UCLA

UCLA Electronic Theses and Dissertations

Title

Prostate Cancer Diagnosis from Multi-parametric Magnetic Resonance Imaging via Deep Learning

Permalink

https://escholarship.org/uc/item/0815r9gk

Author

Cao, Ruiming

Publication Date

2019

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

Prostate Cancer Diagnosis from Multi-parametric Magnetic Resonance Imaging via Deep Learning

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Computer Science

by

Ruiming Cao

2019

© Copyright by

Ruiming Cao

2019

ABSTRACT OF THE THESIS

Prostate Cancer Diagnosis from
Multi-parametric Magnetic Resonance Imaging
via Deep Learning

by

Ruiming Cao
Master of Science in Computer Science
University of California, Los Angeles, 2019
Professor Fabien Scalzo, Chair

Prostate cancer (PCa) is one of the most common cancer-related diseases among men in the United States. Multi-parametric magnetic resonance imaging (mp-MRI) is considered the best non-invasive imaging modality for diagnosing PCa. The core components of mp-MRI include T2-weighted imaging (T2w), diffusion-weighted imaging (DWI), and dynamic contrast-enhanced imaging (DCE), each of which provides distinct anatomical or functional information. However, mp-MRI for PCa diagnosis is currently limited by the qualitative or semi-quantitative interpretation criteria, leading to inter-reader variability and a suboptimal ability to assess lesion aggressiveness. Deep learning is a class of methods designed to automatically learn multi-layer artificial neural networks from the training data for various tasks, including image classification, object detection, and segmentation. Here, deep learning methods specific to multi-parametric imaging were proposed to detect, segment PCa lesion and assess the lesion aggressiveness. In addition, an alternative learning method using unannotated dataset was designed, due to the inaccessibility of accurate annotated dataset in many institutions.

The thesis of Ruiming Cao is approved.

Kyung Hyun Sung

Song-Chun Zhu

Fabien Scalzo, Committee Chair

University of California, Los Angeles

2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This work would not have been possible without the financial support of the Integrated Diagnostics Program, Department of Radiological Sciences & Pathology, David Geffen School of Medicine at UCLA.

I am grateful to all of those with whom I have had the pleasure to work during this and other related projects. Each of the members of my Thesis Committee has provided me extensive personal and professional guidance and taught me a great deal about both scientific research and life in general. I would especially like to thank Dr. Kyung Hyun Sung. As my teacher and mentor, he has taught me more than I could ever give him credit for here. He has shown me, by his example, what a good scientist (and person) should be.

Nobody has been more important to me in the pursuit of this project than the members of my family. I would like to thank my parents, whose love and guidance are with me in whatever I pursue. They are the ultimate role models.

CHAPTER 1

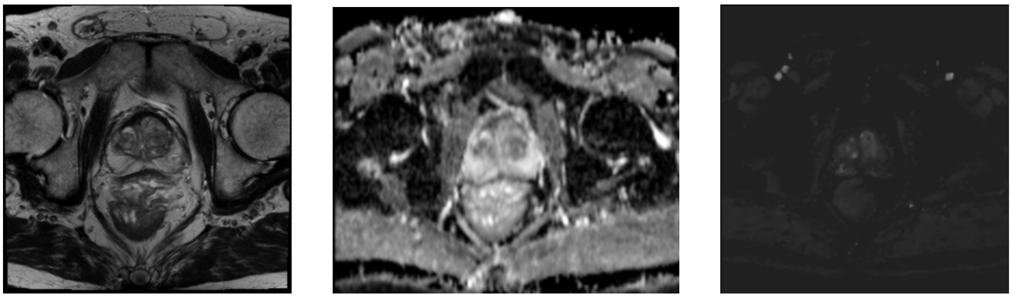
Introduction

Prostate cancer (PCa) is one of the most common cancer-related diseases among men in the United States [SMJ19]. While the incidence is high, PCa presents with a wide range of aggressiveness and in many cases does not develop into life-threatening cancer [EZS16, SMJ19]. The best assessment of lesion aggressiveness is the use of histologically assigned Gleason score (GS) [EEA16]. The general strategy for the diagnosis of PCa relies on non-targeted transrectal ultrasound (TRUS) guided biopsy. TRUS-guided biopsy is an invasive procedure and commonly suffers from over-detections of indolent PCa and under-detections of clinically significant PCa (csPCa) [WFG07, YVM12]. Therefore, there is an urgent need to achieve non-invasive detection and classification of PCa.

With recent advances in medical imaging technologies, multi-parametric magnetic resonance imaging (mp-MRI) provides a powerful combination of anatomical and functional information for PCa [YVM12]. As illustrated in Fig.1.1, the core elements of mp-MRI include T2-weighted (T2w) imaging, diffusion-weighted imaging (DWI), and dynamic contrast-enhanced (DCE) imaging, each of which provides distinct information.

The current diagnostic interpretation of mp-MRI is guided by Prostate Imaging Reporting and Data System version 2 (PI-RADS v2) [WBC16]. In PI-RADS v2 guideline, each element of mp-MRI is accessed and scored from 1 to 5 in a qualitative or semi-quantitative manner which causes inter- and intra-reader variations [SQA13]. Besides, PI-RADS v2 also has limited ability to assess the PCa aggressiveness or to distinguish csPCa from indolent PCa, which does not help for the active surveillance [SPG15].

Computerized analysis of prostate mp-MRI is an active research area to overcome limitations of current interpretation. Previous studies developed machine learning methods



(a) T2-weighted image (b) ADC map (c) K^{trans}

Figure 1.1: The common components of prostate mp-MRI. The apparent diffusion coefficient (ADC) map is calculated from DWI, and the volume transfer constant (K^{trans}) is quantitative measurement from DCE.

to automate the PCa detection and classify for the lesion aggressiveness [OLL10, VBK12, PJY13, LWT13, TKM13, LDB14, FVW15, KXW15, CKH16, KNT17, TLW17, YLW17, PPC17, SZY18, RAE18]. Lematre *et al.* [LMF15] and Wang *et al.* [WBT14] provide good overviews of the previous development. As the interpretation of prostate mp-MRI is highly challenging, the performance from the conventional machine learning models is often suboptimal due to the insufficient information extracted using the predefined features from mp-MRI and conventional machine learning models' limited capacity of representing non-linear relationships.

Deep learning has shown great promise in various tasks [LBH15, Sch15]. In particular, the convolutional neural network (CNN) can automatically learn the optimal image features from the training data, without the need for conventional handcrafted features [LBH15]. The CNN was first proposed in 1989 [LBD89], but at that time it was restricted because of the limited computational capability, the absence of large datasets, and the immature network structure and training algorithm. The renascence of CNN started since 2012 by [KSH12, DDS09], and the CNN soon became popular in computer vision, machine learning, and artificial intelligence communities. Many techniques have been proposed to improve the robustness of CNNs [IS15, SHK14], and CNNs have achieved state-of-the-art performance in object detection [Gir15], segmentation [LSD15], and classification [HZR16] in natural images. Recent works also applied deep learning for computer-aided diagnosis on medical imaging, e.g., X-ray [RIZ17], CT [ACE16, TB16], MRI [SS13, AGH17].

In this thesis, I will focus on improving the existing deep learning techniques in the specific context of prostate cancer diagnosis of mp-MRI to satisfy the unmet clinical needs. This thesis is organized as follows. Firstly, Chapter 2 introduces to use CNN and conditional random field to detect and segment PCa lesion in mp-MRI. Secondly, in Chapter 3, FocalNet, a CNN for ordinal label classification in multi-parametric imaging, is proposed for the joint detection and aggressiveness assessment of PCa. Thirdly, Chapter 4 discusses an alternative deep learning solution for PCa detection when no PCa lesion annotations are available for the training. Lastly, Chapter 5 discusses the potential usage of the proposed systems, limitations of the systems, and future research directions.

CHAPTER 2

Prostate cancer detection and segmentation in multi-parametric MRI via CNN and conditional random field

Multi-parametric MRI (mp-MRI) is a powerful diagnostic tool for prostate cancer (PCa). However, interpreting prostate mp-MRI requires high-level expertise, causing significant inter-reader variations. Convolutional neural networks (CNNs) have recently shown great promise for various tasks. In this study, we propose an improved CNN to jointly detect PCa lesions and segment for accurate lesions contours. Specifically, we adapt focal loss to overcome the imbalance between cancerous and non-cancerous areas for improved lesion detection and design selective dense conditional random field (SD-CRF), a post-processing step to refine the CNN prediction into the lesion segmentation based on a specific imaging component of mp-MRI. We trained and validated the proposed CNN in 5-fold cross-validation using 397 pre-operative mp-MRI exams with whole-mount histopathology-confirmed lesion annotations. In the free-response receiver operating characteristics (FROC) analysis, the proposed CNN achieved 75.1% lesion detection sensitivity at the cost of 1 false positive per patient. In the evaluation for lesion segmentation, the proposed CNN improved the Dice coefficient by 20.6% from the baseline CNN.

2.1 Introduction

Prostate cancer (PCa) is the most diagnosed cancer among men in the United States [SMJ19]. Multi-parametric MRI (mp-MRI) provides a powerful combination of anatomical and func-

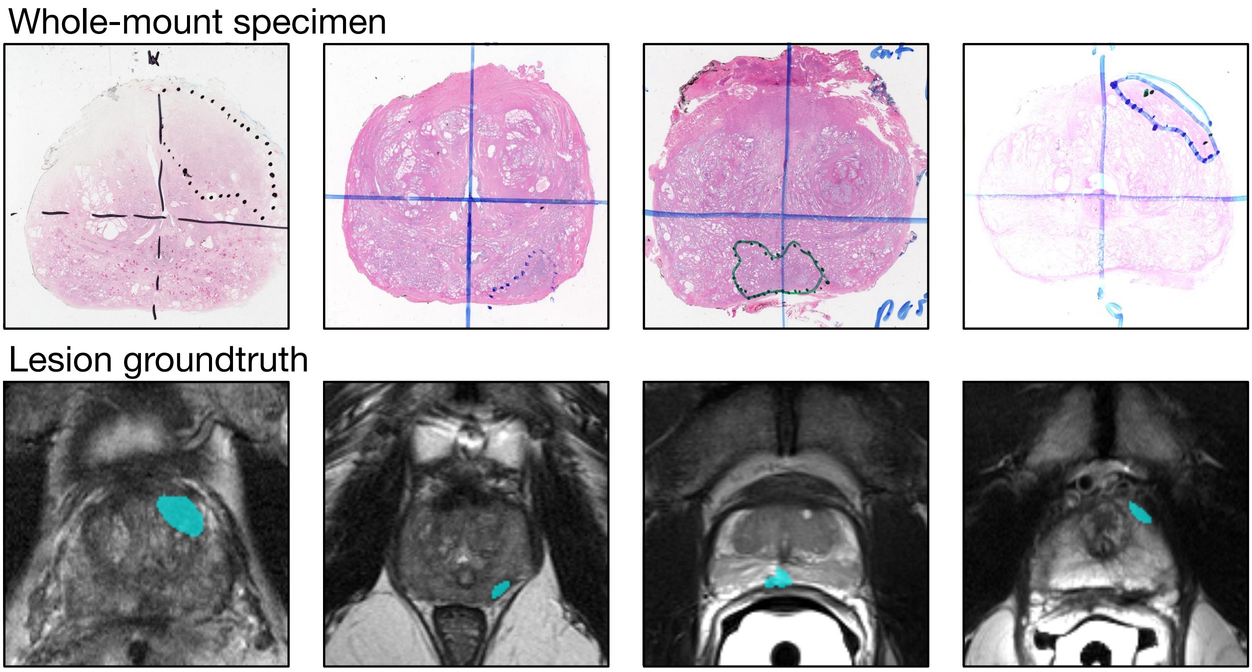


Figure 2.1: The lesion groundtruth was retrospectively annotated using whole-mount histopathology as the reference.

tional information and serves as a non-invasive imaging tool for the diagnosis of PCa. Computerized analysis of mp-MRI is an active research area to complement and overcome limitations of the current qualitative interpretation of mp-MRI [LDB14, WBT14]. Tsehay *et al.* [TLW17] and Kiraly *et al.* [KNT17] demonstrated the improved performance over imaging feature-based methods using convolutional neural networks (CNNs), but the tasks were focused on detection and classification only. Although it is highly desirable to jointly detect and segment the PCa lesions since the location, size, and shape of the lesion play important roles in the diagnosis and treatment planning of PCa [TMA12], the methods that can jointly detect and segment PCa lesions were not well investigated, mainly due to challenges such as the highly imbalanced number of normal and cancerous voxels and difficulties associated with different intensity and contrast patterns in the multi-parametric imaging.

Here, we describe the novel CNN model to jointly detect and segment PCa lesions, overcoming the challenges. Firstly, when the number of normal voxels is much higher than the number of cancerous voxels (e.g., only 1.6% of all voxels are annotated as cancerous in our study), adequate training of CNNs becomes difficult because the total loss is mostly composed of normal voxels. We use focal loss (FL), a modified cross-entropy loss, to adaptively controls the weights for each voxel [LGG17] and thus achieve balanced training from both normal and cancerous voxels. Secondly, as CNNs predict the probability map using multiple imaging components of mp-MRI, when a lesion shows different size or shape across imaging modalities, the probability map does not reflect the same intensity and contrast pattern shown in the component that best defines the lesion. We design the selective dense conditional random field (SD-CRF) 1) to select a certain imaging component in which the lesion is clearly observable, and 2) to fit the predicted probability into lesion segmentation with respect to the intensity pattern of the selected imaging component, instead of simply thresholding the predicted probability map. In addition, mutual finding loss (MFL) is developed and deployed in training, to enable the imaging component selection.

Our contributions are summarized as the proposed CNN model using 1) FL to overcome imbalanced data for improved lesion detection and 2) SD-CRF to fit the CNN prediction into a specific imaging component of mp-MRI for the refined lesion segmentation. The lesion detection is achieved by finding local maxima [LDB14] from the pixel-level lesion probability map, while the segmentation is obtained by combining the probability map into the intensity pattern from mp-MRI imaging components. We train and validate the proposed CNN model using 397 pre-operative mp-MRI exams with whole-mount histopathology-confirmed lesion annotations in 5-fold cross-validation.

2.2 Materials and methods

2.2.1 MRI data & lesion annotation

Under IRB approval, we collected 397 pre-operative MRI exams from patients after radical prostatectomy, satisfying that 1) whole-mount histopathology is available and 2) at least one lesion with Gleason score (GS)>6 or lesion diameter≥10mm was identified in histologic examinations. All imaging was performed in 3T scanners (Siemens Healthcare) using the

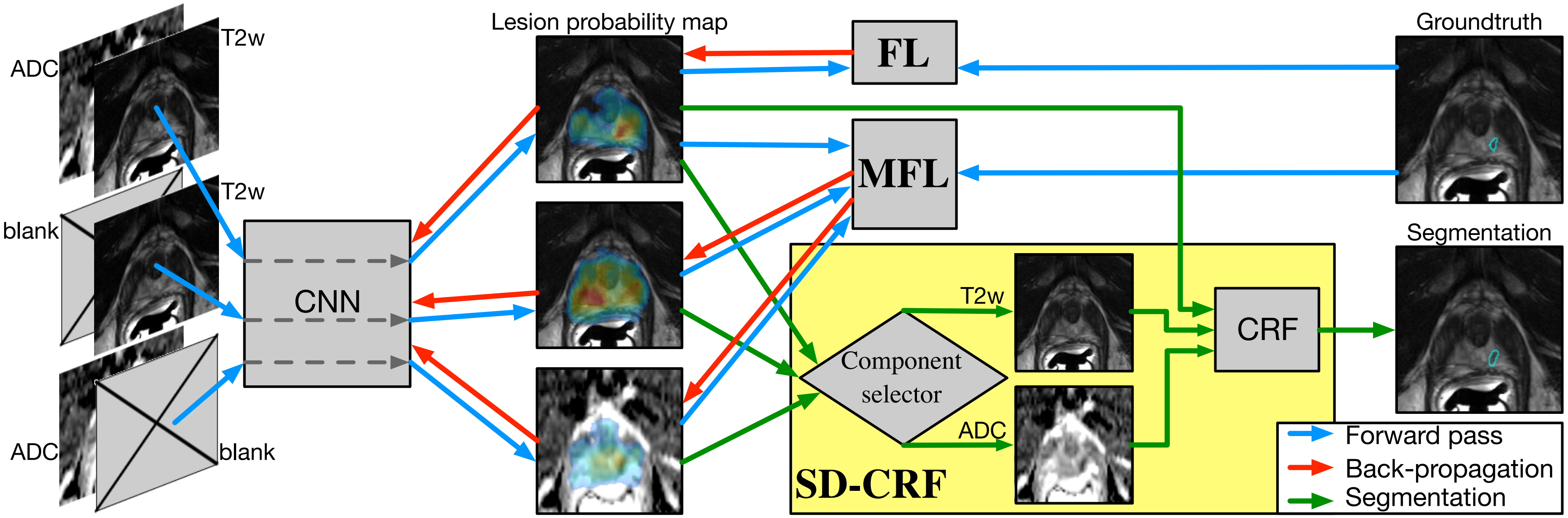


Figure 2.2: The CNN for joint PCa lesion detection and segmentation. ADC and T2w images stack as different channels of the CNN input [TLW17], and FL trains for the lesion probability map as in the top row. Also, T2w and ADC images are individually passed into the CNN, and their probability maps are trained by MFL. SD-CRF determines the specific imaging component and refines the predicted probability map for lesion segmentation using the conditional random field (CRF).

standardized mp-MRI protocol. T2-weighted (T2w) images and apparent diffusion coefficient (ADC) from diffusion-weighted images were used as inputs for the CNN [TLW17, WBC16].

Genitourinary (GU) radiology research fellows, supervised by a senior GU radiologist, retrospectively annotated lesions in mp-MRI with whole-mount specimens available for reference as in Fig. 2.1. Lesions with GS=6 and histologic lesion diameter<10mm were excluded in this study since mp-MRI was reported limited detectability for those lesions [TML15]. In total, we have annotated 546 lesions, including 112 (20%) GS=6 lesions, 266 (49%) GS 3+4 lesions, 109 (20%) GS 4+3 lesions, and 59 (11%) GS≥8 lesions.

2.2.2 Imaging pre-processing

For both training and validation, the intensity of T2w images was linearly normalized to $[0, 1]$ using the intensity value of bladder as the reference for the upper threshold and zero intensity as the lower threshold. Since ADC intensity value was suggestive for cancerous tissues [SNN05], we used fixed thresholds to normalize ADC intensity. The imaging preprocessing pipeline and the CNN were built with 16-bit integers or floating-point numbers so that the numeric precision of imaging intensity was preserved. Moreover, ADC images were registered to T2w images using rigid transformation based on the coordinate information stored in imaging files. An 80mm×80mm window centered at the prostate was cropped for each case [KNT17], and we only used slices with annotated lesions for training and validation as in [KNT17, WLC18].

