


Deep Learning Regression for Prostate Cancer Detection and Grading in Bi-Parametric MRI

Coen de Vente , Pieter Vos , Matin Hosseinzadeh, Josien Pluim, and Mitko Veta 

Abstract—One of the most common types of cancer in men is prostate cancer (PCa). Biopsies guided by bi-parametric magnetic resonance imaging (MRI) can aid PCa diagnosis. Previous works have mostly focused on either detection or classification of PCa from MRI. In this work, however, we present a neural network that simultaneously detects and grades cancer tissue in an end-to-end fashion. This is more clinically relevant than the classification goal of the ProstateX-2 challenge. We used the dataset of this challenge for training and testing. We use a 2D U-Net with MRI slices as input and lesion segmentation maps that encode the Gleason Grade Group (GGG), a measure for cancer aggressiveness, as output. We propose a method for encoding the GGG in the model target that takes advantage of the fact that the classes are ordinal. Furthermore, we evaluate methods for incorporating prostate zone segmentations as prior information, and ensembling techniques. The model scored a voxel-wise weighted kappa of 0.446 ± 0.082 and a Dice similarity coefficient for segmenting clinically significant cancer of 0.370 ± 0.046 , obtained using 5-fold cross-validation. The lesion-wise weighted kappa on the ProstateX-2 challenge test set was 0.13 ± 0.27 . We show that our proposed model target outperforms standard multiclass classification and multi-label ordinal regression. Additionally, we present a comparison of methods for further improvement of the model performance.

Index Terms—Prostate cancer, bi-parametric MRI, Gleason Grade Group, U-Net, deep learning, ordinal regression.

I. INTRODUCTION

PROSTATE cancer (PCa) is the most frequently diagnosed type of cancer among men in most countries [1], accounting for nearly 1 in 5 cancer diagnoses [2]. It was estimated that there were 1.3 million new cases of PCa and 359,000 associated deaths worldwide in 2018 [1]. Methods to detect PCa in an early stage are Prostate Specific Antigen (PSA) measurement and Digital Rectal Examination (DRE) [3]. When these tests indicate the

possibility of PCa, a transrectal ultrasound (TRUS) systematic biopsy is recommended by the guidelines of the European Association of Urology [4]. PSA and DRE, however, have a relatively low sensitivity and specificity [5]. Furthermore, TRUS is invasive, has a relatively low sensitivity and underestimates aggressiveness [6], [7].

Multi-parametric magnetic resonance imaging (mp-MRI) has shown to improve sensitivity, potentially reduce unnecessary biopsies by a quarter, and lower over-diagnosis compared to standard TRUS [8]. PCa analysis with mp-MRI is, nevertheless, a labor intensive procedure and requires a high level of experience [9], [10]. This has inhibited the implementation of MRI guided decision making in most clinics. Automated analyses of these images can potentially overcome these problems, thereby encouraging the use of mp-MRI in more screening environments. However, scanning time for mp-MRI is more than 30 minutes [11]. Bi-parametric magnetic resonance imaging (bp-MRI) has been introduced as a faster alternative without compromising the diagnostic accuracy for PCa [11], [12]. For bp-MRI, only the T2-weighted scan and diffusion weighted imaging (DWI) scans (to compute the apparent diffusion coefficient (ADC) map) are needed. This reduces scanning time to about 17 minutes [11]. It has been shown that bp-MRI has a comparable diagnostic accuracy to mp-MRI [11]–[14].

A system that not only detects PCa, but is also able to predict aggressiveness, can potentially provide the radiologist with more information than a system that only detects PCa. A measure for PCa aggressiveness is the Gleason Grade Group (GGG). It is a predictor of pathological stage and oncological outcome, and can be assigned to potential lesions using the histopathological analysis of biopsies [15]. GGG ranges from 1 to 5, where a GGG of 1 generally requires no treatment, while a GGG of 5 is the most severe type of PCa [15]. Table I provides a description of all GGGs.

Previous works have explored different deep learning methods for PCa detection. Tsehay *et al.* [16] used a convolutional neural network (CNN) with five layers and outputs at every layer to produce a probability map of PCa. Kohl *et al.* [17] used a U-Net [18] with an adversarial loss to segment clinically significant prostate cancer (csPCa) with a Gleason score ≥ 7 , which is equivalent to $GGG \geq 2$.

Other works have focused on the classification of lesions as clinically significant or insignificant. Liu *et al.* [19] trained a 6-layer CNN for classification. Others have used transfer learning for this task [20], [21]. Mehrtaash *et al.* [22] used a 3D CNN for distinguishing these two classes. They used the

Manuscript received March 25, 2020; accepted April 27, 2020. Date of publication May 8, 2020; date of current version January 20, 2021. (Corresponding author: Coen de Vente.)

Coen de Vente is with the Department of Biomedical Engineering, Eindhoven University of Technology 5600 MB, Eindhoven, The Netherlands, and also with the Philips Research 5656, AE, Eindhoven, The Netherlands (e-mail: coendevente@gmail.com).

Pieter Vos is with the Philips Research.

Matin Hosseinzadeh is with the Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center.

Josien Pluim and Mitko Veta are with the Department of Biomedical Engineering, Eindhoven University of Technology.

Digital Object Identifier 10.1109/TBME.2020.2993528

TABLE I
GLEASON GRADE GROUP DESCRIPTIONS

GGG	1	2	3	4	5
Tissue description [29]	Only individual discrete well-formed glands.	Mainly well-formed glands with lesser component of poorly-formed / fused / cribriform glands.	Mainly poorly-formed / fused / cribriform glands with lesser component of well-formed glands.	Only poorly-formed / fused / cribriform glands <i>or</i> mainly well-formed glands and lesser component lacking glands <i>or</i> mainly lacking glands and lesser component of well-formed glands.	Lacks gland formation (or with necrosis) with or without poorly formed / fused / cribriform glands.

dataset provided by the ProstateX challenge, which focused on this binary classification task [23]. More classical machine learning techniques than deep learning have also been used [24].

Multiple works have focused on grading lesions by classifying their GGG from mp-MRI. Jensen *et al.* investigated GGG grading from MRI using a k -nearest neighbor classifier, while merging some GGGs to the same class [25]. Moreover, the goal of the ProstateX-2 challenge was to grade lesions by predicting their GGGs [23]. The winner of the ProstateX-2 challenge [26] used handcrafted texture features as input of a stacked sparse autoencoder, with the five GGG classes as output.

