European Radiology (2020) 30:1243–1253
https://doi.org/10.1007/s00330-019-06417-z

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network

Nader Aldoj[1] · Steffen Lukas[1] · Marc Dewey[1,2] · Tobias Penzkofer[1,2]

Received: 18 March 2019 / Revised: 13 July 2019 / Accepted: 8 August 2019 / Published online: 29 August 2019

Abstract

Objective To present a deep learning–based approach for semi-automatic prostate cancer classification based on multi-parametric magnetic resonance (MR) imaging using a 3D convolutional neural network (CNN).

Methods Two hundred patients with a total of 318 lesions for which histological correlation was available were analyzed. A novel CNN was designed, trained, and validated using different combinations of distinct MRI sequences as input (e.g., T2-weighted, apparent diffusion coefficient (ADC), diffusion-weighted images, and *K*-trans) and the effect of different sequences on the network's performance was tested and discussed. The particular choice of modeling approach was justified by testing all relevant data combinations. The model was trained and validated using eightfold cross-validation.

Results In terms of detection of significant prostate cancer defined by biopsy results as the reference standard, the 3D CNN achieved an area under the curve (AUC) of the receiver operating characteristics ranging from 0.89 (88.6% and 90.0% for sensitivity and specificity respectively) to 0.91 (81.2% and 90.5% for sensitivity and specificity respectively) with an average AUC of 0.897 for the ADC, DWI, and *K*-trans input combination. The other combinations scored less in terms of overall performance and average AUC, where the difference in performance was significant with a *p* value of 0.02 when using T2w and *K*-trans; and 0.00025 when using T2w, ADC, and DWI. Prostate cancer classification performance is thus comparable to that reported for experienced radiologists using the prostate imaging reporting and data system (PI-RADS). Lesion size and largest diameter had no effect on the network's performance.

Conclusion The diagnostic performance of the 3D CNN in detecting clinically significant prostate cancer is characterized by a good AUC and sensitivity and high specificity.

Key Points

- *Prostate cancer classification using a deep learning model is feasible and it allows direct processing of MR sequences without prior lesion segmentation.*
- *Prostate cancer classification performance as measured by AUC is comparable to that of an experienced radiologist.*
- *Perfusion MR images (K-trans), followed by DWI and ADC, have the highest effect on the overall performance; whereas T2w images show hardly any improvement.*

Marc Dewey and Tobias Penzkofer contributed equally to this work.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-019-06417-z) contains supplementary material, which is available to authorized users.

✉ Nader Aldoj
nader.aldoj@charite.de

✉ Marc Dewey
dewey@charite.de

[1] Department of Radiology, Charité – Universitätsmedizin Berlin, Freie Universität Berlin, Humboldt-Universität zu Berlin, Charitéplatz 1, 10117 Berlin, Germany

[2] Berlin Institute of Health (BIH), Anna-Louisa-Karsch-Str. 2, 10178 Berlin, Germany

Keywords Three-dimensional images · Prostate cancer · Multi-parametric MRI · Convolutional neural networks · Deep learning

Abbreviations

ADC	Apparent diffusion coefficient
AUC	Area under the curve
CNN	Convolutional neural network
DCE	Dynamic contrast-enhanced
DWI	Diffusion-weighted imaging
Mp-MRI	Multi-parametric magnetic resonance imaging
MR	Magnetic resonance
PCa	Prostate cancer
PI-RADS	Prostate imaging reporting and data system
PSA	Prostate-specific antigen
ROC	Receiver operating characteristics
T2w	T2-weighted
TRUS	Transrectal ultrasound

Introduction

Prostate cancer (PCa) is the most prevalent type of cancer in men and it has been reported to be the second leading cause of cancer death in men [1–3]. Prostate-specific antigen (PSA) and transrectal ultrasound (TRUS) continue to be the most widely used diagnostic modalities in clinical practice. However, PSA and TRUS have relatively low specificity, giving rise to overdiagnosis and overtreatment [4]. Recently, multi-parametric magnetic resonance imaging (mp-MRI) combining a T2-weighted (T2w), diffusion-weighted (DWI), dynamic contrast-enhanced (DCE) imaging was introduced and established as a more accurate, noninvasive alternative for PCa detection and characterization [5–7]. However, prostate mp-MRI reporting requires extensive training and can lack the necessary sensitivity and specificity [8]. Computer-aided MRI assessment might significantly improve diagnostic capabilities for the detection of PCa.

Several studies have shown that computer-aided systems have a role in PCa detection and diagnostic evaluation [5–10]. The methods proposed so far are based on handcrafted features, using a classifier on top of those to determine whether a PCa lesion is present and to assess its severity by assigning a specific class label to it. Recently, deep convolutional neural networks (CNNs) have shown an impressive performance in various computer vision tasks following training with large image databases [11–14]. In medical applications, training data are limited, and images acquired with diagnostic imaging modalities are very different from other types of images in terms of characteristics and handling procedures. Additional challenges arise from the use of different scanners, imaging protocols, noise levels, and other factors related to image acquisition [15, 16]. Minh Hung Le et al compared three well-known pre-trained algorithms [11–13] on their multi-modal datasets (T2w images, ADC) and combined the features obtained in this way with handcrafted features. Their results showed an improved accuracy compared with some older models [18]. Yang et al explored a co-trained fine-tuned inception-like network [19], showing that the fine-tuned model performed better than the one trained from scratch. All of these studies either analyzed the lesion region only (which required prior manual lesion segmentation) or used 2D images with selection of slices showing the lesion. Recently, Mehrtash et al have demonstrated the feasibility of using 3D images as input for a 3D convolutional neural network. With their approach, which involved feeding a 3D volume centered on the lesion into the network, they achieved an AUC comparable to that of a human reader using PI-RADS (prostate imaging reporting and data system) v1 and v2 [20].

