically significant PCa (CSPCa) [153], while lesion volume is required for lesion scoring, as per Prostate Imaging-Reporting and Data System (PI-RADS) guidelines [24], and for monitoring disease progression in patients placed on active surveillance [166]. In addition to volume estimation, accurate segmentation allows better determination of prostate and lesion extent. Prostate and lesion boundaries are important for planning/conducting whole-gland

or partial prostatectomy [167], for planning/conducting whole-gland or focal radiotherapy [168], for brachytherapy radioactive seed placement [169], and for planning/conducting whole-gland or focal cryotherapy [170]. However, due to the complexity and time-cost of producing segmentations of the prostate and lesions within the prostate, they are not produced by radiologists as part of the clinical routine [112].

Several methods have been described in the literature for the automatic segmentation of organs and lesions on medical images [40]. In particular, deep learning-based segmentation algorithms are producing state-of-the-art performance on medical image segmentation tasks [8, 72, 73]. However, as automatic segmentation approaches have not reached the required robustness for clinical use [45], interactive segmentation methods for medical images are being developed [83, 45, 84].

In this work, we built and evaluated interactive segmentation applications for whole prostate segmentation and prostatic lesion segmentation. The work is based on an interactive segmentation method that uses deep learning introduced by Sakinis et al. [45], named DeepGrow, in which mouse-clicks provided by a user e.g., a radiologist, are used to correct areas of under- or over-segmentation; experiments on multiple organ segmentation on computerised tomography (CT) scans of the abdomen showed that their method generated segmentations in a fast and reliable manner with few clicks and generalised well to unseen structures. Both whole prostate and prostatic lesion segmentation applications are built using MONAI Label (`https://github.com/Project-MONAI/MONAILabel`), which is an open-source repository that enables researchers to build medical image segmentation applications using automatic and interactive segmentation methods. Both segmentation tasks were performed in 3D to ensure user clicks apply in-plane and across slices; a 3D implementation of DeepGrow is available in MONAI Label, building on DeepGrow for 2D segmentation, as presented by Sakinis et al. The whole prostate segmentation task was performed using T2-weighted MRI (T2WI) using the standard MONAI Label application for performing training and inference with DeepGrow, while for prostatic lesion segmentation, the standard MONAI

Label application was extended to accept multimodal input, in this case T2WI, apparent diffusion coefficient (ADC) map, and computed high b-value diffusion-weighted MRI (DWI). In our experiments, we compared the use of Dice loss [74] for training, as used in the work by Sakinis et al., to a hybrid loss composed of Dice loss and Focal loss [154]. In a further experiment, we proposed and evaluated a new click-simulation strategy for training DeepGrow which uses the magnitude of the prediction error associated to false-positive and false-negative regions to determine click-placement, as opposed to the size of false-positive and false-negative regions, as suggested by Sakinis et al. A final experiment compares DeepGrow with DeepEdit, where DeepEdit is a modification of DeepGrow available in MONAI Label. While DeepGrow, as presented by Sakinis et al., requires at least one mouse-click to initiate the generation of a segmentation, DeepEdit is trained to perform click-free segmentation at first, followed by click-based segmentation editing if required. Provided sufficient performance, DeepEdit can reduce the number of mouse clicks required, saving time and effort.

This chapter is organised as follows: In Section 6.2, we describe the technical details of DeepGrow and DeepEdit. In Section 6.3, we describe the pipeline configurations of the whole prostate and prostatic lesion segmentation applications, the training settings, the experiments conducted, and the evaluation methods employed. In Section 6.4, we present the results for segmentation. Finally, in Section 6.5, we conclude by discussing the implications of our results.

6.2 Methods

In this section, we describe the technical details of DeepGrow and DeepEdit. In addition to the standard training strategy for DeepGrow and DeepEdit, a new training strategy is also described.

6.2.1 DeepGrow

DeepGrow refers to the click-based interactive segmentation method proposed by Sakinis et al. [45]. DeepGrow uses an encoder-decoder CNN to segment medical images, using a stacked channel input composed of the image, a foreground

click map, and a background click map; the foreground click map contains the locations of mouse-clicks placed in false-negative regions of the segmentation, while the background click map contains the locations of mouse-clicks placed in false-positive areas of the segmentation. While the original method published by Sakinis et al. segmented 2D images only, MONAI Label has extended DeepGrow for segmentation of 3D images.

6.2.1.1 CNN architecture

The CNN architecture used in the original DeepGrow method published by Sakinis et al. is an encoder-decoder CNN inspired by U-Net [6], while the DeepGrow implementation in MONAI Label uses a dynamic U-Net [6, 7], configured by nnU-Net [8]; nnU-Net reduces the design choices to the very essential ones and automatically infers these choices using a set of heuristic rules.

The U-Net CNN architecture configured by nnU-net follows a U-shaped topology with N + 1 encoding blocks and N decoding blocks. Each encoding block consists of two convolutional layers with leaky rectified linear unit (LReLU) activation (neg. slope 1e-2), and instance normalisation [157]. In all but the first encoder block, strided convolutions are used to perform downsampling. Thirty-two feature maps are output by convolutional layers in the first encoding block, with feature maps doubling in number in each subsequent encoding block. To limit the final model size, the number of feature maps are capped at a maximum of 320. The number of encoder blocks and the sizes and strides of the kernels in each encoder block are determined automatically based on the size and voxel spacing of the input image. Downsampling is terminated at the point at which further downsampling would reduce feature maps to less four voxels in any dimension. For anisotropic 3D data, high resolution axes are downsampled separately until their voxel resolution is within a factor of two of the lower resolution axis. Subsequently, all axes are downsampled simultaneously. Upsampling deconvolution operations are used in the decoding blocks, which receive semantic information from the last encoding block and higher resolution feature maps from skip connections that join encoder blocks to decoder blocks with feature maps of the same spatial size.

6.2.1.2 Inference with user mouse-clicks

In DeepGrow, a user e.g., a radiologist, can click in an area of the predicted segmentation where the CNN has made an error. Mouse-clicks can be placed in false-negative areas of the CNN prediction (henceforth referred to as "foreground clicks") to encourage prediction of the foreground class and mouse-clicks can be placed in false-positive areas of the CNN prediction (henceforth referred to as "background clicks") to discourage prediction of the foreground class in background regions. Foreground and background clicks are converted into foreground and background click maps, which are concatenated with the input image, in the channel dimension, prior to CNN processing. Foreground and background click maps have the same spatial size as the input image and are zero everywhere except voxels corresponding to a click, which take the value one. Prior to CNN processing, the foreground and background click maps are smoothed using a Gaussian filter and normalised to $[0, 1]$ range as in the work by Maninis et al [171]. A schematic representation of inference using DeepGrow is shown in Figure 6.1.

6.2.1.3 Training strategies

DeepGrow can be trained with mouse-clicks provided by a user or using simulated clicks that mimic user mouse-clicks. While training DeepGrow with mouse-clicks provided by a user would incorporate clicks during training that are the most representative of clicks that are likely to be provided during inference, training with mouse-clicks would substantially increase the time required for training, therefore a strategy for simulating clicks was utilised by Sakinis et al. for training DeepGrow. In summary, for each training iteration, an initial simulated click is placed within the structure of interest, to initiate the generation of a segmentation by the CNN. Then, up to K further simulated clicks are added in foreground or background regions to correct false-negative or false positive voxels, respectively. Finally, a backpropagation step is used to update CNN parameters.

When training DeepGrow, the first simulated click is always applied, while the subsequent K simulated clicks happen according to a probability, p_i, which decreases linearly:

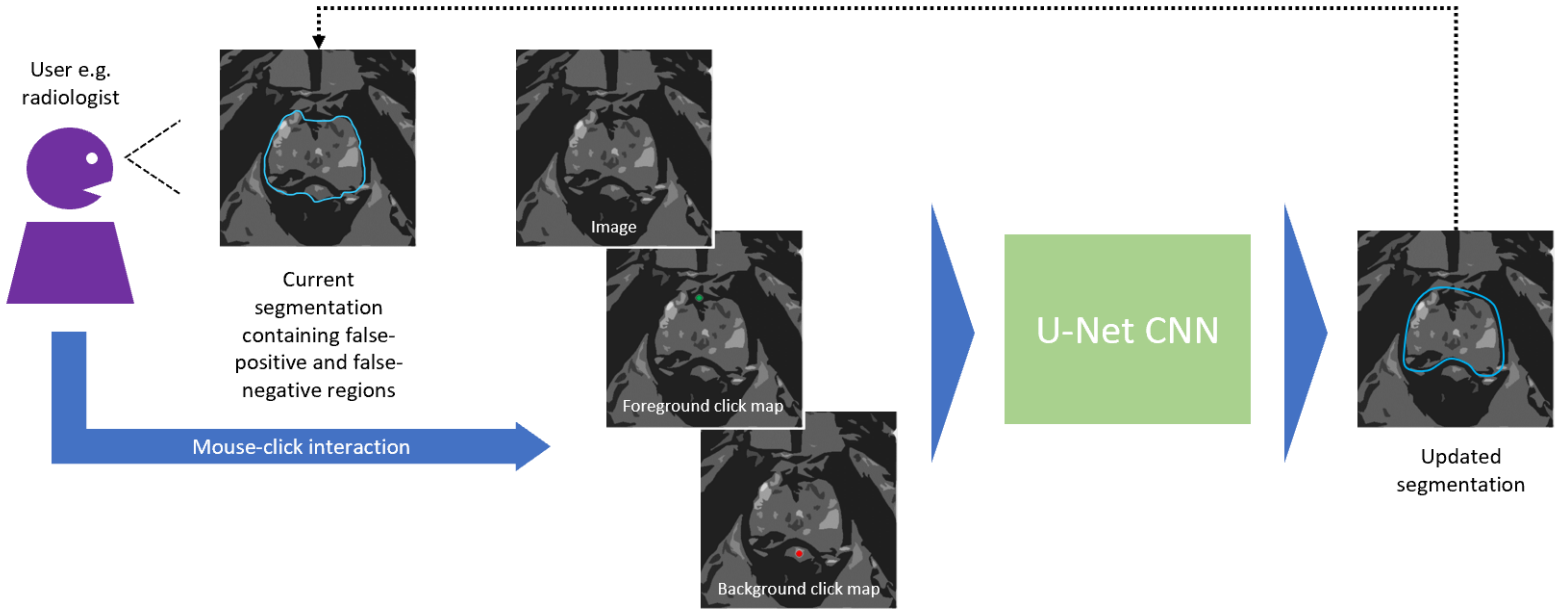


Figure 6.1: Schematic representation of interactive segmentation using DeepGrow. The user, who possesses domain knowledge, observes the current segmentation which may contain false-positive and/or false-negative regions. Subsequently, foreground clicks and background clicks are applied as necessary to produce an updated segmentation.

$$p_i = \frac{K+1-i}{K+1}, \quad i = 1, \ldots, K. \tag{6.1}$$

Therefore, the actual number of simulated clicks beyond the first simulated click will be equal to some value t, such that $1 \leq t \leq K$.

After each simulated click, an intermediate prediction, P, will be output by the CNN through a forward pass with no backpropagation step. P is used to determine the location of the next simulated click, through the computation of a disparity map, D, which is the difference between P and the ground-truth segmentation G:

$$D = G - P = \begin{cases} +1 & \text{if } G = 1 \text{ and } P = 0, \\ -1 & \text{if } G = 0 \text{ and } P = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6.2}$$

The disparity map D can be decomposed into a positive disparity map, D^+, which considers only false-negative regions of D and a negative disparity map, D^-, which considers only false-positive regions of D. The decision of whether a simulated click is placed within a false-negative region or a false-positive region is based on the voxel sums of D^+ and D^-. A simulated click is placed in a false-negative region if the voxel sum of $D^+ \geq D^-$, while a simulated click is placed in a false-positive region if the voxel sum of $D^- > D^+$.

The exact voxel location of a simulated click within a false-negative or false-positive region is determined probabilistically. The approach outlined by Sakinis et al., which we denote "Click-Loc-Method-I", determines the probability of a voxel in a false-negative or false-positive region to receive a simulated click using the volume of the false-negative or false-positive region and the centrality of the voxel within the false-negative or false-positive region. An alternative to "Click-Loc-Method-I" is proposed in this work, hereby referred to as "Click-Loc-Method-II", in which the prediction error is used to inform the voxel location of the simulated click.

- **Click-Loc-Method-I:** The approach taken by Sakinis et al. assigns the high-

est probability of receiving a simulated click to the voxels at the centre of large false-negative or false-positive regions. Using the positive disparity map D^+ and the negative disparity map D^-, Chamfer distance maps, C^+ and C^-, are computed, where the Chamfer distance value for a voxel is the minimum distance to a voxel whose value is zero i.e., the centre of the largest false-negative or false-positive region will have the highest distance value. Assuming 3D input, the probability of a voxel with coordinate (x, y, z) receiving a simulated foreground click or background click is described by R^+ and R^-, respectively, where:

$$R^+(x,y,z) = \frac{\exp(C^+(x,y,z)) - 1}{\Sigma_x \Sigma_y \Sigma_z \exp(C^+(x,y,z)) - 1}, \tag{6.3}$$

$$R^-(x,y,z) = \frac{\exp(C^-(x,y,z)) - 1}{\Sigma_x \Sigma_y \Sigma_z \exp(C^-(x,y,z)) - 1}. \tag{6.4}$$

It should be noted that the first simulated click is generated assuming an initial intermediate prediction P that is zero everywhere. This gives a positive disparity map $D^+ \equiv G$, whose Chamfer distance map C^+ will be greatest at the centre of the foreground object to be segmented.

- **Click-Loc-Method-II:** Rather than using the size of false-negative or false-positive regions to direct simulated click placement, it may be beneficial to direct simulated clicks towards the false-negative or false-positive regions that have the highest prediction error. In this training paradigm, clicks will be directed towards regions in the training set images that are most difficult to classify, which may be representative of difficult-to-classify regions in the test set images/new images to be inferred. Assuming 3D input, the probability of a voxel with coordinate (x, y, z) receiving a simulated foreground or background click is described by R^+ and R^-, respectively, where:

$$R^{+}(x,y,z) = |G - S|, \tag{6.5}$$

$$R^{-}(x,y,z) = |S - G|, \tag{6.6}$$

where G is the ground-truth segmentation and S is the sigmoid probability map output by the CNN, prior to binarisation.

6.2.2 DeepEdit

As described earlier, during each training iteration of DeepGrow, t clicks are simulated, where $1 \leq t \leq K+1$. As a result, DeepGrow is tuned to perform inference optimally when the user makes at least one mouse-click. However, it would be desirable to perform inference fully-automatically initially i.e., without mouse-clicks, followed by "editing" of the segmentation using mouse-clicks, only if required. Therefore, "DeepEdit" has been proposed in MONAI Label. DeepEdit follows DeepGrow in every aspect, except the following during training: for each training iteration, backpropagation will be invoked with zero simulated clicks with some probability pDE or with t simulated clicks with probability $1 - pDE$, where pDE is a hyperparameter to be optimised through experimentation. It should be noted that setting $pDE = 0$ is a special case in which DeepEdit and DeepGrow are equivalent.

6.3 Experimental setup

In this section, we describe the experimental details for the two segmentation tasks investigated in this work, namely, whole prostate segmentation and prostatic lesion segmentation. Experiments were run using the PROSTATEx Challenges training dataset, [44], hereby referred to as the PROSTATEx dataset; a detailed description of the PROSTATEx dataset is given in Chapter 3.

The whole prostate segmentation task concerns the segmentation of the prostate on T2WI. Eleven patients from the PROSTATEx dataset were excluded due to inconsistencies between T2WI and the ground-truth segmentations, leaving a total of 193 patients for use in experiments. The prostatic lesion segmentation

task concerns the segmentation of lesions within the prostate using T2WI, ADC map, and computed high b-value DWI. Since our experiments were conducted using the PROSTATEx dataset, we used the PROSTATEx definition of a lesion i.e., a prostatic lesion is defined as any area of suspicion attributed a PI-RADS score by the expert radiologist (Jelle Barentz) who read and reported PROSTATEx dataset cases; all lesions in the PROSTATEx dataset were scored PI-RADS ≥ 2. Four patients from the PROSTATEx dataset were excluded as they contained no contoured lesions in the ground-truth, leaving a total of 200 patients for use in experiments.

6.3.1 Experiments

For both tasks, experiments were conducted to choose between the following training settings: (i) Dice loss [74] vs a hybrid loss composed of Dice loss and Focal loss [154], (ii) Click-Loc-Method-I vs Click-Loc-Method-II, and (iii) DeepGrow vs DeepEdit with pDE = 0.25 (DeepEdit-0.25) vs DeepEdit with pDE = 0.5 (DeepEdit-0.5).

Ten-fold cross-validations were performed for both tasks. Segmentation quality was assessed using the Dice coefficient. As in the work of Sakinis et al., segmentation performance at inference time was assessed using simulated clicks, as opposed to user mouse-clicks, to objectively assess how segmentation quality improves as clicks are added; segmentation performance was assessed at 0, 1, 2, 5, 10, and 15 simulated inference clicks. To account for variability in simulated inference click placement, the presented results are an average of three repetitions.

6.3.2 Experimental pipeline and settings

6.3.2.1 Whole prostate segmentation task

T2WI were pre-processed by resampling to a common resolution of 0.5 mm $\times$ 0.5 mm $\times$ 3.0 mm, normalisation using per-image whitening, and cropping/padding to a common size of 320 $\times$ 320 $\times$ 32. Using the common size and resolution described above, nnU-Net configured a 3D U-Net with seven encoding blocks and six decoding blocks for use in DeepGrow and DeepEdit. Kernel sizes and strides were determined automatically by nnU-Net, as shown in Table 6.1.

Encoding block	Kernel size	Stride
1	(3,3,1)	(1,1,1)
	(3,3,1)	(1,1,1)
2	(3,3,1)	(2,2,1)
	(3,3,1)	(1,1,1)
3	(3,3,3)	(2,2,1)
	(3,3,3)	(1,1,1)
4	(3,3,3)	(2,2,2)
	(3,3,3)	(1,1,1)
5	(3,3,3)	(2,2,2)
	(3,3,3)	(1,1,1)
6	(3,3,3)	(2,2,2)
	(3,3,3)	(1,1,1)
7	(3,3,3)	(2,2,1)
	(3,3,3)	(1,1,1)

Table 6.1: Kernel size and stride settings configured by nnU-Net for 3D U-Net.

Following CNN processing, the CNN output was transformed to the original T2WI size and resolution using padding and resampling. Subsequently, the raw activations were converted into probabilities using a sigmoid function, followed by binarisation using a threshold equal to 0.5.

DeepGrow and DeepEdit were trained with a learning rate equal to 0.0001, batch size equal to one, and Adam optimisation [148]. During training, in-plane rotation (range: -1 rad to 1 rad), scaling (range: -30% to 40%), and horizontal flip (probability: 0.5) augmentations were applied on-the-fly. In addition, Gaussian noise ($\mu = 0$, $\sigma = 0.05$, probability: 0.15) and intensity scaling (range: -30% to 30%) were also applied on-the-fly. All training runs used a maximum of 10 simulated clicks per training iteration. Simulated clicks were smoothed with a Gaussian kernel with $(\sigma_x, \sigma_y, \sigma_z) = (4, 4, 2/3)$ to account for anisotropic voxel resolution; the kernel settings were determined with reference to the work of Sakinis et al., where a Gaussian kernel with $(\sigma_x, \sigma_y) = (2, 2)$ was used for 2D images with 1 mm isotropic voxel resolution.

DeepGrow was trained for 75 epochs, DeepEdit-0.25 was trained for 100 epochs, and DeepEdit-0.5 was trained for 150 epochs. The maximum epochs for

DeepEdit-0.25 and DeepEdit-0.5 were set to allow the same number of total simulated clicks during training as DeepGrow. Model parameters were saved at an interval of five epochs. The model parameters used for inference were those that gave the highest mean Dice coefficient at five simulated clicks on a validation set; the validation set was chosen at random from the set of nine folds not set-aside for inference.

6.3.2.2 Prostatic lesion segmentation task

For prostatic lesion segmentation, three input modalities were concatenated as input, namely, T2WI, ADC map, and computed high b-value DWI.

A b-value, b = 2000, was selected for computing high b-value DWI as in Verma et al. [28]; computed b2000 (Cb2000) DWI were generated using DWI acquired at lower b-values, extrapolated by assuming a monoexponential model for the per-voxel observed signal [29]. ADC map and Cb2000 DWI were registered to T2WI to account for voluntary/involuntary patient movement between acquisitions and differences in resolution. Registrations were run using NiftyReg (version 1.3; `https://github.com/KCL-BMEIS/niftyreg`), following the approach described in [151]: the registration of ADC map to T2WI employed default parameters for affine registration via symmetric block-matching [158]; subsequently, non-rigid free-form deformation (FFD) registration [135] used a Gaussian kernel with standard deviation equal to 5 mm for local normalised correlation coefficient (LNCC) calculation, control point spacing equal to 10 mm, and bending energy constraint equal to 0.1; following ADC map registration, Cb2000 DWI were registered to T2WI using the same transformation.

T2WI and Cb2000 DWI were normalised by dividing voxel intensities by the interquartile mean of central gland (CG) voxel intensities [151]; CG masks used to identify CG voxels were generated by AutoProstate (see Chapter 5). ADC maps were not normalised as they contain a quantitative measurement.

T2WI, ADC map and Cb2000 DWI, and whole prostate and CG masks output by AutoProstate (see Chapter 5) were resampled to a common resolution of 0.5 mm $\times$ 0.5 mm $\times$ 3 mm. Then, whole prostate masks were used to crop the prostate

region on all MR modalities; a margin was applied in each direction to reduce the likelihood of prostate tissue being discarded. Next, a cropping/padding transformation was used to ensure a common spatial size of $256 \times 256 \times 32$.

Using the common size and resolution described above, nnU-Net configured the same 3D U-Net architecture as described by Table 6.1. The hyperparameters used to train DeepGrow and DeepEdit in the whole prostate segmentation task were also used in the prostatic lesion segmentation task, apart from training epochs which were increased through observation of the training and validation set losses. For prostatic lesion segmentation, DeepGrow was trained for 150 epochs. The same heuristic as used in the whole prostate segmentation task was used to determine the maximum training epochs for DeepEdit-0.25 and DeepEdit-0.5. As in the whole prostate segmentation task, model parameters were saved at an interval of five epochs. The model parameters used for inference were those that gave the highest mean Dice coefficient at five simulated clicks on a validation set; the validation set was chosen at random from the set of nine folds not set-aside for inference.

As in the whole prostate segmentation task, the CNN output was transformed to the original T2WI size and resolution with padding and resampling. Subsequently, the raw activations were converted into probabilities using a sigmoid function, followed by binarisation using a threshold equal to 0.5.

6.4 Results

This section presents the results for whole prostate segmentation and prostatic lesion segmentation. Examples of the 3D Slicer interface for the whole prostate segmentation application and prostatic lesion segmentation application are shown in Figures 6.2 and 6.3, respectively; the 3D Slicer interface for MONAI Label applications is enabled through an extension that is available for download from the 3D Slicer extension server.

6.4.1 Whole prostate segmentation task

A comparison of DeepGrow trained using Dice loss and a hybrid loss composed of Dice loss and Focal loss is shown in Table 6.2. DeepGrow trained using Dice

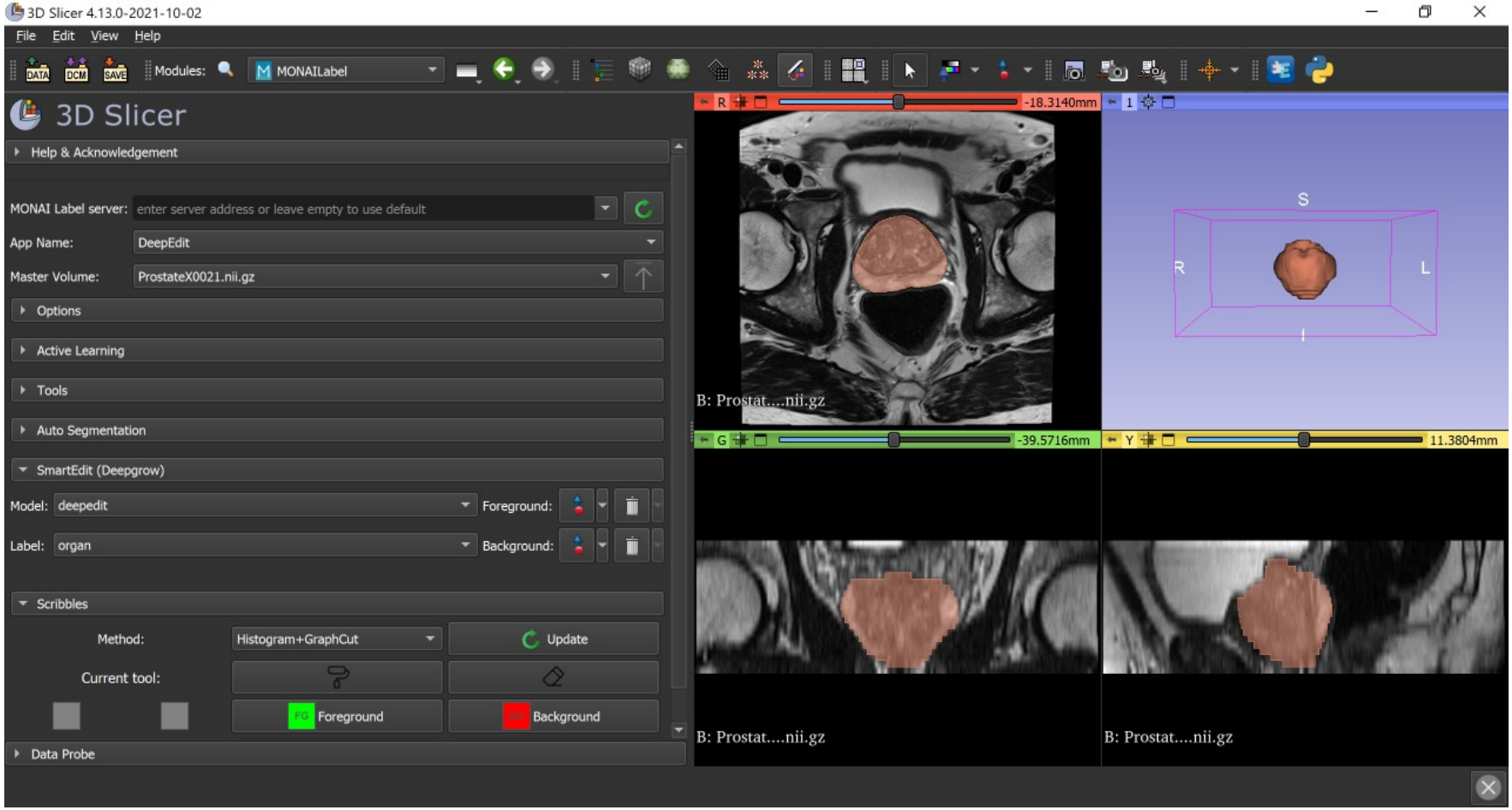


Figure 6.2: 3D Slicer interface for the whole prostate segmentation MONAI Label DeepEdit application.

