Automatic Analysis of Medical Images for Change Detection in Prostate Cancer

Nooshin Ghavami

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

August 10, 2020
I, Nooshin Ghavami, 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 is the most common cancer and second most common cause of cancer death in men in the UK. However, the patient risk from the cancer can vary considerably, and the widespread use of prostate-specific antigen (PSA) screening has led to over-diagnosis and over-treatment of low-grade tumours. It is therefore important to be able to differentiate high-grade prostate cancer from the slowly-growing, low-grade cancer. Many of these men with low-grade cancer are placed on active surveillance (AS), which involves constant monitoring and intervention for risk reclassification, relying increasingly on magnetic resonance imaging (MRI) to detect disease progression, in addition to TRUS-guided biopsies which are the routine clinical standard method to use. This results in a need for new tools to process these images. For this purpose, it is important to have a good TRUS-MR registration so corresponding anatomy can be located accurately between the two. Automatic segmentation of the prostate gland on both modalities reduces some of the challenges of the registration, such as patient motion, tissue deformation, and the time of the procedure.

This thesis focuses on the use of deep learning methods, specifically convolutional neural networks (CNNs), for prostate cancer management. Chapters 4 and 5 investigated the use of CNNs for both TRUS and MRI prostate gland segmentation, and reported high segmentation accuracies for both, Dice Score Coefficients (DSC) of 0.89 for TRUS segmentations and DSCs between 0.84-0.89 for MRI prostate gland segmentation using a range of networks. Chapter 5 also investigated the impact of these segmentation scores on more clinically relevant measures, such as MRI-TRUS registration errors and volume measures, showing that a statistically
significant difference in DSCs did not lead to a statistically significant difference in the clinical measures using these segmentations. The potential of these algorithms in commercial and clinical systems are summarised and the use of the MRI prostate gland segmentation in the application of radiological prostate cancer progression prediction for AS patients are investigated and discussed in Chapter 8, which shows statistically significant improvements in accuracy when using spatial priors in the form of prostate segmentations (0.63 ± 0.16 vs. 0.82 ± 0.18 when comparing whole prostate MRI vs. only prostate gland region, respectively).
Impact Statement

With the increased use of multiparametric-MRI (mp-MRI) for prostate cancer detection, it is important to develop new tools for processing these images and obtaining the relevant information. Differentiation between the high-grade aggressive tumours and low-grade prostate cancer is of huge importance, especially in deciding the treatment paths of these patients, to allow better management through less over- and under-treatment. TRUS-guided biopsies are still the routine clinical standard method to use for the diagnosis of prostate cancer, yet do not provide the accurate tumour location which can be obtained using mp-MRI.

To make use of the information present in both modalities, it is important to have a good TRUS-MR registration so that corresponding anatomy can be located accurately between the two, for example during a biopsy. To aid and reduce the challenges associated with the registration, automatic segmentation of the prostate gland on both modalities should be carried out. Computational algorithms have been quantitatively evaluated for image segmentation, registration, and change detection. The evidence presented in this thesis also supports the importance of segmentation tools, of either gland or lesion, for improved accuracy of progression prediction, and has been presented at conferences and published in several papers.

These findings could aid in the deployment of automatic segmentation methods within clinical workflows and commercial systems for a variety of tasks such as MR-TRUS registration or monitoring patients on an active surveillance cohort over time, as shown in Chapters 4 and 8 respectively. The work from Chapter 4 is already starting to be integrated into the clinical workflow of the SmartTarget system, which is a commercial system for precision targeting of prostate cancer and from which
the data were obtained, for automating the TRUS segmentation which is currently done manually. With further work on the progression prediction of AS patients, described in Chapter 8, there is potential for the use of prediction CNNs to aid radiologists in re-classification of men on an AS cohort.

In regards to academia, the work from this thesis will benefit disciplines such as deep learning and prostate imaging. The work and findings from Chapter 5 of the thesis, which has been published in Medical Image Analysis, serve as a starting point for a shift in the scope of future research and development of automatic segmentation algorithms using CNNs, by encouraging the research community to consider carefully the accuracy of specific downstream tasks of interest within a computational pipeline for the specific clinical application of interest. This is of interest to not only prostate gland segmentation, but other organs where segmentation may be used as an initial step in a clinical workflow.
Acknowledgements

The writing and completion of this thesis has involved a countless number of people.

Firstly, my supervisors Professor Dean Barratt, Dr Yipeng Hu and Professor Caroline Moore for all their expertise and help provided throughout this PhD. Their depth of knowledge in both the technical and clinical fields have been extremely useful along the way, and they have also provided me with constant support throughout the journey. I want to thank Dr Yipeng Hu especially for his endless hours of help and input and for being continuously available even outside of working hours for any small question I may have. I am grateful to the rest of the colleagues in my group, and Dr Ester Bonmati for her guidance and moral support in times of need.

I sincerely want to thank Francesco Giganti and Vasilis Stavrinides for their encouragement, support and helpful suggestions in putting the technical part of this work into perspective. They have helped me with my ideas and problems even when they had a million and one things going on themselves and it has been a great privilege working with them.

It has been a great privilege to work alongside all the other PhDs, Postdocs and academics in the department and UCL Centre for Medical Image Computing (CMIC). I have had the chance to establish many work and personal connections with people in this group, through post seminar drinks and away days. I consider myself very lucky to have formed some very strong friendships with certain individuals in the centre. Thank you to Danny Ramasawmy for all the help and practice with my Matlab skills, and to Marta Ranzini for her discussions on image registration, in addition to the cookies provided during pre-conferences deadlines. Many thanks to Maria Robu, one of the smartest and kindest souls, I hugely owe the com-
pletion of this PhD to our tea times in the office kitchen where she would make sure I don’t give up no matter how hard the situation.

I would like to acknowledge the UCL EPSRC Centre for Doctor Training in Medical Imaging for funding this PhD. Part of this work also presents independent research supported by the HIC Fund (grant no. HICF-T4-310), a parallel funding partnership between the Department of Health and the Wellcome Trust. The views expressed in this thesis are mine and not necessarily those of the Department of Health or the Wellcome Trust.

Many thanks to my close friends and relatives for their moral support. Last but definitely not least, I need to thank my incredible parents and my brother Navid who have given me encouragement and patience and made the tough periods of the PhD more bearable with their endless love. None of this would have been possible without you.
Contents

1 Background and Clinical Motivation ............................................. 24
  1.1 Prostate Anatomy .............................................................. 24
  1.2 Prostate Cancer Grading ..................................................... 25
  1.3 Prostate Imaging .............................................................. 26
    1.3.1 Ultrasound ............................................................... 26
    1.3.2 Magnetic Resonance Imaging ......................................... 29
    1.3.3 Guidelines and Reporting Systems .................................. 33
  1.4 Prostate Cancer Management- Active Surveillance ...................... 33
  1.5 Deep Learning for Medical Imaging ...................................... 35

2 Thesis Objectives ................................................................. 39
  2.1 Clinical Motivations ......................................................... 39
  2.2 Thesis Organisation ......................................................... 41

I Segmentation-based Prostate MRI-TRUS Feature-based Registration ...... 43

3 Background and Literature Review ............................................. 44
  3.1 Image Segmentation ......................................................... 44
    3.1.1 Overview ............................................................... 44
    3.1.2 MRI Segmentation ..................................................... 45
    3.1.3 US Segmentation ..................................................... 47
  3.2 Image Registration ........................................................... 48
4 TRUS Prostate Segmentation

4.1 Introduction ........................................... 55
4.2 Methods .............................................. 57
   4.2.1 Data and Pre-Processing .......................... 57
   4.2.2 CNN Architecture ................................ 58
   4.2.3 Training and Validation .......................... 60
4.3 Results ................................................. 61
   4.3.1 Comparison Between Different Observers ......... 64
   4.3.2 Comparison with Other Prostate Segmentation Techniques . 65
4.4 Conclusions ........................................... 65

5 MRI Prostate Segmentation and Subsequent MRI-TRUS Feature-based Registration ................................................. 67

5.1 Introduction ........................................... 67
5.2 Methods .............................................. 70
   5.2.1 Networks for Comparison .......................... 70
   5.2.2 Segmentation Metrics based on Hold-out Data ....... 71
   5.2.3 Gland Volume Errors and Estimated Target Registration Errors ............................................. 73
   5.2.4 Experiment Design for Network Comparison ......... 75
5.3 Experiments ........................................... 77
   5.3.1 Imaging Data and Ground-Truth Segmentations ....... 77
   5.3.2 Implementation and Network Training ............... 77
   5.3.3 Hyperparameter Configuration ....................... 78
5.4 Results ................................................. 79
   5.4.1 Hyperparameter Searching .......................... 79
   5.4.2 Segmentation Accuracy .............................. 79
   5.4.3 Volume Estimate Errors and Target Registration Errors .... 80
5.5 Discussion and Conclusion ................................................. 83

II Active Surveillance Management ........................................ 87

6 Background and Literature Review ...................................... 89
  6.1 Texture Features for Change Detection .............................. 90
  6.2 Deep Learning for Change Detection ................................ 93

7 Haralick-based Texture-analysis for Cancer Detection in AS Patients ........................................ 95
  7.1 Introduction .................................................................. 95
  7.2 Methods ................................................................... 97
    7.2.1 Data .................................................................. 97
    7.2.2 Key Pipeline Steps .................................................. 98
  7.3 Results ...................................................................... 99
    7.3.1 Comparison Between Dutasteride and Placebo: Manual Contours .............................................. 99
    7.3.2 Comparison Between Dutasteride and Placebo: Automatic Contours .......................................... 101
    7.3.3 Comparison Between Manual and Automatic Measures .......................................................... 101
    7.3.4 Correlation Between Change in Volume and Change in Texture Features .................................. 104
  7.4 Discussion and Conclusion ................................................. 107

8 Predicting Radiological Progression using CNNs ..................... 109
  8.1 Introduction .................................................................. 109
  8.2 Methods ................................................................... 111
    8.2.1 Data and Pre-Processing .......................................... 111
    8.2.2 Neural Network Algorithm ....................................... 112
    8.2.3 Experimental Design ............................................... 113
    8.2.4 Implementation and Cross-validation .......................... 117
  8.3 Results ...................................................................... 118
List of Figures

1.1 Zonal anatomy of the prostate gland 25
1.2 Gleason Pattern Scale [1] 26
1.3 Image depicting a TRUS-guided transperineal biopsy [2] 27
1.4 Example slice of the prostate from a TRUS-guided biopsy 28
1.5 Effect of echo time (TE) on the contrast of the image. Comparing A) TE of 50ms and B) TE of 125ms and C) TE of 800ms shows more signal from molecules that have longer T2 values [3] 31
1.6 Multiparametric MRI of a 67-year old man with a tumour on the right apex with (A) T2-weighted image, (B) ADC map image and (C) coloured perfusion map created from a DCE image [4] 32
1.7 Example of a deep learning model taking in the input data and then at each hidden layer extracting more abstract features from the image, until these features are used to recognise the object and output its identity [5] 36
1.8 Convolutional Neural Network [6] 37
2.1 Workflow of the thesis displaying the separation of the thesis into Part 1 and Part 2, while demonstrating the connection between the two 42
3.1 Gradient descent optimisation used in registration. At each point the gradient is calculated and a step is taken in that direction until it reaches the maximum similarity and gives perfect alignment [7] 50
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.2</td>
<td>MRI baseline (A) and repeat (B) scans of a woman with Alzheimer’s disease and the subtraction images after rigid registration (C) and fluid registration (D) [8].</td>
<td>52</td>
</tr>
<tr>
<td>3.3</td>
<td>Example of an MR prostate image with the prostate and lesion contoured (left) and the corresponding TRUS image with the registered prostate and lesion contour on top (right). The images are acquired from the SmartTarget study [9].</td>
<td>54</td>
</tr>
<tr>
<td>4.1</td>
<td>Example of the ultrasound probe setup with a phantom.</td>
<td>58</td>
</tr>
<tr>
<td>4.2</td>
<td>Proposed network architecture [10].</td>
<td>59</td>
</tr>
<tr>
<td>4.3</td>
<td>Diagram illustrating the experiment for taking different percentage of mid section prostate slices [10].</td>
<td>61</td>
</tr>
<tr>
<td>4.4</td>
<td>Example comparisons between manual (red) and automatic (blue) segmentations. A-D represent the 25th, 50th, 75th and 100th quantile with DSC of 0.84, 0.92, 0.95 and 0.98, respectively [10].</td>
<td>63</td>
</tr>
<tr>
<td>4.5</td>
<td>Differences in the automatically segmented prostate when incorporating different number of neighbouring slices. Manual segmentation (red), automatic segmentation using one adjacent slice (blue), using two adjacent slices (cyan) and three adjacent slices (yellow) overlayed on top of the original prostate slice [10].</td>
<td>64</td>
</tr>
<tr>
<td>4.6</td>
<td>Inter-observer segmentation comparisons for three example slices. Good visual agreement is shown between the manual segmentations from the two observers (red and green) [10].</td>
<td>65</td>
</tr>
<tr>
<td>5.1</td>
<td>Histogram of the results from the PROMISE-12 challenge [11].</td>
<td>68</td>
</tr>
<tr>
<td>5.2</td>
<td>Architecture of the six networks used for this comparison study. Different coloured arrows represent different architecture parts of the networks to visualise similarities and differences between them [11].</td>
<td>73</td>
</tr>
<tr>
<td>5.3</td>
<td>Segmentation-based registration pipeline [11].</td>
<td>75</td>
</tr>
<tr>
<td>5.4</td>
<td>k-fold cross-validation example [11].</td>
<td>76</td>
</tr>
</tbody>
</table>
5.5 Automatic segmentations generated from different CNNs for 6 patients. The two top columns are patients with DSC closest to the 25th percentiles, middle two columns are patients with DSC closest to 50th percentiles and bottom two columns are patients with DSCs closest to the 75th percentiles. Blue shows the segmentation from HighRes3dNet, green from HolisticNet, brown from VNet, magenta the segmentation from DenseVNet, yellow the segmentation from the adapted UNet, cyan from UNet and red the manual segmentation. The last row shows the overlay of all segmentations on top of the original image for each patient [11].

5.6 Box and Whisker plots of different measurement metrics for each of the segmentation networks [11].

5.7 Overlay images of the automatic segmentations from the UNet on top of the original prostate image for two outlier patients producing DSCs of smaller than 0.65 [11].

5.8 Plot showing relationship between DSCs and TREs for patients across all six networks, represented by the six different coloured points [11].

6.1 Calculation of grey-level co-occurrence matrix and subsequent texture features from a grey-scale image, taken from [12]. Step 1 involves the calculation of the corresponding grid of numerical grey-scale values from the original image. Step 2 involves the calculation of the GLCM by counting all possible co-occurrences between different pixel values. Step 3 involves the normalisation of the GLCM to probabilities, and finally Step 4 uses these probabilities for the calculation of different texture features.
7.1 Overall outline of the pipeline used for this chapter, divided into the manual and automatic pathways. In the automatic pathway (top), in step 1, the 6 months scan is registered to the baseline scan using an automatic registration software, and the transformation outputted from the registration is used to propagate the manually drawn lesion on the baseline scan to the space of the 6 months scan, as shown in step 2. In step 3, texture features and volume measurements are computed on both the manually contoured baseline and propagated 6 months lesion, and the differences between the lesion volume and texture features are calculated over time. The same procedure is carried out for the manual pathway (bottom), however, the registration and lesion propagation is excluded, since the lesion is manually contoured on both the baseline and the 6 months scan (step 1 and 2). 100

7.2 Box and Whisker plots of different texture measures calculated from the manual and automatic pipeline. 105

7.3 Comparison of manually contoured and propagated contour regions for six randomly selected patients. The manually drawn contours are shown in blue, while the automatically propagated contours are shown in red, with the centroids of the latter shown as a green dot. 105

7.4 Plot showing relationship between volume and each of the different texture measures, separated in two colours for the dutasteride and placebo group. 106

8.1 PRECISE case report form for reporting of MRI during baseline and follow-up [13]. 110

8.2 The PRECISE checklist [13]. 111
8.3 Baseline and Follow-up images for 6 randomly selected patients, displaying changes in the T2 and ADC images over time, with the red arrow indicating the tumour region. A): Patient showing no progression over time, B): Progression recorded as volume change, C): Progression recorded as conspicuity change, D): Progression recorded as both volume and conspicuity change, E): Progression due to artefact involvement, F): Progression due to appearance of a new lesion. .......................................................... 113

8.4 Network architecture for AS progression prediction. The network architecture is kept the same across all three experiments, however, the input to the network is varied with each experiment, as shown on the left part of the figure. Experiment 1 takes the whole image of each sequence and time point (concatenated) and inputs this into the network. Experiment 2 takes the prostate region of each sequence and time point (concatenated) and inputs this into the network. Finally, experiment 3 takes the bounding box around the lesion region of each sequence and time point (concatenated) and inputs this into the network. .......................................................... 114

8.5 Division of the prostate into different zones/bounding boxes for locating the lesion region. .......................................................... 116

8.6 Overview of our transfer learning architecture. The DenseNet is trained on the PROSTATEx dataset, and the weights of the trained network transferred and fine-tuned to our dataset. .......................... 117
List of Tables

1.1 Gleason Score and Groups [14]. ....................................... 26
1.2 Parameter selection for different image weightings and contrast [15]. 30

4.1 Segmentation metrics obtained from the automatic segmentation results when using different number of neighbouring slices. .......... 62
4.2 Segmentation metrics obtained from the automatic segmentation results when using slow and late fusion methods. .................. 63
4.3 Segmentation metrics obtained from the automatic segmentation results when taking different percentage of middle slices from each patient. ....................................................... 64

5.1 Information regarding the networks chosen for this comparison study. 72
5.2 Different hyperparameter configurations used for the hyperparameters tuning of the different networks. .............................. 79
5.3 The selected hyperparameter configurations for each of the segmentation networks. ............................................................... 80
5.4 Segmentation performance metrics, prostate volume calculations and target registration errors between the manual and automatic segmentation for each network. ................................. 82
5.5 P-values between the different networks for the DSCs, BDs, GVEs and TREs. ................................................................. 83

7.1 Volume measurements and texture feature values at baseline and 6 months for manually drawn lesions. ................................. 102
### List of Tables

7.2 Volume measurements and texture feature values at baseline and 6 months for propagated lesions. ........................................... 103

7.3 Pearson’s correlation coefficient between change in volume and change in each texture feature. ............................................ 106

8.1 Training and testing accuracy for simulated data, using 4 different experiments. ................................................................. 118

8.2 Training and testing accuracy obtained using different input data and learning rates for the AS dataset. Whole Prostate MR is when the whole original image, containing the prostate and surrounding structures is used as input to the network. Only Prostate Region involved inputting only the segmented prostate gland, with the original intensities to the network, while, Only Lesion Bounding describes the experiments where only a bounding box covering the lesion area is used as input to the network ................................ 119

8.3 Training and testing accuracy obtained when pre-trained on the PROSTATEX dataset with different learning rates. ................. 120
Acronyms

**AD**  Alzheimer’s Disease

**ADC**  Apparent Diffusion Coefficient

**ANOVA**  Analysis of Variance

**AS**  Active Surveillance

**AUC**  Area Under Curve

**BD**  Boundary Distance

**BN**  Batch Normalisation

**CNN**  Convolutional Neural Network

**Conv**  Convolution

**CT**  Computed Tomography

**DCE**  Dynamic Contrast Enhanced

**DOF**  Degrees of Freedom

**DSC**  Dice Similarity Coefficient

**DWI**  Diffusion Weighted Images

**FCNN**  Fully Convolutional Neural Network

**FLE**  Fiducial Localisation Error
<table>
<thead>
<tr>
<th>Acronyms</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRE</td>
<td>Fiducial Registration Error</td>
</tr>
<tr>
<td>GPU</td>
<td>Graphic Process Units</td>
</tr>
<tr>
<td>GVE</td>
<td>Gland Volume Error</td>
</tr>
<tr>
<td>LGCPD</td>
<td>Landmark-Guided Coherent Point Drift</td>
</tr>
<tr>
<td>MI</td>
<td>Mutual Information</td>
</tr>
<tr>
<td>MP-MRI</td>
<td>Multiparametric MRI</td>
</tr>
<tr>
<td>MRI</td>
<td>Magnetic Resonance Imaging</td>
</tr>
<tr>
<td>MRS</td>
<td>Magnetic Resonance Spectroscopy</td>
</tr>
<tr>
<td>NCC</td>
<td>Normalised Cross Correlation</td>
</tr>
<tr>
<td>PIRADS</td>
<td>Prostate Imaging Reporting and Data System</td>
</tr>
<tr>
<td>PPV</td>
<td>Positive Predictive Value</td>
</tr>
<tr>
<td>PSA</td>
<td>Prostate Specific Antigen</td>
</tr>
<tr>
<td>ReLU</td>
<td>Rectified Linear Unit</td>
</tr>
<tr>
<td>ResNet</td>
<td>Residual Network Unit</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root Mean Squared Error</td>
</tr>
<tr>
<td>SSD</td>
<td>Sum of Square Differences</td>
</tr>
<tr>
<td>STD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>TA</td>
<td>Texture Analysis</td>
</tr>
<tr>
<td>TRE</td>
<td>Target Registration Error</td>
</tr>
<tr>
<td>TRUS</td>
<td>Trans-Rectal Ultrasound</td>
</tr>
<tr>
<td>US</td>
<td>Ultrasound</td>
</tr>
</tbody>
</table>
Publications


Chapter 1

Background and Clinical Motivation

Since the late 1980s, when PSA screening was embraced, the incidence of prostate cancer has risen, increasing approximately threefold over a 5-year period [16]. Yet, around 22% of these cancers are diagnosed as low-grade cancer [17], meaning that PSA screening leads to over-diagnosis of these low-grade tumours [18], which in turn leads to over-treatment of the tumour. Over-treatment of the patient using prostatectomy or radiation, leads to an outweighing of the side-effects over the benefits for these group of men diagnosed with low-risk cancer [16]. It is therefore hugely important to be able to differentiate high-risk prostate cancer from the slowly-growing, low-risk cancer. For this reason, MRI has emerged as a promising method for localizing the tumour and determining its size and aggressiveness [18], in addition to determining any changes over time.