The aim of the study was to investigate whether prostate cancer classification using deep learning techniques is feasible in an initial clinical evaluation and provides additional value for the clinical workflow. We proposed a semi-automated algorithm for computer-aided diagnosis to assess clinical significance of cancerous lesions from a 3D (processed multi-2D slices) multi-modal image merely requiring the location of the lesion center. No other information is necessary. We used a 3D CNN architecture with multi-modal input comprising of different combinations of sequences. The only lesion-related parameter that needs to be known in our approach is the lesion center, while no further information or any kind of automatic, semi-automatic, or manual segmentation is required. In this study, we compared the performance of the designed network on all possible combinations of sequences, but mainly: group 1: T2w, ADC, DWI, and *K*-trans; group 2: ADC, DWI, and *K*-trans; group 3: T2w and *K*-trans; and group 4: T2w, ADC, and DWI, and discussed the effect of each sequence to the overall network's performance. The *b* value for the used DWI was 800. The classification performance of our network was evaluated by determining the area under the curve (AUC) of the receiver operating characteristics (ROC) [21] and compared with that of an experienced radiologist reported in [17].

Materials and methods

Patients

An imaging dataset of 200 patients with a total number of 318 prostate lesions (243 clinically nonsignificant and 75 significant) was used, among them 175 patients for training and 25 patients for testing our algorithm. Zonal analysis can be seen in Fig. 1, where it shows that the lesions in the peripheral zone have the highest occurrence among all the zones, followed by

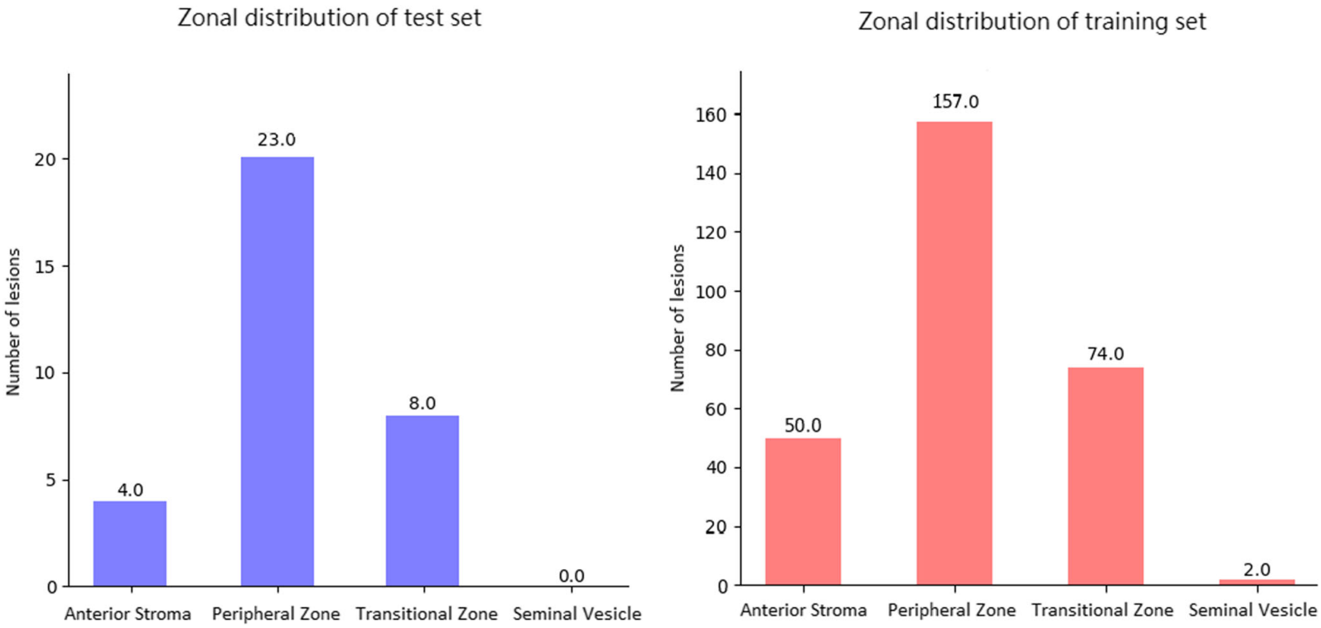

Fig. 1 Zonal distribution of lesions in the training and test sets. It shows that both sets have roughly the same zonal distribution which guarantees that the network's results are not biased

the transitional, anterior stroma and seminal vesicle zone respectively in both training and test sets.

The imaging database used in this study is publicly available and was provided for the SPIE-AAPM-NCI PROSTATEx challenge [22–24]. It comprises images acquired in 204 patients using the following pulse sequences: T2-weighted, apparent diffusion coefficient (ADC), diffusion-weighted (DW), and *K*-trans images that are obtained using dynamic contrast-enhanced (DCE) MR perfusion (see the electronic supplementary material (ESM) for more details about *K*-trans coefficients). All images are provided along with lesion coordinates and biopsy-based histopathological results (clinically significant vs. nonsignificant lesion), serving as the reference standard for the classification task. The significance level was based on the Gleason score where the lesion is considered significant if the Gleason score is 7 or higher and nonsignificant otherwise. Four patients were excluded because of low-resolution imaging.

MR imaging protocol

Prostate MRI examinations were performed on one of two 3T MR Siemens scanners (MAGNETOM Trio and Skyra) using a protocol including a T2-weighted turbo spin echo sequence with a 0.5-mm in-plane resolution and 3.6-mm slice thickness, and a single-shot planar imaging sequence for DWI with a 2-mm in-plane resolution, 3.6-mm slice thickness, and diffusion-encoding gradients in three directions. ADC maps were calculated from three *b* values (50, 400, and 800) using the scanner software. More details about the dataset can be found in [22–24].