2.2.3 CNN with imbalanced data

We build the CNN using the 101-layer deep residual network [HZR16, CPK18] to predict the pixel-level lesion probability map from ADC and T2w images. As in Fig. 2.2, focal loss (FL) [LGG17] is adapted to train for the lesion probability map from both ADC and T2w. FL balances the loss contributed from cancerous and non-cancerous areas by adding a focal weight to the regular cross-entropy loss, such that, for each pixel,

$$\mathrm{FL} = (1-p)^2\, y \log(p) + p^2 (1-y) \log(1-p)\,, \tag{2.1}$$

where $p \in [0, 1]$ is the predicted lesion probability and $y \in \{0, 1\}$ is the groundtruth label. FL adjusts the penalty based on the predicted probability. E.g., for an obvious non-cancerous pixel with $p = 0.05$, the FL is only 1/400 of the corresponding cross-entropy loss; for a cancerous pixel with $p = 0.2$, the FL remain 16/25 of the cross-entropy loss. In this way, the CNN training can concentrate on cancerous or suspicious pixels.

2.2.4 Mutual finding loss for multi-parametric imaging

In addition to FL which trains with respect to both ADC and T2w, mutual finding loss (MFL) is designed to train for the individual imaging components of mp-MRI.

$$\begin{aligned}\mathrm{MFL} = \frac{1}{N} \min\{&d\left(y \otimes f\left(I_{\mathrm{ADC}}, I_{\mathrm{T2w}}\right), y \otimes f\left(I_{\mathrm{ADC}}, \cdot\right)\right), \\ &d\left(y \otimes f\left(I_{\mathrm{ADC}}, I_{\mathrm{T2w}}\right), y \otimes f\left(\cdot, I_{\mathrm{T2w}}\right)\right)\},\end{aligned} \tag{2.2}$$

where d is the L2-distance, f denotes the CNN output, $\otimes$ is the element-wise product, and N is the number of pixels in the image. Specifically, the L2-distance between the CNN output from both imaging components and the output from either ADC or T2w alone is calculated on cancerous areas. MFL selects the individual imaging component, in which lesions are more observable, by choosing the component with the smaller L2-distance. Then, in training, MFL minimizes the L2-distance for the selected component so that the lesions can be equivalently observed from an individual component as from both components. Since MFL aims to train for the individual component, MFL does not back-propagate to the CNN output using both components as in Fig. 2.2.

FL and MFL are combined into the total loss during the training, such that $\mathrm{Loss} = \mathrm{FL} + \lambda \cdot \mathrm{MFL}$, where λ is set to the inverted fraction of cancerous pixels to balance between FL and MFL.

2.2.5 Selective dense conditional random field

Selective dense conditional random field (SD-CRF) is a non-parametric post-processing step to fit the CNN probability map into the intensity pattern of a specific imaging component for the lesion segmentation. SD-CRF first determines whether ADC or T2w defines the lesion

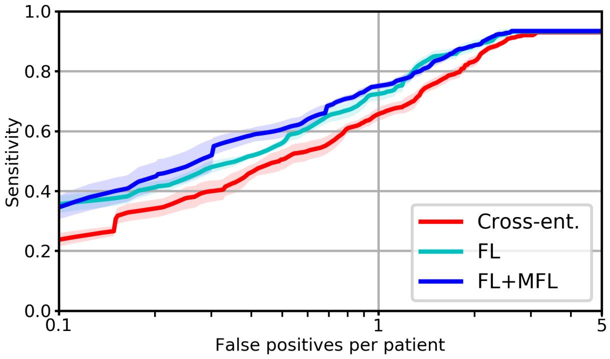


Figure 2.3: The FROC analysis for lesion detection under 5-fold cross-validation. The false positives per patient (x-axis) are shown in log scale.

better using the component selector shown in Fig. 2.2. The component selector compares the L2-distance on cancerous areas with either of the imaging components as in (2.2) and chooses the component resulting in a smaller L2-distance. I.e., the selected component, $I_{\mathrm{Sel}} = \arg\min_{c \in \{I_{\mathrm{ADC}}, I_{\mathrm{T2w}}\}} d\left(\hat{y} \otimes f_{\mathrm{out}}, \hat{y} \otimes f_c\right)$, where f_{out} and f_c are the CNN outputs from both components and from the specific imaging component c. $\hat{y} = \left[f\left(I_{\mathrm{ADC}}, I_{\mathrm{T2w}}\right) > 0.5\right]$ approximates for the groundtruth y, as y is not available in testing.

Then, a conditional random field is built for the refined lesion segmentation y^* with regard to the intensity pattern in the selected imaging component and the CNN output. Specifically, as in [KK11], y^* is inferred by minimizing the energy E such that

$$E\left(y^*\right) = \sum_{i=1}^{N} \phi_u\left(y_i^* | I_{\mathrm{ADC}}, I_{\mathrm{T2w}}\right) + \sum_{i<j}^{N} \phi_p\left(y_i^*, y_j^* | I_{\mathrm{Sel}}\right), \tag{2.3}$$

where ϕ_u is the unary potential from the negative log-likelihood of the CNN predicted probability map, and ϕ_p is the pairwise potential from ith and jth pixels. In particular, the pairwise potential is defined as $\phi_{\mathrm{p}}\left(y_i^*, y_j^* | I_{\mathrm{Sel}}\right) = -\exp\left(-d_{i,j}^2 - \Delta_{i,j}^2\right)$, where $d_{i,j}$ and $\Delta_{i,j}$ respectively are the spatial distance and the intensity difference between the ith and jth pixels. y^* is optimized via the iterative approach in [KK11].

2.3 Results

2.3.1 Experiment setup

Our experiment was performed under 5-fold cross-validation. Each fold contained 317 or 318 cases for training and 79 or 80 cases for validation. We experimented four different settings in the same CNN architecture: focal loss only (*FL*), focal loss and mutual finding loss together (*FL+MFL*), *FL+MFL* with SD-CRF for segmentation (*FL+MFL+CRF*), and the regular cross-entropy loss (*Cross-ent.*) as the baseline. For each setting, pre-trained network weights were used as the weight initialization, and the L2-normalization with a weight of 0.0001 was added to the total loss [HZR16]. Furthermore, common image augmentations, including random image shifting, scaling, and flipping, were applied during the training.

2.3.2 Lesion detection

Lesion detection is evaluated by the free-response receiver operator characteristics (FROC) analysis [LDB14,TLW17]. FROC measures the lesion detection sensitivity versus the number of false positive detections per patient. Given a predicted lesion probability map, detection points are located by finding the local maxima [LDB14]. A detection point is considered to be either true positive if it is in or within 5mm of an annotated lesion contour in the same slice [PNK17], or otherwise false positive.

As shown in Fig. 2.3, FL demonstrated its effectiveness in lesion detection, compared with *Cross-ent.* At 0.5 false positives per patient, *FL* and *FL+MFL* had a sensitivity of 56.0% and 60.6%, 5.5% and 10.1% higher than *Cross-ent.* Similarly, at 1 false positive per patient, *FL* and *FL+MFL* had 72.4% and 75.1% sensitivity, compared to the 65.7% sensitivity from *Cross-ent.* Moreover, for 80% detection sensitivity, *FL* and*FL+MFL* require 1.28 and 1.35 false positives per patient, reducing 28.7% and 25.3% of false positives from *Cross-ent.*

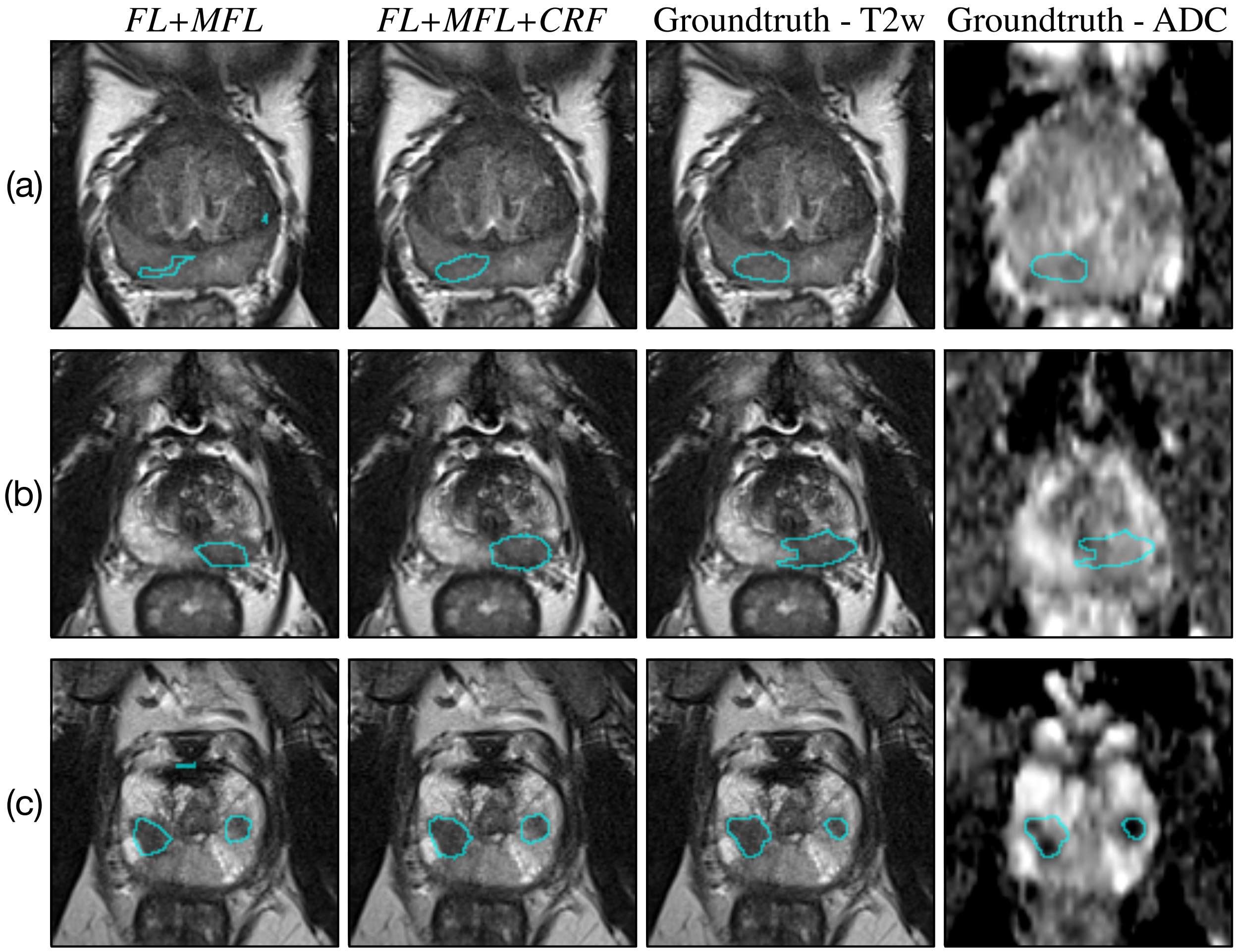


Figure 2.4: Examples for lesion segmentation.

Table 2.1: Evaluation for lesion segmentation under 5-fold cross-validation. Numbers are reported as avg±std.

	Dice(%)	Dice-all(%)	HD-95 (mm)
Cross-ent.	36.7 ± 1.6	18.0 ± 1.4	7.19 ± 0.15
FL	50.0 ± 2.3	34.6 ± 2.1	5.63 ± 0.09
FL+MFL	53.8 ± 1.8	39.7 ± 2.0	**5.24 ± 0.15**
FL+MFL+CRF	**57.3 ± 1.5**	**39.9 ± 1.7**	5.38 ± 0.21

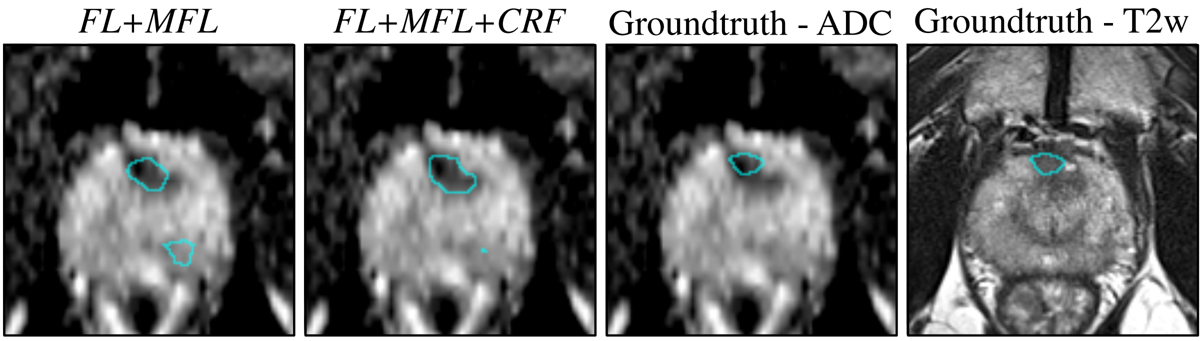


Figure 2.5: Another example of lesion segmentation. The ADC image is used as the background.

2.3.3 Lesion segmentation

Lesion segmentation is evaluated by Dice coefficient (Dice) and Hausdorff-95 distance (HD-95) defined in the brain tumor segmentation challenge [MJB15]. Dice assesses the similarity between the predicted and groundtruth regions, and HD-95 measures the closeness of the two contour boundaries. However, compared with brain tumors, PCa lesions have highly variable detectability [RWB12]. Thus, for the evaluation of lesion segmentation, we only focus on lesions that are roughly on target, to reduce the impact from misdetection. As in [LDB14], we consider predicted lesions, whose centers of mass are within 10mm of centers of mass of any groundtruth lesions, as the predicted lesions on target. Similarly, the groundtruth lesions within 10mm of predicted lesions are considered as groundtruth lesions being targeted. For Dice and HD-95, we evaluate based on predicted lesions on target and groundtruth lesions being targeted, so that the various lesion detection performance (e.g., groundtruth lesions being missed or false positive predicted lesions) does not affect the evaluation for segmentation. We also include Dice-all to evaluate based on all predicted and groundtruth lesions.

The results for lesion segmentation are shown in Fig. 2.4 and TABLE 2.1. SD-CRF improved the segmentation by refining the lesion contours and/or rejecting some noise predictions. As in TABLE 2.1, *FL+MFL+CRF* received higher Dice than *FL+MFL* and *FL*. However, *FL+MFL* had HD-95 marginally lower than *FL+MFL+CRF*. This is because all groundtruth lesion contours were annotated in T2w images, while SD-CRF sometimes fits the prediction into ADC, in which some lesions have different appearance causing this variation as in Fig. 2.5.

2.4 Conclusion

We proposed the improved CNN model to jointly detect PCa lesions and segment lesion contours. The CNN was trained by focal loss (FL) to overcame the imbalance between cancerous and non-cancerous areas, and we designed selective dense conditional random field (SD-CRF) to refine the CNN prediction into the lesion segmentation based on the intensity pattern of a specific imaging component of mp-MRI. We trained and validated the CNN under 5-fold cross-validation using 397 pre-operative mp-MRI exams with groundtruth lesion contours confirmed by whole-mount histopathology. In the experiment, the proposed CNN achieved 75.1% lesion detection sensitivity at 1 false positive per patient, 9.4% higher than the baseline CNN. For lesion segmentation, the proposed CNN received 57.3% Dice coefficient, compared to the baseline CNN with only 36.7% Dice coefficient.

CHAPTER 3

Joint Prostate Cancer Detection and Classification

Multi-parametric MRI (mp-MRI) is considered the best non-invasive imaging modality for diagnosing prostate cancer (PCa). However, mp-MRI for PCa diagnosis is currently limited by the qualitative or semi-quantitative interpretation criteria, leading to inter-reader variability and a suboptimal ability to assess lesion aggressiveness. Convolutional neural networks (CNNs) are a powerful method to automatically learn the discriminative features for various tasks, including cancer detection. We propose a novel multi-class CNN, FocalNet, to jointly detect PCa lesions and predict their aggressiveness using Gleason score (GS). FocalNet characterizes lesion aggressiveness and fully utilizes distinctive knowledge from mp-MRI. We collected a prostate mp-MRI dataset from 417 patients who underwent 3T mp-MRI exams prior to robotic-assisted laparoscopic prostatectomy (RALP). FocalNet is trained and evaluated in this large study cohort with 5-fold cross-validation. In the free-response receiver operating characteristics (FROC) analysis for lesion detection, FocalNet achieved 89.7% and 87.9% sensitivity for index lesions and clinically significant lesions at 1 false positive per patient, respectively. For GS classification, evaluated by the receiver operating characteristics (ROC) analysis, FocalNet received the area under the curve (AUC) of 0.81 and 0.79 for the classifications of clinically significant PCa (GS$\geq$3+4) and PCa with GS$\geq$4+3, respectively. With the comparison to the prospective performance of radiologists using the current diagnostic guideline, FocalNet demonstrated comparable detection sensitivity for index lesions and clinically significant lesions, only 3.4% and 1.5% lower than highly experienced radiologists without statistical significance.

3.1 Introduction

The challenge in diagnosing prostate cancer (PCa) is how to detect and distinguish indolent PCa from potentially clinically significant PCa. The current best assessment of lesion aggressiveness is the use of histologically assigned Gleason score (GS) [EEA16]. The current diagnosis of PCa in general medical practice still relies on non-targeted template driven transrectal ultrasound-guided (TRUS) biopsy, which results in under-detection of clinically significant PCa [YVM12]. 3 Tesla-based multi-parametric MRI (3T mp-MRI) provides a powerful combination of anatomical and functional information for PCa and plays a pivotal role in the diagnosis of PCa by reducing unnecessary biopsies [KRB18] and adding treatment options in active surveillance [DAB12] and focal therapy [VAE14]. The core components of mp-MRI include T2-weighted imaging (T2w), diffusion-weighted imaging (DWI), and dynamic contrast-enhanced imaging (DCE-MRI), each of which provides distinct information. Current diagnostic practice for mp-MRI follows the Prostate Imaging Reporting and Data System: Version 2 (PI-RADS v2) [WBC16], which evaluates radiologic findings in a qualitative or semi-quantitative manner. However, PI-RADS v2 still has limited ability to detect and distinguish between indolent and clinically significant PCa, with a wide range of sensitivity and specificity [RWB12], mainly due to inter-reader variability and suboptimal analysis.

Computer-aided diagnosis (CAD) using mp-MRI for PCa is being actively investigated for lesion detection and classification [OLL10,VBK12,PJY13,LWT13,TKM13,LDB14,WBT14, FVW15,LMF15,KXW15,CKH16,KNT17,TLW17,SZY18,RAE18]. The lesion detection approach typically extracts voxel- and/or region-level features from mp-MRI and predicts either PCa localization points or lesion segmentation masks. With recent advances in deep learning, convolutional neural networks (CNNs) are a powerful tool for image classification [KSH12] and segmentation [LSD15]. Recent studies also showed the feasibility of training CNNs to detect cancer from mp-MRI. Zhang *et al.* [ZSZ18] designed hierarchical coarse-to-fine CNNs to segment voxel-level tumor masks and suggest biopsy locations for breast cancer from DCE-MRI. Song *et al.* [SZY18] built a patch-based CNN to classify between biopsy-proven

PCa lesion and non-lesion regions of interest (ROIs). Kiraly *et al.* [KNT17] proposed to predict voxel-level labels of clinically significant PCa (GS>6) and non-clinically-significant PCa (GS$\leq$6) using CNN with two output channels to enable both detection and classification at the same time.