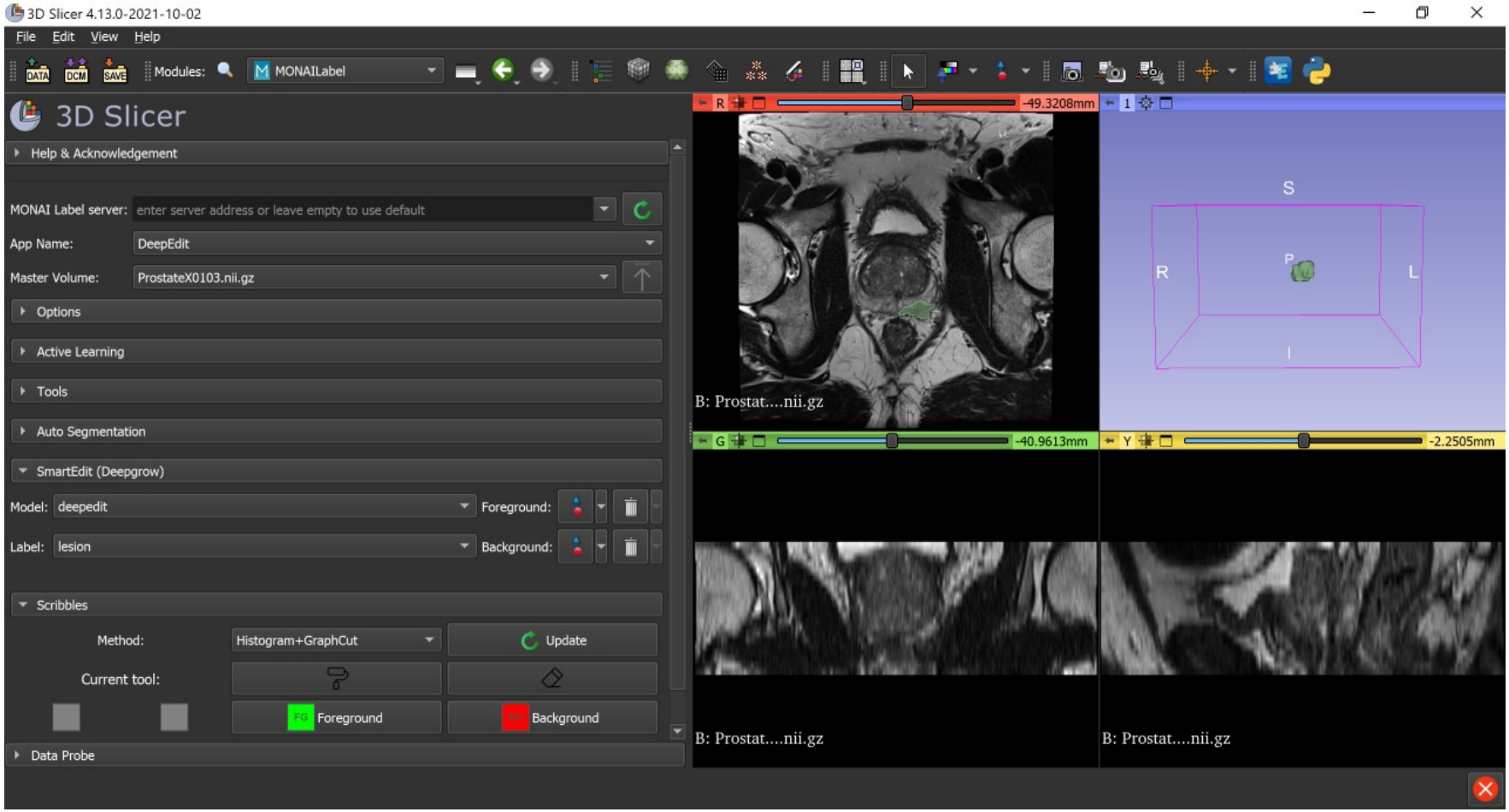


Figure 6.3: 3D Slicer interface for the prostatic lesion segmentation MONAI Label DeepEdit application.

loss and the hybrid loss achieved click-free mean Dice scores of 0.907 ± 0.041 and 0.908 ± 0.052, respectively; the difference was not statistically significant. However, at 1 to 15 simulated inference clicks, DeepGrow trained with Dice loss achieved consistently higher mean Dice scores with statistical significance.

A comparison of DeepGrow trained using Click-Loc-Method-I and Click-Loc-Method-II is shown in Table 6.3. The comparison was unanimous in that Click-Loc-Method-I achieved consistently higher mean Dice scores at each number of simulated inference clicks. However, statistically significant differences were only observed at 5 and 15 simulated inference clicks.

A comparison of DeepGrow, DeepEdit-0.25, and DeepEdit-0.5 is shown in Table 6.4. Furthermore, the distributions of Dice scores are shown in Figure 6.4. DeepEdit-0.5 gave the highest click-free mean Dice score of 0.908, while DeepGrow gave the highest mean Dice scores at 1 to 15 simulated inference clicks. While the Friedman test showed a significant difference between models for each number of simulated clicks, the pairwise comparison showed superiority of a model above both other models, for DeepGrow at 10 and 15 simulated inference clicks only. Alongside the increase in mean Dice with the increase in simulated inference clicks, the minimum Dice also increased, as seen in Figure 6.4. For DeepGrow, DeepEdit-0.25, and DeepEdit-0.5, the Dice score the minimum Dice score increased from 0.554, 0.339, and 0.417 to 0.763, 0.456, and 0.551, to 0.799, 0.706, and 0.783, for 0, 5, and 15 simulated inference clicks, respectively.

6.4.2 Prostatic lesion segmentation task

A comparison of DeepGrow trained using Dice loss and a hybrid loss composed of Dice loss and Focal loss is shown in Table 6.5. DeepGrow trained using Dice loss and the hybrid loss achieved click-free mean Dice scores of 0.166 ± 0.254 and 0.177 ± 0.266, respectively; the difference was not statistically significant. At 1 to 15 simulated inference clicks, DeepGrow trained with Dice loss achieved consistently higher mean Dice scores, though statistical significance was only observed at 10 and 15 simulated inference clicks.

A comparison of DeepGrow trained using Click-Loc-Method-I and Click-Loc-

	Number of simulated clicks					
	0	**1**	**2**	**5**	**10**	**15**
Dice loss	0.907 ± 0.041	$\mathbf{0.915} \pm 0.035$	$\mathbf{0.919} \pm 0.032$	$\mathbf{0.926} \pm 0.024$	$\mathbf{0.932} \pm 0.020$	$\mathbf{0.936} \pm 0.018$
Dice + Focal loss	$\mathbf{0.908} \pm 0.052$	0.911 ± 0.050	0.914 ± 0.048	0.920 ± 0.044	0.928 ± 0.036	0.933 ± 0.034
P-value	0.385	0.024^*	$< 0.001^*$	$< 0.001^*$	0.001^*	0.001^*

Table 6.2: Whole prostate segmentation mean Dice scores $\pm$ one standard deviation, calculated over the 193 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow trained using Dice loss and a hybrid loss composed of Dice loss and Focal loss. Click-Loc-Method-I was fixed for comparison. The highest mean Dice in each column is shown in bold. In addition, P-values, calculated using the Wilcoxon signed-rank test that are less than 0.05 are indicated with an asterisk.

	Number of simulated clicks					
	0	**1**	**2**	**5**	**10**	**15**
Click-Loc-Method-I	**0.907** $\pm$ 0.041	**0.915** $\pm$ 0.035	**0.919** $\pm$ 0.032	**0.926** $\pm$ 0.024	**0.932** $\pm$ 0.020	**0.936** $\pm$ 0.018
Click-Loc-Method-II	0.905 $\pm$ 0.047	0.913 $\pm$ 0.043	0.916 $\pm$ 0.040	0.923 $\pm$ 0.034	0.930 $\pm$ 0.026	0.935 $\pm$ 0.020
P-value	0.265	0.206	0.056	0.012*	0.179	0.004*

Table 6.3: Whole prostate segmentation mean Dice scores $\pm$ one standard deviation, calculated over the 193 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow trained using Click-Loc-Method-I and Click-Loc-Method-II. Dice loss was fixed for comparison. The highest mean Dice in each column is shown in bold. In addition, P-values, calculated using the Wilcoxon signed-rank test that are less than 0.05 are indicated with an asterisk.

	Number of simulated clicks					
	0	**1**	**2**	**5**	**10**	**15**
Mean Dice ± one standard deviation						
DeepGrow	0.907 ± 0.041	**0.915** ± 0.035	**0.919** ± 0.032	**0.926** ± 0.024	**0.932** ± 0.020	**0.936** ± 0.018
DeepEdit-0.25	0.908 ± 0.054	0.912 ± 0.049	0.915 ± 0.048	0.921 ± 0.041	0.929 ± 0.026	0.933 ± 0.024
DeepEdit-0.5	**0.908** ± 0.046	0.911 ± 0.044	0.913 ± 0.042	0.919 ± 0.035	0.926 ± 0.028	0.931 ± 0.022
Friedman ranks						
DeepGrow	1.94	2.09	**2.24**	**2.37**	**2.35**	**2.38**
DeepEdit-0.25	**2.18**	**2.13**	2.08	2.07	2.04	1.98
DeepEdit-0.5	1.89	1.78	1.69	1.56	1.61	1.64
P-value	0.010*	0.001*	< 0.001*	< 0.001*	< 0.001*	< 0.001*
Dunn's pairwise test with Bonferroni correction P-value						
DeepGrow, DeepEdit-0.25	0.059	1.000	0.347	0.110	0.006*	< 0.001*
DeepGrow, DeepEdit-0.5	1.000	0.008*	< 0.001*	< 0.001*	< 0.001*	< 0.001*
DeepEdit-0.25, DeepEdit-0.5	0.013*	0.002*	< 0.001*	< 0.001*	< 0.001*	0.002*

Table 6.4: Whole prostate segmentation mean Dice scores ± one standard deviation for the 193 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow, DeepEdit-0.25, and DeepEdit-0.5. Dice loss and Click-Loc-Method-I were fixed for comparison. The highest mean Dice in each column is shown in bold. The Friedman test for multiple comparisons was used to assess differences between groups, followed by Dunn's pairwise test with Bonferroni correction. P-values less than 0.05 are indicated with an asterisk.

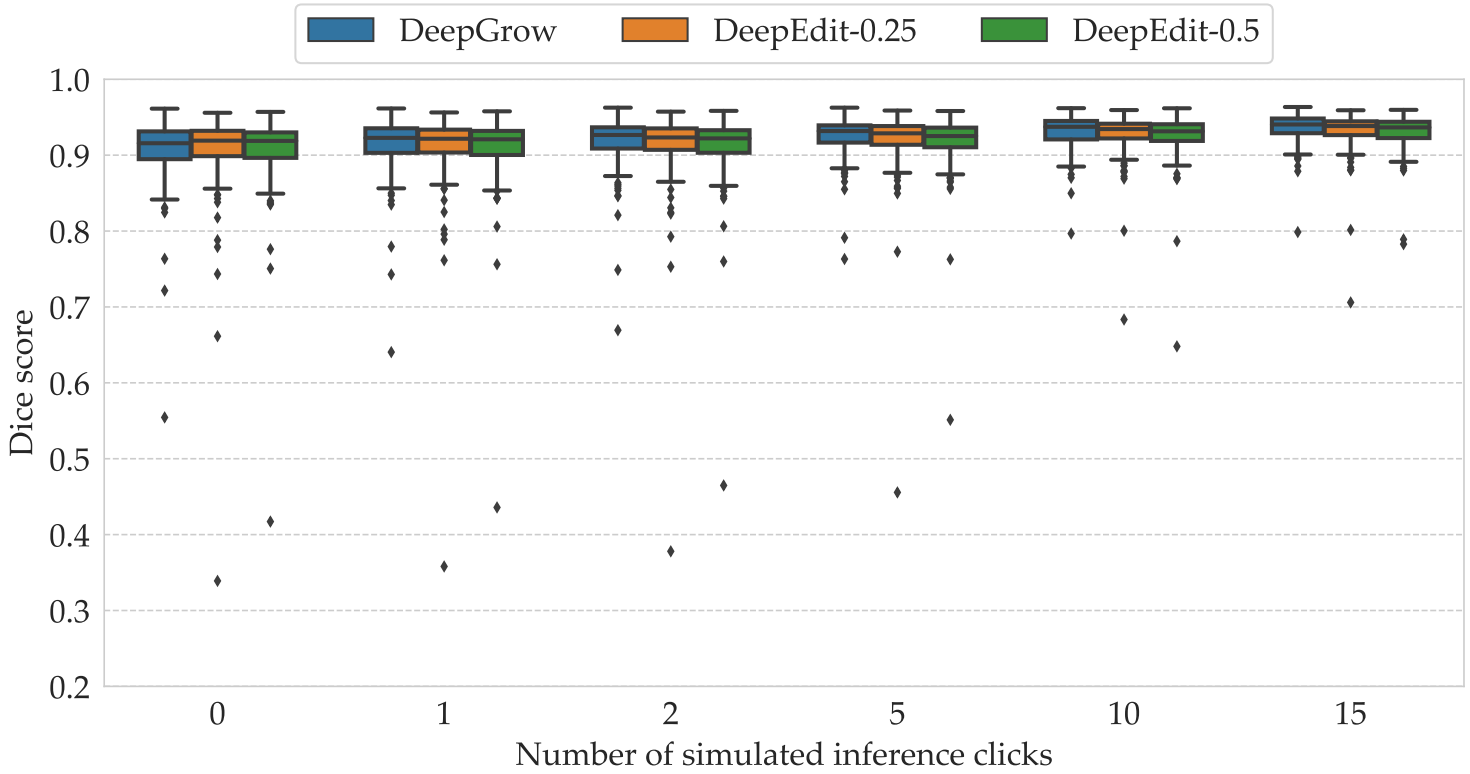


Figure 6.4: Whole prostate segmentation Dice score box plots for the 193 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow, DeepEdit-0.25, and DeepEdit-0.5. Dice loss and Click-Loc-Method-I were fixed for comparison.

	Number of simulated clicks					
	0	**1**	**2**	**5**	**10**	**15**
Dice loss	0.166 ± 0.254	$\mathbf{0.527} \pm 0.166$	$\mathbf{0.592} \pm 0.135$	$\mathbf{0.670} \pm 0.111$	$\mathbf{0.723} \pm 0.095$	$\mathbf{0.749} \pm 0.089$
Dice + Focal loss	$\mathbf{0.177} \pm 0.266$	0.512 ± 0.176	0.579 ± 0.149	0.660 ± 0.122	0.711 ± 0.107	0.741 ± 0.093
P-value	0.778	0.062	0.065	0.161	0.010^*	0.008^*

Table 6.5: Prostatic lesion segmentation mean Dice scores $\pm$ one standard deviation, calculated over the 200 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow trained using Dice loss and a hybrid loss composed of Dice loss and Focal loss. Click-Loc-Method-I was fixed for comparison. The highest mean Dice in each column is shown in bold. In addition, P-values, calculated using the Wilcoxon signed-rank test, less than 0.05 are indicated with an asterisk.

Method-II is shown in Table 6.6. For training DeepGrow, Click-Loc-Method-II outperformed Click-Loc-Method-I for 0 simulated clicks, and vice-versa for 1 to 25 simulated clicks. However, a statistically significant difference was only observed at 10 simulated inference clicks.

A comparison of DeepGrow, DeepEdit-0.25, and DeepEdit-0.5 is shown in Table 6.7. Furthermore, the distributions of Dice scores are shown in Figure 6.5. At 0 simulated inference clicks, DeepGrow, DeepEdit-0.25, and DeepEdit-0.5 achieved mean Dice scores of 0.166 ± 0.254, 0.268 ± 0.271, and 0.272 ± 0.266, respectively. The Friedman test indicated a significant difference between model performances, while the pairwise test showed that the increase in Dice score from DeepGrow to DeepEdit-0.25 was statistically significant, but the increase in Dice score from DeepEdit-0.25 to DeepEdit-0.5 was not. For 1 to 15 simulated inference clicks, DeepGrow consistently achieved the highest mean Dice scores with 0.527 ± 0.166, 0.592 ± 0.135, 0.670 ± 0.111, 0.723 ± 0.095, and 0.749 ± 0.089 at 1, 2, 5, 10, and 15 simulated inference clicks, though the differences in Dice as compared to the DeepEdit models were only statistically significant at 2, 5, 10, and 15 simulated inference clicks, as shown by the pairwise test. A wide distribution of Dice scores was observed for all models at 0 simulated inference clicks, as shown in Figure 6.5; all models had a minimum Dice score of 0.000 and maximum Dice scores of 0.820, 0.816, and 0.810 were observed for DeepGrow, DeepEdit-0.25, and DeepEdit-0.5, respectively. While there was a narrowing of the distribution of Dice scores with the addition of simulated inference clicks, low Dice scores of 0.190, 0.257, and 0.053 for DeepGrow, DeepEdit-0.25, and DeepEdit-0.5, respectively, remained at 15 simulated inference clicks.

6.5 Discussion

Fully-automatic segmentation methods for medical image segmentation have not yet reached the desired robustness for clinical use [45]. Therefore, interactive segmentation methods are being developed [45, 83]. In this work, we investigated interactive segmentation of the whole prostate and prostatic lesions. Click-

	Number of simulated clicks					
	0	**1**	**2**	**5**	**10**	**15**
Click-Loc-Method-I	0.166 ± 0.254	$\mathbf{0.527} \pm 0.166$	$\mathbf{0.592} \pm 0.135$	$\mathbf{0.670} \pm 0.111$	$\mathbf{0.723} \pm 0.095$	$\mathbf{0.749} \pm 0.089$
Click-Loc-Method-II	$\mathbf{0.187} \pm 0.272$	0.525 ± 0.163	0.582 ± 0.144	0.661 ± 0.115	0.712 ± 0.104	0.741 ± 0.095
P-value	0.095	0.938	0.066	0.242	0.032^{*}	0.076

Table 6.6: Prostatic lesion segmentation mean Dice scores $\pm$ one standard deviation, calculated over the 200 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow trained using Click-Loc-Method-I and Click-Loc-Method-II. Dice loss was fixed for comparison. The highest mean Dice in each column is shown in bold. In addition, p-values, calculated using the Wilcoxon signed-rank test, less than 0.05 are indicated with an asterisk.

	Number of simulated clicks					
	0	**1**	**2**	**5**	**10**	**15**
Mean Dice ± one standard deviation						
DeepGrow	0.166 ± 0.254	**0.527** ± 0.166	**0.592** ± 0.135	**0.670** ± 0.111	**0.723** ± 0.095	**0.749** ± 0.089
DeepEdit-0.25	0.268 ± 0.271	0.498 ± 0.174	0.552 ± 0.156	0.632 ± 0.130	0.697 ± 0.114	0.729 ± 0.101
DeepEdit-0.5	**0.272** ± 0.266	0.453 ± 0.197	0.502 ± 0.184	0.592 ± 0.163	0.663 ± 0.145	0.697 ± 0.139
Friedman ranks						
DeepGrow	1.60	**2.24**	**2.39**	**2.41**	**2.36**	**2.33**
DeepEdit-0.25	2.19	2.12	1.99	1.94	2.05	2.04
DeepEdit-0.5	**2.21**	1.65	1.63	1.66	1.60	1.63
P-value	$< 0.001^*$	$< 0.001^*$	$< 0.001^*$	$< 0.001^*$	$< 0.001^*$	$< 0.001^*$
Dunn's pairwise test with Bonferroni correction P-value						
DeepGrow, DeepEdit-0.25	$< 0.001^*$	0.690	$< 0.001^*$	$< 0.001^*$	0.006^*	0.011^*
DeepGrow, DeepEdit-0.5	$< 0.001^*$	$< 0.001^*$	$< 0.001^*$	$< 0.001^*$	$< 0.001^*$	$< 0.001^*$
DeepEdit-0.25, DeepEdit-0.5	1.000	$< 0.001^*$	0.001^*	0.015^*	$< 0.001^*$	$< 0.001^*$

Table 6.7: Prostatic lesion segmentation mean Dice scores ± one standard deviation for the 200 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow, DeepEdit-0.25, and DeepEdit-0.5. Dice loss and Click-Loc-Method-I were fixed for comparison. The highest mean Dice in each column is shown in bold. The Friedman test for multiple comparisons was used to assess differences between groups, followed by Dunn's pairwise test with Bonferroni correction. P-values less than 0.05 are indicated with an asterisk.

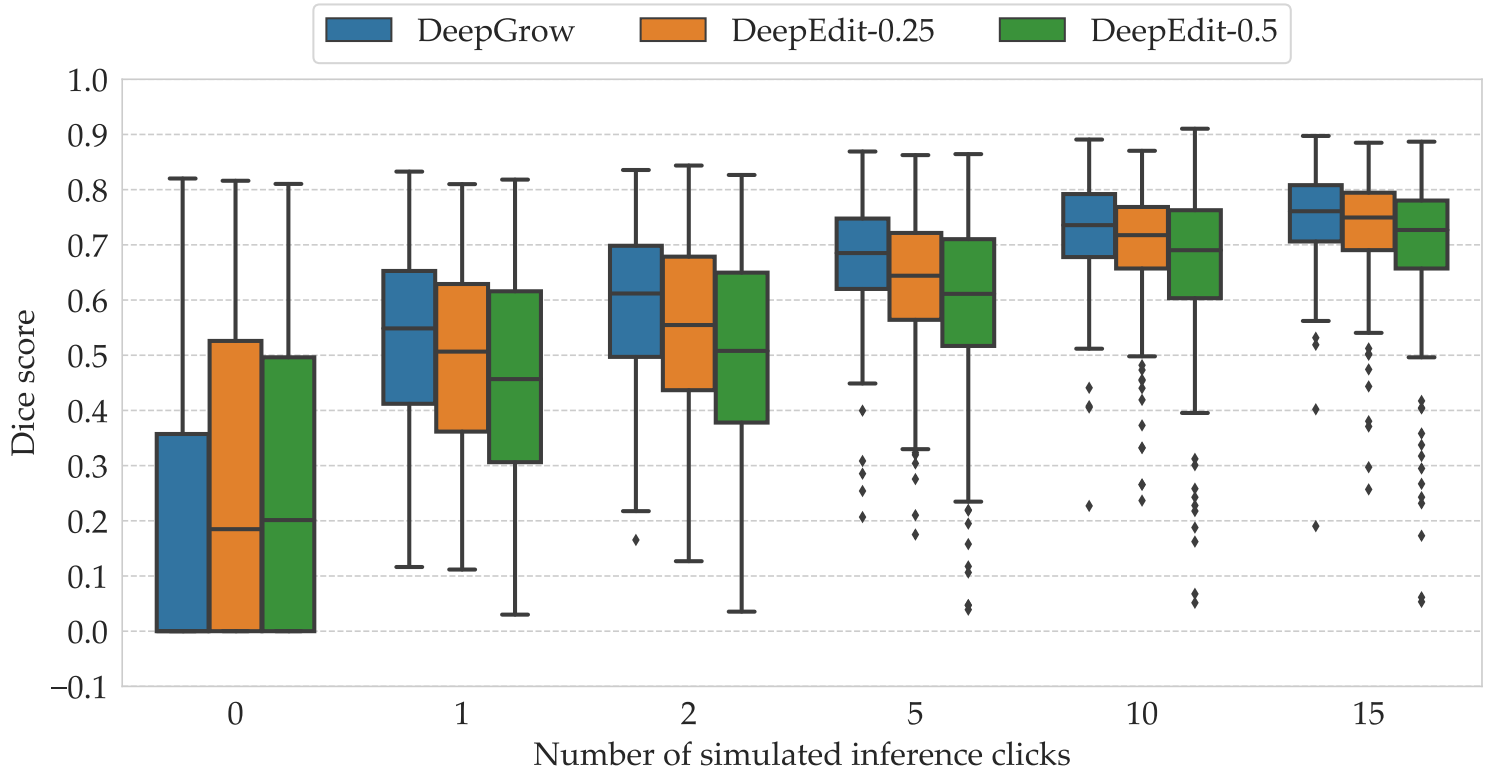


Figure 6.5: Prostatic lesion segmentation Dice score distributions for the 200 PROSTATEx dataset patients used in the ten-fold cross-validation, for DeepGrow, DeepEdit-0.25, and DeepEdit-0.5. Dice loss and Click-Loc-Method-I were fixed for comparison.

based interactive segmentation applications for both tasks were built using MONAI Label (`https://github.com/Project-MONAI/MONAILabel`). Experiments were run using DeepGrow, which is a click-based interactive segmentation method presented in a work by Sakinis et al. [45], and DeepEdit which is a modification of DeepGrow, available in MONAI Label, which intends to improve the initial click-free segmentation, whilst retaining good click-based segmentation editing performance. For both tasks, ten-fold cross-validations were run using the publicly available PROSTATEx dataset.