Problem formulation

The differentiation of clinically significant and nonsignificant lesions is a binary classification task. Specifically, input is a 3D volume comprising the lesion and surrounding tissues, and output is a label of a value either 0 or 1. Let us name the input X and the output o. For a single input from the dataset, we can formulate the loss function as follows in order to minimize the deviation between the predicted value and the real label:

$$F(X,o) = -o\log p(O=1|X) - (l-o)\log p(O=0|X) \qquad (1)$$

where $p(O=i|X)$ represents the probability predicted by the network prediction.

Image pre-processing

The pre-processing step mainly focuses on preparing the images in the dataset before they can be used as input for the neural network. Since the images in the original dataset were of varying voxel size, images were resampled to a common voxel size of 0.5, 0.5, and 3.0 mm in x, y, and z directions respectively using bilinear interpolation. All images acquired with different sequences were spatially aligned using rigid registration. A spherical cropping window with a radius of 20 mm centered on each lesion coordinates was applied. The resulting spherical volume was used in the subsequent steps. Rounding errors during the cropping step were accounted for and lesions bordering to the edge of the volume were padded with black slices.

To limit overfitting (see ESM), images were augmented. The first augmentation consisted of shifting the lesion center by the following vectors (x, y): (0, 7), (0, − 7), (7, 0), (− 7, 0), and (10, 10) mm before reapplying the cropping step. After this step, we had 6 times more images than just taking the lesion at the center of the sphere. Additionally, 6 types of image-based augmentation including image brightening, darkening, adding noise, flipping around x and y, and several rotations in both positive and negative x directions were performed. It is worth mentioning that some of the augmentations were held out for the images with nonsignificant lesion to ensure class balance. Overall, the augmentation steps increased the number of available images from 318 to around 12,0000 images. Further steps included on-the-fly image normalization to [0;1] for the three pulse sequences.

The proposed network architecture

The developed CNN (convolutional neural network) takes different combinations of 3D volumes (e.g., ADC, DWI, and T2w) as input and considers each sequence a separate input channel; the output is the classification of significant vs. nonsignificant lesion. The network comprises 12 convolutional layers arranged in seven blocks with skip connections, two fully connected layers, and one output layer (see ESM for a detailed description of the network substructure). All codes for image processing were done using Python 2.7 and SimpleITK libraries; the neural network was implemented using the TensorFlow library (Version 1.4.0, Google) and run on a TitanXP GPU (Nvidia). The training time of the network for a single model was around 3.2 h. The computation time for a single image was 0.26 s during training and 0.1 during testing.

Network training and statistical analysis

For network training, the AdamOptimizer [25] was employed to minimize loss with a learning rate of 1e−5, a mini-batch of size 50, and a binary cross-entropy loss function. All weights were randomly initialized using truncated normal function with a standard deviation of 0.1. Batch normalization was used after each convolutional layer to speed up convergence. In order to avoid overfitting, a dropout with 0.5 probability in both fully connected layers as well as L2 norm regularization of a β-value of 1e−4 on neuron weights was applied. Early stopping was performed by observing the validation performance and choosing the best model with the highest area under the AUC value.

$$L^{'} = L + \beta\frac{1}{2}\|W\|^2 \tag{2}$$

where $L^{'}$ is the regularized loss, β is the penalization term, and W is the weight matrix.

Eightfold cross-validation was performed, to train and test the model on all possible combinations of the dataset images

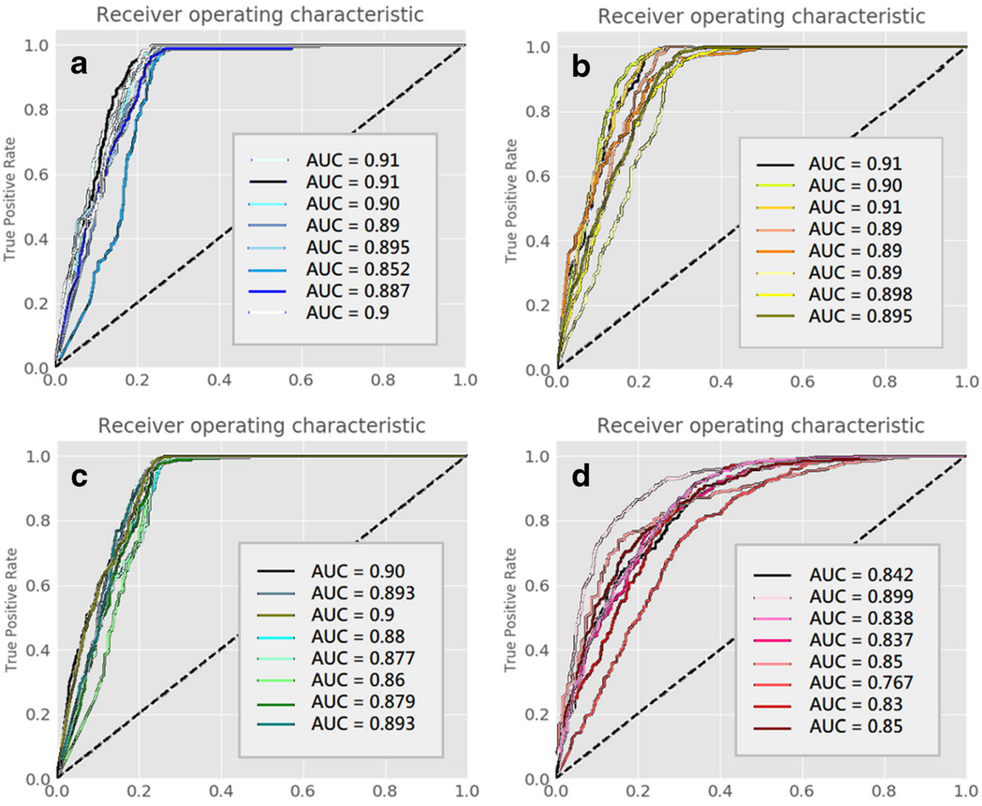

Fig. 2 ROC curves. Performances of the eight different models of the eightfold cross-validation (colors) and the corresponding areas under the ROC curves. Each letter represents the network's performance of different sequence combinations: **a** T2w, ADC, DWI, and *K*-trans; **b** ADC, DWI, and K-trans; **c** T2 and *K*-trans; **d** T2, ADC, and DWI