Interpreting prostate mp-MRI generally requires a high level of expertise as radiologic findings are qualitative, relying on T2 morphology and non-quantitative assessment of diffusion restriction and lesional enhancement [WBC16]. Thus, radiologic findings in one component of mp-MRI are more observable than in others. Common approaches to utilize multiple components of mp-MRI in CNNs are to stack them as different imaging channels (e.g., RGB channels for a color image) [ZLD15, PPA16, KNT17, TLW17, SZY18]. This enables CNNs to learn common knowledge across mp-MRI components from groundtruth annotations but may fail to learn the distinct information from each component of mp-MRI. As a result, some features appearing in only one or certain components of mp-MRI are difficult to be trained, especially when the number of training data is limited. Inspired by the clinical interpretation of prostate mp-MRI [WBC16], we design the mutual finding loss (MFL) to selectively train for different imaging components of mp-MRI. MFL identifies which subset of components would contain more observable information for a given PCa finding and defines the lesion-specific training objective as to observe the PCa finding from only the subset of imaging components.

A stratification of clinically significant PCa becomes important as differentiating between low- and intermediate/high-grade PCa is highly correlated with clinical outcomes [SPS09, DAB12]. The correlation between mp-MRI and GS has been studied [PJY13], but to our knowledge, no prior study has explored the use of mp-MRI to predict fine-grained GS groups via CNNs. Even though multi-class classification using CNN is widely available via one-hot encoding, different classes are usually assumed to be equally distanced, which ignores the progressiveness of GS groups (e.g., the difference between low- and intermediate-grade PCa is assumed to be the same as the difference between low- and high-grade PCa). Instead, we develop the ordinal encoding for different GS groups to adopt the lesion aggressiveness relationship into the encoded vectors. Unlike one-hot encoded vectors, ordinal encoded vectors are not mutually orthogonal and can suggest for the similarities and differences between different GS groups.

Recent CAD systems for PCa are generally trained and validated by using mp-MRI exams with biopsy-confirmed lesion findings [LDB14,KNT17,TLW17,SZY18]. However, the biopsy-confirmed lesion annotations are weighted towards MRI-positive lesions since biopsy cores are mostly based on MRI-positive findings (PI-RADS$\geq$3). As PI-RADS$\geq$3 has a limited ability to detect all PCa lesions [LTS15,RSW16,VHG16], clinically significant lesions can be missed and multi-focal lesions can be highly underestimated at mp-MRI [BGG18, LTS15], resulting in an overestimation of the performance of the CAD systems. Also, there exists a significant risk of the inaccurate lesion annotations since GS between prostate biopsy and radical prostatectomy specimens is occasionally discordant [LSB14, GBM15, EFT12]. Epstein *et al.* reported that more than one-third of the biopsy cases with GS$\leq$6 were upgraded to GS$\geq$7, and one-fourth of GS 3+4 in biopsy were downgraded after checking with whole-mount histopathology [EFT12]. To overcome these limitations, we use pre-operative mp-MRI exams before undergoing robotic-assisted laparoscopic prostatectomy (RALP) for our training and validation. The whole-mount histopathology analysis after RALP would provide the best definition of the GS groups and minimize the underestimation of the multi-focal lesions.

Here, we present a novel multi-class CNN, FocalNet, that jointly detects PCa lesions and predicts their GS. We arrange GS into five fine-grained GS groups [EZS16], i.e., GS 3+3, GS 3+4, GS 4+3, GS=8, and GS$\geq$9. FocalNet encodes six labels, the five GS groups and normal tissue, into ordinal encoded vectors, and predicts the label for each pixel using mp-MRI. FocalNet is also designed to selectively train distinctive features in one or certain imaging components of mp-MRI using mutual finding loss during the training.

We summarize our contributions as follows. Firstly, we propose FocalNet, an improved multi-class CNN to jointly detect PCa lesions and predict their Gleason score groups from mp-MRI. Secondly, in FocalNet, we design ordinal encoding to characterize lesion aggressiveness and mutual finding loss to fully exploit knowledge in the multi-parametric imaging. Thirdly, to our knowledge, this is the first study that trained or validated a CNN-

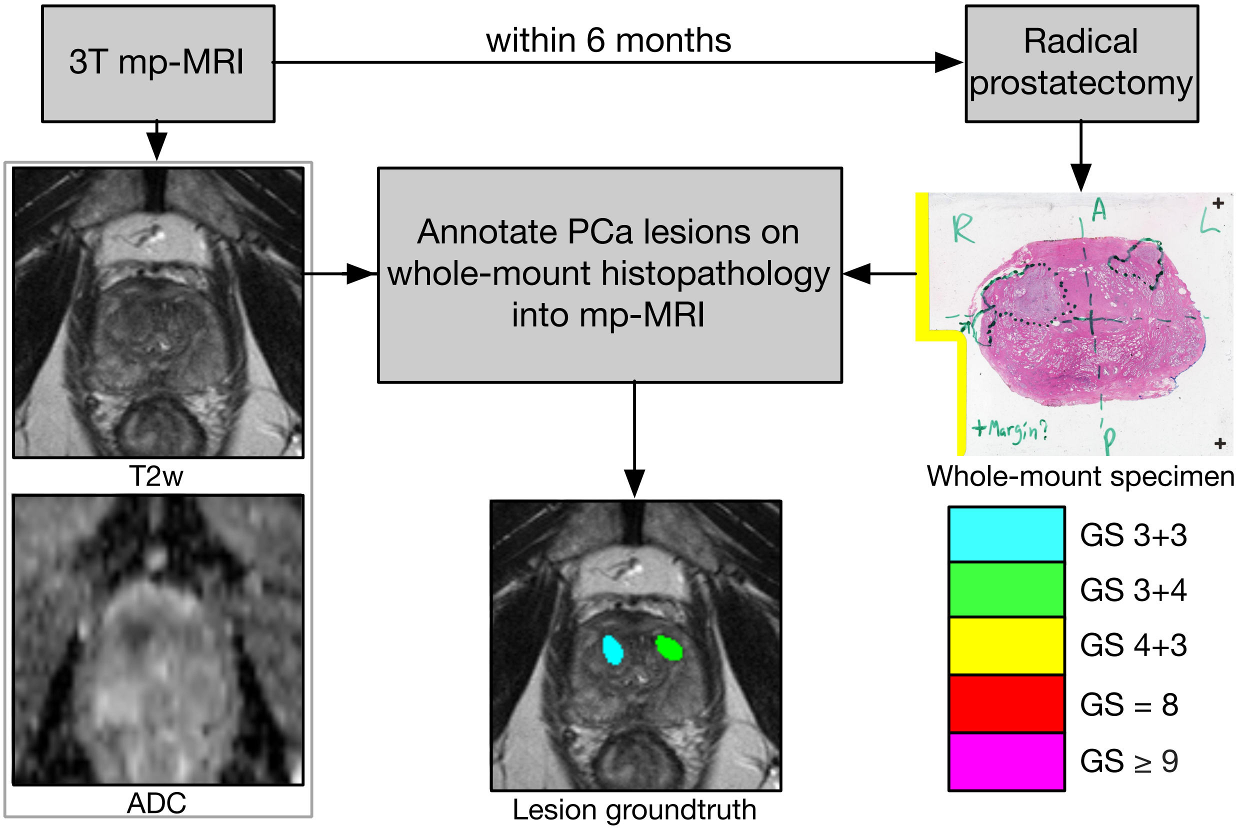


Figure 3.1: Data preparation pipeline. 278 out of 400 prospectively missed (false negative) lesions were retrospectively identified and annotated in mp-MRI, referring to whole-mount histopathology. In the shown example, the lesion in the left anterior (GS 3+4, index lesion) was prospectively missed and retrospectively identified.

based PCa detection and diagnosis system using lesion findings confirmed with whole-mount histopathology in a large study cohort.

This paper is organized as follows: In Section 3.2, we describe the MRI data and annotation process, the technical framework for FocalNet, and the experimental setups for pre-processing, training and validation. Section 3.3 presents PCa lesion detection and GS prediction results. In Section 3.4, we discuss potential implications and extensions of FocalNet, followed by concluding remarks.

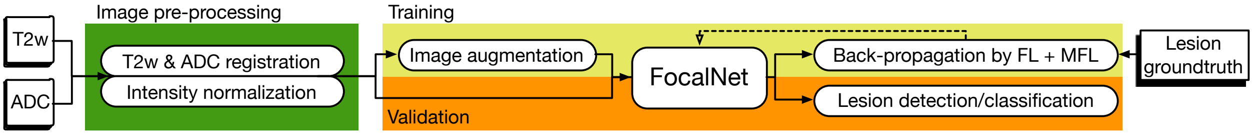


Figure 3.2: The workflow of FocalNet for training and validation. Image registration and intensity normalization are performed with 3D image volumes. As FocalNet operates with 2D images, the corresponding T2w and ADC slices are grouped and fed into FocalNet for pixel-level predictions.

3.2 Materials and Methods

3.2.1 MRI data

Pre-operative mp-MRI exams from 417 patients who later underwent RALP were included in the study. Patients with prior radiotherapy or hormonal therapy were not included.

All imaging was performed on one of the four different 3T scanners (126 patients on Trio, 255 patients on Skyra, 17 patients on Prisma, and 19 patients on Verio; Siemens Healthcare, Erlangen, Germany) with the standardized clinical mp-MRI protocol, including T2w and DWI. We excluded the DCE-MRI for our study because of the limited role in the current diagnostic practice [WBC16, VHG16, KCJ18]. We used axial T2w turbo spin-echo (TSE) imaging and maps of the apparent diffusion coefficient (ADC) using echo-planar imaging (EPI) DWI sequence. For T2w, the repetition time (TR) and echo time (TE) of the T2w TSE were 3800-5040 ms and 101 ms, respectively. With a 14 cm FOV and a matrix size of 256 × 205, we acquired and reconstructed T2w TSE images with 0.55 mm × 0.68 mm in-plane resolution and 3 mm through-plane resolution with no gaps. For DWI, we used TR and TE of 4800 ms and 80 ms. With FOV of 21 cm × 26 cm and matrix of 94 × 160, DWI images were reconstructed with in-plane resolution of 1.6 mm^2 and a slice thickness of 3.6 mm. The ADC maps were obtained 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).

The mp-MRI exams were reviewed by three genitourinary (GU) radiologists (10+ years of clinical prostate MRI reading) as part of the standard clinical care. The findings with PI-RADS score$\geq$3 were reported and considered to be MRI-positive findings. The rest of the findings with PI-RADS$\leq$2 were considered to be MRI-negatives in this study.

3.2.2 Whole-mount histopathology matching & annotation

As in Fig. 3.1, the groundtruth of this study was lesion confirmation on whole-mount histopathology after RALP. The excised prostate was sliced from apex to base with 4-5 mm increment at the approximated mp-MRI orientation. Histopathology examinations of whole-mount specimens were performed by GU pathologists, blinded to all MRI information.

Later, at least one GU radiologist and one GU pathologist re-reviewed mp-MRI and histopathology examinations together at a multidisciplinary meeting scheduled monthly. Each ROI in MRI was matched to the corresponding location on the specimen through visual co-registration. MRI-positive findings were considered to be either true positive if they were in the same quadrant (left and right, anterior and posterior) and in the appropriate segment (base, midgland, and apex) on both mp-MRI and histopathology, or false positive if no corresponding lesions were found on the histopathology.

After the multidisciplinary meeting, GU radiology research fellows, supervised by GU radiologists, retrospectively reviewed each mp-MRI exam, referring to whole-mount histopathology, and annotated all MRI-visible lesions. 69.5% (278 out of 400) of prospectively missed (false negative) lesions were retrospectively identified in the review and were annotated. The MRI non-visible lesions were not included in this study due to the difficulty of the annotation.

Overall, we have annotated 728 lesions, consisting of 286 GS 3+3 lesions, 270 GS 3+4 lesions, 110 GS 4+3 lesions, 30 GS=8 lesions, and 32 GS$\geq$9 lesions. Among these, 93 GS 3+3 lesions, 204 GS 3+4 lesions, 98 GS 4+3 lesions, 26 GS=8 lesions, and 29 GS$\geq$9 lesions were prospectively identified by radiologists. All annotations were on T2w. The index lesion was defined as the lesion with the highest GS or the largest diameter when multiple lesions had the same grade on the histopathology, and clinically significant lesions were lesions with GS$\geq$7 [PEM11].

Table 3.1: Gleason score encoding for multi-class CNNs.

Label	Class	One-hot encoding	Ordinal encoding
Non-lesion	0	1 0 0 0 0 0	0 0 0 0 0
GS 3+3	1	0 1 0 0 0 0	1 0 0 0 0
GS 3+4	2	0 0 1 0 0 0	1 1 0 0 0
GS 4+3	3	0 0 0 1 0 0	1 1 1 0 0
GS = 8	4	0 0 0 0 1 0	1 1 1 1 0
GS $\geq$ 9	5	0 0 0 0 0 1	1 1 1 1 1

3.2.3 FocalNet for PCa detection and Gleason score prediction

FocalNet is an end-to-end multi-class CNN to jointly detect PCa lesions and predict their GS. As shown in Fig. 3.2, FocalNet takes the corresponding T2w and ADC slices into two imaging channels of the input and predicts for the pixel-level labels of the six classes: non-lesion, GS 3+3, GS 3+4, GS 4+3, GS=8, and GS$\geq$9. As in Fig. 3.3, the lesion groundtruth is first converted into a 5-channel groundtruth mask via ordinal encoding, and FocalNet predicts the groundtruth mask via its backbone CNN architecture. FocalNet is trained simultaneously by focal loss (FL) with regard to both T2w and ADC and mutual finding loss (MFL) for PCa features in either of the imaging components.

3.2.3.1 Ordinal encoding for Gleason scores

A conventional multi-class CNN encodes each label into a one-hot vector and predicts the one-hot vector through the multi-channel output [KSH12]. The six different labels can be converted into 6-bit one-hot vectors as in TABLE 3.1. One-hot encoding assumes that different labels are unrelated to each other, and thus the cross-entropy loss penalizes misclassifications equally. However, the progressiveness between different GS, such that the treatment prognosis of a GS 4+4 PCa is more similar to GS 4+3 than to GS 3+3 [EZS16], cannot be accounted for in one-hot encoding. In addition, by dividing lesions into separate classes, the number of samples in each class is very limited.

We instead convert labels from six classes into 5-bit ordinal vectors using ordinal encoding [CWP08, GPS16]. As shown in TABLE 3.1, each bit of an ordinal vector identifies a non-mutually-exclusive condition, such that the k-th bit indicates whether the label is from a class greater or equal to k. In this way, the groundtruth is encoded into a 5-channel mask, e.g., the first channel is the mask for all lesions, the second channel is the mask for clinically significant lesions, etc. Then, the CNN predicts for the encoded mask using the 5-channel output, and a sigmoid function is applied on top of each output channel to normalize the output into the prediction probability from 0 to 1. I.e., the first output channel naturally predicts for lesion detection probabilities.

Given the predicted ordinal encoded vector for a pixel, $\hat{y} = \big(\hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4, \hat{y}_5\big) \in \{0, 1\}$, the predicted class is the highest class k such that $\hat{y}_i = 1\ \forall i \leq k$, or non-lesion if $\hat{y}_i = 0\ \forall i$. The predicted class is written alternatively as $\max_{1 \leq k \leq 5} \left(\prod_{i=1}^{k} \hat{y}_i\right) \left(\sum_{i=1}^{k} \hat{y}_i\right)$.

The ordinal encoding characterizes the relationships between different labels. E.g., GS=8 shares 4 bits in common with GS 4+3, while only 1 bit with non-lesion. The commonness and differences between labels are represented as the shared and distinct bits in ordinal vectors. As a result, ordinal encoding allows the multi-class CNN to learn the commonness of all lesions and the distinctions between different GS at the same time. Besides, even though ordinal encoding does not increase the number of samples directly, it groups different labels so that each channel has a larger joint population of lesions compared with one-hot encoding.

3.2.3.2 Focal loss for ordinal encoding

PCa lesion and non-lesion labels are very imbalanced in the pixel-level groundtruth. In our dataset, non-lesion pixels outnumber lesion pixels by 62:1. After ordinal encoding for GS, the positive bit ratio of the groundtruth mask is only 0.77%. As a result, by accounting for lesion and non-lesion pixels evenly, the cross-entropy loss is occupied by the overwhelming amount of non-lesion terms, many of which are from easily predicted non-lesion pixels. Lesion-related terms, on the other hand, have little emphasis.