For both tasks, experiments were run to choose between the following training settings: (i) Dice loss vs a hybrid loss composed of Dice loss and Focal loss, (ii) Click-Loc-Method-I vs Click-Loc-Method-II, and (iii) DeepGrow vs DeepEdit-0.25 vs DeepEdit with DeepEdit-0.5.

For whole prostate segmentation, the optimal training paradigm was found to be DeepGrow trained with Dice loss and Click-Loc-Method-I. DeepGrow achieved a click-free mean Dice score of 0.907 ± 0.041, which increased to 0.926 ± 0.024 at 5 simulated inference clicks and to 0.936 ± 0.018 at 15 simulated inference clicks. In addition, at 0, 5, and 15 simulated inference clicks, the minimum Dice score increased from 0.554, to 0.763, to 0.799. On inspection, the case with the lowest Dice score at 0 simulated inference clicks, PROSTATEx0192, was found to have bias field corruption on T2WI. Pleasingly, despite the bias field corruption, the Dice score for PROSTATEx0192 increased to 0.791 with 5 simulated clicks, and to 0.879 with 15 simulated clicks. At 15 simulated inference clicks, the case with the lowest Dice score changed to PROSTATEx0200, where a Dice score of 0.799 was observed. On inspection, bladder infiltration of the prostate was found to be reason for limited segmentation quality despite a high number of simulated inference clicks.

The MONAI Label application for click-based interactive whole prostate segmentation has uses in diagnosis, active surveillance, and treatment planning. An accurate whole prostate segmentation can be used to compute prostate volume, as an alternative to the ellipsoid formula which is a crude approximation that is used clinically [152]. Improvements in the accuracy of the calculated prostate volume

will be reflected in downstream calculation of the PSAd, which is an important biomarker for diagnosing CSPCa [24, 25], as well as for monitoring patients placed on active surveillance [172]. An accurate whole prostate segmentation also allows better determination of the prostate boundary. Knowledge of the prostate boundary, including areas of extraprostatic extension (EPE), is important for planning and conducting whole-gland prostatectomy, radiotherapy, and cryotherapy, as well as for determining seed placement for brachytherapy.

For prostatic lesion segmentation, Dice loss and Click-Loc-Method-I were found to be optimal for training DeepGrow. In the comparison of DeepGrow to DeepEdit-0.25 and DeepEdit-0.5, a performance trade-off was observed. DeepEdit-0.5 produced the highest click-free mean Dice score of 0.272, followed by DeepEdit-0.25 with 0.268, followed by DeepGrow with 0.166. However, DeepGrow produced the highest mean Dice scores at 1 to 15 simulated inference clicks with mean Dice scores of 0.527, 0.670, and 0.749 at 1, 5, and 15 simulated inference clicks. It can be observed that the performance difference between DeepGrow and DeepEdit is more pronounced for the prostatic lesion segmentation task than for the whole prostate segmentation task. We postulate that this is due to the increased difficulty of prostatic lesion segmentation as compared to whole prostate segmentation, due to the variability in size and conspicuity of prostatic lesions. The choice of DeepGrow or DeepEdit should be made with the performance trade-off between DeepGrow and DeepEdit in mind.

The MONAI Label application for click-based interactive prostatic lesion segmentation also has potential uses in diagnosis, active surveillance, and treatment planning. Firstly, lesion segmentation is a common first-step in CAD systems for CSPCa lesion classification [15]. Typically, such systems require a manual segmentation of lesions by a radiologist, which can be improved by our interactive approach by reducing the time and effort required to segment lesions. A second use of the interactive prostatic lesion segmentation application is in volume estimation of lesions. Like whole prostate volume estimation, radiologists use the ellipsoid formula to estimate lesion volume, which ignores exact lesion morphology [152].

Determination of lesion volume is important at the time of diagnosis for lesion scoring [24, 25], and for monitoring progression in patients placed on active surveillance [166]. Thirdly, accurate lesion segmentation is important for planning focal treatments e.g., partial prostatectomy [167], focal boosted radiotherapy [173], and focal cryotherapy [170].

This work has some limitations. Firstly, while several different clinical uses of the click-based interactive whole prostate and prostatic lesion segmentation applications have been described, those clinical uses were not examined or evaluated due to the preliminary nature of this work. Another limitation is the lack of external validation, and beyond that, multicenter external validation and prospective validation, all of which should be addressed in future work. A final limitation is the lack of investigation into how active learning impacts performance over time, again due to the preliminary nature of this work; all MONAI Label applications for interactive segmentation contain the functionality to allow active learning.

Chapter 7

Summary and future work

The rise in prostate cancer (PCa) incidence [19], the increasing use of multiparametric magnetic resonance imaging (mpMRI) to identify, score, and stage PCa [26], and a shortfall of specialist radiologists to meet prostate radiology demand is putting significant pressures on the PCa diagnostic pathway [14]. Furthermore, the current diagnostic approach must be improved to reduce the small proportion of men with clinically significant PCa (CSPCa) who are missed by mpMRI, to reduce the large number of men who undergo unnecessary biopsies, and to increase the inter-observer agreement between radiologists [39].

Artificial intelligence (AI) algorithms are being investigated for numerous medical image analysis applications [40]. Notably, the number of AI products for radiology with a CE mark or Food and Drug Administration (FDA) approval has rapidly expanded over the past few years [174, 175]. The work in this thesis investigated applications of a subclass of AI, known as deep learning, to the PCa diagnostic pathway, with a view towards future clinical deployment.

In Chapter 4, we introduced a novel patient classification framework, PCF, that assigns a probability of having CSPCa to patients based on mpMRI and clinical features. PCF extracts features from volumetric mpMRI and derived parameter maps using convolutional neural networks (CNNs) and subsequently, combines imaging features with clinical features through a multi-classifier support vector machine (SVM) scheme. The chief strength of PCF is that it can be trained using patient-level labels only rather than requiring lesion annotations; patient-level la-

bels that indicate the cancer status of patients can be inferred from biopsy findings with minimal additional effort, while drawing lesion contours requires substantial additional effort from radiologists outside of the clinical routine [112]. Another strength of PCF is its ability to combine clinical features with imaging features effectively for enhanced classification performance. In particular, we found that the inclusion of prostate-specific antigen (PSA) density (PSAd) improved classification performance, with statistical significance. On temporal validation using a subset of the "Prostate Imaging Compared to Transperineal Ultrasound-guided biopsy for significant prostate cancer Risk Evaluation" (PICTURE) dataset [33], PCF achieved comparable sensitivity and specificity to a highly experienced radiologist with 10 years' experience of reading and scoring prostate mpMRI. Therefore, deployments of PCF as a triage tool or as a second reader were suggested, following multi-centre and prospective validation.

Prior to multi-centre and prospective validation, technical enhancements to PCF can be made. Firstly, PCF can be enhanced to provide a measure of uncertainty associated to the patient-level probability of having CSPCa output by PCF. In particular, it is important for PCF to be able to indicate when its output may be uncertain due to input data which is outside of the distribution of the data used for training. Several methods have been proposed to capture the epistemic (model-based) and aleatoric (data-based) uncertainty of deep learning systems [155, 176, 177, 84]. A further technical enhancement that may improve performance involves replacing the use of feature selection and SVMs for selection and combination of CNN extracted image features and clinical features, with a transformer-based approach with self-attention [76]. Transformers have gained popularity over the past few years for natural language processing (NLP) tasks in particular, with Google's Bidirectional Encoder Representations from Transformers (BERT) [178] and OpenAI's Generative Pre-trained Transformer 3 (GPT-3) [179] achieving state-of-the-art performance for language translation and sentence prediction tasks. In PCF, a transformer-based approach can improve performance by giving spatial and modality context to each MRI feature and giving context to MRI features using clinical features and vice

versa; this is analogous to context attribution to words in a sentence using neighbouring words, as seen in NLP tasks.

In Chapter 5, we introduced AutoProstate, a deep learning-powered framework for automatic MRI-based PCa assessment and reporting. AutoProstate performs segmentation of the peripheral zone (PZ) and central gland (CG) on T2-weighted MRI (T2WI), and performs segmentation of CSPCa lesions using T2WI, apparent diffusion coefficient (ADC) map, and computed b2000 (Cb2000) diffusion-weighted MRI (DWI). In addition, PZ and CG guidance is provided for CSPCa segmentation since lesion occurrences and appearances depend on their zonal location [159]. Subsequently, patient meta-data and automatic segmentation outputs are used to generate a novel automatic web-based report containing four sections: *Patient Details*, *Prostate Size and PSA Density*, *Clinically Significant Lesion Candidates*, and *Findings Summary*. AutoProstate was trained using the publicly available PROSTATEx dataset [44], and externally validated using the PICTURE dataset [33]. Moreover, the performance of AutoProstate was compared to the performance of an experienced radiologist who prospectively read PICTURE dataset cases. In comparison to the radiologist, AutoProstate showed statistically significant improvements in prostate volume and PSAd estimation. Furthermore, AutoProstate matched the CSPCa lesion detection sensitivity of the radiologist, which is paramount, but produced more false-positive detections. AutoProstate's intended clinical deployment is as a companion system for radiologists to improve diagnostic accuracy and reporting quality, following multi-centre and prospective validation.

Prior to multi-centre and prospective validation, technical enhancements to AutoProstate can be made. First and foremost, the number of false-positive detections must be reduced. False positive detections can lead to unnecessary biopsies and reduced confidence in the system. Increasing the size of the training dataset will likely reduce the number of false-positive detections, while at the same time increasing CSPCa detection sensitivity. The addition of a dedicated false-positive reduction stage to the CSPCa-Segmenter module may also reduce false-positives. For example, in Saha et al. [123], a decoupled residual CNN clas-

sifier was applied to the output of their CSPCa lesion detection CNN, to reduce false-positives by identifying parts of the image unlikely to contain a CSPCa lesion. Alternatively, a dedicated lesion classification system that classifies patches containing CSPCa lesion candidates, detected by CSPCa-U-Net-E, could also reduce false positives; several lesion classification systems have been published [96, 97, 98, 99, 100, 101, 102, 104, 105, 106, 107, 109, 110, 111]. Lesion classification systems do not require training data with a complete histopathological characterisation of the whole prostate through prostatectomy or transperineal template prostate-mapping (TTPM) biopsy, but instead can be trained using patches containing a lesion that has been confirmed using targeted biopsy. Targeted biopsies are performed at a greater frequency in clinical routine [9], therefore lesion classification systems can be trained using a greater number of lesion examples. A further solution for false-positive reduction (and increased CSPCa lesion detection sensitivity) may be replacement of the standard U-Net architecture with a UNEt TRansformers (UNETR) architecture [73]. In UNETR, the U-Net encoder is replaced with a transformer encoder, which the authors claim allows UNETR to capture global multi-scale dependencies, overcoming the locality of convolutions. UNETR has achieved favourable benchmarks on volumetric brain tumour segmentation using multimodal MRI and spleen segmentation using computed tomography (CT). A second enhancement to AutoProstate involves replacing the binary CSPCa lesion segmentation task with a multi-class lesion segmentation task whereby PCa lesions are detected and classified according to a Gleason score, as in the work of Cao et al. [112]. An automatic Gleason score prediction may eradicate the need for biopsy provided sufficient accuracy, and would inform patient selection for treatment, active surveillance, and watchful waiting. A third enhancement to AutoProstate would replace the automatically generated template paragraph in the *Findings Summary* section of the automatic report with automatically generated text from a generative model trained using image and clinical findings text pairs, as in the work of Xue et al. [180], where a multimodal recurrent model with attention was presented for generating high-level conclusive impressions from medical images.

In Chapter 6, we built interactive segmentation pipelines for whole prostate segmentation and prostatic lesion segmentation using MONAI Label. Interactive segmentation methods provide a route to clinical deployment by overcoming the reported lack of robustness associated to automatic segmentation algorithms [45] and allow for continuous optimisation through active learning. Whole prostate segmentation was performed using T2WI, while prostatic lesion segmentation used T2WI, ADC map, and Cb2000 DWI. In both segmentation tasks, a 3D implementation of DeepGrow was used as the baseline framework for click-based interactive segmentation; DeepGrow for 2D segmentation tasks was proposed by Sakinis et al. [45]. Three experiments were run to optimise performance. The first experiment considered whether performance could be improved through a hybrid Dice and Focal loss, as in our work on AutoProstate, as opposed Dice loss which was used by Sakinis et al. in their work. The second experiment investigated a new training strategy in which simulated training clicks were directed to false-negative or false-positive regions with the highest prediction error as opposed to the largest false-negative or false-positive regions, as in the work by Sakinis et al. The third experiment compared the performances of DeepGrow and DeepEdit, where DeepEdit is an alternative to DeepGrow available in MONAI Label for improving the trade-off between click-free and with-click segmentation performance. In our results for whole prostate segmentation, we found that DeepGrow trained using Dice loss and the original training click simulation strategy outlined by Sakinis et al. produced the best overall segmentation performance; the mean Dice score for automatic segmentation was 0.907, which increased to 0.915, 0.919, 0.926, 0.932, and 0.936 with the addition of 1, 2, 5, 10, and 15 simulated inference clicks. For prostatic lesion segmentation, training with Dice loss and the original training click simulation strategy outlined by Sakinis et al. again produced the best overall segmentation performance. However, the comparison of DeepGrow and DeepEdit revealed a performance trade-off. DeepEdit with parameters pDE = 0.25 and pDE = 0.5 gave superior click-free segmentation performance compared to DeepGrow, in terms of mean Dice score, with statistical significance; mean Dice scores of 0.268 and 0.272

were achieved, respectively. However, with one or more simulated inference clicks, DeepGrow achieved higher mean Dice scores than DeepEdit with mean Dice scores of 0.527, 0.592, 0.670, 0.723, and 0.749 at 1, 2, 5, 10, and 15 clicks, though statistical significance was not consistently observed.

As the work on click-based interactive segmentation presented in this thesis was a preliminary study only, several opportunities for future work exist. Firstly, the prostatic lesion segmentation application can be evaluated as the first stage in a lesion classification CAD system, whereby lesions segmented by the interactive application are attributed a Gleason score by a second-stage classification algorithm. Secondly, an extension of the current prostatic lesion segmentation application can be made for segmenting lesions over multiple timepoints, in patients enrolled in active surveillance. In particular, the prostatic lesion segmentation application can undergo patient-specific fine-tuning using all MRI scans acquired to-date ($t = 1, \ldots, n$) and their associated lesion segmentations (also generated with the interactive application), to allow a more accurate segmentation at time $t = n + 1$. Thirdly, extensions of the current whole prostate and prostatic lesion segmentation applications can be developed for whole prostate and focal treatment planning. Finally, work should be undertaken to investigate active learning in the context of the two interactive segmentation applications developed in this thesis. In particular, work can be undertaken to understand the impact of imperfect or "noisy" segmentations, generated in a prospective setting, on active learning. While interactive segmentation methods such as DeepGrow and DeepEdit allow improvements upon an initial automatic segmentation through user-provided clicks, it is likely that the segmentations obtained will not reach the level of accuracy of a manual segmentation produced by a clinician. Rather, user-clicks may be added until the segmentation is "good enough" for its clinical purpose. Therefore, active learning training strategies will need to be developed that are robust to noisy segmentations as ground-truth; several strategies for dealing with noisy ground-truth labels when training deep learning algorithms have been outlined by Karimi et al. [181].

The CAD systems presented in this thesis assume the availability of at least

T2WI and DWI. However, MRI modalities may be missing in clinical practice [182, 183]. Therefore, future development should incorporate strategies to deal with missing modalities. Havaei et al. [182] and Dorent et al. [183] have proposed approaches for segmentation that are robust to missing modalities. Havaei et al. proposed HeMIS: Hetero-Modal Image Segmentation. HeMIS learns, for each modality, an embedding of the modality into a single latent vector space, trained using modality dropout to enable robustness to missing modalities. Points in the latent space are averaged over modalities available at inference time to yield the desired segmentation. Evaluation on neurological MRI datasets revealed state-of-the-art performance when all modalities were available and most importantly, a smooth decline in performance when modalities were removed. Alternatively, Dorent et al. proposed a hetero-modal variational autoencoder (VAE), which learns a shared latent representation, rather than combining latent vectors using averaging, as in Havaei et al. On the task of brain tumour segmentation, their method outperformed the method presented by Havaei et al. and achieved similar performance to subset-specific equivalent networks. A related problem concerns modalities that are acquired with distortion/artifacts, which may be unusable for PCa diagnosis. In this case, rather than missing modality completion, a modality correction problem can be formulated; previous works have looked at correcting MRI motion artifacts [184, 185, 186].

The methodological chapters in this thesis (4, 5, and 6) present CAD systems with varying application suggested. In future, efforts can be allocated to combining the distinct CAD systems into a single modular platform for PCa triage, diagnosis, active surveillance monitoring, and treatment planning. For example, PCF can perform an initial triage of patients who have undergone MRI to rank patients by likelihood of CSPCa/rule-out patients with a low-risk for CSPCa. Subsequently, AutoProstate can perform an automatic assessment of MRI which can be referred to by the diagnostic radiologist to determine a consensus view on biopsy targets. Finally, click-based interactive segmentation can be used to improve the segmentations of the whole prostate and biopsy-confirmed lesions, for the purposes of generating

new training data for AutoProstate for use in active learning, treatment planning, or active surveillance monitoring.

Prior to clinical deployment, the CAD systems introduced in this thesis or a integrated modular platform, as described above, should undergo multi-centre external validation and prospective validation, as recommended in the review by Syer and Mehta et al. [15]. Multi-centre external validation studies reveal the extent to which CAD systems can generalise beyond the data used to train them. A multi-centre external validation of a machine learning-based CAD system was presented by Gaur et al. [113]. In their study, CAD performance was evaluated using images acquired from four countries across three continents, and CAD-assisted radiologist interpretation was compared to CAD-unassisted radiologist interpretation using radiologists from six countries spread across five continents. Unfortunately, evaluation studies of this rigour are rare and to the best of our knowledge, have not been performed for CAD systems that use deep learning. Unlike CAD system validation studies that consider retrospective data in a research setting, prospective validation studies consider performance in the deployment setting. To the best of our knowledge, prospective validation studies for PCa CAD have not been performed. However, an interesting and informative study was published by Beede et al. [187], concerning the prospective validation of a deep learning system for diabetic retinopathy. Screening systems were deployed in 11 clinics across Thailand. They found several human and socio-economic factors that impacted the performance of their deep learning system, which require ample consideration before deployment. Notably, multi-centre external validations and prospective validations may be performed with greater ease in the near future due to the substantial efforts of groups that are working to develop federated learning infrastructures for healthcare [188, 189].

Appendix A

Background theory: magnetic resonance imaging

Magnetic resonance imaging (MRI) is a highly flexible medical imaging technique that can be used to produce detailed anatomical or functional images of parts of the body. MRI relies on the fundamental theory of nuclear magnetic resonance (NMR) and Fourier theory. Using a powerful static magnetic field, a perturbing radiofrequency (RF) field and spatially localising gradient fields, a localised signal can be collected, predominantly from hydrogen nuclei within the body. The signal collected is transformed into the image domain using Fourier transforms. Importantly, through specific pulse sequences and acquisition parameters, images of different contrast are produced.

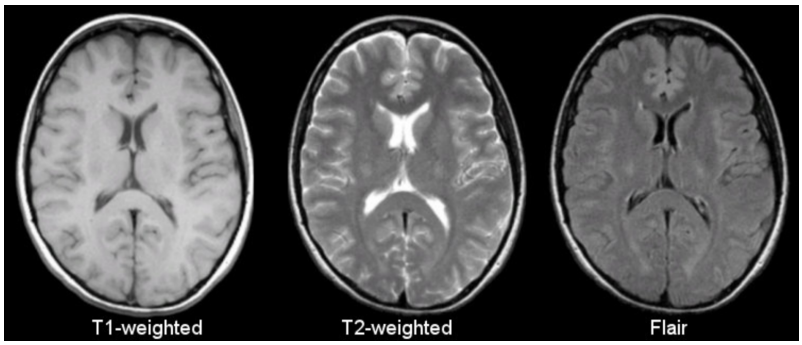


Figure A.1: T1-weighted, T2-weighted, and Flair MRI contrasts for a transverse slice through the brain [1].

A.1 Fundamental nuclear magnetic resonance theory

MRI is based on NMR, which is a physical phenomenon in which nuclei in a magnetic field absorb and re-emit electromagnetic radiation. It was first observed in 1946, separately, by research groups led by Felix Bloch and Edward Purcell. NMR can be described from both quantum and classical physics perspectives. Most of NMR can be described by classical physics, since we consider the bulk effect of many nuclei in tissues, however quantum theory is required to explain some of what is observed.

A.1.1 Spin, magnetic moment, and Larmor precession

Atomic nuclei possess an intrinsic quantum property called spin. The dominant nucleus in MRI is the proton of hydrogen nuclei, due to the abundance of water, fat, and other organic molecules in the human body. The classical interpretation of spin is a particle rotating about its own axis. Extending the classical picture, a rotating charged proton will have a magnetic moment $\vec{\mu} = \gamma\vec{L}$ where γ is the gyromagnetic ratio associated with a proton and $\vec{L}$ is the angular momentum. The rotating charge produces a small magnetic field. MRI is based on the interaction of the nuclear spin with an external magnetic field, $\vec{B_0}$.

When a magnetic moment is exposed to an external magnetic field $\vec{B_0}$, a torque is produced, causing a circular motion about $\vec{B_0}$, known as "precession". This is described by the following differential equation [2][Eq. 2.24]:

$$\frac{d\vec{\mu}}{dt} = \gamma\vec{\mu} \times \vec{B_0}. \tag{A.1}$$

The solution for the above differential equation describes $\vec{\mu}$ precessing about $\vec{B_0}$ at an angular frequency ω_0 where [2][Eq. 1.1]:

$$\omega_0 = \gamma B_0. \tag{A.2}$$

This precession frequency is referred to as the Larmor frequency, named after physi-

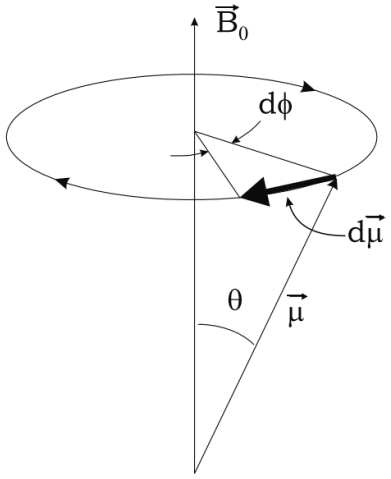


Figure A.2: Clockwise precession of a magnetic moment about an external magnetic field $\vec{B_0}$ [2][Fig 1.1].

cist Joseph Larmor. The gyromagnetic ratio γ for the hydrogen proton in water is roughly 2.68×10^8 rad/s/Tesla or alternatively, $\frac{\gamma}{2\pi}$ is 42.58 MHz/Tesla. The solution to the equation of motion can be shown to be [2][Eq. 2.32]:

$$\vec{\mu}(t) = \mu_x(t)\hat{x} + \mu_y(t)\hat{y} + \mu_z(t)\hat{z}, \tag{A.3}$$

where [2][Eq. 2.33],

$$\mu_x(t) = \mu_x(0)\cos(\omega_0 t) + \mu_y(0)\sin(\omega_0 t), \tag{A.4}$$

$$\mu_y(t) = \mu_y(0)\cos(\omega_0 t) - \mu_x(0)\sin(\omega_0 t), \tag{A.5}$$

$$\mu_z(t) = \mu_z(0). \tag{A.6}$$

A.1.2 Magnetisation

In section A.1.1 above, a single proton was considered in the $\vec{B_0}$ field. However, in MRI, it is the bulk effect of many spins that gives a detectable signal.

Consider a volume element ("voxel") with volume V that contains a large number of protons. The magnetisation is defined as the sum of the individual magnetic moments divided by the total volume [2][Eq. 4.1]:

$$\vec{M} = \frac{1}{V} \sum_{i=\text{protons in V}} \vec{\mu}_i. \tag{A.7}$$

In the absence of an external magnetic field, the magnetic moments will be randomly aligned. Therefore, the net magnetisation vector sum will be approximately 0.

To consider what happens to the net magnetisation when the external magnetic field $\vec{B}_0$ is turned on, the quantum view is useful. In a static external magnetic field, protons take up defined energy states. Spin $\frac{1}{2}$ protons have two energy levels and therefore, two possible alignments. Parallel / "spin up" or anti-parallel / "spin down". The parallel alignment is at the lower energy level, so there exists a small "spin excess" in the parallel alignment [2][Eq. 1.2]:

$$\text{spin excess} \simeq N\frac{\hbar\omega_0}{2kT}, \tag{A.8}$$

where N is the total number of spins present in the sample, $\hbar \equiv \frac{h}{2\pi}$ in terms of Plank's quantum constant h, k is the Boltzmann's constant and T is the absolute temperature. For typical MRI scanner strengths, the spin excess is millions of times smaller than the total number of spins [2][Pg. 5]. However, sufficient signal can be detected from tissue due to the Avogadro number of spins present. As a result, rather than considering individual spins, a net magnetisation vector $\vec{M}$, representing the spin excess is considered, aligned with the $\vec{B}_0$ field.