Recently, Cao *et al.* [27] performed simultaneous grading and detection by training an end-to-end network that predicted Gleason score groups for each voxel in the image. They ordinally encoded the Gleason score groups. One of the novelties of our work is the proposed method for ordinally encoding this model target, which we refer to as soft-label ordinal regression, as an extension of [28]. In this work, we compare their proposed method for encoding the classes, which we refer to as multi-label ordinal regression, to our method for ordinally encoding classes. Moreover, we present a detailed comparison of methods for additional performance improvement. We also propose to include zonal information into the network architecture, which could aid the training process, as lesions have different characteristics in different prostate zones.

The ProstateX-2 challenge was organized by the American Association of Physicists in Medicine (AAPM), together with the Society of Photo-Optical Instrumentation Engineers (SPIE) and the National Cancer Institute (NCI). The goal of this challenge was to grade lesions from mp-MRI, given their coordinates in a volume. Thus, the task was only GGG classification and not detection. The goal of the current work is to simultaneously detect and grade lesions from bp-MRI. In contrast to the methods proposed by the ProstateX-2 contestants, there is no need to manually indicate suspicious regions in the image, which is closer to the potential clinical application.

We compare three different approaches of encoding the GGG in the model target, of which two take into account that the classes are ordinal, i.e., $GGG\ 1 < 2 < \dots < 5$. This is unlike the classes in most other classification problems, where the classes are not ordinal. The manner in which the GGGs are encoded in the model targets vary in these approaches. We use a U-Net [18] with bp-MRI images as input, and segmentations of the lesions that encode the corresponding GGG of each lesion as target.

Furthermore, we explore different methods for the incorporation of zonal information in the network, and investigate the effect of ensemble learning.

II. METHODS

A. Dataset

The ProstateX-2 challenge train set contains 99 patients and 112 lesions. This dataset was used for training and validation. The ProstateX-2 challenge test set contains 63 patients and 70 lesions. The ground truth associated with this test set is not publicly available. All evaluation was done by the challenge organizers.

All scans were acquired at the Radboud University Medical Center. They were read by a radiologist with 20 years of experience. Findings to which the radiologist assigned a PI-RADS score of at least 3 were referred to biopsy.

A pathologist with over 20 years of experience performed analysis of the MRI targeted biopsies and defined the GGGs of the lesions. The challenge dataset contains 36, 41, 20, 8 and 7 lesions of GGG 1, 2, 3, 4 and 5, respectively. The dataset provides coordinates of the lesion centers. We delineated the lesions using in-house software, which is based on a semi-automated region growing technique. The centroid coordinates of the lesions that were acquired from the MRI targeted biopsy were used to make these delineations.

The images were acquired with a 3 T MAGNETOM Trio and Skyra (Siemens Medical Systems) scanner, without an endorectal coil. T2-weighted scans had an in-plane resolution of around 0.5 mm and a slice thickness of 3.6 mm. They were acquired with a turbo spin echo sequence. The ADC scans were calculated using three DWI scans (b-values of 50, 400 and 800). The DWI scans were acquired with a single-shot echo planar imaging sequence with diffusion-encoding gradients in three directions, and had an in-plane resolution of 2 mm and a slice thickness of 3.6 mm.

B. Preprocessing

We used a fixed ROI size of $90 \times 90 \times 80\text{ mm}^3$ to crop the volumes around the image center. We visually verified that the prostate gland was fully contained in the ROIs. The ROIs were resized to $192 \times 192 \times 32$. This size was chosen because this height and width are divisible by two a sufficient number of

times to be able to perform max-pooling and up-sampling six times (which is further discussed in Section II-C).

We registered the ADC scans to the corresponding T2-weighted images, to correct for patient movement during scanning. Most misalignments were expected to be caused by rotation and translation of the patient, as the prostate generally does not deform much. Hence, we performed rigid registration. Since anatomical structures have different intensities in these MRI types, we used mutual information as a metric. Also, we used gradient descent with a learning rate of 1.0 as the optimizer. SimpleITK [30] was used to perform registration.

The images were scaled between 0 and 1. The T2-weighted scans are qualitative, so we normalized these images independently per volume image. In contrast, ADC is quantitative, so the ADC images were normalized over the entire dataset.

In our experiments, we compared 2D to 2.5D methods. In the 2D method, we trained the model on individual slices. The input in the 2.5D approach also contained 2 slices above and 2 slices below the middle slice. The target in the 2.5D method, however, was only the output corresponding to the middle input slice.

C. Network Architecture

The network architectures that we used in this work were all adaptations of U-Net [18]. Fig. 1(a) shows a schematic overview of this architecture. Such a network is a type of encoder-decoder. The final convolution of the network is followed by a 1×1 convolution with sigmoid or softmax activation. This activation function varied per model target, which is explained in Section III-A.

D. Model Target

We compared three different ways to generate the model targets. For all methods, the network assigns a label for each voxel in the input. However, the way the class is represented is different for each method. The multiclass classification target, unlike the other two methods, is not an ordinal approach. One of the ordinal approaches is a common ordinal regression method in literature [27], [31], [32], and the other is our proposed method. Fig. 2 shows an example of a target slice with two lesions, for each of the three model targets.

The ordinal approaches have two advantages. Firstly, these methods penalize a prediction that is further from the target more than a prediction that is closer. In contrast, the multiclass classification method penalizes all incorrect predictions equally. Moreover, when training with the multiclass classification target, the network only learns for one class per sample. In contrast, for the ordinal approaches, the GGGs share output classes, so the model learns for multiple classes with one sample.

The target encodes five different classes. The first class is the background class and consists of clinically insignificant lesions (GGG = 1) and healthy tissue. The other classes correspond to csPCa, where each class is a different GGG.

1) Multiclass Classification: The classes are one-hot encoded. Hence, the output target consists of five output channels, where each output channel corresponds to one class. If a voxel, for example, corresponds to a GGG of 3, the target

is $(0, 0, 1, 0, 0)$. Since each sample is assigned to exactly one class, we use softmax activation after the final convolution. We can express the softmax function S as:

$$S(\mathbf{z})_{i \in 1..K} = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}}, \quad (1)$$

where \mathbf{z} is the input of the activation function, and K is the number of classes. The class with the maximum probability is chosen during test-time.

2) Multi-Label Ordinal Regression: This method was proposed by Cheng *et al.* [32] and used by Cao *et al.* for GGG prediction from mp-MRI [27]. Let K be the number of classes. If the class for a sample is k ($1 \leq k \leq K$), the sample is also assigned to all lower order classes (i.e., all classes $1, \dots, k-1$). Thus, the label for that sample is $(t_1, t_2, \dots, t_{K-1})$, where t_i ($1 \leq i \leq K-1$) is 1 if $k > i$ and 0 otherwise. In this work, $K = 5$, so if, for example, a voxel corresponds to a GGG of 3, the target is $(1, 1, 0, 0)$. Since all channels of one sample could be 0 or multiple channels could be 1, the output classes of one voxel should not always all sum to 1. Hence, sigmoid activation is computed separately for each neuron. The sigmoid function S is defined as:

$$S(\mathbf{z})_{i \in 1..K-1} = \frac{e^{z_i}}{e^{z_i} + 1}, \quad (2)$$

where \mathbf{z} is the input of the activation function. This activation ensures that the output of each class is between 0 and 1.