In the rest of this chapter, I briefly describe the prostate anatomy and grading system, followed by different prostate imaging methods which are currently used. The chapter is ended with a section on prostate cancer management, including Active Surveillance, and a brief introduction to the use of deep learning in medical imaging.

1.1 Prostate Anatomy

The prostate gland is located between the bladder and rectum and surrounds the proximal urethra as it exits the bladder [19]. It is usually described as a ‘walnut-shaped’ organ forming part of the male reproductive system. It is split into three
anatomical regions or ‘zones’ [20]: peripheral zone, central zone and transition zone, as visualised in Fig. 1.1. The peripheral zone constitutes over 70% of the prostate while the central and transition zone makes up the remaining 25% and 5% of the prostate gland, respectively. The anterior fibromuscular stroma, sometimes also known as the fourth zone, is located on the anterior surface of the gland.

1.2 Prostate Cancer Grading

The most common way of grading prostate cancer is using the Gleason Score and Grade groups, determined using samples of cells from a prostate biopsy [14]. The Gleason Score is used to determine the aggressiveness of prostate cancer, and therefore used to choose the appropriate treatment options. The Gleason Score ranges from 1-5 and describes whether the biopsy tissue is healthy tissue (lower score) or abnormal tissue (higher score). Figure 1.2 displays the Gleason pattern scale, where most cancers score a grade of 3 or higher.

Since prostate tumors are often made up of cancerous cells that have different grades, two grades are assigned for each patient. A primary grade is given to describe the cells that make up the largest area of the tumor and a secondary grade is given to describe the cells of the next largest area, where together they make up the total Gleason Score. This combined score is also called the Grade Group, with the different groups described in Table 1.1.
1.3 Prostate Imaging

1.3.1 Ultrasound

Ultrasound (US) is a high-frequency sound wave, used to form images of internal body organs by producing echoes from different tissue boundaries [21]. A B-mode US image is a cross-sectional image showing the organ boundaries and tissues within the body, where each echo will be shown as a point corresponding to the relative position of its origin and the brightness at each point is related to the amplitude of the echo, hence the name B (brightness)-mode [21].

To generate the image, firstly an US transducer is placed in contact with the surface of the structure of interest and short pulses of US are sent into the body.
1.3. Prostate Imaging

Figure 1.3: Image depicting a TRUS-guided transperineal biopsy [2].

along a narrow, beam-shaped path. While travelling through the body these pulses are both reflected and scattered at different boundaries, which lead to echoes. The echoes (waves) which are reflected are detected back at the transducer, and are then used to form the image. The reflection of the waves can be from anything such as boundaries of organs or vessels, or from small irregularities within the tissue [21].

Trans-rectal ultrasound (TRUS) involves an ultrasound probe being inserted into the rectum. A TRUS is used very commonly for guidance and needle placement when surgeons need to perform a transrectal or transperineal biopsy of the prostate (taking small tissue samples). In transrectal biopsies, the biopsy needle passes through the rectal wall, whereas for a transperineal biopsy, the needle is passed through the perineal skin to the prostate [22]. Fig. 1.3 shows more clearly the outline of this procedure for a transperineal biopsy. An example of an US image of the prostate acquired using a TRUS is shown in Fig. 1.4.

A TRUS-guided biopsy has some limitations for prostate cancer detection and classification. Firstly, around 40-60% of prostate cancers are not visible on a con-
1.3. Prostate Imaging

Figure 1.4: Example slice of the prostate from a TRUS-guided biopsy.

Conventional transrectal ultrasound image [4], especially the cancers which are located within the transition and anterior zones. The PROMIS study [23] also showed that for clinically significant cancer, mp-MRI was more sensitive than TRUS-biopsy (93% compared to 48%, respectively). Secondly, even when the cancer is detected by the TRUS-guided biopsy, there can be under grading of clinically significant cancer (due to under sampling), and over diagnosis (due to sampling error). The risk of under diagnosis is estimated to be around 20-30% [4]. Importantly, many of those cancers that are visible on TRUS cannot be reliably distinguished from other features in the image that have a very similar appearance to tumours. Therefore, the specificity of tumour detection is very low. These limitations can be overcome with the use of MRI.
1.3.2 Magnetic Resonance Imaging

MRI is a measurement technique for examining atoms and molecules, based on the interaction between the applied magnetic field and the atoms [24]. Atoms consist of three fundamental particles; protons, neutrons and electrons. The protons and neutrons are located in the ‘core’ of the atom known as the nucleus, and every proton in this nucleus has a quantity known as ‘spin’ which interacts with an applied magnetic field. The spin varies with different protons meaning each proton will be affected differently when placed in the magnetic field [24]. The rate at which the protons spin is associated to the magnetic field strength through the Larmor equation:

\[ w_0 = \gamma \times (B_0) \]

Where \( B_0 \) is the magnetic field strength, \( \gamma \) is known as the gyromagnetic ratio and \( w_0 \) is the Larmor frequency.

After the protons have been excited in the magnetic field, they need to lose their energy once the field is removed. This process is known as relaxation, and the time taken for the protons to go back to their initial state (‘relaxation time’) will depend on the proton type and all the surrounding protons. Therefore, by measuring these relaxation times, the structural information can be obtained [24]. There are two different ways in which the protons relax back to their initial state; T1 and T2 relaxation, given by the two equations below, respectively:

\[ M_z(t) = M_0 (1 - e^{-\frac{t}{T_1}}) \]

\[ M_{xy}(t) = M_0 e^{-\frac{t}{T_2}} \]

where \( M_0 \) represents the initial maximum value of the magnetization, \( t \) represents time, \( M_z \) is the longitudinal component of the magnetization, and \( M_{xy} \) is the transverse component of the magnetization. \( T_1 \) and \( T_2 \) are time constants. The T1- and T2-weighted images mentioned above form the anatomical sequences of
1.3. Prostate Imaging

Table 1.2: Parameter selection for different image weightings and contrast [15].

<table>
<thead>
<tr>
<th>Desired Image Contrast</th>
<th>Parameter Selection</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1-Weighting</td>
<td>Short TR (5-10ms)</td>
<td>T1 contrast would theoretically be better with $\alpha = 60-90^\circ$, but signal would be weak because far from Ernst angle, $\alpha = 30-50^\circ$ is compromise for good signal and T1-weighting at short TR values.</td>
</tr>
<tr>
<td></td>
<td>Short TE (2-5ms)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intermediate $\alpha(30 - 50^\circ)$</td>
<td></td>
</tr>
<tr>
<td>[H]-Weighting</td>
<td>Long TR (100-400ms)</td>
<td>Long TR and small $\alpha$ minimise T1 weighting; short TE minimises T2* effect.</td>
</tr>
<tr>
<td></td>
<td>Short TE (2-5ms)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>small $\alpha(5 - 20^\circ)$</td>
<td></td>
</tr>
<tr>
<td>T2*-Weighting</td>
<td>Long TR (200-800ms)</td>
<td>Long TR and small $\alpha$ minimize T1 weighting; long TE maximises T2* effects.</td>
</tr>
<tr>
<td></td>
<td>Long TE (20-50ms)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>small $\alpha(5 - 20^\circ)$</td>
<td></td>
</tr>
</tbody>
</table>

MR imaging. In addition to these, functional MRI techniques also exist including Dynamic Contrast Enhanced (DCE) and Diffusion-Weighted images (DWI). The different contrasts are obtained by modifying MR scanner parameters, such as the echo time, TE, and the repetition time, TR, accordingly. TE is the time between applying the 90° excitation pulse and measuring the maximum echo (signal), while TR is the time between the successive excitation pulses [24]. If the excitation pulse is not applied at 90°, then varying this angle also controls the contrast. The signal equation for a spin-echo sequence is given below [15], where K is a constant and [H] the spin density. This equation along with Table 1.2 show how and why these parameters affect the contrast. The effects of varying TE on T2-weighted images of the prostate are shown in Fig. 1.5.

$$\text{Signal} = K \cdot [H] \cdot (1 - e^{-\frac{TR}{T_1}}) \cdot e^{-\frac{TE}{T_2}}$$

DWI relies on the concept of diffusion which occurs when there is a concentration gradient and the molecules move from an area of high concentration to low concentration. In DW-MRI, two gradient pulses are applied, leading to a loss in signal from the spins moving during the gradient pulses [24]. The equation for the signal is given by:

$$\text{Signal} \propto e^{-\frac{TE}{T_2}} \cdot e^{-bD}$$
1.3. Prostate Imaging

Figure 1.5: Effect of echo time (TE) on the contrast of the image. Comparing A) TE of 50ms and B) TE of 125ms and C) TE of 800ms shows more signal from molecules that have longer T2 values [3].

with D being the diffusion coefficient and the b value determined by the gradient pulse duration and amplitude. Therefore, tissue or molecules with less motion have smaller D values which leads to smaller signal loss and vice versa [24]. This is the concept behind diffusion-weighted imaging and allows regions with more motion to be distinguished from the regions with very little motion. Diffusion-weighted MR imaging can be useful for detecting the presence and/or location of prostate tumours [16]. This is due to the fact that normal prostate tissue, especially in the peripheral zone, contain glandular structures where water molecules can freely move without much restriction. Cancerous tissue however, contains more tightly packed cells which restrict the diffusion of the water molecules. Therefore, an area with high signal intensity on high b-value DWI often represents an area with restricted diffusion caused by tightly packed cells. DW-MRI is also used to calculate apparent diffusion coefficient (ADC) maps to measure the degree of diffusion in the tissue. Cancerous tissue has higher cellular density and restricted diffusion compared to the surrounding normal tissue, therefore, they appear bright on high b-value DWI and dark on the ADC map with lower ADC tumour values.

The combination of structural and functional MR sequences such as T2, T1 and DWI are being used very often for the prostate, referred to as mp-MRI [4]. For example, in tumour localisation, on T2-weighted images the tumours would look darker due to shorter relaxation times, while on diffusion-weighted images the
1.3. Prostate Imaging

Figure 1.6: Multiparametric MRI of a 67-year old man with a tumour on the right apex with (A) T2-weighted image, (B) ADC map image and (C) coloured perfusion map created from a DCE image [4].

tumours would appear brighter because they have low diffusion values leading to high signal on the diffusion weighted images [3]. An example of a mp-MRI of the prostate is shown in Fig. 1.6, with the tumour located in each sequence.

For prostate cancer specifically, another advantage of mp-MRI includes the fact that it is a non-invasive alternative to TRUS-guided biopsies, which are commonly used, providing a much more powerful tool for the detection and risk stratification of prostate cancer patients. For example, a recent study by Turkbey et al. [25] compared the use of T2-weighted, magnetic resonance spectroscopy (MRS) and DCE imaging for prostate cancer detection. They found a significant improvement in the predictive value when using a combination of the three sequences compared to using each sequence separately. In another study published by the same authors [26], where T2-weighted, DW, DCE and MRS imaging was used, mp-MRI yielded a positive predictive value (PPV) of 98%. This is because this method can localize the prostate tumours and determine factors such as their size and aggressiveness which thereby helps in predicting how they will behave in the future [18]. For example, there has been found to be a negative correlation between the ADC value of a tumour and the Gleason score, so mp-MRI which uses ADC can be used to predict how aggressive the tumour may be [18]. Therefore, adding mp-MRI to the biopsy procedure can lead to an improvement in the detection of prostate cancer by using the lesion(s) which are identified on the MR image to guide the biopsies...
1.4. Prostate Cancer Management- Active Surveillance

[18, 27]. The use of mp-MRI also overcomes the limitations that exist with the random biopsies, most importantly being able to increase the accurate detection of high risk disease while simultaneously decreasing the detection of insignificant, low risk disease [4].

1.3.3 Guidelines and Reporting Systems

Different international standards exist for using MRI in prostate cancer imaging, such as the European Association of Urology (EAU) guidelines [28], the UK National Institute for Health and Care Excellence (NICE) guidelines [29] and the Prostate Imaging- Reporting and Data System (PI-RADS) [27, 30]. In relation to mp-MRIs and biopsies, the NICE guidelines state that mp-MRI should be performed at 12 months, but also anytime in between if there are any additional concerns, to determine if repeat biopsies are needed. Since the interpretation of the scans by the radiologist can vary, it is important to standardise the interpretability of the scans. The Prostate Imaging Reporting and Data System (PIRADS), aims to standardise prostate mp-MRI readings [30]. The scoring system is based on the five-point Likert scale and includes 1) a graphic prostate sector map 2) a separate PIRADS score for each individual lesion and 3) a maximum diameter measure of the largest lesion. All MRI sequences are scored independently (1-5) on a five-point scale for each suspicious lesion within the prostate. In addition, each lesion is given a final overall score (ranging 1-5) according to the probability of clinically significant prostate cancer being present [30].

1.4 Prostate Cancer Management- Active Surveillance

As mentioned in Section 1.1, a large number of men with prostate cancer are diagnosed as low-risk, therefore, instead of performing surgery which has a risk of life-changing side effects e.g. incontinence and impotence, these patients can instead be placed on what is known as active surveillance.

Active surveillance (AS) is the constant monitoring for risk reclassification,
1.4. Prostate Cancer Management- Active Surveillance

with the aim of deferred curative treatment if needed [16]. Traditionally, when men are placed on AS, they are followed with serial PSA assessments, repeat biopsies and sometimes digital rectal examinations (DREs). To be selected for AS, patients are considered with PSA < 10ng/ml, biopsy Gleason score ≤ 6 and three or fewer positive biopsy cores. The tumour may also be considered as clinically insignificant at a volume < 0.5ml with no seminal vesicle involvement [4]. However, even with these mentioned criteria, there is still no standard inclusion criteria for placing patients on AS, with variations in PSA values and Gleason scores between different centres commonly reported. Even more complicated is the lack of criteria for when to switch from monitoring to treatment. While on active surveillance, a third of patients are usually re-classified to higher-risk disease and so will progress to treatment [16]. Most of these re-classifications are due to a repeat biopsy, and around 20% of patients are shown to have a higher-grade cancer on the repeated biopsy compared to the initial [16]. This is most likely the main limitation for patients placed on AS in that the standard use of the TRUS biopsy leads to under-diagnosis, specifically in the anterior zone and apex area. Using mp-MRI in active surveillance, as mentioned briefly in Section 1.2, has a very high negative predictive value (97% on a re-biopsy for clinically significant cancer [16]), which means that this approach can firstly reduce the number of men who require a biopsy and secondly identify those with more clinically significant cancer earlier.

Probably the most important role of mp-MRI for men placed on AS would be distinguishing between the low-grade and high-grade disease, since this is the main reason for monitoring these men in the first place. ADC is the most common mp-MRI sequence used for this differentiation and many studies have suggested using ADC as a marker for selecting patients for AS. A study carried out by Henderson et al. [31] showed that patients who had a lower ADC value calculated from the DWI, progressed earlier than those patients with a higher ADC value at baseline. Patients who had a baseline tumour ADC value below the median (972 mm²/s) had shorter time to treatment by about 7 years, than those with ADC values above the median. This shows the importance of ADC for early risk stratification. In addition to the
grade of the tumour, mp-MRI can also give key information regarding the tumour size and invasiveness which are important factors for deciding whether the patients should be placed on AS or not to begin with. For example, even simple calculations of the tumour volume from the MR images can let the clinician know whether the cancer is low-risk or not, something which is much more difficult to calculate from a TRUS-guided biopsy [4]. Active surveillance will be discussed in more detail in Part 2 of the thesis.

1.5 Deep Learning for Medical Imaging

Deep learning is a powerful and flexible approach within the field of artificial intelligence (AI), and a particular type of machine learning where the problem is represented as a nested hierarchy of concepts. Each of these concepts is defined relative to other simpler concepts [5]. The essential example of a deep learning model is a feedforward deep network, which is essentially a mathematical function mapping a set of input values into output values. Fig. 1.7 [5] shows a simple deep learning classification model.

The perceptron, a basic neural network building block, is one of the earliest neural networks proposed [32] and mathematically formalizes how a biological neuron works. It has been realized that the brain processes information through billions of these interconnected neurons. Perceptrons, which consist of an input layer directly connected to an output node, emulate the biochemical process of neurons in the brain, through an activation function (also referred to as a transfer function) and a few weights. Specifically, it can learn to classify linearly separable patterns by adjusting these weights accordingly. To solve more complex problems, networks with one or more hidden layers of Perceptrons have been introduced [32]. which allows a deep architecture to be built that can express more complex hypotheses.

There are many different deep learning architectures, some of which have increased in popularity over the years. The architectures which are commonly used are CNNs, deep belief networks, deep Autoencoders, recurrent neural networks and deep neural networks [32]. CNNs, in particular, have had a large impact in
1.5. Deep Learning for Medical Imaging

Figure 1.7: Example of a deep learning model taking in the input data and then at each hidden layer extracting more abstract features from the image, until these features are used to recognise the object and output its identity [5].

The medical imaging field, due to the advantages of convolution for image analysis, most importantly being translation invariant, and have increased in popularity over the past 5 years [32], especially because of their ability to be parallelised with graphical processing units (GPUs).

The name of these networks (CNNs) comes from the convolution operator that is an easy way to perform complex operations using convolution filters, and unlike other deep neural networks, they can scale well for inputs with locally correlated data, which is commonly present in images. A CNN consists of a convolution step, following by a pooling/sub-sampling step, from which the output is used as the input of the next convolution step and so on. The convolution step is carried out by multiplication of the input image by a kernal (also known as filter), to produce the resulting convolved image which is know as the ’feature map’. The sub-sampling/pooling
step involves reducing the dimensionality of each feature map, by retaining the most important information. Typical pooling types applied are max or average pooling which compute the maximum or average value of a defined spatial neighbourhood. CNNs usually adopt a fully-connected layer after the final sub-sampling step, to convert the final feature maps into a 1D vector, necessary for classification tasks [32]. A very simple CNN, as proposed in the original paper by Lecun and Bengio [6], is displayed in Fig. 1.8.

For medical imaging in general, CNNs have been applied for segmentation and classification purposes in a variety of fields [33]. For detection and classification problems, a range of work exists for different organs and modalities. From CT scans, CNNs have been used for detection of pulmonary nodules in the chest, detection of polyps from colon and lung disease detection [34]. CNNs have also been used commonly on MRIs for applications such as brain tumour grading, microbleed detection, coronary calcium detection, and Alzheimer’s disease classification, as shown in the review paper by Greenspan et al. [34] and Litjens et al. [35]. Hematoxylin and eosin (HE) staining, has also been a modality where CNNs have been used extensively for applications such as glioma grading, metastases detection in lymph nodes, and classification of colon cancer [35]. Mammography and US are two other modalities where CNNs have been used, especially for breast cancer detection [35]. CNN algorithms for classification and change detection of prostate cancer are described in more detail in Chapters 6 and 8.

CNNs are also used for medical image segmentation, either organ or lesion segmentations. Many well known architectures such as the UNet [36] and VNet
1.5. Deep Learning for Medical Imaging

[37] have already been applied to a range of images. Litjens et al. [35] review many other CNNs for segmentation of a range of medical images, including cell segmentation on HE images, ventricle segmentation on US images, liver tumour segmentation on CT images and brain lesion segmentations on MRIs. A much more thorough review of existing segmentation applications of CNNs can be found in [35]. CNNs for prostate gland segmentation on MRI and TRUS are explored in more detail in Chapters 4-5, with the existing literature described in Chapter 3.
Chapter 2

Thesis Objectives

The aim of this research has been to develop novel computational methods for detecting changes in mp-MRI scans which are predictive of clinically-significant tumour progression in prostate cancer patients on AS. This includes: 1) use of both deep learning methods and texture feature extraction methods for longitudinal change detection and; 2) development of automatic segmentation methods for subsequent MR-TRUS and/or MR-MR registration with the aim of aiding the change detection problem.

The main contributions of this thesis are: 1) a novel CNN architecture proposed for accurate segmentation of TRUS images, validated on a large dataset, 2) a novel and thorough comparison of segmentation networks for prostate MRIs and their impact on volume measurements and MRI-TRUS registration, 3) development and validation of a computational pipeline for comparing ADC-based volume and texture measures between serial prostate mp-MRI scans, combining a registration, label propagation and texture analysis workflow and 4) an automated classification algorithm for predicting longitudinal radiological progression of men on AS.