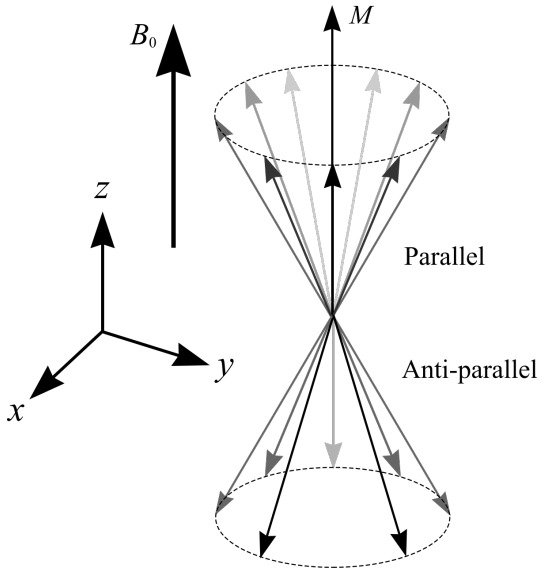


Figure A.3: Net magnetisation vector $\vec{M}$. There is a spin excess in the parallel alignment, giving rise to $\vec{M}$. $\vec{M}$ has no transverse component as the spins lack phase coherence.

A.1.3 Resonance

Since the magnetisation $\vec{M}$ at equilibrium is very small compared to $\vec{B_0}$, it cannot be detected. Therefore, to get a signal from the sample, $\vec{M}$ needs to be "tipped" away from the $\hat{z}$ direction into the transverse plane. The solution is to apply energy to the system through a rotating RF magnetic field $\vec{B_1}$ [2][Eq 3.24]:

$$\vec{B_1} = B_1(\hat{x}\cos(\omega t) - \hat{y}\sin(\omega t)), \tag{A.9}$$

where ω is the oscillation frequency of the RF field. To excite spins, the requirement is [2][Eq 3.30]:

$$\omega = \omega_0 \quad \text{(on-resonance condition)}. \tag{A.10}$$

The $\vec{B_1}$ field is transmitted through RF coils, which are part of the scanner hardware.

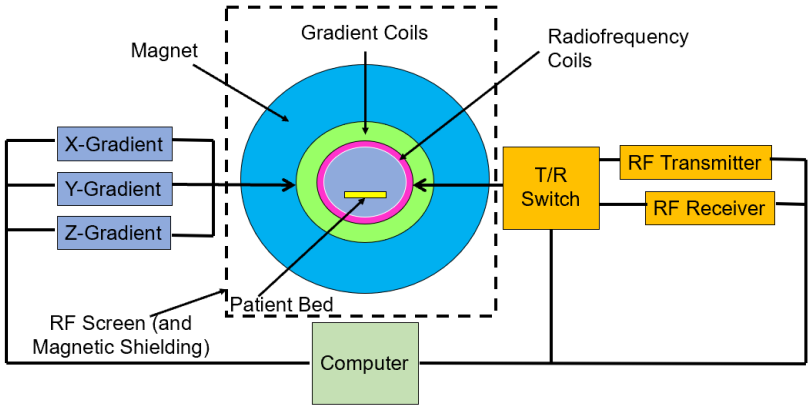


Figure A.4: Block diagram of the key hardware components found in a MRI scanner.

The angle of the flip depends on the strength of the $\vec{B}_1$ field and the time it is on, τ. The formula for the flip-angle is [2][Eq. 3.31]:

$$\Delta\vartheta = \gamma B_1 \tau. \tag{A.11}$$

In MRI, a 90° flip is typically considered, so that all the magnetisation is tipped into the x-y plane. However, a later discussion will show that some pulse sequences utilise flip angles smaller or larger than 90° to increase signal.

The expression for $\vec{B}_1$ given above assumes a cartesian coordinate frame. However, considering a rotating frame can simplify the discussion. The rotating frame rotates clockwise around the z-axis at the Larmor frequency. The coordinate system can be expressed as:

$$\hat{x}' = \hat{x}\cos(\omega_0 t) - \hat{y}\sin(\omega_0 t), \tag{A.12}$$

$$\hat{y}' = \hat{x}\sin(\omega_0 t) + \hat{y}\cos(\omega_0 t), \tag{A.13}$$

$$\hat{z}' = \hat{z}. \tag{A.14}$$

In the rotating frame, the $\vec{B}_1$ field can be expressed [2][Eq. 3.32]:

$$\vec{B}_1 = B_1 \hat{x}'. \tag{A.15}$$

The magnetisation vector is now seen to rotate about the $\hat{x}'$ axis under the influence of the $\vec{B}_1$ field.

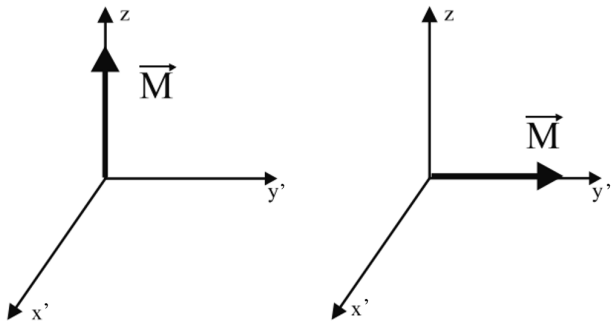


Figure A.5: 90° flip of the net magnetisation vector into the transverse plane [2][Fig. 4.2]

.

A.2 The Bloch equation and relaxation

Following excitation by the rotating $\vec{B}_1$ field, the net magnetisation vector $\vec{M}$ will precess in the transverse plane, giving a signal that is detected by receiver coils. However, due to the interactions of spin magnetic moments with each other and their surroundings, there will be an independent decay of the transverse magnetisation and a recovery of the longitudinal magnetisation back towards equilibrium. This behaviour is summarised by the Bloch equation, which includes the relaxation and decay parameters T_1, T_2, and T_2^* which are later shown to be important in generating image contrast.

A.2.1 Spin-lattice relaxation

Through collisions, rotations, and electromagnetic interactions, magnetic moments lose their magnetic energy to their surroundings (sometimes called the "lattice"). This will facilitate a return to the lower energy equilibrium state of the system (in quantum terms, a small "spin up" excess). As energy leaves the system, the longitudinal magnetisation component M_Z will grow towards the equilibrium magnetisation M_0 at a rate dictated by relaxation parameter T_1. To be precise, T_1 is the empirically determined time it takes for M_Z to reach approximately 63% of its maximal value M_0. T_1 can vary quite significantly by tissue from hundreds of ms

to a few seconds. The recovery of longitudinal magnetisation can be modelled by a differential equation containing the time constant T_1 [2][Eq. 4.11]:

$$\frac{dM_z}{dt} = \frac{1}{T_1}(M_0 - M_z), \tag{A.16}$$

whose solution can be found to be [2][Eq. 4.12]:

$$M_z(t) = M_z(0)\exp(-\frac{t}{T_1}) + M_0(1 - \exp(-\frac{t}{T_1})), \tag{A.17}$$

where $M_z(0)$ is the initial value following the RF pulse.

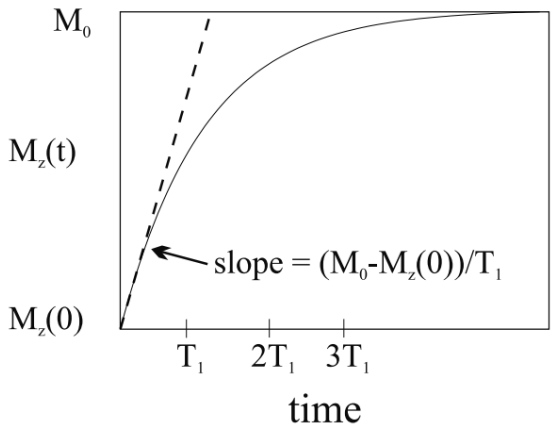


Figure A.6: Longitudinal magnetisation recovery to equilibrium value M_0 [2][Fig. 4.1a].

A.2.2 Spin-spin interaction and transverse decay

Spins experience local fields influenced by the fields of their neighbouring spins. Therefore, the effective field experienced by spins will vary, altering their precession frequency and causing dephasing. This loss of phase coherence can be pictured as a fanning out of spins as they precess at different frequencies. This dephasing leads to a decay of the net transverse magnetisation $M_\perp$, according to the decay parameter T_2. T_2 is the time it takes for $M_\perp$ to fall to approximately 37% of its initial value.

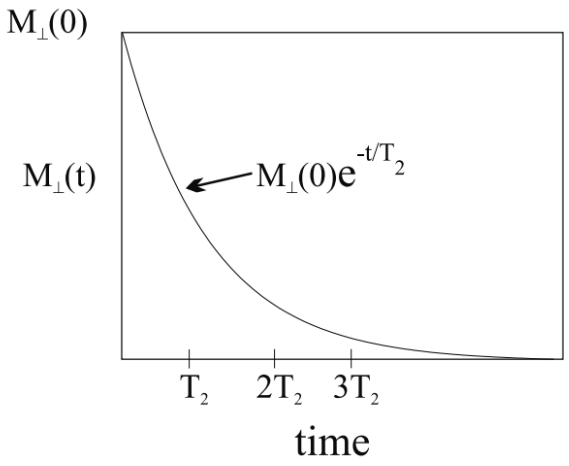


Figure A.7: Transverse magnetisation decay from initial value following 90° flip [2][Fig. 4.1b].

$M_\perp$ can be modelled with the following rotating frame differential equation [2][Eq. 4.14]:

$$\frac{dM_\perp}{dt} = -\frac{1}{T_2}M_\perp, \tag{A.18}$$

with the rotating frame solution [2, eq 4.16][Eq. 4.16]:

$$M_\perp(t) = M_\perp(0)\exp(-\frac{t}{T_2}). \tag{A.19}$$

In liquids, molecules are rapidly tumbling and the magnetic effect of the neighbouring molecules may be cancelled out on the time scale of the MR measurement. This leads to long T_2 values and enables MRI to be performed. Solids on the other hand have very short T_2 values, explaining why bone does not give much signal in MRI.

A.2.3 Magnetic field inhomogeneity

The main magnetic field $\vec{B_0}$ is never perfectly uniform. Therefore, in practice, there is an additional dephasing of the magnetisation. The effect of spin-spin dephasing and external field inhomogeneity is represented by the combined decay parameter T_2^*. This will usually replace T_2 in discussion, however, T_2 is recoverable using "spin-echo" sequences which are discussed in section A.3.2.

A.2.4 The Bloch Equation

The differential equations describing the longitudinal magnetisation recovery and transverse magnetisation decay can be combined into a single differential equation called the Bloch equation [2][Eq. 4.21]:

$$\frac{d\vec{M}}{dt} = \gamma\vec{M} \times B_0\hat{z} + \frac{1}{T_1}(M_0 - M_z)\hat{z} - \frac{1}{T_2}\vec{M}_{\perp}. \tag{A.20}$$

The complete set of solutions (cartesian coordinate frame) is [2][Eq. 4.25, 4.26, 4.27]:

$$M_x(t) = \exp(-\frac{t}{T_2})(M_x(0)\cos(\omega_0 t) + M_y(0)\sin(\omega_0 t)), \tag{A.21}$$

$$M_y(t) = \exp(-\frac{t}{T_2})(M_y(0)\cos(\omega_0 t) - M_x(0)\sin(\omega_0 t)), \tag{A.22}$$

$$M_z(t) = M_z(0)\exp(-\frac{t}{T_1}) + M_0(1 - \exp(-\frac{t}{T_1}). \tag{A.23}$$

Note: T_2^* may replace T_2 depending on the pulse sequence used.

A.3 Signal acquisition

By the action of the rotating $\vec{B}_1$ field, the magnetisation is tipped into the transverse plane. As time evolves, there will be a transverse magnetisation decay and a longitudinal recovery. While the magnetisation has a transverse component, the detection of its precession about $\vec{B}_0$ can be considered through Faraday's law of induction. The bore of the MRI scanner contains receiver coils. As the magnetic field lines of the transverse magnetisation from the sample being imaged sweep across the receiver coils, an electromotive force is induced in the receiver coils. This is referred to as signal.

A.3.1 Free induction decay

After excitation by a 90° (or otherwise) RF pulse, $\vec{B}_1$ is turned off and the RF receiver coil is used to detect the MR signal. The simplest signal from a homogeneous sample is known as a Free Induction Decay (FID). The FID signal is a damped sine

wave, oscillating at the Larmor frequency (ω_0), of the following form:

$$s_0 \exp(-\frac{t}{T_2^*}) \sin(\omega_0 t), \tag{A.24}$$

where s_0 is the initial signal after the flip.

At this point, the notion of a sequence diagram is introduced. These will later become very useful to understand more complex imaging sequences.

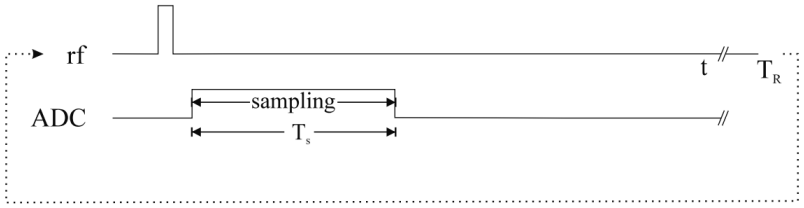


Figure A.8: Sequence diagram for a repeated FID experiment [2][Fig. 8.2].

The sequence diagram is a visual representation of the acquisition process. It shows, in chronological order, the RF pulses used, the sampling time, and for more complex imaging sequences, the gradient fields used to localise (more on this in section A.4.2). Typically, one repeat of the experiment will be shown where T_R is the repeat time.

A.3.2 Spin-echo sequence

The spin-echo sequence is used to reverse the effects of T_2^* decay; T_2^* decay can severely limit the measurable signal. The spin-echo sequence is based on the application of two RF pulses. The usual 90° pulse is followed by a refocusing 180° pulse. Essentially, the dephasing caused by field inhomogeneity is reversed by this additional pulse, so that the spins rephase. When the spins rephase as much as possible, an echo is formed at time T_E (the "echo time"). Spin-echoes allow T_2 to be measured as the effect of field inhomogeneity is removed, but the effect of spin-spin interaction is not.

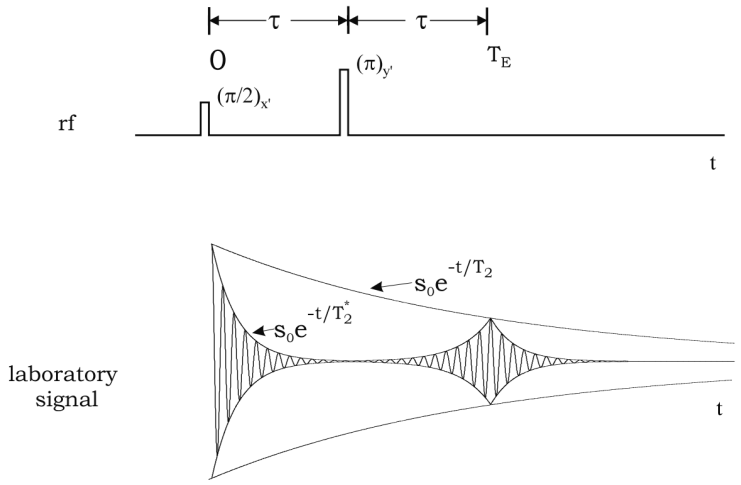


Figure A.9: Sequence diagram and signal representation for a spin-echo sequence [2][Fig. 8.3].

Spin-echo sequences are used extensively in MRI. Another way to form an echo is through the use of gradient echoes which will be introduced during the discussion on spatial localisation in section A.4.2.

A.3.3 Demodulation

MRI returns a signal in the MHz range. By the Nyquist theorem, to prevent aliasing, signals must be sampled at least twice per cycle i.e., the sampling frequency must be greater than $2 \times \omega_0$. This is not practical as samples would have to collected in the nanosecond range. The solution is to sample the demodulated signal, which is similar to viewing the signal in the rotating frame [2][Pg.104] i.e., without the rapid Larmor frequency oscillations.

A.4 Imaging (spatial localisation)

To this point, there has been discussion of the physical principles of NMR and how a signal is generated and collected, but not how signal can be used to form an image. To form an image, spins must be "phase-encoded" using linearly varying gradient fields. For the discussion to follow, the collection of a 2D transverse slice through the body is assumed, though the principles can be extended for other slice orientations or 3D volume imaging.

A.4.1 K-space and Fourier imaging

To produce an image from the collected signal, "k-space" is utilised. K-space is an array of numbers representing the spatial frequencies in an image. In the case of 2D imaging, the k-space array will be a 2D array. The inverse Fourier transform of the k-space produces a MRI image. This relationship is expressed by the Fourier transform pair [2][Eq. 10.7, 10.8]:

$$s(k_x, k_y) = \int\int dx\, dy\, \rho(x, y, z) \exp(-i2\pi(k_x x + k_y y)), \tag{A.25}$$

$$\hat{\rho}(x, y) = \int\int dk_x\, dk_y\, s(k_x, k_y) \exp(i2\pi(k_x x + k_y y)). \tag{A.26}$$

In practice, the discrete Fourier transform is used since k-space is discretely sampled.

(a) Image space object **(b)** k-space signal (magnitude)

Figure A.10: 2D Fourier transform.

To form an image, k-space must be sampled adequately. Different imaging pulse sequences provide different ways of achieving k-space coverage.

A.4.2 Gradient-echo with spatial encoding

The gradient-echo sequence uses echoes induced by gradient fields to collect spatially localised signal. The gradient-echo imaging sequence is shown below and its key features are discussed:

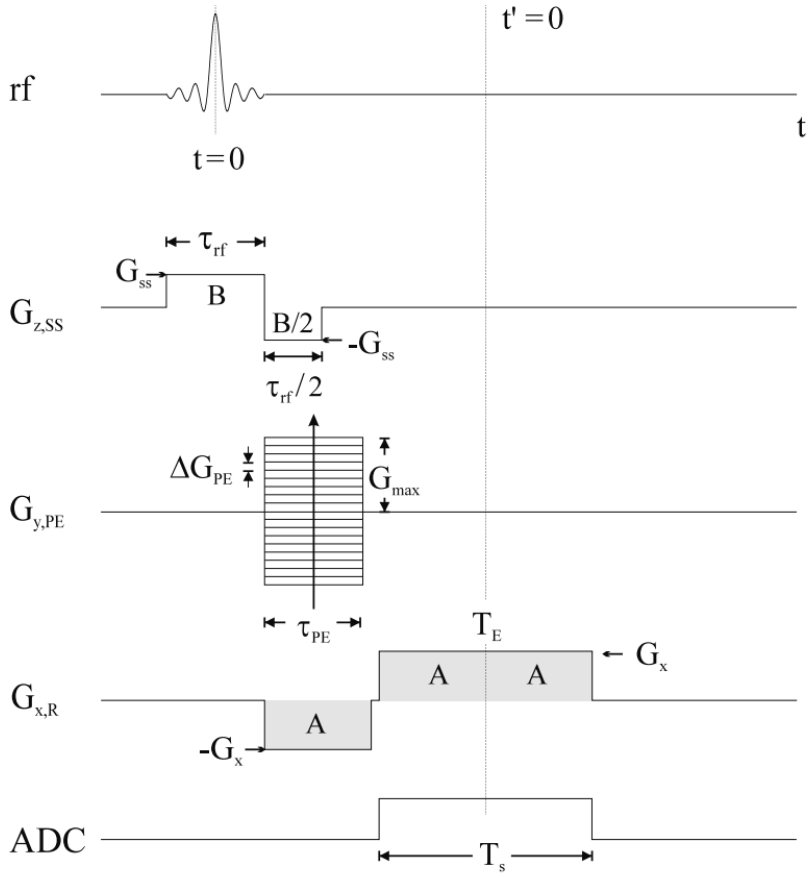


Figure A.11: 2D gradient-echo sequence [2][Fig. 10.14].

- $G_{z,SS}$ is the slice-select gradient. In the case of a transverse slice through the body, the slice-select gradient is applied in the z-direction. By using a sinc RF pulse, only spins in the slice of interest are excited. In the frequency domain, a sinc pulse will give a rectangular distribution of frequencies to excite the spins through a slice orthogonal to the z-axis of thickness TH given by [2][Eq. 10.20]:

$$\mathrm{TH} = \frac{\mathrm{BW_{rf}} 2\pi}{\gamma G_z}. \tag{A.27}$$

  $G_{z,SS}$ is active during the sinc pulse. Following the sinc pulse, a negative gradient is applied to reverse the phase accumulation of the slice-select gradient.

- $G_{x,R}$ is the read (or frequency encoding) gradient in the x-axis direction. The read gradient features a dephasing lobe followed by a rephasing lobe, at the

centre of which an echo forms. The read gradient frequency encodes the spins in the slice. The read gradient allows the collection of one line of k-space data in the k_x direction, with points determined by the formula [2][Eq. 10.9]:

$$\Delta k_x = \frac{\gamma G_x \Delta t}{2\pi}, \tag{A.28}$$

where Δk_x is a step in the k_x direction. Only samples s(k) during the sampling time TS are collected.

- $G_{y,PE}$ is the phase-encoding gradient in the y-axis direction. The phase-encoding gradient allows traversal of k-space in the k_y direction by stepping through a series of gradient steps at each repeat TR. The traversal of k-space in the k_y direction through the action of phase encoding gradients is described the formula [2][Eq. 10.10]:

$$\Delta k_y = \frac{\gamma \Delta G_y \tau_y}{2\pi}, \tag{A.29}$$

where Δk_y is a step in the k_y direction.

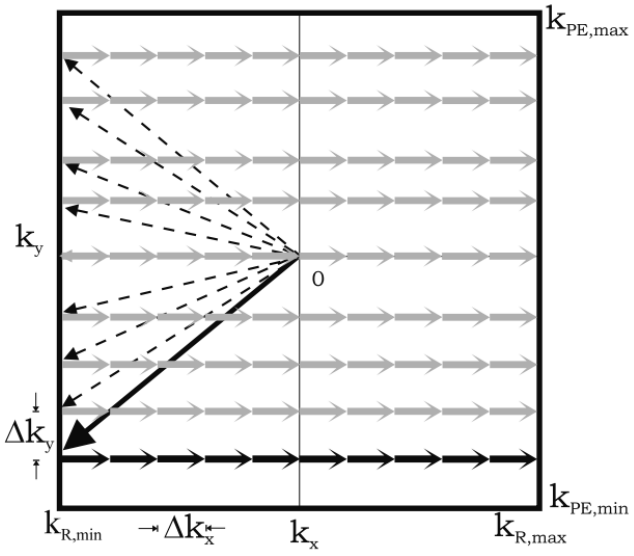


Figure A.12: Traversal of k-space for a typical gradient-echo experiment [2][Fig. 10.15b].

The total acquisition time for a 2D imaging experiment is [2][Eq. 10.12]:

$$T_{acq} = N_y T_R, \tag{A.30}$$

where N_y is the number of phase-encoding steps and TR is the repeat time.

A.4.3 Spin-echo with spatial encoding

To boost the signal that is collected, a 180° pulse can also be incorporated into the imaging sequence. This is referred to as a spin-echo experiment as the requirement of gradients for imaging is implicit.

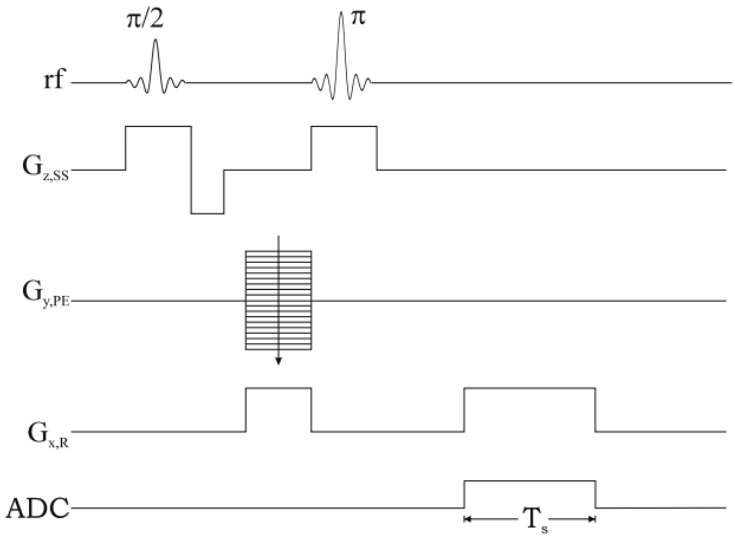


Figure A.13: 2D spin-echo sequence [2][Fig. 10.17].

A.5 Image contrast

MRI is capable of producing multiple different structural and functional contrasts through adjustment of acquisition parameters. A brief description of some of the different image contrasts possible is given below. In addition to those described below, other image contrasts are possible.

A.5.1 Conventional structural contrasts

Contrast	Brief Description	Acquisition
Proton density	Contrast based on density of protons in tissue.	Standard spin-echo (long TR, short TE), Fast spin-echo (long TR, short effective TE), Gradient-echo (short TR, short TE, small flip angle)
T1-weighted	Contrast based on T1 relaxation times differential.	Progressive saturation (short TR/TE spin-echo sequence), Inversion recovery (Inversion pulse preceding short TE, long TR spin-echo sequence)
T2-weighted	Contrast based on T2 relaxation times differential.	Standard spin-echo (long TR, long TE), Fast spin-echo (long TR, long effective TE), Echo-planar imaging with 180° refocusing pulse
T2*-weighted	Contrast based on T2* relaxation times differential.	Echo-planar imaging without 180° refocusing pulse

Table A.1: A brief description of conventional structural MRI contrasts.

A.5.2 Diffusion-weighted MRI

In diffusion-weighted imaging (DWI), contrast is due to the motion (or lack of motion) of water molecules. Water diffusion in tissue can be intracellular, extracellular, and between intracellular and extracellular spaces. A key equation used to understand the motion of water molecules is given by the Einstein relation:

$$< r^2 >= 6Dt, \tag{A.31}$$

where $< r^2 >$ is the mean square displacement of water molecules, D is a diffusion coefficient and t is the time from observation start. For free water at room temperature, D = $2 \times 10^{-3}\,mm^2s^{-1}$. However, in tissue, water motion is impeded, so the measured diffusion coefficient is reduced, giving rise to the apparent diffusion coefficient (ADC), where (ADC $<$ D). The magnitude of the ADC is a reflection of tissue microstructure. Since tissue microstructure can change with pathology, this can be extremely useful.

DWI is possible using a spin-echo sequence with diffusion sensitising gradients

determined by a "b" value.

