Title

Joint Prostate Cancer Detection and Gleason Score Prediction in mp-MRI via FocalNet

Authors

Cao, Ruiming
Mohammadian Bajgiran, Amirhossein
Afshari Mirak, Sohrab
et al.

Publication Date

2019-04-18

DOI

10.1109/TMI.2019.2901928

Peer reviewed

Joint Prostate Cancer Detection and Gleason Score Prediction in mp-MRI via FocalNet

Ruiming Cao, *Member, IEEE*, Amirhossein Mohammadian Bajgiran, Sohrab Afshari Mirak, Sepideh Shakeri, Xinran Zhong, Dieter Enzmann, Steven Raman, and Kyunghyun Sung

***Abstract*—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.**

***Index Terms*—Prostate cancer, magnetic resonance imaging, computer-aided detection and diagnosis, convolutional neural network.**

I. 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) [1]. 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 [2]. 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 [3] and adding treatment options in active surveillance [4] and focal therapy [5]. 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) [6], 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 [7], 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 [8]–[22]. 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 [23] and segmentation [24]. Recent studies also showed the feasibility of training CNNs to detect cancer from mp-MRI. Zhang *et al.* [25] 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.* [21] built a patch-based CNN to classify between biopsy-proven PCa lesion and non-lesion regions of interest (ROIs). Kiraly *et al.* [19] 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 [6]. 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) [19]–[21], [26], [27]. 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.

R. Cao, A. B. Mohammadian, S. M. Afshari, S. Shakeri, X. Zhong, D. Enzmann, S. Raman, and K. Sung are with the Department of Radiology, University of California, Los Angeles, Los Angeles, CA 90095 USA, e-mail: ruimingc@ucla.edu.

R. Cao is also with the Department of Computer Science, University of California, Los Angeles, Los Angeles CA 90095 USA.

This work is supported by funds from the Integrated Diagnostics Program, Department of Radiological Sciences & Pathology, David Geffen School of Medicine at UCLA.

Manuscript received January 18, 2019; revised February 15, 2019; accepted February 19, 2019.

Inspired by the clinical interpretation of prostate mp-MRI [6], 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 [4], [28]. The correlation between mp-MRI and GS has been studied [10], 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 [13], [19]–[21]. 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$3 has a limited ability to detect all PCa lesions [29]–[31], clinically significant lesions can be missed and multi-focal lesions can be highly underestimated at mp-MRI [29], [32], resulting in an overestimation of the performance of the CAD systems. Also, there exists a significant risk of the inaccurate lesion annotations since GS between prostate biopsy and radical prostatectomy specimens is occasionally discordant [33]–[35]. Epstein *et al.* reported that more than one-third of the biopsy cases with GS$\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 [35]. 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 [36], 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-based PCa detection and diagnosis system using lesion findings confirmed with whole-mount histopathology in a large study cohort.

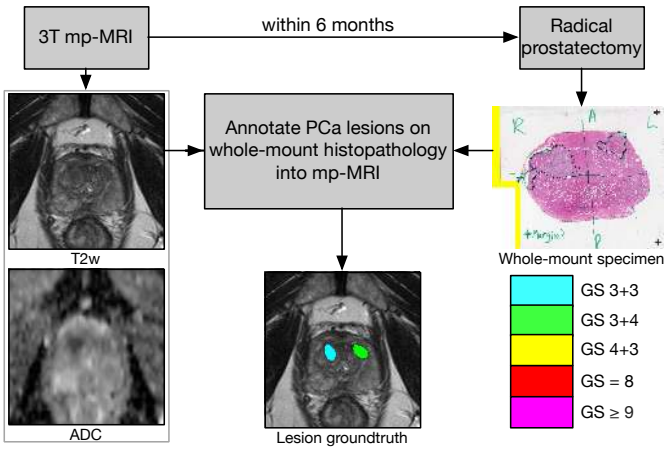


Fig. 1. Data preparation pipeline. 278 out of 400 prospectively missed (false negative) lesions were retrospectively identified and annotated in mp-MRI, referring to whole-mount histopathology. In the shown example, the lesion in the left anterior (GS 3+4, index lesion) was prospectively missed and retrospectively identified.