2.1 Clinical Motivations

A significant proportion of men with low-risk prostate cancer who participate in an AS programme will progress to having clinically significant disease (at least a third of patients during the whole AS period according to Klotz et al. [16] and 15.9% of patients over 39 months as shown in a recent study by Thurtle et al. [38]).
Identifying these men as soon as possible in the AS pathway is important since it allows time for appropriate alternative management options to be considered for the patients, which in turn increases the chance of effective treatment. Serial mp-MRI has been proposed as a useful non-invasive method for monitoring prostate cancer patients and is already incorporated into AS programmes at some centres, such as University College London Hospital (UCLH). Currently, longitudinal images are analysed and compared by a radiologist to determine progression. However, discriminating changes in the tumour appearance associated with clinically-significant tumour progression, such as changes that are indicative of a change in Gleason grade, from changes that occur as a result of the variability in image characteristics between different hospital visits requires considerable skill and experience. In addition, there still exists gaps in knowledge about precisely what changes to look for between longitudinal images. Registration of the images over time allows any changes to be picked up easier, removing the limitations from visual registration, in that many subtle changes between images acquired at different time points cannot be picked out by the eye, as shown by Drew, et al. [39]. However, challenges also exist with registration algorithms, such as differences in the image generation process over time, or between different modalities, substantial tissue deformation and patient motion. Segmentation of the gland as a pre-step to registration allows their relative alignment to be defined and can therefore, reduce the challenges of longitudinal image registration. Using gland and/or lesion segmentation in automated methods, such as those making use of features involved in deep learning algorithms, may allow more detailed representation of images which may not be picked up by a radiologist or clinician. Automation of this radiological reporting process using computational algorithms would, in the first instance, remove the burden of requiring an adequately qualified and experienced radiologist, and potentially would result in reduced inter-observer variability and quicker assessment. Moreover, advanced image analysis techniques may allow the probability of a clinically-significant tumour development to be estimated, thereby providing a tool to predict if an individual patient is likely to progress to a stage where their disease requires treatment...
within a specified time-frame without the need for a biopsy. Currently, no automated image analysis tools exist to support clinical decisions in this way. Although diagnostic software tools to identify clinically-significant prostate cancer from an mp-MRI scan have been developed, these are still the subject of research and have not undergone extensive clinical evaluation on a patient population representative of men on AS. Moreover, such tools have been developed for the diagnosis of men with prostate cancer within a general population, and not specifically for use on AS patients, who have low to intermediate risk disease. Therefore, they may not be well-suited to detecting longitudinal changes in mp-MRI scans on these patients.

2.2 Thesis Organisation

I now proceed with a chapter-by-chapter description of the thesis, pointing out the major contributions. In Part 1, the focus is on the problem of TRUS and MRI prostate segmentation and subsequent TRUS-MRI registration, starting with a background and literature review in Chapter 3 followed by my work on automatic TRUS segmentation in Chapter 4 and concluding with a comparison study for the comparison of automatic MRI segmentation methods on the accuracy of volume measurements and registration accuracy. Part 2 of the thesis focuses on longitudinal change detection of prostate MR images from men on AS, starting with a literature review on the use of both texture analysis and deep learning for detecting changes in Chapter 6. The following two chapters focus on the use of texture analysis (Chapter 7) and deep learning methods (Chapter 8) for longitudinal classification and longitudinal progression prediction of prostate cancer, respectively. The last chapter concludes the thesis and contains a discussion of topics for future research.

A diagram showing the different aspects of the thesis and how they are linked together is displayed below.
Figure 2.1: Workflow of the thesis displaying the separation of the thesis into Part 1 and Part 2, while demonstrating the connection between the two.
Part I

Segmentation-based Prostate MRI-TRUS Feature-based Registration
Chapter 3

Background and Literature Review

3.1 Image Segmentation

3.1.1 Overview

Segmentation is the process by which an image is divided into different regions with similar properties such as grey level, brightness or contrast [40]. In the medical imaging field specifically, the main reasons for carrying out segmentation are: to identify an anatomical structure or region of interest such as a tumour or organ, measure tissue volume for tumour growth measurement and finally in radiation dose calculation to help with treatment planning [41]. Medical image segmentation techniques can be categorised into three groups; grey-level based, texture-feature based, and model based techniques. Each method, and the existing literature on that method, will be described briefly in the sections below.

1. Grey-level feature based techniques: These methods segment an image based on the grey-scale levels of the image(s), and are the simplest segmentation methods available. Some common methods based on this technique include: Amplitude segmentation (histogram-based) which involves segmenting an object by thresholding at a certain value in its histogram distribution. It is most suitable for an image which has an object of a uniform grey-level against a background of another uniform grey-level [41]. Edge-based segmentation which is based on the detection of edges in the image, commonly representing boundaries between different regions [41]. Different gradient-based edge detection functions exist such as Sobel, Canny,
Prewitt and Laplacian. **Region-based segmentation** techniques, which are based on clustering together pixels with similar properties in order to form a homogeneous region [41]. Two different types of this segmentation method exist; region merging—where seeds are placed in the image and the regions are grown by merging together the neighbouring pixels, and region splitting—works exactly the opposite way to the region merging, where the image is continuously split until there is no more region splitting possible [41].

**2. Texture feature based techniques**: These methods segment an image by subdividing it into regions which have different texture properties. The texture features themselves can be extracted in one of three different ways [41]; Statistical methods where the texture features are represented as a vector in the feature space, syntactic/structural methods where the features are spatially organised according to placement rules to generate complete patterns, and finally, spectral methods where the textures are defined by spatial frequencies.

**3. Model-based segmentation**: These methods aim to probabilistically model the organ structure. This probabilistic approach means that there is an a-priori knowledge about its shape and appearance. A model-based segmentation method which is commonly used is the Expectation-Maximisation algorithm [42]. The algorithm will not be described here in much detail; however, the main idea involves finding the maximum likelihood estimates of certain parameters in a probabilistic setting. The method assumes that the model parameter estimation and segmentation can be interleaved by estimating both simultaneously. Model based methods of segmentation also involve active shape and appearance model, and level-set based models.

**3.1.2 MRI Segmentation**

Segmentation of MRIs, is a field which has been explored for many decades, and involves a wide range of applications. Much of the research has focused on segmentation of brain MRIs, where the main goal has been to segment gray matter, white matter and cerebrospinal fluid, in addition to finding areas corresponding to lesion tumours and cysts [41]. Many papers have investigated the use of model based seg-
3.1. Image Segmentation

Image segmentation methods for brain segmentation including Zhang et al. [43] where the authors focus on the Expectation-Maximisation algorithm, Atkins et al. [44] use an integrated approach consisting of active shape models, and Greenspan et al. [45] deploy the use of gaussian mixture models for brain MRI segmentations. Other than model-based segmentation, another commonly used approach for brain MRIs have been atlas-based segmentation [46, 47]. Although the brain is the most commonly used organ for image segmentation, segmentation has also been applied to MRIs of other organs such as breast [48–50] and lungs [51–53].

Segmenting the prostate from MR images has also been an evolving research area, especially in recent years with the increased use of MRI for diagnosis and treatment. Many reasons exist for carrying out the segmentations, such as localisation of prostate boundaries, volume estimations and, most importantly regarding our work, as an initial step for carrying out multi-modal registrations. Many different MR segmentation algorithms exist, as summarised in the PROMISE12 challenge [54]. Most of these also use model based segmentation. For example, Vincent et al. [55] have used active appearance model where they generate dense anatomical landmarks from manually segmented surfaces. These landmarks and their associated images are then used to form the appearance model. A more interactive segmentation method has been proposed by Malmberg et al. [56] where the user segments the prostate or region of interest by sweeping the mouse over the object. The painted areas give immediate feedback to an interaction plane, which in the end produces a 3D segmentation. The user can apply a filter for smoothing the segmentation if further improvement is needed. Yuan et al. [57] propose a convex optimisation approach with a ‘generic star shape’ prior which is a contour evolution segmentation method. The convex optimisation allows the contour to evolve to its optimal position for each time frame, while the star shape prior reduces ambiguity of the prostate segmentation by removing any inconsistent segments. This allows the method to have robustness for poor image quality or artefacts. In addition to the methods submitted to the PROMISE12 challenge, Ghose et al [58], provide a thorough review of MRI prostate segmentation methods, with the majority of papers
3.1. Image Segmentation

using atlas-based [59, 60], grey-level [61, 62], or model-based techniques [63–65].

However, in recent years, with greater computation power and the rise of deep learning and CNNs, the field of medical image segmentation, including prostate MRI segmentation, has progressed from the traditional methods described in section 3.1.1 to deep learning methods, and specifically CNNs. The recent increase in the use of CNNs for prostate segmentation is partly due to the challenges rising from the variations in the shape and size of the prostate, in addition to indistinct prostate boundaries. Recent work using these networks for prostate gland segmentation include works by Milletari et al. [37], Zhu et al. [66], Yu et al. [67], Clark et al. [68], and Tian et al. [69]. Interestingly, in evaluation of the PROMISE12 challenge by Litens et al. [54] published in 2014, the top 10 segmentation methods consisted of either model-based techniques or atlas-based segmentation techniques. However, the current leaderboard, available on the PROMISE12 challenge website, consists of 7 CNN methods in the top 10 best performing segmentation methods.

3.1.3 US Segmentation

Segmentation of US images has also been a wide research area for many different applications [70]. In cardiology, many authors have deployed model-based techniques, including Mishra et al. [71], Mignotte et al. [72], and Bosche et al. [73]. In breast US segmentation, Horsch et al. [74] use grey-level based techniques for segmentation of lesions while Madabhushi et al. [75] combine texture feature based techniques with model-based techniques for lesion segmentation. Similar to application in cardiology, the majority of breast segmentation methods also focus on model-based techniques [76–78].

More specifically for prostate US images, in a survey paper carried out by Ghose et al. [58] on the different segmentation methodologies in US, most of the work in this field carried out model-based segmentation techniques in the form of either a deformable active contour model [79, 80], deformable active shape models [81–84] or active appearance models [85–88]. Support vector machines are also commonly used for prostate US segmentation [89, 90].

Unlike prostate MRIs, not as much work has been carried out on using
CNNs for the application of prostate segmentation from US images, since this is a much newer field with papers only published within the last 3 to 4 years [91–95]. Nonetheless, the majority of these methods show promising results outperforming the state-of-the-art [92, 93, 95] and overcome the boundary incompleteness problem present in many prostate US images [94]. This leads to the first methodological part of the thesis, Chapter 4, introducing a deep learning method using CNNs for automatic segmentation of the prostate gland on TRUS images.

3.2 Image Registration

3.2.1 Overview

Image registration is a process involving the establishment of spatial correspondences between two or more images [96]. Images may be acquired at different times, with different imaging modalities, or from different patients. The simplest case is the pair-wise registration, where two images are used, referred to as the ‘source’ and ‘target’ image. The source image is transformed into the space of the target image using a spatial transformation, in order to form an accurate alignment between the two, after which the accuracy of this alignment is estimated using a similarity measure. Finally, an optimisation is applied to find the best parameters of the transformation function in order to obtain the best similarity measure. These three components (transformation, similarity measure and optimisation), make up the main part of any registration algorithm, and will be described in more detail for both feature-based and intensity-based registration in the sections below.

For the registration work in this thesis, both intensity- and feature-based (in the form of Coherent Point Drift) registration is carried out, and both are described in more detail in the rest of the chapter, in addition to a review of the relevant existing literature.

3.2.1.1 Intensity-based Registration

Intensity-based registration involves using directly the image intensity information to align images, without extraction of any features from the images. The three components of intensity-based registration include the transformation model, similarity
3.2. Image Registration

1. Transformation Model: Consists of rigid, affine and non-rigid transformations. A rigid transformation is composed of only translation and rotation, resulting in 6 degrees of freedom (DOF) for a 3D model [97], while an affine transformation has 12 DOF for a 3D model with the addition of scaling and shearing parameters to the rigid parameters [97]. For many organs and structures in the body however, many more degrees of freedom are needed, and so non-rigid transformations are applied to model deformable changes [98]. The broad range of non-rigid transformation can be categorised into parametric and non-parametric methods depending on whether the deformation field can be parameterised by a function or whether there is direct optimisation of the deformation field respectively [98]. Examples of parametric methods include thin-plate-spline and B-splines, while elastic-, optical flow- and fluid-based methods fall under non-parametric methods.

2. Similarity Measure: The similarity measure is used to judge how well the registration algorithm is working by measuring how much overlap exists between the images after registration. One of the most common of these similarity measures is the sum of square differences (SSD) given by:

$$SSD = \sum_{n} (A_n - B_n(u))^2$$

With $A_n$ representing the reference image and $B_n(u)$ the transformed image [97]. Another measure of similarity is the normalised cross correlation (NCC) which is given by the equation:

$$NCC = \frac{1}{N\sigma_A\sigma_B} \sum_{n} (A_n - \bar{A})(B_n(u) - \bar{B}(u))$$

Where $\bar{A}$ and $\bar{B}$ are the mean intensity of the two images, $N$ is the number of pixels and $\sigma_A$ and $\sigma_B$ are the standard deviations [99]. The final commonly used similarity measure is mutual information (MI) which is an entropy based measurement and is calculated using the equation:
3.2. Image Registration

Figure 3.1: Gradient descent optimisation used in registration. At each point the gradient is calculated and a step is taken in that direction until it reaches the maximum similarity and gives perfect alignment [7].

\[ MI = \sum_{ij} P_{AB}(i,j) \log \frac{P_{AB}(i,j)}{P_A(i)P_B(j)} \]

where \( P_{AB} \) represents the joint probabilities, with intensity \( i \) and \( j \) occurring together and \( P_A \) and \( P_B \) are the marginal probabilities in each image separately [99]. MI is commonly used when registering images from different modalities.

3. Optimisation: This part of the algorithm is usually iterative and finds the transformation that will maximise the chosen similarity measure. Commonly used optimisation methods include Newton’s method, conjugate gradient, and most commonly, gradient descent optimisation. In the gradient descent algorithm small steps are taken in space, where at each step it then computes the similarity and gradient, taking the next step in the direction of the gradient. This is then iterated until the similarity measure converges, reaching a maximum [7], as shown in Fig. 3.1.

Intensity-based registration is used in Chapter 7 of this thesis.

3.2.1.2 Feature-based Registration

Feature-based registration involves finding feature correspondences between images, such as points, lines and anatomical landmarks. Point-based registration is most commonly used, which involves identifying corresponding three-dimensional points in the images to be aligned, registering the points and finding the image transformation that aligns them in a least square sense. The components of feature-based registration are analogous to the intensity-based registration described above, where the only main difference lies between the similarity measures. While for
3.2. Image Registration

intensity-based techniques, the similarity measure is applied directly to the intensities, for feature-based methods, the similarity measures are applied to the extracted features (either points or surfaces). The sum of square differences is commonly used as a similarity measure for feature-based registrations, such as in the iterative closest point (ICP) algorithm [100] and its variants, where the main idea is to update the point correspondence and the transformation alternately until convergence. The coherent point drift (CPD) algorithm is another commonly used feature-based algorithm. In this algorithm it is assumed that points from one dataset can be generated by a gaussian mixture model (GMM), which is centred on the points in the transformed dataset. The expectation maximisation (EM) algorithm is then used to solve this maximum likelihood estimation problem, iteratively updating the point correspondences and transformations in the Expectation and Maximisation steps, respectively. The CPD algorithm is used in Chapter 5 of this thesis.

Due to the lack of gold standards and the ill-posedness of registration, validating the accuracy of image registration is difficult. Nevertheless, some validation methods exist [101]. A few of these methods include, visualising the difference images, measuring an increase in similarity measures and considering the inverse consistency, all of which can allow evaluation of the accuracy [101]. When using point-based rigid registration, different errors can be computed to quantify the accuracy of the registration. These are the Fiducial-Localisation Error (FLE), Fiducial-Registration Error (FRE) and Target-Registration Error (TRE). FLE is the error in the identified position of the point, FRE is the root-mean square displacement between corresponding points and TRE is the displacement between two points which have not been used to derive the registration [102].

3.2.2 Literature Review

Since the early years of registration research, its use has expanded from looking at brain images to many other structures in the body, including: heart, lung and prostate. At the same time, the process has expanded from simple rigid-body registration to affine and non-rigid registration [103]. Initially, registration was largely used for aligning images acquired from the same subject at the same time point, but
over the past 20 years it has advanced substantially into applications for alignment of serial images to monitor change, matching of pre- to post-operative images, and visualising differences across groups [103].

Registration on longitudinal images has been used widely in neuroimaging for many years, especially in investigating the onset and/or progression of Alzheimer’s Disease (AD) from brain MR images [8]. Early diagnosis is very important in this case for the development of therapies and understanding the natural history of the disease. The registrations carried out can be either rigid or non-rigid. After registration, difference images are typically used to visualise the alignment by subtraction of one image from the other [8]. This can be seen in Fig. 3.2 where both a rigid registration (c) and a fluid registration (d) has been carried out. Longitudinal registrations has also been used in other studies such as in fMRI analysis [104], in the field of psychosis [105] and in the field of retinal imaging [106].

The research on the registration of prostate images have focused on both mono
modality registration (MRI-MRI and US-US) and on registering prostate images between different modalities such as MR/CT and MR/US. Lian et al. [107] focused on registering MRI prostate images using an endorectal coil to CT images of the prostate using TPS registration. The mapping between the two was found to lead to a huge improvement compared to non-deformable methods such as a rigid registration. Another study by Yang et al. [108] uses longitudinal registration of the prostate on TRUS images, using the registered images for subsequent training of support vector machines (SVMs). The trained SVM is then used on newly acquired prostate images, leading to accurate and robust segmentations.

However, the process of registering MRI to TRUS images, the main registration application explored in this thesis, is not an easy process. The simplest fusion technique would involve rigid registration. However, this will not take into account the deformation in the shape of the prostate during the biopsy (due to the presence of a probe in the rectum), therefore, leading to a poor registration result. Nonrigid registration techniques take this deformation into account and compensate for it. As described in Section 3.2.1, there are a few different non-rigid registration methods, and two which have been commonly used are surface registrations using an elastic registration method [109] or statistical motion modelling of the prostate deformation [110], both of which produce fast and accurate registration methods. CPD algorithms, for both rigid and non-rigid MRI-TRUS registration, have also been used [111, 112], both papers reporting TREs in the range 2.40-2.42 mm. Typical MR-TRUS registration accuracies reported in other works also fall within the same order of magnitude [113–115]. Fig. 3.3 shows a MR slice of the prostate with the prostate and lesion manually segmented in green and orange respectively, and the outline of both on the TRUS image after MR-TRUS registration, allowing the lesion location to be visualised on the TRUS image, while before registration it is very hard to detect. MR-TRUS registration will be explored in more detail in Chapter 5.

Nevertheless, there has not been extensive work looking at longitudinal MR to MR registration of prostate images [116, 117]. For active surveillance patients, to the best of our knowledge, there is no work on longitudinal MR to MR registration.
Figure 3.3: Example of an MR prostate image with the prostate and lesion contoured (left) and the corresponding TRUS image with the registered prostate and lesion contour on top (right). The images are acquired from the SmartTarget study [9].
Chapter 4

TRUS Prostate Segmentation

In this chapter a novel CNN architecture is developed for automatic segmentation of the prostate gland from TRUS images. The segmentations are useful as an initial step of many TRUS-MRI registration methods. The work from this chapter was published in the SPIE medical imaging conference [91] and a journal extension published in the Journal of Medical Imaging [10].

4.1 Introduction

Recent attempts to address the challenges rising from PSA testing, such as over-diagnosis of patients with low-risk prostate cancer and subsequent over-treatment of these patients, have led to an emergence of TRUS-guided biopsies and focal therapy techniques that are performed in a highly-targeted way. These techniques typically use diagnostic MRI to identify target regions suspected or known to be harbouring clinically significant cancer [118], overcoming the difficulty associated with reliably distinguishing prostate tumours in conventional B-mode TRUS images. However, TRUS remains a safe, low-cost, portable method for guiding the insertion of needles and other instruments into the prostate in real-time; and in recent years a growing number of guidance systems have become available commercially, which spatially register and fuse MRI and TRUS data to aid targeted needle biopsy. Sankineni et al. [119] showed that in 26% of patients with prostate cancer, TRUS-MRI fusion-guided biopsy detected the cancer, whereas a conventional, systematic 12-core biopsy did not. Fully automatic registration of pre-procedural
MRI with intraoperative TRUS images is a challenging problem due to many factors, such as patient motion, soft-tissue deformation, and marked differences in the image intensity characteristics of the different modalities. Consequently, a feature-based approach is typically employed in commercial and research guidance systems in which the prostate is first segmented in the MRI and 3D TRUS images, and the resulting segmentations are aligned using either a rigid or non-rigid (i.e. elastic) registration algorithm. Accurate manual segmentation of the prostate to provide input data when using this approach can be difficult and time-consuming, especially given that the process may need to be repeated multiple times during a procedure to account for prostate motion. Moreover, these segmentations are subject to inter- and intra-observer variability which can introduce variability in the registration accuracy. Automating the image segmentation process provides a way to reduce this variability, thereby improving standardization, and reduce the need for extensive manual interaction during a procedure. Furthermore, although fully-automated registration methods are starting to emerge that do not require explicit segmentation of the input images [120], automated segmentation is still very useful for training, monitoring and evaluation purposes.

Most prior work on prostate image segmentation has focused on the segmentation of T2-weighted MRI images [54, 121]. CNNs have been shown to achieve high accuracy for the segmentation of these images [122–124]. Previous works on automatic segmentation of the prostate from TRUS images have adopted a range of supervised and unsupervised machine learning methods, including texture-feature-extraction methods with SVM [81, 83, 125–127] and neural networks [92–94], where the authors show that CNNs achieve superior performance even for TRUS images compared to other segmentation methods [35]. However, at the time of the first publication from this work, only one other CNN method for TRUS prostate segmentation existed [94], where the authors used data from only 17 patients and obtained high segmentation accuracies. Therefore, this gap of using CNN methods for prostate segmentation on TRUS images, validated on a large dataset, provide the motivation of this work. In this chapter, the accuracy of a CNN-based method
for automatic prostate segmentation was evaluated on clinically acquired TRUS images from 109 patients. In addition to a 2D slice by slice segmentation, neighbouring slices were incorporated into the network to be able to use spatial information for each slice, to take into account 3D information which the human observers often consider in the manual segmentation. Incorporation of spatial information has already shown promising results in fetal ultrasound segmentation [128].