During test-time, the class is determined by counting the number of classes for which the probability is at least 0.5.

3) Soft-Label Ordinal Regression: The target consists of only one output channel. It can be seen as a single soft label, where a higher order class corresponds to a greater probability. In fact, the lowest and highest order classes are assigned to a probability of 0 and 1, respectively. The remaining classes are linearly scaled between these scalar values. Essentially, normalized values of the classes are directly used for the target. The normalization function $N(k)$ can be expressed as $N(k) = \frac{k-1}{K-1}$. For example, if the label is 3, the target will become 0.5. As all output values are between 0 and 1, and there is only one output channel, the sigmoid activation is used in the final layer.

During test-time, the prediction probabilities are converted back to classes. The class that is assigned to a probability is the class for which that probability is closest to the normalized value of that class. For example, if the prediction is 0.45, the closest normalized class is 0.5 (as $N(2) = 0.25$, $N(3) = 0.5$, and $N(4) = 0.75$), which corresponds to class 3.

E. Training

There are many more voxels that are not part of a lesion than voxels that are part of a lesion. Because of this class imbalance, weighted cross-entropy was used as a loss function. For the different model targets, the implementations for weighting the classes were slightly different. In multiclass classification, the cross-entropy loss $\mathcal{L}_{MCC}(p, y)$ for prediction p and label y

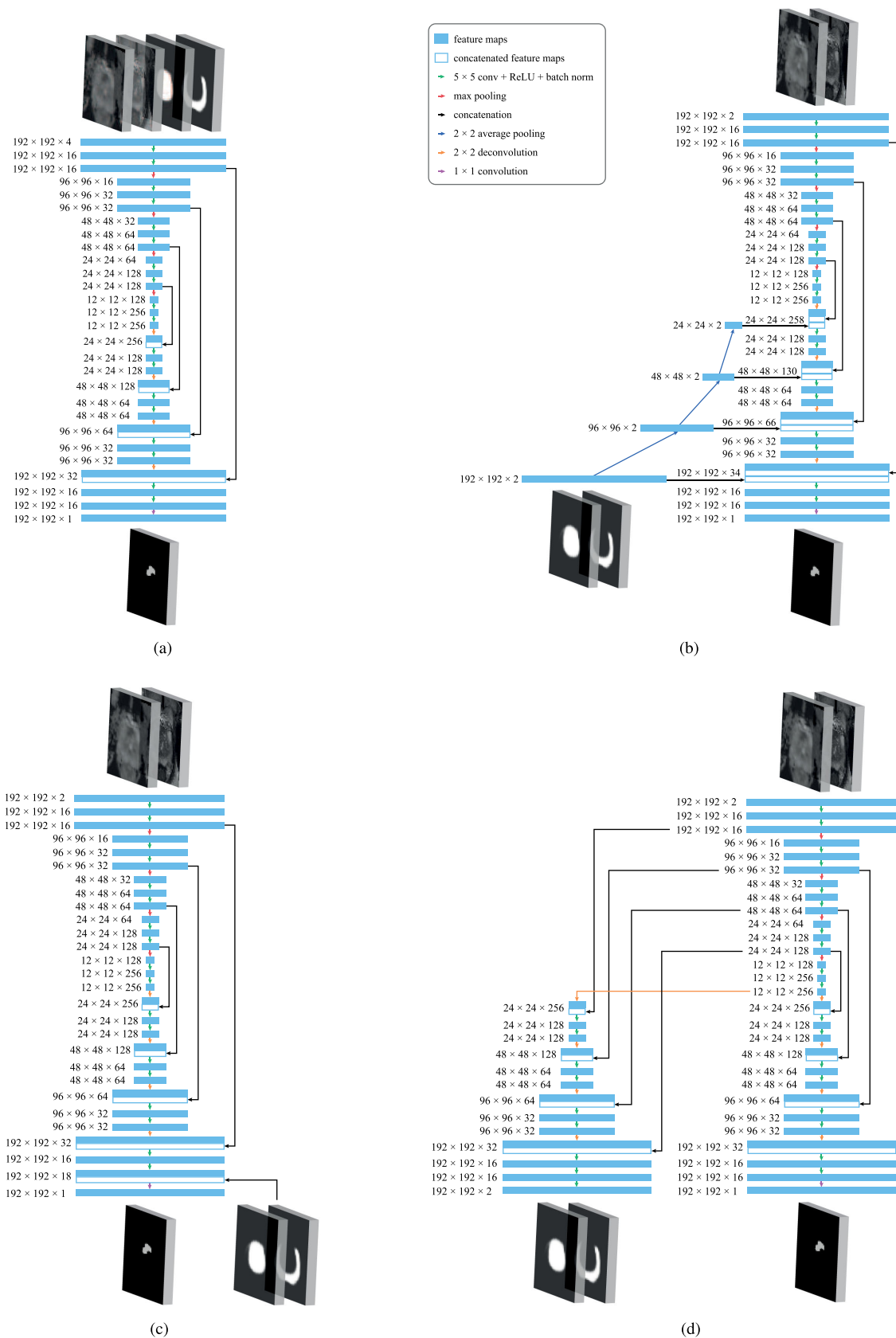


Fig. 1. Network architectures of different approaches to incorporate zonal information. Next to the feature maps, the dimensions of the feature maps are given as *width* \times *height* \times *channels*. (a) Zonal input before first convolution. Note that this figure with only the first two input channels is equivalent to the main processing pipeline. (b) Zonal input in decoder. (c) Zonal input before final convolution. (d) Zonal information as auxiliary outputs.

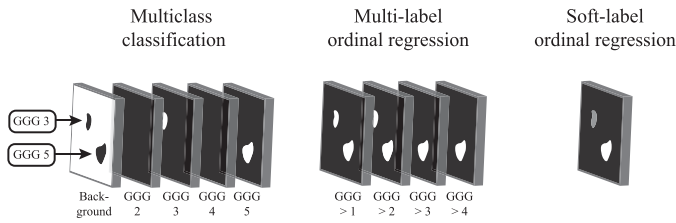


Fig. 2. Different model targets. The same slice with GGG 3 lesion (top left) and GGG 5 lesion (bottom right) is shown.

could be expressed as:

$$\mathcal{L}_{MCC}(p, y) = -y_1 \log(p_1) - \alpha \sum_{k=2}^K y_k \log(p_k), \quad (3)$$

where α is a weighting factor, $p_k \in [0, 1]$ is the prediction of class k , and $y_k \in \{0, 1\}$ is the label of class k . Since the weighting factor is independent of the class, each foreground class is weighted equally.