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

II. Materials and Methods

A. 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 [6], [31], [37]. 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 $\times$ 205, we acquired and reconstructed T2w TSE images with 0.55 mm $\times$ 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 $\times$ 26 cm and matrix of 94 $\times$ 160, DWI images were reconstructed with in-plane

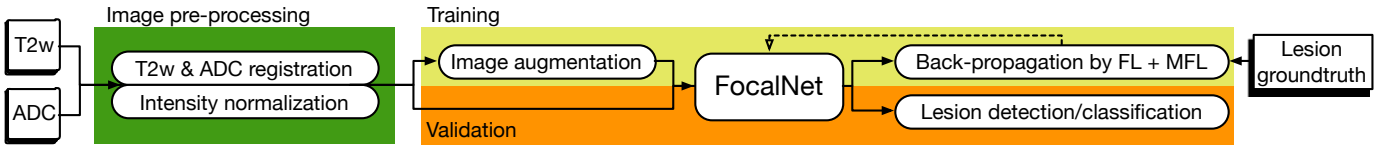


Fig. 2. The workflow of FocalNet for training and validation. Image registration and intensity normalization are performed with 3D image volumes. As FocalNet operates with 2D images, the corresponding T2w and ADC slices are grouped and fed into FocalNet for pixel-level predictions.

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.

B. Whole-mount histopathology matching & annotation

As in Fig. 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 [38].

TABLE I
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

C. 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. 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, 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.

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 [23]. The six different labels can be converted into 6-bit one-hot vectors as in TABLE I. 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 [36], 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 [39], [40]. As shown in TABLE I, 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.

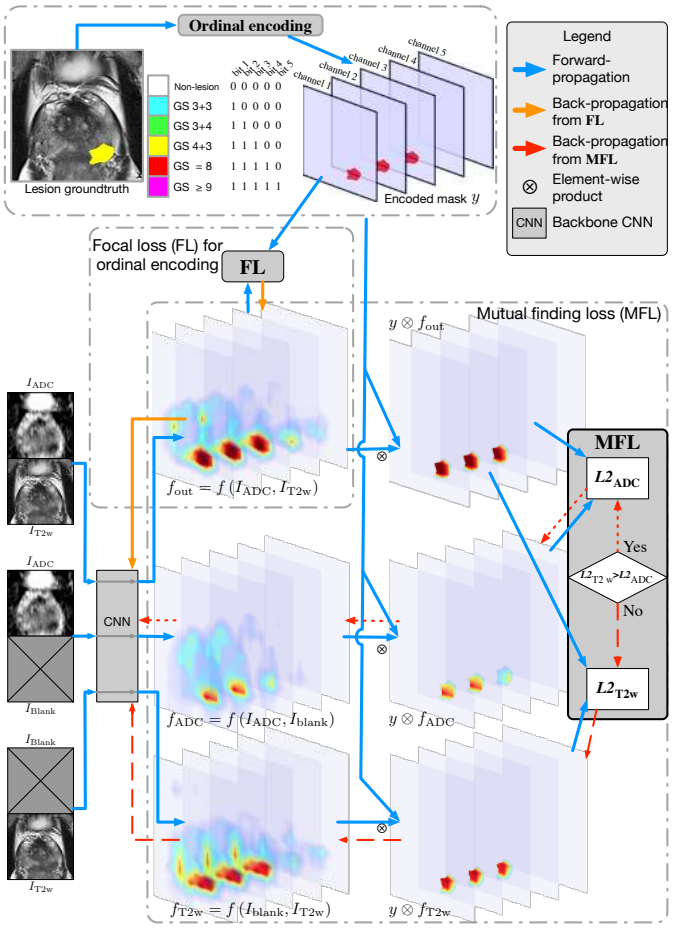


Fig. 3. FocalNet for joint PCa detection and Gleason score prediction. The lesion groundtruth is converted into a 5-channel groundtruth mask using ordinal encoding. The CNN predicts the mask via its multi-channel pixel-level output. Focal loss (FL) trains the CNN with respect to f_{out} using both ADC and T2w inputs. Meanwhile, mutual finding loss (MFL) computes $L2_{\text{ADC}}$ and $L2_{\text{T2w}}$ in the forward-propagation and trains the imaging component of the smaller $L2$.

Given the predicted ordinal encoded vector for a pixel, $\hat{y} = (\hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4, \hat{y}_5) \in \{0,1\}$, 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.