Alternatively, we deploy focal loss (FL) [LGG17] to balance the learning between lesion

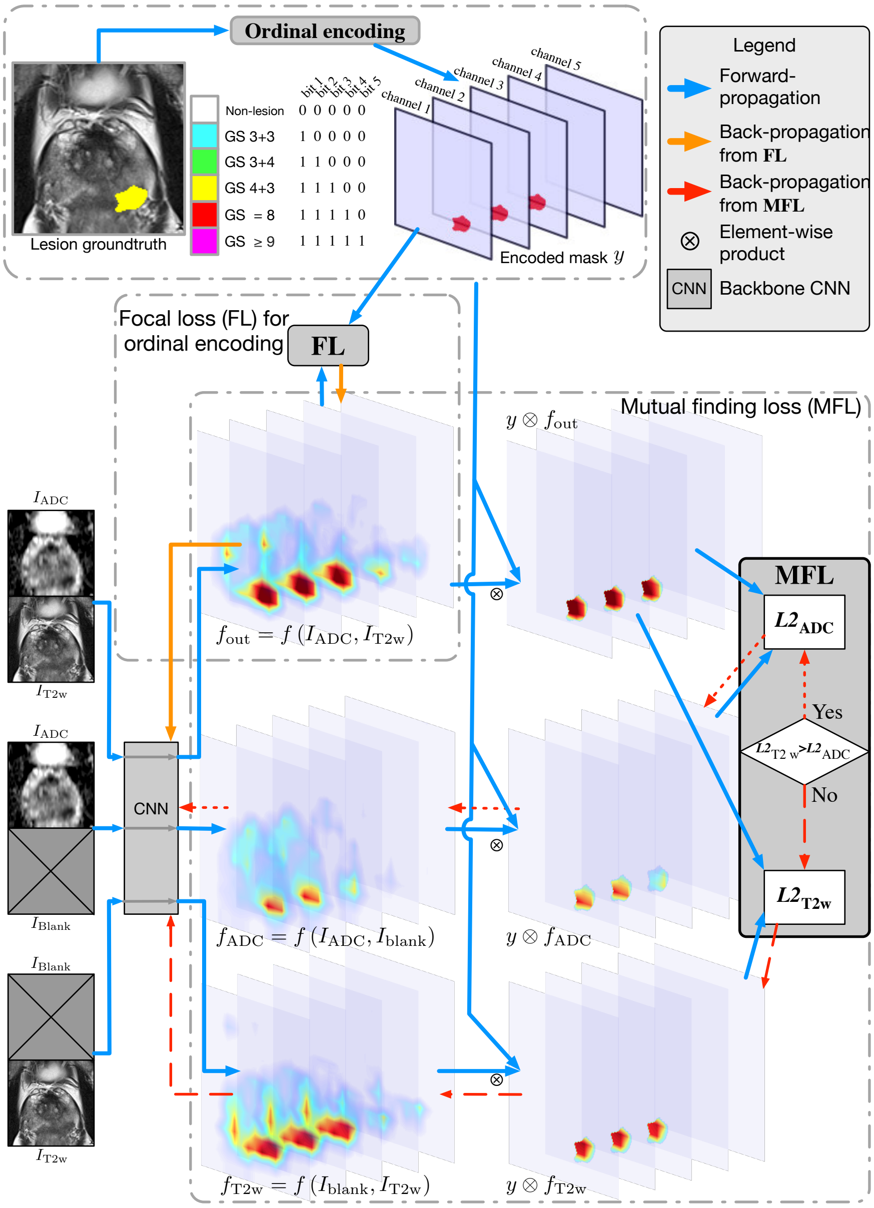

Ordinal encoding
Lesion groundtruth
bit 1 bit 2 bit 3 bit 4 bit 5
Non-lesion 0 0 0 0 0
GS 3+3 1 0 0 0 0
GS 3+4 1 1 0 0 0
GS 4+3 1 1 1 0 0
GS = 8 1 1 1 1 0
GS ≥ 9 1 1 1 1 1
channel 1
channel 2
channel 3
channel 4
channel 5
Encoded mask y
Legend
Forward-propagation
Back-propagation from FL
Back-propagation from MFL
⊗ Element-wise product
CNN Backbone CNN
Focal loss (FL) for ordinal encoding
FL
Mutual finding loss (MFL)
$y \otimes f_{\mathrm{out}}$
I_{ADC}
I_{T2w}
$f_{\mathrm{out}} = f\left(I_{\mathrm{ADC}}, I_{\mathrm{T2w}}\right)$
MFL
$L2_{\mathrm{ADC}}$
Yes
$L2_{\mathrm{T2w}} > L2_{\mathrm{ADC}}$
No
$L2_{\mathrm{T2w}}$
I_{ADC}
CNN
I_{Blank}
$f_{\mathrm{ADC}} = f\left(I_{\mathrm{ADC}}, I_{\mathrm{blank}}\right)$
$y \otimes f_{\mathrm{ADC}}$
I_{Blank}
I_{T2w}
$f_{\mathrm{T2w}} = f\left(I_{\mathrm{blank}}, I_{\mathrm{T2w}}\right)$
$y \otimes f_{\mathrm{T2w}}$

Figure 3.3: FocalNet for joint PCa detection and Gleason score prediction. The lesion groundtruth is converted into a 5-channel groundtruth mask using ordinal encoding. The CNN predicts the mask via its multi-channel pixel-level output. Focal loss (FL) trains the CNN with respect to f_{out} using both ADC and T2w inputs. Meanwhile, mutual finding loss (MFL) computes $L2_{\mathrm{ADC}}$ and $L2_{\mathrm{T2w}}$ in the forward-propagation and trains the imaging component of the smaller $L2$.

and non-lesion pixels. FL adds a focal weight of $(1-p_T)^2$ to the binary cross-entropy loss, where p_T is the prediction probability for the true class. Thereby, true predictions with high confidence contribute much less to the total loss [LGG17]. A common scenario during the training is that a clear non-lesion pixel (e.g., with high ADC intensity, or outside of prostate gland) receives 0.95 prediction probability for being non-lesion, which contributes 0.022 to the standard cross-entropy loss while only 5.6×10^{-5} to FL. By down-weighting easily predicted pixels, the training can be focused on suspicious or hard-to-predict pixels.

FL is further adapted to the ordinal encoding. For a given pixel, let $\vec{y} = \left(y_1, y_2, y_3, y_4, y_5\right) \in \{0,1\}$ be the groundtruth encoded vector corresponding to the 5-channel prediction probability vector $\vec{p} = \left(p_1, p_2, p_3, p_4, p_5\right) \in [0,1]$. Then, the FL for each pixel is

$$\mathrm{FL}\left(\vec{p}\right) = q\left(\vec{p}\right) \sum_{i=1}^{5} -\alpha y_i \log\left(p_i\right) - \left(1-\alpha\right)\left(1-y_i\right) \log\left(1-p_i\right). \tag{3.1}$$

q is the focal weight defined as the largest margin between the prediction probability and the groundtruth among the five channels, such that

$$q\left(\vec{p}\right) = \max_{1 \leq j \leq 5} y_j \left(1-p_j\right)^2 + \left(1-y_j\right) p_j^2. \tag{3.2}$$

In this way, high-grade lesions receive large focal weights if they are missed or downgraded, so that high-grade lesions can receive better attention for lesion detection as well.

Moreover, α is a constant that controls the penalty between false negative and false positive predictions. We find it is desirable to have a smaller penalty for false positives in PCa detection, since benign non-lesion findings, such as benign prostatic hyperplasia and benign adenomas, sometimes have a similar appearance to PCa lesions [MFI07]. Consequently, a large penalty for false positives hinders the learning of true positive PCa features. Besides, a max spatial pooling filter is applied to the focal weight q before the calculation of FL, in order to maintain consistent weights for positive and negative pixels around lesion boundaries. In our practice, α is set to 0.75 for better sensitivity, and the max pooling filter is sized to 3×3.

3.2.3.3 Mutual finding loss for multi-parametric imaging

During the interpretation of prostate mp-MRI, a radiologic finding is initially identified from a single component and later consolidated or rejected after referencing to other imaging components. The PI-RADS v2 score is then assessed primarily based on the finding's suspiciousness in the specific imaging component which describes the finding clearly [WBC16]. Hence, it is desirable for a CAD system to also determine PCa lesions from an individual imaging component as well as from the correspondence between multiple components of mp-MRI.

The underlying challenge is that *different components of mp-MRI capture distinct information and only a portion of the information is shared across all components.* As a result, findings observable in one component may be partially-/non-observable in the others. During the end-to-end training, a CNN with stacked imaging components can effectively learn the common features across components, but there is no mechanism to train for features observable only in a specific imaging component.

Mutual finding loss (MFL) is designed to identify the specific imaging component that contains distinct PCa features and train for the PCa features in the identified component. Firstly, given a training slice, MFL determines whether T2w or ADC alone can provide more information for the groundtruth lesion. As shown in Fig. 3.3, T2w and ADC are individually passed into the same CNN with a blank image with all zeros to substitute for the other component. We compare the CNN prediction output from ADC or T2w alone, $f_{\text{ADC}} = f\left(I_{\text{ADC}}, I_{\text{blank}}\right)$, $f_{\text{T2w}} = f\left(I_{\text{blank}}, I_{\text{T2w}}\right)$, with the output using both components, $f_{\text{out}} = f\left(I_{\text{ADC}}, I_{\text{T2w}}\right)$. The component resulting in a prediction output more similar to f_{out} on the groundtruth lesion region is considered to contain more PCa features. In this way, MFL selects a component to train for each slice.

Then, MFL trains the CNN so that lesion findings can be equivalently observed from the selected imaging component alone. Specifically, MFL minimizes the L2-distance on groundtruth mask y between f_{out} and the output using the selected component. I.e., $L2_{\text{ADC}} = \|y \otimes (f_{\text{out}} - f_{\text{ADC}})\|^2$ or $L2_{\text{T2w}} = \|y \otimes (f_{\text{out}} - f_{\text{T2w}})\|^2$, where $\otimes$ is the element-wise product. The L2-distance is calculated on the groundtruth lesion region while not on non-lesion regions, as MFL aims to train for PCa features. Since non-lesion regions are more likely to have the appearance similar to lesions from the observation of a single component than from both components, enforcing f_{ADC} or f_{T2w} to have the same non-lesion finding of f_{out} may counteract the training for PCa features. Moreover, f_{out} is utilized as a "soft" and adaptive truth reference to train for the specific component, compared with the groundtruth y. When the CNN cannot detect a barely visible lesion even with both components, f_{out} does not expect the CNN to learn the lesion using a single imaging component. Conversely, the CNN is trained for the certain PCa features in a single component if a lesion is clearly detected using both components.

As shown in Fig. 3.3, the process of MFL is summarized into a loss term for the end-to-end learning such that

$$\text{MFL} = \frac{1}{N} \min\{L2_{\text{ADC}}, L2_{\text{T2w}}\}, \tag{3.3}$$

where N is the total number of pixels of an image.

3.2.3.4 FocalNet training

FocalNet is trained by the combined loss from FL and MFL,

$$\text{L} = \mathbb{E}_{\vec{p} \sim S(f_{\text{out}})} \text{FL}\left(\vec{p}\right) + \lambda \cdot \text{MFL}, \tag{3.4}$$

where S is the sigmoid function and $\lambda = \frac{1}{\text{positive bit ratio}}$ is a constant weight to balance between FL and MFL. Besides, as in Fig. 3.3, the orange arrows indicate the back-propagation paths of FL, and the red arrows are back-propagation paths of MFL. MFL does not pass the gradient to f_{out} to train with respect to both imaging components, since f_{out} serves as a truth reference for f_{ADC} or f_{T2w} in MFL.

3.2.4 Pre-processing & Training

3.2.4.1 Registration

ADC images were registered to T2w images via rigid transformation using scanner coordinate information, as in [LWT13]. Since ADC and T2w sequences are temporally close to each other in our scanning protocol, we found minimal patient motion between ADC and T2w. Hence, as suggested in [WBT14], we did not utilize additional non-rigid registration. After the registration, for each patient, an 80 mm $\times$ 80 mm region centered on the prostate was identified manually and later resized to 128 $\times$ 128 pixels [KNT17].

3.2.4.2 Intensity normalization & variation

There are large intensity variations between mp-MRI exams with and without the usage of the endorectal coil, and, as a result, the commonly used normalization via histogram [TLW17] cannot work consistently. Instead, we clip the T2w intensity value by a lower threshold with the intensity of air and an upper threshold based on the intensity of bladder since 1) bladder is easy to locate programmatically, and 2) the intensity of bladder depends on water and is relatively consistent across patients. Then, we linearly normalize the clipped T2w intensity into $[0, 1]$ using the lower and upper thresholds. Moreover, as ADC is quantitative imaging and its intensity value is indicative of lesion detection and classification [VAF11,HSH11], we clip ADC intensity by patient-independent thresholds and normalize to $[0, 1]$. During the training, T2w intensity variation is applied to improve the CNN robustness to variable image intensity caused by the endorectal coil in some scans [RFB15]. The T2w upper-intensity threshold is randomly fluctuated in the estimated range that PCa lesions are detectable after the intensity normalization, which is empirically from -15% to +20%.

3.2.4.3 Implementations

The backbone CNN architecture of FocalNet is implemented using Deeplab [CPK18] with the 101-layer deep residual network [HZR16] on 2D image inputs. In the preliminary experiment, we also tested U-Net [RFB15] as the backbone CNN, but the training with U-Net commonly failed in early stages due to the model diverging, presumably caused by the incompatibility between FL and U-Net skip connections. Furthermore, pre-trained CNN weights from object classification task are applied as a weight initialization [HRG16]. The total loss is optimized by stochastic gradient descent with momentum 0.9 and L2-regularizer of weight 0.0001.

The learning rate starts at 0.001 with 0.7 decay every 2000 steps. The CNN is trained for 200 epochs with batch size 16. In addition to the T2w intensity variation, common image augmentations, including image shifting, scaling, and flipping, are also applied during the training. We did not apply image rotation, as a small angle rotation creates blurriness during interpolation. The image augmentations are performed for each batch of the training images and not for the validation images.

The image registration is implemented using the statistical parametric mapping toolbox [FHW94], and the pre-processing steps take around one minute for the images of each case. FocalNet is implemented using TensorFlow machine learning framework (Google; Mountain View, CA) [ABC16]. The average training time is 3-4 hours for each fold using a NVIDIA Titan Xp GPU with 12GB memory, and the prediction is relatively quick, about 0.5-1 second for each patient, due to the non-iterative nature of CNNs.

3.2.5 Validation

3.2.5.1 Cross-validation

We train and validate this study using 5-fold cross-validation. Each fold consists of 333 or 334 training cases and 84 or 83 cases for validation. In both training and validation, only annotated slices are included as in [KNT17], in order to minimize the chance for miss-annotated lesions. Each case contains 2 to 7 slices, and each fold of training and validation sets has around 1400 and 350 slices, respectively.

3.2.5.2 Lesion localization

For PCa detection, we extract lesion localization points from CNN pixel-level detection output as in [LDB14, WLC18]. For each case, we find 2D local maxima from the detection output of the slices. The trade-off between detection sensitivity and false detections is controlled by thresholding on the detection probabilities of the local maxima.

3.2.5.3 FROC for lesion detection

The lesion detection performance is evaluated through free-response receiver operating characteristics (FROC) analysis due to PCa's multi-focality [LDB14, TLW17]. FROC measures the lesion detection sensitivity versus the number of false positives per patient. True positive detections are localized points in or within 5 mm of lesion ROIs since PCa lesion diameters on the whole-mount specimen are roughly 10 mm greater than the corresponding ROIs in mp-MRI [PNK17]. False positive detections are localized points that are not true positive detections. Since our lesion groundtruth is annotated in 2D slices without the consideration of the 3D lesion volume, a localized point must be in the same slice of an ROI to be considered as a true detection. Lesion detection sensitivity is the number of detected lesions divided by the total number of visible lesions, including both the prospectively identified lesions and the prospectively missed lesions identified in the retrospective review described in Sec. 3.2.2. Because of the availability of whole-mount histopathology, the definition of true or false detection is more accurate than the studies only using biopsy cores.

Moreover, the lesion detection performance is further studied in fine-grained lesion groups as they have different detectabilities, i.e., FROC for lesion detection of each specific GS group. Under this setting, lesion detection sensitivity considers only lesions in a specific GS group. Lesions with GS=8 and GS$\geq$9 are grouped together since 1) either of them have very limited quantity in each fold of validation, and 2) the difference between their treatment is minimal.

3.2.5.4 ROC for Gleason score prediction

The GS prediction is evaluated by receiver operative characteristic (ROC) analysis. We group the multi-class classification into four binary classification tasks [FVW15]: 1) GS$\geq$7 vs. GS$<$7, 2) GS$\geq$4+3 vs. GS$\leq$3+4, 3) GS$\geq$8 vs. GS$<$8 and 4) GS$\geq$9 vs. GS$<$9. A voxel-level ROC is assessed for each task. Specifically, to mimic biopsy setting, twelve detection voxels were sampled for each case by finding the localized points as in Sec. 3.2.5.3. In a joint model for detection and classification, this setting evaluates classification performance without being affected by lesion misdetection, since if a lesion is completely missed by the

Table 3.2: False positives per patient (FP) at given detection sensitivity (Sen) for index lesions. avg±std.

	Index lesions	
	FP@Sen80%	FP@Sen90%
U-Net-Mult	1.19±0.39	1.74±0.49
U-Net-Sing	1.16±0.37	1.61±0.26
Deeplab	1.38±0.40	2.20±0.64
FocalNet	**0.61±0.25**	**1.02±0.37**

Table 3.3: False positives per patient (FP) at given detection sensitivity (Sen) for clinically significant lesions. avg±std.

	Clinically significant lesions	
	FP@Sen80%	FP@Sen90%
U-Net-Mult	1.39±0.36	2.15±0.60
U-Net-Sing	1.21±0.20	1.75±0.55
Deeplab	1.45±0.43	2.44±0.80
FocalNet	**0.65±0.15**	**1.13±0.35**

model, the classification result for the lesion is meaningless as well.

3.2.5.5 Comparison to radiologists

We compare FocalNet with the prospective clinical performance of radiologists for lesion detection. Radiologist performance is assessed on the entire 417 cases grouped by the five validation sets. Radiologist's findings were determined to be true or false positives as described in Sec. 3.2.2. The sensitivity is calculated on the number of true positive findings versus the total number of MRI-visible lesions.

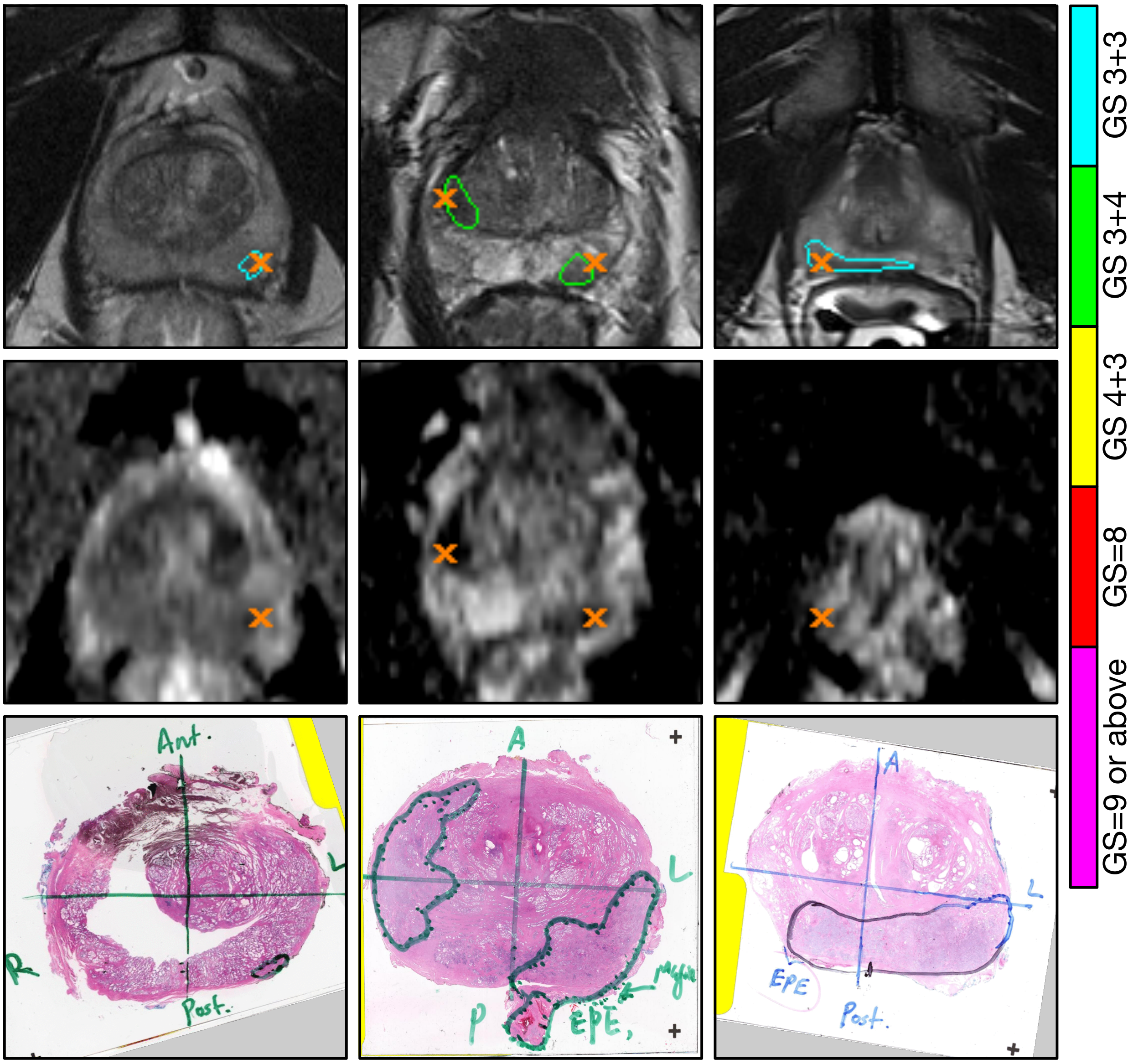


Figure 3.4: From top to down shows T2w images, ADC images, and whole-mount specimens. Lesion detection points from FocalNet are shown as the orange cross signs. Groundtruth lesion contours overlay on T2w images with the colors corresponding to their Gleason score groups.