In multi-label ordinal regression, the loss $\mathcal{L}_{MLOR}(p, y)$ was weighted as follows:

$$\mathcal{L}_{MLOR}(p, y) = \sum_{k=1}^{K-1} -(1 - y_k) \log(1 - p_k) - \alpha y_k \log(p_k). \quad (4)$$

In soft-label ordinal regression, we write the loss $\mathcal{L}_{SLOR}(p, y)$ as:

$$\mathcal{L}_{SLOR}(p, y) = -(1 - y) \log(1 - p) - \alpha y \log(p). \quad (5)$$

p_k and y_k are absent in this formula, as there is only one output channel in this model target.

We used $\alpha = 10$, as we observed that this yielded the best results.

The models were trained until convergence or overfitting occurred by using early stopping with a patience of 15 epochs, which was approximately equal to 10^4 iterations. The networks were implemented in Python using Keras [33] and TensorFlow [34]. The models were trained with a batch size of 4, Adam optimization [35] with a learning rate of 10^{-4} , $\beta_1 = 0.9$, and $\beta_2 = 0.999$. As a means of regularization, we augmented the data on the fly using random elastic deformation, gamma correction with γ between 0.5 and 2.0, rotation between -20° and $+20^\circ$, shearing between -10% and $+10\%$, and flipping along the y-axis with a probability of 50%. To prevent the models from classifying all voxels as background, the model was fed slices with at least one lesion voxel half of the time, and slices without any lesion voxels the other half of the time.

F. Incorporating Zonal Information

We explored four different approaches for incorporating prostate zone information in the network. This information consisted of probabilistic segmentation maps of the prostate zones. One of the reasons for using zonal information is that PCa lesions look differently in MRI images, depending on the zone in which they are located. Secondly, information about

the zones is useful because PCa is more likely to occur in the peripheral zone (PZ) than in the transition zones (TZ) [36], [37]. We did not mask the input with the prostate gland, as tumors can outgrow the prostate, which would lead to the partial exclusion of some lesions from the network input.

The two different prostate zones considered in this work were the PZ and the TZ. We used the non-thresholded softmax output of a zonal segmentation network, as previous work has shown that this improves PCa detection, compared to when the thresholded output is used [38]. The segmentation network was an anisotropic 3D U-Net, trained on 53 T2-weighted MRI volumes, as described by Mooij *et al.* [39]. Fig. 3 shows examples of the segmentation network output.

1) Input Before First Convolution: The probability maps were concatenated with the original input channels. The only difference with the baseline method was that there were four input channels instead of two. The architecture of this method is displayed in Fig. 1(a). Hosseinzadeh *et al.* [38] explored this method of incorporating zonal information in a U-Net for the detection of csPCa.

2) Input in Decoder: We input the zonal probability maps in all U-Net levels of the decoder. Before each layer where feature maps from the encoder are concatenated, downscaled versions of the zonal probability maps were concatenated. The maps were downsampled using average pooling, such that the x- and y-dimensions were equal to the dimensions of the concatenated feature maps. Fig. 1(b) depicts this method.

3) Input Before Final Convolution: Instead of concatenating the zonal probability maps at the start or middle of the network, they were inputted just before the final convolution. This method was explored by Hosseinzadeh *et al.* [38] for csPCa detection. This method is displayed in Fig. 1(c).

4) Auxiliary Loss: To aid the learning of relevant features in the encoder, we used an auxiliary loss term, where the auxiliary target was the two zonal maps. Similarly to the main decoder, a second decoder was connected to the bottleneck with a deconvolution, and skip connections were set from the encoder to this auxiliary decoder (See Fig. 1(d)). The loss function for the auxiliary target was the negative soft Dice score, as proposed by Milletari *et al.* [40]. The weights of the main and auxiliary loss to compute the total loss were 0.8 and 0.2, respectively.

G. Ensemble Learning

We investigated two ensemble techniques to reduce the effect of overfitting on the training set. In Section III-E, different combinations of these techniques are evaluated.

During training, we saved the 10 models with the lowest loss value on the validation set. During test-time, we used these models to make a prediction on the input image. The final prediction was then determined by averaging the output probability maps.

Furthermore, we used test-time augmentation. We augmented the input images during inference by applying the same types of transformations as for the training set, except for random elastic deformations. The inverse of the transformations were then applied to the predictions for each augmentation. The average of the outputs were then used to make the final prediction.

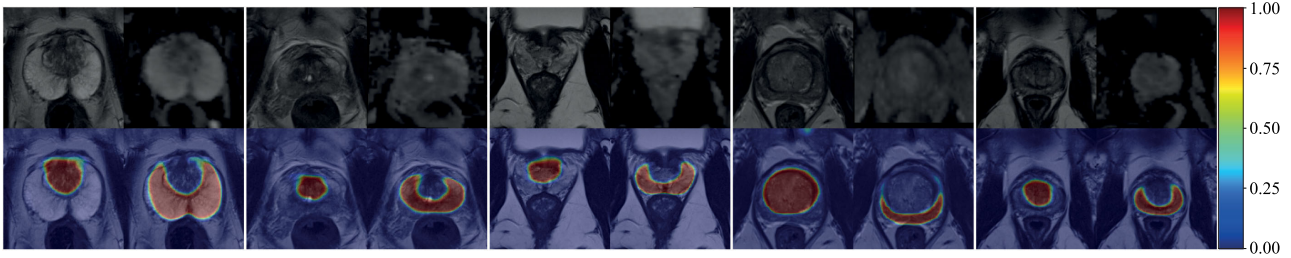


Fig. 3. Examples of the zonal segmentation network output. Each 2×2 block shows the middle slice of the MRI images of one patient. In each block, we show the T2-weighted image (top-left), ADC image (top-right), T2-weighted image overlaid with TZ prediction of the network (bottom-left), T2-weighted image overlaid with PZ prediction of the network (bottom-right). The predictions are the probabilistic network outputs.

H. Evaluation

As there were only a small amount of lesions per GGG in the dataset, simply using a portion of the images as validation set would lead to an unrepresentative validation set. Hence, we evaluated all models with 5-fold cross-validation.

1) Voxel-Wise Predictions: We used the quadratic-weighted kappa score κ_w as a metric for evaluation [41]. This metric has several convenient properties. Firstly, it adjusts for random agreement. This is especially useful in case of class imbalance. Furthermore, this metric penalizes a wrong prediction that is further off from the ground truth more than a prediction that is less far off. Hence, it takes into account that the classes are ordinal. If κ_w is 0, the agreement between the predictions and labels is equal to random chance agreement, κ_w is 1 if there is perfect agreement, and κ_w is -1 if the agreement is exactly opposite.