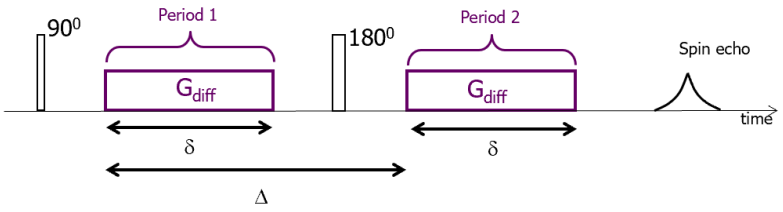


Figure A.14: Spin-echo sequence with diffusion sensitising gradients G_{diff}.

If during period 1, a molecule is in a position x_1 under the influence of G_{diff} and during period 2, the molecule is in a position x_2, again under the influence of G_{diff}, then the net phase accumulated is equal to:

$$\varphi = \gamma G_{diff} \delta (x_2 - x_1). \tag{A.32}$$

From this, it can be seen that water with less mobility will give a higher signal, as motion of water molecules causes dephasing of the transverse magnetisation. Quantitatively, the spin-echo signal magnitude is given by:

$$s = s_0 \cdot \exp(-b \cdot ADC), \tag{A.33}$$

where S_0 is the signal magnitude without applying diffusion sensitising gradients and b is the diffusion weighting (units s/mm^2):

$$b = \gamma^2 G_{diff}^2 \delta^2 (\Delta - \frac{\delta}{3}). \tag{A.34}$$

By repeating the acquisition with different values of b, an ADC map can be calculated, which is clinically useful. For example, ADC maps can be used to image cancerous growths, in which increased cell density will impede the motion of water, giving higher signal than background tissue.

A.5.3 Dynamic contrast-enhanced MRI

In dynamic contrast-enhanced (DCE) MRI, a contrast agent is injected that causes relaxation changes in the surrounding tissue. This has many clinical applications including in cancer diagnosis where angiogenesis of "leaky" vessels around a tumour will cause contrast agent accumulation. The most common contrast agent is Gadolinium, which is paramagnetic. The dominant relaxation effect is T1-shortening, so using a T1-weighted imaging sequence will give increased signal where the Gadolinium contrast agent accumulates.

Appendix B

Background material: machine learning and deep learning basics

Machine learning is based on the idea that algorithms can recognise patterns and make decisions by "learning" from data rather than being explicitly programmed to do so. Mitchell [46] provides a formal definition of learning: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

B.1 Machine learning theory

B.1.1 Supervised and unsupervised learning

Generally, machine learning algorithms can be divided into "supervised" and "unsupervised" learning algorithms. Both are mechanisms by which experience is gained through interaction with a dataset. In supervised learning both the input $\mathbf{x}$ and associated label y are available to the algorithm during optimisation. Generally speaking, the algorithm will have parameters Θ, which are updated, such that the prediction $\hat{y} = f(\mathbf{x}; \Theta)$ minimises a loss function $L(y, \hat{y})$. Popular examples of supervised machine learning algorithms include logistic regression, support vector machine (SVM), random forest, neural network, and convolutional neural network (CNN). In contrast, unsupervised learning algorithms learn patterns in the training data without the use of labels. Common applications of unsupervised learning in-

clude clustering and anomaly detection [46]. The advantages of supervised learning and unsupervised learning are combined by intermediate approaches between the two. For example, semi-supervised learning approaches combine a small quantity of labelled data and a larger quantity unlabelled data for training [190] and self-supervised learning approaches initialise algorithm parameters using a pseudo-task, followed by fine-tuning using labelled training data [191].

B.1.2 Evaluating a learning algorithm

Typically, learning algorithms require a dataset to be split into training, validation, and test (sometimes called inference).

- **Training dataset**: The portion of the dataset used to update the trainable parameters of the algorithm e.g., for a neural network classifier, the trainable parameters are the weights and biases of the neural network.
- **Validation dataset**: The portion of the dataset used to update/tune the hyperparameters of the algorithm e.g., for a neural network classifier, the number of hidden layers is a hyperparameter. Following each training run, the algorithm is evaluated on the validation set. By considering an accuracy or loss-based evaluation measure on the validation set, optimal hyperparameters can be found in an iterative manner.
- **Test dataset**: The portion of the dataset used to obtain an unbiased measure of performance for the trained algorithm on data not used during the training or validation stages. Ultimately, the test dataset is used to measure the ability of the trained algorithm to generalise beyond the data it was trained with.

B.1.3 K-fold cross-validation

A single partitioning of a dataset into training, validation, and test is biased towards the particular partitioning applied. In K-fold cross-validation, the dataset is split into K equally sized portions or "folds". Each of the folds is used as the test set in turn, while the remaining $K-1$ folds are used for training and validation. Values of five or ten for K are typical. There still exists a partitioning bias, as a particular

split into K folds is considered, though this can be reduced by repeating the cross-validation using different partitions into K folds.

B.1.4 Bias and variance

Bias and variance are terms used in statistical theory to describe the properties of a parameter estimator:

$$\text{Bias} = \text{B}(\hat{\vartheta}) = \text{E}(\hat{\vartheta}) - \vartheta \tag{B.1}$$

$$\text{Variance} = \text{V}(\hat{\vartheta}) = \text{E}[(\hat{\vartheta} - \text{E}(\hat{\vartheta})^2] \tag{B.2}$$

For machine learning algorithms, bias and variance can be thought of in terms of underfitting and overfitting:

$$\text{high bias} \leftrightarrow \text{underfitting} \tag{B.3}$$

$$\text{high variance} \leftrightarrow \text{overfitting} \tag{B.4}$$

Underfitting refers to the situation where the machine learning algorithm cannot model the training data or generalise to new data. High error on the training set is indicative of underfitting. On the contrary, overfitting refers to the situation when a machine learning algorithm models the training data too closely i.e., the machine learning algorithm learns both the target function and the noise in the training data, negatively impacting the ability of the trained algorithm to generalise to data unseen during training.

B.2 Deep learning theory

"Deep learning" is a term typically used to describe a subclass of machine learning based on neural networks with multiple layers of processing for extracting progressively higher features from data.

B.2.1 Neural networks representation

The notation used in this section, as well as the next sections on "Training a neural network" and "Optimisation" are from Andrew Ng's "Deep Learning Specialization" on Coursera (`https://www.deeplearning.ai/program/deep-learning-specialization/`).

The basic idea of a neural network is to stack and connect multiple simple nonlinear functions to create a complex nonlinear function. For classification tasks, this will allow us to approximate the complex functional relationship that exists between the inputs and associated labels in a dataset. A neural network is comprised of "neurons" (also called "nodes" or "units"), arranged in "layers". The first layer of a neural network is called the "input layer", while the last layer is called the "output layer". All layers in between the input layer and output layer are referred to as "hidden layers". Several hidden layers constitute a "deep" neural network, giving rise to the term "deep learning". Multilayer perceptrons (MLPs), are the most well-known type of neural network. MLPs have several layers where each neuron in a hidden layer is connected to all neurons in the preceding and following layer.

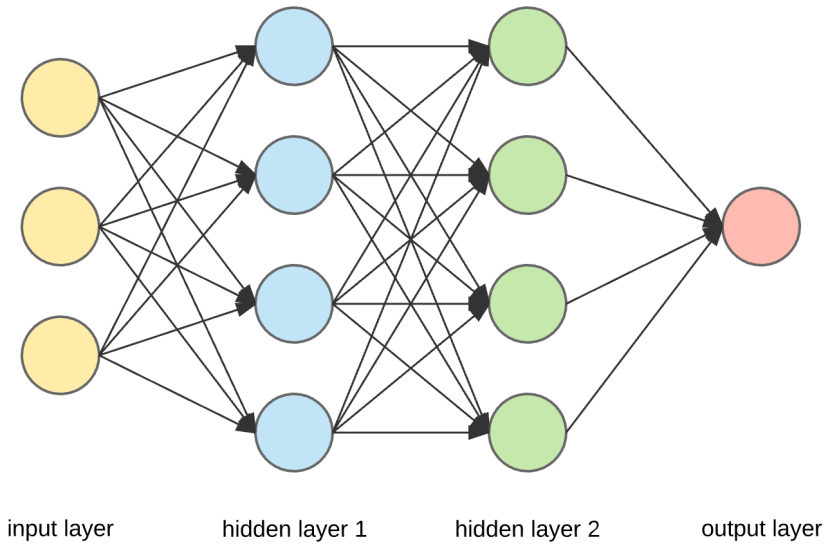


Figure B.1: Diagram of a simple MLP with two hidden layers [3].

In general, for a hidden layer neuron, we can write:

$$z = \mathbf{w}^T\mathbf{x} + b, \tag{B.5}$$

$$a = \sigma(z), \tag{B.6}$$

where $\mathbf{x}$ is the input to the neuron in vector form, $\mathbf{w}$ are the weight parameters that connect the inputs to the neuron in vector form, b is a bias parameter, and $\sigma(\cdot)$ is a activation function. The output of the neuron, a, is called the "activation". It is the weight and bias parameters that are tuned iteratively during the optimisation process called "training".

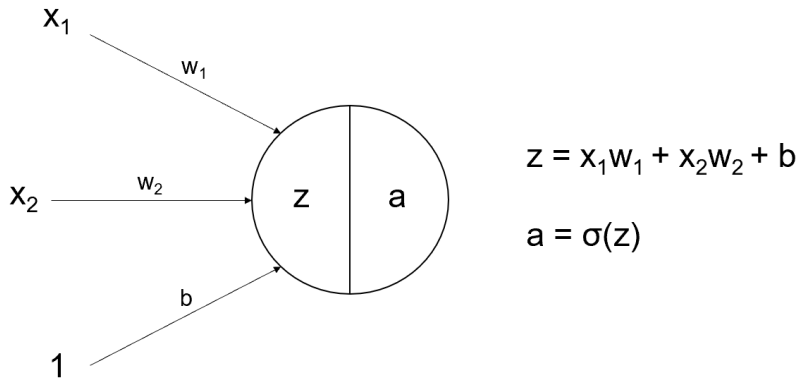


Figure B.2: Operations performed by a neuron with two inputs x_1 and x_2, two weight parameters w_1 and w_2, a bias parameter b, and an activation function $\sigma(\cdot)$.

Rather than considering individual neuron computations, it is desirable to consider layer computations; layer computations are easier to codify. For a hidden layer l, we can write:

$$\mathbf{z}^{[l]} = W^{[l]}\mathbf{x} + \mathbf{b}^{[l]}, \tag{B.7}$$

$$\mathbf{a}^{[l]} = \sigma(\mathbf{z}^{[l]}), \tag{B.8}$$

where:

- $W^{[l]}$ is a matrix of weight parameters of dimension $(n^{[l]}, n^{[l-1]})$ and $n^{[l]}$ is the

number of neurons in layer l;

- $\mathbf{b}^{[l]}$ is a vector biases of dimension $(n^{[l]}, 1)$;
- $\mathbf{z}^{[l]}$ is a vector of linear combinations of dimension $(n^{[l]}, 1)$;
- $\mathbf{a}^{[l]}$ is a vector of nonlinear activations of dimension $(n^{[l]}, 1)$.

When describing the depth of a neural network, layers with tuneable parameters are typically considered i.e., hidden layers and the output layer. For a network with L layers, there will be a total of $L-1$ hidden layers. Using the notation established above, we can write a general expression for the function computed by the hidden layers:

$$f(\mathbf{x}; W^{[1]}, \mathbf{b}^{[1]}, ..., W^{[L-1]}, \mathbf{b}^{[L-1]}) = \sigma(W^{[L-1]} ... \sigma(W^{[1]}\mathbf{x} + \mathbf{b}^{[1]}) ... + \mathbf{b}^{[L-1]}) \quad \text{(B.9)}$$

Following the hidden layer operations, a softmax function is used in the output layer to map the activations $\mathbf{a}$ from the final hidden layer to class probabilities:

$$P(y = i|\mathbf{a}; \mathbf{w}_1, b_1, ..., \mathbf{w}_k, b_k) = \text{softmax}(\mathbf{a}; \mathbf{w}_1, b_1, ..., \mathbf{w}_k, b_k) = \frac{\exp(\mathbf{w}_i^T \mathbf{a} + b_i)}{\sum_{k=1}^{K} \exp(\mathbf{w}_k^T \mathbf{a} + b_k)}, \quad \text{(B.10)}$$

where $\mathbf{w}_i$ is a vector of weight parameters and b_i is a bias parameter associated with the output neuron that corresponds to class i of K classes. The above formulation is specifically for classification tasks.

B.2.2 Training a neural network

"Training" in the context of neural networks refers to the process of iteratively updating the weight and bias parameters of the neural network to improve the neural networks current functional approximation. The key steps in the training process are:

1. Parameter initialisation

2. Forward propagation

3. Loss computation

4. Backward propagation and gradient descent

Steps 2 to 4 are repeated in a loop for some number of iterations. The number of training iterations may be set or based on some stopping criteria. At the end of each iteration, the weight and bias parameters are updated.

B.2.2.1 Initialisation

Initial values must be assigned to weight and bias parameters. These initial values will be updated during training. A simple method is random initialisation, where parameters are set to small random values. However, He initialisation [51] and Xavier initialisation [192] are used in the best performing modern neural networks. In both He and Xavier initialisation, parameters are drawn from a Gaussian distribution, but the variance is adjusted to suit either rectified linear unit (ReLu) activation or hyperbolic tangent (TanH) activation, respectively (see table B.1). For ReLu activation, it has been shown that the variance of the initialised Gaussian distributed weights in a layer l should be equal to $\frac{2}{n^{[l-1]}}$, whereas for TanH activation, the optimal variance has been shown to be $\frac{1}{n^{[l-1]}}$, where $n^{[l-1]}$ is the number of neurons in layer $l-1$.

B.2.2.2 Forward propagation

Forward propagation is the process of passing input data through the network to produce a network output (a distribution of class probabilities). The forward propagation step is necessary in order to produce a network output at the current estimate of the network parameters, which can be compared to ground truth labels to compute a loss.

Vectorised operations can be used to pass the entire dataset through the network at once. However, if the dataset is large, training can be slow even with vectorisation, hence the dataset is typically split into "mini-batches" of size m.

In Section B.2.1 above, we established notation to describe the two operations performed by a hidden layer l of a neural network for a single example in the dataset.

We now extend that notation for a mini-batch of size m. We can write:

$$Z^{[l]} = W^{[l]} A^{[l-1]} + \mathbf{b}^{[l]}, \tag{B.11}$$

$$A^{[l]} = \sigma^{[l]}(Z^{[l]}), \tag{B.12}$$

where:

- $W^{[l]}$ is a matrix of weight parameters of dimension $(n^{[l]}, n^{[l-1]})$, where $n^{[l]}$ is the number of neurons in layer l;
- $\mathbf{b}^{[l]}$ is a vector bias parameters of dimension $(n^{[l]}, 1)$;
- $Z^{[l]}$ is matrix of linear combinations of dimension $(n^{[l]}, m)$, where m is the mini-batch size;
- $A^{[l]}$ is a matrix of nonlinear activations of dimension $(n^{[l]}, m)$;
- $\sigma(\cdot)$ is some nonlinear activation function (see table B.1 for activation functions typically used in neural networks).

Activation function	**Formula**
Linear	$\sigma(x) = x$
Sigmoid	$\sigma(x) = 1/(1+e^{-x})$
Hyperbolic tangent (TanH)	$\sigma(x) = 2/(1+e^{-2x}) - 1$
Rectified linear unit (ReLu)	$\sigma(x) = \max(0, x)$
Leaky ReLu	$\sigma(x) = 1(x < 0)(\alpha x) + 1(x >= 0)(x)$ where α is a small constant

Table B.1: Common activation functions used in neural networks.

At the output layer, the activations are mapped to class probabilities using a softmax function. For a mini-batch of size m, a matrix $\hat{Y}$ of shape (K, m) will be output, where K is the number of classes.

B.2.2.3 Loss and backward propagation

The loss is a measure of how well the trained network approximates the true function that maps the input data to the output labels. The total loss is defined:

$$J(W^{[1]}, \mathbf{b}^{[1]}, ..., W^{[L]}, \mathbf{b}^{[L]}) = \frac{1}{m} \sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)}), \tag{B.13}$$

where m is the mini-batch size, $y^{(i)}$ is the label for training example i and $\hat{y}^{(i)}$ is the network output for training example i. Here the function L computes the loss for a single training example. Two common loss functions used for classification tasks are the squared error loss:

$$J(W^{[1]}, \mathbf{b}^{[1]}, ..., W^{[L]}, \mathbf{b}^{[L]}) = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2, \tag{B.14}$$

and the cross-entropy loss:

$$J(W^{[1]}, \mathbf{b}^{[1]}, ..., W^{[L]}, \mathbf{b}^{[L]}) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \ln \hat{y}^{(i)} + (1 - y^{(i)}) \ln(1 - \hat{y}^{(i)})]. \tag{B.15}$$

Backpropagation is an optimisation technique for loss minimisation by updating the weight and bias parameters based on the derivative of the loss function. This is called "gradient descent". For gradient descent in a layer l, the parameter update rule is defined:

$$W^{[l]} := W^{[l]} - \alpha \frac{dJ}{dW^{[l]}}, \tag{B.16}$$

$$\mathbf{b}^{[l]} := \mathbf{b}^{[l]} - \alpha \frac{dJ}{d\mathbf{b}^{[l]}}. \tag{B.17}$$

where α is a learning rate hyperparameter which dictates the size of the updates at each iteration. Intuitively, the parameters are being updated in the direction that reduces the loss J. Following each forward propagation of data through the network, backpropagation and parameter update will take place.

B.2.3 Optimisation

B.2.3.1 Gradient descent optimisation

There are three common variations of gradient descent that are used for training neural networks. An illustration of the three variations is shown in Figure B.3.

- **Batch gradient descent**: The parameter update rules (B.16) and (B.17) require calculation of the derivative of the loss function with respect to the parameters of the network. If the loss function (B.15) is calculated over ALL training examples rather than a mini-batch, then the derivatives in the parameter update rules are calculated over ALL training examples. This is called batch gradient descent. The advantage of batch gradient descent is stable parameter updates that will approach the minimum of the loss function more directly. However, if the training set is large, the computation of this derivatives is computationally expensive.

- **Online gradient descent**: If the mini-batch size is m = 1, then the loss function (B.15) and its derivative are calculated over a single training example at each iteration. This is called online gradient descent. The advantage of online gradient descent is low computational expense. The disadvantage is unstable parameter updates.

- **Mini-batch gradient descent**: This is the most commonly used form of gradient descent and sits between batch gradient descent and online gradient descent. In mini-batch gradient descent, the loss function (B.15) and its derivative are calculated over a mini-batch of some size $m > 1$. It offers medium stability and medium computational expense.

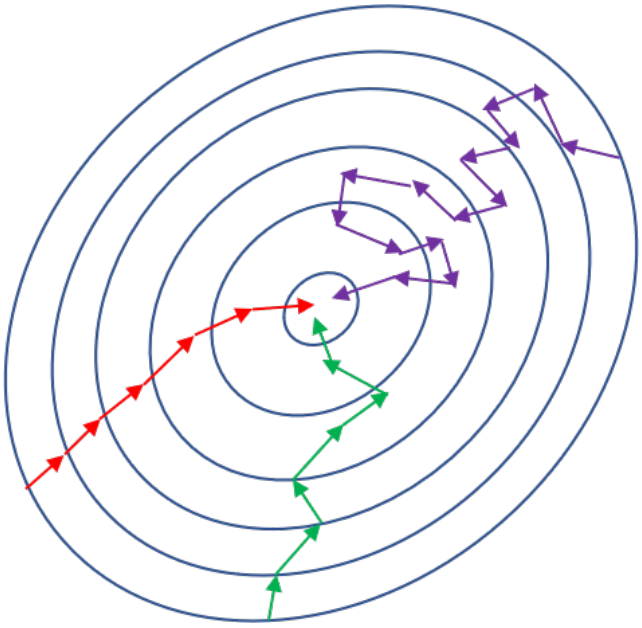


Figure B.3: An illustration of the trajectory of gradient descent towards the minimum of an error function. red: batch, green: mini-batch, purple: online.

B.2.3.2 Alternative optimisers

Alternative optimisers, based on exponentially-weighted moving averages, have been developed to improve upon standard gradient descent. An exponentially-weighted moving average is defined:

$$v_t = \beta v_{t-1} + (1-\beta)\vartheta_t, \tag{B.18}$$

where β is a hyperparameter that controls the number of terms over which the moving average is calculated and ϑ_t is the t-th term in the sequence. Roughly, $\frac{1}{1-\beta}$ equals the number of the past data points being averaged over e.g., if $\beta = 0.9$, then the moving average is computed over the past ten points.

- **Gradient descent with momentum**: The idea is to take an exponentially-weighted moving average of the gradients and use the exponentially-weighted average in the update rule. Using momentum, oscillations in the gradient descent are dampened, providing a more stable, faster, direct path to the minimum. For a training iteration, gradient descent with momentum is computed as follows:

  - Compute $\frac{dJ}{dW}$ and $\frac{dJ}{d\mathbf{b}}$ on the current mini-batch
  - Compute $v_{dW} = \beta v_{dW} + (1-\beta)\frac{dJ}{dW}$
  - Compute $v_{d\mathbf{b}} = \beta v_{d\mathbf{b}} + (1-\beta)\frac{dJ}{d\mathbf{b}}$
  - Compute $W := W - \alpha v_{dW}$ and $\mathbf{b} := \mathbf{b} - \alpha v_{d\mathbf{b}}$

  In the above description, the superscript indicating layer l has been omitted to simplify notation.

- **Root Mean Squared Propagation (RMSprop)**: The idea of RMSprop is to scale the learning rate α based on the recent history of the squared gradients. This will dampen large oscillations early in the training by reducing the gradient descent step size and will aid convergence by increasing the step size as the gradient update becomes small when the minimum is approached. RMSprop for an iteration is computed:

  - Compute $\frac{dJ}{dW}$ and $\frac{dJ}{d\mathbf{b}}$ on the current mini-batch
  - Compute $s_{dW} = \beta s_{dW} + (1-\beta)\frac{dJ}{dW}^2$
  - Compute $s_{d\mathbf{b}} = \beta s_{d\mathbf{b}} + (1-\beta)\frac{dJ}{d\mathbf{b}}^2$
  - Compute $W := W - \alpha \frac{\frac{dJ}{dW}}{\sqrt{s_{dW}+\varepsilon}}$ and $\mathbf{b} := \mathbf{b} - \alpha \frac{\frac{dJ}{d\mathbf{b}}}{\sqrt{s_{d\mathbf{b}}+\varepsilon}}$

  Here, ε is a small number to ensure numerical stability in division.

- **Adaptive Moment (ADAM) optimisation algorithm**: ADAM combines gradient descent with momentum and RMSprop. ADAM for an iteration is computed:

  - Compute $\frac{dJ}{dW}$ and $\frac{dJ}{d\mathbf{b}}$ on the current mini-batch
  - Compute v_{dW}, s_{dW}, $v_{d\mathbf{b}}$, $s_{d\mathbf{b}}$ as above in the descriptions of momentum and RMSProp

– Apply bias correction, so that:
$v_{dW}^{corrected} = \frac{v_{dW}}{1-\beta_1^t}$, $v_{d\mathbf{b}}^{corrected} = \frac{v_{d\mathbf{b}}}{1-\beta_1^t}$,
$s_{dW}^{corrected} = \frac{s_{dW}}{1-\beta_2^t}$, $s_{d\mathbf{b}}^{corrected} = \frac{s_{d\mathbf{b}}}{1-\beta_2^t}$.
Above, β_1 is the momentum β parameter, β_2 is the RMSProp β parameter, and t is the iteration number
– Compute $W := W - \alpha \frac{v_{dW}^{corrected}}{\sqrt{s_{dW}^{corrected} + \varepsilon}}$ and $\mathbf{b} := \mathbf{b} - \alpha \frac{v_{d\mathbf{b}}^{corrected}}{\sqrt{s_{d\mathbf{b}}^{corrected} + \varepsilon}}$

B.2.4 Batch normalisation

"Batch normalisation" [193] is an innovation that helps to optimise deep neural network architectures. The main motivating factor for batch normalisation is "internal covariate shift", which is defined in [193] as a "change in the distribution of network activations due to a change in network parameters during training". Batch normalisation reduces internal covariate shift by applying a normalisation to layer inputs that fixes the mean and variance of the layer inputs.

In subsection B.2.2.2, we described a matrix $Z^{[l]}$ of linear combinations of dimension $(n^{[l]}, m)$ where l is the layer, $n^{[l]}$ is the number of neurons in layer l, and m is the mini-batch size. In batch normalisation, for a layer l, we compute a mean and variance over the columns of this matrix, $\mathbf{z}^{(1)}, ..., \mathbf{z}^{(m)}$:

$$\mu = \frac{1}{m} \sum_i \mathbf{z}^{(i)}, \tag{B.19}$$

$$\sigma^2 = \frac{1}{m} \sum_i (\mathbf{z}^{(i)} - \mu)^2, \tag{B.20}$$

where μ and σ^2 are both vectors of dimension $(n^{[l]}, 1)$. We can then define a normalised quantity and a transformed quantity:

$$\mathbf{z}_{norm}^{(i)} = \frac{\mathbf{z}^{(i)} - \mu}{\sqrt{\sigma^2 + \varepsilon}}, \tag{B.21}$$

$$\tilde{\mathbf{z}}^{(i)} = \gamma \mathbf{z}_{norm}^{(i)} + \beta, \tag{B.22}$$

where ε is a small constant for numerical stability and β and γ are learnable parameter vectors that allow a distribution for the layer input that is not necessarily zero mean and unit variance. The column vectors $\tilde{\mathbf{z}}^{(1)}, ..., \tilde{\mathbf{z}}^{(m)}$ form the matrix $\tilde{Z}^{[l]}$, which is passed through an activation function $\sigma(\cdot)$ to give the output of layer l and the input to layer $l+1$. β and γ are of both of dimension $(n^{[l]}, 1)$ and updated by the rule:

$$\beta^{[l]} := \beta^{[l]} - \alpha \frac{dJ}{d\beta^{[l]}}, \tag{B.23}$$

$$\gamma^{[l]} := \gamma^{[l]} - \alpha \frac{dJ}{d\gamma^{[l]}}. \tag{B.24}$$

B.2.5 Hyperparameter selection in neural networks

One of the biggest challenges faced by practitioners of neural networks is hyperparameter selection. Hyperparameters are aspects of the neural network that are not "learned" during the neural network training process, but rather must be chosen by the practitioner. The key hyperparameters are:

- Number of hidden layers
- Number of nodes in each hidden layer (number of input and output nodes are defined by the problem)
- Activation function
- Initialisation regime
- Learning rate
- Mini-batch size
- Loss function
- Optimiser
- Parameters in the optimisation algorithm e.g., β in momentum

- Parameters in the explicit regularisation approach (see Section B.2.6)

- Parameters associated with other regularisation approaches e.g., dropout probability or data augmentation scaling percentage or rotation angle (see Section B.2.6)

Typically, practitioners will use a combination of experience and grid search to select hyperparameters. In grid search, values of each hyperparameter of interest will be defined along a unique axis. Neural networks will be evaluated with hyperparameters based on each intersecting point of the grid formed by the axes.