4.2 Methods

4.2.1 Data and Pre-Processing

The TRUS images used in this work were acquired as part of the SmartTarget Biopsy Trial [9], and consist of 3D TRUS images of the prostate of 109 patients who underwent targeted transperineal biopsy. The trial was a comparison of visual estimated targeted biopsies compared to non-rigid MR/US image-fusion using an academically developed fusion system [9]. For each patient, a continuous rotational 3D acquisition was used to acquire between 38 and 177 para-sagittal slices to cover the prostate gland. During these acquisitions, the side-firing TRUS probe was rotated slowly whilst being held by a stepper cradle equipped with digital position encoders to measure the rotation. A set up of the probe attached to the cradle for a phantom is shown in Fig. 4.1. After sampling at 3° intervals, up to 59 slices per volume were used, leading to a total of 4034 2D slices. The image slices used in this study had a pixel size of 0.18 \times 0.16 \text{ mm} and an image size of 576 \times 720 pixels.

The ITK-SNAP software [129] was used to carry out the manual delineation of the prostate gland (excluding the seminal vesicles). Manual segmentations were performed independently by two observers, with between 2-4 years of experience in prostate US image analysis. A segmentation of each volume took approximately 20-30 minutes to complete. The segmentations from the first observer (NG-first author of this study) were used as the ground-truth both for training the algorithm and for validating the network segmentation in a cross-validation experiment described in Section 4.2.3, while the segmentations from the second observer were used only for inter-observer comparison purposes.
4.2. Methods

4.2.2 CNN Architecture

The algorithm used for the TRUS segmentation uses a CNN which is based on an adapted U-network architecture [130]. The original network takes as input an ultrasound slice of size $S_0 = [576 \times 720]$ pixels and this is propagated to feature maps of the same size and 16 initial channels ($n_0$) using a convolution (Conv), a batch normalisation (BN) and a nonlinear rectified linear unit (ReLU). A kernel size of $3 \times 3$ is used for the convolutions. For the present work, for each input image slice, a different combination of the neighbouring slices are also considered. The concatenation, between the slice to segment and additional neighbouring slices at either side, act as 3D spatial priors for the network. The resulting feature maps are down-sampled to 4 different resolution levels, where at each level, $k = 1 \ldots 4$, the image size $S_k$ is halved and the number of channels $n_k$ doubled, meaning follow-
ing the 4 downsampling layers the image size is 24 times smaller than the original size. The down-sampling consists of a troika of Conv, BN and ReLU, followed by a max-pooling layer with stride 2. This is then followed by a residual network unit (Resnet) block consisting of two Conv layers with BN and ReLU, including an identity shortcut over these layers. The network architecture is shown in Fig. 4.2. The up-sampling blocks reverse the down-sampling process using transpose convolution layers with stride 2, replacing the max-pooling layers, and output an image-sized logits layer to represent the segmentation. Reverse Resnet blocks are also included with the addition of additive up-sampling shortcut layers after the transpose convolution layers [131]. Summation shortcuts are added before each down-sampling block to the output feature maps from each up-sampling block, which is of a compatible size. Summation shortcuts were used in our network instead of concatenation since they have been shown to provide a more smoothly propagated gradient flow and therefore improving the training efficiency [132], and provided competitive results in segmenting prostate from MR images [67]. The key differences between our proposed method and the original U-Net, include: 1. Incorporation of ResNet shortcuts in both the down-sampling and up-sampling layers, 2. incorporation of summation shortcuts instead of concatenation, 3. addition of additive up-sampling blocks in the decoder.
4.2. Methods

4.2.3 Training and Validation

The network was implemented in Tensorflow\textsuperscript{TM} and trained on a 12 GB NVIDIA TITAN XP GPU for 10,000 iterations, using the Adam optimiser with 64 images in each minibatch. The results presented here were obtained by minimising a negative probabilistic Dice score that is differentiable with an added L2-norm weight-decay on the trainable parameters, the weighting parameter was set to $10e^{-6}$. A 10-fold patient level cross validation was carried in which images from 11 patients were held out for testing, while the remaining patients were used for training the networks. This was repeated until each of all 109 patients were used for evaluation once. For each automatic segmentation, the largest connected component was chosen to eliminate any isolated foreground segmentations, as a simple post-processing step. Segmentation metrics were calculated for each fold, by comparing the automatic segmentations to the manually segmented images (ground-truth) using both the binary Dice similarity coefficient (DSC) and the boundary distance. The boundary distance was defined as the mean absolute value of the distances between all the points from the automatically segmented boundary and the closest boundary points found on the left-out ground-truth segmentation. Additionally, Dice scores were also calculated for 3D volumes on the patient level. This was computed on volumes reconstructed with the slices from the individual patient. The 3D DSC are arguably more relevant to the registration application of interest, which requires 3D prostate TRUS volumes. Furthermore, to evaluate the impact of the number of adjacent slices, the network is tested with different combinations of neighbouring slices (on either side), leading to 3D inputs with 3, 5 and 7 feature maps. As described in the paper by Karpathy et al. [133], there are three different ways of combining the spatial information: early fusion, late fusion and slow fusion. What is described in this chapter is equivalent to the early fusion pattern. For comparison, the slow fusion was also implemented (using 2 adjacent slices on either side), where the neighbouring slices are slowly combined through the network, at each layer combining two sets of neighbouring slices together and in doing so giving access to more global spatial information as we go to deeper layers, and the late fusion which
4.3 Results

Figure 4.3: Diagram illustrating the experiment for taking different percentage of mid section prostate slices [10].

takes two slices (5 slices apart) as two separate networks and only merges these at the fully connected layer. As an additional experiment, the network was also trained by taking a different percentage of the mid-section of the prostate for each patient, as illustrated in Fig. 4.3. The aim of this experiment was to evaluate the effects of removing slices at the base and apex of the prostate, where the boundary of the prostate is generally considered more difficult to identify.

4.3 Results

The computed 2D and 3D DSC and boundary distances averaged over all slices are summarised in Table 4.1 for the three sets of experiments, using 1, 2 and 3 neighbouring slice(s) on each side using the early fusion. From the table, it is shown that taking neighbouring slices leads to an improvement in the 2D DSCs of approximately 0.01 on average. Paired Student’s t-tests ($p = 0.05$) were performed to test statistically significant differences between the network without using neighbouring slices and those using 1, 2 and 3 slices, with results of $p = 0.69$, $p = 0.04$ and $p = 0.49$, respectively. Increasing the number of neighbouring slices seemed to lead
### Table 4.1: Segmentation metrics obtained from the automatic segmentation results when using different number of neighbouring slices.

<table>
<thead>
<tr>
<th>Number of Neighbouring Slices Included on Each Side</th>
<th>2D DSC mean±std</th>
<th>3D DSC mean±std</th>
<th>Boundary Distance mean±std</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.88±0.13</td>
<td>0.88±0.06</td>
<td>1.80±1.68</td>
</tr>
<tr>
<td>1</td>
<td>0.89±0.12</td>
<td>0.89±0.05</td>
<td>1.79±2.05</td>
</tr>
<tr>
<td>2</td>
<td>0.89±0.13</td>
<td>0.88±0.04</td>
<td>1.77±1.46</td>
</tr>
<tr>
<td>3</td>
<td>0.89±0.12</td>
<td>0.88±0.05</td>
<td>1.75±1.77</td>
</tr>
</tbody>
</table>

to a decrease in the calculated boundary distance, with \( p = 0.39 \), \( p = 0.52 \) and \( p = 0.48 \) for 1, 2 and 3, respectively, when compared to using no neighbouring slices. Among these results, only one case showed significant difference with \( p = 0.04 \) when comparing the DSC from the two-neighbouring-slices case.

Comparison between the automatic and manually segmented images for 4 example slices (each representing one of the quartiles from the DSC), taking one neighbouring TRUS slice on either side, is shown in Fig. 4.4. The slices shown are chosen with DSC approximately equal to the four quartile values which are 25th (DSC = 0.84), 50th (DSC = 0.92), 75th (DSC = 0.95) and 100th (DSC = 0.98) quantiles. As illustrated by these examples in Fig. 4.4, the segmentations were generally more accurate when the prostate boundary was more clearly defined, as seen in slices C and D. This supports the issue of boundary incompleteness as described in [126], which is shown to influence the results from the automatic segmentations. This compares well with other work using neural networks for prostate segmentation, reporting a mean DSC value of 0.92 [83], evaluated on 17 subjects. The results presented in Table 4.1 were equivalent to using the early fusion pattern, and as described in Section 4.2.3, the experiment results using 2 and 3 adjacent slices on either side, for the slow and late fusion, respectively, are summarised in Table 4.2. There was no statistically significant difference between the early and slow fusion in terms of the DSC \( (p = 0.34) \), however there was in the boundary distance \( (p = 0.03) \) and for both the DSC and boundary distance when comparing the early and late fusion \( (p < 0.001 \text{ and } p = 0.03, \text{ respectively}) \).

The results of the automatic segmentation for 3 example slices when taking different number of neighbouring slices for each patient are shown in Fig. 4.5.
4.3. Results

Table 4.2: Segmentation metrics obtained from the automatic segmentation results when using slow and late fusion methods.

<table>
<thead>
<tr>
<th>Fusion Method</th>
<th>2D DSC  mean±std</th>
<th>3D DSC mean±std</th>
<th>Boundary Distance mean±std</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slow (2 adjacent slices on each side)</td>
<td>0.89±0.12</td>
<td>0.89±0.05</td>
<td>1.68±1.57</td>
</tr>
<tr>
<td>Late (3 adjacent slices on each side)</td>
<td>0.86±0.12</td>
<td>0.85±0.06</td>
<td>2.15±1.59</td>
</tr>
</tbody>
</table>

Figure 4.4: Example comparisons between manual (red) and automatic (blue) segmentations. A-D represent the 25th, 50th, 75th and 100th quantile with DSC of 0.84, 0.92, 0.95 and 0.98, respectively [10].

Visually similar contours are observed between each of the automatic segmentations (blue, cyan and yellow) and the manual ground-truth (red), which in consistent with the results from Table 4.1. The slice shown in C, however, shows a visual improvement in the yellow contour (using 3 neighbouring slices) when comparing to the ground-truth shown in red, as opposed to the blue and cyan segmentations.

Table 4.3 displays the segmentation metrics obtained when taking different percentage of slices from each patient (full set of slices, 90% slices, 75% slices and 60% slices) whilst incorporating 1 neighbouring slice. There is no statistically significant difference found in the boundary distances ($p = 0.14$, $p = 0.84$, $p = 0.67$) when using 90%, 75% and 60% of the middle slices respectively, nor the 2D DSC ($p = 0.17$, $p = 0.07$), when using 90% and 75% of the slices, however, a significant
4.3. Results

Figure 4.5: Differences in the automatically segmented prostate when incorporating different number of neighbouring slices. Manual segmentation (red), automatic segmentation using one adjacent slice (blue), using two adjacent slices (cyan) and three adjacent slices (yellow) overlayed on top of the original prostate slice [10].

Table 4.3: Segmentation metrics obtained from the automatic segmentation results when taking different percentage of middle slices from each patient.

<table>
<thead>
<tr>
<th>Percentage of Slices</th>
<th>2D DSC mean±std</th>
<th>3D DSC mean±std</th>
<th>Boundary Distance mean±std</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>0.89±0.12</td>
<td>0.89±0.05</td>
<td>1.79±2.05</td>
</tr>
<tr>
<td>90%</td>
<td>0.88±0.12</td>
<td>0.88±0.06</td>
<td>1.90±1.91</td>
</tr>
<tr>
<td>75%</td>
<td>0.89±0.10</td>
<td>0.89±0.05</td>
<td>1.78±1.56</td>
</tr>
<tr>
<td>60%</td>
<td>0.89±0.09</td>
<td>0.89±0.05</td>
<td>1.83±1.42</td>
</tr>
</tbody>
</table>

difference is observed using 60% of the slices with $p = 0.04$. This result, and the reduction in the standard deviation of both the DSC and boundary distance could suggest that slices near the apex and base of the prostate are indeed more challenging to segment.

The addition of the up-sampling shortcuts layer into the network improved the training time from 253 minutes without any up-sampling shortcuts to 161 minutes with these up-sampling shortcuts for 10,000 iterations per fold.

4.3.1 Comparison Between Different Observers

Fig. 4.6 shows examples of manual segmentations from the two different observers, overlaid on top of the original slices for 3 randomly chosen slices. From the figure, good agreement is shown between the two observers (shown in red and green contours), which also agrees with the computed 2D DSC between the two observers of 0.92±0.06. These results, together with those reported above, provide a quantitative reference to compare the inter- and intra-observer variability with variability using
4.3.2 Comparison with Other Prostate Segmentation Techniques

The results obtained using our architecture is compared with a state-of-the-art segmentation technique proposed by Anas et al. [92]. Their architecture uses gated recurrent units with the use of residual convolution for improving the optimization of the network. In addition, the authors use a recurrent interconnection between the feature extraction and upsampling branches which allows the network to incorporate lower level features in the output segmentation. The mean±std DSC reported for this paper on 1017 testing slices is 0.93±0.03 and 1.12±0.79 mm for the DSC and boundary distance, respectively. Based on our implementation, their network needs significantly more GPU memory (approximately >300 times more based on single-slice stochastic gradient descent) than the one proposed in this paper. Previous prostate segmentation techniques using shape models, such as the works by Zhan et al. [126] and Shen et al. [81], report average boundary distances of 0.39±0.05 mm and 1.28±0.03 mm, based on 6 and 8 validation patients’ data, respectively.

4.4 Conclusions

In this chapter I have used CNNs for segmenting the prostate gland from TRUS images [91] and extended this CNN to incorporate spatial information from neigh-
4.4. Conclusions

bouiring TRUS slices and an additive up-sampling shortcuts in the decoder part of the network. Both qualitative and quantitative results show good agreement between the automatic and manually segmented images when taking a range of neighbour- ing slices, but the inclusion of neighbouring TRUS slices with the input image to be segmented was found to make very little or no difference to the segmentation accuracy compared with not including this data for training.

A limitation of this work is that the data used was acquired at a single centre, which does not validate its generalisation to data from different centres. For future work, the network architecture may be improved specifically for slices near the apex and base of the prostate, which are currently the hardest to segment due to the boundary incompleteness, therefore, by taking this problem into account, it may improve segmentation results.
Chapter 5

MRI Prostate Segmentation and Subsequent MRI-TRUS Feature-based Registration

As stated in chapter 4, a common application of TRUS segmentations is their subsequent use in TRUS-MRI registration applications. In this chapter a comparison of six different CNNs was carried out for the task of MR prostate segmentation and their accuracy compared in terms of both segmentation accuracy and, perhaps more importantly, accuracy in terms of clinical measures such as volume measurements and registration accuracy. The work from this chapter was published in the Medical Image Analysis journal [11].

5.1 Introduction

Mp-MRI is emerging as a clinically useful tool for detecting and localising prostate cancer. Results from the recent PROMIS and PRECISION studies, for instance, suggest that mp-MRI may be a valuable triage tool for clinically-significant disease to reduce the number of men needing biopsies [23, 134]. In addition, mp-MRI is increasingly being used to target suspicious regions during biopsy and therapy, with or without the aid of a computer-assisted MRI-ultrasound fusion system [118].

Deep learning methods, especially supervised classification methods based on CNNs, have been successful in the field of medical imaging for segmenting an
anatomy of interest [35]. For example, these networks have produced higher accuracies for automatic prostate segmentations from T2-weighted MR images, compared with other alternative segmentation approaches [35]. Some examples of these networks include the V-Net [37], which was proposed to segment the prostate gland from T2-weighted MR images in 2016, and has since been used for several different applications [135–138]. More recently, other variations of CNNs have also been proposed for prostate segmentation, including [66–69]. At the time of writing, all the top five prostate segmentation algorithms submitted to the PROMISE-12 challenge [54, 139] adopted CNNs, with the highest performing methods generating average Dice scores and boundary distances of 0.90 and 1.71 mm on whole gland segmentation, respectively. Fig. 5.1 shows a histogram of the results from the PROMISE-12 table, with most of the submitted algorithms centred around a score of 82–89. With all these variations of CNNs for prostate MR image segmentation, a direct quantitative comparison of different CNN architectures on a single large data set, especially those with open-source implementations (not a requirement for submitting to the challenge) is important, but to date has not been available to our research community.
Partly limited by the test data size of 30 provided in the PROMISE-12 challenge, the diminishing statistical differences among top performing segmentation algorithms in this challenge [140] can complicate interpreting the difference in segmentation accuracy. Examples of networks which demonstrate these statistical differences between different architectures, include, residual networks [141] and densely connected networks [142]. Both have demonstrated significantly improved results in computer vision tasks, and have been incorporated for medical image segmentation, such as the work proposed by [91] and [135], respectively.

Perhaps more importantly, assessing the value of adopting different CNN architectures in clinical applications requires evaluating their performance within a pipeline of clinical tasks, which arguably have not been turned to research focus for prostate MR segmentation. Segmentation of the prostate from MR images is important for several potential clinical applications. One application includes calculating the gland volume estimation which can be used for measuring drug-induced prostate volume changes [143], for correlation with cancer volume [144] and for detecting significant cancer [145, 146]. Other applications also include for longitudinal analysis of patients undergoing active surveillance [147–151], and as part of segmentation-driven multi-modal registration to support MR-targeted transrectal-ultrasound (TRUS) guided biopsy and therapy [149, 152, 153]. Relating these clinical measures to prostate segmentation networks, for example, those submitted to PROMISE-12 challenge (Fig. 5.1), has not yet been investigated.

However, comparison of deep-learning-based segmentation algorithms also faces significant challenges such as the requirement of test data size, in addition to the dependency on the hyperparameter selection, including initial learning rate, model size (number of layers and feature channels in each layer) and regularisation methods such as weight decay. Cross-validation for hyperparameter searching is effective in resampling the limited data (a common restriction in medical image computing applications) but is likely to produce over-optimistic models due to information bleeding [154]. Arbitrary hyperparameter selection would lead to less clinically meaningful comparison between merely sub-optimally-designed networks,
while marginalising these hyperparameters spaces for architectural comparison is
computationally prohibitive and provides little practical value. Therefore, the data
was split into development and hold-out sets before optimising the hyperparameters
using cross-validation on the development set. The details of the experiment design
and its implementation for prostate segmentation on MR images are provided in
Section 5.2.

In this chapter, the aim was to compare the prostate segmentation accuracy
of six different CNN architectures, in terms of two segmentation metrics, gland
volume estimates and registration errors, the latter two of which are based on the
automatic segmentations, and the differences between these errors. The aims are: 1)
to demonstrate deviations in segmentation accuracy due to varying network archi-
tectures, and 2) to estimate clinically relevant impact that can potentially be caused
by these deviations. In turn, the contributions of this chapter are summarised as
follows: 1) a quantitative comparison of six open-source segmentation algorithms
is carried out, each one adapted to prostate MR segmentation, trained using an ex-
tensive hyperparameters tuning, and tested on an independent hold-out data set; 2)
a comprehensive set of segmentation accuracy results are reported and compared,
over these different networks; 3) clinically interesting results, in gland volume esti-
mation and MR-ultrasound image registration, are reported and compared. It sends
an important message in finding the disagreement between these and the segmenta-
tion accuracy.

5.2 Methods

5.2.1 Networks for Comparison

In this study, six network architectures were chosen: UNet, VNet, HighRes3dNet,
HolisticNet, DenseVNet, and an Adapted UNet. Our inclusion criteria included
relevance, availability and reproducibility, as the implementations of these six net-
works are readily accessible and they have been already applied on the same or
closely-relevant applications. For example (re-)implementations of the first four
are available on the NiftyNet Platform [155] and the Adapted UNet [91] has been
developed in our group, with a minimal adaptation to the original 3D UNet.

The 3D UNet [36] is one of the earliest proposed 3D fully convolutional neural networks originally proposed for segmenting kidney embryos on xenopus and reported an average intersection over union (IoU) of 0.7 for their application. The VNet [37] also adopts a volumetric CNN architecture, focusing on prostate segmentation from MR images by which, an average Dice score and a mean Hausdorff distances of 0.87±0.03 and 5.71±1.20 mm, respectively, was obtained. VNet was evaluated on the PROMISE-12 dataset. HighRes3dNet is an adapted CNN architecture based on dilated convolutions and residual connections [156], proposed for brain structures, achieving an average Dice score of 0.84±0.02. HolisticNet [157] is inspired by previous holistically-nested edge detection algorithms [158], which uses a generalisation of the Dice based on Wasserstein distance as the training loss. HolisticNet was proposed for brain tumour segmentation, reporting an average Dice of 0.89. Based on the VNet architecture, DenseVNet [135] was proposed to incorporate the densely-connected feature stacks. Compared to three other state-of-the-art algorithms, statistically significantly higher Dice scores for spleen, stomach, oesophagus, liver, left kidney, gall bladder and pancreas were achieved. The 3D adapted UNet is based on the original work of segmenting prostate gland from 2D TRUS images [10, 91], with an average Dice score and an average boundary distance of 0.91±0.12 and 1.23±1.46 mm, respectively. The original 2D network was extended to 3D by replacing all 2D operations such as convolution and pooling with the respective 3D operations. Table 5.1 summarises the experiment details of each network used in their original work, while Fig. 5.2 illustrates their network architecture. The readers are referred to the original papers and published code for other network details, which are kept unchanged in this work, for the interest of brevity.