Table 1 Diagnostic performance of all input groups (each with eight cross-validation models) with 95% confidence intervals (CI)

Model number	AUC	CI (95)	Sensitivity (%)	± CI 95%	Specificity (%)	± CI 95%
Group 1: T2w + ADC + DWI + *K*-trans						
1	0.91	0.898–0.93	82.0	± 2.5	90.0	± 2.1
2	0.91	0.897–0.926	76.7	± 3.7	96.0	± 2.0
3	0.90	0.889–0.914	71.1	± 2.3	97.7	± 2.2
4	0.89	0.873–0.909	77.2	± 2.8	92.4	± 2.2
5	0.895	0.872–0.918	87.2	± 2.7	81.6	± 3.3
6	0.852	0.826–0.878	82.5	± 2.6	89.5	± 2.8
7	0.887	0.873–0.903	61.5	± 3.2	97.2	± 1.9
8	0.9	0.891–0.92	65.5	± 3.5	97.0	± 1.8
Average AUC	0.893		75.4		92.6	
Standard deviation	0.018		0.08		0.05	
Group 2: ADC + DWI + *K*-trans						
1	0.91	0.890–0.923	81.2	± 2.9	90.5	± 3.2
2	0.90	0.898–0.920	75.8	± 2.5	9.1	± 1.7
3	0.91	0.899–0.923	71.2	± 3.2	97.1	± 2.3
4	0.89	0.873–0.908	77.2	± 3.6	92.00	± 2.8
5	0.89	0.874–0.916	87.60	± 2.2	80.00	± 2.1
6	0.89	0.874–0.916	87.4	± 2.5	79.3	± 2.9
7	0.898	0.878–0.918	87.6	± 2.4	79.7	± 4.2
8	0.895	0.874–0.916	87.6	± 2.3	79.1	± 2.5
Average AUC	0.897		81.9		86.1	
Standard deviation	0.008		0.06		0.07	
Group 3: T2w + *K*-trans						
1	0.90	0.886–0.92	80.5	± 2.5	89.5	± 2.2
2	0.893	0.877–0.91	74.4	± 3.3	89.3	± 2.5
3	0.90	0.886–0.92	80.4	± 3.3	89.9	± 1.2
4	0.88	0.866–0.902	76.0	± 3.7	95.0	± 2.3
5	0.877	0.851–0.902	87.7	± 2.2	84.4	± 2.5
6	0.86	0.833–0.885	82.9	± 3.0	93.7	± 2.4
7	0.879	0.865–0.895	60.7	± 3.6	95.9	± 2.2
8	0.893	0.877–0.909	63.0	± 2.9	97.5	± 1.7
Average AUC	0.885		75.7		91.9	
Standard deviation	0.013		0.09		0.04	
Group 4: T2w + ADC + DWI						
1	0.842	0.823–0.863	72.7	± 2.8	79.5	± 2.9
2	0.899	0.884–0.916	71.9	± 2.8	91.5	± 2.5
3	0.838	0.818–0.859	73.0	± 3.2	80.4	± 2.5
4	0.837	0.817–0.859	72.9	± 3.0	80.8	± 2.2
5	0.85	0.827–0.872	86.0	± 3.1	66.7	± 3.9
6	0.767	0.739–0.797	73.7	± 3.5	72.1	± 3.5
7	0.83	0.815–0.852	52.1	± 4.3	92.5	± 2.5
8	0.85	0.832–0.872	55.8	± 3.9	90.2	± 2.3
Average AUC	0.839		69.7		81.7	
Standard deviation	0.036		0.10		0.09	
Group 5: DWI + *K*-trans						
1	0.90	0.885–0.915	81.8	± 2.8	90.2	± 3
2	0.90	0.888–0.915	84.3	± 2.9	97.0	± 2
3	0.87	0.866–0.891	67.6	± 2.9	96.7	± 1.9
4	0.86	0.848–0.883	77.0	± 3	94.2	± 2.2
5	0.86	0.84–0.883	87.8	± 2.5	75.7	± 5
6	0.84	0.818–0.864	84.5	± 3	86.8	± 4.2
7	0.867	0.853–0.882	61.1	± 3.5	93.0	± 1.8
8	0.877	0.863–0.892	63.8	± 3.6	96.6	± 1.5
Average AUC	0.87		76		91.2	
Standard deviation	0.02		0.10		0.07	

Table 1 (continued)