B.2.6 Regularisation

Above, we considered the problem of overfitting, where a network fits the training data too well, negatively impacting its ability to generalise. Goodfellow et al. [4] define regularisation as "any modification we make to a learning algorithm that is intended to reduce its generalisation error but not its training error." Several regularisation strategies exist, the most popular of which are described below.

B.2.6.1 Parameter norm penalties

A type of regularisation that involves adding a penalty to the loss J [4]:

$$J(W^{[1]}, \mathbf{b}^{[1]}, ..., W^{[L]}, \mathbf{b}^{[L]}) + \lambda\Omega(W^{[1]}, ..., W^{[L]}), \tag{B.25}$$

where $\lambda \in [0, \infty)$ is a hyperparameter to be set and Ω is a function of the network weights. It should be noted that the regularisation term only contains the weights and not the biases, as regularising bias parameters will cause significant underfitting [4]. One of the most common forms of penalty is the L2 norm penalty, also known as "weight decay", where $\Omega = \frac{1}{2}\mathbf{w}^T\mathbf{w}$ and $\mathbf{w}$ is a 1D vector of all the weights in the neural network. The goal of the L2 norm penalty is to keep the network weights small as overfitting often manifests as some weights in the neural network becoming too large.

B.2.6.2 Dataset augmentation

Neural networks and other machine learning algorithms are less likely to overfit if trained on a large dataset. However, the reality is that data is limited and it is not always possible to have more data. A possible solution that seems to work quite well is to augment the real data to create additional data. The augmented data would be given the same label as its unaugmented counterpart. There are many strategies to augment data e.g., for image data some typical augmentations are flipping the image along its axes, rotation, translation, scaling, and random deformation.

B.2.6.3 Early stopping

Early stopping is the most common form of regularisation in deep learning [4]. Generally, the error J on the training set will tend to decrease the longer the network trains. However, there will be a point at which the generalisability of the network will start to diminish i.e., overfitting. The point at which overfitting begins can be found by monitoring the error on the validation set. Using early stopping, training can be stopped at the point at which overfitting begins.

B.2.6.4 Dropout

As mentioned in Section B.2.6.1, overfitting can be due to some weights in the network becoming very large. Therefore, to reduce overfitting, the magnitude of the weights can be made smaller and more evenly distributed across the network. Dropout is a regularisation strategy that helps to achieve this by removing input and hidden neurons during each training iteration with some probability. Typically, an input unit is removed with probability 0.2 and a hidden unit is removed with probability 0.5 [4]. The input units and hidden units that are omitted are recalculated at every training iteration.

B.2.7 Convolutional neural networks

Convolutional neural networks (CNNs) are neural networks which are specialised for processing image data. CNNs use convolution operations to "learn" the most discriminative features from images rather than requiring "handcrafting" of features. They have been incredibly successful in recent years on a variety of computer

vision tasks.

B.2.7.1 Building blocks

A neural network is a CNN if at least one of its layers is a convolutional layer. A convolution is a mathematical operation of two functions f_1 and f_2, defined as the integral of the product of the two functions where one of the functions is reversed and shifted:

$$f_1 * f_2(x) = \int_{-\infty}^{\infty} f_1(\tau) f_2(x-\tau) d\tau \tag{B.26}$$

$$= \int_{-\infty}^{\infty} f_1(x-\tau) f_2(\tau) d\tau. \tag{B.27}$$

In the terminology of CNNs, the first argument (the function f_1 in this case) is called the "input" and the second argument (the function f_2 in this case) is called the "kernel", while the output is sometimes called a "feature map" [4]. Considering a 2D image input, we must consider discrete 2D convolutions. For a 2D image X and a 2D kernel W, a convolution operation between them is written [4]:

$$X * W(i,j) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} X(m,n) W(i-m, j-n). \tag{B.28}$$

Similarly, a 3D convolution operation is written:

$$X * W(i,j,k) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \sum_{o=-\infty}^{\infty} X(m,n,o) W(i-m, j-n, k-o). \tag{B.29}$$

Convolution kernels are composed of weight parameters, which through training, will learn to detect a feature.

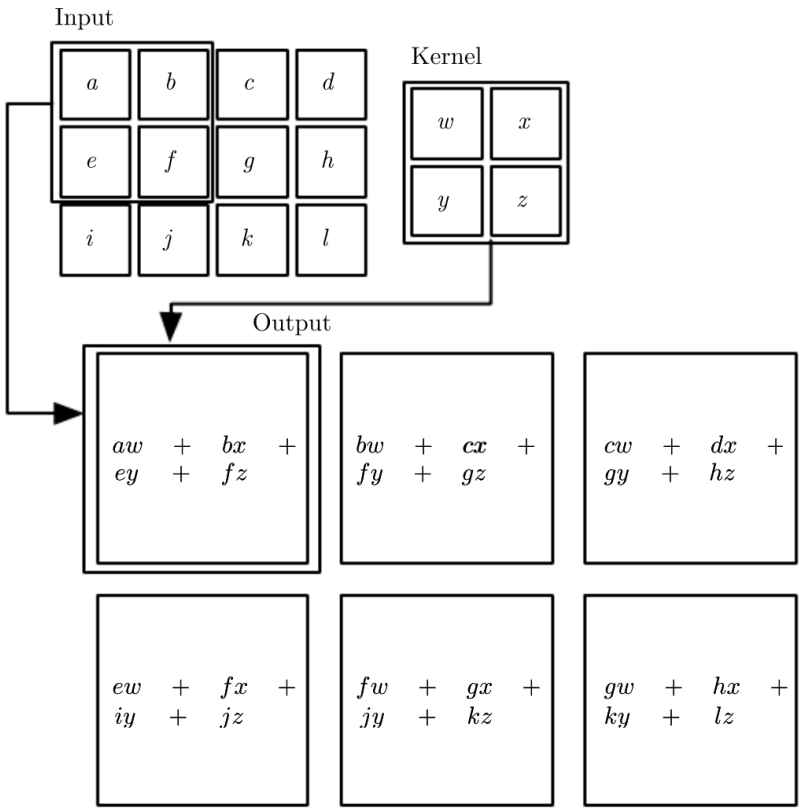


Figure B.4: An illustration of a 2D convolution operation using a single 2×2 kernel [4].

In each convolutional layer, several convolution kernels will be used to learn and detect features. More formally, the input to a layer is convolved with a set of K kernels $\{W_1, W_2, ..., W_K\}$ and added biases $\{b_1, ..., b_k\}$ [4]. Each kernel generates a new feature map X_k. Then, a non-linear function $\sigma(\cdot)$ e.g., ReLu operates on each feature map in an element-wise manner. The same process is repeated for every convolutional layer l:

$$X_k^{[l]} = \sigma(X^{[l-1]} * W_k^{[l]} + b_k^{[l]}). \tag{B.30}$$

For K convolutional kernels, K feature maps are produced. The feature maps are concatenated to produce the output of the layer.

A common point of confusion is whether 2D convolutions or 3D convolutions should be used for a CNN with 2D input and a CNN with 3D input. 2D convolutions are used for CNNs with 2D input and 3D convolutions are used for CNNs with 3D

input. 2D convolution kernels have dimension $f \times f \times d$ where d is the depth of the layer input. The output of such a convolution will have depth equal to one i.e., the output will be a 2D feature map. 3D convolution kernels have dimension $f \times f \times f$ where $f < d$. The output of such a convolution will have depth > 1 i.e., a 3D feature map.

Another non-trivial consideration is computing the size of the output of a convolutional layer. It is dependent on the number of kernels, kernel size, and two items that not yet covered, "stride" and "padding":

- The stride s dictates how many units/pixels/voxels to shift the kernel as the kernel makes its way across the layer input along each of its dimensions. Typically, a stride of 1 or 2 will be used. A stride of one will maintain input size and a stride of two will downsample the input by a factor of two.

- The padding p refers to a border of zeros placed around the layer input. Padding is used to ensure that units/pixels/voxels at the edge of an input get the same kernel exposure as units/pixels/voxels away from the edge.

We formulate the size of the layer output as a function of number of kernels, kernel size, stride, and padding. For a layer l, where the input to the layer has dimensions $n_h^{[l-1]} \times n_w^{[l-1]} \times n_d^{[l-1]} \times n_c^{[l-1]}$, the output of the layer will have dimensions $n_h^{[l]} \times n_w^{[l]} \times n_d^{[l]} \times n_c^{[l]}$, where:

$$n_{h,w,d}^{[l]} = \left\lfloor \frac{n_{h,w,d}^{[l]} + 2p^{[l]} - f^{[l]}}{s^{[l]}} + 1 \right\rfloor, \tag{B.31}$$

and $n_c^{[l]}$ is the number of kernels, $f^{[l]}$ is the kernel size along its height, width and depth, $p^{[l]}$ is the amount of padding, and $s^{[l]}$ is the kernel stride.

Pooling layers are commonly included in the convolutional portion of a CNN to progressively reduce the size of the input, so that higher-level features can be learnt, though convolutional layers with a stride > 1 can also be used to achieve a similar purpose. Pooling layers are placed intermittently between convolutional layers. For CNNs with a 2D input, where the input to the pooling layer consists of

2D feature maps, the most common form of pooling is "max pooling", whereby a max operation with a neighbourhood of size 2×2 is applied with a stride of two to the feature maps, downsampling the height and width of each feature map by a factor of two. In the case described, every max operation would be taking a max over four values. Another common pooling operation is "average pooling", where the average of each item in the neighbourhood is taken rather than the max.

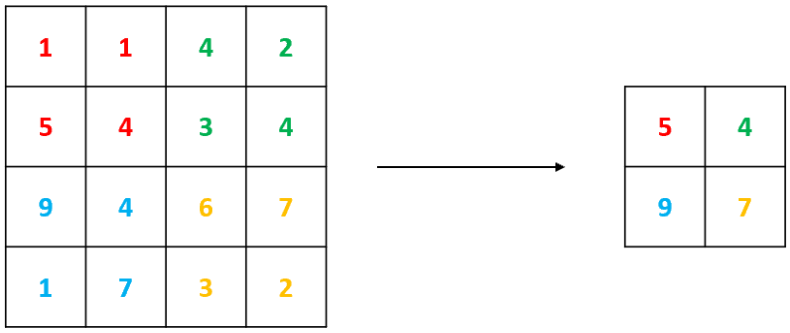


Figure B.5: An illustration of a 2D max pooling operation using a 2×2 neighbourhood and stride 2.

For CNNs with a 3D input, where the input to the pooling layer consists of 3D feature maps, the pooling operation would consider a neighbourhood of size $2 \times 2 \times 2$ and the max or average operation would be over eight values.

B.2.8 Convolutional neural network architectures

In this section, we describe two popular CNN architectures that feature in this thesis, namely ResNet [5] and U-Net [6], for classification and segmentation tasks respectively.

B.2.8.1 Residual Networks

Motivated by the observation that the performance of CNNs degrades as network depth is increased, He et al. [5] proposed Residual Networks (ResNets). In their paper, they showed that ResNets are easier to optimise and can gain accuracy from increased network depth. In their paper, He et al. present ResNets of increasing layer depth, from an 18-layer ResNet (ResNet-18) to a 152-layer ResNet (ResNet-152); ResNet-152 layers won the ImageNet classification challenge in 2015.

The degradation problem was explained as a difficulty in learning the identity mapping. Rather than learn a target function $\mathscr{H}(x)$ directly using a set of nonlinear layers, He et al. hypothesised that it is easier to fit the residual mapping $\mathscr{F}(x) := \mathscr{H}(x) - x$. The target function is recast into $\mathscr{F}(x) + x$, which can be realised by the addition of shortcut connections. A residual network building block is shown in Figure B.6. The shortcut connection skips over two layers in this case, to perform the identity mapping.

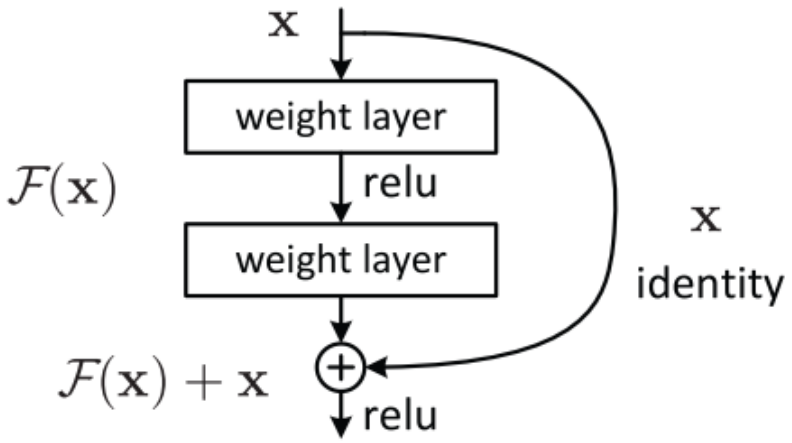


Figure B.6: Residual learning building block [5].

B.2.8.2 U-Net

In 2015, Ronneberger et al. introduced U-Net for biomedical image segmentation [6]. U-Net won the ISBI cell tracking challenge in 2015 by a large margin.

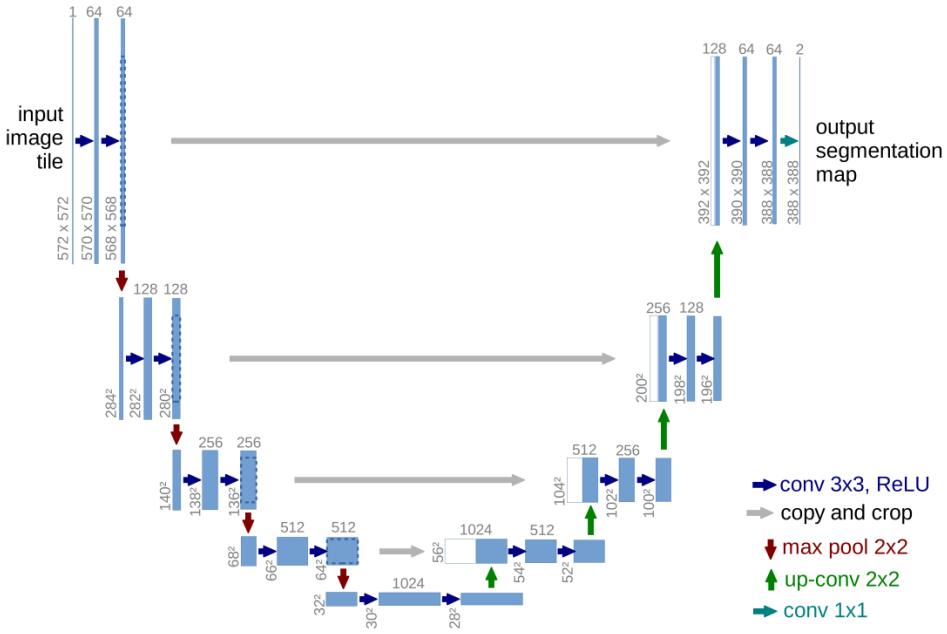


Figure B.7: U-Net architecture [6].

The network architecture is shown in Figure B.7. U-Net consists of an encoder path, a bottleneck block, and a decoder path, with skip connections joining encoder blocks to decoder blocks. The encoder path features four encoder blocks, where each encoder block is a set of two 3×3 convolutional layers, each followed by ReLU, followed by a stride two 2×2 max pooling operation for downsampling. At each downsampling, the number of feature maps are doubled. At the bottom of the U-shape is a bottleneck block which follows the structure of an encoder block with downsampling omitted. The decoder path features four decoder blocks, where each decoder block features an upsampling transposed convolution with a 2×2 kernel that halves the number of feature maps. Skip connections concatenate feature maps from the symmetric encoder block to the corresponding decoder block, followed by two 3×3 convolutional layers, each followed by ReLU. The final layer features 1×1 convolutions to map to output classes, followed by softmax. The idea behind U-Net is to encode the image into feature representations at multiple resolutions, and to decode the feature representations back to pixel space to produce a dense segmentation.

A 3D U-Net was later presented by Cicek et al. [7].

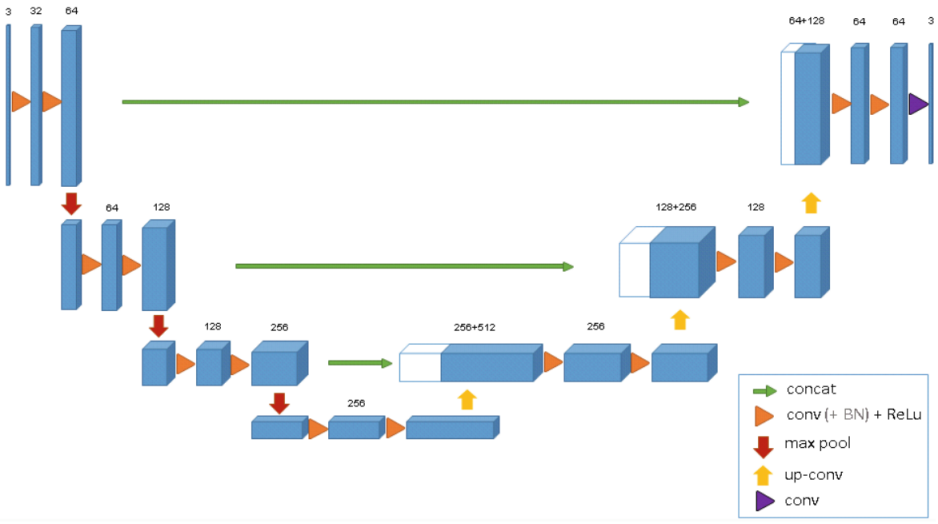


Figure B.8: 3D U-Net architecture [7].

The network architecture is shown in Figure B.8. Like U-Net, 3D U-Net has an encoder and decoder path. However, instead of 3×3 convolutions and 2×2 max pooling in the encoder blocks, 3D U-Net uses $3 \times 3 \times 3$ convolutions and $2 \times 2 \times 2$ max pooling. Furthermore, in the decoder blocks, $2 \times 2 \times 2$ kernels are used for upsampling, followed by $3 \times 3 \times 3$ convolutions, and in the final layers, $1 \times 1 \times 1$ convolutions for class mapping. A further change in 3D U-Net was the addition of batch normalisation [193] before ReLU.

Following the publications of Ronneberger et al. and Cicek et al., both 2D U-Net and 3D U-Net have been adapted either in pre-processing, network architecture, training methodology, or post-processing, for adaption to new tasks and hardware limitations. However, introducing bespoke changes requires high levels of experience, expertise, and manual effort, and does not guarantee optimal results [8]. As a result, Isensee et al. [8] published nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. A diagram of the nnU-Net framework is shown in Figure B.9.

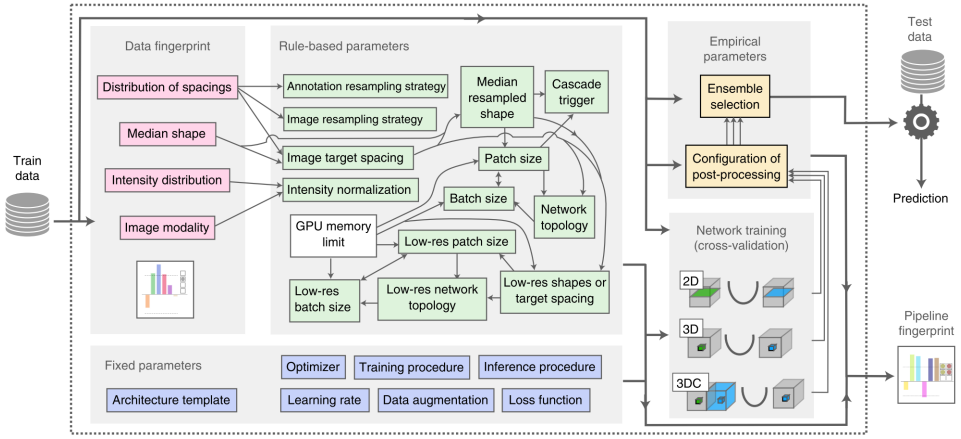


Figure B.9: nnU-Net framework [8].

Given a new segmentation task, properties of the dataset are extracted to give a "dataset fingerprint". The dataset fingerprint is used to inform the choice of pre-processing, model, and post-processing "rule-based parameters". However, some parameters are dataset-agnostic, referred to as "fixed parameters". Then, 2D, 3D, and/or 3D Cascade U-Nets are trained in a five-fold cross-validation. Finally, an empirical selection stage selects the best combination of models based on cross-validation performance. Without manual intervention, nnU-Net surpassed most existing approaches, including highly specialised solutions on 23 public datasets used in international biomedical segmentation competitions [8].

Bibliography

[1] David Preston. Magnetic Resonance Imaging (MRI) of the Brain and Spine: Basics, 2006.

[2] Robert Brown, Yu-Chung Cheng, Mark Haake, Michael Thompon, and Ramesh Venkatesen. *Magnetic Resonance Imaging: Physical Principles and Sequence Design*. John Wiley and Sons, Inc., 2014.

[3] Arden Dertat. Applied Deep Learning - Part 1: Artificial Neural Networks, 2017.

[4] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2015.

[5] Kaiming He, Xiangyu Zhang, Ren Shaoqing, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv*, 1512.03385, 2015.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[7] Ozgun Çiçek, Ahmed Abdulkadir, Soeren Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 424–432, 2016.

[8] Fabian Isensee, Paul F. Jaeger, Simon A.A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 2020.

[9] National Institute for Health and Care Excellence. Prostate cancer: diagnosis and management, 2019.

[10] 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.

[11] The Radiologist Shortage and the Potential of AI. `https://www.aidoc.com/blog/is-radiologist-shortage-real`. Accessed: 2021-07-24.

[12] The NHS AI Lab. `https://www.nhsx.nhs.uk/ai-lab`. Accessed: 2021-07-24.

[13] International Radiology Societies Tackle Radiologist Shortage. `https://www.rsna.org/news/2020/february/international-radiology-societies-and-shortage`. Accessed: 2022-01-10.

[14] Royal College of Radiologists. Clinical Radiology UK Workforce Census 2020 Report, 2021.

[15] Tom Syer, Pritesh Mehta, Michela Antonelli, Sue Mallett, David Atkinson, Sébastien Ourselin, and Shonit Punwani. Artificial intelligence compared to radiologists for the initial diagnosis of prostate cancer on magnetic resonance imaging: A systematic review and recommendations for future studies. *Cancers*, 13(13), 2021.

[16] The Prostate Gland. https://teachmeanatomy.info/pelvis/the-male-reproductive-system/prostate-gland. Accessed: 2021-08-01.

[17] What is Advanced Prostate Cancer? https://www.cancerresearchuk.org/about-cancer/prostate-cancer/advanced-cancer/about-advanced-cancer. Accessed: 2021-08-01.

[18] Prostate Cancer Symptoms. https://www.nhs.uk/conditions/prostate-cancer/symptoms. Accessed: 2021-08-01.

[19] Prostate Cancer Statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer. Accessed: 2021-08-01.

[20] Claudia Allemani, Hannah K. Weir, Helena Carreira, Rhea Harewood, Devon Spika, Xiao Si Wang, Finian Bannon, Jane V. Ahn, Christopher J. Johnson, Audrey Bonaventure, Rafael Marcos-Gragera, Charles Stiller, Gulnar Azevedo E Silva, Wan Qing Chen, Olufemi J. Ogunbiyi, Bernard Rachet, Matthew J. Soeberg, Hui You, Tomohiro Matsuda, Magdalena Bielska-Lasota, Hans Storm, Thomas C. Tucker, Michel P. Coleman, S. Bouzbid, M. Hamdi-Chérif, Z. Zaidi, E. Bah, R. Swaminathan, S. H. Nortje, C. D. Stefan, M. M. El Mistiri, S. Bayo, B. Malle, S. S. Manraj, R. Sewpaul-Sungkur, A. Fabowale, D. Bradshaw, N. I.M. Somdyala, M. Abdel-Rahman, L. Jaidane, M. Mokni, I. Kumcher, F. Moreno, M. S. González, E. Laura, F. V. Pugh, M. E. Torrent, B. Carballo Quintero, R. Fita, D. Garcilazo, P. L. Giacciani, M. C. Diumenjo, W. D. Laspada, M. A. Green, M. F. Lanza, S. G. Ibañez, C. A. Lima, E. Lobo, C. Daniel, C. Scandiuzzi, P. C.F. De Souza, K. Del Pino, C. Laporte, M. P. Curado, J. C. de Oliveira, C. L.A. Veneziano, D. B. Veneziano, T. S. Alexandre, A. S. Verdugo, S. Koifman, J. C. Galaz, J. A. Moya, D. A. Herrmann, A. M. Jofre, C. J. Uribe, L. E.