In this work, we evaluated using the voxel-wise quadratic-weighted kappa score $\kappa_{w,voxel}$ and we always calculated this metric over the entire dataset.

To compare with work that detects and segments clinically significant lesions ($GGG \geq 2$), we also calculated the Dice similarity coefficient (DSC) by considering voxels with prediction $GGG \geq 2$ as the foreground class and the other voxels as the background class.

2) Lesion-Wise Predictions: The task of the ProstateX-2 challenge was to grade lesions, given the image coordinate of that lesion. The described method, however, produces voxel-wise predictions. Hence, to compare with the ProstateX-2 challenge, we converted the voxel-wise to lesion-wise predictions. We computed a GGG prediction p for each lesion provided by the challenge with coordinate \vec{c} from a prediction mask M with a GGG for each voxel.

We thresholded M with $GGG \geq 2$, resulting in a binary mask T . We then considered the connected component in T at location \vec{c} . If there was no connected component in \vec{c} , p was set to 1. Otherwise, we computed the mean of all voxel-wise predictions in M in the selected connected component. p was then set to the rounded value of this statistical measure. We expected this approach to be more robust to noisy predictions than simply taking the voxel prediction at \vec{c} in M .

Once a single GGG prediction was determined for each lesion that the ProstateX-2 challenge proposed, we calculated the lesion-wise quadratic weighted kappa score $\kappa_{w,lesion}$. This metric was also used for ranking the participants in the ProstateX-2

challenge. The previously described $\kappa_{w,voxel}$ is both a metric for grading and localization, while DSC is only a metric for the segmentation quality of csPCa voxels. Unlike $\kappa_{w,voxel}$, $\kappa_{w,lesion}$ provides a measure for the grading performance on a lesion level, instead of on a voxel level.

We based the model selection on $\kappa_{w,voxel}$, as this metric both incorporates localization and grading performance. Furthermore, we observed much lower standard deviations for $\kappa_{w,voxel}$ than for $\kappa_{w,lesion}$, which reduces the probability that performance differences are based on chance.

III. EXPERIMENTS AND RESULTS

In this section, we investigate the influence of hyperparameters and model settings on the performance. We compare different model targets, model sizes by varying the number of U-Net layers, methods for incorporation of zonal information, numbers of model checkpoints used for an ensemble, and amounts of test-time augmentations. A chain of experiments is presented, i.e., except in the first experiment, the chosen best performing hyperparameters of the previous experiments were used. Section III-A to III-E each describe a separate experiment. Results of the experiments are plotted in Fig. 4.

A. Model Target

First, we trained a network for each model target. The performance metrics for each model target are plotted in Fig. 4(a). Soft-label ordinal regression received a higher $\kappa_{w,voxel}$ than the other two model targets. Furthermore, the two ordinal regression approaches outperformed multiclass classification. Soft-label ordinal regression outperformed multiclass classification with statistical significance ($p < 0.05$) in terms of $\kappa_{w,voxel}$. The $\kappa_{w,voxel}$ of multiclass classification, multi-label, and soft-label ordinal regression were 0.225 ± 0.114 , 0.348 ± 0.050 , and 0.391 ± 0.062 , respectively. The differences between multiclass classification and multi-label ordinal regression were not statistically significant. In the remainder of the experiments, soft-label ordinal regression was used as model target.

B. U-Net Layers

We compared different model sizes by varying the number of U-Net layers, which is defined as the unique number of feature map x , y -sizes in the network. Fig. 4(b) shows the performance of networks with 3, 4, 5, and 6 U-Net layers. The U-Net with

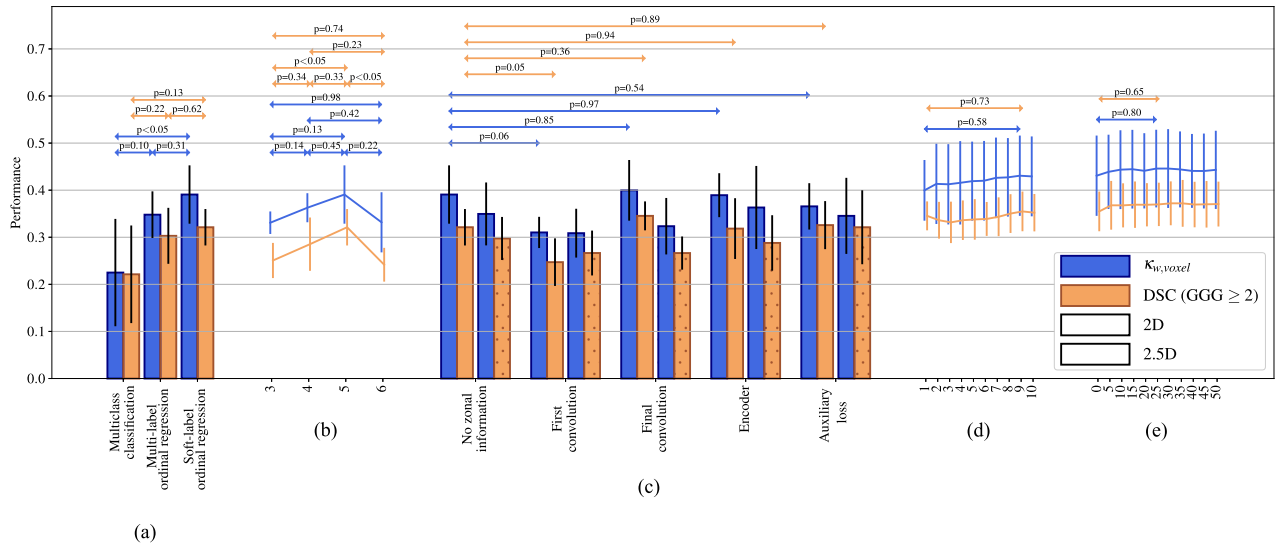


Fig. 4. Results of different experiments. We compare different model targets (a), model sizes by varying the number of U-Net layers (b), methods for the incorporation of zonal information (c), number of model checkpoints used for an ensemble (d), and numbers of test-time augmentations (e). In the zonal information experiment, all model were calculated using both 2D (bars without pattern) and 2.5D (bars with dotted pattern). The error bars are the standard deviations that were calculated from the different cross-validation folds. We determined statistical significance between the different models in each experiment. We calculated the p-values using a two-sided t-test. The p-values of the statistical tests are displayed above the two models between which statistical significance was calculated.

5 layers scored higher in terms of all reported metrics, so in the following experiments we used 5 U-Net layers. The 5 layer network scored a $\kappa_{w,voxel}$ of 0.391 ± 0.062 . None of the differences in $\kappa_{w,voxel}$ were statistically significant. In terms of DSC , however, the 5 layer network performed statistically better than the models with 3 and 6 layers.