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) [41] to balance the learning between lesion 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 [41]. 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 $\vec{y} = (y_1, y_2, y_3, y_4, y_5) \in \{0,1\}$ be the groundtruth encoded vector corresponding to the 5-channel prediction probability vector $\vec{p} = (p_1, p_2, p_3, p_4, p_5) \in [0,1]$. Then, the FL for each pixel is

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

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

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

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 [42]. 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) 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 [6]. 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, 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, $f_{\mathrm{ADC}} = f\left(I_{\mathrm{ADC}}, I_{\mathrm{blank}}\right)$, $f_{\mathrm{T2w}} = f\left(I_{\mathrm{blank}}, I_{\mathrm{T2w}}\right)$, with the output using both components, $f_{\mathrm{out}} = f\left(I_{\mathrm{ADC}}, I_{\mathrm{T2w}}\right)$. 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., $L2_{\mathrm{ADC}} = \|y \otimes (f_{\mathrm{out}} - f_{\mathrm{ADC}})\|^2$ or $L2_{\mathrm{T2w}} = \|y \otimes (f_{\mathrm{out}} - f_{\mathrm{T2w}})\|^2$, 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, the process of MFL is summarized into a loss term for the end-to-end learning such that

$$\mathrm{MFL} = \frac{1}{N} \min\{L2_{\mathrm{ADC}}, L2_{\mathrm{T2w}}\}, \tag{3}$$

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

4) FocalNet training: FocalNet is trained by the combined loss from FL and MFL,

$$\mathrm{L} = \mathbb{E}_{\vec{p} \sim S(f_{\mathrm{out}})} \mathrm{FL}\left(\vec{p}\right) + \lambda \cdot \mathrm{MFL}, \tag{4}$$

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, 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.

D. Pre-processing & Training

1) Registration: ADC images were registered to T2w images via rigid transformation using scanner coordinate information, as in [11]. 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 [14], 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 [19].

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 [20] 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 [43], [44], 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 [45]. 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) Implementations: The backbone CNN architecture of FocalNet is implemented using Deeplab [46] with the 101-layer deep residual network [47] on 2D image inputs. In the preliminary experiment, we also tested U-Net [45] 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 [48]. 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 [49], 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) [50]. 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.

E. Validation

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 [19], 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.

2) Lesion localization: For PCa detection, we extract lesion localization points from CNN pixel-level detection output as in [13], [51]. 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) FROC for lesion detection: The lesion detection performance is evaluated through free-response receiver operating characteristics (FROC) analysis due to PCa's multi-focality [13], [20]. 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 [52]. 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. II-B. 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.

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 [15]: 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. II-E3. 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 model, the classification result for the lesion is meaningless as well.