Model number	AUC	CI (95)	Sensitivity (%)	± CI 95%	Specificity (%)	± CI 95%
Group 6: ADC + *K*-trans						
1	0.90	0.885–0.915	85.0	± 3	82.0	± 2
2	0.88	0.861–0.889	74	± 2.5	93.4	± 1.8
3	0.87	0.866–0.891	67.1	± 3.8	93.8	± 2.4
4	0.853	0.841–0.863	75.6	± 3	90.0	± 2.2
5	0.85	0.838–0.878	87.5	± 2.2	75.3	± 2
6	0.83	0.82–0.852	56.3	± 3.8	90.7	± 2.3
7	0.833	0.812–0.854	56.8	± 3.4	91.2	± 2.4
8	0.87	0.860–0.891	59.8	± 3.9	95.4	± 2
Average AUC	0.86		70.2		88.9	
Standard deviation	0.02		0.12		0.07	
Group 7: ADC + DWI						
1	0.816	0.795–0.839	78.9	± 2.7	73.7	± 3.5
2	0.822	0.81–0.833	74.2	± 3.1	88.6	± 2.3
3	0.80	0.789–0.813	59.9	± 3	85.6	± 1.3
4	0.77	0.758–0.79	67.7	± 4.1	74.3	± 2.6
5	0.754	0.74–0.769	82.9	± 2.6	52.5	± 3.9
6	0.83	0.806–0.855	83.4	± 3.1	79.7	± 3.5
7	0.73	0.718–0.74	77.1	± 3.5	57.0	± 3.2
8	0.75	0.739–0.762	81.8	± 2.7	52.0	± 4.4
Average AUC	0.784		74.8		70.4	
Standard deviation	0.03		0.07		0.14	
Group 8: T2						
1	0.70	0.684–0.731	67.7	± 3.2	60.8	± 3.7
2	0.82	0.811–0.847	72.7	± 2.8	80.0	± 3.7
3	0.72	0.704–0.743	55.7	± 3.4	77.3	± 2.9
4	0.726	0.701–0.749	66.7	± 3.2	68.8	± 2.6
5	0.716	0.688–0.744	81.3	± 3.4	47.1	± 4.4
6	0.71	0.684–0.736	74.7	± 3.4	56.6	± 3.8
7	0.715	0.690–0.736	45.9	± 4.2	80.1	± 2.6
8	0.77	0.750–0.791	52.0	± 3.3	82.9	± 2.1
Average AUC	0.734		64.58		69.2	
Standard deviation	0.04		0.12		0.13	
Group 9: ADC						
1	0.77	0.757–0.8	76.5	± 3	65.2	± 3.2
2	0.786	0.773–0.812	69.6	± 2.5	74.4	± 1.8
3	0.727	0.706–0.744	53.1	± 3.8	77.8	± 3.1
4	0.703	0.679–0.728	62.4	± 3	67.3	± 3.2
5	0.75	0.722–0.774	83.0	± 2.2	53.4	± 3.4
6	0.70	0.675–0.728	75.4	± 3.2	54.8	± 4.0
7	0.683	0.661–0.708	44.5	± 4.4	77.6	± 2.9
8	0.77	0.749–0.793	53.4	± 3.9	84.6	± 2
Average AUC	0.736		64.7		69.4	
Standard deviation	0.03		0.13		0.11	
Group 10: DWI						
1	0.808	0.789–0.828	76.5	± 2.7	72.2	± 3.2
2	0.817	0.800–0.836	71.1	± 3.1	81.8	± 2.3
3	0.815	0.799–0.831	60.4	± 3	87.3	± 1.3
4	0.80	0.787–0.827	72.3	± 3.1	76.0	± 2.6
5	0.795	0.769–0.818	87.8	± 2.6	56.4	± 4.3
6	0.742	0.717–0.767	76.9	± 3.1	60.7	± 3.5
7	0.807	0.789–0.825	57.5	± 3.5	85.9	± 2
8	0.81	0.797–0.830	60.1	± 2.7	87.0	± 1.4
Average AUC	0.79		70.3		75.9	
Standard deviation	0.02		0.10		0.12	

Table 1 (continued)

Model number	AUC	CI (95)	Sensitivity (%)	± CI 95%	Specificity (%)	± CI 95%
Group 11: *K*-trans						
1	0.82	0.799–0.838	76.6	± 2.7	74.4	± 3.1
2	0.80	0.782–0.821	63.8	± 3.1	82.7	± 2.2
3	0.80	0.780–0.814	51.8	± 4	87.6	± 1.5
4	0.80	0.778–0.818	70.3	± 3.1	79.6	± 2.6
5	0.81	0.779–0.83	84.1	± 2.1	66.4	± 3.3
6	0.793	0.768–0.82	80.2	± 3.1	73.1	± 3.5
7	0.79	0.765–0.808	49.5	± 4.1	87.7	± 1.4
8	0.77	0.75–0.79	52.1	± 3.7	72.0	± 2.4
Average AUC	0.798		66.0		77.9	
Standard deviation	0.02		0.13		0.07	

without any overlap between the training and test sets. Cross-validation was also done in a way that the distribution between the two classes in the training and test sets was almost the same, which, in turn, prevents any bias in accuracy. The classification task was assessed by plotting the receiver operating curve (ROC) and calculating the corresponding sensitivity and specificity.

Results

This study presents a network designed to classify MR images of prostate lesions into clinically significant vs. clinically nonsignificant ones. The goal is identification of optimal network design and choice of MR sequences for optimal diagnostic performance of four different input combinations tested to investigate the role of each MR sequence on the overall performance and which input combination performs better. In our eightfold cross-validation, the results show that the network performed the best with the input combination group 2 which resulted an average AUC of 0.897 and 81.9% and 86.1% for sensitivity and specificity respectively which is comparable to that achieved by an experienced radiologist using PI-RADS v2 [17] and higher in comparison to [20] which tackled the same problem on the same dataset and addressed the same hypothesis. Group 1 has an average AUC of 0.893, an average sensitivity of 75.4%, and an average specificity of 92.6%. Groups 3 and 4 have lower average AUC values of 0.885 and 0.839 respectively. The difference in the performance between groups 1 and 2 was not statistically significant with a p value of 0.25. However, p values of 0.02 and 0.00025 were obtained between groups 2 and 3 and groups 2 and 4 respectively which proves the statistical significance in the performance between the aforementioned groups (see details in Fig. 2 and Table 1).

For further testing and analysis, the best model from group 2 was used (all performance's curves and images are the result of this model unless otherwise specified).The best model (from group 2) had an AUC of 0.91 at 81.2% sensitivity and 90.5% specificity. The results are presented in Fig. 2 and compared with the reported results of an experienced radiologists using PI-RADS v2 [17] and a similar study on the same dataset [20, 26].