5.2.2 Segmentation Metrics based on Hold-out Data

For comparison of automatic segmentations with the labelled ground-truth segmentations, two commonly adopted segmentation metrics are used, the DSC and the symmetric boundary distance (BD), given by:
Table 5.1: Information regarding the networks chosen for this comparison study.

<table>
<thead>
<tr>
<th>Network</th>
<th>Data Size</th>
<th>Application</th>
<th>Comparison to Other Methods</th>
<th>Statistical Significance Testing applied in Comparison?</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNet</td>
<td>3 samples (3-fold cross-validation)</td>
<td>Xenopus kidney embryos</td>
<td>2D UNet</td>
<td>No</td>
</tr>
<tr>
<td>VNet</td>
<td>80 subjects (single training-testing-split)</td>
<td>Prostate</td>
<td>Imorphics ScrAutoProstate SBIA Grislies</td>
<td>No</td>
</tr>
<tr>
<td>HighRes3dNet</td>
<td>543 subjects (single training-testing-split)</td>
<td>Brain</td>
<td>Deepmedic 3D UNet VNet</td>
<td>No</td>
</tr>
<tr>
<td>HolisticNet</td>
<td>274 subjects (single training-testing-split)</td>
<td>Brain</td>
<td>None</td>
<td>No</td>
</tr>
<tr>
<td>DenseVNet</td>
<td>90 subjects (9-fold cross-validation)</td>
<td>Abdominal</td>
<td>DEEDS+JLF VNet VoxResNet</td>
<td>Yes</td>
</tr>
<tr>
<td>Adapted UNet</td>
<td>98 (10-fold cross-validation)</td>
<td>2D Prostate</td>
<td>Fine-grained RNN</td>
<td>No</td>
</tr>
</tbody>
</table>

\[ DSC = \frac{2|X \cap Y|}{(|X| + |Y|)} \]

\[ BD = \frac{D(X,Y) + D(Y,X)}{2} \]

respectively, where X and Y are the automatically predicted binary segmentations and the manual ground-truth, respectively. The DSC is an overlap measure with a range of [0, 1]. \( D(X,Y) \) denotes the average Euclidean distance from boundary pixels in X to the closest boundary pixel in Y. These two metrics are adopted to directly measure the network generalisation ability in segmenting regions of interest on unseen hold-out data, here, full gland segmentation of MR images. Both measures were calculated on the largest resampled images with a size of [112, 128, 64] and an isotropic voxel size of [1, 1, 1] mm/voxel. The details of the validation experiment and the ground-truth segmentations used in this chapter are described in Sections 5.2.4 and 5.3.
5.2. Methods

![Architecture diagram of six networks](image)

**Figure 5.2:** Architecture of the six networks used for this comparison study. Different coloured arrows represent different architecture parts of the networks to visualise similarities and differences between them [11].

### 5.2.3 Gland Volume Errors and Estimated Target Registration Errors

As a potential clinical application of prostate MR segmentation, relative gland volume errors (GVEs) were also calculated between the network-segmented prostate gland and the manual ground-truth segmentation in the validation experiments, by counting the positive foreground voxels in the binary masks. GVE is based on the absolute difference between $V(X)$ and $V(Y)$ representing the volumes of the automatic and ground-truth segmentations, respectively:

$$\text{GVE} = |V(X) - V(Y)|$$
5.2. Methods

\[ GVE = \frac{|V(Y) - V(X)|}{V(Y)} \times 100 \]

Although an alternative regression network directly predicting volumes is possible, the GVE results may be useful to demonstrate a non-end-to-end prediction performance in a clinical scenario where, for example, full gland segmentation is required for other tasks such as localising tumours.

The MR-to-TRUS image registration can assist a range of TRUS-guided interventions, such as targeted biopsies and treatments [149]. Many proposed registration methods rely on matching prostate glands from (semi-) automated segmentation methods [111, 120, 159–163]. For the comparison purpose and for the reproducibility of this work, an open-source landmark-guided coherent point drift (LGCPD) algorithm is adopted [164] for deformable registration between the two point sets representing the surfaces of the prostate gland segmentations from MR and ultrasound images. The latter segmentations are obtained from our previous work [91] and remained fixed during all experiments for comparing different MR segmentations. The apex and base points are identified for all the cases, used as guiding landmark pairs with known correspondence in the LGCPD algorithm. The registration produces a non-rigid transformation between the MR and TRUS and this transformation is used to propagate MR landmarks to the space of the TRUS landmarks. Once registered, the root-mean-square distance between the transformed MR landmarks and TRUS landmarks is computed for each case as TRE, for different MR segmentations reported in this study. The landmarks include full gland segmentations, urethra, visible lesions, junctions between the gland, gland zonal separations, vas deference, seminal vesicles, visible lesions, and other patient-specific point landmarks such as calcifications and fluid-filled cysts. For many ROIs with smaller areas, they are good approximation of the corresponding homologous points. For other regions such as a specific section of urethra and boundaries of the zonal boundaries, defined ad hoc for specific cases, the centroids of these areas were used. The fidelity of this approximation would be conditionally affected by factors such as the accuracy of the delineation of the regions and how homogeneous the
deformation was between the two imaging. This, however, may still be considered as a measure of how good the registration was after alignment. A schematic of the registration workflow is displayed in Fig. 5.3.

5.2.4 Experiment Design for Network Comparison

In real-world applications, network hyperparameters are optimised before further clinical testing and adoption. To facilitate a comparison that is informative to clinical practice, it is desirable to find the optimum hyperparameter configurations prior to comparing these six networks described in Section 5.2.1. It is also important to note that estimating segmentation performance directly from a hyperparameter optimisation procedure, e.g. estimated DSCs from a cross-validation, is subject to overfitting, biasing towards the entire data set used for the hyperparameter-optimising cross-validation. Therefore, first the data is separated into development and holdout sets. The development set is used for hyperparameter searching; while the holdout set is used to report independent results on a data set completely unseen in the network development (including searching for hyperparameter values).

An exhaustive grid-search was adopted for tuning hyperparameters based on cross-validation. First, each of the tested hyperparameters is sampled at a uniform interval from a respective pre-defined range; Second, each permutation of these sampled hyperparameters (hereafter referred to as “hyperparameter configuration”) is tested in a $k$-fold (here, $k = 5$) cross-validation experiment. The details of the tested hyperparameter configurations are described in Section 5.3.3; Third, among these hyperpar-
Parameter configurations, segmentation performance is evaluated by the average DSC obtained from the $k$ network-training in the cross-validation; Finally, for each of the six network architectures, the hyperparameter configuration with the highest average DSC is selected. The division of the data used in this procedure is outlined in Fig. 5.4 and the data used in this study is described in Section 5.3.

The networks with the respectively-optimised hyperparameters are then tested on the hold-out data, for comparison purpose. All the segmentation (DSCs and BDs) and clinical measures (GVEs and TREs) described in Sections 5.2.2 and 5.2.3 were computed across all patients in the hold-out set. The two clinical measures (GVEs and TREs) and the DSCs are compared using a one-way analysis-of-variance (ANOVA) test at significance level of 0.05, among those produced by different networks. The ANOVA was followed by a multiple comparison pairwise t-test in each pair of networks to see where the significance in the group means lies, if any significance is obtained using the ANOVA test. This multi-group testing procedure was also performed using non-parametric tests, i.e. using a Kruskal-Wallis (KW) test to examine multiple group distributions. Since the normality of the data has not been tested to see whether there is an underlying statistical distribution, both the ANOVA and KW test were conducted to take into account both the case that the data comes from a normal distribution or not, respectively.
5.3 Experiments

5.3.1 Imaging Data and Ground-Truth Segmentations

The complete data used for this work consisted of T2-weighted MR prostate images taken from three different studies, SmartTarget Biopsy Trial [165], INDEX Trial [166] and the PICTURE Trial [167]. 232 MR image volumes were available from the same number of patients. These trials share the same imaging protocols. Original image size and voxel size range from [256, 256, 25] to [512, 512, 30] and [0.35, 0.35, 3] to [0.86, 0.86, 3.6], respectively. All images were scanned using either a 1.5T or 3T AvantoTM Siemens scanner. Intensity values were normalised to zero-mean and unit-variance intensities for individual volumes.

For all 232 MR image volumes, manual segmentation of the prostate capsule boundary in consecutive transverse slices of each MR volume was carried out by expert clinical observers (either a radiologist or a urologist specialised in MR-targeted procedures, verified by a senior radiologist). These segmentation labels provided ground-truth for segmentation in training and testing in this study.

Among the 232 image and segmentation data, 59 patient data from those taken from the SmartTarget Biopsy Trial [165] were used as the hold-out data set, unused to the hyperparameter searching. These patients had TRUS images available for further testing the subsequent MR-TRUS image registration application, in addition to the volume estimation results.

5.3.2 Implementation and Network Training

From the networks described in Section 5.2.1, for four of these; VNet, DenseVNet, HighRes3dNet and HolisticNet, the source code from NiftyNet [155] were directly used, while I implemented the published UNet [36] and Adapted UNet [91] in TensorFlow™ [168] which is also made publicly available. Each network was trained with a 12GB NVIDIA® Pascal™ TITAN Xp general-purpose graphic process units (GPU) on a high-performance computing cluster. The networks were run for 15000 iterations. During each 5-fold cross-validation for hyperparameter searching, the remaining 173 patients were split into five folds, each containing 33-
5.3. Experiments

35 (20%) patient data. Given a hyperparameter configuration, each of these five folds was left out for testing, with the network trained using the other 138-140 (80%) training data, this was repeated until every patient data was tested once, as shown in Fig. 5.4. This cross-validation procedure was repeated for each hyperparameter configuration (described in Section 5.3.3). Once the best hyperparameter was determined, five segmentations were predicted using the networks trained in the cross-validation on each of the 59 hold-out data. These five segmentations were then combined to generate the final segmentation using majority voting at each voxel, from which the segmentation accuracy and clinical metrics were computed, described in Sections 5.2.2 and 5.2.3, respectively.

5.3.3 Hyperparameter Configuration

To enable a computationally-feasible architecture comparison, four hyperparameters were varied to find the optimum combination of them for each network in this study, including input image size (after resampling from the original MR), initial learning rate of the Adam optimiser, regularisation weight of L2-norm on network parameters (weight decay) and number of initial feature channels. Table 5.2 summarises the four hyperparameters tested in this study, each with four different configurations, leading to a total of 256 hyperparameter configurations for each network.

The detailed values for these configurations are summarised in Table 5.2. The input images were resampled, from the centres of the image volumes, with respect to four different isotropic voxel sizes, [1, 1, 1] mm/voxel, [1.5, 1.5, 1.5] mm/voxel, [2, 2, 2] mm/voxel and [2.5, 2.5, 2.5] mm/voxel, with an empirically-set field of view. This resulted in the four sets of image sizes shown in the first row of Table 5.2. The field-of-views was cropped mainly for computational consideration with an estimate of a fixed physical region that is large enough to contain the entire prostate gland and most surrounding anatomical structures, and is the same for all the data used in this study. The number of initial feature channels represents a measure of network size [169] and, together with input image size, are constrained by GPU memory. Although the minibatch size could also affect the network training [170, 171], it
Table 5.2: Different hyperparameter configurations used for the hyperparameters tuning of the different networks.

<table>
<thead>
<tr>
<th>Training Hyperparameter</th>
<th>Value. 1</th>
<th>Value. 2</th>
<th>Value. 3</th>
<th>Value. 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Image Size</td>
<td>[112, 128, 64]</td>
<td>[80, 96, 48]</td>
<td>[48, 64, 32]</td>
<td>[32, 48, 16]</td>
</tr>
<tr>
<td>Initial Learning Rate</td>
<td>10e-2</td>
<td>10e-3</td>
<td>10e-4</td>
<td>10e-5</td>
</tr>
<tr>
<td>Weight Decay</td>
<td>0</td>
<td>10e-2</td>
<td>10e-4</td>
<td>10e-6</td>
</tr>
<tr>
<td>Number of Initial Channels</td>
<td>4</td>
<td>8</td>
<td>16</td>
<td>32</td>
</tr>
</tbody>
</table>

was found relatively insignificant in our initial experiment. In this study, minibatch sizes, 2, 4, 8 and 16 were fixed according to four decreasing input image sizes, in order to maximise the usage of the GPU memory.

The details of the other hyperparameters model architectures are kept the same as in the originally proposed networks. For the brevity, readers are referred to the respective original publications and open-sourced code.

5.4 Results

5.4.1 Hyperparameter Searching

256 different hyperparameter configurations were tested for the UNet and Adapted UNet, with the 5-fold cross-validation. Initial number of feature maps was not relevant for the other four networks, which had fixed model architectures without considering the change in the number of feature maps. Therefore, 64 hyperparameter configurations were tested for the VNet, HighRes3dNet, HolisticNet and DenseVNet. Based on the highest DSCs obtained from these experiments, the hyperparameter configurations found for each network is listed in Table 5.3. The networks trained with these hyperparameter configurations were used for the subsequent comparison reported in this work. The highest DSCs in addition to the 10th, 50th and 90th percentiles of the obtained DSC values from these experiments are also provided in Table 5.3.

5.4.2 Segmentation Accuracy

Fig. 5.5 shows a comparison of the example images overlaid with typical segmentations automatically generated from the trained networks, illustrating qualitatively different levels of segmentation performance at 25th, 50th and 75th percentiles of
Table 5.3: The selected hyperparameter configurations for each of the segmentation networks.

<table>
<thead>
<tr>
<th>Network</th>
<th>Input Image Size</th>
<th>Initial Learning Rate</th>
<th>Weight Decay</th>
<th>Number of Initial Channels</th>
<th>3D DSC</th>
<th>Max [10th, 50th, 90th] percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>VNet</td>
<td>[32, 48, 16]</td>
<td>10e-4</td>
<td>10e-4</td>
<td>n/a</td>
<td>0.87</td>
<td>[0.84, 0.85, 0.87]</td>
</tr>
<tr>
<td>HighRes3dNet</td>
<td>[32, 48, 16]</td>
<td>10e-2</td>
<td>0</td>
<td>n/a</td>
<td>0.87</td>
<td>[0.73, 0.84, 0.87]</td>
</tr>
<tr>
<td>HolisticNet</td>
<td>[32, 48, 16]</td>
<td>10e-2</td>
<td>10e-6</td>
<td>n/a</td>
<td>0.87</td>
<td>[0.19, 0.68, 0.87]</td>
</tr>
<tr>
<td>DenseVNet</td>
<td>[32, 48, 16]</td>
<td>10e-3</td>
<td>0</td>
<td>n/a</td>
<td>0.85</td>
<td>[0.76, 0.82, 0.85]</td>
</tr>
<tr>
<td>UNet</td>
<td>[48, 64, 32]</td>
<td>10e-2</td>
<td>10e-6</td>
<td>8</td>
<td>0.89</td>
<td>[0.67, 0.85, 0.88]</td>
</tr>
<tr>
<td>Adapted UNet</td>
<td>[48, 64, 32]</td>
<td>10e-3</td>
<td>10e-6</td>
<td>32</td>
<td>0.89</td>
<td>[0.67, 0.85, 0.88]</td>
</tr>
</tbody>
</table>

DSCs. The DSCs and BDs are summarised in Table 5.4 and Fig. 5.6. It shows that a range of the median DSCs between 0.86 and 0.90 and a median BD range between 1.9 mm and 2.4 mm were obtained from the six networks. However, the one-way ANOVA test shows a statistically significant difference between the DSCs, but not for the BDs, with \( p = 0.005 \) and \( p = 0.32 \), respectively, while the non-parametric Kruskal-Wallis test shows a statistically significant difference between both the DSCs and the BDs with \( p < 0.001 \) and \( p < 0.001 \), respectively. The subsequent multiple comparison, based on Tukey’s honest significance test, showed that the difference was caused by the UNet, producing for example, \( p = 0.01, 4 \times 10^{-3} \) and \( 0.03 \), compared with VNet, HighRes3dNet and HolisticNet, respectively, while no significance was found among the other networks (\( p \)-values range from 0.84 to 1.00). The detailed pairwise multiple comparison results are also summarised in Table 5.5.

Investigating further for the seemingly underperformed UNet revealed two outlier cases producing DSCs that were lower than 0.65, as example slices shown in Fig. 5.7. As reported in Section 5.4.1 and Table 5.3, a median DSC of 0.89 was obtained from the UNet training, non-inferior to training errors from other networks. This indicates a clear example of overfitting. A further discussion of the effect from these outliers on the subsequent clinical tasks are discussed in Section 5.4.3.

5.4.3 Volume Estimate Errors and Target Registration Errors

Using the segmentations reported in the previous sections, the relative GVEs and TREs are also summarised in Table 5.4 and Fig. 5.6. These networks estimated the gland volumes with median relative GVEs between 6.5% and 10.4% and median
5.4. Results

Figure 5.5: Automatic segmentations generated from different CNNs for 6 patients. The two top columns are patients with DSC closest to the 25th percentiles, middle two columns are patients with DSC closest to 50th percentiles and bottom two columns are patients with DSCs closest to the 75th percentiles. Blue shows the segmentation from HighRes3dNet, green from HolisticNet, brown from VNet, magenta the segmentation from DenseVNet, yellow the segmentation from the adapted UNet, cyan from UNet and red the manual segmentation. The last row shows the overlay of all segmentations on top of the original image for each patient [11].

TREs all lower than 3 mm. Most interestingly, no statistically significant difference was found among these networks using the one-way ANOVA test, either in GVEs ($p = 0.34$) or in TREs ($p = 0.26$). This lack of significance was also agreed with the non-parametric Kruskal-Wallis test, $p = 0.60$ and $p = 0.39$, in GVEs and TREs, respectively. Additional pairwise multiple comparison results are summarised in Table 5.5.
5.4. Results

Table 5.4: Segmentation performance metrics, prostate volume calculations and target registration errors between the manual and automatic segmentation for each network.

<table>
<thead>
<tr>
<th>Network</th>
<th>3D DSC mean±std [25th, 50th, 75th] percentiles</th>
<th>Boundary Distance (mm) mean±std [25th, 50th, 75th] percentiles</th>
<th>Relative GVE Difference (%) mean±std [25th, 50th, 75th] percentiles</th>
<th>Target Registration Error (mm) mean±std [25th, 50th, 75th] percentiles</th>
<th>Number of Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNet</td>
<td>0.84±0.07 [0.83, 0.86, 0.88]</td>
<td>2.52±1.48 [1.73, 2.07, 2.57]</td>
<td>11.29±9.62 [3.65, 9.83, 16.03]</td>
<td>2.72±0.51 [2.30, 2.82, 3.05]</td>
<td>294k</td>
</tr>
<tr>
<td>VNet</td>
<td>0.88±0.03 [0.87, 0.89, 0.90]</td>
<td>2.45±0.91 [1.78, 2.36, 2.88]</td>
<td>10.71±6.42 [6.32, 10.44, 14.25]</td>
<td>2.84±0.59 [2.43, 2.91, 3.18]</td>
<td>71,044k</td>
</tr>
<tr>
<td>HighRes3dNet</td>
<td>0.89±0.03 [0.88, 0.89, 0.91]</td>
<td>2.33±0.81 [1.71, 2.21, 2.73]</td>
<td>10.15±7.54 [4.77, 8.70, 13.66]</td>
<td>2.86±0.58 [2.36, 2.92, 3.26]</td>
<td>809k</td>
</tr>
<tr>
<td>HolisticNet</td>
<td>0.88±0.12 [0.88, 0.90, 0.92]</td>
<td>2.56±3.22 [1.62, 2.04, 2.50]</td>
<td>9.60±13.49 [2.77, 6.51, 13.66]</td>
<td>2.98±1.25 [2.36, 2.85, 3.20]</td>
<td>4241k</td>
</tr>
<tr>
<td>DenseVNet</td>
<td>0.88±0.03 [0.86, 0.88, 0.90]</td>
<td>2.34±3.66 [2.00, 2.37, 2.92]</td>
<td>10.78±8.85 [4.04, 7.06, 15.80]</td>
<td>2.83±1.57 [2.30, 2.91, 3.18]</td>
<td>867k</td>
</tr>
<tr>
<td>Adapted UNet</td>
<td>0.87±0.03 [0.85, 0.88, 0.90]</td>
<td>1.96±0.61 [1.52, 1.86, 2.22]</td>
<td>8.99±5.61 [4.33, 8.40, 12.44]</td>
<td>2.66±0.45 [2.33, 2.61, 3.02]</td>
<td>9401k</td>
</tr>
</tbody>
</table>

Figure 5.6: Box and Whisker plots of different measurement metrics for each of the segmentation networks [11].

A subject-level comparison of segmentation metric in DSCs versus the corresponding registration performance in TREs is illustrated in Figure 5.8. It shows little visual correlation between these two measures in any tested networks, and a Pearson’s correlation coefficient of 0.015 was obtained between DSCs and TREs. The two outlier cases with the UNet (as reported in Section 5.4.2), which was predominantly responsible for the significant difference in segmentation performance, did not deteriorate registration performance, with TREs being 2.94 mm and 3.27 mm. In these cases, the adverse effect from the relatively poor segmentation was probably mitigated by multiple landmarks and deformation regularisation used in the registration algorithm, therefore proving any potential causal difference may re-
5.5 Discussion and Conclusion

Table 5.5: P-values between the different networks for the DSCs, BDs, GVEs and TREs.