3.3 Results

3.3.1 Baseline methods

Deeplab, *U-Net-Mult*, and *U-Net-Sing* are the three baseline methods in this study. *Deeplab* [CPK18] is the base model of FocalNet, which has the same backbone CNN architecture

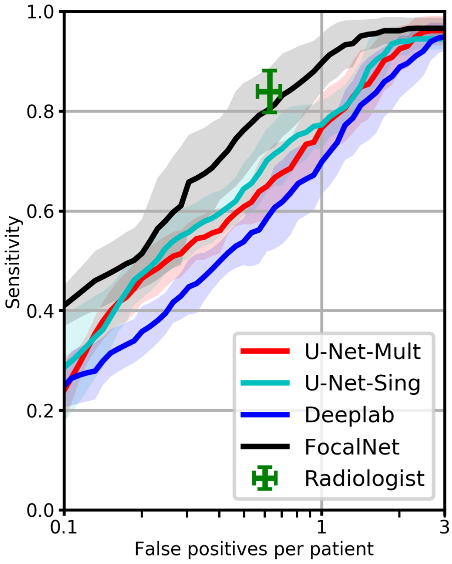


Figure 3.5: FROC analysis for detection sensitivity for index lesions, based on 5-fold cross–validation. The number of false positives per patient (x-axis) is shown on log-scale. The transparent areas are 95% confidence intervals estimated by two times of the standard deviation. The green markers indicate radiologist's performance with a 95% confidence intervals also estimated by two times of the standard deviation.

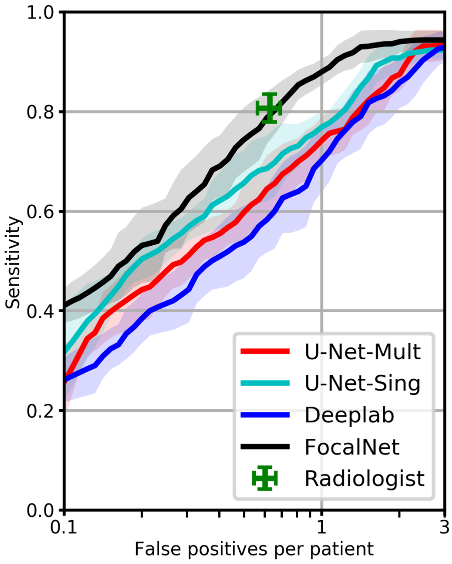


Figure 3.6: FROC analysis for detection sensitivity for clinically significant lesions, based on 5-fold cross-validation.

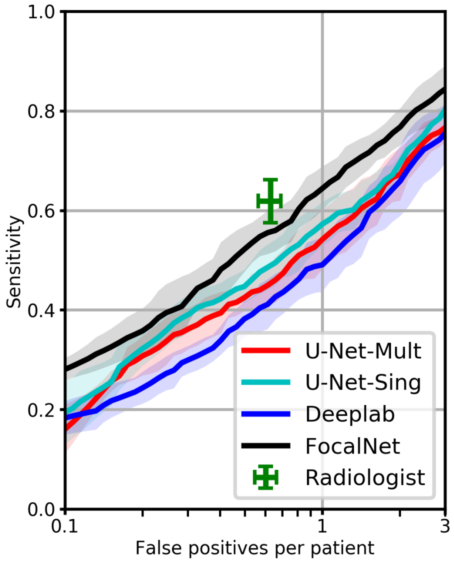


Figure 3.7: FROC analysis for detection sensitivity for all lesions, based on 5-fold cross-validation.

Table 3.4: False positives per patient (FP) at given detection sensitivity (Sen) for all lesions. avg±std.

	All lesions	
	FP@Sen60%	FP@Sen80%
U-Net-Mult	1.38±0.41	3.53±0.41
U-Net-Sing	1.29±0.39	2.98±0.18
Deeplab	1.55±0.46	3.70±1.04
FocalNet	**0.80±0.21**	**2.30±0.61**

of FocalNet with one-hot encoding for six classes, i.e., five GS groups and non-lesion. The same pre-trained weight initialization is applied for *Deeplab* as for FocalNet. U-Net [RFB15] is a popular CNN architecture for various biomedical imaging segmentation tasks. Multi-class U-Net (*U-Net-Mult*) is trained to detect and classify lesions using one-hot encoding as in *Deeplab*. Single-class U-Net (*U-Net-Sing*) is trained for a simplified task to detect lesions only, regardless of their GS. To enable a fair comparison, the training and validation workflows in Fig. 3.2, consisting of image pre-processing, intensity normalization & variation

Table 3.5: False positives per patient (FP) at given lesion detection sensitivity (Sen) for each specific Gleason score group. avg±std.

	GS 3+3		GS 3+4		GS 4+3		GS≥8	
	FP@Sen60%	FP@Sen70%	FP@Sen80%	FP@Sen90%	FP@Sen80%	FP@Sen90%	FP@Sen80%	FP@Sen90%
U-Net-Mult	1.65±0.51	2.16±0.68	1.19±0.32	1.74±0.82	0.12±0.11	0.28±0.26	0.04±0.03	0.10±0.10
U-Net-Sing	1.45±0.27	1.97±1.14	0.86±0.24	1.59±1.48	0.11±0.10	**0.23±0.19**	0.08±0.09	0.210±0.143
Deeplab	1.41±0.81	2.46±1.13	1.13±0.34	1.82±0.50	0.27±0.11	0.40±0.32	0.06±0.02	0.24±0.19
FocalNet	**1.21±0.48**	**1.76±0.63**	**0.58±0.18**	**0.90±0.78**	**0.07±0.11**	0.23±0.14	**0.04± 0.02**	**0.06±0.07**

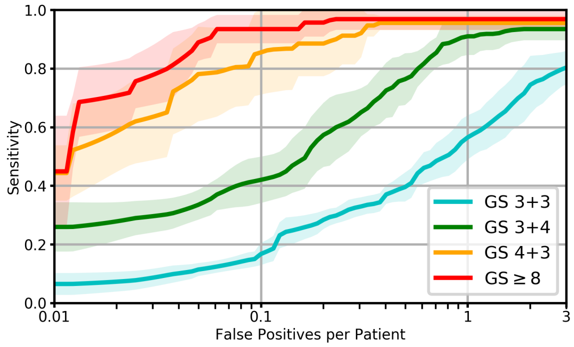
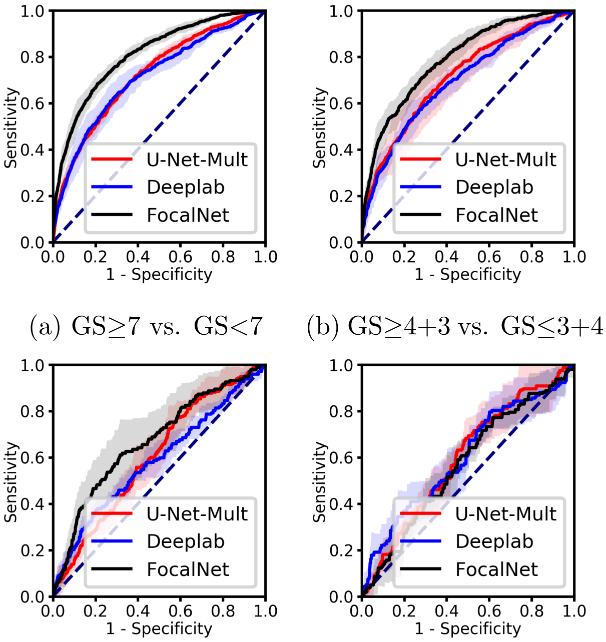


Figure 3.8: FROC analysis for detection sensitivity of FocalNet for each specific Gleason score group. Transparent areas are 95% confidence intervals estimated by two times of the standard deviation. The results of baseline methods are reported in TABLE 3.5.

and image augmentation procedures, are applied equally to all methods. Under the cross-validation setting, the p-values are obtained by two-sample Welch's t-test, with the alpha level adjusted by Bonferroni correction for multiple comparisons.

3.3.2 Lesion detection

Fig. 3.7 shows the FROC analysis for index lesions, clinically significant lesions, and all lesions, respectively, and examples for lesion detection are shown in Fig. 3.4. As in TABLE 3.2, FocalNet achieved 90% sensitivity for index lesion at the cost of 1.02 false positives per patient, while *U-Net-Sing* and *Deeplab* triggered 54.3% and 116.8% more false detections, respectively, for the same sensitivity. Furthermore, as in Fig. 3.2.5.5, FocalNet detected 87.9% clinically significant lesions at 1 false positive per patient, outperforming the best baseline, *U-Net-Sing*, by 11.1%. The partial area under the curve between 0.01 to 1 and 0.1 to 3 false positives per patient for FocalNet are 0.685±0.056 and 2.570±0.101, respectively, which are higher than *U-Net-Sing* (0.596±0.061, 2.402±0.106). Moreover, as in Fig. 3.2.5.5, the sensitivity for all PCa lesions detection is 64.4% at 1 false positive per patient, while *U-Net-Sing* required 1.65 false positives per patient for the same sensitivity. FocalNet reached its maximum sensitivity of 89.3% at 4.64 false positives per patient, in comparison to *U-Net-Sing*'s maximum sensitivity of 84.7% at similar false positives per patient.



(a) GS≥7 vs. GS<7 (b) GS≥4+3 vs. GS≤3+4

(c) GS≥8 vs. GS<8 (d) GS≥9 vs. GS<9

Figure 3.9: ROC analysis for Gleason score classification. Transparent areas are 95% confidence intervals estimated by two times of the standard deviation. *U-Net-Sing* is not in this comparison since *U-Net-Sing* does not classify for Gleason scores.

The radiologist performance is shown in Fig. 3.7 as green markers. Radiologists achieved 83.9% sensitivity for index lesions, 80.7% sensitivity for clinically significant lesions, and 61.8% sensitivity for all lesions, with 0.62 false positives per patient. The radiologist detection sensitivity for index lesions, clinically significant lesions, and all lesions is, respectively, 3.4%, 1.5%, and 6.2% higher than FocalNet at the same false positives per patient.

Lesion detection sensitivity for lesions of each specific GS group is reported in Fig. 3.8 and TABLE 3.5. Both FocalNet and baseline methods had high sensitivity for lesions with GS≥4+3. FocalNet reached 95.3% and 96.8% sensitivity for GS 4+3 and GS≥8 at 0.231 and 0.377 false positives per patient, respectively. FocalNet outperformed baseline methods for the detection of GS 3+4 lesions. At 0.5 and 1 false positive per patient, FocalNet respectively received 76.4% and 91.0% sensitivity for GS 3+4, which are 7.7% and 6.3% higher than *U-Net-Sing*, 15.1% and 16.9% higher than *U-Net-Mult*, and 16.1% and 14.3% higher than *Deeplab*.

3.3.3 Gleason score prediction

Fig. 3.9a and Fig. 3.9b show the ROC analysis for GS≥7 vs. GS<7 and GS≥4+3 vs. GS≤3+4. FocalNet achieved ROC area under the curve (AUC) 0.81±0.01 and 0.79±0.01, respectively in 5-fold cross-validation, in comparison to *U-Net-Mult* (0.72±0.01 and 0.71±0.03) and *Deeplab* (0.71±0.02 and 0.70±0.02). FocalNet achieved AUC significantly higher than *U-Net-Mult* (p<0.0005) and *Deeplab* (p<0.01) for clinically significant lesion (GS≥7) classification. However, as in Fig. 3.9c and Fig. 3.9d, both FocalNet and baseline methods exhibited limited capabilities of classifying GS≥8 vs. GS<8 and GS≥9 vs. GS<9. FocalNet has ROC AUC 0.67±0.04, and 0.57±0.02 respectively, not significantly different from *U-Net-Mult* (0.60±0.03, and 0.60±0.03) and *Deeplab* (0.59±0.01, and 0.60±0.04).

3.3.4 Loss contribution

We trained FocalNet with different loss combinations to understand their contributions to PCa detection performance. Under the same setting, we specifically compared three different losses: cross-entropy loss (*CE*), focal loss (*FL*), and the combined loss from FL and MFL (*FL+MFL*) described in 3.2.3.4. As shown in Fig. 3.10, *CE* had only 62.9% lesion detection sensitivity at 1 false positive per patient, as the cross-entropy loss was dominated by non-cancerous pixels during the training. *FL* showed its effectiveness for the imbalanced labels and improved the detection sensitivity by more than 15% from *CE* in range of 0.05 to 1.42 false positives per patient. The combination of FL and MFL (*FL+MFL*) further improved the lesion detection sensitivity from *CE* and *FL* respectively by 30.3%, 14.2% at 0.5 false positives per patient and by 25.0%, 8.1% at 1 false positive per patient. We also noted that the detection performance of *CE* was marginally lower than *Deeplab* reported in Fig. 3.2.5.5, as the ordinal encoding strategy caused the labels to become more imbalanced for *CE*.

3.3.5 Image augmentation

As image augmentation is non-trivial for training a CNN when the number of training data is limited, we compared three different augmentation strategies in the context of PCa detection:

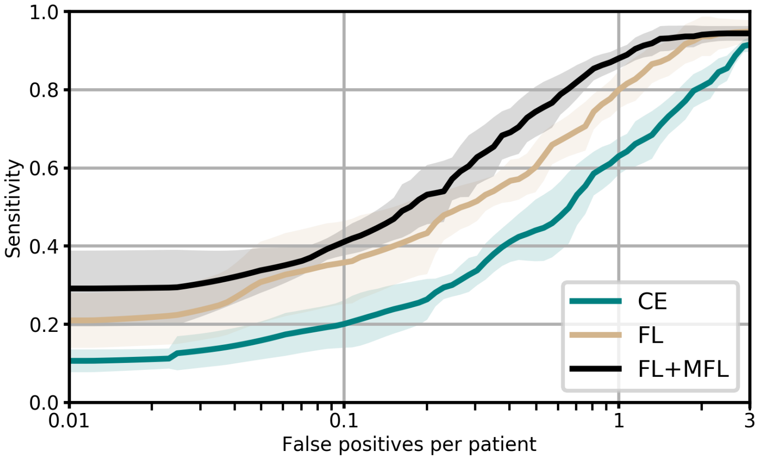


Figure 3.10: FROC analysis for the detection of clinically significant lesions using three different loss combinations during the training: cross-entropy loss (*CE*), focal loss (*FL*), and the combined loss from focal loss and mutual finding loss (*FL+MFL*). The number of false positives per patient (x-axis) is shown on log-scale. The transparent areas are 95% confidence intervals estimated by two times of the standard deviation.

training without augmentation, with basic augmentation, and with advanced augmentation. The basic augmentation included image shifting, scaling, and flipping, while the advanced augmentation additionally includes intensity variation as described in Sec. 3.2.4.2. As shown in Fig. 3.11, the advanced augmentation strategy became effective as false positives per patient become higher (>0.24), and the basic augmentation was ineffective when the number of false positives per patient was greater than 0.75. The sensitivity with the advanced augmentation strategy was 9.8% higher than the one with the basic augmentation at 1 false positive per patient. This suggests that applying random intensity variation during training improves the detection of hard-to-spot lesions rather than easy-to-spot lesions. This would be particularly important when there exist strong intensity variations caused by the endorectal coil.

3.4 Discussion

We compared FocalNet with the prospective clinical performance of radiologists for lesion detection and did not find differences with statistical significance. The radiologists following PI-RADS v2 achieved 83.9% and 80.7% sensitivity for the detection of histopathology-proven index lesions and clinically significant lesions. FocalNet had slightly lower, 80.5% and 79.2% sensitivity at the same false positives per patient, which were not significantly different from the radiologist performance (p=0.53 and p=0.66). Our prostate mp-MRI exams were interpreted and scored by expert GU radiologists who have 10+ years of post-fellowship experience and read more than 1,000 prostate MRI exams yearly. Hence, the reported radiologist performance is expected to reflect or to be close to the upper limit of prostate MRI reading quality under the current guideline. As prostate MRI reading quality largely varies according to reader's experience [RWB12], FocalNet can potentially assist less experienced readers or augment the PCa detection task for non-experts. In addition, the direct numerical comparisons between FocalNet and the radiologist performance may include some bias due to their different definitions for true and false detection. The true positives for FocalNet are defined as localized detection points in or within 5mm of the lesion ROIs, while the true positives for the radiologist performance are defined as lesions in the same quadrant and in the appropriate segment, as described in Sec. 3.2.2. This is mainly because PI-RADS is designed for the clinical interpretation, not for the specific detection task.

The handling of multi-parametric imaging information was previously explored. Wang *et al.* [WLC18] proposed to use separate CNNs for individual imaging components of mp-MRI and enforced the consistency between different outputs of the imaging components. Fidon *et al.* [FLG17] designed the ScaleNet block to extract multi-component features and single-component features. In comparison, MFL does not rely on the strong assumption of the consistency across all imaging components. Instead, inspired by the clinical interpretation of prostate mp-MRI, MFL identifies the most distinctive imaging features from one or certain components of mp-MRI and trains the CNN together with FL for both single and multiple imaging component knowledge at the same time, with minimal changes to the existing CNNs

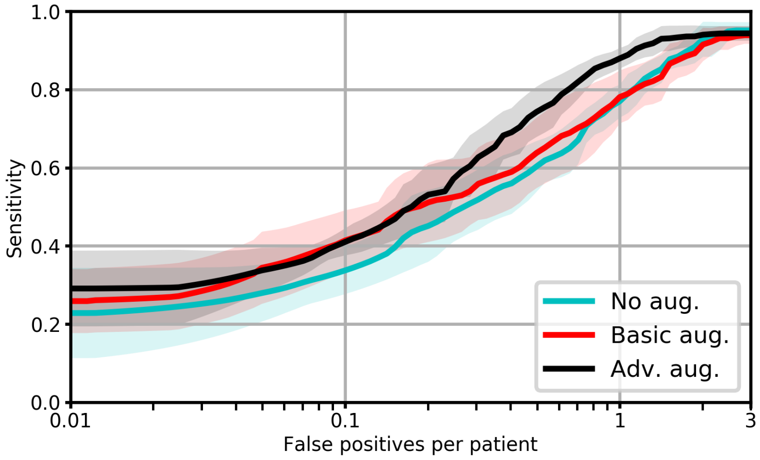


Figure 3.11: FROC analysis for the detection of clinically significant lesions under three different augmentation strategies during the training: with no augmentation, with basic augmentation (image shifting, scaling and flipping), and with advanced augmentation (basic augmentation + intensity variation). Transparent areas are 95% confidence intervals estimated by two times of the standard deviation.

and no additional parameters to train.