Figure 3 shows four different test cases along with the statistical category. The lesion volumes in the test set can be seen in Fig. 4. Details of the results achieved with each of the cross-validation models are compiled in Table 1. Figure 5 shows the training and testing accuracy of the network along the iteration time together with the prediction accuracy according to the types of lesions, which gives an intuition of the number of correctly vs. incorrectly predicted lesions. Examples of some of the incorrectly predicated cases can be seen in Fig. 6.

Discussion

In this study, we present a newly developed 3D CNN architecture for the analysis of mp-MR images. The tool analyzes prostate lesions, classifying them into clinically significant vs. nonsignificant cancers. We trained and tested this tool on a high-quality dataset of mp-MRI examinations of the prostate. For all cases included, a biopsy-based diagnosis was available.

Group 2 showed the best performance's results in term of average AUC, sensitivity, and specificity. As can been seen from the aforementioned results, different input combinations affected the network's performance. T2w can be seen as the least MR sequence to introduce any significant improvement (see the difference between groups 1 and 2 in Table 1) on the performance (p value is 0.25 and the difference is not significant at $p < 0.05$). In contrast, we can see that *K*-trans is the highest MR sequence which affects the performance and drove the results to be more accurate (see the difference between groups 2 and 4 in Table 1) where p value is 0.00025 and the difference is significant at $p < 0.05$. The effect of using DWI images (ADC and b value) is also significant when comparing groups 2 and 3 where p value is 0.02. These

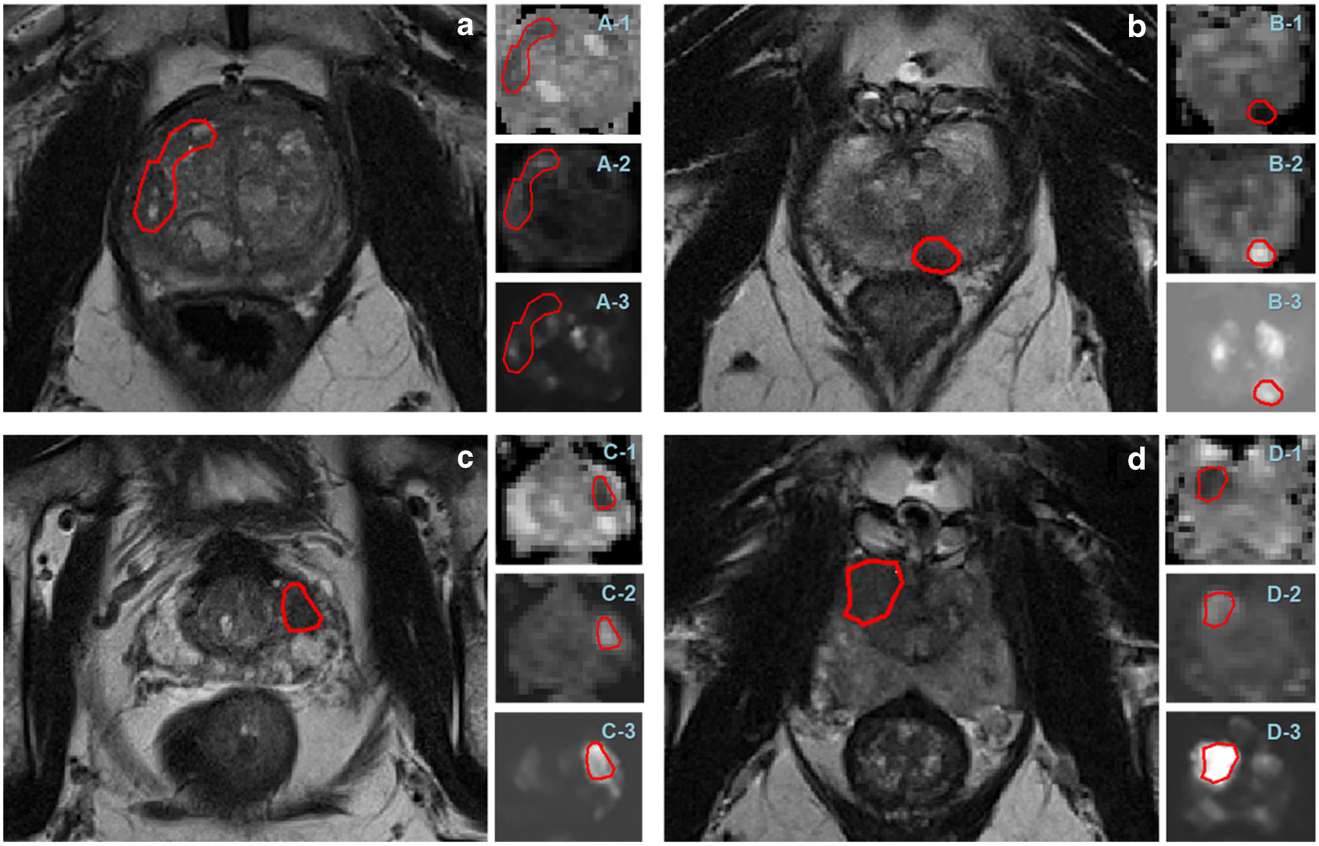

Fig. 3 Multi-parametric MR images for all diagnostic categories. **a**–**d** Example images for true positive (TP, **a**), false positive (FP, **b**), false negative (FN, **c**), and true negative (TN, **d**), with the corresponding T2w (1), ADC (2), and DWI images. The red line highlights the extent of the lesion. T2w is used for visualization purposes

results are also supported by more detailed testing of further combinations of MRI sequences and standalone sequences (see Table 1).