<table>
<thead>
<tr>
<th>Networks</th>
<th>DSC p-value (tukey-kramer)</th>
<th>BD p-value (tukey-kramer)</th>
<th>GVEs p-value (tukey-kramer)</th>
<th>TRE p-value (tukey-kramer)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNet vs VNet</td>
<td>0.01</td>
<td>1.00</td>
<td>0.96</td>
<td>0.95</td>
</tr>
<tr>
<td>UNet vs HighRes3dNet</td>
<td>3.97e-3</td>
<td>0.99</td>
<td>0.83</td>
<td>0.92</td>
</tr>
<tr>
<td>UNet vs HolisticNet</td>
<td>0.03</td>
<td>1.00</td>
<td>0.29</td>
<td>0.42</td>
</tr>
<tr>
<td>UNet vs Dense VNet</td>
<td>0.03</td>
<td>1.00</td>
<td>0.43</td>
<td>0.97</td>
</tr>
<tr>
<td>UNet vs Adapted UNet</td>
<td>0.15</td>
<td>0.38</td>
<td>0.64</td>
<td>1.00</td>
</tr>
<tr>
<td>VNet vs HighRes3dNet</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>VNet vs HolisticNet</td>
<td>1.00</td>
<td>1.00</td>
<td>0.83</td>
<td>0.92</td>
</tr>
<tr>
<td>VNet vs Dense VNet</td>
<td>1.00</td>
<td>1.00</td>
<td>0.93</td>
<td>1.00</td>
</tr>
<tr>
<td>VNet vs Adapted UNet</td>
<td>0.96</td>
<td>0.54</td>
<td>0.99</td>
<td>0.77</td>
</tr>
<tr>
<td>HighRes3dNet vs HolisticNet</td>
<td>0.99</td>
<td>0.97</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>HighRes3dNet vs Dense VNet</td>
<td>1.00</td>
<td>1.00</td>
<td>0.99</td>
<td>1.00</td>
</tr>
<tr>
<td>HighRes3dNet vs Adapted UNet</td>
<td>0.84</td>
<td>0.79</td>
<td>1.00</td>
<td>0.71</td>
</tr>
<tr>
<td>HolisticNet vs Dense VNet</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.89</td>
</tr>
<tr>
<td>HolisticNet vs Adapted UNet</td>
<td>0.99</td>
<td>0.30</td>
<td>1.00</td>
<td>0.19</td>
</tr>
<tr>
<td>Dense VNet vs Adapted UNet</td>
<td>0.99</td>
<td>0.48</td>
<td>1.00</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Figure 5.7: Overlay images of the automatic segmentations from the UNet on top of the original prostate image for two outlier patients producing DSCs of smaller than 0.65 [11].

quire substantially more data, or it would not make a difference in clinical use of the segmentation such as MR-to-TRUS registration tested in this study.

5.5 Discussion and Conclusion

In this chapter, six recently-proposed CNNs were compared to segment prostate glands in MR images. Segmentation performance was reported in terms of overlap measures and boundary distance in DSCs and BDs, respectively, based on a 232-patient data set with expert labels. Although the purpose of the work is not to recommend any network architectures over the others, an extensive experiment
including 3840 model training was designed to ensure a practically-feasible fair comparison on generalisation. The results were reported on a hold-out data set, after a parameter searching based on cross-validation. Furthermore, two real clinical applications were tested, in which the automatically-predicted segmentations were used for volume estimation and multimodal image registration. The results, in relative GVEs and TREs, were also compared statistically among all the networks. I believe that: 1) it is the first time that a comparison experiment based on a single data set of this size is reported for MR prostate segmentation; 2) it is the first study to investigate the values due to the difference in MR prostate segmentation in subsequent applications where they are used clinically; and 3) it is of reference value for wider medical image analysis research fields.

For example, one potential application for calculating the prostate volume, also briefly mentioned in Section 5.1, is for investigating the effect of a drug on the prostate over time, such as the work proposed by Moore et al. [143] which showed a 15% reduction in prostate volume and 36% reduction in tumour volume in patients given dutasteride, a drug for treating prostatic hyperplasia. In this case, the reported median relative GVEs between 6.5% and 10.4% (Table 5.4) cannot be considered insignificant for this application, and it will not be improved by changing network architectures among those tested in this study.

In the MR-to-TRUS registration application, all networks tested showed median TREs between 2.6 mm and 2.9 mm, comparable to other segmentation-based
registration methods [111, 152, 163]. As shown by Van et al. [172], for detecting a clinically significant tumour volume, a TRE of 3.1 mm is required. Results in Table 5.4, showed that the percentage of patients with a TRE smaller than 3.1 mm were 80%, 64%, 71%, 67%, 73% and 82% for UNet, VNet, HighRes3dNet, HolisticNet, Dense VNet and Adapted UNet respectively.

Moreover, I believe results from this work also corroborate with the findings in a number of previous studies in which the value of segmentation metrics, and the resulting league table positions in segmentation challenges, was cautioned [140, 173]. With evidence from the prostate segmentation in MR images, it is demonstrated that statistically significant difference in segmentation accuracy, does not necessarily lead to any detectable impact in the following clinical applications that use these segmentations.

Our conclusions need to be considered with limitations such as data size, access to segmentation networks that are designed for these clinical applications and the choice of method using these segmentations. For instance, the registration algorithm used in this work is an open-source algorithm that produced acceptable registration results, but it may be interesting to compare with other methods with or without using segmentations.

With the use of deep learning in medical imaging, especially with networks which are proposed to be used in clinics, representing model uncertainty is of huge importance. The uncertainty can arise from noisy data, uncertainty in the model parameters and structure uncertainty [174]. Although outside the current scope of this work, it will be interesting to investigate uncertainty information in future work to further quantify our confidence about each of the networks’ performance.

Segmentation metrics are unquestionably useful in evaluating the segmentation performance of different network architectures, and in being adapted as loss function for training learning-based algorithms. An architecture producing statistically significantly better segmentation results perhaps remains an important engineering goal. Rather than suggesting one network over others or implying any unimportance of network architectures, the results reported in this work suggest that future
research shall update the focus of improving network architecture from sole segmentation performance to one which also considers its value in the clinical application of interest. This work serves as a starting point for this shift by demonstrating that any found statistical significance cannot be generalised to downstream clinical scenarios without further validation.
Part II

Active Surveillance Management
In this second part of the thesis, I focus on the longitudinal change detection of prostate MR images of men on AS. Detecting longitudinal change is very important, especially for the AS population where the sooner the changes leading to progression of cancer are found, the earlier the patients can be re-assessed and offered treatment if significant progression has occurred. However, as mentioned in Chapter 2, it is often difficult to visually interpret these changes as many subtle changes may not be picked up by the eye, particularly between consecutive scans over the years. Registering the MR images over time may allow these changes to be picked out more clearly since any differences, which may be missed by simple visual registration [39], can be easier to see with the images aligned. Segmentation of the prostate gland allows a better localisation of the prostate, and as discussed in both Chapters 4 and 5, can be used as an initial step for prostate image registration, both between MRI-TRUS and MRI-MRI, the latter which may then be followed with longitudinal change detection. Therefore, the work carried out thus far, mainly the prostate gland segmentation on both MRI and TRUS images, can be used as an initial step for the longitudinal analysis and change detection (Fig. 2.1), where either feature extraction or deep learning methods may be used to look for these changes in the images (both methods are discussed in Chapters 7 and 8, respectively).
Chapter 6

Background and Literature Review

AS is an increasingly used management option for low-risk prostate cancer, as an alternative to radical surgery or radiotherapy in order to address the over-diagnosis and reduce over-treatment for this group of men. The exact selection criteria for placing men on AS varies between centres and countries, but most use variations of the D’Amico classification [175] in order to identify low-risk disease, where the risk is defined based on the following criteria: (low-risk: PSA $\leq 10\, \text{ng/mL}$, Gleason score $\leq 6$, stage T1-T2a; intermediate risk: $10 < \text{PSA} < 20\, \text{ng/mL}$, Gleason score = 7, stage T2b; high-risk: $\text{PSA} > 20\, \text{ng/mL}$, Gleason score $\geq 8$, stage T2c-T3a). Patients placed on AS undergo regular clinical examinations, PSA testing and prostate biopsies. In order to see whether the patient has progressed over time, a set of clinical thresholds are established to see if there is progression from insignificant to clinically significant disease. For the routine reassessment of the AS cohort, MRI has great potential as a tool in deciding which patients to include or exclude [18]. Different reporting systems exist, the most common of which are the Likert and PIRADS score which help standardize the process of diagnosing clinically significant cancers. While the Likert provides the radiologist an overall impression of the prostate to inform the score, PIRADS has a rules based approach focusing more on the lesion appearance itself. Using these reporting systems, many studies have shown the use of mpMRI for selecting AS candidates [176]. Marliere et al. [177] and Ouzzane et al. [149] showed that men who were previously diagnosed with a standard systematic approach were reclassified with rates ranging between
6.1 Texture Features for Change Detection

10-59% when using mpMRI targeted biopsy. Studies have also shown that lesions on mpMRI images with PIRADS scores of 4-5, i.e. highly suspicious lesions, are associated with clinically significant disease [178–180].

Perhaps more importantly, mpMRI is of significant interest for detecting changes predictive of radiological progression in men on AS over time. Although there is no well-defined schedule for when and how often these mpMRI scans need to be carried out, which is often determined by factors such as the baseline risk and the size and/or visibility of the lesion, a study by Rais-Bahrami et al. [181] proposed a monitoring interval of at least 2 years between the baseline and follow-up scans. Studies which have investigated the use of mpMRI for change detection include the study by Walton-Diaz et al. [148] where the authors concluded that using mpMRI reduced the average number of patients needing a biopsy for detecting pathological progression from 8.75 to 2.89. DWI, specifically the ADC maps, have shown association with pathological progression where a 10% reduction in tumour ADC indicated progression with sensitivity and specificity of 93% and 40%, respectively [148]. In another study by Felker et al. [182], while investigating the impact of radiological progression on pathological progression, the authors found that the addition of mpMRI parameters to models based on clinical variables increased the predictive value of the model from an AUC of 0.87 to 0.91. More recently, the Prostate Cancer Radiological Estimation of Change in Sequential Evaluation (PRECISE) guidelines [13] have been published for reporting change in MRIs in men on AS. These guidelines will be discussed in more detail in Chapter 8.

6.1 Texture Features for Change Detection

The texture of an image, refers to the appearance, structure and arrangement of parts of an object within an image [183]. Essentially, texture features are mathematical parameters which can be computed from the distribution of the pixels within the image and characterize the underlying structure of the objects within the image. Texture analysis has many applications in the field of medical imaging, such as segmentation of structures within the image, diagnosis of skeletal muscle dystro-
6.1. Texture Features for Change Detection

Texture analysis can be categorized into four different types [183], each of which is explained briefly below: 1) Structural Methods: In this type, the textures are represented using primitives, where primitives is a group of pixels representing the simplest subpattern [184], therefore providing a good symbolic description of the image. 2) Model-based Methods: In this type, mathematical models are used to represent the texture, with the model parameters being used for the image analysis. This approach is computationally expensive and requires complicated parameter estimation. 3) Statistical Methods: With statistical approaches, the texture is represented using the properties related to the distribution of the grey-level values in the image. 4) Transform Methods: Here the image is converted into another space, such as frequency space, and the texture properties are evaluated within this space. The most commonly used transform for texture analysis is the wavelet transform.

The texture methods most commonly used are the statistical methods [183], such as intensity histograms, gradients, the run-length matrix and the co-occurrence matrix. The co-occurrence matrix is the most commonly used in prostate cancer, our clinical application of interest. A co-occurrence matrix allows information to be extracted regarding the distribution of a pair of pixels in the image [12]. By specifying a distance and a direction (i.e. pixels to right, left, above or below), only pixels separated by this distance, and in this specific direction will be considered. An example of the calculation of such a grey level co-occurrence matrix (GLCM) is shown in Fig. 6.1 [12], where Haralick texture features are extracted from the image. A very simple grey-level image is shown at the beginning, followed by a grid of numerical values, corresponding to the grey-level intensity values. From the grid, the GLCM is calculated by considering the relationship between every pixel and the pixel to its right, and then counting the times that each two pixel values co-occur. All of the co-occurrences are added up and recorded in a table as shown in step 2 of Fig. 6.1. The GLCM is then normalised, and these probabilities are used to
6.1. Texture Features for Change Detection

Figure 6.1: Calculation of grey-level co-occurrence matrix and subsequent texture features from a grey-scale image, taken from [12]. Step 1 involves the calculation of the corresponding grid of numerical grey-scale values from the original image. Step 2 involves the calculation of the GLCM by counting all possible co-occurrences between different pixel values. Step 3 involves the normalisation of the GLCM to probabilities, and finally Step 4 uses these probabilities for the calculation of different texture features.

calculate different texture features shown in step 4. Each of these features provides information about the underlying relationship of the pixels: The energy captures the extent of similarity between the voxels; the entropy captures the amount of variation of the voxels and is a measure of the amount of disorder in the signal intensities; the correlation indicates how the different voxels are correlated with each other; homogeneity, as the name suggests, is a measure of how similar pairs of voxels are; and the contrast indicates the amount of dissimilarity between voxel pairs and measures the variation in intensities [12].

Many studies have shown that, as well as providing a basis for discriminating malignant tumour from normal tissue [185, 186], measures of ADC from DW MRI images may be useful in classifying disease aggressiveness in terms of Gleason grade [18]. ADC-based texture measures, such as those resulting from a Haralick texture analysis [187], may be useful. In this method, measures such as energy, correlation, homogeneity and contrast characterise the spatial variation of the grey
levels throughout an ADC map. In a recent study by Wibmer et al. [12], inertia was found to be significantly higher in cancerous regions whilst energy, correlation and homogeneity were significantly lower. Khalvati, F. et al. [188] used an extension of the Haralick texture features to include also Gabor and Kirsch filter features for detecting prostate cancer and found that this texture feature model produced a sensitivity of 86% and specificity of 88%. Other studies have also looked at using texture parameters for differentiation of healthy versus cancerous tissues [188, 189] while some have been considering absolute ADC values [185, 186]. These features will be explored in our work in Chapter 7.

6.2 Deep Learning for Change Detection

Recently, neural networks have been applied for change detection in a variety of applications, both in medical fields and others [190]. The range of applications include optical remote sensing [191], change detection in coastal landscapes [192], background subtraction from videos [193] and, for detecting change in synthetic aperture radar images [194–196]. More specifically, in medical imaging, convolutional neural networks have been used as change detection tools for tumours in CT liver scans [197], for quantifying aortic calcifications in abdominal CT [198], for detecting and staging the severity of meniscus and patellofemoral cartilage lesions in patients with osteoarthritis using MRI scans [199], for detecting changes in the retina in patients with diabetes [200], for the diagnosis of alzheimer’s disease [201], and investigating longitudinal change for the detection of tumours in mammography [202].

For prostate cancer, there are many reports in the literature on the use of MRI-related markers for classification of prostate cancer on MR images, using both deep learning and more traditional methods [203–207]. Yet, there exists less work on the use of these markers for predicting upgrading or progression of the cancer over time, especially on AS cohorts. One such work investigated the addition of MRI findings to current clinical factors and found its potential in improving the prediction of prostate cancer upgrading in AS patients, specifically MRI suspicious score and
total lesion density [151]. Morgan et al. [208] presented changes in tumour volume of up to 60% for patients on AS in a 1 year period, with changes in ADC being shown to correlate with changes in clinically significant volume growth.

However, none of the works mentioned above use CNNs for the temporal change detection on the AS cohort, for which, as discussed, detecting change as early as possible is very important. The changes can be investigated by analysing patterns and textures in the MRIs at different time points, and comparing differences and similarities over time for different groups of patients. In the remaining two chapters both non deep learning, in the form of texture features, and deep learning methods are used to investigate the temporal changes in men on AS as detected from MR images. Chapter 7 focuses on the use of texture features to observe the differences over time for patients on AS when given dutasteride compared to those placed on placebo, over a 6 month period. In Chapter 8 a CNN is used for predicting radiological progression over time on another AS cohort, using T2-weighted and ADC maps of the prostate.
Chapter 7

Haralick-based Texture-analysis for Cancer Detection in AS Patients

In this chapter texture features and volume measurements derived from ADC maps were analysed on an AS cohort on a dutasteride study to distinguish between the percentage of men given dutasteride and those given placebo. Observations were made both between the dutasteride and placebo group, and also between calculations computed from an automatic pipeline and those computed manually. The manually computed results were also compared to the results published in the literature.

7.1 Introduction

Mp-MRI has revolutionised the detection and management of early prostate cancer, and there is increasing interest in the role of this modality for evaluating prostate cancer aggressiveness. Mp-MRI is especially important for AS, where men are usually followed with serial MRI images [18, 209]. Recent studies have indicated that, as well as providing a basis for discriminating malignant tumour from normal tissue [185, 186], measures of ADC from DW MRI images may be useful in classifying disease aggressiveness in terms of Gleason grade [18]. However, determining tumour changes relies on significant skill and experience to identify often subtle morphological and image-intensity changes between scans.

Tumour heterogeneity is a crucial factor in predicting a tumour’s malignant po-
tential at a cellular level. Texture analysis (TA) provides a means to quantify signal heterogeneity in images by analysis of the regularity and coarseness of pixel/voxel value spatial distributions not visually perceptible to the human eye [210]. The revised PI-RADS 2 guidelines now advocate the use of tumour signal homogeneity on mp-MRI to grade disease [27, 30]. In particular, Haralick texture analysis [211] describes how often one grey tone will appear in a specified spatial relationship to another grey tone on the image [187]. A range of quantitative parameters (‘texture features’) are generated that characterize the local spatial variation of grey levels throughout an image. Successful applications of texture analysis have been documented in a variety of fields [211–215], including a number of works in the field of prostate cancer [12, 216–219].

To date, there has been very little research on automating the detection of changes in MRI images in prostate cancer patients on AS. Given the evidence from the studies highlighted above, it is reasonable to hypothesise that changes in ADC-based measures within prostate tumour regions over time might provide a useful additional biomarker to monitor cancer regression or progression in patients on AS. Such measures may be especially useful in cases where morphological changes are too small to detect reliably. Manual radiological analysis of DW MRI scans to compute these measures on longitudinal images typically requires manual delineation of tumour regions in each scan, which is labour-intensive. Therefore, automated computation of ADC-related measures from serial mp-MRI scans may be particularly helpful in the clinical setting.

In this chapter, the effects of dutasteride, a drug inhibiting the enzyme 5 alpha-reductase that converts testosterone to dihydrotestosterone (DHT), is investigated on an AS population. After puberty, DHT may be considered a “bad” hormone, as it is responsible for prostate growth, in addition to other problems. Therefore the role of the 5-alpha-reductase in the later stages of a man’s life is problematic as it may cause pathological prostate growth [220]. Dutasteride was developed in the 90s, and in 1998, early-phase clinical trial results were published. This inhibitor was found to lower DHT serum levels significantly [220]. Dutasteride has also been
shown to decrease the conspicuity of prostate cancer, related to the mean ADC value of the tumour [221], and therefore it is interesting to explore the effect of this drug on other intensity-based measures. To investigate the effects of this drug, a computational pipeline is proposed incorporating longitudinal ADC non-rigid registration for lesion propagation, in addition to the calculation of volume and texture features on ADC maps. Two different hypothesis were investigated: firstly comparison of the ADC-derived measures (including both volume measurements and textural measures) was made between patients receiving dutasteride and those receiving placebo, and secondly, comparison was also made between the resulting ADC-based measures computed using propagated tumour region segmentations and those computed using a manual segmentation of the tumour region by a radiologist.

7.2 Methods

7.2.1 Data

The data for this work came from the MAPPED study [222] which consisted of 42 men with biopsy-confirmed low-intermediate risk prostate cancer who had chosen AS. After an initial 3T mp-MRI scan, these men had two repeated scans at 3 and 6 months respectively [222]. Patients were randomised to either receiving dutasteride or placebo with equal allocation to each arm. Dutasteride has been shown to reduce the prostate volume by around 25% within 6 months [222], and the MAPPED study assessed its effect on MRI-visible prostate cancer over a 6 month study period. Patients on dutasteride were found to have a reduction in tumour volume and a reduction in tumour conspicuity over time, where conspicuity is defined as the mean ADC of the PZ divided by the mean ADC of the tumour [221]. Each of the 40 patients used in this work (2 patients left the study) had three mp-MRI scans including T2-weighted, DW (from which a corresponding ADC map is calculated) and DCE imaging sequences. In the analysis of this chapter, the focus was only on the ADC images. Using the ADC maps, a radiologist with 7 years of experience in mp-MRI interpretation (including DW imaging), manually contoured the cancerous region, blinded to clinical information and treatment allocation, for all 40 patients
on the three time points.

7.2.2 Key Pipeline Steps

The steps of the pipeline are summarised in Fig. 7.1 with each part described in more detail in the following subsections:

7.2.2.1 Image Registration and Contour Propagation

Automatic registration of ADC images was performed between the scan acquired at 6 month to the baseline scan using cubic B-spline non-rigid registration with normalised mutual information as the similarity measure (implemented as reg f3d in the ‘NiftyReg’ software [223]), shown in step 1 of the automatic analysis in Fig. 7.1. The transformation output from the non-rigid registration was then used to propagate the tumour region on the baseline scan to the space of each follow-up scan, at 3 and 6 months respectively (step 2 of the automatic analysis in Fig. 7.1). The interpolation scheme used for the propagation was nearest-neighbour. Validation of the automatic registration algorithm was carried out both visually, and by estimating the TREs for registered ADC images for a sampled subset of 5 patients using 7 corresponding anatomical landmark features visible in each registered image pair. Landmark placement was only done for 5 of the patients due to the time-consuming nature of the process, with each of the 7 landmarks taking time to locate on each scan. The landmarks used to estimate TRE for ADC images were uniquely identifiable features visible in both registered images. Four of the landmarks were chosen on the boundary and 3 were internal features within the prostate, including the entry and exit point of the urethra and cysts that appeared as small, discrete bright white spots that appeared in each image. The TRE was estimated as the post registration root-mean-squared error (RMSE) between corresponding manually identified anatomical landmarks on image pairs.

