Title
Advancing Segmentation and Classification Methods in Magnetic Resonance Imaging via Artificial Intelligence

Author
Liu, Yongkai

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Advancing Segmentation and Classification Methods in Magnetic Resonance Imaging via Artificial Intelligence

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Physics & Biology in Medicine

by

Yongkai Liu

2022

ABSTRACT OF THE DISSERTATION

Advancing Segmentation and Classification Methods in Magnetic Resonance Imaging via Artificial Intelligence

by

Yongkai Liu

Doctor of Philosophy in Physics & Biology in Medicine

University of California, Los Angeles, 2022

Professor Kyunghyun Sung, Chair

Magnetic Resonance Imaging (MRI) is a non-invasive imaging technology that produces detailed anatomical images, which has provided a giant leap forward in medical diagnosis. MRI segmentation and classification play an essential role in disease assessment and detection, volume measurement, and biopsy. Methods relying on feature engineering have traditionally been used to perform MRI segmentation and classification and usually produce sub-optimal results. With the fast growth of artificial intelligence, deep learning has achieved great success and outperformed these traditional methods. However, current deep learning models, especially in prostate MRI segmentation and classification, may provide an insufficient representative power, and lack prediction uncertainty information and prior knowledge such as cancer heterogeneity, and usually

require large-scale and generalizability evaluation. Primarily with prostate MRI, this dissertation concerns several advanced deep learning-based MRI segmentation and classification methods or applications to address the above issues. The contributions of this thesis are as follows:

1. A new deep learning method with feature pyramid attention to enhance multi-scaled and high-level feature extraction was developed for automated prostate zonal segmentation. The proposed method outperformed state-of-art deep learning-based prostate zonal segmentation method such as U-Net.
2. The attention mechanism improves the deep learning-based segmentation by focusing more on relevant information to the region of interest, and Bayesian statistics equips deep learning with uncertainty measurement. An attentive Bayesian deep learning network was developed for the prostate zonal segmentation with uncertainty estimation. The proposed method was superior to the first method developed above on prostate zonal segmentation. Uncertainties produced between different prostate zones at three prostate locations were consistent with the actual model performance.
3. Texture provides the prior knowledge that can quantitatively describe the tumor heterogeneity. Texture-based deep learning (Textured-DL), which can be potentially used in a small dataset due to the exploitation of tumor prior information, was proposed for the prostate cancer classification. The Textured-DL showed superior performance to the radiologist-based classification, conventional machine learning, and deep learning methods.
4. A previously developed deep learning model in the second method above, attentive Bayesian deep learning network, was evaluated for the whole prostate gland segmentation using a large patient cohort. In the qualitative evaluation, the deep learning method demonstrated acceptable

or excellent segmentation quality in most cases. The deep learning method was superior to the state-of-art deep learning methods in the quantitative evaluation.

5. A previously developed deep learning model in the second method above, attentive Bayesian deep learning network, was tailored and used for the placental segmentation on longitudinal MRI to investigate the model's generalizability for other biomedical image applications. The deep learning model can automatically segment the placenta with high accuracy. In addition, placental volume measurement with the deep learning-based and manual segmentation can be used interchangeably.

In summary, the deep learning model with feature pyramid attention and attentive Bayesian deep learning method achieved superior prostate zonal segmentation performance; enriching the image prior knowledge to the deep learning enhances the prostate cancer classification; large-scale and generalizability evaluation further demonstrated the segmentation model's outstanding segmentation and generalizability abilities. Future studies will explore the prior knowledge that can enhance the segmentation performance, study the contour-based fast segmentation using graph convolution, explore the non-textured and clinical features that could enhance the classification performance, and analyze the effect of data size on the Textured-DL's classification performance.

The dissertation of Yongkai Liu is approved.

Michael Albert Thomas

Holden H. Wu

Peng Hu

Guang Yang

Kyunghyun Sung, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENT

First, I would like to thank my advisor, Professor Kyunghyun Sung, who provided considerable support and comprehensive mentorship for the past five years of my Ph.D. Without him as an exemplar, I cannot achieve what I have done so far. He not only provided knowledgeably, propound, insightful advice in my research but also shared his wisdom on work-life balance and communication. No matter how busy he was, he is always warm and kind to me and greeted me with a smile whenever I needed his help in my research and life.

Secondly, I would like to express my gratitude and appreciation to Dr. Thomas Albert, Dr. Peng Hu, Dr. Holden H. Wu, and Dr. Guang Yang for being on my committee. Their support helped me refine the research projects and form the foundation of the dissertation.

I would like to thank my collaborators in MRRL. I want to express my special gratitude to Dr. Fatemeh Zabihollah, who encouraged me and provided valuable comments on my dissertation. I also want to thank Ran Yan, who provided great comments and advice on my dissertation organization. I also want to thank all the past and current members in MRRL for all the informative and exciting discussions and meticulous supports. I want to give thanks to Xinzhou, Jiahao, Le, Zhaohuan, Kai, Alex, Shufu, Andres, Haoxin, Chang, Caroline, Jiaxin, Tess, Fadil and Alibek.

I would also like to thank my collaborators outside MRRL. Thank you, Dr. Steven Raman, Dr. Miao Qi, Dr. Sohrab, and Dr. Melina for providing professional clinical perspectives on my projects. I still remember Dr. Qi's finger getting hurt during the large-cohort evaluation study. Thank you for the spirit of working hard and sacrificing. I am looking incredibly forward to continuing working with you after graduation.

Last and most importantly, I would like to express my sincere thanks to my parents and siblings. They provided endless support to all aspects of my life. Whenever I need help, they are

always the first to provide help. They encouraged me when I felt sad during the bad times. They taught me to be humble when I succeeded in my research projects.

VITA

Education

2014 – 2017	Tsinghua University	Beijing, China
	M.S., Biomedical Engineering	
2010 – 2014	South China Agricultural University	Guangzhou, China
	B.S., Electronic Information Science and Technology	

First Author Peer Reviewed Publications

1. **Yongkai Liu**, Haoxin Zheng, Zhengrong Liang, Qi Miao, Wayne G. Brisbane, Leonard S. Marks, Steven S. Raman, Robert E. Reiter, Guang Yang, and Kyunghyun Sung. "Textured-Based Deep Learning in Prostate Cancer Classification with 3T Multiparametric MRI: Comparison with PI- RADS-Based Classification." Diagnostics 11, no. 10 (2021): 1785
2. **Yongkai Liu**, Qi Miao, Haoxing Zheng, Guang Yang, Steven S.Raman, and Kyunghyun Sung; "Deep Learning Enables Prostate MRI Segmentation: A Large Cohort Evaluation with Inter-rater Variability Analysis." Frontiers in Oncology 11 (2021): 801876-801876.
3. **Yongkai Liu**, Guang YANG, Kyung Sung et al. "Automatic Prostate Zonal Segmentation Using Fully Convolutional Network With Feature Pyramid Attention." IEEE Access 7 (2019) 163626-163632.
4. **Yongkai Liu**, Guang YANG, Kyung Sung et al., "Exploring Uncertainty Measures in Bayesian Deep Attentive Neural Networks for Prostate Zonal Segmentation", IEEE Access, vol. 8, p151817-151828, 2020.

5. **Yongkai Liu**, Ran Yan, Guang Yang, Haoxing Zheng, Steven S. Raman, Kyunghyun Sung; "Evaluation of Spatial Attentive Deep Learning for Automatic Placental Segmentation on MRI", Journal of Magnetic Resonance Imaging (In Review)

Co-Author Peer Reviewed Publications

6. Yao Jin, Guang Yang, Ying Fang, Ruipeng Li , Xiaomei Xu, **Yongkai Liu**, Xiaobo Lai. "PBV-Net: An Automated Prostate MRI Data Segmentation Method ." Computers in Biology and Medicine. 2021 Jan;128:104160.
7. Zhang, Wenbo, Yang, Guang, Huang, He, Yang, Weiji, Xu, Xiaomei, **Yongkai Liu**, Lai, Xiaobo; "ME-Net: Multi-Encoder Net Framework for Brain Tumor Segmentation." International Journal of Imaging Systems and Technology. 2021 Dec; 31(10)
8. Haoxin Zheng, Qi Miao, **Yongkai Liu**, Steven S. Raman, Fabien Scalzo, Kyunghyun Sung; "Integrative Machine Learning Prediction of Prostate Biopsy Results from Negative Multiparametric MRI." Journal of Magnetic Resonance Imaging. 2022 Jan; 55(1):100-110.
9. Haoxin Zheng, Qi Miao, **Yongkai Liu**, Steven S. Raman, Fabien Scalzo, Kyunghyun Sung; "Multiparametric MRI-Based Radiomics Model to Predict Pelvic Lymph Node Invasion for Patients with Prostate Cancer." European Radiology (2022)
10. Chang Gao, Shu-Fu Shih, Holden, H. Wu, **Yongkai Liu**, Marcel Dominik Nickel, Thomas Vahle, Brian Dale, Victor Sai, Ely Felker, Qi Miao, Xiaodong Zhong, Peng Hu; "Undersampling artifact reduction for free-breathing 3D stack-of-radial MRI based on a deep adversarial learning network." Magnetic Resonance in Medicine. (In Review)

Chapter 1 Introduction

Magnetic resonance imaging (MRI) is a non-invasive imaging technology that produces 3-dimensional anatomical images. MRI provides excellent soft tissue contrast and does not have the ionization radiation that can cause damage to the human body, thus playing a significant role in disease diagnosis. MRI segmentation and classification play an essential role in disease assessment and detection, volume measurement, and biopsy. For example, since prostate cancer in different prostate zones exhibits different morphological and functional characteristics on multi-parametric MRI (mpMRI), prostate zonal segmentation is crucial in interpreting mpMRI for prostate cancer assessment. T2-weighted, and apparent diffusion coefficient MRI images are usually used for primary interpretation of lesions in the peripheral and transitional zones of the prostate[1]. Also, whole prostate gland (WPG) segmentation is critical to enable MRI-targeted transrectal ultrasound fusion (MRI-fusion) biopsy[2] and prostate volume measurement. Placenta MRI segmentation is the critical first step required toward accuracy in the abnormalities detection that can affect maternal and fetal health[3,4].

Manual segmentation and expert-based disease interpretation are highly dependent on reader experience and expertise, usually suffer from significant intra-and inter-reader variability and are also time-consuming and laborious[3,5,6]. Traditionally, various methods relying on feature [7–10] have been exploited to automate MRI segmentation and classification. In recent years, deep learning has achieved great success and demonstrated superior capabilities in various segmentation and classification tasks to the conventional methods due to its capacities for extracting higher-level representative features with the need for handcrafted feature extraction[11]. However, current deep learning models, particularly in prostate MRI segmentation and classification, may provide an insufficient representative power, and lack prediction uncertainty information and prior knowledge

such as cancer heterogeneity, and usually require large-scale and generalizability evaluation. Primarily with prostate MRI, this dissertation covers several advanced deep learning-based MRI segmentation and classification methods or applications to address the above issues. Specific issues of current deep learning-based segmentation and classification methods and potential improvements are detailed below.

Deep learning-based methods, such as U-Net[12] and its variants, have recently been developed to perform prostate zonal segmentation. However, high-level semantic and multiple-scaled information captured by U-Net may not be sufficient to describe the heterogeneous anatomic structures of the prostate and indiscernible borders between the prostate zones, resulting in inconsistent and sub-optimal performance. Deep learning incorporating a feature pyramid attention network can enhance capturing abilities of relevant features, especially multiple-scaled features[13]. The first aim of the dissertation is to develop and evaluate a deep learning method with feature pyramid attention for prostate zonal segmentation.

Outcomes from the current deep learning networks are deficient in acquiring uncertainties of the MRI segmentation. Segmentation uncertainty produced by the model can allow human experts to intervene to enhance current segmentation workflows by improving the highly uncertain cases. Also, the attention mechanism improves the deep learning-based segmentation by making the model focus more on semantic information related to the regions of interest. The second aim of the dissertation is to develop and evaluate an attentive Bayesian deep learning network for prostate zonal segmentation with uncertainty estimation. This network has a great potential to enhance the prostate zonal segmentation method of the first aim.

Heterogeneity is commonly regarded as a footprint and ecology of tumor evolution, which can be described by the image texture. Enriching texture into deep learning could enhance the

model's classification performance and potentially make the model able to be used in a small dataset due to the exploitation of tumor prior information such as heterogeneity. The third aim of the dissertation is to evaluate a texture-based deep learning model (Textured-DL) for differentiating between clinically significant prostate cancer (csPCa) and non-csPCa and compare the Textured-DL model with conventional deep learning and radiologist-based classification.

In addition, current deep learning-based segmentation methods were commonly evaluated by a relatively small sample size, limiting the ability to test the deep models in the clinical setting. The fourth aim of the dissertation is to evaluate the attentive Bayesian deep learning network, a deep learning-based segmentation model developed in the second aim, for whole prostate gland segmentation by using a large, continuous cohort of prostate 3T MRI scans, including 3,360 MRI scans.

Deep learning-based segmentation methods above were developed and evaluated in the prostate MRI. To assess the deep model's generalizability for other biomedical image applications, such as the placenta segmentation, the fifth aim of this dissertation is to evaluate a deep learning method, which was adapted from the attentive Bayesian deep learning network, a previously developed in the second aim, for the placental segmentation in longitudinal MRI.

This dissertation aims to develop and evaluate advanced deep learning methods for MRI segmentation and classification, primarily in the prostate MRI. The specific aims are 1) to develop and evaluate a deep learning with feature pyramid attention for prostate zonal segmentation; 2) to develop and evaluate the attentive Bayesian deep learning network for prostate zonal segmentation with uncertainty estimation; 3) to evaluate Textured-DL for differentiating between csPCa and non-csPCa and compare Textured-DL's performance with conventional deep learning and radiologist-based classification; 4) to evaluate the attentive Bayesian deep learning network, a deep

learning-based segmentation model developed in the second aim, for whole prostate gland segmentation by using a large, continuous cohort of prostate 3T MRI scans, including 3,360 MRI scans; 5) to evaluate the generalizability of a deep learning method, which was adapted from the attentive Bayesian deep learning network that was previously developed method in the second aim.

1.1 Outline

The structure of this dissertation goes as follows:

Chapter 2: Background

Chapter 2 makes the basic introduction of medical image segmentation, deep learning, Bayesian deep learning, texture-based deep learning methods, magnetic resonance imaging, and evaluation metrics for segmentation and classification tasks.

Chapter 3: Automatic Prostate Zonal Segmentation Using Fully Convolutional Network with Feature Pyramid Attention

Chapter 3 describes the work of a deep learning-based method with feature pyramid attention for automated prostate zonal segmentation in detail. This chapter developed and evaluated a deep learning-based method equipped with the feature pyramid attention mechanism for prostate zonal segmentation. The model was assessed separately on two testing datasets to investigate the model's expansibility to different datasets. Performance discrepancy across different sections of the prostate was also investigated in the chapter. Finally, this chapter compared the model's performance with the inter-reader consistency that incorporates two independent expert-based manual segmentations.

Chapter 4: Exploring Uncertainty Measures in Bayesian Deep Attentive Neural Networks for Prostate Zonal Segmentation

Chapter 4 developed and evaluated an attentive Bayesian deep learning method for automated prostate zonal segmentation with uncertainty estimation. The proposed method was superior to the method developed in the Chapter 3. Similar to Chapter 3, performance discrepancy of the proposed method in two separate testing datasets was investigated in this chapter. This chapter also examined the importance of the attention mechanism in the proposed model and calculated the average prostate zonal segmentation uncertainty maps at apex, middle and base. Based on uncertainty maps, primary patterns of prostate zonal segmentation uncertainty at the three prostate sections were summarized.

Chapter 5: Texture-based Deep Learning for Prostate Cancer Classification

Chapter 5 describes the work of Textured-DL for prostate cancer classification. In this chapter, Textured-DL was developed and evaluated for automated PCa classification of suspicious prostate lesions on a 3T mpMRI dataset with whole-mount histopathology (WMHP) correlation. After a lesion was detected and contoured as part of the clinical interpretation, Textured-DL was developed to further improve the classification of PCa for any positive MRI findings. The model performance was evaluated by an independent testing set and compared with the conventional deep learning and radiologist-based classification. Performance difference of the Textured-DL on lesions with different locations and types (solitary and multi-focal) were also investigated.

Chapter 6: Deep Learning Enables Prostate MRI Segmentation: A Large Cohort Evaluation with Inter-rater Variability Analysis

Chapter 6 describes the work of the large cohort evaluation of a deep learning method for automated whole prostate gland segmentation. The deep learning method is modified from the one in Chapter 4 by adding the coronal-view segmentation assistance. The large cohort evaluation includes a qualitative, a quantitative assessment, and a volume measurement evaluation.

Chapter 7: Evaluation of Spatial Attentive Deep Learning for Automated Placental Segmentation on Longitudinal MRI

This chapter describes a deep learning-based segmentation method, spatial attention deep learning method (SADL), for fully automated placental segmentation on longitudinal MRI. Deep learning method from Chapter 4 forms the basic structure of SADL. Different from the method in Chapter 4, SADL used the criss-cross spatial attention instead of the conventional spatial attention, which could relieve the issue of large GPU memory. SADL-based automated volume measurement is also assessed in this Chapter.

Chapter 8: Summary and Future Work

This chapter summarizes the five advanced deep learning-based methods or applications for MRI segmentation and classification in this dissertation. Also, potential directions for future research are briefly discussed.

Chapter 2 Background

This Chapter provides the basic introduction of medical image segmentation, deep learning, Bayesian deep learning, texture-based deep learning methods (Textured-DL), magnetic resonance imaging (MRI), and evaluation metrics for segmentation and classification tasks.

2.1 Medical Image Segmentation

Medical image segmentation is the task of contouring the regions of interest like organs or tumors in a digital medical image such as CT or MRI. Medical image segmentation plays a critical role in assisting various medical image-related applications, such as volume measurement, diagnosis, and treatment. For example, prostate zonal segmentation is an essential step to interpret multi-parametric MRI for prostate cancer assessment[14,15]. Some prostate computer-aided diagnosis (CADx) systems[16] require the segmentation of the whole prostate gland before feeding the medical image data to the systems. Segmentation enables the organ volume estimation, which is important for the disease treatment. For example, liver volume is one of the key considerations when assessing the suitability for surgeries related to liver[17]. In the treatment, volume measurement also facilitates the treatment response assessment. Furthermore, segmentation has been shown to enhance image-guided radiotherapy[18]. By segmenting the tumor, the beam of radiation can be delivered to the correct target area[18].

2.1.1 Image Segmentation Methods

Manual segmentation suffers from great inter-rater variability and is typically time-consuming, laborious, and thus, inadequate for large-scale applications enabled by segmentation[19]. Various methods have been proposed to perform semi-automatic or automatic segmentation.

The region-based image segmentation method[20] is a simple method that groups pixels with similar attributes into unique regions based on the similarities between adjacent pixels. Thresholding and region growing are the two most popular region-based image segmentation methods. The thresholding method assumes that images consist of regions with different intensity values. Threshold values are usually detected in the peaks and valleys from the histogram, which can be used to segment the images into different regions. Since the thresholding method does not consider spatial information, it will be sensitive to noise and intensity in the homogeneous region. The region-growing method[21] requires some seed points to be initialized and compares the adjacent pixels with the initial seed points. The pixels will be part of the same region with the seed points if a similarity criterion is met.

The clustering method is another common algorithm for image segmentation, classifying data and patterns into categories. Clustering method is an unsupervised algorithm, which does not require labelled data. It assumes that pixels in the same region have a higher similarity than those in different regions. The K-means algorithm[22] is one of most common clustering methods, which uses an iterative approach to segment an image into k regions.

Edge-based Segmentation relies on edges detection in an image. Edges are the discontinuities of intensity or texture in an image. By acquiring knowledge from the image gradients, various edge detection operators, such as Sobel edge operator and Robert edge operator, can help locate the edges. Segmentation achieved by edge segmentation is usually an intermediate segmentation result. Other segmentation methods, such as region-based methods, can be used to further improve it to obtain a complete segmentation.

Partial differential equation-based segmentation, such as level set method[23], is a popular image segmentation method based on the curve propagation. The basic idea to evolve an initial

curve to the actual contour when a cost function, which represents a task to be solved, reaches the lowest potential.

The methods mentioned above rely on the human efforts of engineering. Human engineering, which is usually lack of domain knowledge, might use less knowledge than human to achieve the segmentation. Trainable segmentation methods, such as deep neural network, can help address such an issue. Specifically, these models learn the domain knowledge directly from the human segmentation via training. U-Net is one of most popular trainable segmentation methods for medical image segmentation. A more detailed introduction of U-Net is given in section 2.2.2.

2.2 Deep Learning

Deep learning has achieved great success in natural and medical image tasks such as classification, detection, and segmentation in recent years. By gradually extracting higher-level features with convolutional layers, deep learning learns representation from raw data[24]. Deep learning relieves the burden of human expert-enabled feature engineering, making it easy to exploit the huge amount of accessible data. Convolutional neural network is one of the most popular deep neural networks.

2.2.1 Convolutional Neural Network

The convolutional neural network (CNN) is one specialized type of neural networks designed to process image data, primarily incorporating three layers: convolutional layer, pooling layer, and fully connected layer. Convolutional layers are the primary building blocks in convolutional neural networks. The convolution operation is the application of a filter to an image, which abstracts an image to a feature map. A convolutional layer applies a convolution operation to the input image and passes the feature maps to the next layer. The purpose of pooling layers is

to reduce the dimension of the feature maps, resulting in reducing the number of parameters to learn and the amount of computation performed in the network. Fully connected layers usually form the last few layers of the CNN for the classification tasks. Each neuron in a fully connected layer connects all the neurons coming from the previous layer.

2.2.2 Deep Learning-based segmentation method

U-Net is a deep learning-based model that gained the most popularity for biomedical image segmentation[12], which comprises a contracting path as an encoder to extract high-level features and an expansive path as a decoder to recover spatial resolution compromised in the encoder. Encoder, consisting of multiple convolutions, each followed by a rectified linear unit (ReLU) and a max-pooling operation, is used to extract the high semantic information. After the encoder, the spatial resolution decreased, and semantic information increased. Decoder recovers spatial resolution by a sequence of up-convolutions and concatenations with high-resolution features from the encoder.

2.3 Bayesian Deep Learning

2.3.1 Bayesian probability theory

Probability refers to the chance that a given event will happen. In Bayesian probability theory, probability of an event is calculated by first specifying a prior probability and then updating the prior probability based on Bayes' theorem after obtaining new data. The updated prior probability is called posterior probability. The whole process can be described by the formula:

$$P(M \mid N) = \frac{P(N \mid M)\, P(M)}{P(N)} \qquad (2\text{-}1)$$

M represents an event. $P(M)$ is the prior probability. N represents the new data or new event. The conditional probability of M given new data N is $P(M \mid N)$, which is also called posterior probability. $P(N \mid M)$ is the probability of observing new data N when the hypothesis M is fixed. Prior probability $P(M)$ will be updated to the posterior probability $P(M \mid N)$ based on the Bayes' formula after taking the new data into the consideration.

2.3.2 Bayesian deep learning

Outcomes from current deep learning are typically deterministic; there is a lack of knowledge on the confidence of the model parameters[25]. Bayesian deep learning is a probabilistic model that applies Bayesian inference to a deep learning network structure, which can provide the uncertainty measurement of the model prediction.

Model uncertainty estimation is commonly motivated and mathematically supported by Bayesian statistics. For example, Bayesian neural network (BNN) is one of the early examples that applies Bayesian inference to the neural network[26]. However, issues such as intractable integrals from Bayesian inference hinder the development of Bayesian-based models. Variational distributions, e.g., Gaussian distribution, have been studied to approximate the posterior in the Bayesian inference by minimizing the Kullback-Leibler (KL) divergence between the actual posterior and the variational distribution[27]. Nevertheless, number of parameters within the model will significantly increase by use of the variational distribution, which will make model computationally expensive. By randomly shutting down weights in the training, dropout usually serves as one of the efficacious regularization techniques to avoid overfitting in deep learning. Recently, studies have shown that performing dropout on the weights of a deep learning model is equivalent to placing the variational distribution - Bernoulli distribution over the weights[28]. Also, the effect of minimizing the cross-entropy loss is the same as the minimization of the KL-

divergence. Therefore, training with dropout allows the approximation of posterior. These dropouts are also required to be kept active during the testing. In the testing, by performing stochastic forward passes through a trained deep learning network using dropout, Monte Carlo samples were taken from the posterior distribution. Compared to the use of variational Bayesian inference, dropout-based approaches[27] can help deep learning to produce uncertainty estimations in a cheap way by avoiding the unnecessary computation cost.

2.4 Texture-based Deep Learning

2.4.1 Image Texture

Image texture provides the information related to the spatial arrangement of intensities in the image. Due to the heterogenous nature of the cancer, texture pattern usually correlates with the cancer risk[29].

2.4.2 Gray-Level Co-occurrence Matrix

As a most used approach to statistically examine the image texture, the gray-level co-occurrence matrix (GLCM)[30] is a matrix that calculates the frequency of pixel/voxel pairs with different spatial orientations and specific gray-level values, which is formulated as:

$$C_{d,\theta}(\mathrm{x},\mathrm{y}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\begin{cases} 1 & if\ I(i,j) = x\ and\ I(i+di, j+dj) = y \\ 0 & Otherwise \end{cases} \qquad (2\text{-}2)$$

, where (di, dj) is a displacement between the point (i, j) and another point along the direction θ, I is the image data, (i, j) is a pixel location in the image I, and $I(i, j)$ is the pixel value at (i, j). In general, GLCM of a flat image will be a diagonal matrix. The more pixel intensities variate, the larger the off-diagonal values in the GLCM[31]. Since GLCM is usually sparse and large, Haralick

texture features, such as correlation, contrast, homogeneity, and energy, computed from the GLCM, are commonly the descriptors to represent the image texture[32]. A classification algorithm, e.g., support vector machine (SVM) or random forest, is used to classify the image using the Haralick texture features.

2.4.3 Texture-based Deep Learning

In the workflow of texture-based deep learning, a deep learning model learns on GLCM directly to conduct the image classification without feature engineering. Due to the characteristic such as parameter sharing scheme, deep learning such as CNN is very apt to and fits the image processing. Using the deep learning to learn on raw images will suffer from the resizing operation of the input image. GLCM will have a fixed size when the image is scaled to the same gray level. Therefore, CNN model demonstrates a great potential to learn the more useful and informative features from the GLCM than the hand-crafted Haralick features[30] and raw image data.

2.5 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a non-invasive medical imaging technique that generates three-dimensional detailed anatomical images using a powerful magnetic field, magnetic field gradients, and radio waves. Unlike CT and PET, MRI does not have the ionizing radiation that can cause damage to the human body. MRI provides superior imaging contrast of soft tissue such as the brain or prostate compared to other imaging modalities.

MRI is commonly formed by the signals produced by using magnetization properties of the hydrogen protons, which are abundant in water and fat of the human body. In everyday situations, hydrogen protons in the body spin around the randomly oriented axes, while when the body is in a strong magnetic field, the hydrogen protons will align with the magnetic field to create

a net magnetic moment parallel to the magnetic field. Hydrogen protons will absorb the radiofrequency (RF) energy when the radiofrequency pulse is temporarily applied to the patient, tilting the net magnetic moment away from the magnetic field. Once the RF pulse disappears, the net magnetic moment will realign to the magnetic field and return to equilibrium, and hydrogen protons will lose energy by emitting the RF signal. During the process, longitudinal (T1) and transverse (T2) relaxations will simultaneously occur. Transverse relaxation refers to the decaying process of transverse components of magnetization. Longitudinal relaxation is the process in which net magnetization return to the initial value. Characteristics of various tissues in the body can be reflected by longitudinal and transverse magnetizations; therefore, the differentiation between tissues can be based on these relaxations. Specifically, an MRI sequence can be designed to be T1-weighted, T2-weighted, or proton-density weighted. A magnetic field gradient can achieve spatial encoding by varying the frequency of hydrogen protons as a function of position along the gradient direction. Finally, the Fourier transform transforms the encoded image into the spatial image.

2.6 Evaluation Metric

2.6.1 Evaluation Metrics for Segmentation Task

Segmentation tasks can be evaluated both quantitatively and qualitatively. Metrics for quantitative evaluation include dice similarity coefficient (DSC) and average Hausdorff distance (HD) in this dissertation.

DSC is an index ranging from 0 to 1 to describe the overlapping ratio between the automated and manual segmentation. 0 and 1 indicate no and complete spatial overlapping between the two segmentations. DSC is formulated as:

$$DSC = \frac{2|X \cap Y|}{|X|+|Y|} \quad (2\text{-}3)$$

where X and Y are the automated and manual 3D segmentation.

Patient-wise HD is an another commonly used metric to assess the performance of automated medical segmentation methods by measuring the longest distance between the automated and manual segmentation, which is formulated as:

$$\mathrm{HD}(\mathrm{X},\mathrm{Y}) = \max(\mathrm{h}(\mathrm{X},\mathrm{Y}),\mathrm{h}(\mathrm{Y},\mathrm{X})) \quad (2\text{-}4)$$

, where $\mathrm{h}(\mathrm{X},\mathrm{Y})$ is the directed HD, which is given by $\mathrm{h}(\mathrm{X},\mathrm{Y}) = \max_{\mathrm{x}\in\mathrm{X}} \min_{\mathrm{y}\in\mathrm{Y}} \|\mathrm{x}-\mathrm{y}\|$, X and Y are the point sets on the automated and manual 3D segmentation.

In the qualitative evaluation, visual grading was adopted to evaluate the segmentation performance. Specifically, human experts assign a visual grade based on the level to which the segmentation can be accepted in the clinic setting, to score the segmentation performance.

2.6.2 Evaluation Metrics for Classification Task

Evaluation metrics for classification task include area under the receiver operating characteristic curve (ROC) (AUC), sensitivity, and specificity. ROC is a plot to show the performance of a classification model at each decision threshold. Each point on the ROC curve corresponds to a pair of sensitivity (true positive rate) and specificity (false positive rate) at a decision threshold. AUC represents the aggregate measure of the classification performance across all the decision thresholds.

True positive rate (TPR) also called sensitivity is the percentage of positive identifications that is correct, formulated as:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \qquad (2\text{-}5)$$

, where TP and FN represent the true positive and false negative.

True negative rate (TNR) also called specificity is the percentage of negative identifications that is correct, formulated as:

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \qquad (2\text{-}6)$$

, where TN and FP represent the true negative and false positive.

Chapter 3: Automatic Prostate Zonal Segmentation Using Fully Convolutional Network with Feature Pyramid Attention

In this Chapter, a deep learning with feature pyramid attention is developed and evaluated for the automated prostate zonal segmentation. The proposed deep learning method is compared with the previous deep learning-based prostate zonal segmentation, U-Net. Also, proposed method is also compared with the inter-reader consistency between the two independent expert-based manual segmentations.