Most earlier approaches proposed to classify prostate lesions used 2D CNNs with segmentations of the suspicious lesion [18, 19]. Lui et al [26] used a 3D network to classify significant prostate lesions. However, such approaches require time-consuming preparation, which may preclude routine clinical use. Therefore, a 3D approach requiring little or no manual pre-processing is desirable. Mehrtash et al [20] tackled

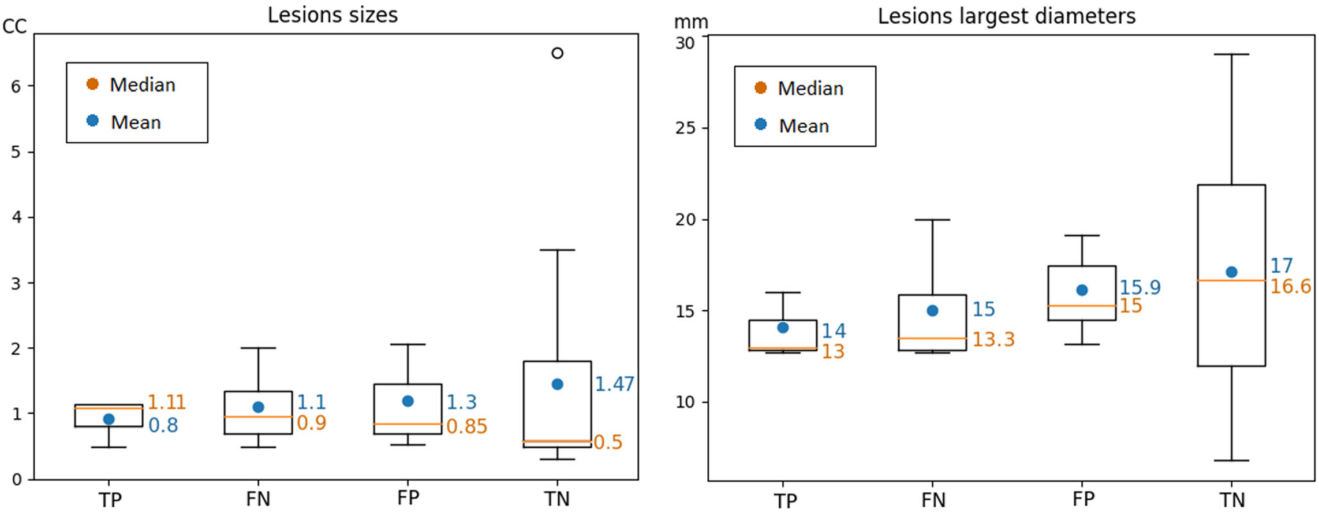

Fig. 4 Largest diameters and sizes of the lesions. Boxplot on the left side shows the volumes of the lesions and boxplot on the right side shows the largest diameters in the four diagnostic categories. Numbers with orange color indicate the mean value while the numbers with blue represent the median value

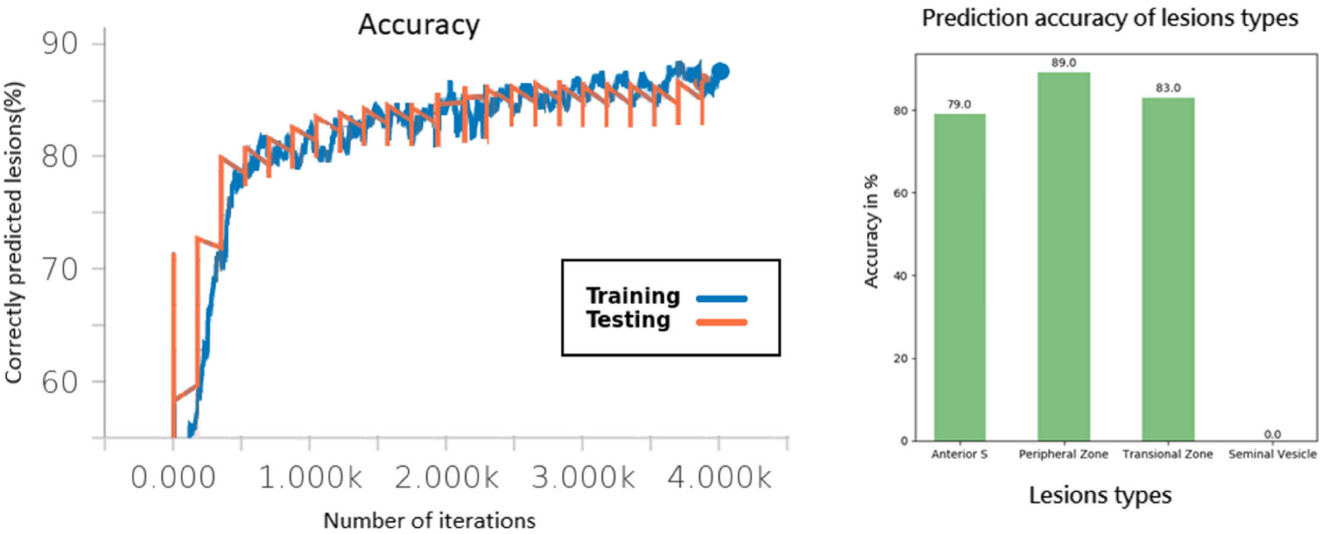

Fig. 5 Network's accuracy. Left side, the percentage of correctly predicted lesions during training (blue curve) and testing (orange curve). Right side, the prediction accuracy of different types of lesions

the problem in a similar fashion as we did and achieved an AUC of 0.80 using a 3D CNN with 3 parallel pipelines. Our method, in contrast, is a 3D-based single pipeline network and requires the least amount of preparation, which includes only resampling all pulse sequences used as input and lesion location(s), and the values are higher to those achieved in [20, 26] and an experienced human reader reported in [17], in terms of AUC, sensitivity (was higher in [26]), and specificity.