C. Zonal Information

Next, we show the results of adding zonal information in Fig. 4(c). Since zonal segmentations could have been under-segmented in the slice direction, we also trained each network with a 2.5D variant. $\kappa_{w,voxel}$ increased from 0.391 ± 0.062 to 0.400 ± 0.064 when merging the zonal feature maps before the final convolution. Thus, this method for zonal information incorporation was used for the remaining experiments. None of the other methods of using zonal information improved the baseline. Moreover, none of the zonal information incorporation approaches underperformed or outperformed the baseline with statistical significance. $\kappa_{w,voxel}$ dropped in all cases when going from 2D to 2.5D.

D. Model Checkpoint Ensemble

Fig. 4(d) shows the effect of using multiple model checkpoints for making an ensemble. We evaluated 10 ensembles, where the i -th ensemble included the i best performing models on the validation set. Except for when going from 2 to 3 and going from 9 to 10 checkpoints, $\kappa_{w,voxel}$ increased. Optimal performance was reached when using 9 checkpoints for the ensemble. However, this was not statistically significantly different from only using the best model. $\kappa_{w,voxel}$ increased from 0.400 ± 0.064 to 0.431 ± 0.085 . Hence, in the next experiment, 9 model checkpoints were used.

E. Test-Time Augmentations

In Fig. 4(e), the effect of test-time augmentation is shown. We evaluated the performance for 0 to 50 augmentations, with a step size of 5. $\kappa_{w,voxel}$ increased most with 25 augmentations (from 0.431 ± 0.085 without augmentations to 0.446 ± 0.082). The largest performance gain in one step occurred when adding the first few augmentations, as the performance curves in Fig. 4(e) are steepest when going from 0 to 5 augmentations.

F. Results

The score for lesion-wise prediction of the best performing model in the previously described chain of experiments was $\kappa_{w,lesion} = 0.172 \pm 0.169$. The model from the cross-validation iteration with the highest validation performance was used for making predictions on the challenge test set, on which this method achieved a $\kappa_{w,lesion}$ of 0.13 ± 0.27 . This mean and standard deviation were calculated using bootstrapping with 1 k iterations.

Qualitative results of the model are shown in Fig. 5. The top row shows examples for which the model performed well, while the bottom row shows examples where the model predictions were not close to the manual delineations. Common false positives include hypo-intense tissues that do not correspond to clinically significant tissue according to the ground truth. Furthermore, the effect of image registration is demonstrated with several examples in Fig. 6.

IV. DISCUSSION

Soft-label ordinal regression performed statistically significantly better than multiclass classification. This could be explained by the fact that the former method uses the information

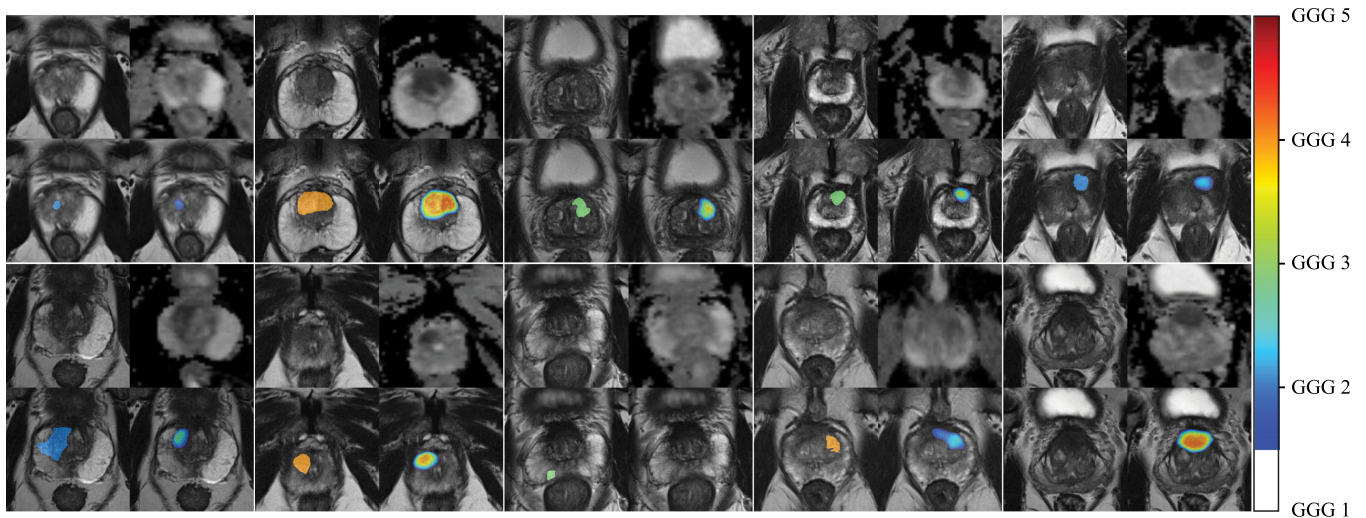


Fig. 5. Qualitative validation results of our approach. Each 2×2 shows an example of the same slice, where the top-right image is the ADC map and the other three are T2-weighted images. The bottom-left is overlaid with the ground truth and the bottom-right is overlaid with the model prediction. This model prediction is the output of the soft-label ordinal regression, which is between 0 (GGG 1) and 1 (GGG 5). The heatmap is transparent for voxels where the assigned class was the first class (healthy tissue or GGG 1). The top row shows five lesions with correct localization and grading. The bottom row shows examples where our method failed.

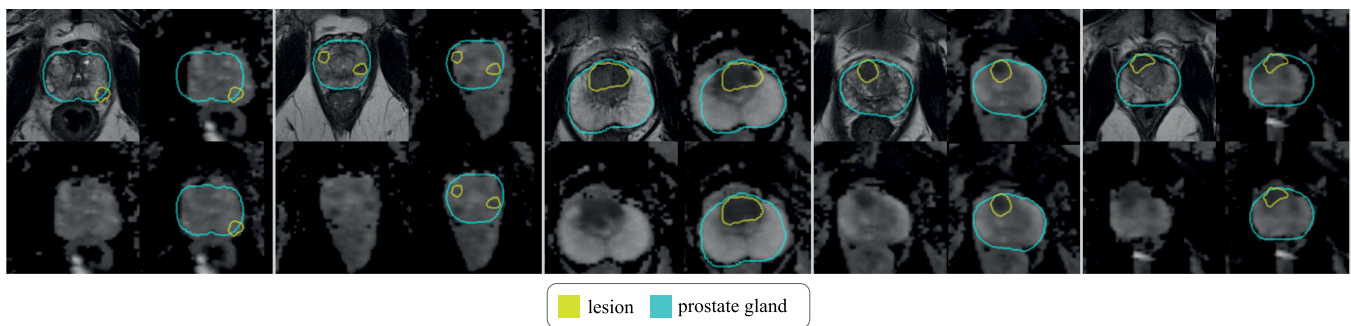


Fig. 6. Examples of registrations. Each 2×2 block shows a slice of one patient. The blocks show the T2-weighted image in the top-left, unregistered ADC in the top-right and bottom-left, and registered ADC in the bottom-right. To make it easier to observe the effect of the registration, three images are overlaid with prostate gland and lesion segmentations that are based on the T2-weighted image.

that the classes can be ordered and that not all incorrect predictions are equally wrong. No statistical difference was found between multi-label ordinal regression and soft-label ordinal regression. However, the latter did score a higher performance on average.