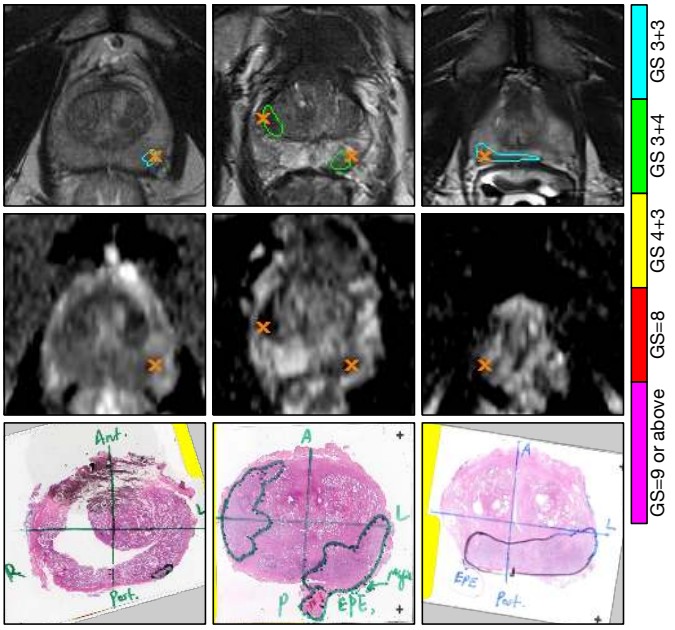


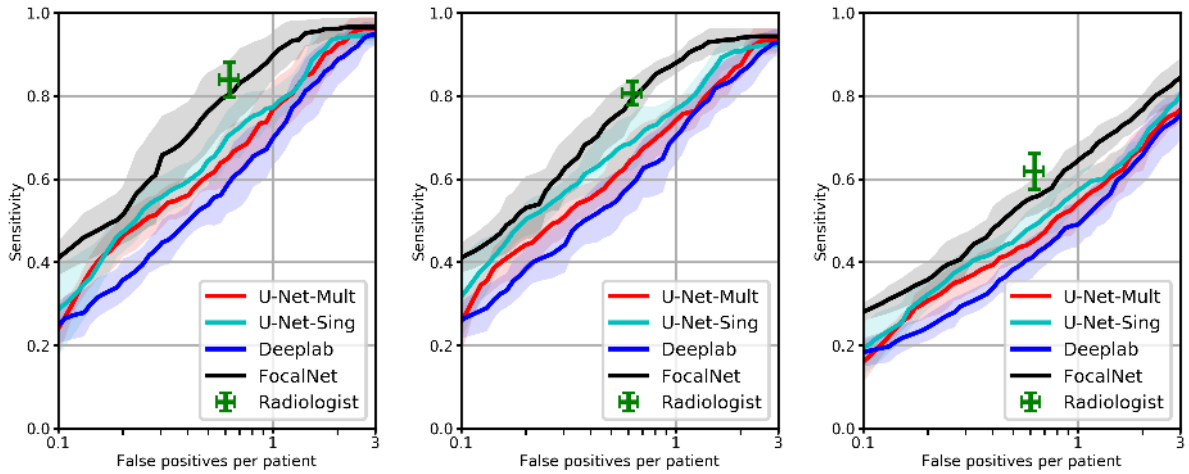
Fig. 4. From top to down shows T2w images, ADC images, and whole-mount specimens. Lesion detection points from FocalNet are shown as the orange cross signs. Groundtruth lesion contours overlay on T2w images with the colors corresponding to their Gleason score groups.

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. II-B. The sensitivity is calculated on the number of true positive findings versus the total number of MRI-visible lesions.

III. Results

A. Baseline methods

Deeplab, *U-Net-Mult*, and *U-Net-Sing* are the three baseline methods in this study. *Deeplab* [46] is the base model of FocalNet, which has the same backbone CNN architecture 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 [45] 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. 2, consisting of image pre-processing, intensity normalization & variation 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.



(a) Index lesions (b) Clinically significant lesions (c) All lesions

Fig. 5. FROC analysis for detection sensitivity on index lesions, clinically significant lesions, and all 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.

TABLE II
FALSE POSITIVES PER PATIENT (FP) AT GIVEN LESION DETECTION SENSITIVITY (SEN). AVG±STD.

	Index lesions		Clinically significant lesions		All lesions	
	FP@Sen80%	FP@Sen90%	FP@Sen80%	FP@Sen90%	FP@Sen60%	FP@Sen80%
U-Net-Mult	1.194±0.387	1.741±0.491	1.386±0.363	2.150±0.596	1.384±0.411	3.525±0.412
U-Net-Sing	1.161±0.373	1.613±0.260	1.211±0.202	1.753±0.550	1.287±0.389	2.982±0.184
Deeplab	1.375±0.401	2.201±0.637	1.454±0.427	2.442±0.802	1.553±0.459	3.698±1.044
FocalNet	**0.610±0.246**	**1.015±0.369**	**0.651±0.149**	**1.130±0.345**	**0.804±0.210**	**2.296±0.608**

TABLE III
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	1.651±0.514	2.161±0.675	1.189±0.316	1.738±0.822	0.122±0.109	0.284±0.258	0.042±0.028	0.097±0.104
U-Net-Sing	1.450±0.273	1.974±1.135	0.860±0.236	1.585±1.476	0.111±0.096	**0.230±0.194**	0.078±0.091	0.210±0.143
Deeplab	1.410±0.806	2.458±1.132	1.131±0.335	1.821±0.500	0.273±0.112	0.399±0.324	0.061±0.020	0.244±0.186
FocalNet	**1.211±0.483**	**1.763±0.631**	**0.577±0.180**	**0.899±0.779**	**0.071±0.108**	0.231±0.143	**0.035± 0.018**	**0.055±0.065**

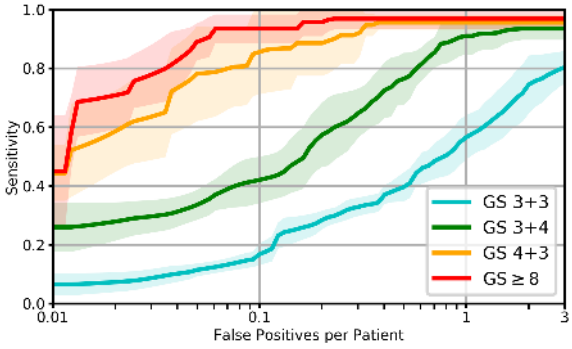
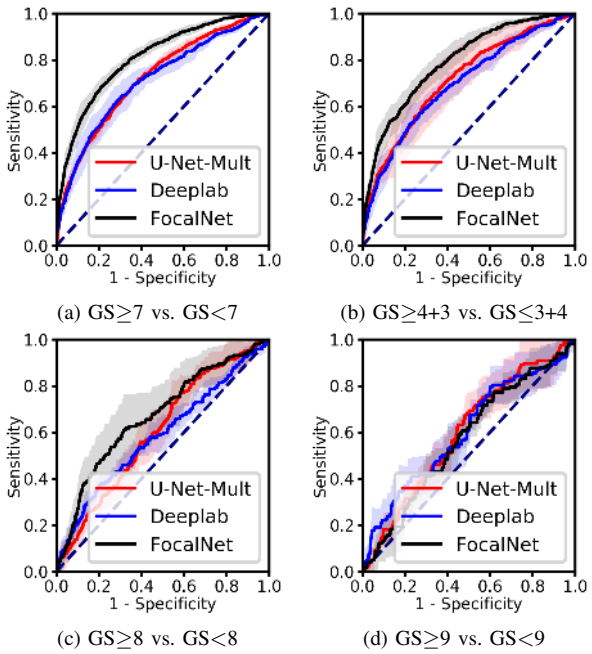