3.1 Introduction

Prostate cancer (PCa) is the most common solid noncutaneous cancer in American men[33]. Multiparametric MRI (mpMRI), including T2, diffusion weighted imaging (DWI) and T1 dynamic contrast enhanced imaging (DCE) has shown promising results for the detection and staging for clinically significant PCa (csPCa)[34,35]. Previous studies have reported that PCa in transition and peripheral zones exhibit different morphological and functional characteristics on mpMRI. The Prostate Imaging Reporting and Data System version 2.1 (PI-RADSv2.1), an expert guideline for performance and interpretation of mpMRI for PCa detection[14,15], T2 and DWI images are used for primary interpretation of lesions in the PZ and TZ respectively for assigning a PI-RADS score to lesions detected on mpMRI[1]. A robust method for reproducible, automatic segmentation of

prostate zones (ASPZ) may enable the consistent assignment of mpMRI lesion location since manual segmentation of prostate zones is a dependent time-consuming process, dependent on reader experience and expertise. A robust ASPZ may also help relieve clinician's cognitive workload[36].

Atlas based methods were previously proposed to segment the prostate zones[37]. Deep learning (DL) based methods, such as U-Net[12] and its variants[1,38–41], have recently been developed to perform prostate AS. U-Net, an architecture based on fully convolutional networks (CNN), contains encoder and decoder sub-networks, where the encoder module is used to capture the higher semantic information, and the decoder module recovers spatial information. U-Net can classify pixels of the two zones and effectively localize and segment TZ and PZ. However, semantic information captured by U-Net may not be sufficient to describe the heterogeneous anatomic structures of the prostate and indiscernible borders between TZ and PZ, resulting in inconsistent and sub-optimal ASPZ performance.

In this study, we propose a new DL based method for automatic segmentation of prostate zones by developing a fully CNN with a novel feature pyramid attention mechanism. In particular, the proposed CNN consisted of three sub-networks, comprised of an improved deep residual network (based on the ResNet50)[42], a pyramid feature network with attention[13], and a decoder. We incorporated the ResNet50 to cope with heterogeneous prostate anatomy with high level semantic features and the pyramid network with attention is designed to capture information at multiple scales. The proposed DL model was evaluated using both internal testing and external testing datasets on axial mpMRI slices. In addition, we compared the proposed method with inter-reader consistency using two independent expert based manual segmentations.

3.2 Materials and Methods

3.2.1 MRI Datasets

With approval from the institutional review board (IRB), this study was carried out in compliance with the United States Health Insurance Portability and Accountability Act (HIPAA) of 1996. The MRI datasets were collected from two centers: 1) The Cancer Imaging Archive (TCIA) for SPIE-AAPM-NCI PROSTATEX (PROSTATEX) challenge[7] for development and internal testing of the model (n=250 and 63) and 2) a U.S. tertiary academic medical center with a highly curated mpMRI dataset with whole mount histopathology (WMHP) correlation for external testing of the model (n=46; age 45 to 73 years and weight 68 to 113 kg). Axial T2 turbo spin-echo (TSE) slices (Table 3-1) were used for segmentation. For the PROSTATEX data, both TZ and PZ were segmented in OsiriX (Pixmeo SARL, Bernex, Switzerland) by two MRI research fellows, where the contours were later cross-checked by both genitourinary (GU) radiologists (10-15 years of post-fellowship experience interpreting over 1,000 prostate mpMRI) and clinical research fellows. For the single institutional data, the pre-operative mpMRI scans performed between October 2017 and December 2018 on one of the three 3T MRI scanners (Skyra (n=38) on, Prisma (n=1), and Vida (n=7); (Siemens Healthineers, Erlangen, Germany)). Two clinical GU research fellows, supervised by expert GU radiologists, independently contoured TZ and PZ in a blinded fashion.

Table 3-1 Detailed T2w TSE Protocols Used for Two MRI Datasets

Datasets	Internal Testing dataset (ITD)	External Testing dataset (ETD)
Spatial Resolution	0.5x0.5x3.6mm^3	0.65x0.65x3.6mm^3

Flip angle	160°	160°
Matrix Size	380x380	320x320
Field-of-View	190x190 mm^2	208x208 mm^2
Repetition Time/Echo Time	5660 ms / 104 ms	4000 ms / 109 ms

3.2.2 Proposed Deep Learning Model for Automatic Prostate Segmentation

The structure of the proposed fully convolutional network is shown in Figure 3-1. The network consists of three separate sub-networks, including the improved ResNet50 for encoding of rich semantic information from original images, a feature pyramid attention network to help capture the information at multiple scales, and the naïve decoder network to recover the spatial information. The three sub-networks are connected to be an end-to-end prostate zonal segmentation pipeline.

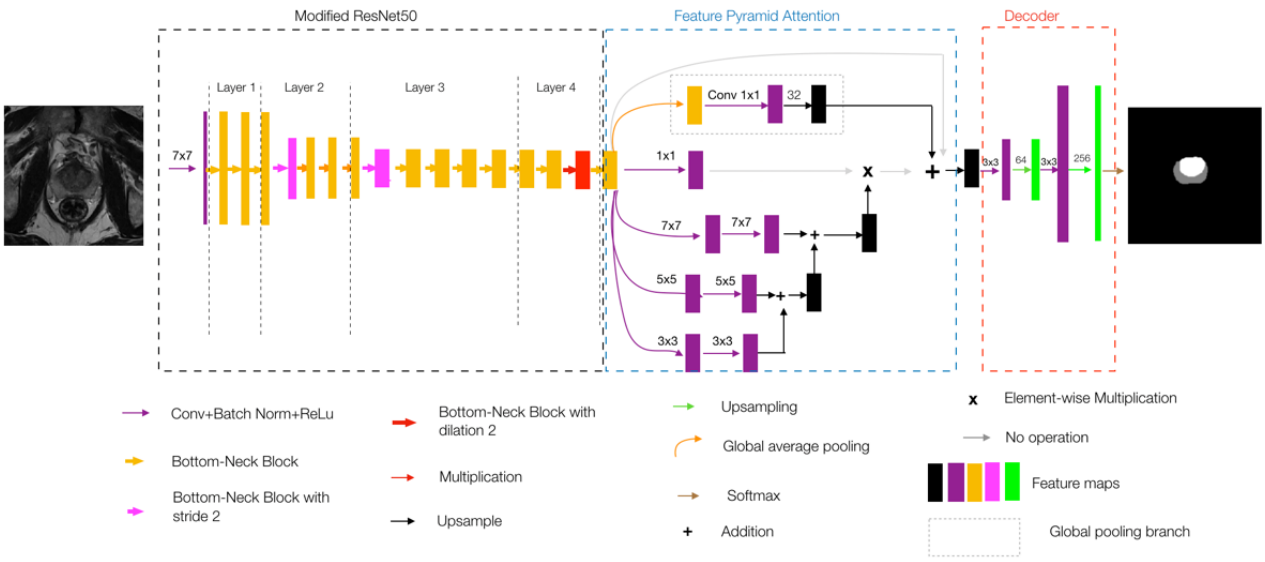

Figure 3-1 An overall structure of the proposed algorithm, where the input is a 2D slice of T2w MRI, and output is a mask showing the result of PZ and TZ segmentation (white – TZ and gray – PZ). The algorithm consists of three sub-networks - improved ResNet50 (a), feature pyramid attention (b), and decoder (c).

ResNet50 utilizes skip connections to avoid vanishing gradients problems so that more convolutional layers can be added to the network. We improved the ResNet50 by removing the initial max pooling layer and using the regular block instead of the bottleneck block at stride 1 as the first block in the 4th layer, as shown in Figure 3-1 (a). The dilated bottleneck block was employed as the second block in the 4th layer to remain the size of the receptive field. This can minimize any potential loss to the spatial information and alleviate the burden of the decoder.

Feature pyramid attention was added after modified ResNet50 for better sensing fine details at different scales (Figure 3-1 (b)). The 3×3, 5×5, 7×7 convolutions in the pyramid structure were used to extract features from different scales. The features from different scales were integrated progressively for more precise incorporation of adjacent scales of context features and then were multiplied by the features from the improved ResNet50 after a 1×1 convolution operation for the global prior attention. The output features will be then added with features from both the global pooling branch and the modified ResNet50. The decoder network consisted of two convolutions and two upsampling layers to recover the image dimensions to the original size (Figure 3-1 (c)). The final output was fed into a multi-class soft-max classifier for simultaneous segmentation of TZ and PZ.

We used cross entropy (CE) as the loss function for the proposed algorithm. For each given pixel, the cross entropy was defined as,

$$CE = \frac{1}{3}\sum_{i=0}^{3} -y_i \log(p_i) - (1 - y_i)\log(1 - p_i) \qquad (3\text{-}1)$$

where $y_i \in \{0,1\}$ is the ground-truth binary indicator, corresponding to the 3-channel predicted probability vector $p_i \in [0,1]$. All the training and evaluation were performed on a desktop computer with a 64-Linux system with Titan Xp GPU with 12 GB GDDR5 RAM based on PyTorch. Learning rate was initially set to 2.5e-3, with momentum 0.9 and weight decay 0.0001. The model was trained for 100 epochs with batch size 48 and stochastic gradient descent. Since prostate areas are always in the middle, a central region ($93mm \times 93\ mm$) was automatically cropped from original images before segmentation. Data augmentation methods were applied to increase the training data size, including flipped horizontally, rotated randomly between [−5°, 5°] and elastic transformations.

3.2.3 Model Development and Testing

A total of 250 patients' MRI from PROSTATEX were used for model development. Within the development dataset, 5-fold cross validation was adopted for model hyperparameter tuning. For internal testing (internal testing dataset (ITD)) the remaining 63 MRI datasets from PROSTATEX were used. For external testing (external testing dataset (ETD)) 47 MRI datasets from the large, U.S. tertiary academic medical center were used. For evaluation of segmentation, the Dice Similarity Coefficient (DSC) was used, formulated as:

$$DSC = \frac{2|X \cap Y|}{|X|+|Y|} \qquad (3\text{-}2)$$

where X is the predicted 3D zonal segmentation and Y is the ground-truth of 3D zonal contours on the slices. From superior to inferior, prostate MRI slices were categorized into three levels, composed of base-end (includes mostly TZ), middle (includes mostly both TZ and PZ), and apex-

end (includes mostly PZ), as shown in Figure 3-2. Both the prostate base-end and apex-end slices were identified when manual segmentation was performed, typically including one or two end slices of the prostate gland with only one prostate zone. A representative example of different prostate MRI slices is shown in Figure 3-3. DSCs were calculated considering different 3D zonal segmentation results, such as all slices (includes false positives), prostate slices (excludes false positives), base-end, middle, and apex-end slices. To assess the inter-reader consistency, we computed DSCs between two contours of TZ and PZ performed by two independent experts in a blinded fashion. The corresponding imaging slices were used for the inter-reader agreement assessment.

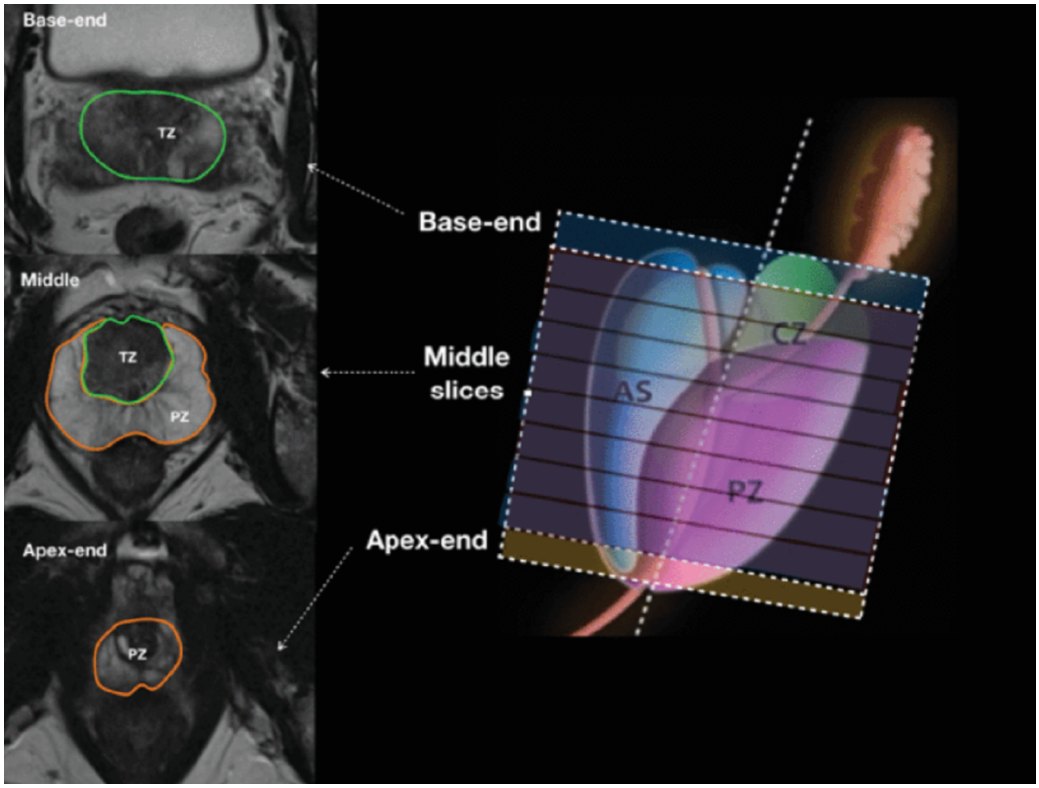

Figure 3-2 Representative examples of slices of prostate MRI. In left side, base-end slice (Only TZ exists), middle slice (both PZ and TZ exist) and apex-end slice (only PZ exists) are shown from top to bottom. The regions are encircled by green (TZ) and orange (PZ) boundaries.

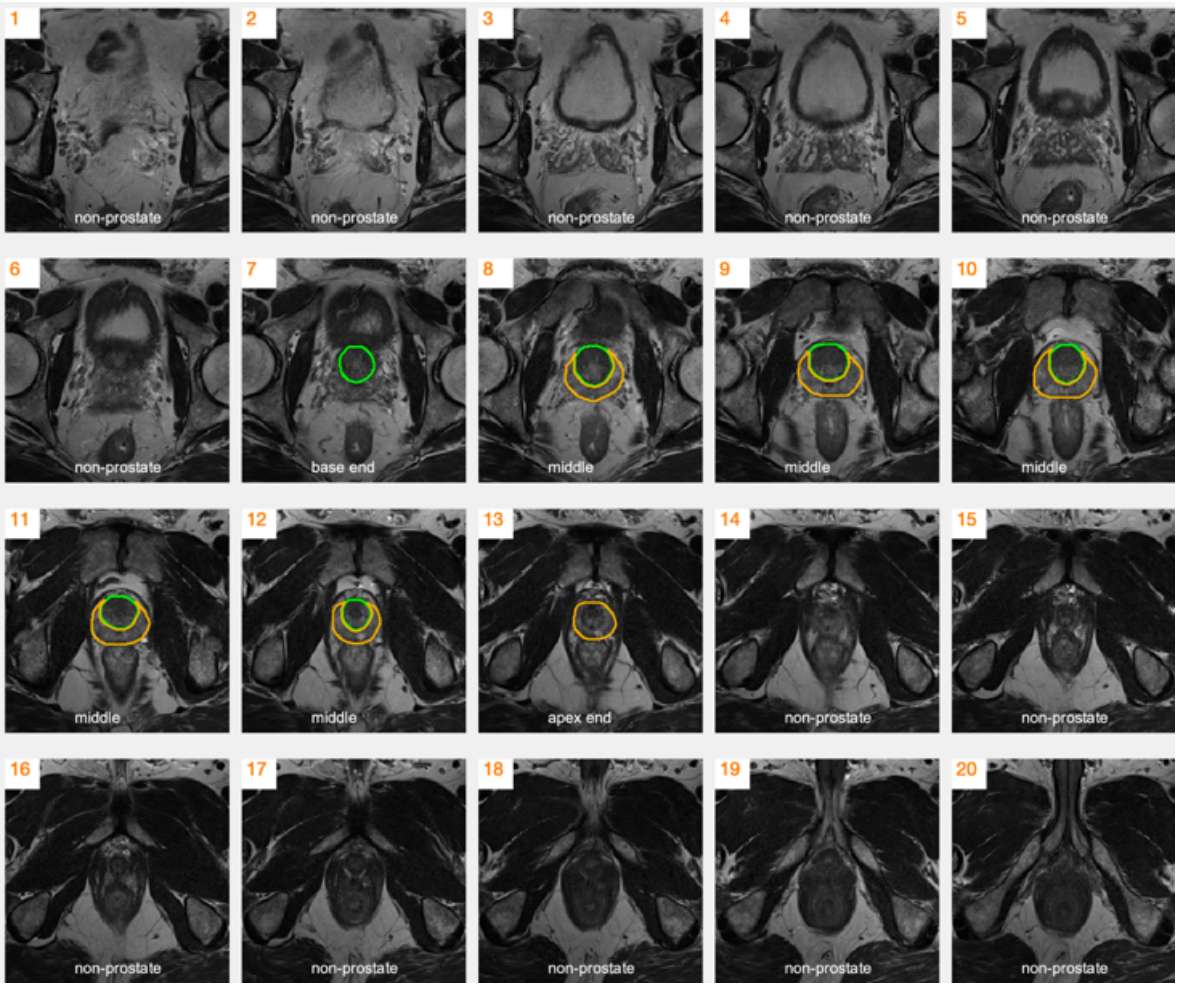

Figure 3-3 MRI slices from superior to inferior (slice 1 – 20). An example of non-prostate (slice 1-6, slice 14-20), base-end (slice 7), middle (slice 8-12) and apex-end (slice 13) slices is shown. Regions encircled by orange, green boundaries are PZ and TZ, respectively.

3.2.4 Statistical Analysis

Mean and standard deviation (SD) were used to summarize the distribution of DSCs. We performed the following three comparisons. First, the performance of the proposed method was compared to the baseline method – U-Net on the ITD dataset by using Wilcoxon rank-sum test. Second, the performance of the proposed method on the ITD was compared to the ETD by also using Wilcoxon rank-sum test. Third, the performance of proposed method was compared with the

inter-reader agreement (Expert 1 vs. Expert 2) by using the Wilcoxon signed-rank test. P values less than 0.05 were considered statistically significant.

3.3 Result

3.3.1 Model Testing Using Internal Testing Dataset (ITD) and External Testing Dataset (ETD)

Two representative examples of automatic prostate zonal segmentation on ITD and ETD by our proposed method and U-Net are shown in Figure 3-4.

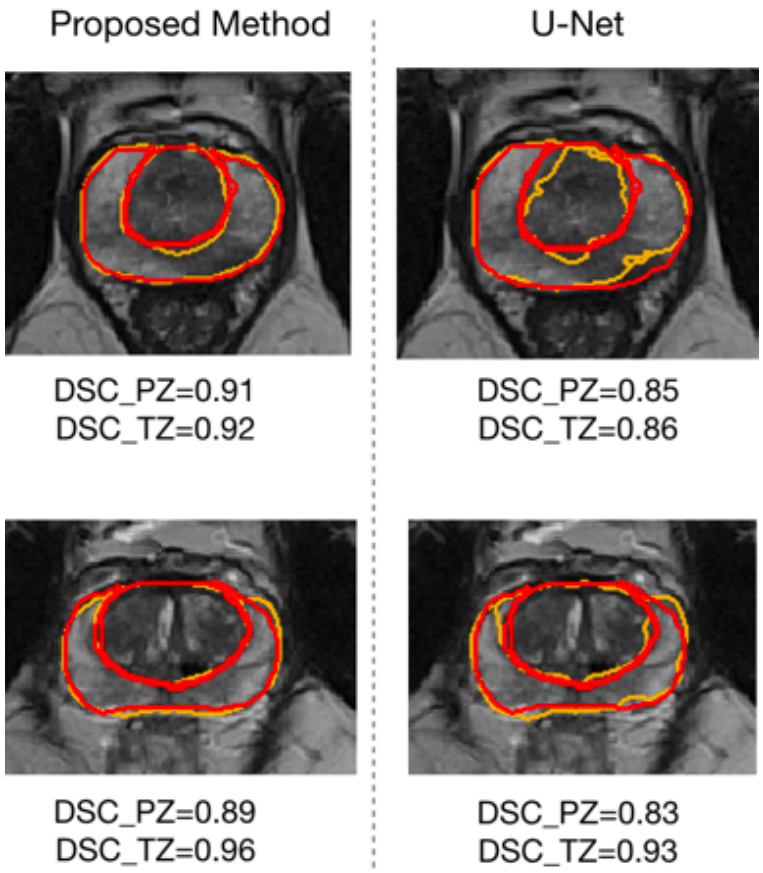

Figure 3-4 Representative examples of the automatic segmentation by the proposed method (orange lines) and U-Net in comparison with manual segmentation (red lines). DSCs are shown below the figures.

Table 3-2 includes mean and standard deviation of DSCs for PZ and TZ. Our proposed method achieved the mean DSC of 0.74 and 0.86 for PZ and TZ on ITD, mean DSC of 0.74 and 0.79 for PZ and TZ on ETD, which are all significantly larger than U-Net's results.

Table 3-2 Performance of the Proposed Algorithm on both Internal (ITD) and External Testing Dataset (ETD). P Values are the Comparisons Between the Proposed Model's Performance and the U-Net on Internal Testing Dataset

Datasets	ITD		ETD	
	PZ	TZ	PZ	TZ
U-Net	0.69 ± 0.10	0.83 ± 0.09	0.67 ± 0.09	0.76 ± 0.10
Proposed Method	0.74 ± 0.08 P<0.05	0.86 ± 0.07 P<0.05	0.74 ± 0.07 P<0.05	0.79 ± 0.12 P<0.05

Table 3-3 shows the performance of prostate zonal segmentation by the proposed model with Max-Pool and without Max-Pool on the ITD. After adding the Max-Pool in the ResNet50, mean DSCs for PZ and TZ are 0.72 and 0.84, which are smaller than the DSCs of proposed method (No Max-Pool in the ResNet50). This proves Max-Pool compromises the segmentation performance of prostate zones.

Table 3-3 Performance Comparison Between the Proposed Model with Max-Pool and Without Max-Pool Under ITD. In Our Proposed Method, the Max-Pool was Removed in ResNet50.

	DSC	
	PZ	TZ
Proposed Method	0.74 ± 0.08	0.86 ± 0.07

Add the Max-Pool in the proposed method	0.72 ± 0.08	0.84 ± 0.07

3.3.2 Comparison of Model Performance on Internal Testing Dataset (ITD) and External Testing Dataset (ETD)

In Table 3-4, we show the performance of the proposed algorithm in the ITD and ETD. There was no significant difference of model's DSC between the ITD and the ETD for PZ. However, for TZ, there was a small difference between model's DSC on the ITD and the ETD. The DSC differences of proposed algorithm on the ITD compared to the validation dataset were 8%.

Table 3-4 Performance of the Proposed Algorithm on ITD and the ETD. P Values of Model's Performance on ITD Relative to ETD are Given and Were Obtained by Using Wilcoxon Rank-Sum Test

Datasets	PZ	TZ
ITD	0.74 ± 0.08 P=0.14	0.86 ± 0.07 P<0.05
ETD	0.74 ± 0.07	0.79 ± 0.12

3.3.3 Comparison Between Proposed Model and Experts Under ETD

Examples of automatic segmentation results of slices by our proposed method, Expert 1 and Expert 2, at the base-end, middle, and apex-end on ETD are shown in Figure 3-5.

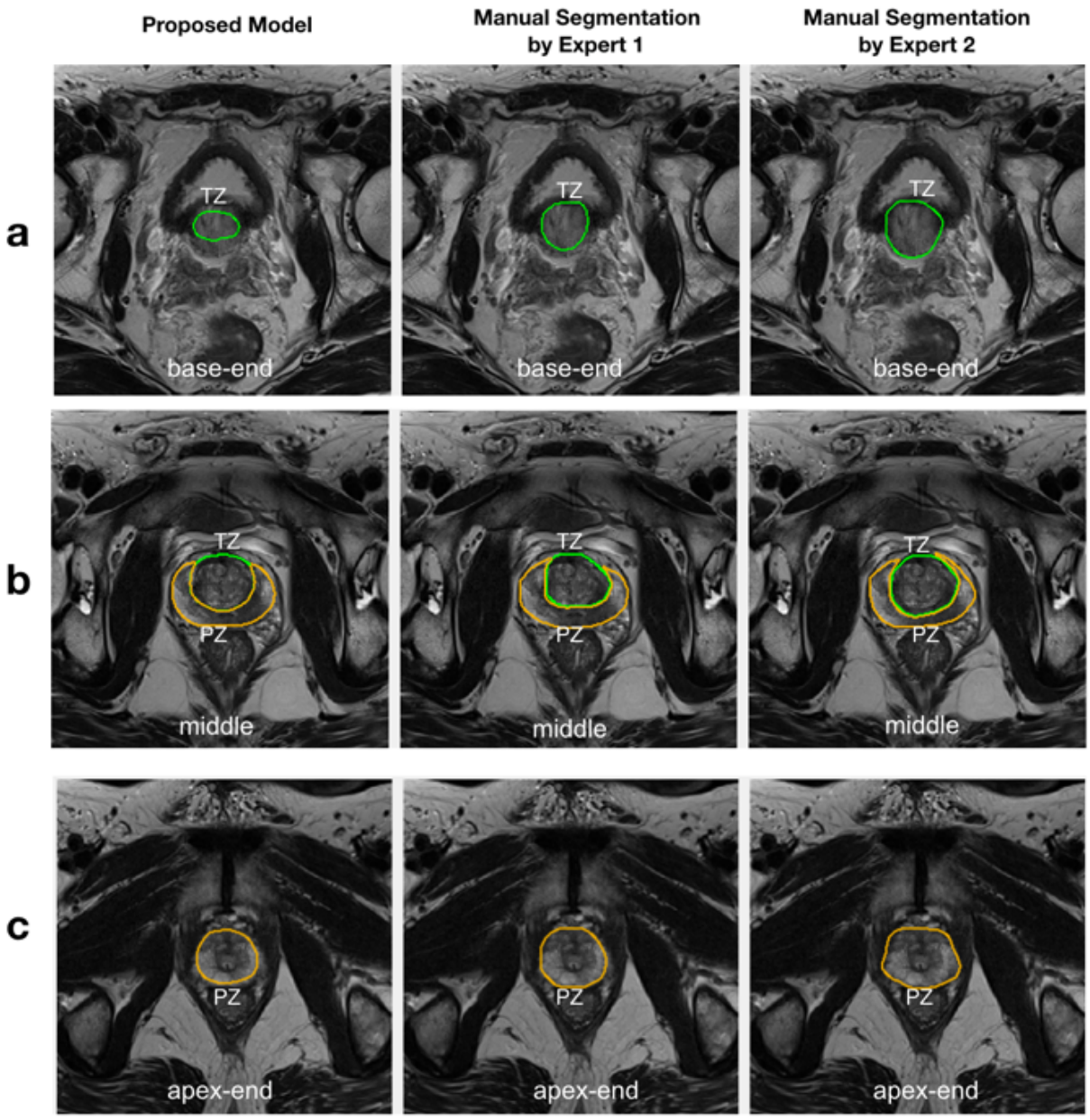

Figure 3-5 Representative examples of the automatic segmentation for testing, in comparison with manual segmentations by Expert 1 and 2. TZ is colored as green, and PZ is colored as orange. From superior to inferior, base-end (a), middle (b), and apex-end (c) slices are shown with segmentations of the prostate zones.

In Table 3-4 the DSCs of the proposed algorithm on different types of slices in the ETD are shown with inter-reader agreement between Expert 1 and Expert 2. The proposed model's

DSCs for both PZ and TZ are significantly larger than the inter-reader consistency in all slices, middle slices, apex-end slices, and base-end slices when taking Expert 1's annotations as the ground-truth.

3.4 Discussion

In this study we proposed and validated a novel fully convolutional network-based model with feature pyramid attention for the automatic segmentation of the two prostate zones. The proposed model performed consistently on both the ITD and ETD. We observed slight differences between ITD and ETD, particularly in segmenting TZ. We believe this can be potentially due to 1) differences in the imaging sequences, such as in-plane resolution and T2 contrast, 2) discrepancies in the zonal annotations since different experts independently segmented the prostate zones for ITD and ETD. We also found that the manual PZ segmentation was less consistent than the manual TZ segmentation, measured by DSCs between two experts (Table 3-5). This may be due to the more complex shape and structure of PZ as its boundaries are sometimes not well discerned due to a variety of factors such as prostate or patient motion. Similarly, Meyer et. al,[36] reported that the PZ segmentation had worse inter-reader consistency than TZ segmentation with three different experts (first urologist, second urologist with the help of a medical student, and an assistant radiologist). Meyer et al. also utilized three orthogonal planes of the T2w MRI, i.e., sagittal, coronal and axial, to automatically determine the bounding box for the prostate before performing the segmentation. The bounding box approach could be added as pre-processing to improve both the segmentation performance and inter-reader consistency by minimizing the false positives.

Table 3-5 DSCs of the Proposed Algorithm on Different Types of Slices in the External Testing Dataset. P Values Relative to Inter-Reader Agreement (Expert 1 vs. Expert 2) are Given in the Table for Each and Were Obtained by Using Wilcoxon Signed-Rank Test.

Comparisons	All slices		Middle Slices		Apex-End	Base-End
Zone	PZ	TZ	PZ	TZ	PZ	TZ
Model vs. Expert 1	0.74 ±0.07 P<0.05	0.79 ±0.12 P<0.05	0.75 ±0.07 P<0.05	0.83 ±0.09 P<0.05	0.84 ±0.11 P<0.05	0.77 ±0.21 P<0.05
Expert 1 vs. Expert 2	0.71 ±0.13	0.75 ±0.14	0.71 ±0.13	0.81 ±0.12	0.76 ±0.21	0.65 ±0.27

In the ETD, when only considering middle slices for testing, mean DSCs were higher than considering all slices. This may be because: 1) the features for the differentiation of PZ and TZ are more distinct in the middle slices than the other slices. 2) when only considering middle slices, some false positives from adjacent non-prostate slices, apex-end slices and base-end slices can be avoided. Besides, we also found mean PZ DSC for apex-end slices is larger than the PZ DSC for middle slices, but to the contrary, TZ DSC for base-end slices is smaller than the TZ DSC for middle slices. The large standard deviations and low DSC of TZ for base-end slices indicated some significant discrepancies between two experts at the base-end. This indicates that it's hard to recognize TZ in the base-end slices, which may explain why the proposed method got a low TZ DSC at the base-end.