Furthermore, adding zonal information to the network at a late stage in the network slightly increased the performance, compared to omitting zonal information. During the same experiment, we observed that when using 2.5D instead of 2D input, performance consistently dropped. Apparently, there was no useful information for the network in adjacent slices. These slices likely only confused the network during training.

We also used an ensemble of 9 model checkpoints and test-time augmentation, which increased the performance. By evaluating the amount of augmentations during test-time, we observed that at a certain point adding more augmentations had no positive effect on the performance.

Performance scores of the approaches proposed by us and other challenge participants were relatively low in general, which illustrates that the task of grading cancer from MRI data is difficult. A reason for this could be the definition of GGG. This grading system is based on the most common and second-most common type of cancer tissue in the lesion. Hence, if the most common type of cancer has a low grade (e.g., 3), but the second-most common type has a much higher grade (e.g., 5), the Gleason score will be $3 + 5 = 8$, thus this lesion will be assigned to GGG 4. However, since most of the cancer tissue is low grade, the lesion will probably appear as unaggressive in an MRI scan. Other studies have shown that PCa grading can lead to large disagreement between raters. For example, a study of Ruprecht *et al.* [10] showed that the inter-rater agreement between radiologists was $\kappa = 0.0129$ (unweighted) for distinguishing two stages of PCa from MRI. Moreover, even when grading from tissue biopsies, pathologists scored an

unweighted κ for inter- and intra-agreement of 0.54 and 0.66, respectively [42]. These agreement scores form a target for a clinically desirable performance.

Other works that segmented csPCa from MRI are Kohl *et al.* [17], Artan *et al.* [43] and Chung *et al.* [44], who report a *DSC* of 0.41, 0.46, and 0.39, respectively. This is comparable to our approach that scored a *DSC* of 0.37 for this task. Our grading method with $\kappa_{w,lesion} = 0.13$ did not outperform the best scoring submission to the ProstateX-2 challenge with $\kappa_{w,lesion} = 0.27$ [23]. However, we did perform a more difficult and more clinically relevant task. After all, the goal of the ProstateX-2 was GGG classification, while our proposed method includes grading as well the detection of prostate lesions in bp-MRI.

In future work, ROI selection could be based on a prostate gland segmentation, instead of taking a fixed sized box in the middle of the image. This would lead to the prostate always being in the middle of ROI, which could aid the training process. Furthermore, a larger dataset is likely to improve the performance. Especially for high grade cancer, there is only a small amount of lesions in the training dataset. Moreover, in the current work, lesion segmentations are based on MRI. Thus, improvements could also be achieved with the usage of histopathology based delineations. Cao *et al.* [27] uses both a larger dataset and reference segmentations that are based on histopathology. However, that dataset is not publicly available, so we cannot directly evaluate our proposed method on their dataset.

Another way this approach could be improved is by splitting the task up into multiple networks. For example, a csPCa segmentation network could be developed, followed by a classification network that grades the segmented regions. We performed initial experiments using Mask R-CNN [45], which is an example of such a workflow. However, we have not achieved promising results using this architecture. This was most likely caused by the small amount of lesions in the training population, resulting in overfitting. Approaches such as those proposed by Abraham *et al.* [26], where not the image, but image features are used as input, are potentially less prone to overfitting on such a small amount of training samples.

V. CONCLUSION

In conclusion, we have shown that soft-label ordinal regression improves the performance of PCa grading and detection from bp-MRI over other methods. We have presented a comparison of methods for improving the model performance, including ensembling techniques and the use of zonal information. However, it remains an open question whether it is feasible to predict GGGs from MRI with the expert performance that is reached by pathologists when grading from histopathology images.

ACKNOWLEDGMENT

The authors would like to thank the ProstateX-2 challenge organizers for providing the data used in this research, and for calculating the test results of our submission.