Fig. 6. FROC analysis for detection sensitivity of FocalNet for each specific Gleason score group. Transparent areas are 95% confidence intervals estimated by two times of the standard deviation. The results of baseline methods are reported in TABLE III.

B. Lesion detection

Fig. 5 shows the FROC analysis for index lesions, clinically significant lesions, and all lesions, respectively, and examples for lesion detection are shown in Fig. 4. As in TABLE II, 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. 5b, 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. 5c, 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.

The radiologist performance is shown in Fig. 5 as green markers. Radiologists achieved 83.9% sensitivity for index



(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

Fig. 7. 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.

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. 6 and TABLE III. 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*.

C. Gleason score prediction

Fig. 7a and Fig. 7b 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. 7c and Fig. 7d, 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).

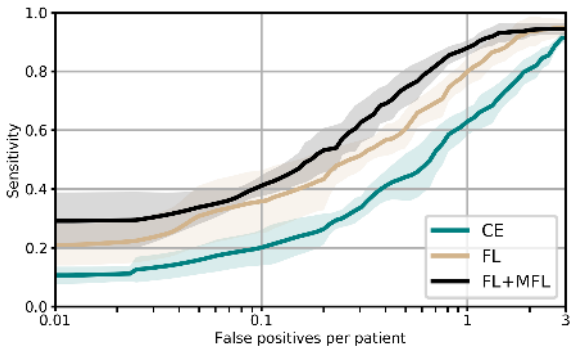


Fig. 8. FROC analysis for the detection of clinically significant lesions using three different loss combinations during the training: cross-entropy loss (*CE*), focal loss (*FL*), and the combined loss from focal loss and mutual finding loss (*FL+MFL*). The number of false positives per patient (x-axis) is shown on log-scale. The transparent areas are 95% confidence intervals estimated by two times of the standard deviation.

D. 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 II-C4. As shown in Fig. 8, *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. 5b, as the ordinal encoding strategy caused the labels to become more imbalanced for *CE*.

E. 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: 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. II-D2. As shown in Fig. 9, 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

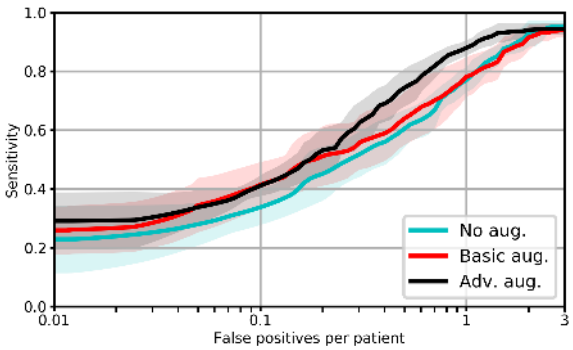


Fig. 9. FROC analysis for the detection of clinically significant lesions under three different augmentation strategies during the training: with no augmentation, with basic augmentation (image shifting, scaling and flipping), and with advanced augmentation (basic augmentation + intensity variation). Transparent areas are 95% confidence intervals estimated by two times of the standard deviation.

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.

IV. 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 [7], 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. II-B. 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.* [51] 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.* [53] 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 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

$$\mathrm{MFL} = \min_{1 \le i \le m} \|y \otimes (f_{\mathrm{out}} - f_i)\|^2, \tag{5}$$

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. 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 predetermined combinations of imaging components, similar to PI-RADS v2 [6], 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 [8]. 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 [54], 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 [12], [15], 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 [10], [43]. Further study is needed to consolidate the correlation between mp-MRI and PCa aggressiveness, particularly with available molecular subtypes of PCa [55].

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 [8], [14], [17], which could potentially include some inaccuracies because of the discrepancy between prostate biopsy and radical prostatectomy in histologic findings. Recently, the ProstateX Challenge [13] 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 [33] 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.