Compared with the DSCs to that of Meyer et al.,[36] our method's DSCs for both PZ and TZ are slightly lower. This may be related to: 1) Difference in sample sizes for the evaluation. In our

method, 63 patient datasets were used for the testing data set, in compared with their testing data set of only 20 patients. 2) Discrepancy in manual annotations for both PZ and TZ. 3) Inherent differences in methods. 4) Differences of preprocessing. In their method, before the segmentation, the bounding box for the prostate was determined to reduce the false positives.

Our study also has a few limitations. Firstly, the same MRI vendor was used for both ITD and ETD. Also, in-plane resolution of the ITD is very close to that of the ETD. Datasets from different vendors and with considerable different in-plane resolutions will be incorporated into future related studies. Secondly, the proposed algorithm is a 2D-based FCN model, which is still deficient in capturing inter-slice correlation information compared to 3D-based models. In the future, we will explore ways of improving the capturing of inter-slice correlation information in our proposed model. Thirdly, the number of experts involved in obtaining inter-reader consistency in the paper is two. In the future, more experts will be added in the study to get more robust inter-reader consistency.

3.5 Conclusion

This Chapter proposed a novel deep learning algorithm for the automatic segmentation of the two prostate zones using T2w MRI. The proposed algorithm outperforms the U-Net on automatic segmentation of PZ and TZ. The difference between the proposed method's performance is similar on the ITD and ETD, especially for the segmentation of PZ. Moreover, the performance of the proposed method is comparable to the experts in the external testing dataset.

Chapter 4: Exploring Uncertainty Measures in Bayesian Deep Attentive Neural Networks for Prostate Zonal Segmentation

This chapter describes a Bayesian deep learning method for automated prostate zonal segmentation. The deep learning method from Chapter 3 forms the bone structure of the Bayesian deep learning method, such as the encoder and decoder. Apart from the bone structure, the Bayesian deep learning method also equips with the dropout layers to help generate the uncertainty, and spatial attention mechanism to further enhance the segmentation performance.

4.1 Introduction

Prostate cancer (PCa) is the most common solid organ malignancy and is among the most common causes of cancer-related death among men in the United States[33]. Multi-parametric MRI (mpMRI) is the most widely available non-invasive and sensitive tool for the detection of clinically significant PCa (csPCa), 70% and 30% of which are located in the peripheral zone (PZ) and transition zone (TZ) respectively[14,43]. The clinical reporting of mpMRI relies on a qualitative expert consensus-based structured reporting scheme (Prostate Imaging-Reporting and Data System (PI-RADS)). The interpretation is based primarily on diffusion-weighted imaging (DWI) in the peripheral zone (PZ) and T2-weighted (T2w) imaging in the transitional zone (TZ) since csPCa lesions have different primary imaging features[14,43].

Accurate segmentation of PZ and TZ within the 3T mpMRI is essential for localization and staging of csPCa to enable MR targeted biopsy and guide and plan further therapy such as radiation, surgery, and focal ablation[44]. Segmentation of the prostate zones on mpMRI is typically done manually, which can be time-consuming and sensitive to readers' experience, resulting in significant intra- and inter-reader variability[5]. Automated segmentation of prostatic zones (ASPZ) is reproducible and beneficial for consistent location assignment of PCa lesions[36]. ASPZ also enables automated quantitative imaging feature extraction related to prostate zones and can be used as a pre-processing step to improve the computer-aided diagnosis (CAD) of PCa[45].

ASPZ was previously proposed by the atlas-based method[46]. Later, Zabihollahy et al.[41] proposed a U-Net-based method for ASPZ. Clark et al.[39] developed a staged deep learning architecture, which incorporated a classification into U-Net, to segment the whole prostate gland and TZ. However, the U-Net-based segmentation sometimes resulted in inconsistent performance because the anatomic structure of the prostate can be less distinguishable, and the boundaries between PZ and TZ may distort semantic features[5]. Liu et al.[5] recently improved the encoder of the U-Net by using the residual neural network, ResNet50[42], followed by feature pyramid attention to help capture the information at multiple scales. Furthermore, Rundo et al.[47] proposed an attentive deep learning network for ASPZ via incorporating the squeeze and excitation (SE) blocks into U-Net. SE adaptively recalibrated the channel-wise features to potentially improve inconsistencies in the segmentation performance. Moreover, segmentation outcomes from ASPZ are typically deterministic; there is a lack of knowledge on the confidence of the model[25]. Providing uncertainties of the model can improve the overall segmentation workflow since it easily allows refining uncertain cases by human experts[25]. The uncertainty can be estimated by the Bayesian deep learning model, which not only produces predictions but also provides the

uncertainty estimations for each pixel. This can be done by adopting probability distributions of weights rather than the deterministic weights of the model.

In this study, we propose an ASPZ with an estimation of pixel-wise uncertainties using a spatial attentive Bayesian deep learning network. Different from Rundo et al.[47], we adopt a spatial attentive module (SAM), which models the long-range spatial dependencies between PZ and TZ by calculating the pixel level response from the image[48]. The proposed model incorporates four sub-networks, including SAM, an improved ResNet50 with dropout, a multiple-scaled feature pyramid attention module (MFPA)[5], and a decoder. The SAM forces the entire network focusing on specific regions that have more abundant semantic information related to prostatic zones. We use the improved ResNet50 to handle the heterogeneous prostate anatomy with semantic features. The MFPA is designed to enhance the multi-scale feature capturing. Finally, the spatial resolution is recovered by the decoder. We also implement the Bayesian model through both training the proposed model with dropout and Monte Carlo (MC) samples of the predictions during the inference, inspired by prior work by Gal and Ghahramani[28]. The dropout can be regarded as using Bernoulli's random variables to sample the model weights[28]. We evaluate the proposed model's performance using internal and external testing datasets and compared it with previously developed ASPZ methods. The segmentation performance is compared to investigate the discrepancy between two MRI datasets. The importance of each individual module within the proposed method is also examined. Finally, the overall prostate zonal segmentation at apex, middle, and base slices are computed to illustrate the uncertainty of segmentation at different positions of the prostate.

4.2 Materials and Methods

This study was carried out in compliance with the United States Health Insurance Portability and Accountability Act (HIPAA) of 1996 with approval by the local institutional review board (IRB). The MRI datasets were acquired from two sources. For model development and internal testing (n = 259 and n = 45)—internal testing dataset (ITD)—we used the Cancer Imaging Archive (TCIA) data from the SPIE-AAPM-NCI PROSTATE X (PROSTATE X) challenge[7]. For independent model testing, we used an external testing dataset (ETD) (n = 47; age 45 to 73 years and weight 68 to 113 kg) retrieved from our tertiary academic medical center. For the ETD, the pre-operative MRI scans, which were acquired between October 2017 and December 2018 using one of the three 3T MRI scanners (Skyra (n = 39), Prisma (n = 1), and Vida (n = 7); (Siemens Healthineers, Erlangen, Germany)) were collated. For both ITD and ETD data, both PZ and TZ were contoured using OsiriX (Pixmeo SARL, Bernex, Switzerland) by MRI research fellows. Then, two genitourinary radiologists (10-19 years of post-fellowship experience interpreting over 10,000 prostate MRI) cross-checked the contours. The axial T2 TSE (turbo spin-echo) MRI sequence was used for both ITD and ETD segmentation (Table 4-1). Prior to the training and testing, all the images in both datasets were normalized to an interval of [0, 1] and were also resampled to the common in-plane resolution (0.5×0.5 mm).

Table 4-1 Detailed T2w TSE Protocols Used for Two MRI Datasets

Datasets	Internal Testing dataset (ITD)	External Testing dataset (ETD)
Spatial Resolution	0.5x0.5x3.6mm^3	0.65x0.65x3.6mm^3
Flip angle	160°	160°
Matrix Size	380x380	320x320
Field-of-View	190x190 mm^2	208x208 mm^2

Repetition Time/Echo Time	5660 ms / 104 ms	4000 ms / 109 ms

4.3 Methods

4.3.1 Proposed Model for Automatic Prostatic Zonal Segmentation

The overall workflow of the proposed network is shown in Figure 4-1, which consists of four sub-networks. By joining the four sub-networks together, a fully end-to-end prostatic zonal segmentation workflow was formed. Both PZ and TZ segmentations were done simultaneously using a single network.

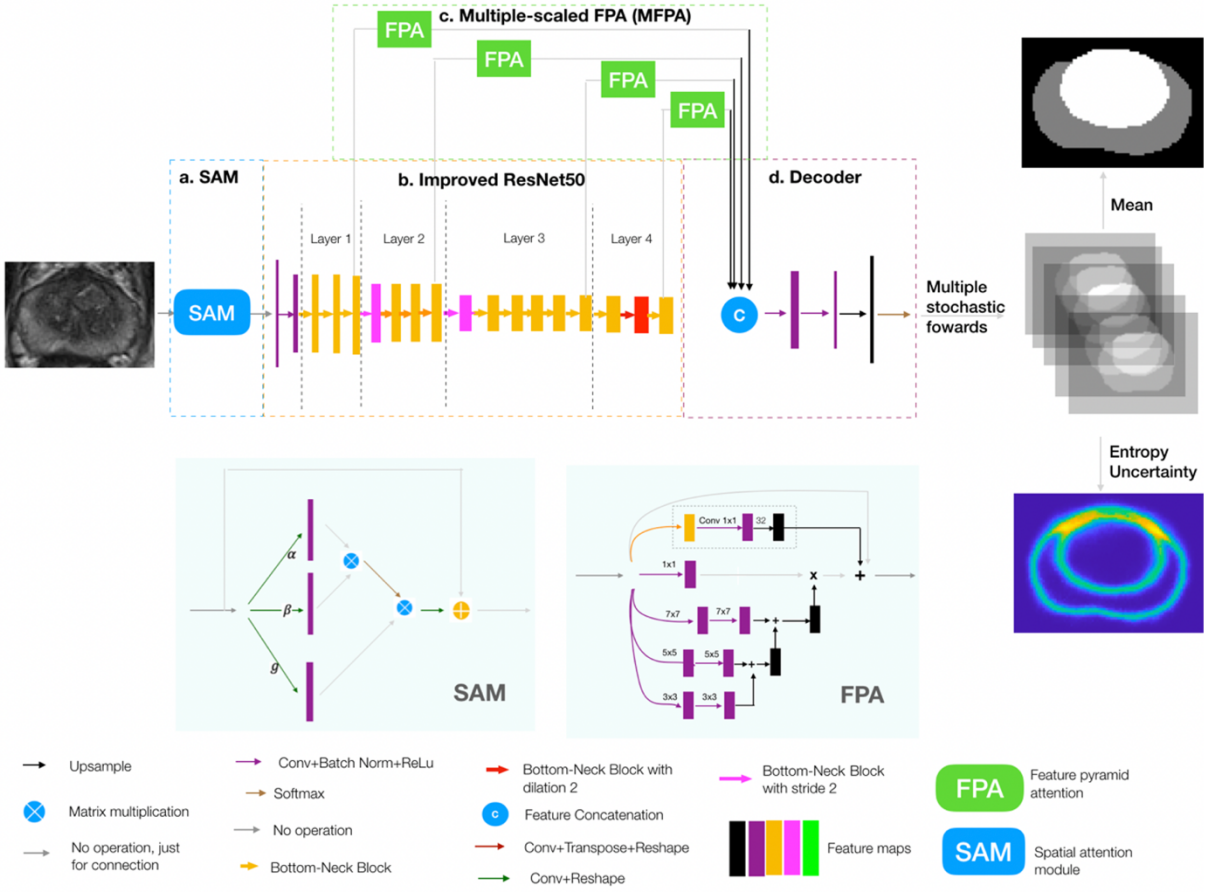

Figure 4-1 A whole workflow of the proposed model. Input is a 2D T2w MRI slice, and output is a segmentation mask, which has the PZ and TZ segmentation result (Gray and white colors indicate PZ and TZ, respectively), and a pixel-wise uncertainty map (yellow pixel indicates large uncertainty and blue indicates low uncertainty). There are four sub-networks in the network, which are (a) spatial attention

module (SAM), (b) improved ResNet50, (c) multiple-scaled feature pyramid attention (FPA), and (d) decoder.

Inspired by Wang et al.[48], the SAM was designed to make the network intelligently pay attention to the regions, which had more semantic features associated with PZ and TZ (shown in Figure 4-1 (a)). Inside the images, there existed some spatial dependencies of PZ and TZ pixels. For instance, TZ was always surrounded by PZ in the bottom of the prostate, and the urinary bladder region was always above PZ and TZ in the image and TZ was usually in the image center. SAM helped the network to model such spatial dependent information through global features. Specifically, the response at each pixel was computed by considering all the pixels in the image. Higher priorities were then adaptively assigned to the pixels, which had more informative semantic features. Detailed processes regarding spatial attention are shown in the left bottom of Figure 4-1. After going through a convolution layer and reshaping, three kinds of vectors - query vector $\alpha(x)$, key vector $\beta(x)$ and representative vector $g(x)$, were formed. Then, we performed the matrix multiplication between the transpose of the query vector and the key vector, and after that, we applied a soft-max layer to compute the weight matrix which models the spatial relationship between any two pixels of the features. Next, we again performed a matrix multiplication between the weight matrix and the representative vector and reshaped the result to the size of original features. These processes can be formulated by:

$$y = softmax\big(\alpha(x)^T * \beta(x)\big) * g(x) \qquad (4\text{-}1)$$

where x and y represent the raw image and attentive map of the raw image, respectively. * means matrix multiplication. Finally, an element-wise sum operation between the result above

and the original features was performed to obtain the final result which reflected the long-range dependencies.

Improved ResNet50 (shown in Figure 4-1 (b)) was served as the bone structure of the network. ResNet50 in this paper was improved by the following three steps, which followed the methods of Liu et al.[5]. First, the initial max-pool was removed since it was proved to compromise the performance of segmentation. Bottleneck block at stride one as the first block in the 4th layer was replaced with the regular block. Then, we used the dilated bottleneck to serve as the second block in the 4th layer so as to minimize the potential loss to the spatial information. Finally, the dropout layer was inserted after each block within the improved ResNet50 to transform the current neural network to the bayesian neural network[49]. 3) Multi-Scaled Feature Pyramid Attention (MFPA) Feature pyramid attention (FPA) module (shown in the bottom right of Figure 4-1) was applied after each layer within Resnet50 to help capture the features from the multiple scales. Next, feature maps after each FPA were then upsampled to the same size and then concatenated in the decoder.

The decoder (Figure 4-1 (d)) was used to recover feature maps' spatial resolution. In the decoder, the total features calculated in the 3) went through two 3×3 convolutional layers and one 1×1 convolutional layer, followed by an up-sampling (by a factor of 4). In the end, the multi-class softmax classifier was performed for the simultaneous segmentation of TZ and PZ.

4.3.2 Uncertainty Estimation for Prostate Zonal Segmentation

Figure 4-1 shows the uncertainty estimation workflow by the proposed method. Monte Carlo dropout[28] was served as the method for approximate inference. Usually, a posterior distribution $p(W|X,Y)$ placed over weights W of the neural network is computed to capture the uncertainty in the model, where X is the training samples, and Y is the corresponding ground truth

labels of prostate zones[27]. However, it is intractable to compute the posterior. The posterior can be approximated by the variational distribution q(W) , which minimizes the Kullback-Leibler (KL) divergence between the actual posterior and the variational distribution: $\mathrm{KL}\big(q(W)\big)||p(W|X,Y)$ that performing dropout on a hidden layer is equivalent to placing the variational distribution – Bernoulli distribution over the weights of that layer[28]. Also, the effect of minimizing the cross-entropy loss is the same as the minimization of the KL-divergence. Therefore, training with dropout allows the approximate inference. These dropouts are also required to be kept active during the testing. As the dropout is the same as placing a Bernoulli distribution over the network weights, the sample from a dropout network's outputs can be used to approximate the posterior. A Monte Carlo sample from the posterior distribution is produced by performing a stochastic forward pass through a trained dropout network. There are two types of uncertainties: epistemic uncertainty - caused by the ineptitude of the model because of the lack of training data; aleatoric uncertainty - caused by the noisy measurements in the data[28]. Epistemic uncertainty can be mitigated by increasing the training samples. Aleatoric uncertainty can be restrained by increasing the sensor precision. Aleatoric uncertainty occurs during measuring the inherent noise in the samples and is reflected in the uncertainty over the model's parameters[50]. A model with the precise set of parameters will lower down the aleatoric uncertainty[50]. The combination of aleatoric and epistemic uncertainty forms the predictive uncertainty[27]. In this paper, we focused on the exploring of predictive uncertainty for the prostate zonal segmentation, which can be measured by the entropy of the predictive distribution[27,50] and is formulated as:

$$-\sum_{c=1}^{C}\frac{1}{T}\sum_{t=1}^{T}p(y=c|x,w_t))\log\left(\frac{1}{T}\sum_{t=1}^{T}p(y=c|x,w_t)\right) \qquad (4\text{-}2)$$

where y is the output variable, T is the number of stochastic forward passes (50 was chosen is the experiments (Figure 4-1)), C is the number of classes (C=3, for background, PZ and TZ), $p(y = c|x, w_t)$ is the soft-max probability of input x being in class c, w_t represents model's parameters on the t_{th} forward pass.

4.3.3 Average Uncertainty Maps for the Prostate Zonal Segmentation

The average uncertainty map tells the overall zonal uncertainty in different positions on the prostate image. Figure 4-2 shows the processes of obtaining the average uncertainty map.

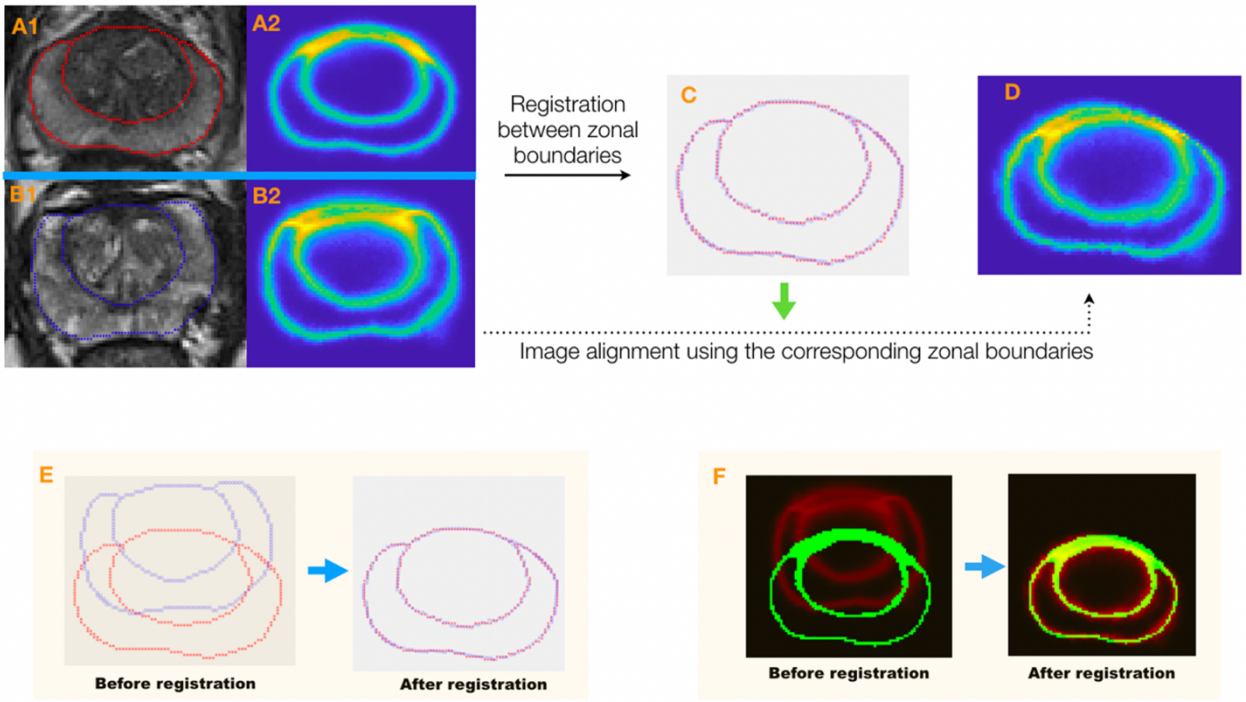

Figure 4-2 The overall workflow for the registration of the sample (one of the non-templates) uncertainty map to the template uncertainty map. A1 and A2 are a template image and its uncertainty map. B1 and B2 are a sample image and its uncertainty map, respectively. C shows the result after the zonal boundary registration between the sample and the template. Red and blue points represent the zonal boundaries on the template and the sample images, respectively. D is the warped uncertainty map based on the corresponding zonal points after the registration. E and F show the overlapping of zonal boundary points and uncertainty maps before and after registration.

In order to obtain the average uncertainty map at the prostate apex, middle portion, and base, three template prostate images at the three sections were chosen by a radiologist after inspecting all the prostate images. Next, for each prostate section, zonal boundary points on non-template prostate images (sample images) were then registered to those on the prostate template image within the section using a non-rigid coherent point drift method (CPD)[51]. Within non-rigid CPD, alignment of two-point sets was thought of as a probability density estimation problem where one point set serves as the centroids of the gaussian mixture model (GMM), and the other represents the data points. By maximizing the likelihood, GMM centroids were then fitted to the data. Also, GMM centroids were forced to move coherently to preserve the topological structure by regularizing the displacement field and utilizing the variational calculus to obtain the optimal transformation. The thin plate spline (TPS) method[52] was then used to warp the sample uncertainty maps to the template uncertainty map based on the corresponding zonal boundary points (Figure 4-2). In doing so, the average was computed among all the warped sample prostate uncertainty maps, including the template uncertainty map, yielding an average uncertainty map in this prostate section. In the end, three average uncertainty maps were obtained for the prostate apex, middle portion, and base. In addition, the prostate zonal average uncertainty score for each prostate section was calculated by averaging all of the pixels' uncertainties in the zone.

4.3.4 Model Development and Testing

Cross entropy (CE) served as the loss function to train the proposed model. For each given pixel, cross-entropy was formulated as,

$$CE = \frac{1}{3}\sum_{i=0}^{2} -y_i \log(p_i) - (1 - y_i)\log(1 - p_i) \qquad (4\text{-}3)$$

where $y_i \in \{0,1\}$ is the ground-truth binary indicator, corresponding to the 3-channel predicted probability vector $p_i \in [0,1]$. Training and evaluation were performed on a desktop computer with a 64-Linux system with 4 Titan Xp GPU of 12 GB GDDR5 RAM. Pytorch was used for the implementation of algorithms. The learning rate was initially set to 1e-3. The model was trained for 100 epochs with batch size 8. The loss was optimized by stochastic gradient descent with momentum 0.9 and L2-regularizer of weight 0.0001. The central regions $80mm \times 80mm$ were automatically cropped from the original images of the prostate. This is because prostate areas are always located in the middle. On-the-fly data augmentation approaches included random rotation between $[-3°, 3°]$, flipped horizontally, and elastic transformations. For the elastic transformation, there are three steps: 1) A coarse displacement grid with a random displacement for each grid point was generated. 2) Displacement for each pixel (deformation field) in the input image was computed via a thin plate spline (TPS) method on the coarse displacement grid. 3) The input image and the corresponding segmentation mask were deformed according to the deformation field. (Bilinear and nearest-neighbor interpolation methods were used to handle the non-integer pixel locations on the warped input image and segmentation mask). Totally, we used 308 unique subject MRIs from PROSTATE X for model development and internal testing. The model was trained by 70% (N = 218) of the dataset, with 15% (N = 45) held out for validation and 15% (N = 45) for internal testing (internal testing dataset (ITD)). For external testing (external testing dataset (ETD)), 47 unique subject MRI from the large U.S. tertiary academic medical center were used. No endorectal coil was used in the study. Patient-wised Dice Similarity Coefficient (DSC) was employed to evaluate the segmentation performance and to compare with baseline methods, which is formulated as:

$$DSC = \frac{2|A \cap B|}{|A|+|B|} \quad (4\text{-}4)$$

where A is the predicted 3D zonal segmentation, which is stacked by the 2D algorithmic prostate zonal segmentation and B is the ground-truth of 3D zonal segmentation stacked by the 2D manual segmentation on the prostate slices.

Patient-wise Hausdorff Distance (HD) was also used to evaluate the segmentation performance, which is formulated as:

$$HD(X,Y) = max(h(X,Y), h(Y,X)) \quad (4\text{-}5)$$

where $h(X,Y)$ is the directed HD, which is given by $h(X,Y) = \max_{x \in X} \min_{y \in Y} \|x - y\|$, X and Y are the point sets on the A and B (defined in the Patient-wised DSC).

4.3.5 Statistical Analysis

The distribution of DSCs was described by the mean and standard deviation. Paired sample t-test test was used to compare the performance difference between the proposed method and baselines on both ITD and ETD. The performance difference of the proposed method was also tested by paired sample t-test.

4.4 Result

4.4.1 Performance Using Internal Testing Dataset (ITD) and External Testing Dataset (ETD)

Figure 4-3 shows two typical examples of prostate zonal segmentation results by the proposed method and the three comparison methods, including Deeplab V3+[53], USE-Net[47], U-

Net[12], Attention U-Net[54] and R2U-Net[55]. USE-Net was proposed by Rundo et al for the prostate zonal segmentation, which embeds the squeeze-and-excitation (SE) block into the U-Net and enables the adaptive channel-wise feature recalibration. Attention U-Net, proposed by Ozan et al, which incorporates attention gates into the standard U-Net architecture to highlight salient features that passes through the skip connections. Deeplab V3+[53] is one of the state-of-art deep neural networks for image semantic segmentation, which takes the encoder-decoder architecture to recover the spatial information and utilizes multi-scale features by using atrous spatial pyramid pooling (ASPP). Convolutional features at multiple scales are probed by ASPP via applying several parallel atrous convolutions with different rates. R2U-Net is an extension of standard U-Net using recurrent neural network and residual neural networks.

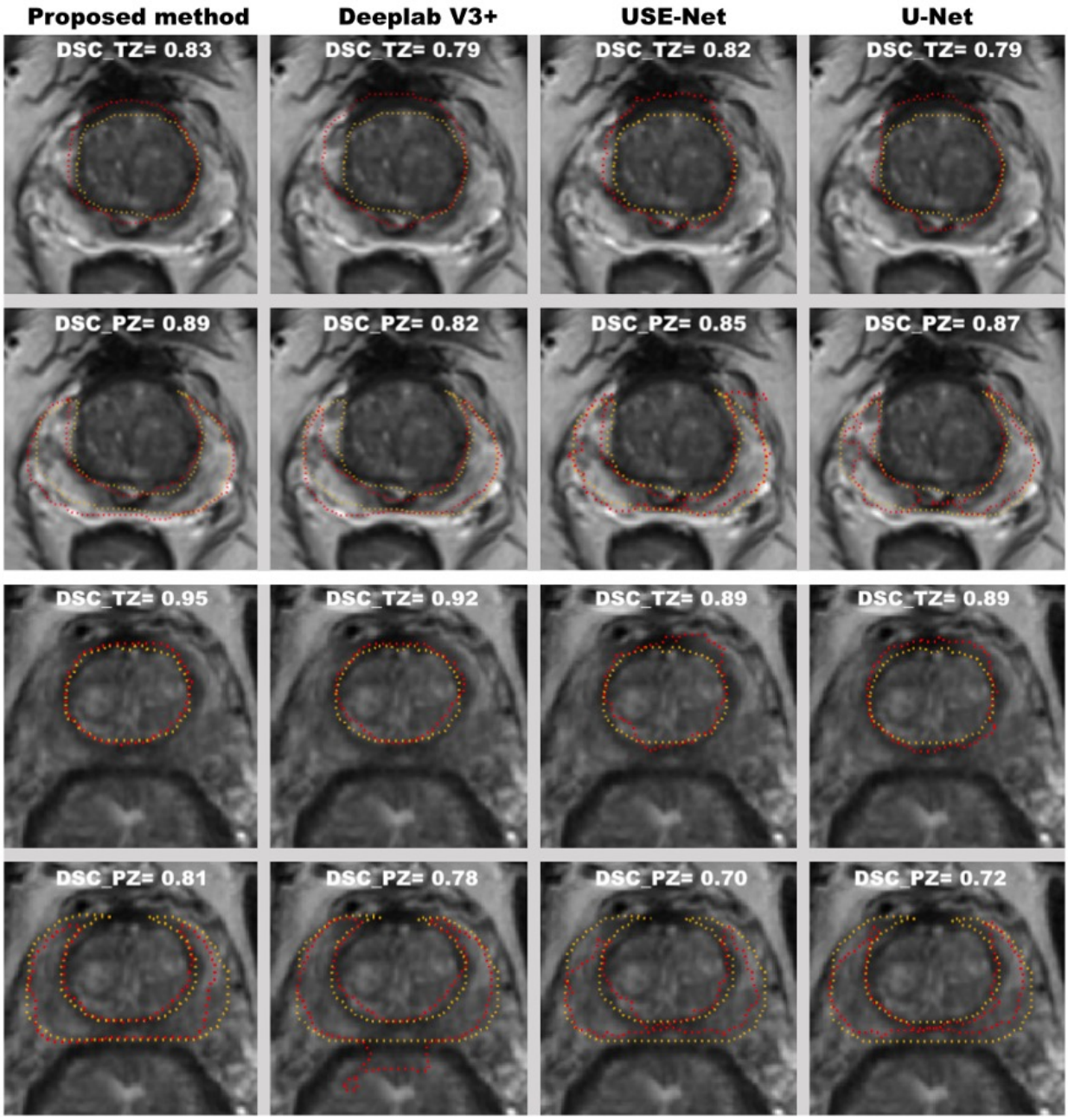

Figure 4-3 Two representative examples of the zonal segmentation by the proposed method, DeeplabV3+ USE-Net, U-Net. Yellow lines are the manually annotated zonal segmentation, and the red lines are algorithmic results. The top two and bottom two rows represent the segmentation examples from two different subjects.

Means and standard deviations of DSCs for PZ and TZ on ITD and ETD are shown in Table 4-2. Mean DSCs for PZ and TZ of the proposed method were 0.80 and 0.89 on ITD, 0.79 and 0.87 on ETD, which were all higher than the results obtained by the comparison methods with significant difference.