We demonstrated FocalNet with two imaging components of mp-MRI. MFL can be extended to multiple imaging components, such that

$$\mathrm{MFL} = \min_{1 \leq i \leq m} \|y \otimes (f_{\mathrm{out}} - f_i)\|^2, \tag{3.5}$$

where f_{out} is the CNN output with all components, m is the number of imaging component subsets, and f_i is the CNN output using the i-th subset of imaging components. However, each additional imaging component will require extra GPU memory and create considerable computation overhead during the training, since every imaging component subset requires one forward-propagation of the CNN for the calculation of MFL as shown in Fig. 3.3. It is hence impractical to account for a large number of imaging components. An alternative approach to reducing the computational cost would be to utilize pre-determined combinations of imaging components, similar to PI-RADS v2 [WBC16], and to consider only these as possible subsets of imaging components to train with MFL.

Furthermore, FocalNet can be adapted for the PCa lesion segmentation task [OLL10].

As the first output channel of the FocalNet predicts for lesion vs. non-lesion, additional post-processing methods (e.g., simple thresholding, fully-connected conditional random field [KK11], etc.) can be applied on the predicted probabilities for the lesion segmentation.

We used a 2D CNN instead of a 3D CNN for prostate mp-MRI since 1) the imaging is non-isotropic in our protocol, 2) 3D PCa lesion annotations are error-prone due to the difficulty of prostate mp-MRI interpretation, and 3) a 3D CNN has more parameters and thus requires more training samples. Nevertheless, FocalNet is not limited to 2D CNNs. In some other domains (e.g., brain imaging), 3D CNNs may be more suitable for lesion detection or segmentation as 3D CNNs can fully benefit from the volumetric spatial information.

FocalNet can be further improved by combining the voxel-level predictions with a region-level GS classifier. Similar to previous works [TKM13,FVW15], we can build the region-level classification models to classify GS for candidate regions provided by the output from FocalNet's lesion detection. This hybrid approach can potentially improve the GS classification performance since region-based classifiers provide additional robustness to pixel-level classifications.

The prediction of fine-grained GS groups is an early attempt to apply multi-class CNN models to explore the correlation between mp-MRI and PCa aggressiveness. The ordinal encoding for GS is used under the assumption that different PCa aggressiveness on microscopic tumor structure exhibit both similarities and distinctions in mp-MRI as suggested by [PJY13,VAF11]. Further study is needed to consolidate the correlation between mp-MRI and PCa aggressiveness, particularly with available molecular subtypes of PCa [LLH04].

The accurate groundtruth lesion annotation is one of the key challenges for PCa CAD systems. Many studies used mp-MRI exams with biopsy-confirmed lesion findings as the groundtruth [OLL10, WBT14, KXW15], which could potentially include some inaccuracies because of the discrepancy between prostate biopsy and radical prostatectomy in histologic findings. Recently, the ProstateX Challenge [LDB14] has attempted to improve the inaccurate lesion annotations by using MR-guided biopsy as the groundtruth. This will reduce the chances of lesion misdetection and GS upgrading/downgrading due to the biopsy needle misplacement, but the MR-guided biopsy confirmations may still include the inaccurate histologic finding [LSB14] and do not provide the information of the exact shape, location, and size of the lesions. Here, we annotated lesions based on whole-mount histopathology specimens from radical prostatectomy, providing the most accurate lesion characterizations.

Our study did not include MRI non-visible lesions because 1) they are difficult to annotate via visual co-registration from whole-mount histopathology, and 2) it is hard to confirm whether the imaging plane sufficiently contains the lesion information at the time of MRI scan. Future study may investigate rigid registration between whole-mount slices and mp-MRI imaging, which enables a direct correlation between histopathology and mp-MRI. The discovery of lesions not detectable by human eyes from mp-MRI can further extend the utility of machine learning in clinical practice.

In conclusion, we proposed a novel multi-class CNN, FocalNet, consisting of mutual finding loss to fully utilize distinctive knowledge from multi-parametric MRI and ordinal encoding to preserve the progressiveness between labels in a multi-class CNN. We used FocalNet to jointly detect prostate cancer and predict the fine-grained Gleason score groups. We trained and validated FocalNet under 5-fold cross-validation using 417 pre-operative mp-MRI exams with annotations of all MRI-visible PCa lesions on whole-mount histopathology. For the detection of histopathology-proven index lesions and clinically significant lesions, FocalNet achieved 89.7% and 87.9% sensitivity at 1 false positive per patient and received sensitivity only 3.4% and 1.5% lower than experienced radiologists using PI-RADS v2. FocalNet also outperformed all three CNN-based baseline methods, with an AUC of 0.809 for the classification of clinically significant PCa.

CHAPTER 4

Prostate Cancer Inference via a Large Collection of Negative Prostate Multi-parametric MRI

Multi-parametric MRI (mp-MRI) is the best non-invasive diagnostic tool for prostate cancer (PCa). However, the computer-aided diagnosis systems for PCa are often constrained by the limited access to accurate lesion groundtruth annotations for training. Here, we propose the baseline MRI model to alternatively learn the appearance of mp-MRI that is negative for PCa. The baseline MRI model is trained by MRI-negative scans only, without any PCa annotations. After training, the baseline MRI model synthesizes specified image regions based on an input image, and the synthesized regions are negative for PCa. We then utilize the baseline MRI model to infer the pixel-wise suspiciousness map of PCa for the testing image. We trained the baseline MRI model with 1095 negative prostate mp-MRI scans. For the evaluation, we collected a separated dataset of 116 pre-operative mp-MRI scans with annotated lesion regions of interest (ROIs) confirmed with post-surgical whole-gland specimens. The suspiciousness map was evaluated by receiver operating characteristic (ROC) analysis for PCa lesions versus non-PCa regions classification and free-response receiver operating characteristic (FROC) analysis for PCa localization. Our proposed method received 0.84 area under the ROC curve and 77.0% sensitivity at 1 false positive per patient in FROC analysis.

4.1 Introduction

Prostate cancer (PCa) is one of the most common cancer-related diseases among men in the United States [SMJ19]. Multi-parametric MRI (mp-MRI) is a powerful, non-invasive diagnosis tool for PCa, and T2-weighted imaging (T2W) and diffusion-weighted imaging (DWI) are the key components containing structural and functional information for the PCa diagnosis. However, the interpretation of mp-MRI is highly challenging because of the qualitative or semi-quantitative assessment of the imaging [WBC16].

Recent studies have explored quantitative interpretations of mp-MRI by training machine learning models [LDB14, SZY18, TLW17, WLC18, FVW15]. Most of the models were trained under strong supervision using the lesion annotations as the groundtruth, and thus the performance of the models is dependent on both quantity and quality of training data associated with groundtruth annotations. However, the findings from mp-MRI are not easy to be fully integrated with histologic findings due to misregistration or insufficient histologic information, resulting in a limited number or quality of groundtruth annotations available. Litjens *et al.* used MR-guided biopsy dataset to identify biopsy-confirmed lesions in MRI [LDB14], and Fehr *et al.* annotated PCa region of interest (ROI) using post-surgical whole-gland specimens as a reference [FVW15]. Nevertheless, both of these studies use limited numbers of cases (348 and 147 cases, respectively) due to the availability of histologic results.

Despite the difficulty of obtaining accurate lesion annotations, the number of mp-MRI scans has been increased in years as mp-MRI gains clinical acceptance for PCa diagnosis. A large portion of mp-MRI scans are reported as negative for PCa, and the MRI-negative is shown to be reliable without the need for histologic confirmations [HDW15]. Thus, the collection of negative prostate mp-MRI scans in a large quantity is more plausible than having a large number of MRI-positive scans with accurate lesion annotations. While the MRI-negative scans are vastly available, the existing PCa detection models cannot solely learn from the MRI-negative scans since they need to be trained with PCa lesion annotations to differentiate between PCa lesions and non-PCa regions.

Alternatively, we propose the baseline MRI model that learns the appearance of the prostate from the MRI-negative scans. The baseline MRI model is implemented as a convolutional neural network (CNN), and it synthesizes a region of an image using the rest of the image as the input. As the baseline MRI model is trained with only MRI-negative scans, the synthesized region is also MRI-negative for PCa given any input image. The region to synthesize is specified by a collection of ROI candidates, which describes the common locations and shapes of PCa lesions, so that the baseline MRI model can focus on the normal tissue appearance in the areas prone to PCa. Then, we use the trained baseline MRI model to infer the PCa suspiciousness map. Given a testing image, the baseline MRI model synthesizes for different regions from the collection of ROI candidates, and then the suspiciousness map is summarized by comparing the original image regions and the synthesized image regions.

We summarize our contributions as follows. We proposed the baseline MRI model to infer the pixel-wise suspiciousness map of prostate cancer in mp-MRI through unsupervised learning, without the need for PCa annotations during training. We trained the baseline MRI model using 1095 mp-MRI scans negative for PCa, which were identified from 3127 total collected mp-MRI scans. We evaluated the proposed prostate cancer inference in a separate dataset of highly curated 116 pre-operative mp-MRI scans with annotated PCa ROI confirmed with histologic whole-gland specimens. Our method achieved similar lesion localization performance as the previously reported fully-supervised methods.

4.2 Materials and methods

4.2.1 Negative prostate MRI dataset

With IRB approval, we collected 3127 3 Tesla (3T) prostate mp-MRI scans from 2016 to 2018 at a single institution. The mp-MRI scans were prospectively reported by genitourinary (GU) radiologists following the standardized clinical guideline. We excluded the scans 1) with the endorectal coil, 2) immediately after biopsy, and 3) for patients underwent prior radiotherapy, hormonal therapy, or surgery. We parsed the plain text radiology reports and identified MRI-negative cases having 1) no suspicious target in the finding section and 2) no more than mildly suspicious in the impression section. We manually examined a random subset of reports to ensure the correctness of the identification. A total of 1261 MRI-negative scans were identified, and we divided them into training, validation, and testing sets containing 1095, 50, and 116 cases, respectively.

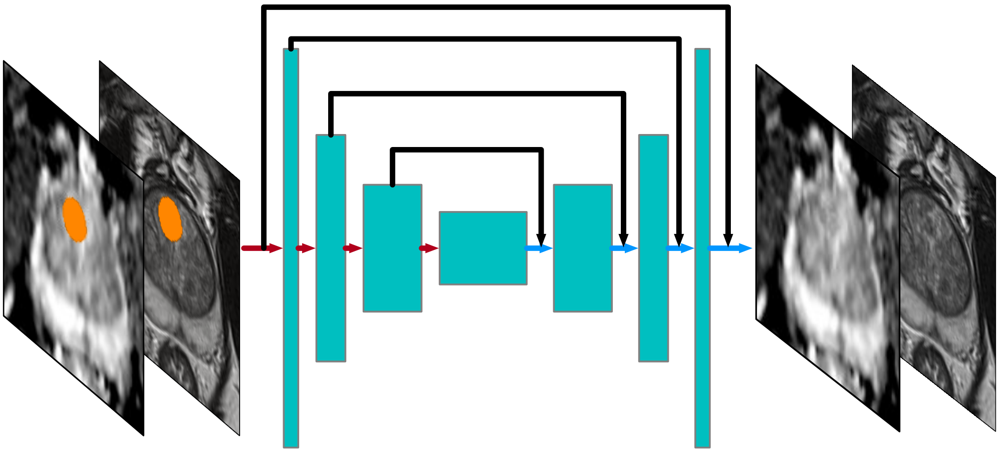


Figure 4.1: The baseline MRI model synthesizes for the region M (shown in orange) unobserved in the input.

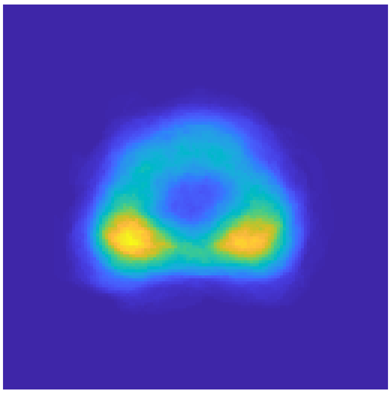


Figure 4.2: The prevalence map of PCa from the collection of ROI candidates [JRM18].

For each scan, the axial turbo spin-echo (TSE) T2W (TR/TE, 3800-5040/101ms; FOV, 14×14cm^2; matrix, 256×205; slice thickness, 3 mm; no gap) and maps of apparent diffusion coefficient (ADC) single-shot echo-planar imaging (SS-EPI) DWI (TR/TE, 4800/80ms; FOV, 21×26cm^2; matrix, 130×160; slice thickness, 3.6 mm; b-values, 0/100/400/800 s/mm^2) were used. ADC was registered into T2W, with 0.625×0.625mm^2 in-plane resolution and 3mm through-plane resolution. Both T2W and ADC were cropped into a small field-of-view (8×8cm^2) to improve the model convergence. Four consecutive slices around mid and base gland were selected for each scan, resulting in the total of 4380 slices for training.

4.2.2 Baseline MRI model

The proposed baseline MRI model aims to synthesize the mp-MRI negative for PCa, with respect to an input image. Instead of generating for an entire MRI-negative image at once, each time the baseline MRI model, f, synthesizes for a specified region of an image, MI, using the rest of the image, $(1 - M)\, I$, observed as the input, where M is the region in binary mask form and I is the input image. Specifically, as in Fig. 4.2.1, $Mf\big((1 - M)\, I; \theta\big) \rightarrow MI$, where $I = (I_{\mathrm{T2W}}, I_{\mathrm{ADC}})$ is the combined image of the corresponding T2W and ADC stacked as in different image channels, and θ is the trainable weights of the baseline MRI model. A collection of ROI candidates is used to specify M. We train the baseline MRI model using only MRI-negative scans. In this way, the baseline MRI model learns the various appearance of mp-MRI negative for PCa in training.

We use a U-Net CNN structure [RFB15] for the baseline MRI model since the encoder-decoder design of U-Net helps to summarize the global anatomical information [DGS18], and the skip connections from U-Net simplify the training for observed regions, s.t., the observed input feeds directly into the last decoding layer without going through the encoder-decoder. Besides, we use partial convolutional layers instead of conventional convolutional layers to compensate for the unobserved input region during encoding [LRS18]. The baseline MRI model operates with 2D inputs and outputs due to the non-isotropic resolution.

We want the baseline MRI model to focus on learning the negative mp-MRI appearance of the PCa-prone areas, rather than irrelevant areas in the images (e.g., muscle, fat, bone). We collected 1055 annotated 2D ROIs for PCa (without the corresponding mp-MRI) from a separate study cohort without any case overlapping [JRM18]. For each 2D ROI, the in-plane location relative to the center of the prostate is maintained, and the through-plane position is ignored. Each ROI is converted into a binary mask for the baseline MRI model as an ROI candidate to specify a region M to synthesize. As all the ROI candidates are considered in one plane, the collection of ROI candidates, $\mathcal{M}$, can account for the common locations and shapes of PCa. $\mathcal{M}$ is visualized as a prevalence map, P, s.t., $P = \sum_{M \in \mathcal{M}} M$, as shown in Fig. 4.2.1.

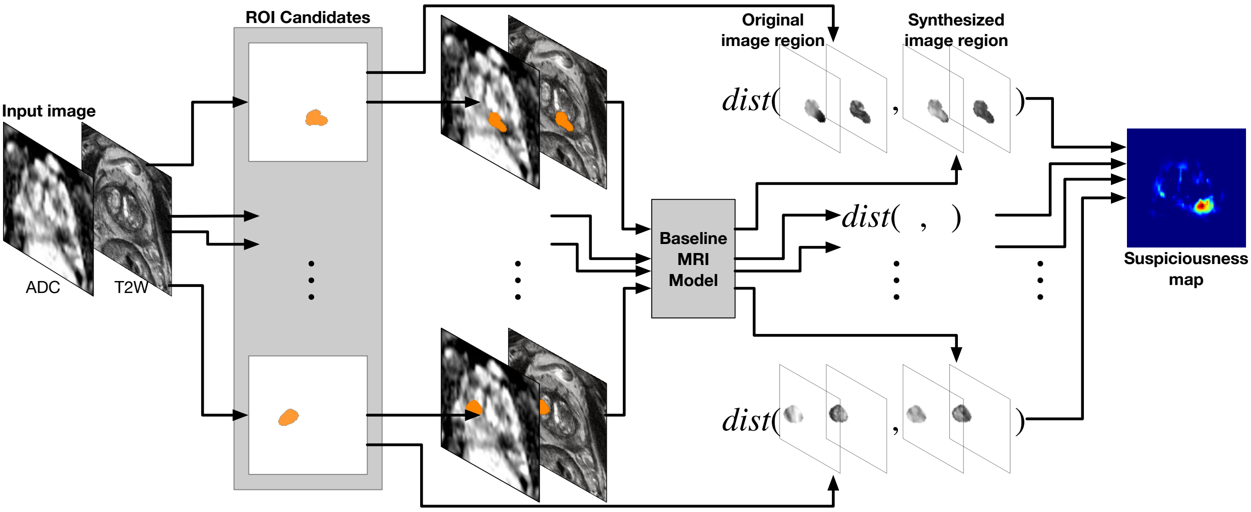


Figure 4.3: The inference of the PCa suspiciousness map using the trained baseline MRI model given an input testing image. The baseline MRI model synthesizes regions specified from the collection of ROI candidates. *dist* is the distance function for the original image region and the synthesized image region.

4.2.2.1 Training for the baseline MRI model

We train the baseline MRI model using the combination of L1 loss, perceptual loss, and style loss [GEB16]. The VGG-19 network pre-trained for image classification is used for the calculation of perceptual loss and style loss. We only take the feature map from the first convolutional layer for perceptual loss and style loss, since the network is trained for natural images and the higher-level features are not applicable to our context. The same weighting for loss terms is used as in [LRS18].

The baseline MRI model is trained for 4000 epochs using a mini-batch of eight 128×128 training images. The learning rate is set to 0.0002 in first 1000 epochs, and it is reduced to 0.00005 in the remaining 3000 epochs with the batch normalization for the encoder turned off as suggested in [LRS18]. Common image augmentations, including shifting, left-right flipping, and gray value variations [RFB15], are applied. We also randomly combine multiple ROI candidates together to accelerate training. The training took two days using one NVIDIA Titan Xp GPU.

4.2.3 Inference via the baseline MRI model

The trained baseline MRI model is utilized to infer the pixel-wise PCa suspiciousness map given a testing image. Since the baseline MRI model synthesizes a specified region negative for PCa, the synthesized image region is expected to be similar to the region in the original image if it is MRI-negative for PCa. Conversely, if the specified region in the testing image is MRI-positive for PCa, the synthesized image region will be different from the original image region. In other words, the region is considered to be suspicious when the difference between the synthesized image region and the original region is nontrivial.