REFERENCES

[1] J. I. Epstein, L. Egevad, M. B. Amin, B. Delahunt, J. R. Srigley, and P. A. Humphrey, "The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma," *The American journal of surgical pathology*, vol. 40, no. 2, pp. 244–252, 2016.

[2] J. H. Yacoub, S. Verma, J. S. Moulton, S. Eggener, and A. Oto, "Imaging-guided prostate biopsy: conventional and emerging techniques," *Radiographics*, vol. 32, no. 3, pp. 819–837, 2012.

[3] V. Kasivisvanathan, A. S. Rannikko, M. Borghi, V. Panebianco, L. A. Mynderse, M. H. Vaarala, A. Briganti, L. Budäus, G. Hellawell, R. G. Hindley *et al.*, "Mri-targeted or standard biopsy for prostate-cancer diagnosis," *The New England Journal of Medicine*, vol. 378, no. 19, pp. 1767–1777, 2018.

[4] M. A. DallEra, P. C. Albertsen, C. Bangma, P. R. Carroll, H. B. Carter, M. R. Cooperberg, S. J. Freedland, L. H. Klotz, C. Parker, and M. S. Soloway, "Active surveillance for prostate cancer: a systematic review of the literature," *European Urology*, vol. 62, no. 6, pp. 976–983, 2012.

[5] M. Valerio, H. U. Ahmed, M. Emberton, N. Lawrentschuk, M. Lazzeri, R. Montironi, P. L. Nguyen, J. Trachtenberg, and T. J. Polascik, "The role of focal therapy in the management of localised prostate cancer: a systematic review," *European urology*, vol. 66, no. 4, pp. 732–751, 2014.

[6] J. C. Weinreb, J. O. Barentsz, P. L. Choyke, F. Cornud, M. A. Haider, K. J. Macura, D. Margolis, M. D. Schnall, F. Shtern, C. M. Tempany *et al.*, "Pi-rads prostate imaging–reporting and data system: 2015, version 2," *European Urology*, vol. 69, no. 1, pp. 16–40, 2016.

[7] O. Ruprecht, P. Weisser, B. Bodelle, H. Ackermann, and T. J. Vogl, "Mri of the prostate: interobserver agreement compared with histopathologic outcome after radical prostatectomy," *European journal of radiology*, vol. 81, no. 3, pp. 456–460, 2012.

[8] S. Ozer, D. L. Langer, X. Liu, M. A. Haider, T. H. van der Kwast, A. J. Evans, Y. Yang, M. N. Wernick, and I. S. Yetik, "Supervised and unsupervised methods for prostate cancer segmentation with multispectral mri," *Medical Physics*, vol. 37, no. 4, pp. 1873–1883, 2010.

[9] P. Vos, J. Barentsz, N. Karssemeijer, and H. Huisman, "Automatic computer-aided detection of prostate cancer based on multiparametric magnetic resonance image analysis," *Physics in Medicine & Biology*, vol. 57, no. 6, p. 1527, 2012.

[10] Y. Peng, Y. Jiang, C. Yang, J. B. Brown, T. Antic, I. Sethi, C. Schmid-Tannwald, M. L. Giger, S. E. Eggener, and A. 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*, vol. 267, no. 3, pp. 787–796, 2013.

[11] P. Liu, S. Wang, B. Turkbey, K. Grant, P. Pinto, P. Choyke, B. J. Wood, and R. 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*, vol. 8670. International Society for Optics and Photonics, 2013, p. 86701G.

[12] P. Tiwari, J. Kurhanewicz, and A. Madabhushi, "Multi-kernel graph embedding for detection, gleason grading of prostate cancer via mri/mrs," *Medical image analysis*, vol. 17, no. 2, pp. 219–235, 2013.

[13] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, "Computer-aided detection of prostate cancer in mri," *IEEE Transactions on Medical Imaging*, vol. 33, no. 5, pp. 1083–1092, 2014.

[14] S. Wang, K. Burtt, B. Turkbey, P. Choyke, and R. M. Summers, "Computer aided-diagnosis of prostate cancer on multiparametric mri: a technical review of current research," *BioMed Research International*, vol. 2014, 2014.

[15] D. Fehr, H. Veeraraghavan, A. Wibmer, T. Gondo, K. Matsumoto, H. A. Vargas, E. Sala, H. Hricak, and J. O. Deasy, "Automatic classification of prostate cancer gleason scores from multiparametric magnetic resonance images," *Proceedings of the National Academy of Sciences*, vol. 112, no. 46, pp. E6265–E6273, 2015.