Table 4-2 Performance (DSC) of the Proposed Method and Baselines on Internal Testing Dataset (ITD) and External Testing Dataset (ETD). P Values are the Comparisons Between the Proposed Methods and Baselines in ITD and ETD

Datasets	ITD		ETD	
	PZ	TZ	PZ	TZ
Proposed Method	**0.80±0.05**	**0.89±0.04**	**0.79±0.06**	**0.87±0.07**
Deeplab V3+	0.74±0.06 P<0.05	0.87±0.05 P<0.05	0.71±0.09 P<0.05	0.82±0.06 P<0.05
Attention U-Net	0.75±0.08 P<0.05	0.87±0.04 P<0.05	0.75±0.07 P<0.05	0.82±0.08 P<0.05
R2U-Net	0.70±0.10 P<0.05	0.85±0.05 P<0.05	0.69±0.08 P<0.05	0.78±0.10 P<0.05
USE-Net	0.72±0.10 P<0.05	0.86±0.06 P<0.05	0.72±0.08 P<0.05	0.80±0.08 P<0.05

U-Net	0.71±0.09 P<0.05	0.85±0.06 P<0.05	0.72±0.07 P<0.05	0.81±0.06 P<0.05

Means and standard deviations of Hausdorff Distance (HD) are shown in Table 4-3. The proposed method achieved the lowest mean HD among all the methods for both PZ and TZ segmentation.

Table 4-3 Average Hausdorff Distance (mm) of the Proposed Method and Baselines on Internal Testing Dataset (ITD) and External Testing Dataset (ETD). P Values are the Comparisons Between the Proposed Methods and Baselines in ITD and ETD.

Datasets	ITD		ETD	
	PZ	TZ	PZ	TZ
Proposed Method	**4.77±2.86**	**3.52±1.81**	**5.96±3.13**	**4.92±2.73**
Deeplab V3+	5.48±2.55 P=0.26	5.33±4.50 P<0.05	7.72±4.47 P<0.05	7.45±5.36 P<0.05
Attention U-Net	5.79±3.96 P=0.14	5.27±4.28 P<0.05	7.60±4.82 P=0.06	5.92±4.02 P=0.13
R2U-Net	5.46±2.76 P=0.23	6.24±4.76 P<0.05	7.89±4.64 P<0.05	10.01±8.54 P<0.05
USE-Net	8.61±6.83	7.90±6.27	9.74±7.06	10.96±9.03

	P<0.05	P<0.05	P<0.05	P<0.05
U-Net	8.88±7.63 P<0.05	11.38±7.63 P<0.05	9.72±6.06 P<0.05	11.55±6.54 P<0.05

Figure 4-4 showed the superior and inferior cases for the PZ and TZ segmentation. The superior case had DSC > 0.90 for PZ segmentation and DSC > 0.95 for TZ segmentation. DSCs of the inferior case were lower than 0.60 and 0.50 for the PZ and TZ segmentations, respectively.

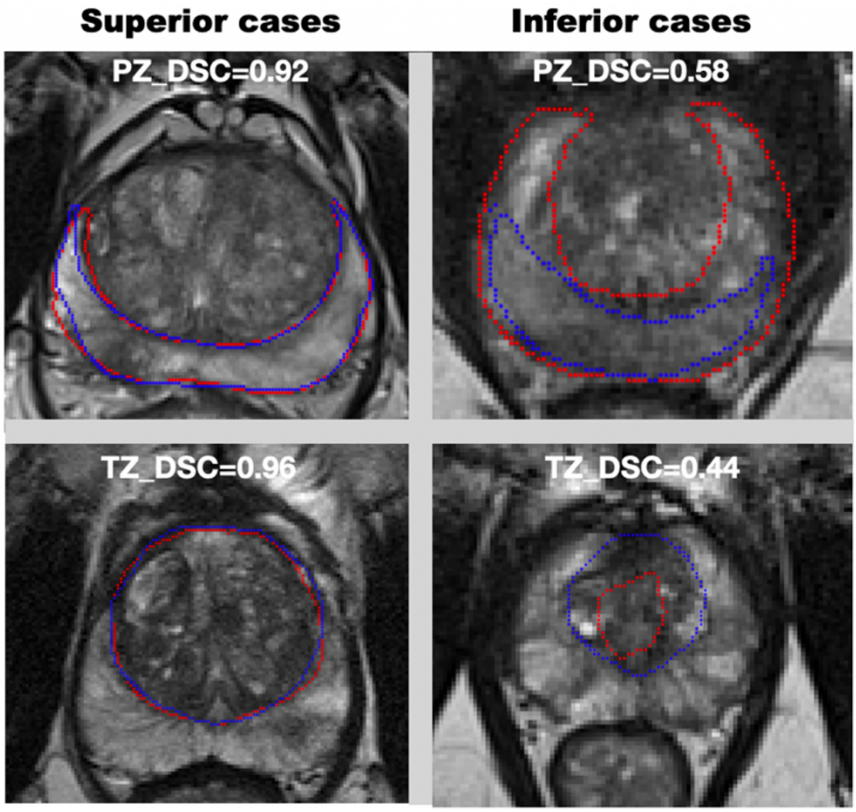

Figure 4-4 Superior and inferior cases for PZ and TZ segmentation. Superior and inferior cases for PZ and TZ are shown in the first and second row.

4.4.2 Performance Discrepancy Between the Internal Testing Dataset (ITD) and External Testing Dataset (ETD)

There was no significant difference ($p<0.05$) between ITD and ETD for the performance of PZ segmentation for the proposed method. However, there was a 2.2% difference for the TZ segmentation (Table 4-2).

4.4.3 Performance Investigation for Each Individual Module in the Proposed Method

We carried out the following ablation studies to investigate the importance of each module within the proposed network. Table 4-4 indicates which module was used (a checkmark) or not used (a cross) in each experiment. We showed that the best model performance is achieved when both SAM and MFPA are used in the model for the zonal segmentation.

Table 4-4 Performance Investigation for Each Individual Module of the Proposed Method. Average DSCs With Standard Deviation are Shown in the Table. SAM is the Spatial Attention Module. MFPA is the Multi-Scale Feature Pyramid Attention. Apart From the Proposed Method, There are Two Additional Independent Experiments, Where √ and × Under Each Row Indicates Whether the Experiment Contains the Module or Not

Experiments	SAM	MFPA	ITD		ETD	
			PZ	TZ	PZ	TZ
Proposed method	√	√	**0.80 ±0.05**	**0.89 ±0.04**	**0.79 ±0.06**	**0.87 ±0.07**
Experiment 1	×	√	0.79 ±0.07	0.88 ±0.04	0.77 ±0.07	0.85 ±0.07

Experiment 2	√	×	0.78 ±0.07	0.88 ±0.04	0.77 ±0.06	0.85 ±0.07

In experiment 1, DSCs for both zones on ITD and ETD decreased and were lower than the proposed model when SAM was removed from the proposed model, which proved that SAM helped improve the overall segmentation performance. In experiment 2, DSCs for PZ on ITD, and both zones on ETD decreased when MFPA was removed from the proposed model, indicating that that GFM was essential within the model.

4.4.4 The Overall Uncertainty for the Prostate Zonal Segmentation of the Proposed Method

Figure 4-5 and Table 4-5 shows the overall uncertainties of the proposed method for the prostate zonal segmentation. The pixel-by-pixel uncertainty maps showed that the zonal boundaries had higher uncertainties than the interior areas at the three prostate locations (apex, middle, and base slices). Also, highest uncertainties were observed at the intersection between the PZ, TZ and the AFS.

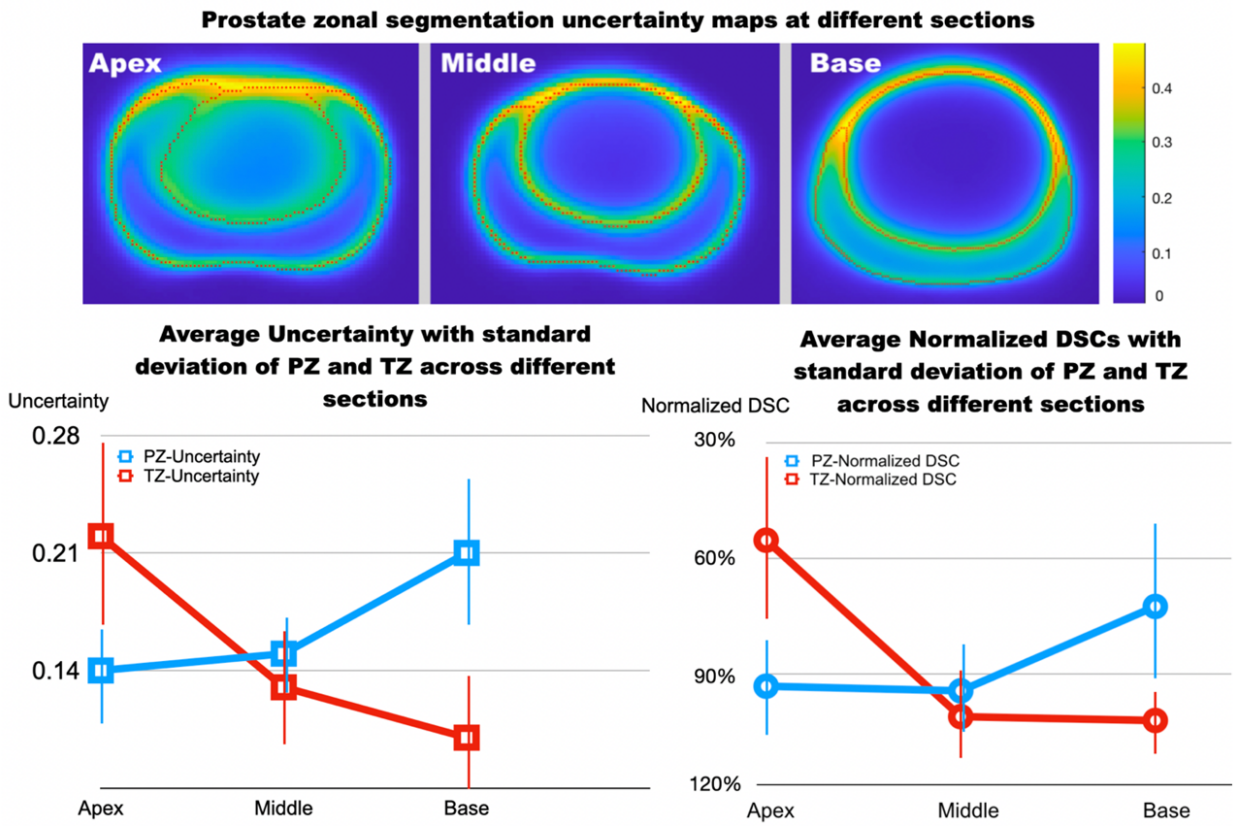

Figure 4-5 The pixel-by-pixel uncertainty estimation of the zonal segmentation at the apex, middle, and base slices of the prostate (top). The orange color indicates high uncertainties, and blue color indicates low uncertainties. Bottom: Average uncertainty scores (bottom left) and average normalized DSCs (bottom right; normalized by TZ DSC– 0.87 shown in in Table 4-4) with the standard deviation at the apex, middle, and base slices of the prostate (x-axis).

Table 4-5 Performance investigation for each individual module of the proposed method. Average DSCs with standard deviation are shown in the table. SAM is the spatial attention module. MFPA is the multi-scale feature pyramid attention. Apart from the proposed method, there are two additional independent experiments, where √ and × under each row indicates whether the experiment contains the module or not.

Experiments	SAM	MFPA	ITD		ETD	
			PZ	TZ	PZ	TZ

Proposed method	√	√	0.80 ±0.05	0.89 ±0.04	0.79 ±0.06	0.87 ±0.07
Experiment 1	×	√	0.79 ±0.07	0.88 ±0.04	0.77 ±0.07	0.85 ±0.07
Experiment 2	√	×	0.78 ±0.07	0.88 ±0.04	0.77 ±0.06	0.85 ±0.07

The TZ segmentation had lower overall uncertainties than the PZ segmentation, and the proposed method achieved better segmentation in TZ (DSC=0.87) compared to PZ (0.79). We used a normalized DSC (DSC_{norm}, normalized by TZ DSC – 0.87) to show relative differences at different locations of the prostate. For PZ segmentation, the highest overall uncertainty was observed at base, consistent with the worst model performance at base (DSC_{norm}=72.4 %). For TZ segmentation, the highest overall uncertainty was observed at the apex, matched with the worst segmentation performance of the model at apex (DSC_{norm}=55.2 %). Figure 4-5 bottom-left shows the average uncertainty estimation at different prostate locations, and the trend is well matched with the actual model performance (Fig. 4-5 bottom-right).

4.5 Discussion

In this study, an attentive Bayesian deep learning model that accounts for long-range spatial dependencies between TZ and PZ with an estimation of pixel-wise uncertainties of the model was proposed. The performance discrepancy between ITD and ETD of the proposed model was

minimal. There was no difference in PZ segmentation between ITD and ETD, and a 2.2% discrepancy in TZ segmentation. The average uncertainty estimation showed lower overall uncertainties for TZ segmentation than PZ, consistent with the actual segmentation performance difference between TZ and PZ. We attribute this to the complicated and curved shapes of PZ. The PZ boundaries generally have bilateral crescentic shapes, while the TZ boundaries are ellipsoid in shape. SAM aided the model to focus on certain spatial areas in the zonal segmentation. This was done by the modeling of spatial dependencies with the help of global features. Since spatial attention was inserted adjacent to the raw images, large GPU memory was required to obtain the global spatial features during the training and evaluation. The SAM can be inserted into other positions within the network, but we observed that the zonal segmentation performed the best when the SAM followed directly after the raw image. There exist high segmentation uncertainties on the zonal boundaries. This may be explained by the inconsistent manual annotations since the boundaries between TZ and PZ are hard to be defined precisely due to partial volume artifact. This resembles the "random error", which persists throughout the entire experiment, so we call such uncertainty "random uncertainty" in the prostate zone segmentation. The areas with the highest uncertainty are located at the junction of AFS, PZ and TZ. One possible reason is that it is hard for the MRI to distinguish the tissue around the junction. There is probably a significant reduction of signal by the more severe partial volume artifacts caused by PZ with the high pixel intensity, TZ with the intermediate pixel intensity and AFS with lower pixel intensity. The overall uncertainties were higher at apex slices than those at base slices for the TZ segmentation. This may be caused by the fact that the size of TZ gradually increases from apex to base slices, making it hard to recognize the zone for the model. In contrast, the overall uncertainties for PZ were higher at the

base slices than at the apex and middle slices. Similar to TZ, we attributed the low uncertainties to the large PZ structure between apex and middle slices (Figure 4-6).

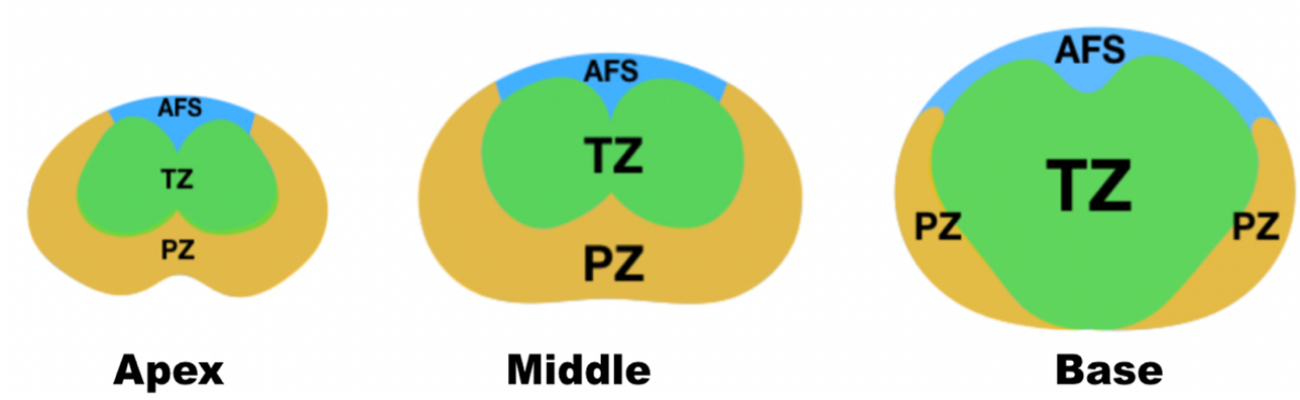

Figure 4-6 Prostate zonal anatomy at apex, middle, and base slices of the prostate.

The estimation of pixel-wise uncertainties of the prostate zonal segmentation would provide confidence and trust in an automatic segmentation workflow, which allows a simple rejection or acceptance based on a certain uncertainty level. This can be implemented as a partial or entire rejection of the automatic segmentation results when presenting to experts, and future research will be needed to determine the level of uncertainties to be acceptable to experts. We believe that this additional confidence would enable more natural adaption or acceptance of the automatic prostate segmentation than the one without it when the prostate segmentation is integrated into the downstream analysis decision. We observed that simple incorporation of the inter-slice information by 3D U-Net was not sufficient to improve the segmentation performance. Our prostate MRI data had a lower through-plane resolution (3-3.6 mm) than the in-plane resolution (0.5-0.65 mm), resulting in a conflict between the anisotropism of the 3D images and isotropism of the 3D convolutions[56,57]. This may be the main reason that the model's generalization was compromised. Specifically, voxels in the x-z plane will correspond to the structure with

different scales along x- and z-axes after the 3D convolution[58]. Moreover, the performance was more significantly different when both ITD and ETD were used for testing, potentially due to the difference in the imaging protocol. Further study may be needed to investigate advanced approaches that incorporate the inter-slice information into the 3D convolution when there exists a difference between in-plane and through-plane resolutions while minimizing sensitivities to different imaging protocols. The significant effect of including SAM and MFPA was investigated in the ablation study. The average DSCs of the proposed method were higher than the experiments in the ablation study for PZ and TZ in both datasets. However, there were no significant differences between DSCs obtained by the experimental methods and the proposed method for both zones in the ablation study when a paired t-test was used. Based on the power analysis, we need 100, 253, 143, and 194 cases for Experiment 1 in Table 4-4 (when SAM is removed) and 394, 253, 143, and 194 cases for Experiment 2 (when MFPA is removed) to achieve 80% power with alpha = 0.05. We also compared the uncertainty of the proposed method and that of the U-Net. We found that average uncertainty scores of the proposed method for both PZ and TZ at three different prostate locations are all smaller than U-Net (Table 4-6).

Table 4-6 Row 2 - 4: Average Uncertainty Scores for all Prostate, Apex, Middle, and Base Slices in PZ and TZ Under the Proposed Method; Row 5 - 7: Average Uncertainty Scores for all Prostate, Apex, Middle, and Base Slices in PZ and TZ Under U-Net.

Zone	All	Apex	Middle	Base
Proposed Method				
PZ - Uncertainty	0.16±0.02	0.14±0.03	0.15±0.04	0.21±0.06
TZ - Uncertainty	0.13±0.05	0.22±0.08	0.13±0.06	0.10±0.05

U-Net				
PZ - Uncertainty	0.26±0.03	0.25±0.04	0.26±0.04	0.31±0.06
TZ - Uncertainty	0.15±0.04	0.26±0.06	0.15±0.04	0.15±0.03

This study still has a few limitations. First, the training time was long due to small batch sizes to extract the global features which also required a large GPU memory. Second, all MR images were acquired without the use of an endorectal coil in the study. This mirrors general clinical use since the use of endorectal coil is decreasing due to patients' preference. Also, studies showed no significant difference for the detection of PCa between MR images acquired with and without the endorectal coil[56,57] due to the increased signal-to-noise ratios (SNRs) and spatial resolution of 3T MRI scanners, compared to 1.5T. We can apply pixel-to-pixel translation techniques such as cycle-GAN to handle the cases with an endorectal coil since the images with the endorectal coil contain large signal variations near the coil. Third, the study considered the slices that contain the prostate, which could potentially reduce the false positives of the non-prostate slices and increase the overall segmentation performance.

4.6 Conclusion

This Chapter proposed a spatial attentive Bayesian deep learning model for the automatic segmentation of prostatic zones with pixel-wise uncertainty estimation. The proposed method is superior to the state-of-art methods on the segmentation of two prostate zones, such as TZ and PZ. Both spatial attention and multiple-scale feature pyramid attention modules had their merits for the prostate zonal segmentation. Also, the overall uncertainties by the Bayesian model

demonstrated different uncertainties between TZ and PZ at three prostate locations (apex, middle and base), which was consistent with the actual model performance evaluated by using internal and external testing data sets.

Chapter 5: Textured-Based Deep Learning in Prostate Cancer Classification with 3T Multiparametric MRI: Comparison with PI-RADS-Based Classification

This chapter described a texture-based deep learning (Textured-DL) for prostate cancer classification on MRI and compared it with the radiologist-based classification and conventional deep learning methods. Sub-analyses of the proposed method on the classification of lesions on different prostate zones, such as peripheral zone and transition zone, and index types (solidary and multi-focal lesions), were performed.

5.1 Introduction

Multi-parametric MRI (mpMRI) has shown the ability to acquire anatomical details to assess the aggressiveness of PCa[34]. Over the last three years, 3T mpMRI has been integrated into guidelines for the diagnosis of prostate cancer (PCa)[93,94]. The current standardized scheme for the interpretation of mpMRI is the Prostate Imaging Reporting and Data System version 2.1 (PI-RADS v2.1)[15]. The PI-RADS scoring system has been widely adopted, and studies have shown increased clinically significant PCa (csPCa) diagnostic accuracy[95–97]. However, PI-RADS requires a high level of expertise and exhibits a significant degree of inter-reader and intra-reader variability[98], likely reflecting inherent ambiguities in the classification scheme. Also, there exist limited abilities to use the PI-RADS suspicious score in assessing the spectrum of cancer[99,100]. In particular, several

studies reported only 15% to 35% were biopsy positive among the PI-RADS 3 lesions when identifying csPCa[101,102].

Image texture analysis[103,104] provides the spatial arrangement of intensities in the image and can be used to quantitatively describe the tumor heterogeneity, which can be the primary feature of csPCa[105]. Automated classification of csPCa using texture analysis[106] may overcome the current challenges associated with PI-RADS but commonly suffers from the tedious designing process, including handcrafted feature engineering, to fully capture the underlying imaging information. Alternatively, with the development of deep learning in medical imaging[5,107,108], convolutional neural networks (CNNs) with the texture analysis[30] may further improve the accuracy of csPCa classification without handcrafted feature engineering.

In this study, we designed a texture-based deep learning (textured-DL) model for the automated prostate cancer classification of the suspicious lesion on MRI. After a lesion was detected and contoured as part of the clinical interpretation, our deep learning model was developed to further improve the classification of csPCa for any positive MRI findings (PI-RADS ≥ 3). The model performance was tested by an independent testing set and compared with the conventional deep learning and PI-RADS-based classification[99,100].

5.2 Materials and Methods

5.2.1 Study population and MRI datasets

With approval from the institutional review board (IRB), this retrospective study was carried out in compliance with the United States Health Insurance Portability and Accountability Act (HIPAA) of 1996. A total of 402 patients who later underwent robotic-assisted laparoscopic prostatectomy (RALP) between October 2010 and June 2018 were enrolled in this study. Detailed

characteristics of the overall patients and tumors are shown in Table 5-1. Pre-operative prostate mpMRI scans were acquired using a standardized protocol based on the recommendation from PI-RADS. Specifically, the MRI protocol included axial T2w turbo spin-echo (TSE) imaging (TR = 3800-5040 ms, TE = 101 ms, FOV = 20 cm, matrix size = 320 × 310, in-plane resolution = 0.6 mm × 0.6 mm, slice thickness = 3 mm) and echo-planar diffusion-weighted imaging (EP-DWI) (TR = 3300-4800 ms, TE = 60-80 ms, FOV = 26 cm × 21 cm, matrix = 160 × 94, in-plane resolution = 1.6mm × 1.6 mm, slice thickness = 3.6 mm). The ADC maps were calculated by using linear least squares curve fitting of pixels (in log scale) in the four diffusion-weighted images against their corresponding b values (0/100/400/800 s/mm^2). Both axial T2w TSE and ADC were used as input to the textured-DL model.

Three fellowship-trained genitourinary (GU) radiologists (each had interpreted 1,000-3,000 prostate mpMRI scans with 10+ years of experience) identified suspicious lesions for PCa on mpMRI. Each suspicious lesion was contoured with an assigned PI-RADS suspicious score by the radiologists. For MRI scans interpreted before the adoption of PI-RADS v2 (2010-2015), an abdominal imaging fellow (in postgraduate year 6) and the fellowship-trained GU radiologist retrospectively reviewed and assigned PI-RADS v2 to each ROI, blinded to the pathological findings and clinical information at the time of the interpretation. Any lesions with PI-RADS ≥ 3 were reported as positive findings.

Table 5-1 Patient and tumor lesion characteristics.

Characteristics	Overall	Train/Validation	Test
Patient Number	402	281	121
csPCa lesions / all lesions	303 / 466	225 / 324	78 / 142
Non-csPCa lesions / all lesions	163 / 466	99 / 324	64 / 142

Age (yr.)		61 (56-66)	61 (56-67)	61 (58-66)
Weight (kg.)		87 (77-95)	87 (77-96)	86 (78-94)
PSA (ng/ml)		8.3 (4.7-8.7)	8.7 (4.7-8.9)	7.4 (4.6-8.7)
Tumor Volume (cm^3)		1.1 (0.3-1.2)	1.1 (0.3-1.2)	0.9 (0.2-1.1)
Gleason score	False Positive	80	49	31
	3+3=6	83	50	33
	3+4=7	168	126	42
	4+3=7	79	57	22
	>7	56	42	14
PI-RADS	3	119	76	43
	4	197	138	59
	5	150	110	40
Prostate zone	PZ	364	255	109
	TZ	99	67	32
	AFS	3	2	1
Focality	Solitary	190	137	53
	Multi-focal	276	187	89

Blinded to MRI, two genitourinary (GU) pathologists (each had interpretated up to 1,000 prostate wholemount histopathologic reports) identified and outlined tumors on whole-mount histopathology (WMHP) following RALP. On each section, individual prostate cancer lesion size, location, and Gleason Score (primary and secondary Gleason grade) were reported. Next, at a separate monthly meeting, a multidisciplinary research team consisting of GU radiologists, GU pathologists, and urologists reviewed each case to match the pathologically detected lesion with its corresponding lesion on mpMRI through visual co-registration. Each lesion detected by mpMRI was defined as a true-positive if it corresponded to the same quadrant (left, right, anterior, or posterior) and level (base, midgland, or apex) as the lesion from WMHP, or a false positive if no corresponding lesions existed on WMHP. Lesions with false-negative findings were lesions from

WMHP that lacked a corresponding lesion on mpMRI. The index tumor was defined as the most extensive tumor area in the surgical specimen, more specifically, the lesion with the highest Gleason Score or the largest diameter when multiple lesions had the same Gleason Score. csPCa was defined as a lesion with GS≥3+4. After the meeting, all csPCa, indolent lesions (GS=3+3), and false positives were retrospectively contoured on T2w and ADC images using OsiriX (Pixmeo SARL, Bernex, Switzerland). ADC images were registered to T2w images by using a non-rigid registration method[109]. Table 5-1 summarizes the overall patient and lesion characteristics, stratified by Gleason score, PI-RADS, prostate zones, and lesion focality.

5.2.2 Texture-based deep learning model

Figure 5-1 showed the overall workflow of the proposed texture-based deep learning (textured-DL) model, consisting of a 3D gray-level co-occurrence matrix (GLCM) extractor and a CNN. As part of the clinical MRI interpretation, a suspicious lesion was identified by the PI-RADS suspicious score. The positive finding (PI-RADS ≥ 3) or a suspicious PCa lesion contoured on T2w and ADC was assumed be an input data format to the proposed model. For a given suspicious PCa lesion, the volumetric patches of T2w and ADC were first cropped based on the lesion segmentation as input to textured-DL. The rectangular patch was selected to closely surround the lesion and normalized to 0-255 prior to the textured-DL model. In the workflow, 3D GLCM descriptors were used to extract the texture features from each volumetric patch (T2w and ADC), and the 3D GLCM features were combined as an input to CNN. The output of textured-DL is the probability of being a csPCa lesion for a given positive MRI finding (PI-RADS≥3). Note that the conventional deep learning approach would exclude the 3D texture extraction (i.e., the rectangular patches are directly used as input to CNN)[110]. We implemented this as the baseline model (CNN).

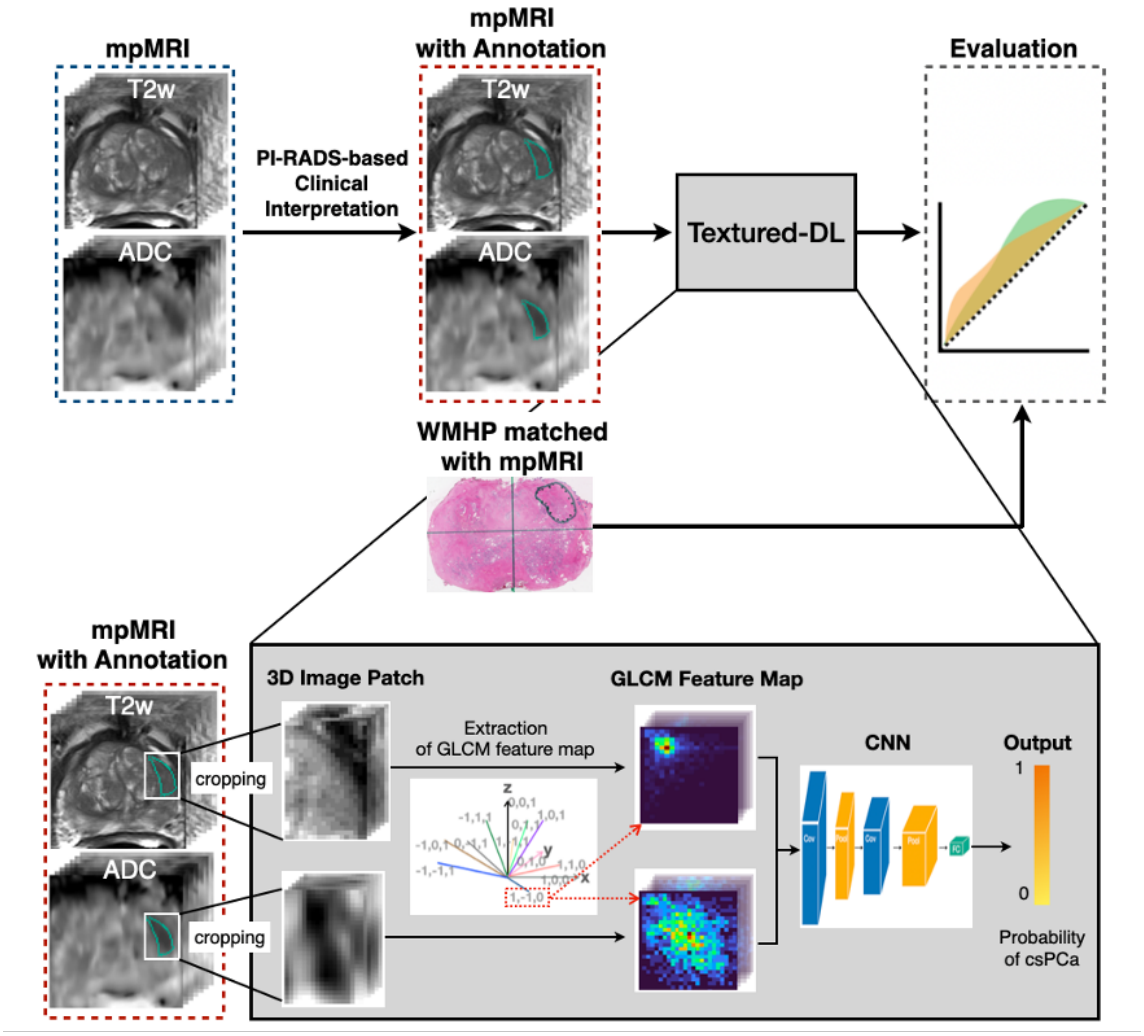

Figure 5-1 Overall workflow of the proposed textured-DL model for the prostate cancer classification. Suspicious lesions were firstly detected by the PI-RADS. Then, PI-RADS scores were assigned to the detected lesions. Lesions with PI-RADS score ≥ 4 were considered as clinically significant prostate cancers (csPCa). After the manual segmentation of the prostate lesion, 3D rectangular patches of the prostate lesion were cropped from the T2w and ADC images, and gray-level co-occurrence matrices (GLCM) were extracted from two patches. Next, the two GLCMs were concatenated and fed into CNN to generate the probability of clinically significant prostate cancer (one being the highest probability of csPCa). ROC curve, sensitivity and specificity were adopted to evaluate and compare the performance of csPCa classification by the PI-RADS and Textured-DL, confirmed by the histopathological findings.