REFERENCES

- [1] F. Bray *et al.*, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.
- [2] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2019," *CA: A Cancer J. Clinicians*, vol. 69, no. 1, pp. 7–34, 2019.
- [3] R. A. Smith *et al.*, "Cancer screening in the United States, 2018: A review of current American Cancer Society guidelines and current issues in cancer screening," *CA: A Cancer J. Clinicians*, vol. 68, no. 4, pp. 297–316, 2018.
- [4] N. Mottet *et al.*, "EAU-ESTRO-SIOG guidelines on prostate cancer. Part 1: Screening, diagnosis, and local treatment with curative intent," *Eur. Urol.*, vol. 71, no. 4, pp. 618–629, 2017.
- [5] W. J. Catalona *et al.*, "Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate cancer: Results of a multicenter clinical trial of 6,630 men," *J. Urol.*, vol. 151, no. 5, pp. 1283–1290, 1994.
- [6] R. Kvåle *et al.*, "Concordance between Gleason scores of needle biopsies and radical prostatectomy specimens: A population-based study," *BJU Int.*, vol. 103, no. 12, pp. 1647–1654, 2009.
- [7] R. T. Divrik *et al.*, "Increasing the number of biopsies increases the concordance of Gleason scores of needle biopsies and prostatectomy specimens," in *Urologic Oncology: Seminars and Original Investigations*, vol. 25. New York, NY, USA: Elsevier, 2007, pp. 376–382.
- [8] H. U. Ahmed *et al.*, "Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): A paired validating confirmatory study," *Lancet*, vol. 389, no. 10071, pp. 815–822, 2017.
- [9] H. K. Lim *et al.*, "Prostate cancer: Apparent diffusion coefficient map with T2-weighted images for detection—a multireader study," *Radiology*, vol. 250, no. 1, pp. 145–151, 2009.
- [10] O. Ruprecht *et al.*, "MRI of the prostate: Interobserver agreement compared with histopathologic outcome after radical prostatectomy," *Eur. J. Radiol.*, vol. 81, no. 3, pp. 456–460, 2012.
- [11] K. C. D. Thestrup *et al.*, "Biparametric versus multiparametric MRI in the diagnosis of prostate cancer," *Acta Radiologica Open*, vol. 5, no. 8, 2016, Art. no. 2058460116663046.
- [12] A. Stanzone *et al.*, "Biparametric 3 T Magentic Resonance Imaging for prostatic cancer detection in a biopsy-naïve patient population: A further improvement of PI-RADS v2?" *Eur. J. Radiol.*, vol. 85, no. 12, pp. 2269–2274, 2016.
- [13] C. K. Kuhl *et al.*, "Abbreviated biparametric prostate MR imaging in men with elevated prostate-specific antigen," *Radiology*, vol. 285, no. 2, pp. 493–505, 2017.
- [14] P. De Visschere *et al.*, "Dynamic contrast-enhanced imaging has limited added value over T2-weighted imaging and diffusion-weighted imaging when using PI-RADSv2 for diagnosis of clinically significant prostate cancer in patients with elevated PSA," *Clin. Radiol.*, vol. 72, no. 1, pp. 23–32, 2017.
- [15] J. I. Epstein *et al.*, "A contemporary prostate cancer grading system: A validated alternative to the Gleason score," *Eur. Urol.*, vol. 69, no. 3, pp. 428–435, 2016.
- [16] Y. K. Tsehay *et al.*, "Convolutional neural network based deep-learning architecture for prostate cancer detection on multiparametric magnetic resonance images," in *Proc. Med. Imag. Comput.-Aided Diagnosis*, 2017, vol. 10134, Art. no. 1013405.
- [17] S. Kohl *et al.*, "Adversarial networks for the detection of aggressive prostate cancer," 2017, *arXiv:1702.08014*.
- [18] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241.
- [19] S. Liu *et al.*, "Prostate cancer diagnosis using deep learning with 3D multiparametric MRI," in *Proc. Med. Imag. Comput.-Aided Diagnosis* 2017, vol. 10134, Art. no. 1013428.
- [20] Y. Yuan *et al.*, "Prostate cancer classification with multi-parametric MRI transfer learning model," *Med. Phys.*, vol. 46, pp. 756–765, 2019.
- [21] J. C. Seah, J. S. Tang, and A. Kitchen, "Detection of prostate cancer on multiparametric MRI," in *Proc. Med. Imag. Comput.-Aided Diagnosis*, 2017, vol. 10134, Art. no. 1013429.
- [22] A. Mehrtash *et al.*, "Classification of clinical significance of MRI prostate findings using 3d convolutional neural networks," in *Proc. Med. Imag. Comput.-Aided Diagnosis*, 2017, vol. 10134, Art. no. 101342A.
- [23] S. G. Armato *et al.*, "PROSTAT Ex challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images," *J. Med. Imag.*, vol. 5, no. 4, 2018, Art. no. 044501.

- [24] P. C. Vos *et al.*, "Computer-assisted analysis of peripheral zone prostate lesions using T2-weighted and dynamic contrast enhanced T1-weighted MRI," *Phys. Med. Biol.*, vol. 55, no. 6, pp. 1719–1734, 2010.
- [25] C. Jensen *et al.*, "Assessment of prostate cancer prognostic Gleason grade group using zonal-specific features extracted from biparametric MRI using a KNN classifier," *J. Appl. Clin. Med. Phys.*, vol. 20, pp. 146–153, 2019.
- [26] B. Abraham and M. S. Nair, "Computer-aided classification of prostate cancer grade groups from MRI images using texture features and stacked sparse autoencoder," *Computerized Med. Imag. Graph.*, vol. 69, pp. 60–68, 2018.
- [27] R. Cao *et al.*, "Joint prostate cancer detection and gleason score prediction in MP-MRI via FocalNet," *IEEE Trans. Med. Imag.*, vol. 38, no. 11, pp. 2496–2506, Nov. 2019.
- [28] C. de Vente *et al.*, "Simultaneous detection and grading of prostate cancer in multi-parametric MRI," *Med. Imag. Deep Learn.*, 2019.
- [29] J. I. Epstein *et al.*, "The 2014 International Society of Urological Pathology (ISUP) consensus conference on Gleason grading of prostatic carcinoma," *Amer. J. Surgical Pathol.*, vol. 40, no. 2, pp. 244–252, 2016.
- [30] B. C. Lowekamp *et al.*, "The design of SimpleITK," *Frontiers Neuroinform.*, vol. 7, pp. 45–58, 2013.
- [31] Z. Niu *et al.*, "Ordinal regression with multiple output CNN for age estimation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4920–4928.
- [32] J. Cheng *et al.*, "A neural network approach to ordinal regression," in *Proc. IEEE Int. Joint Conf. Neural Netw. World Congr. Comput. Intell.*, 2008, pp. 1279–1284.
- [33] F. Chollet *et al.*, "Keras," 2015. [Online]. Available: <https://keras.io>
- [34] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. software available from tensorflow.org. [Online]. Available: <http://tensorflow.org/>
- [35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. on Learning Representations*, 2015.
- [36] J. E. McNeal *et al.*, "Zonal distribution of prostatic adenocarcinoma. Correlation with histologic pattern and direction of spread," *Amer. J. Surgical Pathol.*, vol. 12, no. 12, pp. 897–906, 1988.
- [37] J. C. Weinreb *et al.*, "PI-RADS prostate imaging—reporting and data system: 2015, version 2," *Eur. Urol.*, vol. 69, no. 1, pp. 16–40, 2016.
- [38] M. Hosseinzadeh, P. Brand, and H. Huisman, "Effect of adding probabilistic zonal prior in deep learning-based prostate cancer detection," *Med. Imag. Deep Learn.*, 2019.
- [39] G. Mooij, I. Bagulho, and H. Huisman, "Automatic segmentation of prostate zones," 2018, *arXiv:1806.07146*.
- [40] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. IEEE 4th Int. Conf. 3D Vision*, 2016, pp. 565–571.
- [41] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit," *Psychological Bull.*, vol. 70, no. 4, pp. 213–220, 1968.
- [42] J. Melia *et al.*, "A UK-based investigation of inter-and intra-observer reproducibility of Gleason grading of prostatic biopsies," *Histopathology*, vol. 48, no. 6, pp. 644–654, 2006.
- [43] Y. Artan *et al.*, "Prostate cancer localization with multispectral MRI using cost-sensitive support vector machines and conditional random fields," *IEEE Trans. Image Process.*, vol. 19, no. 9, pp. 2444–2455, Sep. 2010.
- [44] A. G. Chung *et al.*, "Prostate cancer detection via a quantitative radiomics-driven conditional random field framework," *IEEE Access*, vol. 3, pp. 2531–2541, 2015.
- [45] K. He *et al.*, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2961–2969.