[16] G. Lemaître, R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau, "Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric mri: a review," *Computers in biology and medicine*, vol. 60, pp. 8–31, 2015.

[17] J. T. Kwak, S. Xu, B. J. Wood, B. Turkbey, P. L. Choyke, P. A. Pinto, S. Wang, and R. M. Summers, "Automated prostate cancer detection using t2-weighted and high-b-value diffusion-weighted magnetic resonance imaging," *Medical Physics*, vol. 42, no. 5, pp. 2368–2378, 2015.

[18] A. Cameron, F. Khalvati, M. A. Haider, and A. Wong, "Maps: a quantitative radiomics approach for prostate cancer detection," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 6, pp. 1145–1156, 2016.

[19] A. P. Kiraly, C. A. Nader, A. Tuysuzoglu, R. Grimm, B. Kiefer, N. El-Zehiry, and A. Kamen, "Deep convolutional encoder-decoders for prostate cancer detection and classification," in *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 489–497.

[20] Y. Tsehay, N. Lay, X. Wang, J. T. Kwak, B. Turkbey, P. Choyke, P. Pinto, B. Wood, and R. 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*. IEEE, 2017, pp. 642–645.

[21] Y. Song, Y.-D. Zhang, X. Yan, H. Liu, M. Zhou, B. Hu, and G. Yang, "Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric mri," *Journal of Magnetic Resonance Imaging*, 2018.

[22] I. Reda, B. O. Ayinde, M. Elmogy, A. Shalaby, M. El-Melegy, M. A. El-Ghar, A. A. El-fetouh, M. Ghazal, and A. El-Baz, "A new cnn-based system for early diagnosis of prostate cancer," in *Proceedings of IEEE International Symposium on Biomedical Imaging*. IEEE, 2018, pp. 207–210.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of Advances in Neural Information Processing Systems*, 2012, pp. 1097–1105.

[24] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision Pattern Recognition*, 2015, pp. 3431–3440.

[25] J. Zhang, A. Saha, Z. Zhu, and M. A. Mazurowski, "Hierarchical convolutional neural networks for segmentation of breast tumors in mri with application to radiogenomics," *IEEE Transactions on Medical Imaging*, 2018.

[26] W. Zhang, R. Li, H. Deng, L. Wang, W. Lin, S. Ji, and D. Shen, "Deep convolutional neural networks for multi-modality isointense infant brain image segmentation," *NeuroImage*, vol. 108, pp. 214–224, 2015.

[27] S. Pereira, A. Pinto, V. Alves, and C. A. Silva, "Brain tumor segmentation using convolutional neural networks in mri images," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1240–1251, 2016.

[28] J. R. Stark, S. Perner, M. J. Stampfer, J. A. Sinnott, S. Finn, A. S. Eisenstein, J. Ma, M. Fiorentino, T. Kurth, M. Loda *et al.*, "Gleason score and lethal prostate cancer: does 3+ 4= 4+ 3?" *Journal of Clinical Oncology*, vol. 27, no. 21, p. 3459, 2009.

[29] J. D. Le, N. Tan, E. Shkolyar, D. Y. Lu, L. Kwan, L. S. Marks, J. Huang, D. J. Margolis, S. S. Raman, and R. E. Reiter, "Multifocality and prostate cancer detection by multiparametric magnetic resonance imaging: correlation with whole-mount histopathology," *European Urology*, vol. 67, no. 3, pp. 569–576, 2015.

[30] J. P. Radtke, C. Schwab, M. B. Wolf, M. T. Freitag, C. D. Alt, C. Kesch, I. V. Popeneciu, C. Huettenbrink, C. Gasch, T. 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*, vol. 70, no. 5, pp. 846–853, 2016.

[31] H. Vargas, A. Hötker, D. Goldman, C. Moskowitz, T. Gondo, K. Matsumoto, B. Ehdaie, S. Woo, S. Fine, V. 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*, vol. 26, no. 6, pp. 1606–1612, 2016.

[32] S. Borofsky, A. K. George, S. Gaur, M. Bernardo, M. D. Greer, F. V. Mertan, M. Taffel, V. Moreno, M. J. Merino, B. J. Wood *et al.*, "What are we missing? false-negative cancers at multiparametric mr imaging of the prostate," *Radiology*, vol. 286, no. 1, pp. 186–195, 2018.

[33] J. D. Le, S. Stephenson, M. Brugger, D. Y. Lu, P. Lieu, G. A. Sonn, S. Natarajan, F. J. Dorey, J. Huang, D. J. Margolis *et al.*, "Magnetic resonance imaging-ultrasound fusion biopsy for prediction of final prostate pathology," *The Journal of urology*, vol. 192, no. 5, pp. 1367–1373, 2014.