Bravo, G. Lopez Guarnizo, D. M. Jurado, M. C. Yepes, Y. H. Galán, P. Torres, F. Martínez-Reyes, L. Jaramillo, R. Quinto, P. Cueva, J. Yépez, C. R. Torres-Cintrón, G. Tortolero-Luna, R. Alonso, E. Barrios, C. Russell, L. Shack, A. J. Coldman, R. R. Woods, G. Noonan, D. Turner, E. Kumar, B. Zhang, F. R. McCrate, S. Ryan, H. Hannah, R. A.D. Dewar, M. MacIntyre, A. Lalany, M. Ruta, L. Marrett, D. E. Nishri, K. A. Vriends, C. Bertrand, R. Louchini, K. I. Robb, H. Stuart-Panko, S. Demers, S. Wright, J. George, X. Shen, J. T. Brockhouse, D. K. O'Brien, L. Almon, J. L. Young, J. Bates, R. Rycroft, L. Mueller, C. Phillips, H. Ryan, J. Walrath, A. Schwartz, F. Vigneau, J. A. MacKinnon, B. Wohler, R. Bayakly, K. C. Ward, K. Davidson-Allen, S. Glaser, D. West, M. D. Green, B. Y. Hernandez, C. F. Lynch, K. M. McKeen, B. Huang, D. Deapen, L. Liu, M. C. Hsieh, X. C. Wu, K. Stern, S. T. Gershman, R. C. Knowlton, G. Copeland, G. Spivak, D. B. Rogers, D. Lemons, L. L. Williamson, M. Hood, H. Jerry, G. M. Hosain, J. R. Rees, K. S. Pawlish, A. Stroup, C. Key, C. Wiggins, A. R. Kahn, M. J. Schymura, G. Leung, C. Rao, L. Giljahn, B. Warther, A. Pate, M. Patil, D. K. Shipley, M. Esterly, R. D. Otto, J. P. Fulton, D. L. Rousseau, T. A. Janes, S. M. Schwartz, S. W. Bolick, D. M. Hurley, R. A. Tenney, M. A. Whiteside, A. Hakeneweerth, M. A. Williams, K. Herget, C. Sweeney, J. Martin, S. Wang, M. G. Harrelson, M. B. Keitheri Cheteri, A. G. Hudson, R. Borchers, L. Stephenson, J. R. Espinoza, B. K. Edwards, N. Wang, L. Yang, J. S. Chen, G. H. Song, X. P. Gu, P. Zhang, H. M. Ge, D. L. Zhao, J. H. Zhang, F. D. Zhu, J. G. Tang, Y. Shen, J. Wang, Q. L. Li, S. P. Yang, J. M. Dong, W. W. Li, L. P. Cheng, J. G. Chen, Q. H. Huang, S. Q. Huang, G. P. Guo, K. Wei, W. Q. Chen, H. Zeng, A. V. Demetriou, P. Pavlou, W. K. Mang, K. C. Ngan, A. C. Kataki, M. Krishnatreya, P. A. Jayalekshmi, P. Sebastian, S. D. Sapkota, Y. Verma, A. Nandakumar, E. Suzanna, L. Keinan-Boker, B. G. Silverman, H. Ito, M. Hattori, H. Sugiyama, M. Utada, K. Katayama, S. Natsui, Y. Nishino, T. Koike, A. Ioka, K. Nakata, K. Kosa, I. Oki, A. Shibata, O. Nimri, A. Ab Manan, N. Bhoo Pathy, C. Ochir, S. Tuvshingerel,

A. M. Al Khater, H. Al-Eid, K. W. Jung, Y. J. Won, S. Park, C. J. Chiang, M. S. Lai, K. Suwanrungruang, S. Wiangnon, K. Daoprasert, D. Pongnikorn, S. L. Geater, H. Sriplung, S. Eser, C. I. Yakut, M. Hackl, N. Zielonke, H. Mühlböck, W. Oberaigner, M. Piñeros, A. A. Zborovskaya, K. Henau, L. Van Eycken, N. Dimitrova, Z. Valerianova, M. Šekerija, A. Znaor, M. Zvolský, G. Engholm, T. Aareleid, M. Mägi, N. Malila, K. Seppä, M. Velten, E. Cornet, X. Troussard, A. M. Bouvier, J. Faivre, A. V. Guizard, V. Bouvier, G. Launoy, P. Arveux, M. Maynadié, M. Mounier, A. S. Woronoff, M. Daoulas, J. Clavel, S. Le Guyader-Peyrou, A. Monnereau, B. Trétarre, M. Colonna, S. Delacour-Billon, F. Molinié, S. Bara, D. Degré, O. Ganry, B. Lapôtre-Ledoux, P. Grosclaude, J. M. Lutz, A. Belot, J. Estève, D. Forman, F. Sassi, R. Stabenow, A. Eberle, A. Nennecke, J. Kieschke, E. Sirri, H. Kajueter, K. Emrich, S. R. Zeissig, B. Holleczek, N. Eisemann, A. Katalinic, H. Brenner, R. A. Asquez, V. Kumar, E. J. Ólafsdóttir, L. Tryggvadóttir, H. Comber, P. M. Walsh, H. Sundseth, T. Dal Cappello, G. Mazzoleni, A. Giacomin, M. Castaing, S. Sciacca, A. Sutera, M. Corti, G. Gola, S. Ferretti, D. Serraino, A. Zucchetto, R. Lillini, M. Vercelli, S. Busco, F. Pannozzo, S. Vitarelli, P. Ricci, V. Pascucci, M. Autelitano, C. Cirilli, M. Federico, M. Fusco, M. F. Vitale, M. Usala, R. Cusimano, F. Vitale, M. Michiara, P. Sgargi, C. Sacerdote, R. Tumino, L. Mangone, F. Falcini, L. Cremone, M. Budroni, R. Cesaraccio, A. Madeddu, F. Tisano, S. Maspero, R. Tessandori, G. Candela, T. Scuderi, S. Piffer, S. Rosso, R. Zanetti, A. Caldarella, E. Crocetti, F. La Rosa, F. Stracci, P. Contiero, G. Tagliabue, P. Zambon, P. Baili, F. Berrino, G. Gatta, M. Sant, R. Capocaccia, R. De Angelis, A. Verdecchia, E. Liepina, A. Maurina, G. Smailyte, D. Agius, N. Calleja, S. Siesling, S. Larønningen, B. Møller, A. Dyzmann-Sroka, M. Trojanowski, S. Gózdz, R. Mezyk, M. Gradalska-Lampart, A. U. Radziszewska, J. Didkowska, U. Wojciechowska, J. Blaszczyk, K. Kepska, G. Forjaz, R. A. Rego, J. Bastos, L. Antunes, M. J. Bento, A. M. da Costa Miranda, A. Mayer-da Silva, D. Coza, A. I. Todescu, A. Krasilnikov, M. Valkov, J. Adamcik,

C. Safaei Diba, M. Primic Žakelj, T. Žagar, J. Stare, E. Almar, A. Mateos, M. V. Argüelles, J. R. Quirós, J. Bidaurrazaga, N. Larrañaga, J. M. Díaz García, A. I. Marcos, M. L. Vilardell Gil, E. Molina-Portillo, M. J. Sánchez, M. Ramos Montserrat, M. D. Chirlaque, C. Navarro, E. Ardanaz, S. Felipe Garcia, R. Peris-Bonet, J. Galceran, S. Khan, M. Lambe, B. Camey, C. Bouchardy, M. Usel, S. M. Ess, C. Hermann, F. G. Levi, M. Maspoli-Conconi, C. E. Kuehni, V. R. Mitter, A. Bordoni, A. Spitale, A. Chiolero, I. Konzelmann, S. I. Dehler, R. I. Laue, D. Meechan, J. Poole, D. Greenberg, J. Rashbass, E. Davies, K. Linklater, E. Morris, T. Moran, A. Gavin, R. J. Black, D. H. Brewster, M. Roche, S. McPhail, J. Verne, M. Murphy, D. W. Huws, C. White, G. Lawrence, C. Brook, J. Wilkinson, P. Finan, N. Sanz, X. S. Wang, R. Stephens, J. Butler, M. Peake, E. Chalker, L. Newman, D. Baker, C. Scott, B. C. Stokes, A. Venn, H. Farrugia, G. G. Giles, T. Threlfall, D. Currow, C. Lewis, and S. A. Miles. Global surveillance of cancer survival 1995-2009: Analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (CONCORD-2). *The Lancet*, 385(9972):977–1010, 2015.

[21] Should I have a PSA test? `https://www.nhs.uk/conditions/prostate-cancer/should-i-have-psa-test`. Accessed: 2021-08-01.

[22] Leen Naji, Harkanwal Randhawa, Zahra Sohani, Brittany Dennis, and Jason Profetto. Digital rectal examination for prostate cancer screening in primary care: A systematic review and meta-analysis. *Annals of Family Medicine*, 16(2):149–154, 2018.

[23] American College of Radiology. PI-RADS Version 2, 2015.

[24] American College of Radiology. PI-RADS Version 2.1, 2019.

[25] Mrishta Brizmohun Appayya, Jim Adshead, Hashim U. Ahmed, Clare Allen, Alan Bainbridge, Tristan Barrett, Francesco Giganti, John Graham, Phil

Haslam, Edward W. Johnston, Christof Kastner, Alexander P.S. Kirkham, Alexandra Lipton, Alan McNeill, Larissa Moniz, Caroline M. Moore, Ghulam Nabi, Anwar R. Padhani, Chris Parker, Amit Patel, Jacqueline Pursey, Jonathan Richenberg, John Staffurth, Jan van der Meulen, Darren Walls, and Shonit Punwani. National implementation of multi-parametric magnetic resonance imaging for prostate cancer detection – recommendations from a UK consensus meeting. *BJU International*, 122(1):13–25, 2018.

[26] 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, Alex P. Kirkham, Robert Oldroyd, Chris Parker, and Mark Emberton. 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.

[27] Caroline Hoeks, Jelle Barentsz, Thomas Hambrock, Derya Yakar, Diederik Somford, Stijn Heijmink, Tom Scheenen, Pieter Vos, Henkjan Huisman, Inge van Oort, J Alfred Witjes, Arend Heerschap, and Jurgen Fütterer. Prostate cancer: multiparametric MR imaging for detection, localization, and staging. *Radiology*, 261(1):46–66, 2011.

[28] Sadhna Verma, Saradwata Sarkar, Jason Young, Rajesh Venkataraman, Xu Yang, Anil Bhavsar, Nilesh Patil, James Donovan, and Krishnanath Gaitonde. Evaluation of the impact of computed high b-value diffusion-weighted imaging on prostate cancer detection. *Abdominal Radiology*, 41(5):934–945, 2016.

[29] Matthew D Blackledge, Martin O Leach, David J Collins, Dow-Mu Koh, Imaging May, Improve Tumor, Matthew D Blackledge, Martin O Leach, and David J Collins. Computed Diffusion-weighted MR Imaging May Improve Tumor Detection. *Radiology*, 261(2):573–581, 2011.

[30] 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*, 211(5):226–241, 2018.

[31] Louise Dickinson, Hashim U Ahmed, Clare Allen, Jelle O Barentsz, Brendan Carey, Jurgen J Futterer, Stijn W Heijmink, Peter J Hoskin, Alex Kirkham, Anwar R Padhani, Raj Persad, Philippe Puech, Shonit Punwani, Aslam S Sohaib, Bertrand Tombal, Arnauld Villers, Jan Van Der Meulen, and Mark Emberton. Magnetic Resonance Imaging for the Detection , Localisation , and Characterisation of Prostate Cancer: Recommendations from a European Consensus Meeting. *European Urology*, 59(4):477–494, 2011.

[32] TNM Staging. `https://www.cancerresearchuk.org/about-cancer/prostate-cancer/stages/tnm-staging`. Accessed: 2021-08-03.

[33] Lucy A.M. Simmons, Abi Kanthabalan, Manit Arya, Tim Briggs, Dean Barratt, Susan C. Charman, Alex Freeman, James Gelister, David Hawkes, Yipeng Hu, Charles Jameson, Neil McCartan, Caroline M. Moore, Shonit Punwani, Navin Ramachandran, Jan Van Der Meulen, Mark Emberton, and Hashim U. Ahmed. The PICTURE study: Diagnostic accuracy of multiparametric MRI in men requiring a repeat prostate biopsy. *British Journal of Cancer*, 116(9):1159–1165, 2017.

[34] Donald Gleason. Classification of Prostatic Carcinomas. *Cancer Chemotherapy Reports*, 50:125–128, 1966.

[35] Jonathan Epstein, Lars Egevad, Amin Mahul, Brett Delahunt, John Srigley, and Peter Humphrey. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma. *American Journal of Surgical Pathology*, 40(2):244–252, 2016.

[36] Prostate Cancer Treatment. `https://www.nhs.uk/conditions/prostate-cancer/treatment`. Accessed: 2021-08-04.

[37] Rafael Rocha Tourinho-Barbosa, Victor Srougi, Igor Nunes-Silva, Mohammed Baghdadi, Gregory Rembeyo, Sophie S. Eiffel, Eric Barret, Francois Rozet, Marc Galiano, Xavier Cathelineau, and Rafael Sanchez-Salas. Biochemical recurrence after radical prostatectomy: What does it mean? *International Braz J Urol*, 44(1):14–21, 2018.

[38] Channing J. Paller and Emmanuel S. Antonarakis. Management of biochemically recurrent prostate cancer after local therapy: Evolving standards of care and new directions. *Clinical Advances in Hematology and Oncology*, 11(1):14–23, 2013.

[39] Giorgio Brembilla, Paolo Dell'Oglio, Armando Stabile, Anna Damascelli, Lisa Brunetti, Silvia Ravelli, Giulia Cristel, Elena Schiani, Elena Venturini, Daniele Grippaldi, Vincenzo Mendola, Paola Maria Vittoria Rancoita, Antonio Esposito, Alberto Briganti, Francesco Montorsi, Alessandro Del Maschio, and Francesco De Cobelli. Interreader variability in prostate MRI reporting using Prostate Imaging Reporting and Data System version 2.1. *European Radiology*, 30:3383–3392, 2020.

[40] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, 42:60–88, 2017.

[41] Arnaldo Stanzione, Andrea Ponsiglione, Gianluca Armando Di Fiore, Stefano Giusto Picchi, Martina Di Stasi, Francesco Verde, Mario Petretta, Massimo Imbriaco, and Renato Cuocolo. Prostate Volume Estimation on MRI: Accuracy and Effects of Ellipsoid and Bullet-Shaped Measurements on PSA Density. *Academic Radiology*, 2020.

[42] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan H. Heckers, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv*, 1902.09063, 2019.

[43] Philipp Steiger and Harriet C. Thoeny. Prostate MRI based on PI-RADS version 2: How we review and report. *Cancer Imaging*, 16(1), 2016.

[44] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. ProstateX Challenge data, 2017.

[45] Tomas Sakinis, Fausto Milletari, Holger Roth, Panagiotis Korfiatis, Petro Kostandy, Kenneth Philbrick, Zeynettin Akkus, Ziyue Xu, Daguang Xu, and Bradley J. Erickson. Interactive segmentation of medical images through fully convolutional neural networks. *arXiv*, 1903.08205, 2019.

[46] Tom Mitchell. *Machine Learning*. McGraw Hill, 1997.

[47] What Is Machine Learning? `https://uk.mathworks.com/discovery/machine-learning.html`. Accessed: 2021-08-23.

[48] What Is Deep Learning? `https://www.ibm.com/cloud/learn/deep-learning`. Accessed: 2021-08-23.

[49] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[50] Vinod Nair and Geoffrey E Hinton. Rectified Linear Units Improve Restricted Boltzmann Machines. In *International Conference on Machine Learning*, 2010.

[51] Kaiming He, Xiangyu Zhang, Ren Shaoqing, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *arXiv*, 1502.01852, 2015.

[52] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1409.4842, 2015.

[53] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. *arXiv*, 1709.01507, 2017.

[54] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. *36th International Conference on Machine Learning, ICML 2019*, pages 10691–10700, 2019.

[55] Riyaj Uddin Khan, Mohammad Tanveer, and Ram Bilas Pachori. A novel method for the classification of Alzheimer's disease from normal controls using magnetic resonance imaging. *Expert Systems*, 38(1), 2021.

[56] C. Jongkreangkrai, Y. Vichianin, C. Tocharoenchai, and H. Arimura. Computer-aided classification of Alzheimer's disease based on support vector machine with combination of cerebral image features in MRI. *Journal of Physics: Conference Series*, 694(1), 2016.

[57] Ramon Casanova, Fang Chi Hsu, and Mark A. Espeland. Classification of Structural MRI Images in Alzheimer's Disease from the Perspective of Ill-Posed Problems. *PLoS ONE*, 7(10), 2012.

[58] Jayadeva, Reshma Khemchandani, and Suresh Chandra. *Twin Support Vector Machines*. Springer, 2017.

[59] Ehsan Hosseini-Asl, Georgy Gimel'farb, and Ayman El-Baz. Alzheimer's Disease Diagnostics by a Deeply Supervised Adaptable 3D Convolutional Network. *Frontiers in Bioscience*, 23:584–596, 2016.

[60] Joseph Antony, Kevin McGuinness, Noel E. O'Connor, and Kieran Moran. Quantifying radiographic knee osteoarthritis severity using deep convolutional neural networks. *Proceedings - International Conference on Pattern Recognition*, 0:1195–1200, 2016.

[61] Walter H L Pinaya, Ary Gadelha, Orla M Doyle, Cristiano Noto, André Zugman, Quirino Cordeiro, Andrea P Jackowski, Rodrigo A Bressan, and João R Sato. Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Nature Publishing Group*, 6(June), 2016.

[62] Jun Shi, Xiao Zheng, Yan Li, Qi Zhang, and Shihui Ying. Multimodal Neuroimaging Feature Learning With Multimodal Stacked Deep Polynomial Networks for Diagnosis of Alzheimer ' s Disease. *IEEE Journal of Biomedical and Health Informatics*, 22(1):173–183, 2018.

[63] Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Geert Litjens, Paul Gerke, Colin Jacobs, Sarah J. Van Riel, Mathilde Marie Winkler Wille, Matiullah Naqibullah, Clara I. Sanchez, and Bram Van Ginneken. Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks. *IEEE Transactions on Medical Imaging*, 35(5):1160–1169, 2016.

[64] Balazs Harangi. Skin lesion classification with ensembles of deep convolutional neural networks. *Journal of Biomedical Informatics*, 86(January):25–32, 2018.

[65] Marta Cullell-Dalmau, Sergio Noé, Marta Otero-Viñas, Ivan Meić, and Carlo Manzo. Convolutional Neural Network for Skin Lesion Classification: Understanding the Fundamentals Through Hands-On Learning. *Frontiers in Medicine*, 8(March), 2021.

[66] M Jorge Cardoso. Segmentation [lecture notes, UCL Medical Imaging CDT, 2018], 2018.

[67] Stuart P. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[68] H. P. Ng, S. H. Ong, K. W.C. Foong, P. S. Goh, and W. L. Nowinski. Medical image segmentation using k-means clustering and improved watershed algorithm. *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation*, 2006:61–65, 2006.

[69] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1977.

[70] Daozheng Chen. Expectation-Maximization Algorithm and Image Segmentation, 2008.

[71] Goo Rak Kwon, Dibash Basukala, Sang Woong Lee, Kun Ho Lee, and Moonsoo Kang. Brain image segmentation using a combination of expectation-maximization algorithm and watershed transform. *International Journal of Imaging Systems and Technology*, 26(3):225–232, 2016.

[72] Yufan He, Dong Yang, Holger Roth, Can Zhao, and Daguang Xu. DiNTS: Differentiable Neural Network Topology Search for 3D Medical Image Segmentation. *arXiv*, 2103.15954, 2021.

[73] Ali Hatamizadeh, Dong Yang, Holger Roth, and Daguang Xu. UNETR: Transformers for 3D Medical Image Segmentation. *arXiv: 2103.10504*, abs/2103.10504, 2021.

[74] Fausto Milletari, Nassir Navab, and Seyed-ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *3DV*, 2016.

[75] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, AnnetteKopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Bram van Ginneken, Michel Bilello, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc J. Gollub, Stephan H. Heckers, Henkjan Huisman, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Jennifer S. Goli Pernicka, Kawal Rhode, Catalina Tobon-Gomez, Eugene Vorontsov, Henkjan Huisman, James A. Meakin, Sebastien Ourselin, Manuel Wiesenfarth, Pablo Arbelaez, Byeonguk Bae, Sihong Chen, Laura Daza, Jianjiang Feng, Baochun He, Fabian Isensee, Yuanfeng Ji, Fucang Jia, Namkug Kim, Ildoo Kim, Dorit Merhof, Akshay Pai, Beomhee Park, Mathias Perslev, Ramin Rezaiifar, Oliver Rippel, Ignacio Sarasua, Wei Shen, Jaemin Son, Christian Wachinger, Liansheng Wang, Yan Wang, Yingda Xia, Daguang Xu, Zhanwei Xu, Yefeng Zheng, Amber L. Simpson, Lena Maier-Hein, and M. Jorge Cardoso. The Medical Segmentation Decathlon. *arXiv*, abs/2106.05735, 2021.

[76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume 2017, pages 6000–6010, 2017.

[77] Yuri Boykov and Gareth Funka-Lea. Graph cuts and efficient N-D image segmentation. *International Journal of Computer Vision*, 70(2):109–131, 2006.

[78] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[79] Zeynettin Akkus, Jiri Sedlar, Lucie Coufalova, Panagiotis Korfiatis, Timothy L. Kline, Joshua D. Warner, Jay Agrawal, and Bradley J. Erickson. Semi-automated segmentation of pre-operative low grade gliomas in magnetic resonance imaging. *Cancer Imaging*, 15(1), 2015.

[80] Leo Grady, Thomas Schiwietz, Shmuel Aharon, and Rüdiger Westermann. Random walks for interactive organ segmentation in two and three dimensions: Implementation and validation. In *Medical Image Computing and Computer-Assisted Intervention*, pages 773–780, 2005.

[81] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang. Deep Interactive Object Selection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 373–381, 2016.

[82] Eirikur Agustsson, Jasper R. Uijlings, and Vittorio Ferrari. Interactive full image segmentation by considering all regions jointly. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June(I):11614–11623, 2019.

[83] Guotai Wang, Wenqi Li, Maria A. Zuluaga, Rosalind Pratt, Premal A. Patel, Michael Aertsen, Tom Doel, Anna L. David, Jan Deprest, Sebastien Ourselin, and Tom Vercauteren. Interactive Medical Image Segmentation Using Deep Learning with Image-Specific Fine Tuning. *IEEE Transactions on Medical Imaging*, 37(7):1562–1573, 2018.

[84] Guotai Wang, Maria A. Zuluaga, Wenqi Li, Rosalind Pratt, Premal A. Patel, Michael Aertsen, Tom Doel, Anna L. David, Jan Deprest, Sebastien Ourselin, and Tom Vercauteren. DeepIGeoS: A Deep Interactive Geodesic Framework for Medical Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1559–1572, 2019.

[85] Michela Antonelli, M. Jorge Cardoso, Edward W. Johnston, Mrishta Brizmohun Appayya, Benoit Presles, Marc Modat, Shonit Punwani, and Sebastien

Ourselin. GAS: A genetic atlas selection strategy in multi-atlas segmentation framework. *Medical Image Analysis*, 52:97–108, 2019.

[86] X Yang, Y Lei, T Wang, X Jiang, A Jani, H Mao, W Curran, P Patel, T Liu, and B Wang. 3D prostate segmentation in MR image using 3D deeply supervised convolutional neural networks. *Medical Physics*, 45(6):582–583, 2018.

[87] Nader Aldoj, Federico Biavati, Florian Michallek, Sebastian Stober, and Marc Dewey. Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like U-net. *Scientific Reports*, 10(1), 2020.

[88] Renato Cuocolo, Arnaldo Stanzione, Anna Castaldo, Davide Raffaele De Lucia, and Massimo Imbriaco. Quality control and whole-gland, zonal and lesion annotations for the PROSTATEx challenge public dataset. *European Journal of Radiology*, 138:109647, 2021.

[89] Zhiqiang Tian, LiZhi Liu, and Baowei Fei. A fully automatic multi-atlas based segmentation method for prostate MR images. In *SPIE Medical Imaging*, volume 9413, 2015.

[90] Yangming Ou and Jimit Doshi. Multi-atlas segmentation of the prostate: A zooming process with robust registration and atlas selection. In *MICCAI Grand Challenge: Prostate MR Image Segmentation*, 2012.

[91] G Litjens, R Toth, W van de Ven, C Hoeks, S Kerkstra, B van Ginneken, G Vincent, G Guillard, N Birbeck, J Zhang, R Strand, F Malmberg, Y Ou, C Davatzikos, M Kirschner, F Jung, J Yuan, W Qiu, Q Gao, P E Edwards, B Maan, F van der Heijden, S Ghose, J Mitra, J Dowling, D Barratt, H Huisman, and A Madabhushi. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. *Medical Image Analysis*, 18(2):359–373, 2014.

[92] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[93] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv*, 1606.02147, 2016.

[94] Dong Kyu Lee, Deuk Jae Sung, Chang-Su Kim, Yuk Heo, Jeong Yoon Lee, Beom Jin Park, and Min Ju Kim. Three-Dimensional Convolutional Neural Network for Prostate MRI Segmentation and Comparison of Prostate Volume Measurements by Use of Artificial Neural Network and Ellipsoid Formula. *American Journal of Roentgenology*, 214(6):1229–1238, 2020.

[95] Nasr Makni, Nacim Betrouni, and Olivier Colot. Introducing spatial neighbourhood in Evidential C-Means for segmentation of multi-source images: Application to prostate multi-parametric MRI. *Information Fusion*, 19(1):61–72, 2014.