5.2.3 3D GLCM Extractor

Gray-level image was discretized into 64 gray-level bins, yielding the 3D gray-level images whose voxels' values ranged from 1 to 64. Then, we generated the 3D GLCM by calculating the frequency of voxel pairs with different spatial orientations and specific gray-level values. Unlike the 2D GLCM, which only considers the in-plane pixel adjacency, 3D GLCM also considers the through-plane voxel adjacency. For each direction θ, the corresponding GLCM is calculated as follows:

$$C_{d,\theta}(\mathrm{x},\mathrm{y}) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\begin{cases}1 & if\ f(i,j,k) = x\ and\ f(i+di, j+dj, k+dk) = y \\ 0 & Otherwise\end{cases} \quad (5\text{-}1)$$

, where (di, dj, dz) is a displacement between the point (i, j, k) and another point along the direction θ, f is the 3D image data, (i, j, k) is a pixel location in the image f, and $f(i, j, k)$ is the pixel value at (i, j, k).

After 3D GLCMs (size is $13 \times 64 \times 64$ for both modalities) for T2w and ADC were computed, the two feature maps were concatenated (final size is $26 \times 64 \times 64$). Then, the feature maps went through the CNN, which consists of two convolutional layers with kernel sizes of 3×3 and stride of 1, two pooling layers with a filter size of 2×2, and two fully connected layers, to perform the classification of csPCa and non-csPCa. Input and output channel sizes are (26, 32) and (32, 64) for the first and second convolutional layer, respectively. Each convolutional layer was equipped with batch normalization (BatchNorm) and Rectified Linear Unit (ReLU).

5.2.4 Model development and comparison

The patient cohort was randomly split into three sub-datasets, including training (n=239; 60%), validation (n=42; 10%) and testing (n=121; 30%) datasets. The model was trained on the training dataset, and hypermeter tuning, and best model selection were performed on the validation dataset. Weighted cross-entropy was used as the loss function, which was optimized by the Adam optimizer [111] with the default parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). The learning rate was set to 1e-5 with a momentum of 0.9. The model was trained for 200 epochs with a batch size of 10.

We used PI-RADS as an expert reader baseline to compare classification performance with textured-DL. Lesions with PI-RADS larger or equal than four were considered as prediction for csPCa [99,100] in the model comparison. In addition, we included a radiomics-based machine learning workflow (Radiomics-ML)[112] to compare classification performance with texture-DL. To make a fair comparison with our proposed textured-DL, within each prostate lesion, a total of 104 GLCM-based radiomic texture features on both the T2w and ADC were calculated. The radiomic features were then fed into an ensemble learning method, random forest (RF)[113], for the csPCa classification. Furthermore, we also compared our proposed textured-DL with a relatively deep convolutional neural network (DCNN)[114] that was inspired by VGG-Net[115]. Since PI-RADS 3 lesions are variable in the diagnosis of prostate cancer, we conducted the sub-analysis of the PI-RADS 3 lesions compared to PI-RADS 4-5 lesions for the performance of Textured-DL to diagnose the prostate cancer.

In addition, we also performed the sub-analyses on the classification of csPCa lesions on different prostate zones, such as peripheral zone (PZ) and transition zone (TZ), and index types (solidary and multi-focal lesions) between textured-DL and PI-RADS. This is due to that 1) there exist significant differences in morphological appearance and cancer prevalence between tumors in PZ and TZ, and the assignment of the PI-RADS for each lesion utilizes different imaging

sequences according to zonal anatomy[14]. 2) The aggressiveness of the index tumor aggressiveness is clinically important for treatment decisions, pre-biopsy planning, and pre-surgical planning.

5.2.5 Statistical analysis

All models (PI-RADS, CNN, Radiomics-ML, DCNN, and textured-DL) were evaluated on the testing dataset by the area under the ROC (AUC) curve, sensitivity, and specificity. The 95% confidence interval (CI) of the AUC was computed by bootstrapping with 1000 samples. The Wald method[116] was used to calculate 95% CI of the sensitivity and specificity. The model sensitivity and specificity were selected by the Youden index[117]. Statistical significance was defined as a p-value <0.05. DeLong test[118] was used to perform the AUC comparisons between the baseline methods and the proposed textured-DL. P-values for statistical comparisons of sensitivity and specificity were provided by the Mcnemar's test[119].

5.3 Results

5.3.1 Model Performance in Comparison with PI-RADS for All Tumors

Figure 5-2 displayed two representative examples of mpMRI findings matched with WMHP and predictions of the csPCa classification using textured-DL. Figure 5-2 (top) represents the imaging for a 56-year-old man with a serum prostate-specific antigen (PSA) of 12.2 ng/m. A PI-RADS 4 lesion and Gleason score 3+3 were shown on both MRI and WMHP images. The textured-DL predicted the lesion as a non-csPCa while this would have been considered as csPCa when PI-RADS 4 was used as a cutoff. Figure 5-2 (bottom) represents the imaging for a 72-year-old man with a PSA of 8.8 ng/m. A PI-RADS 3 lesion and Gleason score 4+3 were shown on both MRI and WMHP images. Similarly, textured-DL predicted correctly, which would have been missed in PI-RADS-based classification.

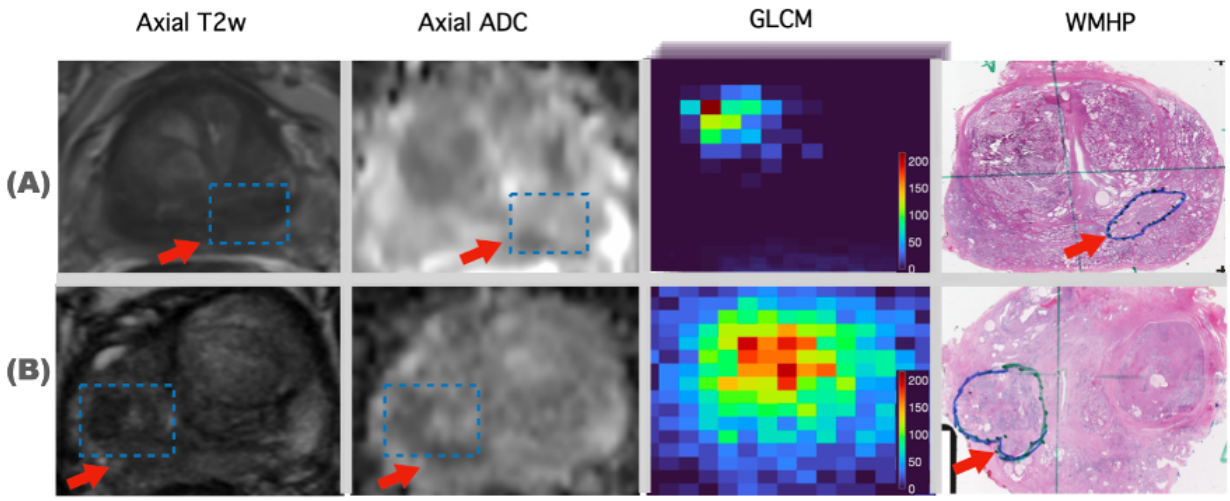

Figure 5-2 Two examples of prostate lesion classification are shown in row (A) and (B), respectively. In each row, from left to right, axial T2w and axial ADC, gray-level co-occurrence matrix (GLCM), and matched whole-mount histopathology (WMHP) are shown. A) Imaging for a man with a 56-year-old man with a PSA of 12.2 ng/m. A lesion (blue rectangular box pointed by a red arrow) with PI-RADS 4 and GS 3+3 was shown on both the axial T2w and ADC images. The proposed textured-DL predicted this lesion as a non-clinically significant lesion. B) Imaging for a 72-year-old man with a PSA of 8.8 ng/m. A lesion (blue rectangular box pointed by a red arrow) with PI-RADS 3 and GS 4+3 was shown on both the axial T2w and ADC images. The proposed textured-DL predicted this lesion as a clinically significant lesion.

Figure 5-3 showed the overall classification performance among textured-DL, baseline CNN, Radiomics-ML, and PI-RADS. For all lesions, textured-DL achieved an AUC of 0.85, significantly higher than CNN (AUC of 0.74; $p<0.01$), radiomics-ML (AUC of 0.78; $p=0.04$), and PI-RADS (AUC of 0.73; $p<0.01$). The textured-DL model also demonstrated the significantly higher sensitivity than CNN, radiomics-ML, and PI-RADS (all p values<0.05), with a comparable specificity to CNN and radiomics-ML (p values are 0.82, 0.76) and a significantly higher specificity than PI-RADS (p value<0.01).

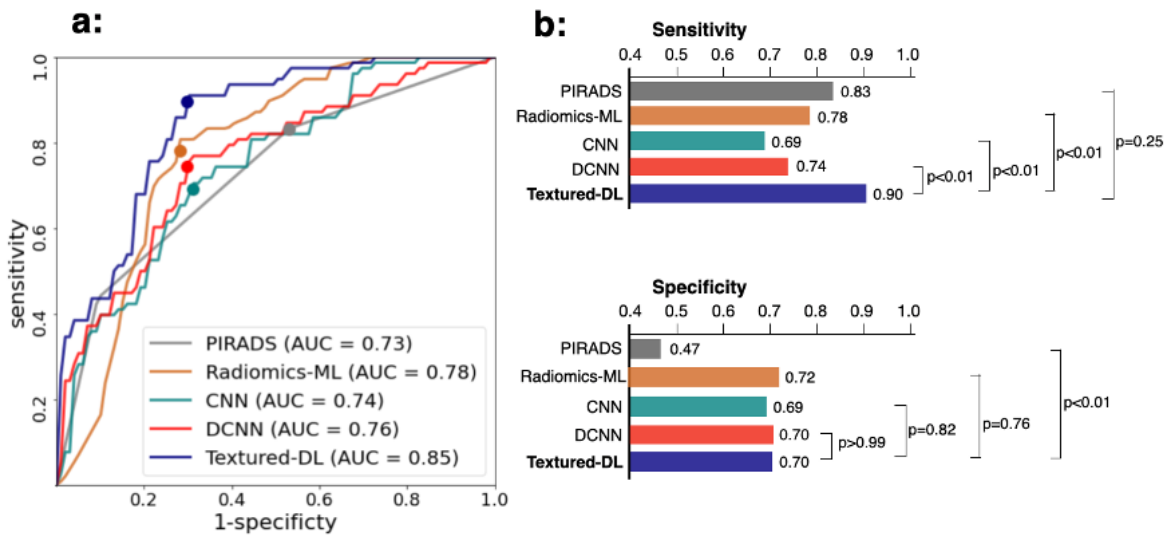

Figure 5-3 Comparisons of ROC, sensitivity, and specificity between PI-RADS, Radiomics ML, conventional CNN, DCNN, and textured-DL on the classification of csPCa in the overall tumor lesions.

5.3.2 Classification Performance for Tumors on Different Prostate Zone

We further conducted the secondary analysis on the different lesion locations, such as peripheral zone (PZ) and transition zone (TZ), for the classification performance using the same model (Figure 5-4). In PZ, we found that our textured-DL achieved an AUC of 0.88, higher than PI-RADS (AUC of 0.72; $p<0.01$). The specificity of textured-DL (0.78) was also significantly higher than that of PI-RADS ($p<0.01$). In TZ, textured-DL achieved a higher AUC than that of the PI-RADS with higher sensitivity and similar specificity.

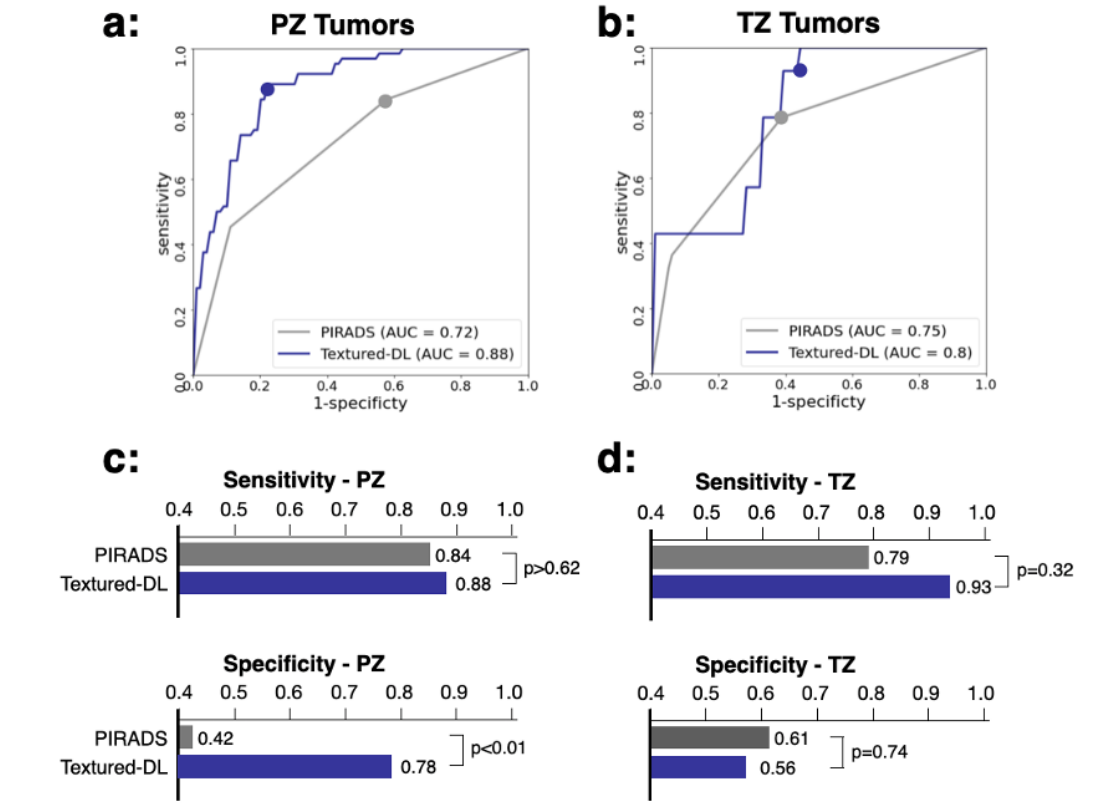

Figure 5-4 Comparisons of ROC, sensitivity, and specificity between PI-RADS and textured-DL in the tumor lesions on different prostate zones, transition, and peripheral zones.

5.3.3 Classification Performance for Solidary and Multi-focal Tumors

Figure 5-5 shows another secondary analysis with different tumor types, solitary and multi-focal tumors. In solitary tumors, textured-DL demonstrated a significantly higher AUC and specificity than those of PI-RADS ($p<0.01$ and $p=0.01$). Similarly, in multi-focal tumors, we observed that the specificity of textured-DL was significantly higher than those of PI-RADS ($p=0.04$). However, the sensitivity was similar between PI-RADS and textured-DL.

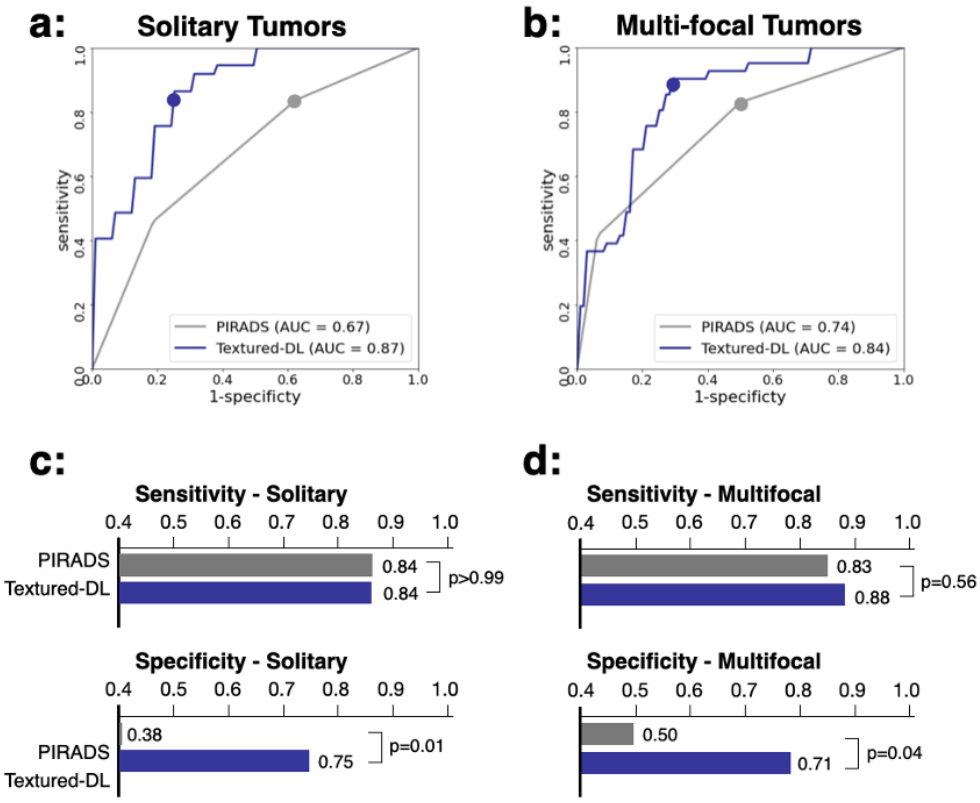

Figure 5-5 Comparisons of ROC, sensitivity, and specificity between PI-RADS and textured-DL in the solitary and multi-focal tumors.

5.3.4 Classification Performance for Tumors of different PI-RADS categories

We also separately compared classification performance between PI-RADS and textured-DL in PI-RADS category 3, 4, and 5 lesions (Table 5-2). We found that the proposed textured-DL achieved consistent classification performance in AUC, sensitivities, and specificities across different PI-RADS-categorized lesions. There exist 13 csPCa lesions in PI-RADS 3 lesions, which would have been missed if a threshold of PI-RADS 4 is used for the classification of csPCa (PI-RADS≥4), while 11 of those correctly classified by textured-DL. Also, there exist 28 non-csPCa in PI-RADS 4 lesions, which would have included as positive lesions if a threshold of PI-RADS 4

is used, and 20 of those were correctly classified by the proposed textured-DL. Four out of six non-csPCa lesions with PI-RADS 5 were also correctly classified by textured-DL.

Table 5-2 Classification performance of textured-DL on the tumor lesions with different PI-RADS categories.

Lesion Type	csPCa (%)	Non csPCa (%)	Method	AUC (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)
PI-RADS 3	13 (30)	30 (70)	Textured-DL	0.81 (0.68-0.94)	0.85 (0.65-1)	0.73 (0.58-0.89)
PI-RADS 4	31 (53)	28 (47)	Textured-DL	0.84 (0.74-0.94)	0.87 (0.75-0.99)	0.71 (0.55-0.88)
PI-RADS 5	34 (85)	6 (15)	Textured-DL	0.84 (0.66-1)	0.88 (0.77-0.99)	0.67 (0.29-1)

5.3.5 Classification Performance for Index Tumors

We further carried out the sub-analysis with the index tumor lesions only (Table 5-3). The index tumors were divided into three groups according to the PSA values ($PSA < 4$, $4 \leq PSA < 10$, and $10 \leq PSA$). For the group of index tumors with PSA<10 (i.e., low-to-average risk group), textured-DL achieved a higher sensitivity in detecting csPCa than PI-RADS, while textured-DL achieved the perfect specificity for the group of index tumors with $PSA \geq 4$ (i.e., average-to-high risk groups).

Table 5-3 Performance comparison between PI-RADS and textured-DL on the classification of the index tumors with different PSA levels.

Tumor Type	csPCa (%)	Non csPCa (%)	Method	AUC (95% CI)	p-value	Sensitivity (95% CI)	p-value	Specificity (95% CI)	p-value
All index	77 (84)	15 (16)	PI-RADS	0.65 (0.52-0.78)	0.32	0.83 (0.75-0.91)	0.25	0.27 (0.04-0.49)	0.18
			Textured-DL	0.73 (0.59-0.88)		0.90 (0.83-0.96)		0.47 (0.21-0.72)	
Index with PSA<4	10 (77)	3 (23)	PI-RADS	0.72 (0.41-1)	0.06	0.7 (0.42-0.98)	0.32	0.67 (0.13-1)	>0.99
			Textured-DL	0.87 (0.58-1)		0.9 (0.71-1)		0.67 (0.13-1)	
Index with 4 ≤PSA<10	53 (87)	8 (13)	PI-RADS	0.56 (0.37-0.75)	0.35	0.87 (0.78-0.96)	0.03	0.12 (0-0.35)	0.56
			Textured-DL	0.43 (0.20-0.65)		0.98 (0.94-1)		0.25 (0-0.55)	
Index with PSA≥10	14 (78)	4 (22)	PI-RADS	0.79 (0.58-0.99)	0.17	0.79 (0.57-1)	>0.99	0.25 (0-0.67)	0.08
			Textured-DL	0.93 (0.80-1)		0.79 (0.31-0.83)		1 (1-1)	

5.4 Discussion

A novel textured-DL method for the automated prostate cancer classification was proposed by combining CNN with the texture analysis[30]. Compared with conventional image texture analysis and deep learning, Textured-DL utilized the spatial arrangement of intensities in the MRI images without handcrafted feature engineering, which can be used to describe the tumor heterogeneity. Furthermore, textured-DL may alleviate the requirement of large datasets for

training as CNN was trained on predefined textured-based features. Compared to conventional CNN and PI-RADS-based classification, textured-DL achieved significantly higher AUC than both methods.

The training and testing for the model were based on the patient cohort who underwent 3T mpMRI prior to radical proctectomy. Although the testing dataset contained the similarly distributed lesions (78 csPCa vs. 64 non-csPCa), the results may not be directly translatable for the biopsy planning patient cohort, including biopsy naïve and prior negative biopsy patients, due to lower rates of csPCa. However, our findings in the PI-RADS-based classification were consistent with the previous multi-center, multi-reader study[120], and the proposed model consistently achieved higher sensitivities and specificities than the PI-RADS-based classification. We believe that the proposed model can be adopted as an additional means to reduce the overdiagnosis of csPCa in conjunction with radiologists. Future studies, including the biopsy planning cohort for model testing, will further solidify our findings.

The clinical significance of PI-RADS 3 lesions is considered to be equivocal. The range of positive biopsy rates in PI-RADS 3 lesions is between 15% and 35%[101,102]. Our method achieved the AUC of 0.81 in differentiating csPCa and non-csPCa among PI-RADS 3 lesions. Of 30 non-csPCa with PI-RADS category 3, 73% were correctly classified by textured-DL, and of 13 csPCa lesions with PI-RADS category 3, 85% were correctly classified by textured-DL. There are still no standardized strategies to predict the risks associated with PI-RADS 3 lesions, but PSA density (PSAD)[102] is commonly used as the reference. Table 5-4 includes a comparison between PSAD-based classification and textured-DL. The textured-DL model performed better than the PSAD-based predictions among PI-RADS 3 lesions by having a high negative predictive value (NPV) while maintaining a high positive predictive value (PPV). This indicated that textured-DL could

potentially serve as an additional tool to predict risks associated with PI-RADS 3 lesions and further to reduce unnecessary biopsies for PI-RADS 3 lesions.

For feature engineering-based workflow, handcrafted radiomic features such as grey-level co-occurrence matrix-based texture features (correlation, contrast, energy, and homogeneity) were firstly extracted. Afterwards, the hand-crafted features were input to the machine learning algorithms such as random forest (RF). However, the handcrafted features do not necessarily represent the whole of the texture within the GLCM. For the textured CNN workflow, instead of computing the hand-crafted features from GLCM, we directly input the GLCM to the CNN. Therefore, textured CNN can get rid of the designing process of handcrafted texture features and, meanwhile, could potentially probe more into the texture within the GLCM.

Deep learning has already been used in the prostate cancer classification. For example, Zhong et.al[110] and Yuan.et al[121] proposed a transfer deep learning workflow for the prostate cancer classification. For these existing CNN workflows, raw image patches must be resized to a fixed size before feeding them into CNN, which could compromise the scale information of the tumors. However, within our proposed textured CNN, GLCM, which is a fix-sized matrix, was input to the CNN. In addition, texture describes the tumor heterogeneity, which can be the primary feature of csPCa. Texture information from the GLCM provided the prior knowledge of prostate cancer to the whole workflow.

This study included a few limitations: 1) the patient cohort was based on an MRI dataset at a single academic center. In the future, model evaluation using multi-center MRI datasets can be conducted to test the generalizability of the proposed model. 2) Our study included T2w and ADC for the model. The inclusion of other MRI sequences/components, such as high b-value DWI, and dynamic contrast-enhanced (DCE) MRI, into the model is expected to further improve the

prostate cancer classification in the future. 3) The number of patients in the independent testing dataset was not large, particularly for all sub-analyses. Although we observed interesting findings in different tumor locations, types, and PI-RADS categories, larger testing datasets would provide further detailed comparisons between PI-RADS and textured-DL. 4) Our study mainly focused on showing the benefit of using a combination of GLCM-based texture information and CNN in the classification of prostate cancer. We believe that other clinical and demographic information, such as PSA, PSA density, age, location of the lesion, patients' inheritance, BMI, etc., can be combined with our model to improve the performance in the future.

5.5 Conclusion

This Chapter proposed a novel texture-based deep learning (textured-DL) method for the automated prostate cancer classification using 3T mpMRI. The proposed textured-DL outperformed PI-RADS in the classification of clinically significant prostate cancer. The textured-DL showed superior performance in specificities for the PZ and solitary tumors, compared with PI-RADS-based classification, and demonstrated a sensitivity of 0.85 and a specificity of 0.73 among the PI-RADS 3 lesions.

Chapter 6: Deep Learning Enables Prostate MRI Segmentation: A Large Cohort Evaluation with Inter-rater Variability Analysis

This chapter describes a large patient cohort evaluation of a previously deep learning method for automated whole prostate gland segmentation. The deep learning method evaluated is primarily based on the one from Chapter 4 but with one improvement by adding the coronal-view segmentation assistance. The large cohort evaluation includes a qualitative, a quantitative assessment, and a volume measurement evaluation.

6.1 Introduction

Whole-prostate gland (WPG) segmentation plays an important role in prostate volume measurement, biopsy, and surgical planning[59]. Magnetic resonance imaging (MRI)-targeted transrectal ultrasound fusion (MRI-fusion) biopsy has shown increased detection of clinically significant PCa and reduced identification of clinically insignificant PCa[60], where the WPG segmentation is critical to enable the MRI-fusion biopsy[2]. Also, prostate volume measurement via WPG segmentation can be used to quantify the progression of benign prostatic hyperplasia [59] and to assist surgical planning[61].

Manual segmentation of WPG is time-consuming and laborious and commonly suffers from inter-rater variability[3], making it inadequate for large-scale applications[19]. Deep learning (DL) [62–65] has increasingly been utilized for the automatic segmentation of WPG. Zhu et al.[66] proposed

a deeply supervised convolutional neural network (CNN) using the convolutional information to segment the prostate from MR images. Cheng et al.[63] developed a DL model with holistically nested networks for prostate segmentation on MRI. Jia et al.[67] proposed an atlas registration and ensemble deep CNN-based prostate segmentation. In addition, attentive DL[68] models were introduced to enhance DL by paying attention to the particular regions of interest in an adaptive way and thus, have achieved better segmentation performance than other DL-based models. However, to the best of our knowledge, the evaluation of these methods was currently limited by relatively small sample size, ranging from tens to hundreds of MRI scans. It is relatively expensive to create large, continuous samples with manual segmentation of WPG, which limits the ability to test the DL models in a clinical setting.

In this paper, we evaluated a previously developed DL-based automatic segmentation model, deep attentive neural network (DANN)[68], using a large, continuous cohort of prostate 3T MRI scans acquired between 2016 and 2020. The WPG segmentation by DANN was evaluated both quantitatively and qualitatively. The quantitative evaluation was performed by using independent testing set with manual segmentation as a ground-truth on a small dataset (n=100). The dice similarity coefficient (DSC)[69] was used to measure the segmentation performance, compared with other baseline DL methods. For qualitative evaluation, the segmentation performance was evaluated by two abdominal radiologists independently via visual grading since the ground-truth manual segmentation was not available for the large cohort (n=3,210). Inter-rater agreement between the two radiologists was evaluated to check the consistency of the visual grading. We further investigated the segmentation on different anatomical locations (i.e., apex, midgland, and base) as a secondary analysis. Finally, we conducted the volume measurement using

DANN-based segmentation on a small cohort (n=50) (DANN-enabled volume measurement) and compared it with the manual volume measurement.

6.2 Materials and methods

6.2.1 MRI Datasets

With approval from the institutional review board (IRB), this retrospective study was carried out in compliance with the United States Health Insurance Portability and Accountability Act (HIPAA) of 1996. After excluding MRI scans with severe artifacts and patients with prior surgery history and Foley catheter, a total of 3,695 MRI scans, acquired on 3 T scanners (Skyra, Prisma, and Vida, Siemens Healthineers, Erlangen, Germany), from January of 2016 to August of 2020, were included in the study. Axial and coronal T2-weighted (T2W) Turbo spin-echo (TSE) images were used. Table 6-1 shows the characteristics of the T2W MRI scan in the study.