In each time, we specify a region to synthesize from the collection of ROI candidates, $M \in \mathcal{M}$, and the synthesized image region from the baseline MRI model is $Mf\left(\left(1-M\right)I^t;\theta\right)$ where $I^t = (I^t_{\mathrm{T2W}}, I^t_{\mathrm{ADC}})$ is the testing image. By synthesizing different image regions with different ROI candidates, we can obtain the suspiciousness map by

$$Susp\left(I^t\right) = \frac{1}{P}\sum_{M\in\mathcal{M}} dist\left(MI^t, Mf\left(\left(1-M\right)I^t;\theta\right)\right), \tag{4.1}$$

where $dist\left(I^{ori}, I^{syn}\right)$ is the distance function measuring the pixel-wise difference between the original image region and the synthesized image region, and P is the prevalence map to normalize the suspiciousness map.

Two distance functions are tested individually: *T2W SSIM* and *ADC Increment*. Firstly, since T2W mainly contains structural information, we evaluate the variation of T2W by *T2W SSIM*, s.t., $dist\left(I^{ori}, I^{syn}\right) = 1 - SSIM\left(I^{ori}_{\mathrm{T2W}}, I^{syn}_{\mathrm{T2W}}\right)$, where $SSIM$ is the structural similarity. Secondly, ADC is quantitative imaging, and PCa lesion usually has lower ADC intensity than normal tissues [PJY13]. The suspicion for PCa is high if the ADC intensity in the original region is lower than in the synthesized MRI-negative region. Hence, we measure the ADC intensity increment of the synthesized region compared with the original region. *ADC Increment* is defined as $dist\left(I^{ori}, I^{syn}\right) = \max\left(I^{syn}_{\mathrm{ADC}} - I^{ori}_{\mathrm{ADC}}, 0\right)$,

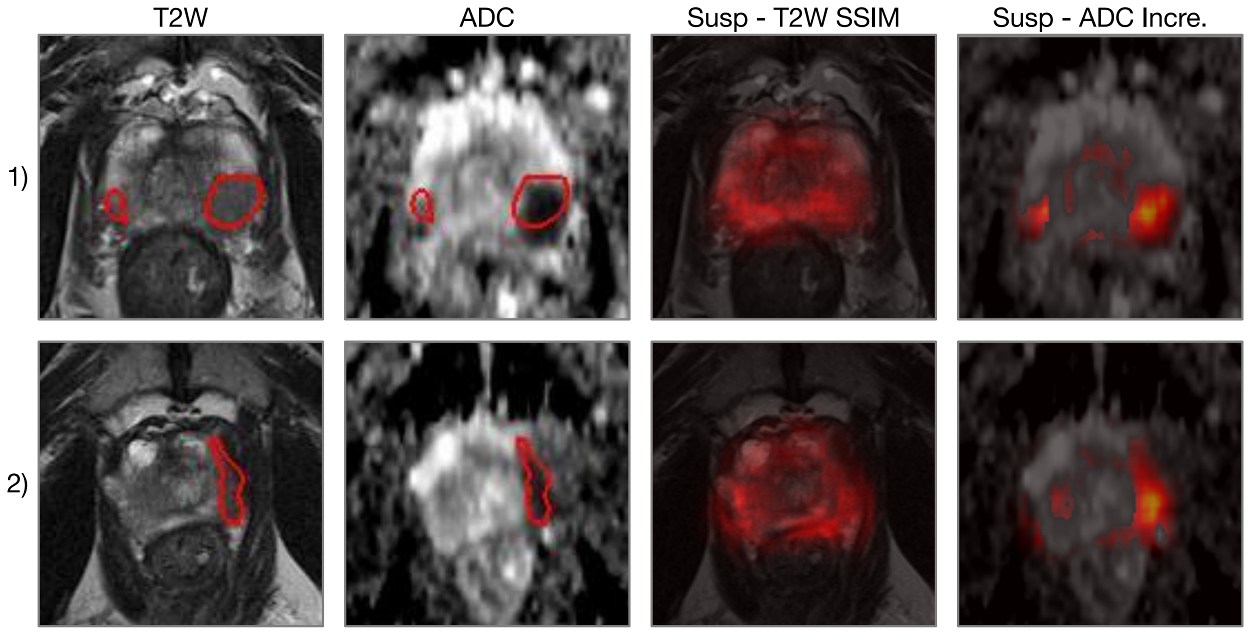


Figure 4.4: The PCa suspiciousness maps with different distance functions for testing images. The red contours on T2W and ADC are the groundtruth ROIs.

4.3 Experiments

4.3.1 Evaluation dataset

A separate dataset was collected with PCa lesion annotations for the evaluation. The evaluation dataset consisted of pre-operative 3T mp-MRI exams prior to prostatectomy from 2013 to 2015, and patients with prior treatment or scanned with endorectal coil were excluded. For the 116 eligible cases, the FOV and slice were determined in the same way as in 4.2.1. Clinical research fellows used whole-gland surgical specimens and pathology reports to retrospectively identify confirmed clinically significant PCa lesions (Gleason Score$\geq$3+4) in mp-MRI and annotated their groundtruth ROIs on T2W. Apart from these MRI-positive cases, the 116 testing cases from the negative prostate MRI dataset were used as the MRI-negative testing cases.

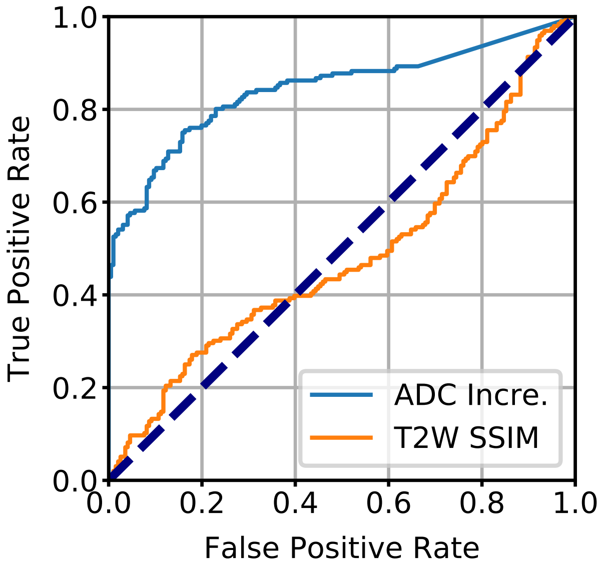


Figure 4.5: ROC analysis for the classification between PCa lesions and non-PCa regions.

4.3.2 Evaluation metrics

The suspiciousness map by the baseline MRI model is used to distinguish between PCa lesions and non-PCa regions [SZY18, WLC18]. The PCa lesions are given by the groundtruth ROIs, and non-PCa regions are defined as the same groundtruth ROIs in the MRI-negative testing cases. The average value over the region on the suspiciousness map is calculated as the predictive value for each ROI. The performance is evaluated by receiver operating characteristic (ROC) analysis.

We also evaluate the lesion localization performance using free-response receiver operating characteristic (FROC) analysis [LDB14, WLC18]. The PCa localization points are determined by the local maximums of the suspiciousness map [WLC18]. A localization point is considered as a true positive if it is within 5mm of a groundtruth lesion ROI, or it is otherwise a false negative [PNK17]. FROC measures the lesion detection sensitivity versus the average number of false positives for each patient.

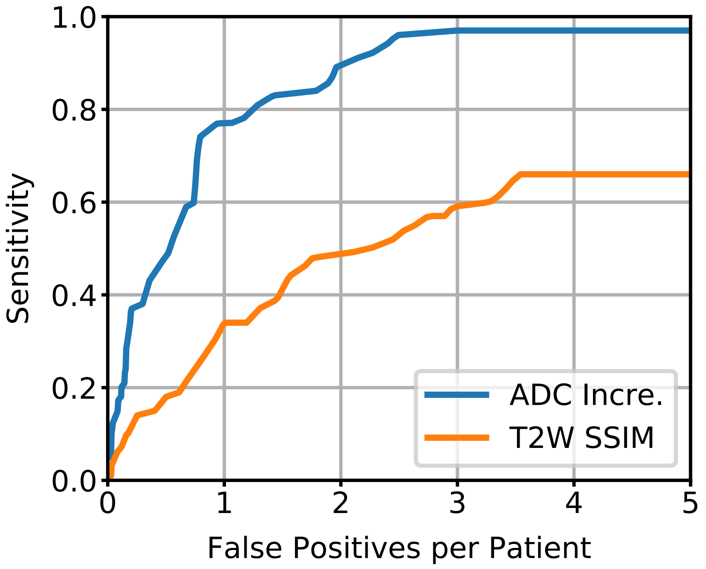


Figure 4.6: FROC analysis for lesion localization performance.

4.3.3 Results

Fig. 4.4 shows representative examples of the inferred suspiciousness map. The ROC analysis for the classification between PCa lesions and non-PCa regions is shown in Fig. 4.3.2. *ADC Increment* (*ADC Incre.*) achieved the area under the curve (AUC) of 0.84, while the suspiciousness map using *T2W SSIM* exhibited limited predictability for PCa. Compared with ADC, T2W has a more diverse appearance for the normal tissues, causing the inconsistent inference for the suspiciousness map.

The FROC analysis for lesion localization is shown in Fig. 4.3.2. *ADC Increment* and *T2W SSIM* had 77.0% and 33.8% detection sensitivity for clinically significant PCa lesions with 1 false positive per patient, respectively, and 89.5% and 48.8% detection sensitivity at 2 false positives per patient. *ADC Increment* received 95% sensitivity at 2.44 false positives per patient, and *T2W SSIM* reached its maximum sensitivity of 66.0% at 3.54 false positives per patient.

4.4 Discussion

The PCa detection systems from previous studies reported lesion detection sensitivity from 38.8% to 89.8% at 1 false positive per patient in FROC analysis [LDB14, TLW17, WLC18]. Despite the difference in the dataset and lesion definition, our proposed unsupervised learning method, without using lesion annotations in training, shows similar performance to the fully-supervised methods. Compared with the fully-supervised methods trained with lesion annotations, the proposed method requires only MRI-negative scans in training, which is more approachable to institutions without a large annotated prostate MRI collection and suitable for multi-site, multi-vendor collaborations.

In conclusion, we proposed the baseline MRI model for the unsupervised inference of prostate cancer in multi-parametric MRI without using any PCa annotations. The baseline MRI model was trained using 1095 negative mp-MRI scans. In the evaluation using a separate dataset consisting of 116 mp-MRI scans with histologically confirmed lesion annotations, the proposed method achieved 0.84 AUC in ROC analysis and 77.0% detection sensitivity at 1 false positive per patient in FROC analysis.

CHAPTER 5

Discussion & Future Work

5.1 Potential usage

A deep learning-based detection system can act as a preliminary reader or quality checker complementary to radiologists under the current workflow of mp-MRI interpretation. As a preliminary reader, the system will propose a number of lesion detection points or regions based on a certain sensitivity threshold set by the reader prior to the interpretation. During the reading, readers will confirm or reject each of the detections. Since the systems in Chapter 2 and Chapter 3 received high lesion detection sensitivity with a relax threshold setting, readers can only check the proposed detection candidates instead of reading the entire volumetric imaging, which potentially will boost the interpretation efficiency. Nevertheless, the system needs to go through strict validations under the clinical settings as well as the approval from oversight committees and regulators for this usage into the practice. It is hard to predict how long it would take to realize this usage, but this is unlikely to happen in a short period of time.

On the other hand, using proposed detection systems as quality checkers could be potentially easier to translate into clinical practice. In this scenario, the readers will interpret prostate mp-MRI using the existing workflow. The system meanwhile will find out detection points with high confidence for the higher specificity. Right after the reader's interpretation, the detection system will check if the system's detection with high confidence is contained in the reader's findings. In case if a detection point with high confidence is not in the findings, the system will notify the reader to re-check the detected region by the system to make sure it is not a mis-detection.

I want to note that the usage of detection systems is highly dependent on the clinical history and clinical objective, which is often the case for radiologists' interpretation. By setting different sensitivity threshold values, the user will receive detection proposals customized by clinical considerations.

In addition, the detection or diagnosis systems can potentially work together with radiologists. I.e., the systems' outputs can be utilized as a quantitative measurement incorporated into the existing PI-RADS system if the systems can thoroughly demonstrate their effectiveness in clinical evaluations. In particular to this usage, as mentioned in [CLP19], it is critical to investigate the development of user interface and user interaction. The systems need to be embedded into the commercial image viewers. Although the computation load is not very heavy for the predictions or inferences, parallel computing or cloud computing is highly desirable due to the increasing total number of scans. Also, the processing time for a system needs to be short to ensure the minimum delay for users. Real-time detection will be important to the potential usage of automated targeting for in-bore biopsy.

5.2 Future works

This thesis has extensively discussed the deep learning-based methods for PCa detection, segmentation, and classification from mp-MRI. During my study, I have identified some future directions to further improve the proposed systems for PCa diagnosis.

Firstly, although the proposed systems were evaluated using the existing metrics, clinical validation of the detection and classification is needed. For the clinical validation, a testing set of cases should be prospectively collected with proper IRB approval. Retrospectively collected cases, such as those in this thesis, shall be used as the training and validation sets. The model needs to be fixed before evaluated using the testing set, and the evaluation on the testing set can only be performed once to avoid model's implicit overfitting to the testing data [CLP19]. In addition to the FROC and ROC analyses in 3.2.5.3 and in 3.2.5.4, the systems can also be scored qualitatively by experienced radiologists with a well-defined condition and scoring rubric.

Secondly, besides the detection and classification of PCa lesions, the identification of PCa-negative cases is also important due to a large number of negative mp-MRI scan caused by the low specificity of PSA test [GP03]. The total radiologists' workload could be reduced if some negative cases can be identified by a system with high confidence. Identifying PCa-negative cases is a task related to PCa detection, i.e., an mp-MRI scan can be identified as PCa-negative if a detection system does not detect any lesion in the scan. As the evaluation cohort for the detection systems only consists of patients underwent prostatectomy in this thesis, further investigations are needed with a study cohort similar to the actual screening population.

Thirdly, multi-modal information, such as electronic medical record and genomics, can be incorporated into the system in addition to mp-MRI. Lab results (e.g. PSA level) and previous prostate mp-MRI scans can be retrieved from the electronic medical record, to stratify patient's risk level before the interpretation of the current image. This will be especially helpful for patients in active surveillance [KVS14]. Also, genomics information was shown to be related to PCa [RVW15]. As there exist associations between genomics and mp-MRI features [SPT16], adding genomics information into the system can potentially make the system more robust.

REFERENCES

[ABC16] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. “Tensorflow: a system for large-scale machine learning.” In *USENIX Symposium on Operating Systems Design and Implementation*, volume 16, pp. 265–283, 2016.

[ACE16] Marios Anthimopoulos, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. “Lung pattern classification for interstitial lung diseases using a deep convolutional neural network.” *IEEE transactions on medical imaging*, **35**(5):1207–1216, 2016.

[AGH17] Zeynettin Akkus, Alfiia Galimzianova, Assaf Hoogi, Daniel L Rubin, and Bradley J Erickson. “Deep learning for brain MRI segmentation: state of the art and future directions.” *Journal of digital imaging*, **30**(4):449–459, 2017.

[BGG18] Samuel Borofsky, Arvin K George, Sonia Gaur, Marcelino Bernardo, Matthew D Greer, Francesca V Mertan, Myles Taffel, Vanesa Moreno, Maria J Merino, Bradford J Wood, et al. “What are we missing? False-negative cancers at multiparametric MR imaging of the prostate.” *Radiology*, **286**(1):186–195, 2018.

[CKH16] Andrew Cameron, Farzad Khalvati, Masoom A Haider, and Alexander Wong. “MAPS: a quantitative radiomics approach for prostate cancer detection.” *IEEE Transactions on Biomedical Engineering*, **63**(6):1145–1156, 2016.

[CLP19] Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng. “How to develop machine learning models for healthcare.” *Nature materials*, **18**(5):410, 2019.

[CPK18] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **40**(4):834–848, 2018.

[CWP08] Jianlin Cheng, Zheng Wang, and Gianluca Pollastri. “A neural network approach to ordinal regression.” In *IEEE International Joint Conference on Neural Networks*, pp. 1279–1284. IEEE, 2008.

[DAB12] Marc A DallEra, Peter C Albertsen, Christopher Bangma, Peter R Carroll, H Ballentine Carter, Matthew R Cooperberg, Stephen J Freedland, Laurence H Klotz, Christopher Parker, and Mark S Soloway. “Active surveillance for prostate cancer: a systematic review of the literature.” *European Urology*, **62**(6):976–983, 2012.

[DDS09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A large-scale hierarchical image database.” In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

[DGS18] Adrian V. Dalca, John Guttag, and Mert R. Sabuncu. “Anatomical Priors in Convolutional Networks for Unsupervised Biomedical Segmentation.” *Proceedings of IEEE Conference on Computer Vision Pattern Recognition*, pp. 9290–9299, 2018.

[EEA16] Jonathan I Epstein, Lars Egevad, Mahul B Amin, Brett Delahunt, John R Srigley, and Peter A Humphrey. “The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma.” *The American journal of surgical pathology*, **40**(2):244–252, 2016.

[EFT12] Jonathan I Epstein, Zhaoyong Feng, Bruce J Trock, and Phillip M Pierorazio. “Upgrading and downgrading of prostate cancer from biopsy to radical prostatectomy: incidence and predictive factors using the modified Gleason grading system and factoring in tertiary grades.” *European Urology*, **61**(5):1019–1024, 2012.

[EZS16] Jonathan I Epstein, Michael J Zelefsky, Daniel D Sjoberg, Joel B Nelson, Lars Egevad, Cristina Magi-Galluzzi, Andrew J Vickers, Anil V Parwani, Victor E Reuter, Samson W Fine, et al. “A contemporary prostate cancer grading system: a validated alternative to the Gleason score.” *European Urology*, **69**(3):428–435, 2016.

[FHW94] Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. “Statistical parametric maps in functional imaging: a general linear approach.” *Human brain mapping*, **2**(4):189–210, 1994.

[FLG17] Lucas Fidon, Wenqi Li, Luis C Garcia-Peraza-Herrera, Jinendra Ekanayake, Neil Kitchen, Sebastien Ourselin, and Tom Vercauteren. “Scalable multimodal convolutional networks for brain tumour segmentation.” In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 285–293. Springer, 2017.

[FVW15] Duc Fehr, Harini Veeraraghavan, Andreas Wibmer, Tatsuo Gondo, Kazuhiro Matsumoto, Herbert Alberto Vargas, Evis Sala, Hedvig Hricak, and Joseph O Deasy. “Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images.” *Proceedings of the National Academy of Sciences*, **112**(46):E6265–E6273, 2015.

[GBM15] Marietta Garmer, Martin Busch, Serban Mateiescu, David E Fahlbusch, Birgit Wagener, and Dietrich HW Grönemeyer. “Accuracy of MRI-targeted in-bore prostate biopsy according to the Gleason score with postprostatectomy histopathologic controla targeted biopsy-only strategy with limited number of cores.” *Academic radiology*, **22**(11):1409–1418, 2015.