[34] M. Garmer, M. Busch, S. Mateiescu, D. E. Fahlbusch, B. Wagener, and D. H. 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*, vol. 22, no. 11, pp. 1409–1418, 2015.

[35] J. I. Epstein, Z. Feng, B. J. Trock, and P. 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*, vol. 61, no. 5, pp. 1019–1024, 2012.

[36] J. I. Epstein, M. J. Zelefsky, D. D. Sjoberg, J. B. Nelson, L. Egevad, C. Magi-Galluzzi, A. J. Vickers, A. V. Parwani, V. E. Reuter, S. W. Fine *et al.*, "A contemporary prostate cancer grading system: a validated alternative to the gleason score," *European Urology*, vol. 69, no. 3, pp. 428–435, 2016.

[37] P. Kozlowski, S. D. Chang, E. C. Jones, and S. L. 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*, vol. 31, no. 3, p. e3882, 2018.

[38] G. Ploussard, J. I. Epstein, R. Montironi, P. R. Carroll, M. Wirth, M.-O. Grimm, A. S. Bjartell, F. Montorsi, S. J. Freedland, A. Erbersdobler *et al.*, "The contemporary concept of significant versus insignificant prostate cancer," *European Urology*, vol. 60, no. 2, pp. 291–303, 2011.

[39] J. Cheng, Z. Wang, and G. Pollastri, "A neural network approach to ordinal regression," in *IEEE International Joint Conference on Neural Networks*. IEEE, 2008, pp. 1279–1284.

[40] P. A. Gutierrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, and C. Hervas-Martinez, "Ordinal regression methods: survey and experimental study," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 127–146, 2016.

[41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of IEEE International Conference on Computer Vision*. IEEE, 2017, pp. 2999–3007.

[42] H. Miao, H. Fukatsu, and T. Ishigaki, "Prostate cancer detection with 3-t mri: comparison of diffusion-weighted and t2-weighted imaging," *European Journal of Radiology*, vol. 61, no. 2, pp. 297–302, 2007.

[43] H. A. Vargas, O. Akin, T. Franiel, Y. Mazaheri, J. Zheng, C. Moskowitz, K. Udo, J. Eastham, and H. Hricak, "Diffusion-weighted endorectal mr imaging at 3 t for prostate cancer: tumor detection and assessment of aggressiveness," *Radiology*, vol. 259, no. 3, pp. 775–784, 2011.

[44] T. Hambrock, D. M. Somford, H. J. Huisman, I. M. van Oort, J. A. Witjes, C. A. Hulsbergen-van de Kaa, T. Scheenen, and J. O. Barentsz, "Relationship between apparent diffusion coefficients at 3.0-t mr imaging and gleason grade in peripheral zone prostate cancer," *Radiology*, vol. 259, no. 2, pp. 453–461, 2011.

[45] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[46] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision Pattern Recognition*, 2016, pp. 770–778.

[48] S. Hoo-Chang, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, p. 1285, 2016.

[49] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. Frackowiak, "Statistical parametric maps in functional imaging: a general linear approach," *Human brain mapping*, vol. 2, no. 4, pp. 189–210, 1994.

[50] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, "Tensorflow: a system for large-scale machine learning." in *USENIX Symposium on Operating Systems Design and Implementation*, vol. 16, 2016, pp. 265–283.

[51] Z. Wang, C. Liu, D. Cheng, L. Wang, X. Yang, and K.-T. 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*, vol. 37, no. 5, pp. 1127–1139, 2018.

[52] A. Priester, S. Natarajan, P. Khoshnoodi, D. J. Margolis, S. S. Raman, R. E. Reiter, J. Huang, W. Grundfest, and L. 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*, vol. 197, no. 2, pp. 320–326, 2017.

[53] L. Fidon, W. Li, L. C. Garcia-Peraza-Herrera, J. Ekanayake, N. Kitchen, S. Ourselin, and T. Vercauteren, "Scalable multimodal convolutional networks for brain tumour segmentation," in *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 285–293.

[54] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Proceedings of Advances in Neural Information Processing Systems*, 2011, pp. 109–117.

[55] J. Lapointe, C. Li, J. P. Higgins, M. Van De Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim *et al.*, "Gene expression profiling identifies clinically relevant subtypes of prostate cancer," *Proceedings of the National Academy of Sciences*, vol. 101, no. 3, pp. 811–816, 2004.