Out of 3,695 3T MRI scans, 335 MRI scans (9%) were used as a training set, and the remaining 3,360 (91%) MRI scans were used as a testing set. Training and testing datasets were randomly chosen from the whole dataset. The testing set included a qualitative evaluation set (n=3,210), a quantitative evaluation set (n=100), and a volume measurement evaluation set (n=50). Table 5-2 shows the data characteristics for each dataset. Training, quantitative, and volume measurement evaluation sets required manual prostate contours as the segmentation reference standard. The manual annotation was prepared by an abdominal radiologist (Q.M.) with more than five years of experience in the interpretation of prostate MRI. In the training set, prostate contours were drawn on all axial T2W images from all MRI scans, and on four mid-coronal T2W images (8^{th} to 11^{th} out of twenty slices) from a subset of 100 MRI scans. In the quantitative and volume measurement evaluation sets, prostate contours were drawn on all axial T2W images.

Table 6-1 T2-weighted TSE MRI sequence parameters in the study.

View	Axial	Coronal
Matrix size	320×320	320×320
Flip angle	160^{o}	147^{o}
Resolution	$0.625 \times 0.625 \times 3.6$	$0.625 \times 0.625 \times 3.6$
Field of View (mm^2)	200×200	200×200
Repetition Time (ms)	3000-7480	2880-7200
Echo Time (ms)	97-112	97-109
Number of slices	20	20
Scan Time (s)	200	200

ms: Millisecond; s: second; mm: millimeter;

Table 6-2 Data characteristics in the training, qualitative, and quantitative evaluation.

		Training Dataset	Qualitative Evaluation Dataset	Quantitative Evaluation Dataset	Volume Evaluation Dataset
Number of MRI scans		335	3,210	100	50
Number of patients with Endo-Rectal Coil		3	84	0	0
MRI scans with different vendors	Skyra	295	2,806	93	45

	Prisma	10	145	4	3
	Vida	30	259	3	2

6.2.2 DL-based Whole Prostate Gland Segmentation Model

Figure 6-1 shows the overall workflow of the automatic WPG segmentation with DANN[68]. We added the segmentation on the coronal plane to assist the selection of axial slices, reducing the inference time of segmentation on the axial plane. During the testing, the workflow went through the following steps. First, a $DANN_{cor}$, responsible for segmenting coronal slices, was adopted to segment the prostate on the two-middle coronal images (9^{th} and 10^{th} slices out of twenty slices) for each MRI scan in the entire testing set. The segmented coronal images were used to automatically select the axial T2W images that contained the prostate gland. This would provide proper through-plane coverage of the prostate in the axial slices. Next, $DANN_{ax}$ was used to perform the WPG segmentation on the selected axial T2W images for each MRI scan in all the testing sets.

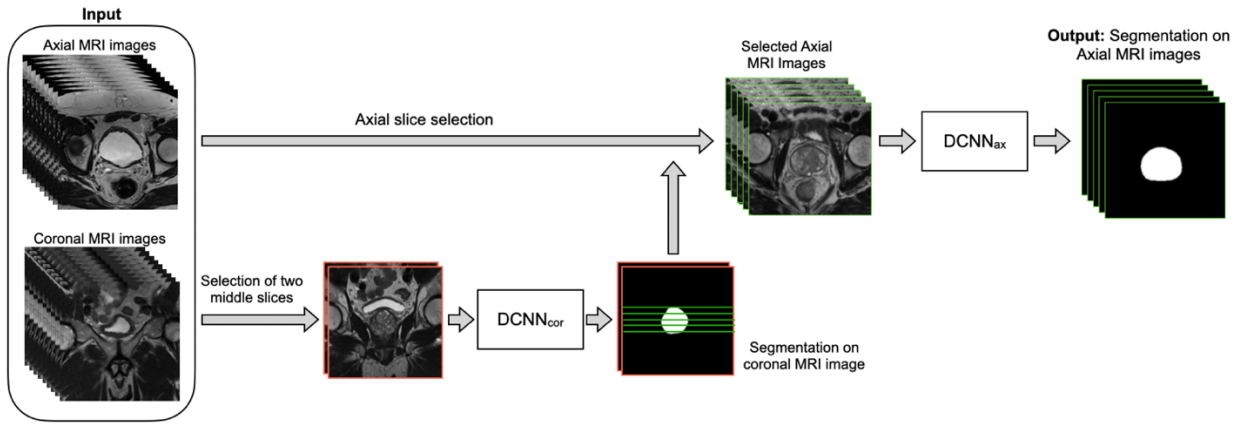

Figure 6-1 The overall workflow of the automatic WPG segmentation with DANN. Both axial and coronal T2W images were used as input, where the coronal images were used to assist the selection of certain axial images containing the prostate gland. $DANN_{cor}$ was firstly performed on the two middle coronal images, indicated by images with the red border. Next, green lines selected by the prostate segmentation on the coronal images were used to determine the selection of axial slices (images with green borders). Once the axial images were selected, $DANN_{ax}$ was performed on the axial MRI slices for the segmentation of WPG.

Both $DANN_{ax}$ and $DANN_{cor}$ were trained independently using the training set (n=335). First, a subset of the training data (n=100) was used for training of $DANN_{cor}$, and four-middle coronal slices (8th to 11th slices out of twenty slices) were used to make use of as many samples as possible. Once the initial training of $DANN_{cor}$ was finished, two middle coronal slices were used as input to $DANN_{cor}$ for the rest of the training data. The segmented coronal slices by $DANN_{cor}$ were used to select certain axial slices, and $DANN_{ax}$ was trained using all the selected axial slices in the entire training set. Training and inferencing were conducted on a desktop computer with a 64-Linux system with 4 Titan Xp GPU of 32 GB GDDR5 RAM. All the networks were trained with stochastic gradient descent as the optimizer, with binary cross-entropy as the loss function. Pytorch was used to implement all the DL networks. The models were initially trained using 80% of the training dataset, and the rest of the training dataset was used as the validation dataset. After the optimal hypermeters were found, we re-trained the models using the whole training dataset. The learning rate was initially set to 2.5e-3. All the networks were trained for 100 epochs with batch size 12.

6.2.3 Evaluation of Segmentation Performance

Table 6-3 Description of each visual grade for qualitative segmentation evaluation.

Score	Visual scoring description
3	The segmentation is excellent. The vast majority (>90%) of the prostate region has been correctly segmented and the percentage of prostate slices with the failure segmentation is less than 10%.
2	The segmentation is acceptable. Most of the region (>70%) is correctly segmented, and the percentage of prostate slices that the method fails to segment is less than 30%.
1	The segmentation is unacceptable. More than 30% of the prostate region has been not correctly segmented or wrongly segmented, and the percentage of prostate slices that the method fails to segment is larger than 30%.

We adopted the visual grading, similar to the literature[70], to qualitatively evaluate the WPG segmentation. Two abdominal radiologists (M.Q. and C.S; each has over five years of experience in prostate MRI interpretation) assigned a visual grade, ranging from 1 to 3, to evaluate the segmentation performance, focusing on the whole prostate and sub-portions of the prostate (e.g., apex, midgland, and base). 1, 2, and 3 indicate unacceptable, acceptable, and excellent segmentation performance, respectively. Table 3 shows the details when assigning the visual grade. Typical examples associated with each visual grade are shown in Figure 6-2. The readers independently ranked the segmentation quality. In addition, inter-rater reliability was assessed. To further investigate the segmentation at sub-portions of the prostate, we performed the sub-analysis for MRI scans without excellent segmentation performance agreed by both radiologists. Also, the segmentation performance for MRI scans with and without endorectal coil (ERC) was compared.

3D DSC[5] was also used to quantitatively evaluate and compare the segmentation performance in the quantitative evaluation set (n=100). The manual segmentations (M) were prepared by the radiologist on all axial slices as ground truths. DSC measures the overlapping between M and method-based (N) segmentation of the WPG volume and can be formulated as:

$$DSC = \frac{2|M \cap N|}{|M| \cup |N|} \quad (5\text{-}1)$$

where ∩ and ∪ indicate the intersection and union, respectively.

We further evaluated the performance of DANN-enabled volume measurements. After the radiologist manually drew the WPG contour on all axial slices, Pyradiomics[71] was used to calculate the prostate volume in the volume measurement evaluation set (n=50). The prostate volume from the DANN-based segmentation was compared with the manual volume measurement. The Bland-Altman plot[72] was used to analyze the agreement between manual and DANN-enabled WPG volume measurements.

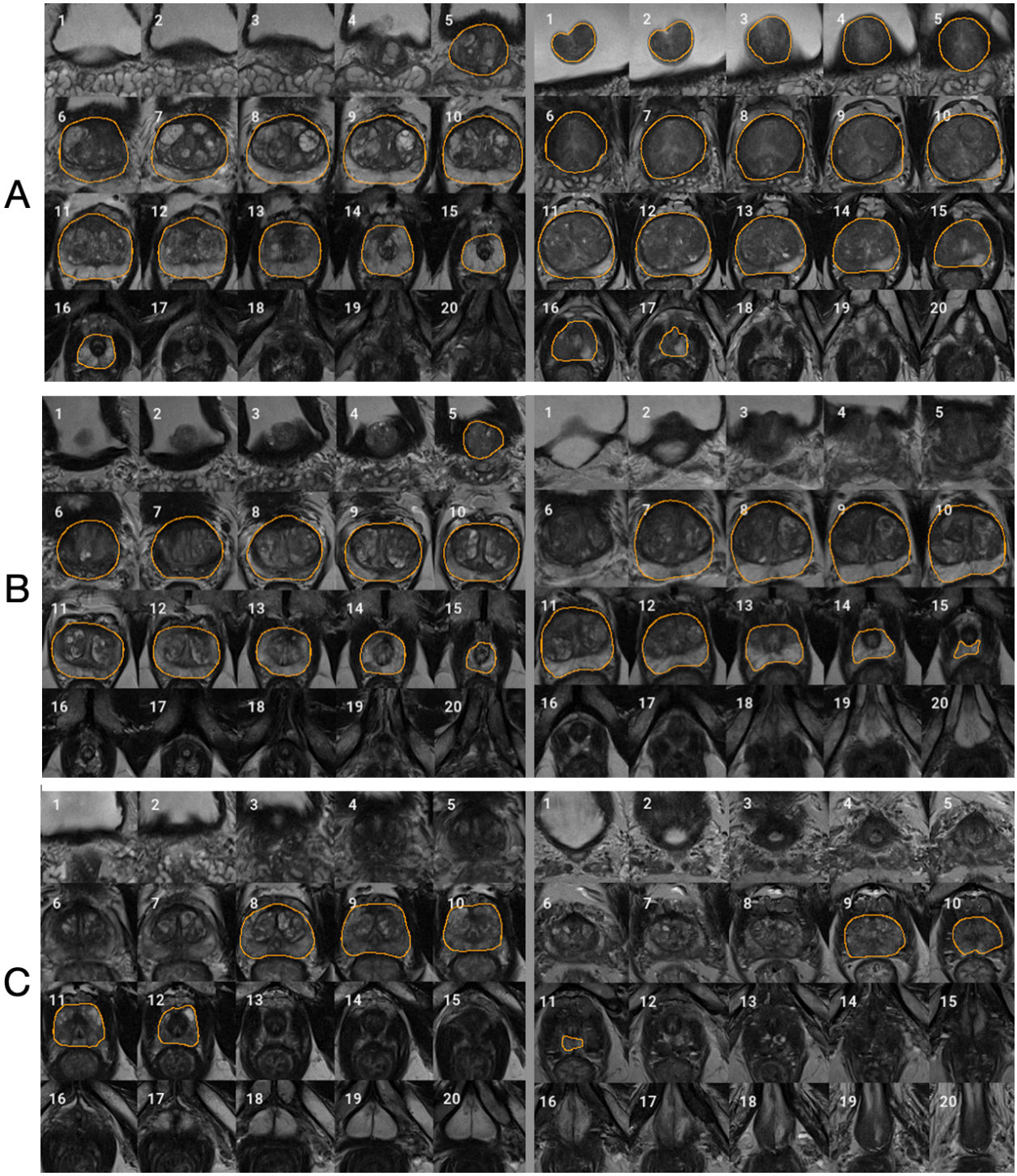

Figure 6-2 Typical examples for each visual grade. Row A, B, and C represent two segmentation examples with visual grades 3 (excellent), 2 (acceptable), and 1 (unacceptable), respectively. Slice 1-20 represents MRI slices from superior to inferior. Regions encircled by organ boundary are the prostate whole gland.

6.2.4 Statistical Analysis

Mean and standard deviation were used to describe the distribution of DSC. The quantitative segmentation performance difference between the DANN and the baselines was compared using a paired sample t-test[73]. P values < 0.05 were considered statistically significant. Inter-rater reliability between two radiologists was measured by using the κ statistic[74]. The relationship between the value of κ and inter-rater reliability is listed as below, $\kappa<0$: pool agreement; $0<\kappa<0.2$: slight agreement; $0.21<\kappa<0.4$: fair agreement; $0.41<\kappa<0.6$: moderate agreement; $0.61<\kappa<0.8$: substantial agreement; $0.81<\kappa<1.0$: almost perfect agreement.

6.3 Result

6.3.1 Qualitative Evaluation of WPG Segmentation

Figure 6-3 shows the proportion of acceptable or excellent segmentation quality in all MRI scans on the qualitative evaluation set at the whole prostate, or each sub-portion (apex, midgland, or base) of the prostate. The DANN method exhibited an acceptable or excellent segmentation performance in more than 96% of the MRI scans on the whole prostate or each sub-portion of the prostate. The segmentation at the midgland portion had achieved the best segmentation performance, while performed the worst at the base portion.

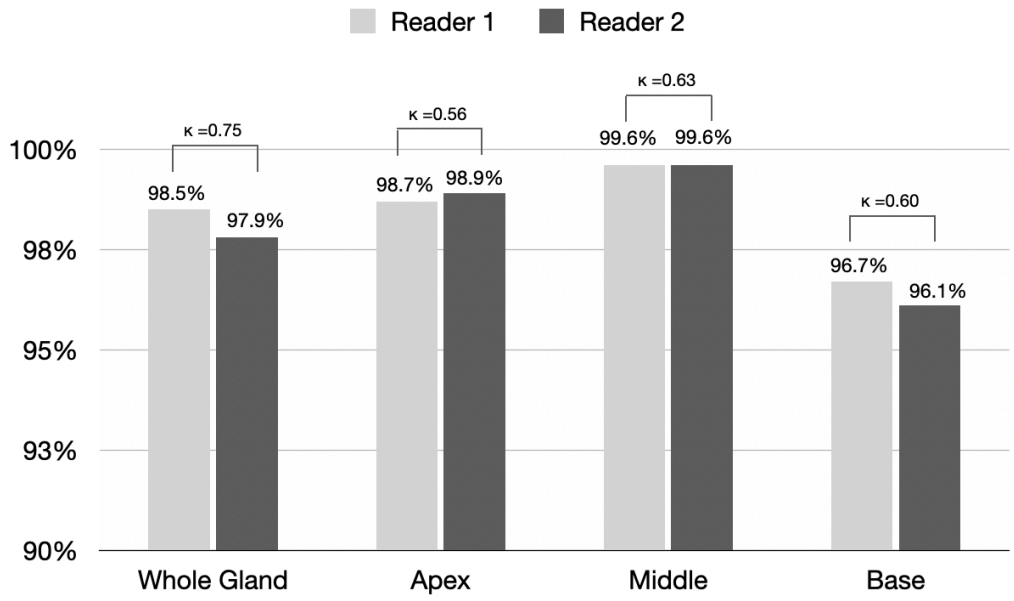

Figure 6-3 The proportion of segmentation with acceptable or excellent performance evaluated by radiologists 1 and 2 among all MRI scans (n=3210). Kappa statistics between the two readers were also provided in the figure.

For WPG segmentation, 97.9% (n=3,141) and 93.2% (n=2,992) of the MRI scans were graded as having acceptable or excellent segmentation performance. Table 6-4 includes the confusion matrix to show the inter-rater variability of the visual grading. Overall, two readers reached a substantial consensus on the visual grading in 95.8% of the patients (κ =0.74). When readers differed on the grading, the discrepancy in grading was less than one. 94.6% of segmentation results were unanimously considered as acceptable or excellent. Moreover, 91.5% of the MRI scans (n=2,861) were graded as having excellent segmentation performance according to the two radiologists. Unacceptable segmentation performance occurred only in 1.2% of the MRI scans (n=39), agreed by the two radiologists.

Table 6-4 Confusion matrices between the visual grades assigned by two readers. Kappa coefficient (κ) is used to measure the inter-rater variability between the two readers.

All	Reader 2				Kappa (κ)
Reader 1	Visual grade	1	2	3	
	1	47 (1.5)	1 (0.0)	0 (0.0)	Substantial agreement (κ =0.75)
	2	22 (0.7)	99 (3.1)	49 (1.5)	
	3	0 (0.0)	63 (2.0)	2,929 (91.3)	

We conducted the sub-analysis related to each sub-portion of the prostate (apex, midgland, or base) when the WPG segmentation was not excellent. The MRI scans with excellent segmentation agreed by two readers were excluded (n=2,929), and the rest of the MRI scans were used for the analysis (n=281). Figure 6-4 shows the confusion matrices of each sub-portion of the prostate on the rest of the MRI scans. 46.3% of the MRI scans achieved the acceptable (or better) segmentation quality at the base slices, while 94.3% and 83.3% of the MRI scans achieved the acceptable (or better) segmentation quality at the midgland and apex slices.

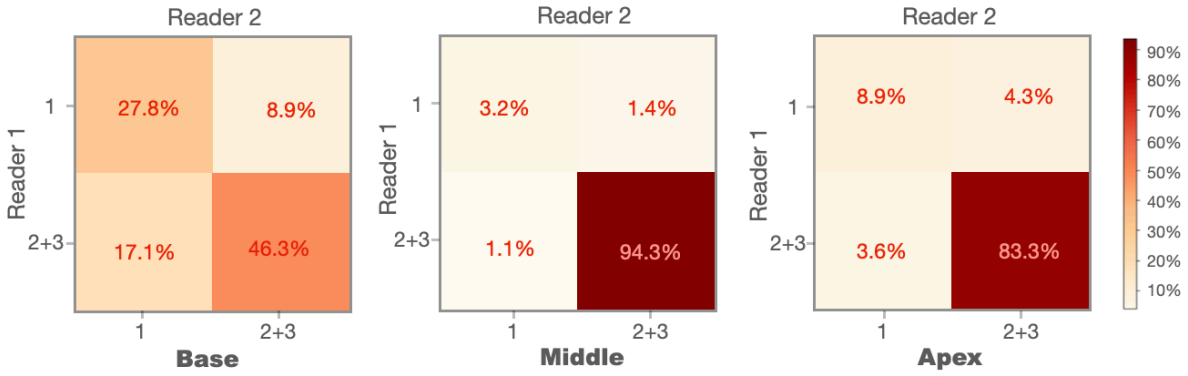

Figure 6-4 Confusion matrices of the prostate base, mid-gland, and apex for the cases without excellent segmentation (n=281).

We compared the WPG segmentation quality for the MRI scans with and without ERC[75]. Figure 6-5 shows the confusion matrices of the visual grades of segmentation on MRI scans with and without ERC. There were substantial agreements (κ =0.64 and 0.85) between the two radiologists on WPG segmentation of MRI scans with and without ERC. When considering the inter-rater agreement of WPG segmentation, DANN demonstrated acceptable WPG performance in more than 95.5% of MRI scans with ERC compared to 84.3% of those without ERC. MRI scans with ERC had a larger proportion of unacceptable WPG segmentation compared to those without ERC (12.1% vs. 2.2%).

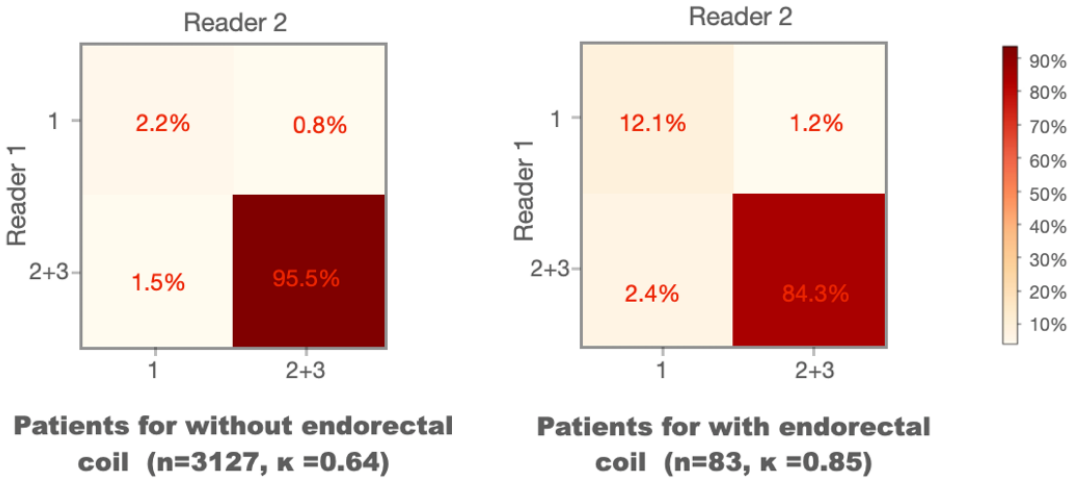

Figure 6-5 Confusion matrices of the visual grades of segmentation on MRI scans with and without endo-rectal coils. Kappa coefficient (κ) is used to measure the inter-rater variability between the two readers.

6.3.2 Quantitative Evaluation of WPG Segmentation

The quantitative performance of the DANN was compared to the other two baseline methods, including Deeplab v3+[53] and UNet[12]. Table 6-5 shows the comparisons of DSCs between DANN and the baseline methods. The DANN achieved a DSC of 0.93, which was higher than those of Deeplab v3+ and UNet with significant differences (both p values < 0.05).

Table 6-5 Quantitative DSC comparisons with baseline methods

Methods	DSC
Proposed Method	0.93±0.02
Deeplab v3+	0.92±0.02 P<0.05
UNet	0.91±0.03 P<0.05

6.3.3 Evaluation of Volume Measurement

Figure 6-6 shows the agreement between manual and DANN-enabled volume measurements in the Bland-Altman plot. The mean difference between the two-volume measurements was calculated as an estimated bias. Standard deviation (SD) of the differences and 95% limits of agreement (average difference ± 1.96 SD) were calculated to assess the random fluctuations around this mean. 48 out of 50 cases (96%) had the volume measurement differences within 95% limits of agreement, indicating that the manual and DANN-enabled volume measurements can be potentially used interchangeably.

Table 6-6 Inference time estimation and DSCs obtained with and without coronal segmentation assistance

	Without coronal segmentation assistance	With coronal segmentation assistance
Overall inference time estimation in the qualitative evaluation	16.4 minutes (67,775)	12.6 minutes (45,713)

DSCs obtained in the quantitative evaluation	0.93	0.93

( ) indicates the total amount of MRI slices the method needed to segment.

6.4 Discussion

A deep attentive neural network[68], DANN, for the automatic WPG segmentation was evaluated on a large, continuous patient cohort. In the qualitative evaluation, DANN demonstrated that the segmentation quality is either acceptable or excellent in most cases. Two radiologists exhibited a substantial agreement for the qualitative evaluation. In the quantitative evaluation, DANN exhibited a significantly higher DSC than the baseline methods, such as UNet and Deeplab v3+. Also, 96% of the testing cohort had volume measurement differences within 95% limits of agreement.

We found that DANN demonstrated worse segmentation performance at the prostate base than at the apex and midgland slices. This may be due to the fact that the anatomical structure of the prostate base is relatively more complex than other prostate portions. The prostate base is in continuity with the bladder and seminal vesicles, and thus the boundary may contain partial volume effects and mild movement artifacts.

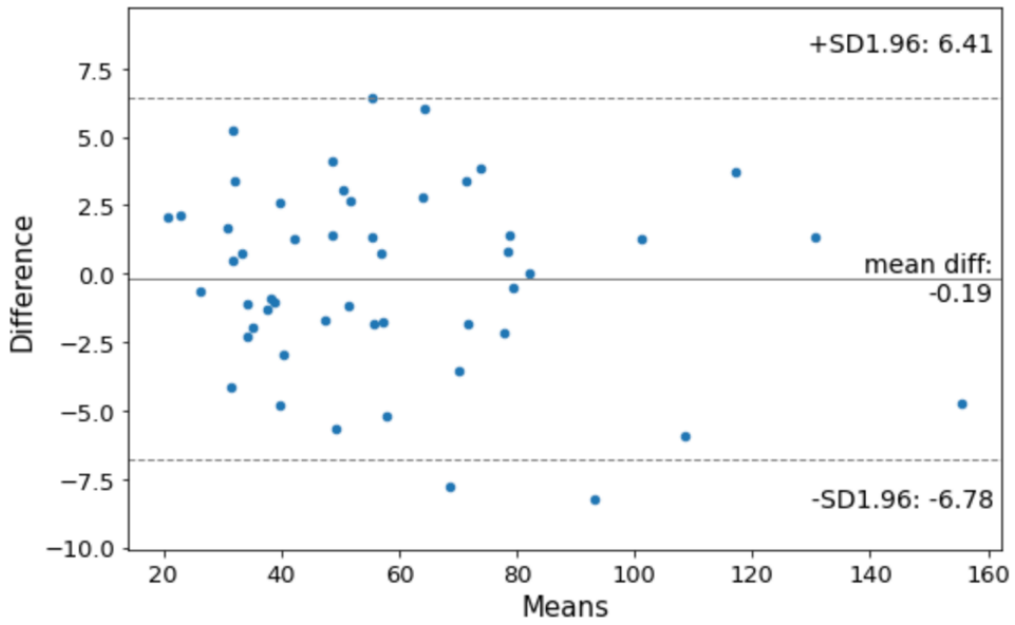

Figure 6-6 Bland–Altman plot to show the agreement between manual and DANN-enabled WPG volume measurements.

We observed that the segmentation performance was somewhat limited when MRI scans were acquired with an ERC. We believe that this may be because there were only three MRI scans with ERC in the training dataset. A large training data with ERC may allow the model to learn representative features related to the prostate MRI with ERC. In addition, images often exhibit large intensity variation compared to the MRI scans without ERC as ERC is close to the prostate. This may require including an even larger training dataset to account for these intensity variations than those without ERC.

We refined DANN by adding the coronal segmentation to assist the selection of axial slices for WPG segmentation. With assistance from the coronal segmentation, the axial model conducted the segmentation only on the selected axial slices instead of applying it to all axial slices, which reduces the inference time. Table 6-6 contains the inference time between the segmentation with and without coronal segmentation. The total inference time in a combination of coronal and axial

slices was 25% less than the inference time without assisting the selection of axial slices (12.6 min vs. 16.4 min). In addition, we observed that DSC was not different when adding the coronal segmentation in the quantitative evaluation.

Compared with quantitative evaluation, qualitative evaluation includes unique characteristics and benefits. The DSC-based evaluation often overlooks the segmentation performance on small regions when they were combined with larger regions. Prostate at apex or base slices is relatively smaller than the one in the middle, and therefore, the quantitative evaluation may not be sensitive enough to illustrate limitations at these locations when 3D DSC is used for the evaluation. Also, the DSC-based evaluation is not directly associated with clinical implications, while qualitative evaluation categorized the segmentation results based on the quality to which segmentation can be acceptable clinically.

Our study still has a few limitations: 1) the MRI scans in this study were acquired from three 3T MRI scanners at a single medical center. Prostate MRI sequence parameters are generally well-standardized by the Prostate Imaging–Reporting and Data System (PI-RADS) guidelines[15], but future studies would include testing DANN at multiple institutions. 2) the inter-rater variability was tested between two radiologists. We will include more radiologists to evaluate comprehensive inter-rater variability. 3) large GPU memory was required during the training and testing since DANN included the spatial attention mechanism that caused considerable computational complexity. In the future, we will explore the criss-cross attention module[76] that uses the contextual information of all the pixels on the criss-cross path for each pixel, which has shown the potential to reduce the GPU memory.

6.5 Conclusion

The proposed deep learning-based prostate segmentation (DANN) could generate segmentation of the prostate with sufficient quality in a consistent manner when a large, continuous cohort of prostate MRI scans was used for evaluation. The qualitative evaluation conducted by two abdominal radiologists showed that 95% of the segmentation results were either acceptable or excellent with a great inter-rater agreement. In the quantitative evaluation, DANN was superior to the state-of-art deep learning methods, and the difference between manual and DANN-enabled volume measurements was subtle in most cases. The method has a great potential to serve as a tool to assist prostate volume measurements, and biopsy and surgical planning in a clinically relevant setting.

Chapter 7: Evaluation of Spatial Attentive Deep Learning for Automated Placental Segmentation on Longitudinal MRI

To investigate the deep learning's generalizability for other biomedical image applications except prostate segmentation, this chapter describes an end-to-end deep learning-based segmentation method, spatial attention deep learning method (SADL), for automated placental segmentation on the longitudinal MRI. SADL improves the deep learning method from Chapter 4 by adding the criss-cross spatial attention that could relieve the issue of large GPU memory required for conventional spatial attention. SADL-based automated volume measurement is also evaluated by comparing it with the manual volume measurement.

7.1 Introduction

The placenta is a critical intrauterine organ necessary for the maintenance of pregnancy [77]. Abnormal placental development can adversely affect maternal health and interfere with nutrient and oxygen transport to the developing fetus. Collectively, aberrant placental development contributes toward perinatal morbidity and mortality through the development of preeclampsia (PE) in the mother with or without fetal growth restriction (FGR)[77,78]. Magnetic resonance imaging (MRI), a reliable imaging modality that offers significantly higher resolution than ultrasound, has previously been studied in the detection of placental dysfunction related to placental volume and

perfusion[78,79]. Segmentation of the placenta by MRI is the critical first step required toward accuracy in the detection of volumetric abnormalities that can affect maternal and fetal health[3,4]. Manual segmentation of the placenta involves comprehensive delineation of multiple 2D MRI placental slices that contribute toward the ultimate construction of the entire placenta volumetrically. This process is time-consuming and is wrought with significant inter-and intra-individual reader variability[6]. Automated segmentation enables the process of rapid tissue segmentation and overcomes subjectivity offered by the inter- and intra-user variability observed with manual segmentation.