[96] Ahmad Algohary, Rakesh Shiradkar, Shivani Pahwa, Andrei Purysko, Sadhna Verma, Daniel Moses, Ronald Shnier, Anne Maree Haynes, Warick Delprado, James Thompson, Sreeharsha Tirumani, Amr Mahran, Ardeshir R. Rastinehad, Lee Ponsky, Phillip D. Stricker, and Anant Madabhushi. Combination of peri-tumoral and intra-tumoral radiomic features on bi-parametric mri accurately stratifies prostate cancer risk: A multi-site study. *Cancers*, 12(8), 2020.

[97] M Antonelli, E W Johnston, N Dikaios, K K Cheung, H S Sidhu, M B Appayya, F Giganti, L A M Simmons, A Freeman, C Allen, H U Ahmed, D Atkinson, S Ourselin, and S Punwani. Machine learning classifiers can predict Gleason pattern 4 prostate cancer with greater accuracy than experienced radiologists. *European Radiology*, 29(9):4754–4764, 2019.

[98] D Bonekamp, S Kohl, M Wiesenfarth, P Schelb, J P Radtke, M Gotz, P Kickingereder, K Yaqubi, B Hitthaler, N Gahlert, T A Kuder, F Deister, M Freitag, M Hohenfellner, B A Hadaschik, H P Schlemmer, and K H Maier-Hein. Radiomic Machine Learning for Characterization of Prostate Lesions with MRI: Comparison to ADC Values. *Radiology*, 289(1):128–137, 2018.

[99] Nikolaos Dikaios, Jokha Alkalbani, Harbir Singh Sidhu, Taiki Fujiwara, Mohamed Abd-Alazeez, Alex Kirkham, Clare Allen, Hashim Ahmed, Mark Emberton, Alex Freeman, Steve Halligan, Stuart Taylor, David Atkinson, and Shonit Punwani. Logistic regression model for diagnosis of transition zone prostate cancer on multi-parametric MRI. *European Radiology*, 25(2):523–532, 2015.

[100] A H Dinh, C Melodelima, R Souchon, P C Moldovan, F Bratan, G Pagnoux, F Mege-Lechevallier, A Ruffion, S Crouzet, M Colombel, and O Rouviere. Characterization of Prostate Cancer with Gleason Score of at Least 7 by Using Quantitative Multiparametric MR Imaging: Validation of a Computer-aided Diagnosis System in Patients Referred for Prostate Biopsy. *Radiology*, 287(2):525–533, 2018.

[101] T Hambrock, P C Vos, C A Hulsbergen-Van De Kaa, J O Barentsz, and H J Huisman. Prostate cancer: Computer-aided diagnosis with multiparametric 3-T MR imaging - Effect on observer performance. *Radiology*, 266(2):521–530, 2013.

[102] Y Iyama, T Nakaura, K Katahira, A Iyama, Y Nagayama, S Oda, D Utsunomiya, and Y Yamashita. Development and validation of a logistic regression model to distinguish transition zone cancers from benign prostatic hyperplasia on multi-parametric prostate MRI. *European Radiology*, 27(9):3600–3608, 2017.

[103] Mou Li, Ling Yang, Yufeng Yue, Jingxu Xu, Chencui Huang, and Bin Song. Use of Radiomics to Improve Diagnostic Performance of PI-RADS v2.1 in Prostate Cancer. *Frontiers in Oncology*, 10, 2021.

[104] G J Litjens, J O Barentsz, N Karssemeijer, and H J Huisman. Clinical evaluation of a computer-aided diagnosis system for determining cancer aggressiveness in prostate MRI. *European Radiology*, 25(11):3187–3199, 2015.

[105] Emilie Niaf, Carole Lartizien, Flavie Bratan, Laurent Roche, Muriel Rabilloud, Florence Mège-Lechevallier, and Olivier Rouvière. Prostate focal peripheral zone lesions: Characterization at multiparametric MR imaging-influence of a computer-aided diagnosis system1. *Radiology*, 271(3):761–769, 2014.

[106] X K Niu, Z F Chen, L Chen, J Li, T Peng, and X 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.

[107] S Transin, R Souchon, C Gonindard-Melodelima, R de Rozario, P Walker, M Funes de la Vega, R Loffroy, L Cormier, and O Rouviere. Computer-aided diagnosis system for characterizing ISUP grade ¿= 2 prostate cancers at multiparametric MRI: A cross-vendor evaluation. *Diagnostic and Interventional Imaging.*, 100:801–811, 2019.

[108] J Wang, C J Wu, M L Bao, J Zhang, X N Wang, and Y D Zhang. Machine learning-based analysis of MR radiomics can help to improve the diagnostic performance of PI-RADS v2 in clinically relevant prostate cancer. *European Radiology*, 27(10):4082–4090, 2017.

[109] David Jean Winkel, Hanns Christian Breit, Bibo Shi, Daniel T. Boll, Hans Helge Seifert, and Christian Wetterauer. Predicting clinically significant prostate cancer from quantitative image features including compressed

sensing radial MRI of prostate perfusion using machine learning: Comparison with PI-RADS v2 assessment scores. *Quantitative Imaging in Medicine and Surgery*, 10(4):808–823, 2020.

[110] Piotr Woźnicki, Niklas Westhoff, Thomas Huber, Philipp Riffel, Matthias F. Froelich, Eva Gresser, Jost von Hardenberg, Alexander Mühlberg, Maurice Stephan Michel, Stefan O. Schoenberg, and Dominik Nörenberg. Multiparametric MRI for prostate cancer characterization: Combined use of radiomics model with PI-RADS and clinical parameters. *Cancers*, 12(7), 2020.

[111] X Zhong, R Cao, S Shakeri, F Scalzo, Y Lee, D R Enzmann, H H Wu, S S Raman, and K Sung. Deep transfer learning-based prostate cancer classification using 3 Tesla multi-parametric MRI. *Abdominal Radiology*, 44(6):2030–2039, 2019.

[112] 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.

[113] S Gaur, N Lay, S A Harmon, S Doddakashi, S Mehralivand, B Argun, T Barrett, S Bednarova, R Girometti, E Karaarslan, A R Kural, A Oto, A S Purysko, T Antic, C Magi-Galluzzi, Y Saglican, S Sioletic, A Y Warren, L Bittencourt, J J Futterer, R T Gupta, I Kabakus, Y M Law, D J Margolis, H Shebel, A C Westphalen, B J Wood, P A Pinto, J H Shih, P L Choyke, R M Summers, and B Turkbey. Can computer-aided diagnosis assist in the identification of prostate cancer on prostate MRI? A multi-center, multi-reader investigation. *Oncotarget*, 9(73):33804–33817, 2018.

[114] V Giannini, S Mazzetti, E Armando, S Carabalona, F Russo, A Giacobbe, G Muto, and D Regge. Multiparametric magnetic resonance imaging of the

prostate with computer-aided detection: experienced observer performance study. *European Radiology*, 27(10):4200–4208, 2017.

[115] Matthew D. Greer, Nathan Lay, Joanna H. Shih, Tristan Barrett, Leonardo Kayat Bittencourt, Samuel Borofsky, Ismail Kabakus, Yan Mee Law, Jamie Marko, Haytham Shebel, Francesca V. Mertan, Maria J. Merino, Bradford J. Wood, Peter A. Pinto, Ronald M. Summers, Peter L. Choyke, and Baris Turkbey. Computer-aided diagnosis prior to conventional interpretation of prostate mpMRI: an international multi-reader study. *European Radiology*, 28(10):4407–4417, 2018.

[116] 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.

[117] Sherif Mehralivand, Stephanie Harmon, Joanna Shih, Clayton Smith, Nathan Lay, Burak Argun, Sandra Bednarova, Ronald Baroni, Abdullah Canda, and Baris Turkbey. Multicenter Multireader Evaluation of an Artificial Intelligence–Based Attention Mapping System for the Detection of Prostate Cancer With Multiparametric MRI. *Genitourinary Imaging*, 215:903–912, 2020.

[118] Patrick Schelb, Simon Kohl, Jan Philipp Radtke, Manuel Wiesenfarth, Philipp Kickingereder, Sebastian Bickelhaupt, Tristan Anselm Kuder, Albrecht Stenzinger, Markus Hohenfellner, Heinz-Peter Schlemmer, Klaus H. Maier-Hein, and David Bonekamp. Classification of Cancer at Prostate MRI: Deep Learning versus Clinical PI-RADS Assessment. *Radiology*, 293(3):607–617, 2019.

[119] Patrick Schelb, Xianfeng Wang, Jan Philipp Radtke, Manuel Wiesenfarth, Philipp Kickingereder, Albrecht Stenzinger, Markus Hohenfellner, Heinz Peter Schlemmer, Klaus H. Maier-Hein, and David Bonekamp. Simulated clinical deployment of fully automatic deep learning for clinical prostate MRI assessment. *European Radiology*, 2020.

[120] Anika Thon, Ulf Teichgräber, Cornelia Tennstedt-Schenk, Stathis Hadjidemetriou, Sven Winzler, Ansgar Malich, and Ismini Papageorgiou. Computer aided detection in prostate cancer diagnostics: A promising alternative to biopsy? A retrospective study from 104 lesions with histological ground truth. *PLoS ONE*, 12(10), 2017.

[121] Lina Zhu, Ge Gao, Yi Liu, Chao Han, Jing Liu, Xiaodong Zhang, and Xiaoying Wang. Feasibility of integrating computer-aided diagnosis with structured reports of prostate multiparametric MRI. *Clinical Imaging*, 60(1):123–130, 2020.

[122] Dominik Deniffel, Nabila Abraham, Khashayar Namdar, Xin Dong, Emmanuel Salinas, Laurent Milot, Farzad Khalvati, and Masoom A. Haider. Using decision curve analysis to benchmark performance of a magnetic resonance imaging–based deep learning model for prostate cancer risk assessment. *European Radiology*, 30:6867–6876, 2020.

[123] Anindo Saha, Matin Hosseinzadeh, and Henkjan Huisman. End-to-end Prostate Cancer Detection in bpMRI via 3D CNNs: Effects of Attention Mechanisms, Clinical Priori and Decoupled False Positive Reduction. *Medical Image Analysis*, 73:102155, 2021.

[124] Valentina Giannini, Simone Mazzetti, Giovanni Cappello, Valeria Maria Doronzio, Lorenzo Vassallo, Filippo Russo, Alessandro Giacobbe, Giovanni Muto, and Daniele Regge. Computer-aided diagnosis improves the detection of clinically significant prostate cancer on multiparametric-mri: A multi-observer performance study involving inexperienced readers. *Diagnostics*, 11(6), 2021.

[125] Mark Brown. UK prostate cancer screening programme "could be running in three years", 2021.

[126] David A. Bluemke, Linda Moy, Miriam A. Bredella, Birgit B. Ertl-Wagner, Kathryn J. Fowler, Vicky J. Goh, Elkan F. Halpern, Christopher P. Hess,

Mark L. Schiebler, and Clifford R. Weiss. Assessing radiology research on artificial intelligence: A brief guide for authors, reviewers, and readers-From the Radiology Editorial Board. *Radiology*, 294(2):487–489, 2020.

[127] Lucy A.M. Simmons, Hashim Uddin Ahmed, Caroline M. Moore, Shonit Punwani, Alex Freeman, Yipeng Hu, Dean Barratt, Susan C. Charman, Jan Van der Meulen, and Mark Emberton. The PICTURE study - prostate imaging (multi-parametric MRI and Prostate HistoScanning™) compared to transperineal ultrasound guided biopsy for significant prostate cancer risk evaluation. *Contemporary Clinical Trials*, 37(1):69–83, 2014.

[128] Nancy N Wang, Richard E Fan, John T Leppert, Pejman Ghanouni, Christian A Kunder, James D Brooks, Benjamin I Chung, and Geoffrey A Sonn. Performance of multiparametric MRI appears better when measured in patients who undergo radical prostatectomy. *Research and Reports in Urology*, 10:233–235, nov 2018.

[129] Peter Steenbergen, Karin Haustermans, Evelyne Lerut, Raymond Oyen, Liesbeth De Wever, Laura Van Den Bergh, Linda G.W. Kerkmeijer, Frank A. Pameijer, Wouter B. Veldhuis, Jochem R.N. Van Der Voort Van Zyp, Floris J. Pos, Stijn W. Heijmink, Robin Kalisvaart, Hendrik J. Teertstra, Cuong V. Dinh, Ghazaleh Ghobadi, and Uulke A. Van Der Heide. Prostate tumor delineation using multiparametric magnetic resonance imaging: Inter-observer variability and pathology validation. *Radiotherapy and Oncology*, 115(2):186–190, 2015.

[130] Pritesh Mehta, Michela Antonelli, Hashim U. Ahmed, Mark Emberton, Shonit Punwani, and Sébastien Ourselin. Computer-aided diagnosis of prostate cancer using multiparametric MRI and clinical features: A patient-level classification framework. *Medical Image Analysis*, 73:102153, 2021.

[131] Samuel Borofsky, Arvin K. George, Sonia Gaur, Marcelino Bernardo, Matthew D. Greer, Francesca V. Mertan, Myles Taffel, Vanesa Moreno,

Maria J. Merino, Bradford J. Wood, Peter A. Pinto, Peter L. Choyke, and Baris Turkbey. What Are We Missing? False-negative Cancers at Multiparametric MR Imaging of the Prostate. *Radiology*, 286(1):186–195, 2017.

[132] Wenqi Li, Guotai Wang, Lucas Fidon, Sebastien Ourselin, M. Jorge Cardoso, and Tom Vercauteren. On the compactness, efficiency, and representation of 3D convolutional networks: Brain parcellation as a pretext task. In *Information Processing in Medical Imaging (IPMI 2017)*, volume 10265, pages 348–360, 2017.

[133] Massimo De Luca, Valentina Giannini, Anna Vignati, Simone Mazzetti, Christian Bracco, Michele Stasi, Enrico Armando, Filippo Russo, Enrico Bollito, Francesco Porpiglia, and Daniele Regge. A fully automatic method to register the prostate gland on T2-weighted and EPI-DWI images. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 8029–8032. IEEE, 2011.

[134] Marc Modat, David M. Cash, Pankaj Daga, Gavin P. Winston, John S. Duncan, and Sébastien Ourselin. Global image registration using a symmetric block-matching approach. *Journal of Medical Imaging*, 1(2), 2014.

[135] Marc Modat, Gerard R. Ridgway, Zeike A. Taylor, Manja Lehmann, Josephine Barnes, David J. Hawkes, Nick C. Fox, and Sébastien Ourselin. Fast free-form deformation using graphics processing units. *Computer Methods and Programs in Biomedicine*, 98(3):278–284, 2010.

[136] Pascal Cachier, Eric Bardinet, Didier Dormont, Xavier Pennec, and Nicholas Ayache. Iconic feature based nonrigid registration: The PASHA algorithm. *Computer Vision and Image Understanding*, 89:272–298, 2003.

[137] Bashar Zelhof, Martin Lowry, Greta Rodrigues, Sigurd Kraus, and Lindsay Turnbull. Description of magnetic resonance imaging-derived enhancement variables in pathologically confirmed prostate cancer and normal peripheral zone regions. *BJU International*, 104(5):621–627, 2009.

[138] Olga A. Kubassova, Roger D. Boyle, and Aleksandra Radjenovic. Quantitative Analysis of Dynamic Contrast-Enhanced MRI Datasets of the Metacarpophalangeal Joints. *Academic Radiology*, 14(10):1189–1200, 2007.

[139] Jussi Toivonen, Ileana Montoya Perez, Parisa Movahedi, Harri Merisaari, Marko Pesola, Pekka Taimen, Peter J. Boström, Jonne Pohjankukka, Aida Kiviniemi, Tapio Pahikkala, Hannu J. Aronen, and Ivan Jambor. Radiomics and machine learning of multisequence multiparametric prostate MRI: Towards improved non-invasive prostate cancer characterization. *PLoS ONE*, 14(7), 2019.

[140] László G. Nyúl, Jayaram K. Udupa, and Xuan Zhang. New variants of a method of MRI scale standardization. *IEEE Transactions on Medical Imaging*, 19(2):143–150, 2000.

[141] Kaiming He, Xiangyu Zhang, Ren Shaoqing, and Jian Sun. Deep Residual Learning for Image Recognition. *arXiv*, 1512.03385, 2015.

[142] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. *arXiv*, 1603.05027, 2016.

[143] M Efroymson. Stepwise regression—a backward and forward look. *Eastern Regional Meetings of the Institute of Mathematical Statistics*, 1966.

[144] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[145] Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[146] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[147] Andrzej S. Kosinski. A weighted generalized score statistic for comparison of predictive values of diagnostic tests. *Statistics in Medicine*, 32(6), 2013.

[148] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*, 2015.

[149] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of Machine Learning Research*, 2017.

[150] N F Haq, P Kozlowski, E C Jones, S D Chang, S L Goldenberg, and M Moradi. A data-driven approach to prostate cancer detection from dynamic contrast enhanced MRI. *Computerized Medical Imaging & Graphics*, 41:37–45, 2015.

[151] Pritesh Mehta, Michela Antonelli, Saurabh Singh, Natalia Grondecka, Edward W. Johnston, Hashim U. Ahmed, Mark Emberton, Shonit Punwani, and Sébastien Ourselin. AutoProstate: Towards Automated Reporting of Prostate MRI for Prostate Cancer Assessment Using Deep Learning. *Cancers*, 13(23), 2021.

[152] Arnaldo Stanzione, Andrea Ponsiglione, Gianluca Armando Di Fiore, Stefano Giusto Picchi, Martina Di Stasi, Francesco Verde, Mario Petretta, Massimo Imbriaco, and Renato Cuocolo. Prostate Volume Estimation on MRI: Accuracy and Effects of Ellipsoid and Bullet-Shaped Measurements on PSA Density. *Academic Radiology*, 28(8):219–226, 2021.

[153] Florian A Distler, Jan P Radtke, David Bonekamp, Claudia Kesch, Heinz-Peter Schlemmer, Kathrin Wieczorek, Marietta Kirchner, Sascha Pahernik, Markus Hohenfellner, and Boris A Hadaschik. The Value of PSA Density in Combination with PI-RADS™ for the Accuracy of Prostate Cancer Prediction. *The Journal of urology*, 198(3):575–582, sep 2017.

[154] Wentao Zhu, Yufang Huang, Liang Zeng, Xuming Chen, Yong Liu, Zhen Qian, Nan Du, Wei Fan, and Xiaohui Xie. AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Medical Physics*, 46(2):576–589, 2019.

[155] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *33rd International Conference on Machine Learning, ICML 2016*, volume 48, pages 1651–1660, 2016.

[156] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

[157] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv*, 1701.02096, 2016.

[158] S. Ourselin, A. Roche, G. Subsol, X. Pennec, and N. Ayache. Reconstructing a 3D structure from serial histological sections. *Image and Vision Computing*, 19:25–31, 2001.

[159] Matin Hosseinzadeh, Patrick Brand, and Henkjan Huisman. Effect of Adding Probabilistic Zonal Prior in Deep Learning-based Prostate Cancer Detection. In *Medical Imaging with Deep Learning (MIDL) 2019*, 2019.

[160] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian SegNet : Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[161] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699, 2002.

[162] Tsung-yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[163] N Ghavami, Y Hu, E Gibson, E Bonmati, M Emberton, C M Moore, and D C Barratt. Automatic segmentation of prostate MRI using convolutional neural networks: Investigating the impact of network architecture on the accuracy of volume measurement and MRI-ultrasound registration. *Medical Image Analysis*, 58, 2019.

[164] Daniel J. Morgan, Lisa Pineles, Jill Owczarzak, Larry Magder, Laura Scherer, Jessica P. Brown, Chris Pfeiffer, Chris Terndrup, Luci Leykum, David Feldstein, Andrew Foy, Deborah Stevens, Christina Koch, Max Masnick, Scott Weisenberg, and Deborah Korenstein. Accuracy of practitioner estimates of probability of diagnosis before and after testing. *JAMA Internal Medicine*, 181:747–755, 6 2021.

[165] Andres Diaz-Pinto, Pritesh Mehta, Sachidanand Alle, Muhammad Asad, Richard Brown, Vishwesh Nath, Alvin Ihsani, Michela Antonelli, Daniel Palkovics, Csaba Pinter, Ron Alkalay, Steve Pieper, Holger R. Roth, Daguang Xu, Prerna Dogra, Tom Vercauteren, Andrew Feng, Abood Quraini, Sebastien Ourselin, and M. Jorge Cardoso. Deepedit: Deep editable learning for interactive segmentation of 3d medical images. volume 13567. Springer Nature Switzerland, 2022.

[166] Vasilis Stavrinides, Francesco Giganti, Mark Emberton, and Caroline M. Moore. MRI in active surveillance: a critical review. *Prostate Cancer and Prostatic Diseases*, 2018.

[167] Jose Marenco, Clement Orczyk, Tom Collins, Caroline Moore, and Mark Emberton. Role of MRI in planning radical prostatectomy: what is the added value? *World Journal of Urology*, 37(7):1289–1292, 2019.

[168] Luke Nicholls, Yae-eun Suh, Ewan Chapman, Daniel Henderson, Caroline Jones, Kirsty Morrison, Aslam Sohaib, Helen Taylor, Alison Tree, and Nicholas Van As. Clinical and Translational Radiation Oncology Stereotactic radiotherapy with focal boost for intermediate and high-risk prostate cancer

: Initial results of the SPARC trial. *Clinical and Translational Radiation Oncology*, 25:88–93, 2020.

[169] Alexandru Patriciu, Doru Petrisor, Michael Muntener, Dumitru Mazilu, Michael Schar, and Dan Stoianovici. Automatic Brachytherapy Seed Placement Under MRI Guidance. *IEEE Transactions on Biomedical Engineering*, 54(8):1499–1506, 2007.

[170] M J Connor, M A Gorin, H U Ahmed, and R Nigam. Focal therapy for localized prostate cancer in the era of routine multi-parametric MRI. *Prostate Cancer and Prostatic Diseases*, 23:232–243, 2020.

[171] K. K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool. Deep Extreme Cut: From Extreme Points to Object Segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 616–625, 2018.

[172] Alberto Briganti, Nicola Fossati, James W.F. Catto, Philip Cornford, Francesco Montorsi, Nicolas Mottet, Manfred Wirth, and Hendrik Van Poppel. Active Surveillance for Low-risk Prostate Cancer: The European Association of Urology Position in 2018. *European Urology*, 74(3):357–368, 2018.

[173] Linda G. W. Kerkmeijer, Veerle H. Groen, Floris J. Pos, Karin Haustermans, Evelyn M. Monninkhof, Robert Jan Smeenk, Martina Kunze-Busch, Johannes C. J. de Boer, Jochem van der Voort van Zijp, Marco van Vulpen, Cédric Draulans, Laura van den Bergh, Sofie Isebaert, and Uulke A. van der Heide. Focal boost to the intraprostatic tumor in external beam radiotherapy for patients with localized prostate cancer: Results from the flame randomized phase iii trial. *Journal of Clinical Oncology*, 39(7):787–796, 2021. PMID: 33471548.

[174] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Mesko. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *npj Digital Medicine*, 3(118), 2020.

[175] Kicky van Leeuwen, Steven Schalekamp, Matthieu Rutten, Bram van Ginneken, and Maarten de Rooij. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *European Radiology*, 31:3797–3804, 2021.

[176] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems (NIPS 2017)*, pages 5575–5585, 2017.

[177] Zach Eaton-Rosen, Felix Bragman, Sotirios Bisdas, Sébastien Ourselin, and M. Jorge Cardoso. Towards safe deep learning: Accurately quantifying biomarker uncertainty in neural network predictions. *arXiv*, 1806.08640, 2018.

[178] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv*, 1810.04805, 2019.

[179] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Sam McCandish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *arXiv*, 2005.14165, 2020.

[180] Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang B. Multimodal Recurrent Model with Attention for Automated Radiology Report Generation. In *MICCAI*, volume 10433, pages 457–466. Springer International Publishing, 2017.

[181] Davood Karimi, Haoran Dou, Simon K. Warfield, and Ali Gholipour. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. *arXiv*, 1912.02911, 2020.

[182] Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. HeMIS: Hetero-modal image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI*, 2016.

[183] Reuben Dorent, Samuel Joutard, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Hetero-Modal Variational Encoder-Decoder for Joint Modality Completion and Segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI*, 2019.

[184] Thomas Kustner, Karim Armanious, Jiahuan Yang, Bin Yang, Fritz Schick, and Sergios Gatidis. Retrospective correction of motion-affected MR images using deep learning frameworks. *Magnetic Resonance in Medicine*, 82:1527–1540, 2019.

[185] Jongyeon Lee, Byungjai Kim, and HyunWook Park. MC2- Net: motion correction network for multi- contrast brain MRI. *Magnetic Resonance in Medicine*, 86:1077–1092, 2020.

[186] Ben A. Duffy, Lu Zhao, Farshid Sepehrband, Joyce Min, Danny JJ Wang, Yonggang Shi, Arthur W. Toga, and Hosung Kim. Retrospective motion artifact correction of structural MRI images using deep learning improves the quality of cortical surface reconstructions. *NeuroImage*, 230, 2021.

[187] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In *Conference on Human Factors in Computing Systems*, 2020.

[188] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-hein, Sébastien Ourselin, Micah Sheller, and Ronald M Summers. The future of digital health with federated learning. *npj Digital Medicine*, 3(1), 2020.

[189] Santiago Silva, Andre Altmann, Boris Gutman, Marco Lorenzi, Santiago Silva, Andre Altmann, Boris Gutman, and Marco Lorenzi. Fed-BioMed : A general open-source frontendframework for federated learning in healthcare To cite this version : HAL Id : hal-02966789 framework for federated learning in healthcare. In *MICCAI*, 2020.

[190] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2006.

[191] Xiaohua Zhai, Lucas Beyer, and Google Brain. Revisiting Self-Supervised Visual Representation Learning. *arXiv*, 1901.09005, 2019.

[192] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 249–256, 2010.

[193] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv*, 1502.03167, 2015.