7.2.2.2 Texture and Volume Analysis

Texture features were analysed based on the work carried out by Khalvati et al. [188]. A total of 40 features were used, which included 4 first-order statistical features (mean, standard deviation (STD), skewness and kurtosis), 16 second-order
7.3 Results

7.3.1 Comparison Between Dutasteride and Placebo: Manual Contours

Table 7.1 summarizes the manually calculated lesion volume and texture features computed on the baseline and 6 months, for both the dutasteride and placebo arms of the patient population. The percentage change in each feature over the 6 months is also displayed in the table, in addition to whether there is any statistically signifi-
7.3. Results

Figure 7.1: Overall outline of the pipeline used for this chapter, divided into the manual and automatic pathways. In the automatic pathway (top), in step 1, the 6 months scan is registered to the baseline scan using an automatic registration software, and the transformation outputted from the registration is used to propagate the manually drawn lesion on the baseline scan to the space of the 6 months scan, as shown in step 2. In step 3, texture features and volume measurements are computed on both the manually contoured baseline and propagated 6 months lesion, and the differences between the lesion volume and texture features are calculated over time. The same procedure is carried out for the manual pathway (bottom), however, the registration and lesion propagation is excluded, since the lesion is manually contoured on both the baseline and the 6 months scan (step 1 and 2).

cant difference between the dutasteride and placebo groups. The change in volume showed a statistically significant difference ($p = 0.0039$) between the dutasteride and placebo group. Investigating the percentage changes of each feature between the two groups, it is clear from Table 7.1 that, there are differences between the changes of the dutasteride patients compared to those on placebo. For example, for both the Energy and Gabor feature, there is a reduction in the feature over time for patients on dutasteride, whereas an increase is observed for the placebo group (-2.91 ± 39.44% vs. 9.59 ± 39.00% and -17.66 ± 42.46% vs. 38.22 ± 123.35% for Energy
and Gabor change respectively). The opposite effect is also observed, i.e. where there is an increase over time for patients on dutasteride yet a decrease for those on placebo, as shown by the change in mean (2.02 ± 15.85% vs. -0.17 ± 2.83%) and change in Entropy (15.84 ± 60.72% vs. -4.28 ± 33.33%). However, none of these differences are statistically significant.

7.3.2 Comparison Between Dutasteride and Placebo: Automatic Contours

The corresponding mean ± STD TREs when registering the 6 month scan to the baseline scan for the subset of 5 patients, using the automatic pipeline, was 3.7±2.1 mm. Table 7.2 summarizes the calculated lesion volumes and texture features using the automatic pipeline, reporting measures on the baseline and 6 months scan for both the dutasteride and placebo arms, similarly to Table 7.1. For the propagated contours, the analysis between the two groups at the different time points was done on only 37 of the patients, due to the propagated lesion becoming too small on the 6 months scan for three of the patients to compute any of the measures. From the last column in this table we can see that none of the features showed a statistically significant difference using the t-test between the dutasteride and placebo groups. Larger percentage changes were observed for most of the features compared to Table 7.1. Possible reasons for this are discussed in the following section.

7.3.3 Comparison Between Manual and Automatic Measures

The boxplots in Fig. 7.2 show a comparison of the change in texture features and volume between the results obtained using the manual and automatic pipeline for the patients receiving dutasteride. A difference between the two pipelines can be observed from the figure, especially for measures such as STD, Skewness, Gabor and Kirsch features, where there is a difference in both the mean values and the range of values observed. For other features such as Kurtosis, Contrast, Energy, Entropy and Homogeneity, the automatic pipeline produces a smaller range of values across the 40 patients, compared to the manual pipeline. Comparison between the manual and automatic pipelines was also made by computing the $p$-value using the
Table 7.1: Volume measurements and texture feature values at baseline and 6 months for manually drawn lesions.

<table>
<thead>
<tr>
<th></th>
<th>Dutasteride (mean ± std)</th>
<th>Placebo (mean ± std)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline (Volume)</strong></td>
<td>0.0172 ± 0.0132</td>
<td>0.0143 ± 0.0099</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Volume)</strong></td>
<td>0.0121 ± 0.0103</td>
<td>0.0151 ± 0.0079</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Volume)</strong></td>
<td>-25.66 ± 33.56</td>
<td>31.39 ± 69.34</td>
<td>0.0039</td>
</tr>
<tr>
<td><strong>Baseline (Mean)</strong></td>
<td>1.02e3 ± 1.61e2</td>
<td>1.04e3 ± 2.60e2</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Mean)</strong></td>
<td>1.04e3 ± 1.76e2</td>
<td>9.93e2 ± 2.14e2</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Mean)</strong></td>
<td>2.02 ± 15.85</td>
<td>-0.17 ± 2.83</td>
<td>0.77</td>
</tr>
<tr>
<td><strong>Baseline (STD)</strong></td>
<td>123.60 ± 39.27</td>
<td>123.54 ± 34.20</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (STD)</strong></td>
<td>128.28 ± 63.49</td>
<td>130.62 ± 65.99</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (STD)</strong></td>
<td>6.42 ± 44.86</td>
<td>14.78 ± 64.98</td>
<td>0.64</td>
</tr>
<tr>
<td><strong>Baseline (Skewness)</strong></td>
<td>0.49 ± 0.59</td>
<td>0.068 ± 0.73</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Skewness)</strong></td>
<td>0.53 ± 0.59</td>
<td>0.64 ± 0.46</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Skewness)</strong></td>
<td>39.62 ± 258.58</td>
<td>1.2e3 ± 5.6e3</td>
<td>0.34</td>
</tr>
<tr>
<td><strong>Baseline (Kurtosis)</strong></td>
<td>3.00 ± 1.46</td>
<td>3.03 ± 1.54</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Kurtosis)</strong></td>
<td>2.79 ± 1.46</td>
<td>3.04 ± 1.07</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Kurtosis)</strong></td>
<td>8.47 ± 67.99</td>
<td>18.53 ± 62.38</td>
<td>0.64</td>
</tr>
<tr>
<td><strong>Baseline (Contrast)</strong></td>
<td>8.49 ± 2.24</td>
<td>10.55 ± 3.25</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Contrast)</strong></td>
<td>9.92 ± 4.26</td>
<td>9.13 ± 3.37</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Contrast)</strong></td>
<td>32.36 ± 80.74</td>
<td>2.58 ± 50.97</td>
<td>0.12</td>
</tr>
<tr>
<td><strong>Baseline (Energy)</strong></td>
<td>0.60 ± 0.13</td>
<td>0.54 ± 0.08</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Energy)</strong></td>
<td>0.56 ± 0.18</td>
<td>0.57 ± 0.15</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Energy)</strong></td>
<td>-2.91 ± 39.44</td>
<td>9.59 ± 39.00</td>
<td>0.33</td>
</tr>
<tr>
<td><strong>Baseline (Entropy)</strong></td>
<td>0.77 ± 0.22</td>
<td>0.86 ± 0.13</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Entropy)</strong></td>
<td>0.81 ± 0.33</td>
<td>0.80 ± 0.26</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Entropy)</strong></td>
<td>15.84 ± 60.72</td>
<td>-4.28 ± 33.33</td>
<td>0.22</td>
</tr>
<tr>
<td><strong>Baseline (Homogeneity)</strong></td>
<td>0.83 ± 0.04</td>
<td>0.79 ± 0.07</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Homogeneity)</strong></td>
<td>0.80 ± 0.09</td>
<td>0.82 ± 0.07</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Homogeneity)</strong></td>
<td>-2.96 ± 13.49</td>
<td>4.40 ± 13.26</td>
<td>0.10</td>
</tr>
<tr>
<td><strong>Baseline (Gabor)</strong></td>
<td>1.48e3 ± 1.35e3</td>
<td>1.16e3 ± 7.20e2</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Gabor)</strong></td>
<td>1.15e3 ± 1.10e3</td>
<td>1.18e3 ± 8.10e2</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Gabor)</strong></td>
<td>-17.66 ± 42.46</td>
<td>38.22 ± 123.35</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>Baseline (Kirsch)</strong></td>
<td>1.26e4 ± 2.20e3</td>
<td>1.27e4 ± 3.18e3</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Kirsch)</strong></td>
<td>1.28e4 ± 2.63e3</td>
<td>1.23e4 ± 2.91e3</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Kirsch)</strong></td>
<td>2.24 ± 16.93</td>
<td>1.03 ± 31.21</td>
<td>0.88</td>
</tr>
</tbody>
</table>
Table 7.2: Volume measurements and texture feature values at baseline and 6 months for propagated lesions.

<table>
<thead>
<tr>
<th></th>
<th><strong>Dutasteride (mean ± std)</strong></th>
<th><strong>Placebo (mean ± std)</strong></th>
<th><strong>p-value</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline (Volume)</strong></td>
<td>0.0172 ± 0.0132</td>
<td>0.0143 ± 0.0099</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Volume)</strong></td>
<td>0.0144 ± 0.0109</td>
<td>0.0132 ± 0.0091</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Volume)</strong></td>
<td>-12.27 ± 20.80</td>
<td>-7.39 ± 15.38</td>
<td>0.45</td>
</tr>
<tr>
<td><strong>Baseline (Mean)</strong></td>
<td>1.03e3 ± 1.70e2</td>
<td>1.05e3 ± 2.66e2</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Mean)</strong></td>
<td>1.09e3 ± 1.91e2</td>
<td>1.00e3 ± 3.23e2</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Mean)</strong></td>
<td>7.51 ± 17.63</td>
<td>-2.44 ± 26.65</td>
<td>0.20</td>
</tr>
<tr>
<td><strong>Baseline (STD)</strong></td>
<td>119.02 ± 36.79</td>
<td>127.19 ± 34.53</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (STD)</strong></td>
<td>142.03 ± 57.08</td>
<td>147.13 ± 58.37</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (STD)</strong></td>
<td>21.91 ± 38.13</td>
<td>19.82 ± 51.08</td>
<td>0.89</td>
</tr>
<tr>
<td><strong>Baseline (Skewness)</strong></td>
<td>0.54 ± 0.60</td>
<td>0.12 ± 0.75</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Skewness)</strong></td>
<td>0.17 ± 0.61</td>
<td>0.51 ± 0.61</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Skewness)</strong></td>
<td>34.27 ± 185.20</td>
<td>87.43 ± 192.96</td>
<td>0.42</td>
</tr>
<tr>
<td><strong>Baseline (Kurtosis)</strong></td>
<td>3.15 ± 1.47</td>
<td>3.07 ± 1.63</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Kurtosis)</strong></td>
<td>2.72 ± 0.90</td>
<td>3.22 ± 1.59</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Kurtosis)</strong></td>
<td>-0.95 ± 49.78</td>
<td>17.63 ± 61.85</td>
<td>0.34</td>
</tr>
<tr>
<td><strong>Baseline (Contrast)</strong></td>
<td>8.34 ± 2.28</td>
<td>10.61 ± 3.42</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Contrast)</strong></td>
<td>10.15 ± 3.84</td>
<td>12.44 ± 4.11</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Contrast)</strong></td>
<td>23.11 ± 42.64</td>
<td>20.25 ± 32.30</td>
<td>0.83</td>
</tr>
<tr>
<td><strong>Baseline (Energy)</strong></td>
<td>0.61 ± 0.13</td>
<td>0.54 ± 0.08</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Energy)</strong></td>
<td>0.53 ± 0.16</td>
<td>0.47 ± 0.11</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Energy)</strong></td>
<td>-11.40 ± 17.44</td>
<td>-11.41 ± 15.16</td>
<td>0.99</td>
</tr>
<tr>
<td><strong>Baseline (Entropy)</strong></td>
<td>0.76 ± 0.23</td>
<td>0.86 ± 0.14</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Entropy)</strong></td>
<td>0.818 ± 0.28</td>
<td>0.97 ± 0.17</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Entropy)</strong></td>
<td>16.93 ± 28.33</td>
<td>13.63 ± 17.46</td>
<td>0.69</td>
</tr>
<tr>
<td><strong>Baseline (Homogeneity)</strong></td>
<td>0.85 ± 0.04</td>
<td>0.81 ± 0.06</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Homogeneity)</strong></td>
<td>0.82 ± 0.07</td>
<td>0.78 ± 0.07</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Homogeneity)</strong></td>
<td>-3.79 ± 6.74</td>
<td>-3.96 ± 6.30</td>
<td>0.93</td>
</tr>
<tr>
<td><strong>Baseline (Gabor)</strong></td>
<td>1.54e3 ± 1.42e3</td>
<td>1.14e3 ± 6.92e2</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Gabor)</strong></td>
<td>1.44e3 ± 1.28e3</td>
<td>1.07e3 ± 6.58e2</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Gabor)</strong></td>
<td>2.05 ± 26.40</td>
<td>-4.62 ± 20.34</td>
<td>0.42</td>
</tr>
<tr>
<td><strong>Baseline (Kirsch)</strong></td>
<td>1.26e4 ± 2.32e3</td>
<td>1.29e4 ± 3.22e3</td>
<td></td>
</tr>
<tr>
<td><strong>6 Months (Kirsch)</strong></td>
<td>1.36e4 ± 2.36e3</td>
<td>1.27e4 ± 3.39e3</td>
<td></td>
</tr>
<tr>
<td><strong>% Change (Kirsch)</strong></td>
<td>9.42 ± 18.44</td>
<td>0.29 ± 21.19</td>
<td>0.19</td>
</tr>
</tbody>
</table>
paired t-test, where only the percentage change in energy and homogeneity showed a statistically significant difference between the manually and automatically computed values ($p = 0.01$ and $p = 0.03$ respectively). The reason for the discrepancies between the two pipelines is most likely due to the registration part of the automatic pipeline, not accounting for the shrinkage and displacement of the lesion over time. The lesion contour on the 6 months scan using both the manual and automatic pipeline is visualised in Fig. 7.3. The image shows the contoured image for both the manually drawn (shown in blue) and propagated contours (shown in red). Clear differences are seen between the two contours, especially in the shape and size of the lesion, where for the images A-C, the segmentation propagation (arising from the registration algorithms) leads to much smaller lesion contours than those drawn by the radiologist. For each of the six images, the centroid of the propagated lesion is also shown as a green dot. While for the first four cases (A-D), the centroid of the propagated lesion falls within the manually drawn lesion, this is not the case for the last two images (E-F), where there is a larger displacement between the manually drawn and propagated contour. Therefore the location to which the lesion is propagated may not be the ‘correct’ location of the lesion. In addition to this qualitative analysis, the DSC between the manual and propagated contours was computed at a mean value of 0.38, supporting the lack of alignment shown in Fig. 7.3. Additionally, since the automatically propagated 6 month lesion does not cover the complete ‘true’ lesion area, when computing the texture features on the propagated contour, pixels which are not within the lesion are most likely taken into account too. This leads to differences in the computed feature values, i.e. the computed texture feature values are computed over a region containing both cancerous and healthy tissue. The differences in size seen in Fig. 7.3 also explains the difference in the computed lesion volumes.

### 7.3.4 Correlation Between Change in Volume and Change in Texture Features

The original published paper on the MAPPED study showed a reduction in the apparent tumour volume for the patients receiving dutasteride [143], as was confirmed
### 7.3. Results

<table>
<thead>
<tr>
<th>Change in Volume from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Mean from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image2" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in STD from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Skewness from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image4" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Kurtosis from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image5" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Contrast from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image6" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Energy from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image7" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Entropy from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image8" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Homogeneity from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image9" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Gabor from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image10" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Kirsch from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image11" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Volume from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image12" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Mean from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image13" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in STD from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image14" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Skewness from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image15" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Kurtosis from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image16" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Contrast from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image17" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Energy from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image18" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Entropy from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image19" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Homogeneity from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image20" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Gabor from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image21" alt="" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Change in Kirsch from Manual and Automatic Pipeline</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image22" alt="" /></td>
</tr>
</tbody>
</table>

**Figure 7.2:** Box and Whisker plots of different texture measures calculated from the manual and automatic pipeline.

**Figure 7.3:** Comparison of manually contoured and propagated contour regions for six randomly selected patients. The manually drawn contours are shown in blue, while the automatically propagated contours are shown in red, with the centroids of the latter shown as a green dot.

From our results in Table 7.1. However, the paper did not analyse the effect of dutasteride on any of the texture features which were computed in this chapter. A subject-level comparison of the change in volume and change in the different texture measures is illustrated in Fig. 7.4. Homogeneity and Gabor filter features show
7.3. Results

Figure 7.4: Plot showing relationship between volume and each of the different texture measures, separated in two colours for the dutasteride and placebo group.

Table 7.3: Pearson’s correlation coefficient between change in volume and change in each texture feature.

<table>
<thead>
<tr>
<th>Patient</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.0930</td>
</tr>
<tr>
<td>STD</td>
<td>0.0687</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.2483</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>0.2042</td>
</tr>
<tr>
<td>Contrast</td>
<td>-0.4924</td>
</tr>
<tr>
<td>Energy</td>
<td>0.3935</td>
</tr>
<tr>
<td>Entropy</td>
<td>-0.3218</td>
</tr>
<tr>
<td>Homogeneity</td>
<td>0.5744</td>
</tr>
<tr>
<td>Gabor</td>
<td>0.5906</td>
</tr>
<tr>
<td>Kirsch</td>
<td>0.1336</td>
</tr>
</tbody>
</table>

the strongest correlation, out of all the features, with volume, with Pearson’s correlation coefficients of 0.57 and 0.59, respectively. On the other hand, the mean and STD show the weakest correlation with volume, with coefficients of 0.09 and 0.07, respectively. The coefficients for the rest of the features are summarised in Table 7.3.
7.4 Discussion and Conclusion

In this chapter, texture feature analysis and volume measurements were used to observe changes between patients given dutasteride compared to those given a placebo, where the lesion size of the patients given the dutasteride drug is expected to decrease in size. Calculations were done on both manually drawn lesion regions and regions automatically propagated using a segmentation-based label propagation. Using the manual pipeline, there was found to be a statistically significant difference in volume change between the patients on dutasteride and placebo, however, no statistically significant differences was found in longitudinal changes of any of the texture features. Unfortunately, these results were not supported by the automatic pipeline, which showed differences in results in both the volume and texture feature calculations.

Our conclusions need to be considered with limitations such as data size, since 40 patients may be a very small number to find patterns from and conclude significances from these patterns. A bigger dataset would be useful to validate these results. Another limitation arises from the registration accuracy, as studies in the literature (including those on MRI-ultrasound registration) suggest a lower bound on the TRE achievable using existing registration algorithms on the order 2-3 mm. However, the average registration errors calculated in this work were higher and were part of the reason for the differences in the computed measures between the manual and automatic pipeline. Therefore, improving the registration accuracy could lead to improvements in the rest of the analysis of the automatic pipeline.

Analysing longitudinal change, especially for patients on AS are unquestionably important, and the sooner these changes are detected the better. Although texture features have been shown to be useful in distinguishing cancerous vs. non-cancerous regions [12, 218], from the work presented in this chapter, they have not shown much promise in distinguishing between changes in different groups over time, therefore it may be concluded that dutasteride does not have any effect on the texture of the lesion over time. However, the manual selection of these features, in addition to the small data size mentioned in the limitations, may have impacted the
fact that not many detectable changes were observed. This leads to the following, and final chapter where instead of hand-crafted features, a deep learning method was used, with a larger dataset, for the task of change detection.
Chapter 8

Predicting Radiological Progression using CNNs

In this chapter, the use of CNNs is investigated for predicting longitudinal progression on an AS cohort, consisting of two or more follow-up visits. While the previous chapter also used an AS dataset, the focus was on the effects of a drug on changes in the lesion, whereas, in this chapter the predictive value of CNNs for radiological progression of men with low-risk cancer is investigated.

8.1 Introduction

The key idea in AS is the avoidance of treatment for low-grade, low-risk cancers unless evidence of disease progression is seen [224]. However, this idea involves two very important challenges. Firstly, accurate and early distinction of patients who have low-grade, low-risk cancers from those needing immediate treatment, for example, based on the D’Amico classification [175] described in Chapter 6, is needed. The second challenge is the accurate recognition of disease progression such that treatment options can be considered and delivered in a timely manner. The use of mp-MRI for monitoring men on AS varies between different healthcare centres and countries and there are few published data to inform of specific radiological changes for defining longitudinal progression. Last year the (PRECISE) recommendations were published [13, 225], establishing reporting standards for mpMRI in men on AS. The guidelines resulted in a PRECISE checklist outlining the key
8.1. Introduction

Figure 8.1: PRECISE case report form for reporting of MRI during baseline and follow-up [13].

Information that should be reported for men having an MRI on an AS cohort. A case report form is also designed for radiologists to report an MRI, either at baseline or follow-up [13]. The case report and checklist are displayed in Figs. 8.1 and 8.2, respectively. Although these recommendations allow for standardised reporting protocols for radiologists, manual reporting and assessment of true change can be time-consuming. Automating the process of change estimation over time allows for a faster and more accurate process, and there has been very little research focused on automating this process thus far. The only work available is by Francesco et al. [225], in which a software is used for automated comparison between sequential scans on AS, however, with the manual radiological contouring and reporting still necessary.

The main aims of the work in this chapter were to investigate the automation of the process of sequential change estimation from mp-MRI for men on AS using CNNs on two different sequences, ADC and T2-weighted images. The two questions which are investigated in this chapter are: 1. Can an automated method be used to produce the same results as a radiologist in predicting progression? 2. Can
the addition of a spatial prior regarding the prostate and lesion region improve the prediction performance?