In recent years, machine learning and deep learning (DL) have demonstrated superior capabilities in the medical image segmentation[5,12,68,77,80–87]. For example, Wang et al.[46] developed an online learning-based placental segmentation method in MRI images[88]. Alansary et al.[86] implemented a 3D multi-scale convolutional neural network (CNN) with 3D dense conditional random fields for placental segmentation. Wang et al.[80,87] developed a DL-based interactive framework that integrated user interaction with CNN for placenta MRI segmentation. Han et al.[81] and Shahedi et.al[77] evaluated the U-Net variants for the placental segmentation in MRI. However, these studies were either evaluated in a single and late gestational age (GA) or incapable of providing fully automated workflows for placental segmentation. The placenta is constantly evolving and growing throughout pregnancy. An automated segmentation model working in the case of late gestation pregnancy MRI may not transfer effectively for use during early pregnancy MRI. In addition, a single placental MRI scan could comprise a multitude of image slices, that have to be appropriately managed in the proper sequential order. Manual operations and interactions such as selecting image slices containing the placenta could contribute toward significant subjectivity and increase the cognitive workload on experts.

In this study, we aimed to evaluate an end-to-end, fully automated segmentation workflow, spatial attention deep learning method (SADL), for placental segmentation using MRIs obtained during early pregnancy. Two temporal measurements were used (14-18 weeks and 19-24 weeks of gestation), and the placenta segmentation with SADL was evaluated by comparing the segmentation performance with the state-of-the-art DL-based method, U-Net. We further compared placenta volume measurements obtained by manual and automated volume assessments.

7.2 Materials and Methods

7.2.1 Subject Population and MRI Dataset

This study was carried out according to the United States Health Insurance Portability and Accountability Act (HIPAA) of 1996 with approval from the institutional review board (IRB), and all subjects provided written informed consent. We approached all eligible pregnant women entering prenatal care in the first trimester of pregnancy at the local antenatal clinic without pre-selection. Inclusion criteria were a gestational age of fewer than 14 weeks, age more than 18 years old, pregnancy with a single fetus, the absence of fetal chromosomal or structural abnormalities, no treatment with aspirin, heparin, or antihypertensive drugs before enrollment, the ability to provide consent, a non-smoker, and planning to deliver at the same local institution. Exclusion criteria included abortion (spontaneous or planned termination), loss of follow-up, withdrawal from the study, and a history of diabetes mellitus. A total of 154 pregnant women who completed two MRI scans during the second trimester were recruited between 2016 and 2019. The longitudinal MRI scans were acquired at 14-18 weeks (first MRI) and 19 to 24 weeks (second MRI) gestational age. Gestational age was confirmed by a dating ultrasound scan in the first trimester of pregnancy. The summary of the study subjects' characteristics is shown in Table 7-1.

Table 7-1 Summary of the characteristics of the subjects with pregnancies.

		Total	Training	Validation	Testing
No. of Patients		154	108	15	31
Age, yr. (IQR)		32.9[1] (30-35)[2]	33 (30-35)	33 (30-35)	32 (30-34)
Weight, kg. (IQR)		67.2 (58.5-73.0)	67.1 (58.3-72.7)	67.0 (58.3-72.1)	67.5 (59.1-74.5)
GA[3] at the first MRI, weeks (IQR)		15.7 (15.0-16.3)	15.6 (14.9-16.3)	15.8 (15.3-16.5)	15.8 (15.1-16.0)
GA at the second MRI, weeks (IQR)		20.7 (19.9-21.3)	20.7 (19.9-21.3)	20.6 (19.9-20.8)	20.9 (20.1-21.3)
MRI scans	No. of first MRIs	154	108	15	31
	No. of second MRIs	154	108	15	31
No. of MRI slices		42,553	29,896	4,162	8,495

[1]Mean; [2]Interquartile range; [3]Gestational age

A T2-weighted Half-Fourier Single Shot Turbo Spin Echo (T2 HASTE)[89] sequence was used to acquire placental MRI on one of the two 3.0 T MRI scanners (Prisma and Skyra; Siemens Healthcare, Erlangen, Germany). Detailed sequence parameters for T2-HASTE are described in Table 7-2. The T2 HASTE MRI images were acquired in three orthogonal imaging planes (axial, sagittal, and coronal). The image analysis was performed using the open-source image analysis software OsiriX MD software package (Pixmeo SARL, Geneva, Switzerland). 3D regions of interest (ROIs) were first manually drawn in each imaging plane to cover the entire placenta as

ground truth. A clinical fellow supervised by the MRI scientist manually defined ROIs on each of the T2-HASTE MRI slices. The clinical fellow was also supervised by an obstetrician-gynecologist who is a specialist in maternal and fetal medicine whenever the placental anatomy was considered challenging to segment. Figure 7-1 shows a representative example of a placental MRI image with manual placental segmentation in the three planes. The placenta volumes measured by three orthogonal imaging planes were averaged to minimize potential error due to low through-plane resolution on T2 HASTE MRI images.

Table 7-2 Detailed T2-HASTE MRI sequence parameters.

Parameter	**Value**
TE/TR (msec)	92 / 3000
Flip Angle (degree)	150
Bandwidth (Hz/pixel)	390
Resolution ($mm_x \times mm_y$)	0.976 × 0.976
Slice Thickness (mm)	5
Echo Train Length	70
Matrix ($N_x \times N_y$)	272 × 512
Field of View ($mm_x \times mm_y$)	265 × 500

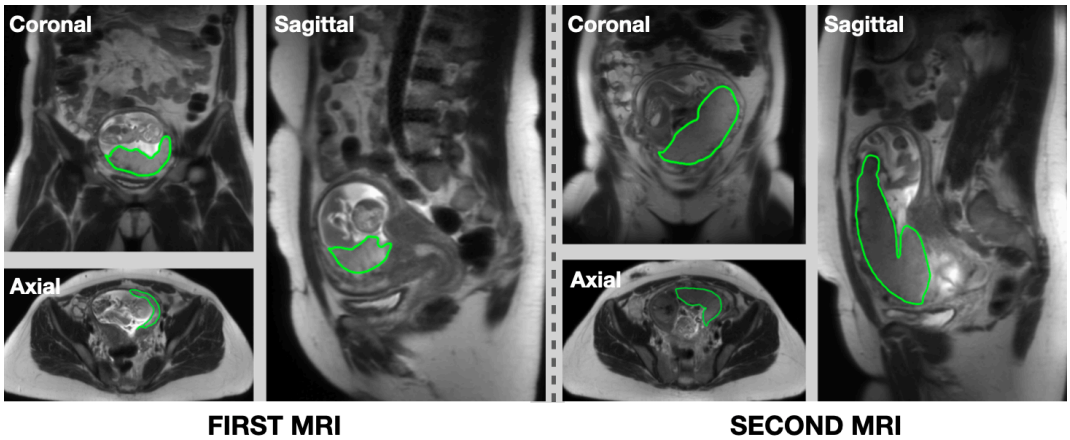

Figure 7-1 Representative placental MRI images in three imaging planes at the first MRI at 15.3 weeks (left; volume = 119cm^3) and second MRI at 21.3 weeks (right; volume = 270cm^3). The placenta was manually contoured and shown as the green line.

7.2.2 Proposed Spatial Attentive Deep Learning

The structure of the proposed SADL is shown in Figure 7-2. The whole network is comprised of a spatial attentive deep residual network (based on the ResNet50)[42] as the encoder, a feature pyramid attention module[13] to enhance capturing of multi-scaled information, and a naïve decoder network to recover spatial resolution. Inside the encoder, the modifications were twofold. 1) Criss-cross (CC) spatial attention module[76] was added at the beginning of the ResNet50, which helped the network emphasize areas with more semantic features of the placenta by modeling spatial dependent information via the global features. Specifically, each pixel's response was obtained by considering all the pixels so that more importance was adaptively given to pixels with more semantic information. 2) MaxPool was removed from the original ResNet50. Several studies[5] have proved that the inclusion of MaxPool compromised image segmentation performance. Next, a feature pyramid attention (FPA) network was added after the encoder, thereby furthering the enhancement of multi-scaled feature extraction. Finally, a naïve decoder was connected to the FPA to recover the spatial resolution.

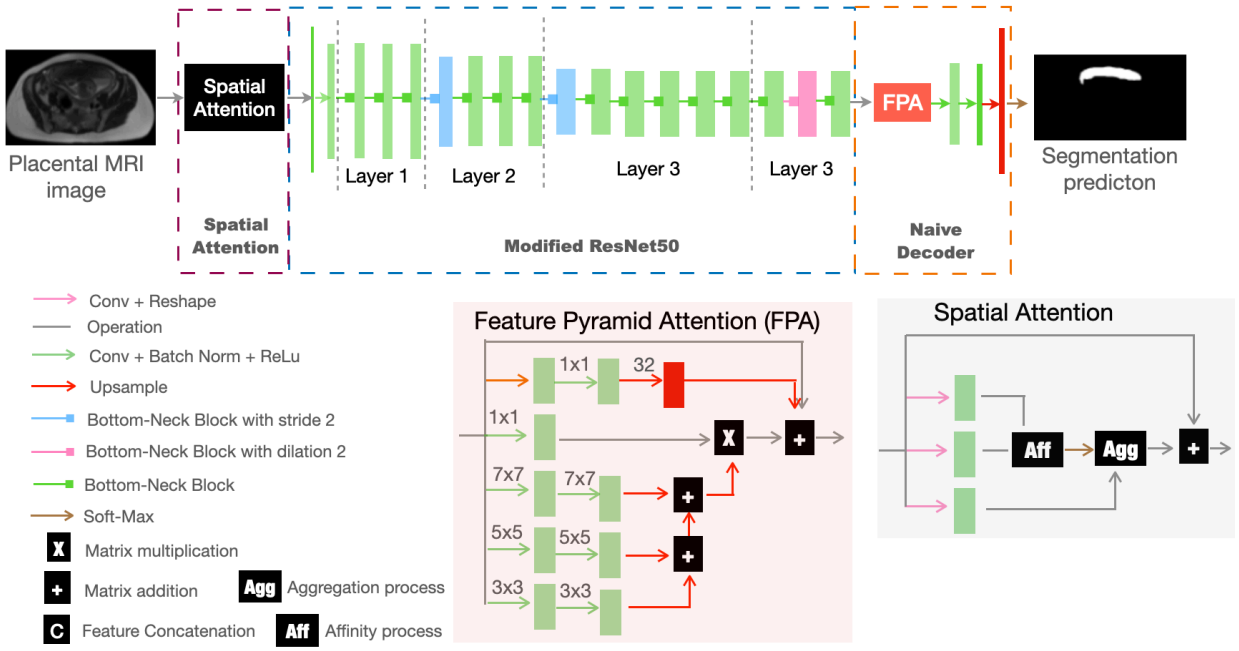

Figure 7-2 An overall structure of the proposed SPDL network. The network consists of 4 sub-networks: a spatial attention module, an improved attentive ResNet50, a feature pyramid attention, and a naïve decoder. The input and output are a 2D placental MRI slice and a placental segmentation prediction. Aggregation and affinity processes were defined in the literature[76].

7.2.3 Experimental Setups – Training and Testing

All deep learning models were implemented using Pytorch, and the volume was calculated using Pyradiomics[90]. We divided the study cohort into training (n=108; 70%), validation (n=15; 10%), and testing (n=31; 20%) sets. We used stochastic gradient descent (SGD) as an optimizer and a binary cross-entropy as the loss function for the deep learning model training. The network was trained for 200 epochs and the model with the lowest validation loss was selected as the optimal model for placenta segmentation. Finally, the optimal model was tested using the testing set of MRIs. The image slices in the three views were cropped by a matrix size of 256 × 256 in the central region. The bounding box contained all placental structures in the images obtained from each view. In-fly data augmentation techniques included random rotations between [−5°, 5°],

elastic transformations, random contrast adjustment, and random horizontal flip. We also performed the image normalization to reduce skewing. Each placental MRI scan contained three imaging planes (axial, coronal, and sagittal). To make the model capable of segmenting placental MRI images in different imaging planes (axial, coronal, and sagittal), we included the images from all three views to train the model. All slices, including those with and without the placenta, were fed into the models to help the network learns the placenta span.

7.2.4 Evaluation metrics and Statistical Analysis

3D Dice Similarity Coefficient (DSC)[69] was used to measure the segmentation performance in the testing set, formulated as:

$$DSC = \frac{2|A \cap B|}{|A \cup B|} \quad (7\text{-}1)$$

where A and B are automated and manual segmentation of 3D placental, respectively. DSC of each MRI scan was calculated by averaging the DSCs from the three orthogonal imaging planes (axial, sagittal, and coronal). We further evaluated the automated volume measurement by SADL, compared to the manual volume measurement. The Bland–Altman plot[72] was used to analyze the agreement between manual and automatic placental volume measurements. The significant difference between DSC obtained using SADL and the baseline method was investigated using a paired sample t-test at the 95% level of confidence.

7.3 Results

Figure 7-3 shows a representative example of automated placenta segmentation by SADL and U-Net. Table 7-3 shows the DSC comparison between SADL and U-Net in the testing dataset.

In the first and second MRI, SADL achieved the average DSCs of 0.83 ± 0.06 and 0.84 ± 0.05, which were significantly higher than those of U-Net (both p-values <0.05). We also found that SADL performed similarly between the first and second MRI (0.83 ± 0.06 vs. 0.84 ± 0.05). Representative examples of excellent and poor placental segmentation of MRI images using SADL are shown in Figure 7-4.

Table 7-3 DSC comparisons between SADL and U-Net in the testing dataset.

Methods	DSC	
SADL	0.83 ± 0.06	
	First MRI	Second MRI
	0.83 ± 0.06	0.84 ± 0.05
U-Net	DSC	
	0.76 ± 0.09 p<0.05	
	First MRI	Second MRI
	0.77 ± 0.08 p<0.05	0.76 ± 0.10 p<0.05

P values are the comparisons between the SADL and the U-Net.

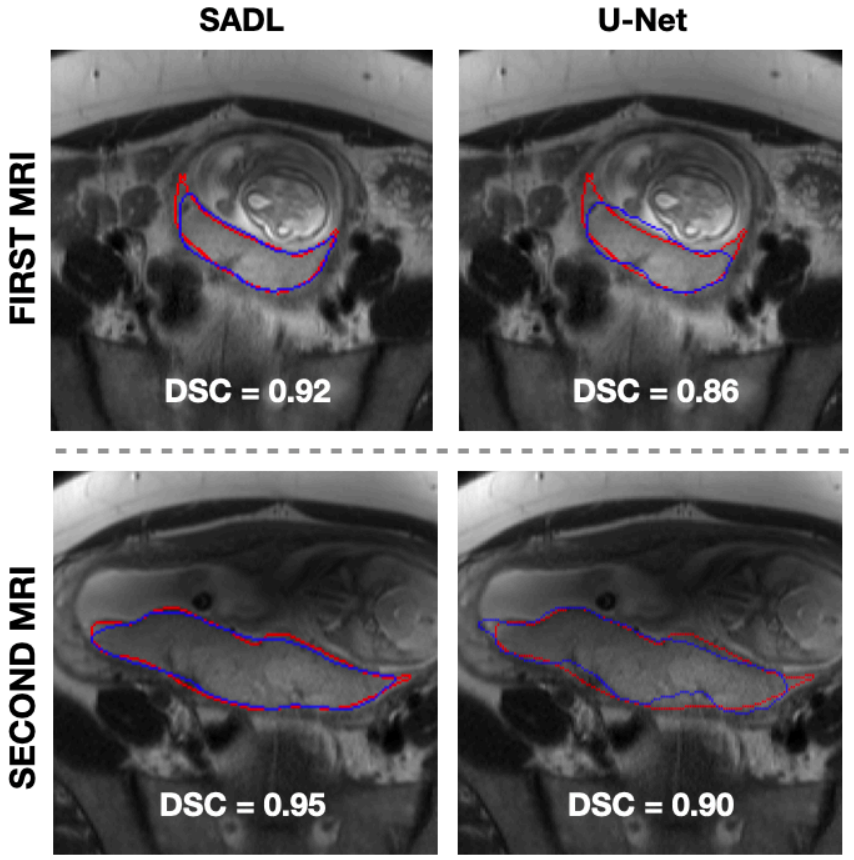

Figure 7-3 Representative example of automated segmentation by SADL and U-Net (blue lines) compared to the manual segmentation (red lines) at the first MRI (*GA = 15 weeks and 1 day*) and the second MRI (*GA = 19 weeks and 4 days*). DSCs are shown below.

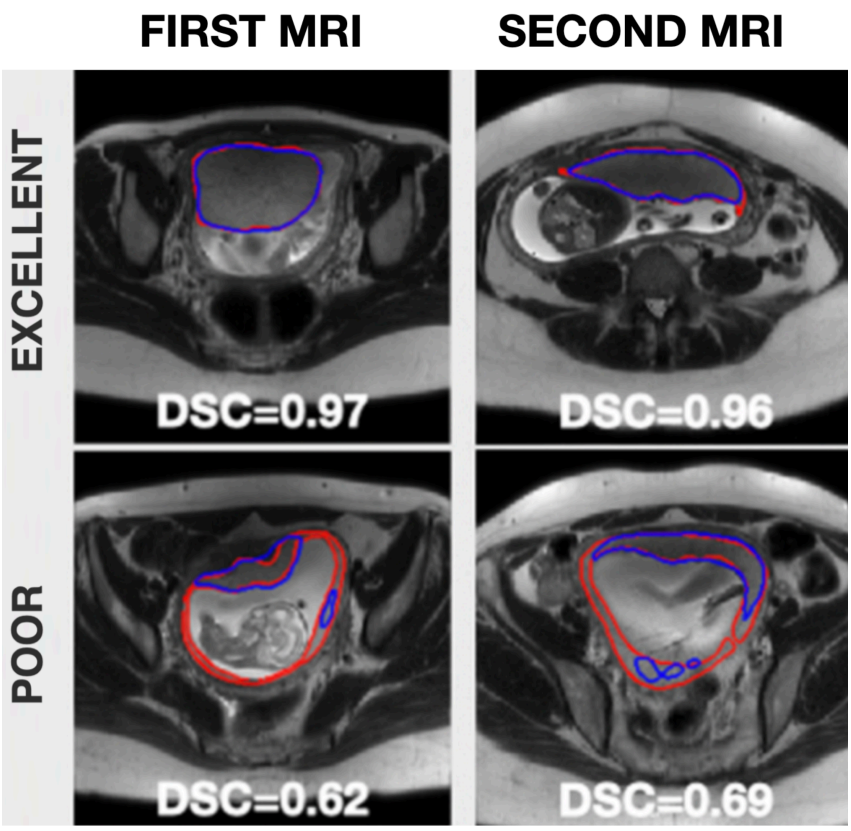

Figure 7-4 Representative examples of excellent and poor automated placental segmentation at the first MRI scan (GA between 14-18 weeks) and the second MRI scan (GA between 14-18 weeks) by the proposed method. Red and blue lines are manual and automated segmentation.

Table 7-4 shows the DSC by SADL and U-Net across the three different orthogonal planes. At the first MRI, SADL achieved similar DSCs across the three orthogonal imaging planes (0.83, 0.83, and 0.82 for axial, coronal, and sagittal planes). However, at the second MRI, the SADL achieved a DSC of 0.87 ± 0.03 at the axial plane, which was slightly higher than those of other imaging planes (0.82 ± 0.10 and 0.83 ± 0.07 for coronal and sagittal planes). In addition, we found SADL outperformed U-Net across each imaging plane at either MRIs.

Table 7-4 Segmentation Performance of SADL in the three orthogonal views in the testing dataset.

Methods		Ax	Cor	Sag

SADL	First MRI	0.83 ± 0.06	0.83 ± 0.07	0.82 ± 0.12
	Second MRI	0.87 ± 0.03	0.82 ± 0.10	0.83 ± 0.07
U-Net	First MRI	0.79 ± 0.06 p<0.5	0.78 ± 0.09 p<0.5	0.74± 0.15 p<0.5
	Second MRI	0.84 ± 0.04 p<0.5	0.77 ± 0.10 p<0.5	0.67 ± 0.25 p<0.5

Cor, Sag and Ax are abbreviated for the coronal, sagittal and Axial planes, respectively.

Figure 7-5 shows the agreement between manual and SADL-based automated volume measurements in the Bland–Altman plot. The mean difference between the two-volume measurements was calculated as an estimated bias. Standard deviation (SD) of the differences, and 95% limits of agreement (average difference ± 1.96 SD) were calculated to assess the random fluctuations around this mean. It turned out that 6 out of 62 MRI scans (9.5%) had the volume measurement differences that were beyond the 95% limits of agreement, which supports the validity of the automated volume measurements even during early pregnancy to be at least as reliable as the manual measurements, and perhaps can be used interchangeably.

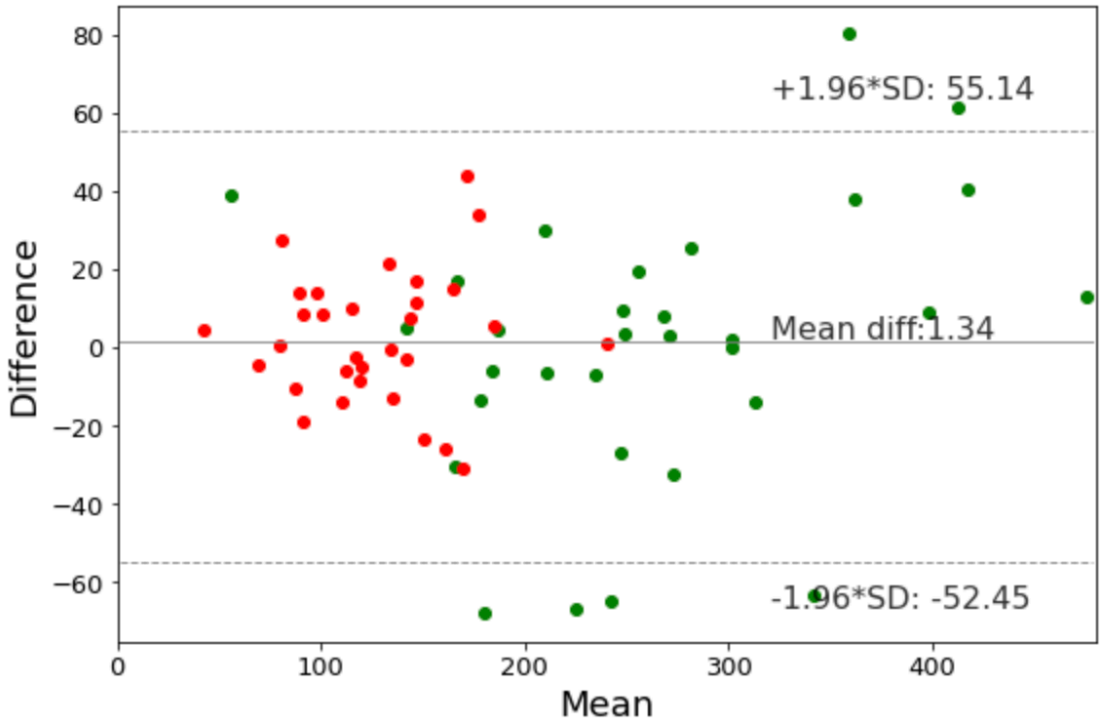

Figure 7-5 A Bland-Altman plot showing the agreement between the automated and manual placental volume measurement. Red and green points represent the first and second MRIs, respectively.

Figure 7-6 shows the linear regression models between the placental volume size and the gestational ages for SADL-based automated (Figure 7-6 (A)) and manual (Figure 7-6 (B)) volume calculations. Automated and manual volume calculations shared similar linear regression models that characterized the relationship between the volume and gestational ages.

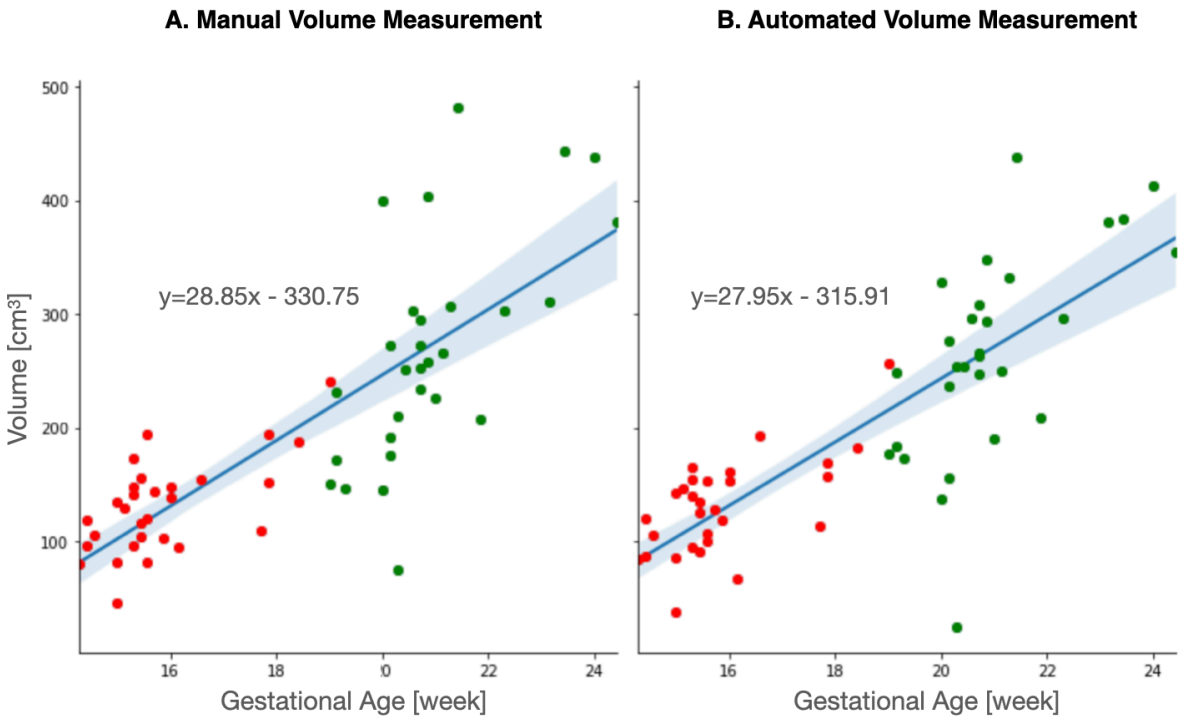

Figure 7-6 Linear regression models between placental volume and gestational age with the manual (A) and automated (B) segmentation. Red and green points are the volume measurements for the first and second MRIs. Blue lines represent the linear regression models between placental volume and gestational age.

7.3 Discussion

We developed a novel method, SADL, for automated segmentation of placenta from two longitudinal MRI scans, taken at 14-18 and also at 19-24 weeks of gestation. Our results demonstrated that accurate contouring of placenta on both MRIs is feasible, which is a prerequisite step for detecting abnormalities in this organ. The proposed method is fully automated; thus, the results are reproducible. The evaluation results showed that placental volume computed from manual and automatic segmentations can be used interchangeably. Our suggested methodology outperformed to the most-state-of-the-art methods for placenta delineation from MRI.

Our model is equipped with a CC spatial attentive module[76] that requires less GPU memory than the regular spatial attentive module. In the CC spatial attentive module, the response of each pixel only considers the pixels on its criss-cross path rather than all the pixels in the image, which reduces a significant amount of GPU memory. CC spatial attentive module could provide an alternative to the regular spatial attention module when the GPU memory available is limited.

Previously some deep learning-based techniques presented for placenta segmentation, among them only the method introduced by Han et al.[81] employed a U-Net to perform placental segmentation on axial, coronal, and sagittal MRI. Although we used a U-Net-based architecture , our work is different from theirs in the following aspects: 1) We had a larger data set containing MRI of 154 subjects, from which 32 images were used for test; 2) Our model is fully automated where the whole images is fed to the network for placental segmentation while in the method proposed by Han et al., first user needs to determine the extent of the placenta on MRI; 3) Our study conducted a patient-wise segmentation evaluation compared to the image-wise segmentation evaluation in their study; 4) Dataset used in our study included MRI scans obtained at two gestational ages, which we used to explore the performance of the segmentation model across multiple gestational ages;

Table 7-5 presented the testing result of SADL when different combinations of MRI served as the training set. We found that the model trained using both MRIs (first and second) achieved better segmentation performance than the one trained only using a single MRI. Since the placenta is a temporary human organ that varies substantially during early gestation, using the first and second MRI together during the training increased the amount and type of training samples of placental MRI, which could have improved the model's ability to recognize different-sized placental regions. The consistent segmentation performance in the MRI at two different early GAs

also suggests that the model potentially provide robust segmentation across differing longitudinal MRI scans during early gestation.

Table 7-5 Testing of SADL trained with different combinations of MRI

	Model Trained using the first MRI	Model Trained using the second MRI	Model Trained using both MRI
Testing on the first MRI	0.81 ± 0.06	0.77 ± 0.09	0.83 ± 0.06
Testing on the second MRI	0.81 ± 0.06	0.76 ± 0.13	0.84 ± 0.05

Our findings indicated that the relationship between volume and gestational age was maintained between manual and automated volume calculations. This could benefit future studies that require such relationships for the detection of placental volume-related disease models. Examples such as gestational diabetes mellitus or ischemic placental disease that lend themselves to changes in placental volumes[91,92] could perhaps be accurately detected prior to the development of clinical and biochemical symptomatology.

Our study has some limitations. First, the SADL model is a 2D-based deep learning model, which does not retain the inter-slice correlation information as in 3D placental images. We will explore ways to develop a 3D-based model to better capture the inter-slice correlation information in the future. Second, although we used a relatively large dataset, all images were obtained from a single medical center, that may introduce population bias to this study. Moreover, the same placental MRI protocol with a single vendor was used for acquiring placental MRI images for all scans. In the future, datasets from multiple institutions and vendors will be integrated to test the generalizability of our developed automated placenta segmentation method.

7.4 Conclusion

This Chapter proposed a spatial attentive deep learning network - SADL for automated segmentation of the placenta during the second trimester. SADL can automatically segment the placenta with high accuracy in placenta MRI at different gestational ages during the second trimester. In addition, the difference between automated placental volume measurement with the SADL-based segmentation and manual volume measurement was subtle.

Chapter 8 Summary and Future Work

This dissertation covers several improvements for current deep learning-based methods for MRI segmentation and classification, particularly in the prostate MRI. Specifically, for prostate zonal segmentation, Chapter 3 describes a deep learning model that incorporates the feature pyramid attention module to enhance the network's segmentation abilities, and Chapter 4 describes an attentive Bayesian deep learning method to enhance Chapter 3's result and meanwhile produce the uncertainty measurements of the automated segmentation; Chapter 5 describes a texture-based deep learning method (Textured-DL), which enriches tumor prior information into the deep learning, to improve the prostate cancer classification of the suspicious lesion on MRI; Chapter 6 describes a large patient cohort evaluation of a deep learning method that was previously developed in Chapter 4, for whole prostate gland segmentation; To evaluate the generalizability of the deep learning model for other biomedical image applications, Chapter 7 evaluated a deep learning-based segmentation method, which was adapted from the deep learning model developed in Chapter 4, for the placental segmentation on longitudinal MRI.