[GEB16] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. “Image Style Transfer Using Convolutional Neural Networks.” In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition*, 2016.

[Gir15] Ross Girshick. “Fast r-cnn.” In *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[GP03] Matthew B Gretzer and Alan W Partin. “PSA markers in prostate cancer detection.” *The Urologic clinics of North America*, **30**(4):677–686, 2003.

[GPS16] Pedro Antonio Gutierrez, Maria Perez-Ortiz, Javier Sanchez-Monedero, Francisco Fernandez-Navarro, and Cesar Hervas-Martinez. “Ordinal regression methods: survey and experimental study.” *IEEE Transactions on Knowledge and Data Engineering*, **28**(1):127–146, 2016.

[HDW15] Esther H.J. Hamoen, Maarten De Rooij, J. Alfred Witjes, Jelle O. Barentsz, and Maroeska M. Rovers. “Use of PI-RADS for prostate cancer detection with mp-MRI.” *European Urology*, **67**(6):1112–1121, 2015.

[HRG16] Shin Hoo-Chang, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. “Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning.” *IEEE Transactions on Medical Imaging*, **35**(5):1285, 2016.

[HSH11] Thomas Hambrock, Diederik M Somford, Henkjan J Huisman, Inge M van Oort, J Alfred Witjes, Christina A Hulsbergen-van de Kaa, Thomas Scheenen, and Jelle O Barentsz. “Relationship between apparent diffusion coefficients at 3.0-T MR imaging and Gleason grade in peripheral zone prostate cancer.” *Radiology*, **259**(2):453–461, 2011.

[HZR16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition.” In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition*, pp. 770–778, 2016.

[IS15] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” *arXiv preprint arXiv:1502.03167*, 2015.

[JRM18] David C. Johnson, Steven S. Raman, Sohrab A. Mirak, Lorna Kwan, Amirhossein M. Bajgiran, William Hsu, Cleo K. Maehara, Preeti Ahuja, Izak Faiena, Aydin Pooli, Amirali Salmasi, Anthony Sisk, Ely R. Felker, David S.K. Lu, and Robert E. Reiter. “Detection of Individual Prostate Cancer Foci via Multiparametric Magnetic Resonance Imaging.” *Eur. Urol.*, 2018.

[KCJ18] Piotr Kozlowski, Silvia D Chang, Edward C Jones, and S Larry Goldenberg. “Assessment of the need for DCE MRI in the detection of dominant lesions in the whole gland: Correlation between histology and MRI of prostate cancer.” *NMR in Biomedicine*, **31**(3):e3882, 2018.

[KK11] Philipp Krähenbühl and Vladlen Koltun. “Efficient inference in fully connected crfs with gaussian edge potentials.” In *Proceedings of Advances in Neural Information Processing Systems*, pp. 109–117, 2011.

[KNT17] Atilla P Kiraly, Clement Abi Nader, Ahmet Tuysuzoglu, Robert Grimm, Berthold Kiefer, Noha El-Zehiry, and Ali Kamen. "Deep Convolutional Encoder-Decoders for Prostate Cancer Detection and Classification." In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 489–497. Springer, 2017.

[KRB18] Veeru Kasivisvanathan, Antti S Rannikko, Marcelo Borghi, Valeria Panebianco, Lance A Mynderse, Markku H Vaarala, Alberto Briganti, Lars Budäus, Giles Hellawell, Richard G Hindley, et al. "MRI-targeted or standard biopsy for prostate-cancer diagnosis." *The New England Journal of Medicine*, **378**(19):1767–1777, 2018.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." In *Proceedings of Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[KVS14] Laurence Klotz, Danny Vesprini, Perakaa Sethukavalan, Vibhuti Jethava, Liying Zhang, Suneil Jain, Toshihiro Yamamoto, Alexandre Mamedov, and Andrew Loblaw. "Long-term follow-up of a large active surveillance cohort of patients with prostate cancer." *Journal of Clinical Oncology*, **33**(3):272–277, 2014.

[KXW15] Jin Tae Kwak, Sheng Xu, Bradford J Wood, Baris Turkbey, Peter L Choyke, Peter A Pinto, Shijun Wang, and Ronald M Summers. "Automated prostate cancer detection using T2-weighted and high-b-value diffusion-weighted magnetic resonance imaging." *Medical Physics*, **42**(5):2368–2378, 2015.

[LBD89] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. "Backpropagation applied to handwritten zip code recognition." *Neural computation*, **1**(4):541–551, 1989.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature*, **521**(7553):436, 2015.

[LDB14] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. "Computer-aided detection of prostate cancer in MRI." *IEEE Transactions on Medical Imaging*, **33**(5):1083–1092, 2014.

[LGG17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. "Focal Loss for Dense Object Detection." In *Proceedings of IEEE International Conference on Computer Vision*, pp. 2999–3007. IEEE, 2017.

[LLH04] Jacques Lapointe, Chunde Li, John P Higgins, Matt Van De Rijn, Eric Bair, Kelli Montgomery, Michelle Ferrari, Lars Egevad, Walter Rayford, Ulf Bergerheim, et al. "Gene expression profiling identifies clinically relevant subtypes of prostate cancer." *Proceedings of the National Academy of Sciences*, **101**(3):811–816, 2004.

[LMF15] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C Vilanova, Paul M Walker, and Fabrice Meriaudeau. "Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review." *Computers in biology and medicine*, **60**:8–31, 2015.

[LRS18] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting Chun Wang, Andrew Tao, and Bryan Catanzaro. "Image Inpainting for Irregular Holes Using Partial Convolutions." In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[LSB14] Jesse D Le, Samuel Stephenson, Michelle Brugger, David Y Lu, Patricia Lieu, Geoffrey A Sonn, Shyam Natarajan, Frederick J Dorey, Jiaoti Huang, Daniel JA Margolis, et al. "Magnetic resonance imaging-ultrasound fusion biopsy for prediction of final prostate pathology." *The Journal of urology*, **192**(5):1367–1373, 2014.

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In *Proceedings of IEEE Conference on Computer Vision Pattern Recognition*, pp. 3431–3440, 2015.

[LTS15] Jesse D Le, Nelly Tan, Eugene Shkolyar, David Y Lu, Lorna Kwan, Leonard S Marks, Jiaoti Huang, Daniel JA Margolis, Steven S Raman, and Robert E Reiter. "Multifocality and prostate cancer detection by multiparametric magnetic resonance imaging: correlation with whole-mount histopathology." *European Urology*, **67**(3):569–576, 2015.

[LWT13] Peter Liu, Shijun Wang, Baris Turkbey, Kinzya Grant, Peter Pinto, Peter Choyke, Bradford J Wood, and Ronald M Summers. "A prostate cancer computer-aided diagnosis system using multimodal magnetic resonance imaging and targeted biopsy labels." In *Medical Imaging 2013: Computer-Aided Diagnosis*, volume 8670, p. 86701G. International Society for Optics and Photonics, 2013.

[MFI07] Huadong Miao, Hiroshi Fukatsu, and Takeo Ishigaki. "Prostate cancer detection with 3-T MRI: comparison of diffusion-weighted and T2-weighted imaging." *European Journal of Radiology*, **61**(2):297–302, 2007.

[MJB15] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. "The multimodal brain tumor image segmentation benchmark (BRATS)." *IEEE transactions on medical imaging*, **34**(10):1993, 2015.

[OLL10] Sedat Ozer, Deanna L Langer, Xin Liu, Masoom A Haider, Theodorus H van der Kwast, Andrew J Evans, Yongyi Yang, Miles N Wernick, and Imam S Yetik. "Supervised and unsupervised methods for prostate cancer segmentation with multispectral MRI." *Medical Physics*, **37**(4):1873–1883, 2010.

[PEM11] Guillaume Ploussard, Jonathan I Epstein, Rodolfo Montironi, Peter R Carroll, Manfred Wirth, Marc-Oliver Grimm, Anders S Bjartell, Francesco Montorsi, Stephen J Freedland, Andreas Erbersdobler, et al. "The contemporary concept of significant versus insignificant prostate cancer." *European Urology*, **60**(2):291–303, 2011.

[PJY13] Yahui Peng, Yulei Jiang, Cheng Yang, Jeremy Bancroft Brown, Tatjana Antic, Ila Sethi, Christine Schmid-Tannwald, Maryellen L Giger, Scott E Eggener, and Aytekin Oto. "Quantitative analysis of multiparametric prostate MR images: differentiation between prostate cancer and normal tissue and correlation with Gleason score—a computer-aided diagnosis development study." *Radiology*, **267**(3):787–796, 2013.

[PNK17] Alan Priester, Shyam Natarajan, Pooria Khoshnoodi, Daniel J Margolis, Steven S Raman, Robert E Reiter, Jiaoti Huang, Warren Grundfest, and Leonard S Marks. "Magnetic resonance imaging underestimation of prostate cancer geometry: use of patient specific molds to correlate images with whole mount pathology." *The Journal of Urology*, **197**(2):320–326, 2017.

[PPA16] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. "Brain tumor segmentation using convolutional neural networks in MRI images." *IEEE Transactions on Medical Imaging*, **35**(5):1240–1251, 2016.

[PPC17] Nestor Andres Parra, Alan Pollack, Felix M Chinea, Matthew C Abramowitz, Brian Marples, Felipe Munera, Rosa Castillo, Oleksandr N Kryvenko, Sanoj Punnen, and Radka Stoyanova. "Automatic detection and quantitative DCE-MRI scoring of prostate cancer aggressiveness." *Frontiers in oncology*, **7**:259, 2017.

[RAE18] Islam Reda, Babajide O Ayinde, Mohammed Elmogy, Ahmed Shalaby, Moumen El-Melegy, Mohamed Abou El-Ghar, Ahmed Abou El-fetouh, Mohammed Ghazal, and Ayman El-Baz. "A new CNN-based system for early diagnosis of prostate cancer." In *Proceedings of IEEE International Symposium on Biomedical Imaging*, pp. 207–210. IEEE, 2018.

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

[RIZ17] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." *arXiv preprint arXiv:1711.05225*, 2017.

[RSW16] Jan P Radtke, Constantin Schwab, Maya B Wolf, Martin T Freitag, Celine D Alt, Claudia Kesch, Ionel V Popeneciu, Clemens Huettenbrink, Claudia Gasch, Tilman Klein, et al. "Multiparametric magnetic resonance imaging (MRI) and MRI–transrectal ultrasound fusion biopsy for index tumor detection: correlation with radical prostatectomy specimen." *European Urology*, **70**(5):846–853, 2016.

[RVW15] Dan Robinson, Eliezer M Van Allen, Yi-Mi Wu, Nikolaus Schultz, Robert J Lonigro, Juan-Miguel Mosquera, Bruce Montgomery, Mary-Ellen Taplin, Colin C Pritchard, Gerhardt Attard, et al. "Integrative clinical genomics of advanced prostate cancer." *Cell*, **161**(5):1215–1228, 2015.

[RWB12] Oliver Ruprecht, Philipp Weisser, Boris Bodelle, Hanns Ackermann, and Thomas J Vogl. "MRI of the prostate: interobserver agreement compared with histopathologic outcome after radical prostatectomy." *European journal of radiology*, **81**(3):456–460, 2012.

[Sch15] Jürgen Schmidhuber. "Deep learning in neural networks: An overview." *Neural networks*, **61**:85–117, 2015.

[SHK14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting." *The Journal of Machine Learning Research*, **15**(1):1929–1958, 2014.

[SMJ19] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. "Cancer statistics, 2019." *CA: a cancer journal for clinicians*, 2019.

[SNN05] Chiho Sato, Shinji Naganawa, Tatsuya Nakamura, Hisashi Kumada, Shunichi Miura, Osamu Takizawa, and Takeo Ishigaki. "Differentiation of noncancerous tissue and cancer lesions by apparent diffusion coefficient values in transition and peripheral zones of the prostate." *Journal of Magnetic Resonance Imaging*, **21**(3):258–262, 2005.

[SPG15] Ivo G Schoots, Neophytos Petrides, Francesco Giganti, Leonard P Bokhorst, Antti Rannikko, Laurence Klotz, Arnauld Villers, Jonas Hugosson, and Caroline M Moore. "Magnetic resonance imaging in active surveillance of prostate cancer: a systematic review." *European urology*, **67**(4):627–636, 2015.

[SPS09] Jennifer R Stark, Sven Perner, Meir J Stampfer, Jennifer A Sinnott, Stephen Finn, Anna S Eisenstein, Jing Ma, Michelangelo Fiorentino, Tobias Kurth, Massimo Loda, et al. "Gleason score and lethal prostate cancer: does 3+ 4= 4+ 3?" *Journal of Clinical Oncology*, **27**(21):3459, 2009.

[SPT16] Radka Stoyanova, Alan Pollack, Mandeep Takhar, Charles Lynne, Nestor Parra, Lucia LC Lam, Mohammed Alshalalfa, Christine Buerki, Rosa Castillo, Merce Jorda, et al. "Association of multiparametric MRI quantitative imaging features with prostate cancer gene expression in MRI-targeted prostate biopsies." *Oncotarget*, **7**(33):53362, 2016.

[SQA13] L Schimmöller, M Quentin, C Arsov, RS Lanzman, A Hiester, R Rabenalt, G Antoch, P Albers, and D Blondin. "Inter-reader agreement of the ESUR score for prostate MRI using in-bore MRI-guided biopsies as the reference standard." *European radiology*, **23**(11):3185–3190, 2013.

[SS13] Heung-Il Suk and Dinggang Shen. "Deep learning-based feature representation for AD/MCI classification." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 583–590. Springer, 2013.

[SZY18] Yang Song, Yu-Dong Zhang, Xu Yan, Hui Liu, Minxiong Zhou, Bingwen Hu, and Guang Yang. "Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric MRI." *Journal of Magnetic Resonance Imaging*, 2018.

[TB16] Gijs van Tulder and Marleen de Bruijne. "Combining generative and discriminative representation learning for lung CT analysis with convolutional restricted boltzmann machines." *IEEE transactions on medical imaging*, **35**(5):1262–1272, 2016.

[TKM13] Pallavi Tiwari, John Kurhanewicz, and Anant Madabhushi. "Multi-kernel graph embedding for detection, Gleason grading of prostate cancer via MRI/MRS." *Medical image analysis*, **17**(2):219–235, 2013.

[TLW17] Yohannes Tsehay, Nathan Lay, Xiaosong Wang, Jin Tae Kwak, Baris Turkbey, Peter Choyke, Peter Pinto, Brad Wood, and Ronald M Summers. "Biopsy-guided learning with deep convolutional neural networks for Prostate Cancer detection on multiparametric MRI." In *Proceedings of IEEE International Symposium on Biomedical Imaging*, pp. 642–645. IEEE, 2017.

[TMA12] Baris Turkbey, Haresh Mani, Omer Aras, Ardeshir R Rastinehad, Vijay Shah, Marcelino Bernardo, Thomas Pohida, Dagane Daar, Compton Benjamin, Yolanda L McKinney, et al. "Correlation of magnetic resonance imaging tumor volume with histopathology." *The Journal of urology*, **188**(4):1157–1163, 2012.

[TML15] Nelly Tan, Daniel J Margolis, David Y Lu, Kevin G King, Jiaoti Huang, Robert E Reiter, and Steven S Raman. "Characteristics of detected and missed prostate cancer foci on 3-T multiparametric MRI using an endorectal coil correlated with whole-mount thin-section histopathology." *American Journal of Roentgenology*, **205**(1):W87–W92, 2015.

[VAE14] Massimo Valerio, Hashim U Ahmed, Mark Emberton, Nathan Lawrentschuk, Massimo Lazzeri, Rodolfo Montironi, Paul L Nguyen, John Trachtenberg, and Thomas J Polascik. "The role of focal therapy in the management of localised prostate cancer: a systematic review." *European urology*, **66**(4):732–751, 2014.

[VAF11] Hebert Alberto Vargas, Oguz Akin, Tobias Franiel, Yousef Mazaheri, Junting Zheng, Chaya Moskowitz, Kazuma Udo, James Eastham, and Hedvig Hricak. "Diffusion-weighted endorectal MR imaging at 3 T for prostate cancer: tumor detection and assessment of aggressiveness." *Radiology*, **259**(3):775–784, 2011.

[VBK12] PC Vos, JO Barentsz, N Karssemeijer, and HJ Huisman. "Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis." *Physics in Medicine & Biology*, **57**(6):1527, 2012.

[VHG16] HA Vargas, AM Hötker, DA Goldman, CS Moskowitz, T Gondo, Kazuhiro Matsumoto, B Ehdaie, S Woo, SW Fine, VE Reuter, et al. "Updated prostate imaging reporting and data system (PIRADS v2) recommendations for the detection of clinically significant prostate cancer using multiparametric MRI: critical evaluation using whole-mount pathology as standard of reference." *European radiology*, **26**(6):1606–1612, 2016.

[WBC16] Jeffrey C Weinreb, Jelle O Barentsz, Peter L Choyke, Francois Cornud, Masoom A Haider, Katarzyna J Macura, Daniel Margolis, Mitchell D Schnall, Faina Shtern, Clare M Tempany, et al. "PI-RADS prostate imaging–reporting and data system: 2015, version 2." *European Urology*, **69**(1):16–40, 2016.

[WBT14] Shijun Wang, Karen Burtt, Baris Turkbey, Peter Choyke, and Ronald M Summers. "Computer aided-diagnosis of prostate cancer on multiparametric MRI: a technical review of current research." *BioMed Research International*, **2014**, 2014.

[WFG07] H Gilbert Welch, Elliott S Fisher, Daniel J Gottlieb, and Michael J Barry. "Detection of prostate cancer via biopsy in the medicare–SEER population during the PSA era." *Journal of the National Cancer Institute*, **99**(18):1395–1400, 2007.

[WLC18] Zhiwei Wang, Chaoyue Liu, Danpeng Cheng, Liang Wang, Xin Yang, and Kwang-Ting Cheng. "Automated Detection of Clinically Significant Prostate Cancer in mp-MRI Images Based on an End-to-End Deep Neural Network." *IEEE Transactions on Medical Imaging*, **37**(5):1127–1139, 2018.

[YLW17] Xin Yang, Chaoyue Liu, Zhiwei Wang, Jun Yang, Hung Le Min, Liang Wang, and Kwang-Ting Tim Cheng. "Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric MRI." *Medical image analysis*, **42**:212–227, 2017.

[YVM12] Joseph H Yacoub, Sadhna Verma, Jonathan S Moulton, Scott Eggener, and Aytekin Oto. "Imaging-guided prostate biopsy: conventional and emerging techniques." *Radiographics*, **32**(3):819–837, 2012.

[ZLD15] Wenlu Zhang, Rongjian Li, Houtao Deng, Li Wang, Weili Lin, Shuiwang Ji, and Dinggang Shen. "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation." *NeuroImage*, **108**:214–224, 2015.

[ZSZ18] Jun Zhang, Ashirbani Saha, Zhe Zhu, and Maciej A Mazurowski. "Hierarchical Convolutional Neural Networks for Segmentation of Breast Tumors in MRI with Application to Radiogenomics." *IEEE Transactions on Medical Imaging*, 2018.