8.2 Methods

8.2.1 Data and Pre-Processing

In this work, T2-weighted images and ADC-maps from 75 men on AS, acquired at two different time points, were used. These men were included if they had a follow-up scan between 5 and 24 months after the initial baseline MRI. All the images were
resampled to a voxel size of $0.5 \times 0.5 \times 3\text{mm}$ and output size of $128 \times 128 \times 82\text{ mm}$. Images were also resampled to zero mean and unit variance. The two sequences and time points were stacked together to form a 4-dimensional matrix for each patient, where the first dimension is of length 4 and represents the different sequences at each time point, i.e. Visit1—ADC, Visit1—T2, Visit2—ADC and Visit2—T2. Binary images representing progression or no-progression based on radiological reports at the second time point were used as the ground-truth labels. In addition to the whole MR images, prostate contours and lesion bounding boxes were also drawn on for each patients using the ITK-SNAP software [129]. From this data, 56% of the men (42/75) showed no progression between the two time points, while for the other 44% who showed progression, different types of changes were present. For 15% of the men (5/33), progression was in the form of an increase in volume, for 21% (7/33), progression was a change in the conspicuity of the lesion, for another 15% (5/33), both volume and intensity changes were present, while for 36% of the patients (12/33) progression was recorded as the appearance of a new lesion. In 12% of the men (4/33) progression was recorded due to resolved artefacts or technical differences. An example of all these progression types, for both T2 and ADC images, is shown in Fig. 8.3 below.

### 8.2.2 Neural Network Algorithm

The problem was formulated as a classification problem using a fully convolutional neural network (FCNN) with end-to-end training. The CNN used in this work is the DenseNet [142], which is used frequently for natural image classification tasks. The network takes a 4D volume of size $S_0 = [256 \times 256 \times 27 \times 4]$ as input, representing the image volume size $(x, y, z)$ followed by the initial number of feature channels (representing the two different sequences at the two time points). The Dense block consists of a batch normalisation, ReLU and convolutions, where a kernel size of $3 \times 3 \times 3$ is used for the convolution. Two classes are considered, progression vs. no progression, a 2-element vector defined where [1, 0] represents progression and [0, 1] represents no progression over time. The loss function used for training the networks was the cross-entropy and the model was optimised using the Adam opti-
Figure 8.3: Baseline and Follow-up images for 6 randomly selected patients, displaying changes in the T2 and ADC images over time, with the red arrow indicating the tumour region. A): Patient showing no progression over time, B): Progression recorded as volume change, C): Progression recorded as conspicuity change, D): Progression recorded as both volume and conspicuity change, E): Progression due to artefact involvement, F): Progression due to appearance of a new lesion.

miser. The network is shown in Fig. 8.4.

8.2.3 Experimental Design

8.2.3.1 Simulated Data

Firstly, I trained and tested the network on simulated data to make sure that the network was performing as it should before using the real image datasets.

To begin with, a binary prostate segmentation from one patient from the dataset was chosen as the ‘baseline’ prostate, and 200 different affine transformations were applied to this image, resulting in 200 different binary prostate images. Within each prostate, a lesion was also included, making sure the lesions are smaller than 0.5cm³, to represent clinically insignificant tumours [4]. Four different experiments were carried out, where for each experiment, half of the images (100) were sim-
8.2. Methods

Figure 8.4: Network architecture for AS progression prediction. The network architecture is kept the same across all three experiments, however, the input to the network is varied with each experiment, as shown on the left part of the figure. Experiment 1 takes the whole image of each sequence and time point (concatenated) and inputs this into the network. Experiment 2 takes the prostate region of each sequence and time point (concatenated) and inputs this into the network. Finally, experiment 3 takes the bounding box around the lesion region of each sequence and time point (concatenated) and inputs this into the network.

ulated as progressing while for the other half no progression was occurring. For the first experiment, progression was simulated by an increase in the volume of the lesion, where the lesion was dilated by 2 pixels on either side (on the slice which the lesion was most visible on). The reason for choosing 2 pixels as the dilation size of progression is due to the fact that, as shown by Morgan et al. [208], there was an increase of 52% in the mean lesion volume over a two-year period, and since we are considering follow-up visits between 5-24 months, the same percentage increase was used. In the second experiment, progression was simulated by an increase in conspicuity of the lesion. Giganti et al. [221] define conspicuity as the mean ADC of the PZ divided by the mean ADC of the tumour, and showed an increase in conspicuity from baseline to follow-up of 1.56 to 1.67, respectively. Since the intensity of the PZ of the prostate was kept constant, the intensity of the lesion
on the follow-up visit needs to be $\frac{1.67}{1.56} = 1.07$ times smaller. Therefore, in this experiment progression was simulated by decreasing the intensity of the lesion to 0.93% of its original value. In the third experiment, the change in volume and change in conspicuity were combined for the progression patients. Finally, in the fourth experiment in addition to the changes in volume and conspicuity, for a subset of patients, progression was simulated by the appearance of a new lesion, a factor which also occurs often in real data, as described in the previous section.

After simulating progression vs. no-progression using the binary segmentation of the prostate, in order to be more realistic, in the next part of the simulation, real data was used. However, apart from the lesion region, the rest of the prostate and surrounding area was kept constant between the baseline and follow-up visits, meaning the lesion region of the follow-up scan was used, but the same background as the baseline scan, so that any changes occurring are representative of the lesion region only. This was done by masking out and ‘replacing’ the voxel values corresponding to the lesion region on the baseline scan by the actual lesion values from the follow-up. For this simulation experiment, the same number of data as the real dataset was used, (75). All simulation results were carried out with a learning rate of 0.001.

8.2.3.2 AS Data

For the patient AS dataset, I carried out three set of experiments, as shown in the left hand side of Fig. 8.4. Firstly the whole MR images were used as input data to the neural network, which as described in Section 8.2.1, consisted of two sequences at two different time points. The second set of experiments involved inputting only the prostate region to the network, calculated by multiplying the binary mask of the prostate by the actual image. This resulted in an image with zero intensity where there is no prostate and original intensity values inside the prostate. The third and final experiment inputted only a bounding box around the lesion region to the network. To draw the bounding box, the prostate was divided into approximately four parts containing different zonal regions of the prostate, as shown in Fig. 8.5. Based on which region the lesion was placed, a bounding box was drawn around
8.2. Methods

Figure 8.5: Division of the prostate into different zones/bounding boxes for locating the lesion region.

that whole region. As in the second experiment, the binary mask of the bounding box was multiplied with the original image and this bounding box was used as the input to the network.

8.2.3.3 Transfer Learning with the PROSTATEx Dataset

Finally, in order to overcome the problem of the limited data, transfer learning was carried out using the data from the PROSTATEx challenge. The dataset contains 203 patients used for training the DenseNet network for 2000 iterations. The trained model was saved and the weights of this model were used to fine-tune the final layer of the DenseNet for our dataset. The early layers in the network extract only low-level features (such as edges), which would be similar between our dataset and the PROSTATEx dataset, whereas the upper layers are more task-specific and learn the high-level features for classification. Therefore, the last layer of the network is fine-tuned for the new task to learn the appropriate task-specific features. Since the PROSTATEx dataset contains only one time point for both the T2 and ADC images, these were used as the second time point, and the first time point was initialised as a black (zero-intensity) image. The transfer learning was carried out only with the whole image data as input, an overview is presented in Fig. 8.6.
Figure 8.6: Overview of our transfer learning architecture. The DenseNet is trained on the PROSTATEx dataset, and the weights of the trained network transferred and fine-tuned to our dataset.

8.2.4 Implementation and Cross-validation

The network was implemented in Tensorflow and trained using a 12GB Nvidia® Pascal™ TITAN Xp general-purpose GPU on a high-performance computing cluster for 15000 iterations with 8 volumes in each minibatch. The weight decay was set to 0 and the initial learning rate was varied from 0.01 to 0.0001. A 10-fold patient level cross validation was carried in which images from 7-8 patients were held out for testing, while the remaining patients were used for training the networks. This was repeated until each of all 75 patients were used for evaluation once. The measures reported below is the accuracy between the prediction and ground-truth of the held-out data.
8.3 Results

### 8.3.1 Simulated Data

Table 8.1 summarises the training and testing accuracy of the network for the four different experiments carried out using the simulated data from the binary prostate segmentation, in addition to the simulated real data experiment. Between 98-100% accuracy is obtained during the training across all experiments, and a mean testing accuracy across all folds ranging from 0.72-0.99 between the five experiments is obtained. From the table, it is clear that the first four experiments, which simulated different types of progression changes using the binary prostate segmentation, resulted in very high testing accuracies. Experiment 4, where progression was simulated using the appearance of a new lesion, resulted in the smallest prediction accuracy of $0.95 \pm 0.02$, producing significantly lower accuracies compared to the simplest cases of simulating changes in volume or intensity, with $p < 0.01$ and $p = 0.03$, respectively. Simulations using the real data resulted in testing accuracies of $0.72 \pm 0.12$, significantly smaller than the simulations using the binary prostate segmentations, and therefore, showing the effect of real lesion and background intensities on the prediction ability of the network. Although the network is able to learn well the different changes corresponding to lesion progression, the performance is compromised by the true intensities and more subtle changes visible in real data.

### 8.3.2 AS Dataset

The training and testing accuracy from the three different sets of experiments carried out on the AS dataset is displayed in Table 8.2. For all three experiments,

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Training Accuracy (mean ± std)</th>
<th>Testing Accuracy (mean ± std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulated Data Binary Prostate Experiment1</td>
<td>1.00 ± 0.00</td>
<td>0.99 ± 0.00</td>
</tr>
<tr>
<td>Simulated Data Binary Prostate Experiment2</td>
<td>0.98 ± 0.01</td>
<td>0.97 ± 0.01</td>
</tr>
<tr>
<td>Simulated Data Binary Prostate Experiment3</td>
<td>0.98 ± 0.01</td>
<td>0.96 ± 0.01</td>
</tr>
<tr>
<td>Simulated Data Binary Prostate Experiment4</td>
<td>0.98 ± 0.01</td>
<td>0.95 ± 0.02</td>
</tr>
<tr>
<td>Simulated Data Real Images</td>
<td>0.98 ± 0.01</td>
<td>0.72 ± 0.12</td>
</tr>
</tbody>
</table>
Table 8.2: Training and testing accuracy obtained using different input data and learning rates for the AS dataset. Whole Prostate MR is when the whole original image, containing the prostate and surrounding structures is used as input to the network. Only Prostate Region involved inputting only the segmented prostate gland, with the original intensities to the network, while, Only Lesion Bounding describes the experiments where only a bounding box covering the lesion area is used as input to the network.

<table>
<thead>
<tr>
<th>Data</th>
<th>Training Accuracy (mean ± std)</th>
<th>Testing Accuracy (mean ± std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole Prostate MR (LR = 0.001)</td>
<td>1.00 ± 0.00</td>
<td>0.53 ± 0.17</td>
</tr>
<tr>
<td>Whole Prostate MR (LR = 0.0001)</td>
<td>1.00 ± 0.00</td>
<td>0.56 ± 0.21</td>
</tr>
<tr>
<td>Whole Prostate MR (LR = 0.00001)</td>
<td>1.00 ± 0.00</td>
<td>0.63 ± 0.16</td>
</tr>
<tr>
<td>Only Prostate Region (LR = 0.001)</td>
<td>1.00 ± 0.00</td>
<td>0.66 ± 0.15</td>
</tr>
<tr>
<td>Only Prostate Region (LR = 0.0001)</td>
<td>1.00 ± 0.00</td>
<td>0.70 ± 0.16</td>
</tr>
<tr>
<td>Only Prostate Region (LR = 0.00001)</td>
<td>1.00 ± 0.00</td>
<td>0.82 ± 0.18</td>
</tr>
<tr>
<td>Only Lesion Bounding Box (LR = 0.001)</td>
<td>1.00 ± 0.00</td>
<td>0.66 ± 0.17</td>
</tr>
<tr>
<td>Only Lesion Bounding Box (LR = 0.0001)</td>
<td>1.00 ± 0.00</td>
<td>0.67 ± 0.13</td>
</tr>
<tr>
<td>Only Lesion Bounding Box (LR = 0.00001)</td>
<td>1.00 ± 0.00</td>
<td>0.70 ± 0.15</td>
</tr>
</tbody>
</table>

and the range of learning rates used, the training achieves accuracy of minimum 100%. The testing accuracy however is not as high, with a range of mean testing accuracy between 0.53-0.82. Using the optimum learning rate for each experiment, inputting the prostate region to the network (experiment 2), produces a significantly higher accuracy compared to experiment 1 ($p = 0.03$). Similarly the same pattern is observed when using the lesion bounding box as input (experiment 3), producing higher accuracy compared to experiment 1 ($p = 0.4$), however, a smaller mean accuracy compared to experiment 2 is obtained (0.70 compared to 0.82). It is also interesting to observe that comparing the first row of Table 8.2, to the last row of Table 8.1 which simulated progression on the real data, a statistically significantly higher testing accuracy is obtained using the simulation experiment ($p = 0.02$), showing the effect of the natural changes occurring outside of the lesion, within the prostate, on the prediction accuracy.

### 8.3.3 PROSTATEx Transfer Learning

By training the network on the PROSTATEx dataset and fine-tuning the fully convolutional layer on the AS data, the average training and testing accuracies obtained using the different learning rates are displayed in Table 8.3. Comparing the results
Table 8.3: Training and testing accuracy obtained when pre-trained on the PROSTATEX dataset with different learning rates.

<table>
<thead>
<tr>
<th>Data</th>
<th>Training Accuracy (mean ± std)</th>
<th>Testing Accuracy (mean ± std)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole Prostate MR (LR = 0.001)</td>
<td>0.98 ± 0.00</td>
<td>0.57 ± 0.13</td>
</tr>
<tr>
<td>Whole Prostate MR (LR = 0.0001)</td>
<td>0.98 ± 0.00</td>
<td>0.59 ± 0.18</td>
</tr>
<tr>
<td>Whole Prostate MR (LR = 0.00001)</td>
<td>0.98 ± 0.01</td>
<td>0.64 ± 0.13</td>
</tr>
</tbody>
</table>

in Table 8.3 and Table 8.2, there is an improvement on the average testing accuracy for all learning rates, for the experiments using the 'Whole Prostate MR’, when carrying out the transfer learning. None of the differences, however, were statistically significant.

8.4 Discussion and Conclusion

In this chapter, the concept of longitudinal change detection was extended using a CNN for predicting radiological progression on MRI of men on an AS cohort. The accuracy of the predictions was calculated using firstly the whole images, secondly using images containing only the prostate region, and finally, images containing a bounding box around the lesion area. With the addition of spatial priors in the form of prostate and lesion regions, an improvement was observed in the progression accuracy of the network.

Although the data size used in this chapter is larger than what was used in Chapter 7, it may still be insufficient, especially for a deep learning framework. This is because the model will not be able to learn the true pattern in the signal with a small dataset, and instead, can lead to overfitting of the model (as shown in our results with the high training yet smaller testing accuracies). In future work, a bigger dataset will be collected, which will be representative of a larger variety of longitudinal changes, allowing for better analysis. Ideally, with a larger dataset, the different patterns of progression, in terms of changes in volume and intensity and conspicuity, will be able to be modelled better by the network, therefore leading to higher testing accuracy results. Since this chapter presents early work on the topic of using CNNs for change detection, a simple architecture was used to begin with,
with only two layers (limited by the data size), which may have limited the network to learning only features such as edges and shapes, not enough to generalise to new data. For future work, other more complex and deeper architectures may be used, achievable by cropping or resampling the data to a smaller size, for improving the performance by learning more high-level features for better generalisation. Another limitation of this work lies in the limited number of time points used. Two time points may not be enough for analysing any change/progression in the patients, therefore, future work can also focus on the addition of more time points to the network.

This work, and research topic in general, will have the potential to be used as a tool for accurate change detection of men on AS, and more importantly, inform clinicians on progression to clinically-significant disease as early as possible for the most appropriate intervention.
Chapter 9

Conclusions and Future work

9.1 Contributions of Thesis

Two main contributions are presented in this thesis:

Firstly, a novel segmentation method for the task of prostate gland segmentation on TRUS images is developed and described in Chapter 4, based on convolutional neural networks. At the time of writing this was one of the first CNNs developed for the task of TRUS prostate segmentation. The performance of the network on the 109 patient data was investigated and promising results, comparable to inter-observer accuracy, were reported. The segmentation network was used also for segmenting the prostate gland on MR images, within a comparison study presented in Chapter 5. The segmentation performance of our method, in addition to 5 other publicly available CNNs, were compared both in terms of segmentation metrics (Dice and boundary distance) and clinical metrics related to their subsequent clinical use, such as gland volume measurements and target registration errors. This allows a new way of evaluating the performance of segmentation networks, by considering not just the segmentation, but more importantly the end goal use of these segmentations.

Secondly, a pipeline has been proposed for analysing longitudinal change for men on AS, using both Haralick texture feature analysis and CNN based analysis, described in Chapters 7 and 8, respectively. These include both detecting changes over time for groups of men given an active drug (dutasteride vs. placebo), and
Future Work

The work carried out on TRUS prostate segmentation in Chapter 4, has shown very promising results and is in the process of being implemented into the SmartTarget device, a state of the art system used for precision targeting of prostate cancer, for automating the TRUS segmentation part of their workflow. At the moment the segmentation of the prostate on the TRUS images is carried out by manual point placement on the clinical system, therefore, by automating this process it will remove the time and burden of manually selecting points.

Regarding the comparison study in Chapter 5, we hope that this serves as a starting point for a shift in the medical image segmentation community to consider carefully the specific downstream tasks of interest within a computational pipeline for the specific clinical application. Going forward, many more emerging CNNs can be added into the comparison to allow a larger variety of networks to be compared, including more complicated networks such as attention networks and region proposal based networks. In addition to comparing segmentation methods, different registration networks can also be compared in future work.

The most important future direction of the thesis will be on the work regarding the AS patients. During the final year of the PhD where this work was carried out, we managed to collect 138 patients, of which 75 were used in our investigations, however, the database contains around 554 men all with longitudinal mp-MRI. Therefore, firstly the task will be to keep collecting more data over time, increasing the datasize used in the CNN which will allow more patterns and longitudinal variations to be learnt from the data. In addition to collecting more raw data, the manual delineations of the tumour also need to be obtained from the radiologist, since at the moment I have used boundary boxes around the regions of the
lesion, but not the exact lesion delineation itself. Getting access to all the images and delineations is a long procedure which will be continuous for another year at least.

Another important question regarding this dataset which can be investigated in the future, would be prediction of the time of progression, rather than accurately predicting radiological progression as is done at the moment. For example, the subset of patients who had no lesion present at the baseline MRI scan but showed progression after the 3rd/4th scan, would be chosen. Using the baseline and the first follow-up scan, it will be investigated whether any characteristics could be picked up which are predictive of the progression outcome of these men. This question is of huge clinical importance, since the clinician wants to make sure that the men who progress or need treatment are given this treatment as soon as possible while reducing the surveillance and any unnecessary treatment for those who have no progression. Therefore, it is important to use these deep learning methods to investigate whether relevant changes can be picked up earlier than is done by a radiologist. There are also many ways in which the texture analysis work carried out in Chapter 7 can be fed into the work of Chapter 8, for example by taking the group of men who did have a visible lesion on the baseline scan and comparing the lesion characteristics, between the subgroup of these men who progressed to treatment and those who did not. This is so that we can understand whether any differences can be picked up within or surrounding the lesion which may allow the clinicians to determine whether the cancer would be progressing as early as the baseline scan. Finally, correlation between MR images and biopsy reports may also be investigated, since it is important to find the correlation between information obtained on MR images and any clinical significance shown on the biopsy findings. This is important in order to be able to find any progression of the tumour from low-grade, clinically insignificant to high-grade disease as early on as possible for more efficient treatment.
Bibliography


[13] Caroline M Moore, Francesco Giganti, Peter Albertsen, Clare Allen, Chris Bangma, Alberto Briganti, Peter Carroll, Masoom Haider, Veeru Kasivis-


[57] Jing Yuan, Wu Qiu, Eranga Ukwatta, Martin Rajchl, Yue Sun, and Aaron Fenster. An efficient convex optimization approach to 3d prostate mri seg-


[63] Philip D Allen, James Graham, David C Williamson, and Charles E Hutchinson. Differential segmentation of the prostate in mr images using combined


[78] Ruey-Feng Chang, Wen-Jie Wu, Woo Kyung Moon, Wei-Ming Chen, Wei Lee, and Dar-Ren Chen. Segmentation of breast tumor in three-dimensional...


[114] Andriy Fedorov, Siavash Khallaghi, C Antonio Sánchez, Andras Lasso, Sidney Fels, Kemal Tuncali, Emily Neubauer Sugar, Tina Kapur, Chenxi Zhang,


[121] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henk-


[128] Yuan Gao and J Alison Noble. Detection and characterization of the fetal heartbeat in free-hand ultrasound sweeps with weakly-supervised two-


[148] Annerleim Walton Diaz, Nabeel Ahmad Shakir, Arvin K George, Soroush Rais-Bahrami, Baris Turkbey, Jason T Rothwax, Lambros Stamatakis,


[171] Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le.


[190] Richard J Radke, Srinivas Andra, Omar Al-Kofahi, and Badrinath Roysam.


[210] Fergus Davnall, Connie SP Yip, Gunnar Ljungqvist, Mariyah Selmi, Francesca Ng, Bal Sanghera, Balaji Ganeshan, Kenneth A Miles, Gary J Cook, and Vicky Goh. Assessment of tumor heterogeneity: an emerging