In this chapter, summaries of technical developments covered in this dissertation were described in the first part of the chapter and potential future work are discussed in the second part of this chapter.

8.1 Summary of Technical Development

8.1.1 Automatic Prostate Zonal Segmentation Using Fully Convolutional Network with Feature Pyramid Attention

As described in chapter 3, the proposed deep learning method was superior to the typical deep learning method - U-Net, for prostate zonal segmentation, in two separate testing datasets.

Besides, performance difference between the two datasets for prostate zonal segmentation was subtle for the proposed method. In addition, segmentation performance achieved in the middle slices was better than in other slices, such as apex-end and base-end slices. The reason could be the more conspicuous image features in the middle slices than in other slices. Moreover, the proposed deep learning-based model has a comparable segmentation performance to the human experts. Nevertheless, only two experts were involved in the performance comparison with the model. The study will recruit more experts to acquire more robust comparisons.

8.1.2 Exploring Uncertainty Measures in Bayesian Deep Attentive Neural Networks for Prostate Zonal Segmentation

As described in chapter 4, the proposed attentive Bayesian deep learning can automatically produce the prostate zonal segmentation with the uncertainty measurement. For the prostate zonal segmentation, the proposed method is superior to the state-of-art methods on the prostate zonal segmentation including the method in the Chapter 3. For the average uncertainty measurement at the three prostate locations (apex, middle, and base slices), zonal boundaries exhibit higher segmentation uncertainties than interior areas. Organ boundaries are usually challenging to define precisely due to the partial volume effects, possibly leading to inconsistent manual annotations and higher segmentation uncertainty. Also, the model has the highest uncertainties at the intersection between the prostate zones. The overall uncertainties by the Bayesian model demonstrated different uncertainties between prostate zones at three prostate locations (apex, middle and base), which was consistent with the actual model performance.

8.1.3 Textured-Based Deep Learning in Prostate Cancer Classification with 3T Multiparametric MRI: Comparison with PI-RADS-Based Classification

As described in Chapter 5, the proposed texture-based deep learning model (Textured-DL) outperformed the radiologist-based classification, radiomics texture-based machine learning, and conventional deep learning methods. In addition, Textured-DL demonstrated a better classification performance for the PZ tumors than the TZ tumors. Also, Textured-DL achieved better classification performance for lesions with PI-RADS scores of 4 or 5 than those with a PI-RADS score of 3. The superior classification performance of Textured-DL over PSAD-based predictions among PI-RADS 3 lesions indicated that Textured-DL could potentially serve as an additional tool to predict risks associated with PI-RADS 3 lesions and further reduce unnecessary biopsies for these lesions.

8.1.4 Deep Learning Enables Prostate MRI Segmentation: A Large Cohort Evaluation with Inter-rater Variability Analysis

In Chapter 6, a previously developed deep learning-based segmentation model in Chapter 4, attentive Bayesian deep learning network, was evaluated for whole prostate gland segmentation using a large, continuous cohort of prostate 3T MRI scans. In qualitative evaluation, the deep learning model demonstrated either acceptable or excellent segmentation performance in most of the MRI scans. In the quantitative evaluation, the deep learning model demonstrated a dice similarity coefficient of 0.93, which outperformed other baseline deep learning methods, such as DeepLab v3+ and UNet. Deep learning-enabled volume measurement can be used interchangeably with manual volume measurement in evaluating the volume measurement.

8.1.5 Evaluation of Spatial Attentive Deep Learning for Automated Placental Segmentation on Longitudinal MRI

In Chapter 7, a deep learning network, which was adapted from a previously developed deep learning in Chapter 4, was evaluated for the placental segmentation on longitudinal MRI to

investigate the model's generalizability for other biomedical image applications except the prostate segmentation. The deep learning model can automatically segment the placenta with high accuracy. In addition, the difference between automated placental volume measurement with the deep learning-based segmentation and manual volume measurement was small. Furthermore, automated and manual volume calculations shared similar linear regression models of characterizing the relationship between the volume size and gestational ages.

8.1.6 Overall Summary

Advanced deep learning models covered in this dissertation demonstrated outstanding segmentation and classification performance and outperformed state-of-art deep learning methods. The deep learning model with feature pyramid attention achieved superior prostate zonal segmentation performance to the state-of-art deep learning method. This dissertation's advanced deep learning models demonstrated outstanding segmentation and classification performance and outperformed state-of-art deep learning methods. The deep learning model with feature pyramid attention achieved superior prostate zonal segmentation performance to the state-of-art deep learning method. The attentive Bayesian deep learning model outperformed more state-of-art methods, including the deep learning model with feature pyramid attention, and generated uncertainties consistent with the actual model performance. Deep learning-based prostate cancer classification can be enhanced by enriching prior knowledge into the deep learning. The deep learning-based segmentation could generate segmentation of the prostate with sufficient quality when a large cohort of MRI scans was used for evaluation. High accuracy segmentation performance of the deep learning on the longitudinal MRI demonstrated the model's superb generalizability abilities on other biomedical applications apart from the prostate segmentation.

8.2 Future work

In the segmentation, one of the future works is to explore the prior knowledge that can enhance the MRI segmentation performance. Fusing the prior knowledge into the deep learning model can potentially reduce the deep learning's need for a large dataset in training. Gradient and texture images can be the prior knowledge for the segmentation. Especially for tumor segmentation, texture image will play a critical role since texture can quantify a tumor's prior knowledge, such as heterogeneity. A combination of prior knowledge and the input image can directly input into the model to enhance the segmentation performance in a small dataset.

Another future work in segmentation is fast segmentation. We can utilize contour-based deep learning instead of pixel-based deep learning to perform the automated segmentation. In the contour-based method, graph convolution achieves a regression task of pixel-wise offsets to deform the initial contour to the reference contour of the target region. Unlike pixel-based segmentation, the contour-based method's computation is primarily around the contour, which can vastly improve the segmentation speed.

As described in Chapter 6, evaluation of deep learning-based segmentation can be done using quantitative, and qualitative evaluation. Using segmentation-based downstream applications such as volume measurement and radiomic features can also evaluate segmentation performance. In Chapters 5 and 6, the discrepancy between automated and manual volume measurements has been adopted to measure model's segmentation performance. In the future, the difference between radiomic features extracted from the automated and manual segmented regions will also be used to estimate the model's segmentation performance. Segmentation evaluation using radiomic features difference assesses to which level the model can be involved in automating the extraction of features, which is significant to large-scale studies related to radiomics features that usually

relies heavily on the automated process.

In Chapter 5, Textured-DL used T2W and ADC as the input to classify a suspicious prostate cancer. Although the model has achieved the satisfactory result, inclusion of other MRI sequences/components, such as high b-value DWI, dynamic contrast-enhanced (DCE) MRI, and oxygen-enhanced MRI, into the model could provide more useful information for the prostate cancer classification. Besides, Textured-DL relied on the texture information from the GLCM to classify the suspicious lesions. Although texture information is the key to cancer, other radiomics features such as shape-based features and first-order statistics could still be essential in the classification. Also, the clinical and demographic information, such as PSA, PSA density, age, location of the lesion, patients' inheritance, BMI, etc., can be used to improve the performance. In the future, the study will also include the non-textured radiomics features and clinical and demographic information into the model to provide a more representative and robust characterization of prostate cancer.

In Chapter 5, Textured-DL uses tumor prior knowledge for the classification, thus making it potentially able to be used in a small dataset. However, this Chapter did not evaluate the effect of data size on the Textured-DL classification performance. Future studies will perform a thorough comparison of the amount of data required by Textured-DL and deep learning and investigate the minimal data size used to sustain Textured-DL classification performance.

REFERENCE

1. Asvadi NH, Afshari Mirak S, Mohammadian Bajgiran A, et al. 3T multiparametric MR imaging, PIRADSv2-based detection of index prostate cancer lesions in the transition zone and the peripheral zone using whole mount histopathology as reference standard. *Abdom Radiol*. 2018;43(11):3117-3124. doi:10.1007/s00261-018-1598-9
2. Kasivisvanathan V, Rannikko AS, Borghi M, et al. MRI-targeted or standard biopsy for prostate-cancer diagnosis. *N Engl J Med*. 2018;378(19):1767-1777.
3. Wenger E, Mårtensson J, Noack H, et al. Comparing manual and automatic segmentation of hippocampal volumes: reliability and validity issues in younger and older brains. *Hum Brain Mapp*. 2014;35(8):4236-4248.
4. Maldjian C, Adam R, Pelosi M, Pelosi III M, Rudelli RD, Maldjian J. MRI appearance of placenta percreta and placenta accreta. *Magn Reson Imaging*. 1999;17(7):965-971.
5. Liu Y, Yang G, Mirak SA, et al. Automatic Prostate Zonal Segmentation Using Fully Convolutional Network With Feature Pyramid Attention. *IEEE Access*. 2019;7:163626-163632.
6. Dahdouh S, Andescavage N, Yewale S, et al. In vivo placental MRI shape and textural features predict fetal growth restriction and postnatal outcome. *J Magn Reson Imaging*. 2018;47(2):449-458. doi:10.1002/jmri.25806
7. Litjens G, Debats O, Barentsz J, Karssemeijer N, Huisman H. Computer-aided detection of prostate cancer in MRI. *IEEE Trans Med Imaging*. 2014;33(5):1083-1092.
8. Litjens G, Debats O, Ven W van de, Karssemeijer N, Huisman H. A pattern recognition approach to zonal segmentation of the prostate on MRI. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. ; 2012:413-420.

9. Cameron A, Khalvati F, Haider MA, Wong A. MAPS: a quantitative radiomics approach for prostate cancer detection. *IEEE Trans Biomed Eng*. 2015;63(6):1145-1156.

10. Fehr D, Veeraraghavan H, Wibmer A, et al. Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. *Proc Natl Acad Sci*. 2015;112(46):E6265--E6273.

11. Gu J, Wang Z, Kuen J, et al. Recent advances in convolutional neural networks. *Pattern Recognit*. 2018;77:354-377.

12. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. ; 2015:234-241.

13. Li H, Xiong P, An J, Wang L. Pyramid Attention Network for Semantic Segmentation. Published online May 25, 2018. Accessed July 18, 2019. http://arxiv.org/abs/1805.10180

14. Weinreb JC, Barentsz JO, Choyke PL, et al. PI-RADS prostate imaging–reporting and data system: 2015, version 2. *Eur Urol*. 2016;69(1):16-40.

15. Turkbey B, Rosenkrantz AB, Haider MA, et al. Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2. *Eur Urol*. Published online March 2019. doi:10.1016/j.eururo.2019.02.033

16. Wang S, Burtt K, Turkbey B, Choyke P, Summers RM. Computer aided-diagnosis of prostate cancer on multiparametric MRI: a technical review of current research. *Biomed Res Int*. 2014;2014.

17. Lim MC, Tan CH, Cai J, Zheng J, Kow AWC. CT volumetry of the liver: where does it stand in clinical practice? *Clin Radiol*. 2014;69(9):887-895.

18. Dawson LA, Jaffray DA. Advances in image-guided radiation therapy. *J Clin Oncol*.

2007;25(8):938-946.

19. Yuan Y, Li B, Meng MQ-H. Bleeding frame and region detection in the wireless capsule endoscopy video. *IEEE J Biomed Heal informatics*. 2015;20(2):624-630.

20. Karthick S, Sathiyasekar K, Puraneeswari A. A survey based on region based segmentation. *Int J Eng Trends Technol*. 2014;7(3):143-147.

21. Hojjatoleslami SA, Kittler J. Region growing: a new approach. *IEEE Trans Image Process*. 1998;7(7):1079-1084.

22. Dhanachandra N, Manglem K, Chanu YJ. Image segmentation using K-means clustering algorithm and subtractive clustering algorithm. *Procedia Comput Sci*. 2015;54:764-771.

23. Li C, Huang R, Ding Z, Gatenby JC, Metaxas DN, Gore JC. A level set method for image segmentation in the presence of intensity inhomogeneities with application to MRI. *IEEE Trans image Process*. 2011;20(7):2007-2016.

24. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444.

25. Nair T, Precup D, Arnold DL, Arbel T. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Med Image Anal*. 2020;59:101557.

26. Kononenko I. Bayesian neural networks. *Biol Cybern*. 1989;61(5):361-370.

27. Gal Y. Uncertainty in deep learning. *Univ Cambridge*. 2016;1:3.

28. Gal Y, Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: *International Conference on Machine Learning*. ; 2016:1050-1059.

29. Wanders JOP, van Gils CH, Karssemeijer N, et al. The combined effect of mammographic texture and density on breast cancer risk: a cohort study. *Breast Cancer Res*. 2018;20(1):1-10.

30. Tan J, Gao Y, Liang Z, et al. 3D-GLCM CNN: A 3-Dimensional Gray-Level Co-Occurrence Matrix-Based CNN Model for Polyp Classification via CT Colonography. *IEEE Trans Med Imaging*. 2019;39(6):2013-2024.

31. Bharati MH, Liu JJ, MacGregor JF. Image texture analysis: methods and comparisons. *Chemom Intell Lab Syst*. 2004;72(1):57-71.

32. Haralick RM, Shanmugam K, Dinstein IH. Textural features for image classification. *IEEE Trans Syst Man Cybern*. 1973;(6):610-621.

33. Ghafoori M, Alavi M, Aliyari Ghasabeh M. MRI in Prostate Cancer. *Iran Red Crescent Med J*. 2013;15(12). doi:10.5812/ircmj.16620

34. Hoeks CMA, Barentsz JO, Hambrock T, et al. Prostate Cancer: Multiparametric MR Imaging for Detection, Localization, and Staging. *Radiology*. 2011;261(1):46-66. doi:10.1148/radiol.11091822

35. Junker D, Schäfer G, Kobel C, et al. Comparison of Real-Time Elastography and Multiparametric MRI for Prostate Cancer Detection: A Whole-Mount Step-Section Analysis. *Am J Roentgenol*. 2014;202(3):W263-W269. doi:10.2214/AJR.13.11061

36. Meyer A, Rakr M, Schindele D, et al. Towards Patient-Individual PI-Rads v2 Sector Map: Cnn for Automatic Segmentation of Prostatic Zones From T2-Weighted MRI. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. ; 2019:696-700. doi:10.1109/ISBI.2019.8759572

37. Padgett K, Swallen A, Nelson A, Pollack A, Stoyanova R. SU-F-J-171: Robust Atlas Based Segmentation of the Prostate and Peripheral Zone Regions On MRI Utilizing Multiple MRI System Vendors. *Med Phys*. 2016;43(6Part11):3447.

38. Mooij G, Bagulho I, Huisman H. Automatic segmentation of prostate zones. *arXiv Prepr*

arXiv180607146. Published online 2018.

39. Clark T, Zhang J, Baig S, Wong A, Haider MA, Khalvati F. Fully automated segmentation of prostate whole gland and transition zone in diffusion-weighted MRI using convolutional neural networks. *J Med Imaging*. 2017;4(4):41307.

40. de Gelder A, Huisman H. Autoencoders for Multi-Label Prostate MR Segmentation. *arXiv Prepr arXiv180608216*. Published online 2018.

41. Zabihollahy F, Schieda N, Krishna Jeyaraj S, Ukwatta E. Automated segmentation of prostate zonal anatomy on T2-weighted (T2W) and apparent diffusion coefficient (ADC) map MR images using U-Nets. *Med Phys*. Published online 2019.

42. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. ; 2016:770-778.

43. Turkbey B, Rosenkrantz AB, Haider MA, et al. Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2. *Eur Urol*. Published online 2019.

44. Sonn GA, Chang E, Natarajan S, et al. Value of targeted prostate biopsy using magnetic resonance--ultrasound fusion in men with prior negative biopsy and elevated prostate-specific antigen. *Eur Urol*. 2014;65(4):809-815.

45. Hosseinzadeh M, Brand P, Huisman H. Effect of Adding Probabilistic Zonal Prior in Deep Learning-based Prostate Cancer Detection. In: *International Conference on Medical Imaging with Deep Learning -- Extended Abstract Track*. ; 2019.

46. Wang G, Zuluaga MA, Pratt R, et al. Slic-seg: Slice-by-slice segmentation propagation of the placenta in fetal MRI using one-plane scribbles and online learning. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and*

Lecture Notes in Bioinformatics). Vol 9351. Springer Verlag; 2015:29-37. doi:10.1007/978-3-319-24574-4_4

47. Rundo L, Han C, Nagano Y, et al. USE-Net: Incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets. *Neurocomputing*. 2019;365:31-43.

48. Wang X, Girshick R, Gupta A, He K. Non-local neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. ; 2018:7794-7803.

49. Mukhoti J, Gal Y. Evaluating bayesian deep learning methods for semantic segmentation. *arXiv Prepr arXiv181112709*. Published online 2018.

50. Kendall A, Gal Y. What uncertainties do we need in bayesian deep learning for computer vision? In: *Advances in Neural Information Processing Systems*. ; 2017:5574-5584.

51. Myronenko A, Song X. Point set registration: Coherent point drift. *IEEE Trans Pattern Anal Mach Intell*. 2010;32(12):2262-2275.

52. Donato G, Belongie S. Approximate thin plate spline mappings. In: *European Conference on Computer Vision*. ; 2002:21-31.

53. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 11211 LNCS. ; 2018:833-851. doi:10.1007/978-3-030-01234-2_49

54. Oktay O, Schlemper J, Folgoc L Le, et al. Attention u-net: Learning where to look for the pancreas. *arXiv Prepr arXiv180403999*. Published online 2018.

55. Alom MZ, Hasan M, Yakopcic C, Taha TM, Asari VK. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv Prepr*

arXiv180206955. Published online 2018.

56. Barth BK, Rupp NJ, Cornelius A, et al. Diagnostic Accuracy of a MR Protocol Acquired with and without Endorectal Coil for Detection of Prostate Cancer: A Multicenter Study. *Curr Urol*. 2018;12(2):88-96.

57. Mirak SA, Shakeri S, Bajgiran AM, et al. Three Tesla Multiparametric Magnetic Resonance Imaging: Comparison of Performance with and without Endorectal Coil for Prostate Cancer Detection, PI-RADS™ version 2 Category and Staging with Whole Mount Histopathology Correlation. *J Urol*. 2019;201(3):496-502.

58. Chen J, Yang L, Zhang Y, Alber M, Chen DZ. Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In: *Advances in Neural Information Processing Systems*. ; 2016:3036-3044.

59. Garvey B, Türkbey B, Truong H, Bernardo M, Periaswamy S, Choyke PL. Clinical value of prostate segmentation and volume determination on MRI in benign prostatic hyperplasia. *Diagnostic Interv Radiol*. 2014;20(3):229.

60. Ahmed HU, Bosaily AE-S, Brown LC, et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *Lancet*. 2017;389(10071):815-822.

61. Oelke M, Bachmann A, Descazeaud A, et al. EAU guidelines on the treatment and follow-up of non-neurogenic male lower urinary tract symptoms including benign prostatic obstruction. *Eur Urol*. 2013;64(1):118-140.

62. Jin Y, Yang G, Fang Y, et al. 3D PBV-Net: An automated prostate MRI data segmentation method. *Comput Biol Med*. 2021;128:104160. doi:https://doi.org/10.1016/j.compbiomed.2020.104160

63. Cheng R, Roth HR, Lay NS, et al. Automatic magnetic resonance prostate segmentation by deep learning with holistically nested networks. *J Med imaging*. 2017;4(4):41302.

64. Checcucci E, Autorino R, Cacciamani GE, et al. Artificial intelligence and neural networks in urology: current clinical applications. *Minerva Urol e Nefrol Ital J Urol Nephrol*. 2019;72(1):49-57.

65. Checcucci E, De Cillis S, Granato S, et al. Applications of neural networks in urology: a systematic review. *Curr Opin Urol*. 2020;30(6):788-807.

66. Zhu Q, Du B, Turkbey B, Choyke PL, Yan P. Deeply-supervised CNN for prostate segmentation. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. ; 2017:178-184.

67. Jia H, Xia Y, Song Y, Cai W, Fulham M, Feng DD. Atlas registration and ensemble deep convolutional neural network-based prostate segmentation using magnetic resonance imaging. *Neurocomputing*. 2018;275:1358-1369.

68. Liu Y, Yang G, Hosseiny M, et al. Exploring Uncertainty Measures in Bayesian Deep Attentive Neural Networks for Prostate Zonal Segmentation. *IEEE Access*. 2020;8:151817-151828.

69. Dice LR. Measures of the Amount of Ecologic Association Between Species. *Ecology*. 1945;26(3):297-302. doi:10.2307/1932409

70. Kiri\csli HA, Schaap M, Klein S, et al. Evaluation of a multi-atlas based method for segmentation of cardiac CTA data: a large-scale, multicenter, and multivendor study. *Med Phys*. 2010;37(12):6279-6291.

71. Kitzing YX, Prando A, Varol C, Karczmar GS, Maclean F, Oto A. Benign conditions that mimic prostate carcinoma: MR imaging features with histopathologic correlation.

Radiographics. 2016;36(1):162-175.

72. Giavarina D. Understanding bland altman analysis. *Biochem medica*. 2015;25(2):141-151.

73. Semenick D. Tests and measurements: The T-test. *Strength \& Cond J*. 1990;12(1):36-37.

74. McHugh ML. Interrater reliability: the kappa statistic. *Biochem medica*. 2012;22(3):276-282.

75. Turkbey B, Merino MJ, Gallardo EC, et al. Comparison of endorectal coil and nonendorectal coil T2W and diffusion-weighted MRI at 3 Tesla for localizing prostate cancer: correlation with whole-mount histopathology. *J Magn Reson Imaging*. 2014;39(6):1443-1448.

76. Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W. Ccnet: Criss-cross attention for semantic segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. ; 2019:603-612.

77. Shahedi M, Dormer JD, Tensingh Rajan Thanga Kani AD, et al. Segmentation of uterus and placenta in MR images using a fully convolutional neural network. In: *Event: SPIE Medical Imaging*. Vol 2020. ; 2020:59. doi:10.1117/12.2549873

78. Liu D, Shao X, Danyalov A, et al. Human placenta blood flow during early gestation with pseudocontinuous arterial spin labeling MRI. *J Magn Reson Imaging*. 2020;51(4):1247-1257.

79. Sorensen AV, Hutter JM, Grant EP, Seed M, Gowland P. T2* weighted placental MRI: basic research tool or an emerging clinical test of placental dysfunction? *Ultrasound Obstet Gynecol*. Published online 2019.

80. Wang G, Li W, Zuluaga MA, et al. Interactive Medical Image Segmentation Using Deep Learning with Image-Specific Fine Tuning. *IEEE Trans Med Imaging*. 2018;37(7):1562-

1573. doi:10.1109/TMI.2018.2791721

81. Han M, Bao Y, Sun Z, et al. Automatic Segmentation of Human Placenta Images with U-Net. *IEEE Access*. 2019;7:180083-180092. doi:10.1109/ACCESS.2019.2958133

82. Milletari F, Navab N, Ahmadi SA. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*. Institute of Electrical and Electronics Engineers Inc.; 2016:565-571. doi:10.1109/3DV.2016.79

83. Havaei M, Davy A, Warde-Farley D, et al. Brain tumor segmentation with Deep Neural Networks. *Med Image Anal*. 2017;35:18-31. doi:10.1016/j.media.2016.05.004

84. Looney P, Stevenson GN, Nicolaides KH, et al. Fully automated, real-time 3D ultrasound segmentation to estimate first trimester placental volume using deep learning. *JCI insight*. 2018;3(11). doi:10.1172/jci.insight.120178

85. Hu R, Singla R, Yan R, Mayer C, Rohling RN. Automated Placenta Segmentation with a Convolutional Neural Network Weighted by Acoustic Shadow Detection. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. Institute of Electrical and Electronics Engineers Inc.; 2019:6718-6723. doi:10.1109/EMBC.2019.8857448

86. Alansary A, Kamnitsas K, Davidson A, et al. Fast fully automatic segmentation of the human placenta from motion corrupted MRI. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 9901 LNCS. Springer Verlag; 2016:589-597. doi:10.1007/978-3-319-46723-8_68

87. Wang G, Zuluaga MA, Li W, et al. DeepIGeoS: A Deep Interactive Geodesic Framework

for Medical Image Segmentation. *IEEE Trans Pattern Anal Mach Intell*. 2019;41(7):1559-1572. doi:10.1109/TPAMI.2018.2840695

88. Wang G, Zuluaga MA, Pratt R, et al. Slic-Seg: A minimally interactive segmentation of the placenta from sparse and motion-corrupted fetal MRI in multiple views. *Med Image Anal*. 2016;34:137-147. doi:10.1016/j.media.2016.04.009

89. Patel MR, Klufas RA, Alberico RA, Edelman RR. Half-fourier acquisition single-shot turbo spin-echo (HASTE) MR: Comparison with fast spin-echo MR in diseases of the brain. *Am J Neuroradiol*. 1997;18(9):1635-1640.

90. Van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104--e107.

91. Metzenbauer M, Hafner E, Hoefinger D, et al. Three-dimensional ultrasound measurement of the placental volume in early pregnancy: Method and correlation with biochemical placenta parameters. *Placenta*. 2001;22(6):602-605. doi:10.1053/plac.2001.0684

92. Adams T, Yeh C, Bennett-Kunzier N, Kinzler WL. Long-term maternal morbidity and mortality associated with ischemic placental disease. In: *Seminars in Perinatology*. Vol 38. ; 2014:146-150.

93. Bjurlin MA, Carroll PR, Eggener S, et al. Update of the standard operating procedure on the use of multiparametric magnetic resonance imaging for the diagnosis, staging and management of prostate cancer. *J Urol*. 2020;203(4):706-712.

94. Mottet N, Bellmunt J, Bolla M, et al. EAU-ESTRO-SIOG guidelines on prostate cancer. Part 1: screening, diagnosis, and local treatment with curative intent. *Eur Urol*. 2017;71(4):618-629.

95. Woo S, Suh CH, Kim SY, Cho JY, Kim SH. Diagnostic performance of prostate imaging

reporting and data system version 2 for detection of prostate cancer: a systematic review and diagnostic meta-analysis. *Eur Urol*. 2017;72(2):177-188.

96. Padhani AR, Weinreb J, Rosenkrantz AB, Villeirs G, Turkbey B, Barentsz J. Prostate imaging-reporting and data system steering committee: PI-RADS v2 status update and future directions. *Eur Urol*. 2019;75(3):385-396.

97. Tewes S, Mokov N, Hartung D, et al. Standardized reporting of prostate MRI: comparison of the prostate imaging reporting and data system (PI-RADS) version 1 and version 2. *PLoS One*. 2016;11(9):e0162879.

98. Purysko AS, Bittencourt LK, Bullen JA, Mostardeiro TR, Herts BR, Klein EA. Accuracy and interobserver agreement for prostate imaging reporting and data system, version 2, for the characterization of lesions identified on multiparametric MRI of the prostate. *Am J Roentgenol*. 2017;209(2):339-349.

99. Girometti R, Giannarini G, Greco F, et al. Interreader agreement of PI-RADS v. 2 in assessing prostate cancer with multiparametric MRI: A study using whole-mount histology as the standard of reference. *J Magn Reson Imaging*. 2019;49(2):546-555.

100. Seo JW, Shin S-J, Taik Oh Y, et al. PI-RADS version 2: detection of clinically significant cancer in patients with biopsy gleason score 6 prostate cancer. *Am J Roentgenol*. 2017;209(1):W1--W9.

101. van der Leest M, Cornel E, Israël B, et al. Head-to-head Comparison of Transrectal Ultrasound-guided Prostate Biopsy Versus Multiparametric Prostate Resonance Imaging with Subsequent Magnetic Resonance-guided Biopsy in Biopsy-naïve Men with Elevated Prostate-specific Antigen: A Large Prospective M. *Eur Urol*. 2019;75(4):570-578. doi:10.1016/j.eururo.2018.11.023

102. Venderink W, van Luijtelaar A, Bomers JGR, et al. Results of targeted biopsy in men with magnetic resonance imaging lesions classified equivocal, likely or highly likely to be clinically significant prostate cancer. *Eur Urol*. 2018;73(3):353-360.

103. Manjunath BS, Ma W-Y. Texture features for browsing and retrieval of image data. *IEEE Trans Pattern Anal Mach Intell*. 1996;18(8):837-842.

104. Partio M, Cramariuc B, Gabbouj M, Visa A. Rock texture retrieval using gray level co-occurrence matrix. In: *Proc. of 5th Nordic Signal Processing Symposium*. Vol 75. ; 2002.

105. Gatenby RA, Grove O, Gillies RJ. Quantitative imaging in cancer evolution and ecology. *Radiology*. 2013;269(1):8-14.

106. Avanzo M, Wei L, Stancanello J, et al. Machine and deep learning methods for radiomics. *Med Phys*. 2020;47(5):e185--e202.

107. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118.

108. Liu H, Li H, Habes M, et al. Robust Collaborative Clustering of Subjects and Radiomic Features for Cancer Prognosis. *IEEE Trans Biomed Eng*. Published online 2020.

109. Heinrich MP, Jenkinson M, Bhushan M, et al. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Med Image Anal*. 2012;16(7):1423-1435.

110. Zhong X, Cao R, Shakeri S, et al. Deep transfer learning-based prostate cancer classification using 3 Tesla multi-parametric MRI. *Abdom Radiol*. 2019;44(6):2030-2039.

111. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv Prepr arXiv14126980*. Published online 2014.

112. Bonekamp D, Kohl S, Wiesenfarth M, et al. Radiomic machine learning for characterization

of prostate lesions with MRI: comparison to ADC values. *Radiology*. 2018;289(1):128-137.

113. Ho TK. Random decision forests. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol 1. ; 1995:278-282.

114. Song Y, Zhang Y-D, Yan X, et al. Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric MRI. *J Magn Reson Imaging*. 2018;48(6):1570-1577.

115. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv Prepr arXiv14091556*. Published online 2014.

116. Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. *Am Stat*. 1998;52(2):119-126.

117. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. *Biometrical J J Math Methods Biosci*. 2005;47(4):458-472.

118. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. Published online 1988:837-845.

119. Eliasziw M, Donner A. Application of the McNemar test to non-independent matched pair data. *Stat Med*. 1991;10(12):1981-1991.

120. Gaur S, Lay N, Harmon SA, et al. Can computer-aided diagnosis assist in the identification of prostate cancer on prostate MRI? a multi-center, multi-reader investigation. *Oncotarget*. 2018;9(73):33804.

121. Yuan Y, Qin W, Buyyounouski M, et al. Prostate cancer classification with multiparametric MRI transfer learning model. *Med Phys*. 2019;46(2):756-765.