An ideal model for medical diagnosis should have a high AUC, sensitivity, and specificity. However, sensitivity is more relevant than specificity as the overarching aim always is to have a low false negative rate. An experienced human radiologist using PI-RADS v2 is reported to achieve an AUC of 0.83 with 77% sensitivity and 81% specificity [17]. While Mehrtash et al [20] reported a value of 0.80, and Lui et al [26] reported 0.84 for AUC, our model (model 1 from group 2) achieved a higher AUC value than [17, 20, 26] and was less sensitive and more specific than [26]. However, the model at this stage of performance cannot replace the radiologist; yet it can help the radiologist and facilitate the decision-making process. The model's superior specificity can potentially augment the radiologist's specificity, so that the overall combined performance could improve diagnostic accuracy in patients.

It is very important for the radiologist to have a reliable model prediction that can augment his or her decision, ultimately resulting in a better diagnosis. However, along with the prediction, radiologists would be interested in how the model has arrived at the prediction and on which structure in the mp-MR image the network has based its decision. In 2D models, this information is easily provided

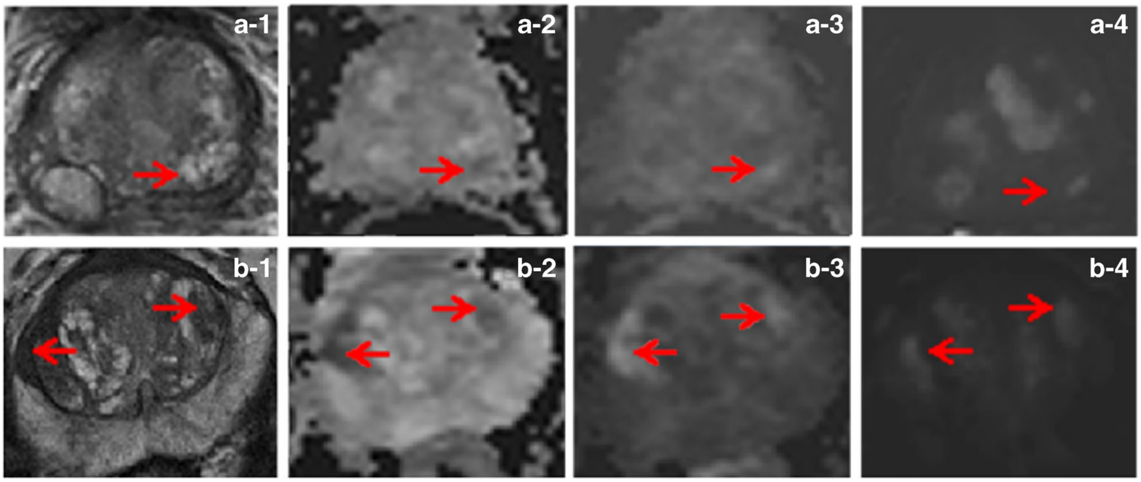

Fig. 6 Incorrectly predicted lesions. **a** and **b** represent different cases, where lesions "A" and "B-left" are nonsignificant, yet they were predicted as significant lesions, while lesion "B-right" is significant and was predicted as nonsignificant lesion. (1–4) represent the sequences of T2w, ADC, DWI, and *K*-trans respectively, where arrows indicate then lesion's location and T2w is used for visualization purposes

by visualizing the class activation map, which acts as a heat map and shows the part of the image most markedly involved in the classification process. Unfortunately, since our model is 3D, there are currently no reliable mapping methods available which could visualize such an activation pattern and show the radiologists the part of the image that has driven the classification procedure. As an alternative, we showed the performance of the network in terms of correctly predicted lesions and plotted the accuracy during training and testing to give an impression of the network behavior (see Fig. 6).

The network performance depends on many aspects such as the quality of the input images as well as the intrinsic features of the network (architecture and parameters, etc.). Figure 6 shows some examples of incorrectly predicted lesions where each row represents a different case and each column represents a different MR sequence. As we can see from case "A," the lesion is not significant, yet it was predicted as a significant lesion. The reason behind this could be that the lesion is visible and has a higher contrast in all MR sequences that were used as an input which in turn enhanced the presence of the lesion and made its contrast more pronounced. In case "B," there are two lesions; the one on the left side is significant while the one on the right is not. However, the network predicted that both lesions are not significant. This error could be due to the fact that both lesions have almost the same level of contrast with the surrounding tissues and furthermore, they are visible in all input sequences that were used as an input. In addition, the network in this stage is not perfect; some modifications might be necessary to optimize the network performance and this could be addressed in future work.

Figure 3 might suggest that correct classification of prostate lesions by the network varies with lesion size or diameter and is better for larger lesions, which are more conspicuous. However, Fig. 4 shows that there is no correlation between lesion size or lesion largest diameter and network performance. Hence, the network's decision apparently does not depend on lesion size nor the largest diameter, and lesion size, in turn, does not lead to correct prediction or better performance.

Our study has limitations. A dataset of 200 patients (318 individual 3D volumes) was used, which is a small number of training examples for a neural network. Although the images used were augmented multiple times using different augmentation methods, the resulting images did not differ much from the original images, potentially not reducing overfitting as effectively as desired; more images would be needed. Furthermore, the network only predicted the clinical significance of prostate lesions but not where they were located; lesion detection, additionally to lesion characterization, would naturally be desirable.

In conclusion, automated prostate cancer classification using the proposed multi-channel 3D deep convolutional neural network is feasible in our initial clinical evaluation. The only required parameter in our approach is the lesion location; no further manual or semi-automatic segmentation is necessary. The network takes different combinations of 3D volumes of distinct MR sequences (e.g., T2w, ADC, and DWI) at the same time. The diagnostic performance of this CNN is comparable to that achieved by experienced radiologists, indicating that the network is a very promising tool to improve computer-aided diagnostic capabilities for PCa classification as well as clinical workflow.

Funding This work was funded by the German Research Foundation (Graduate Program BIOQIC, GRK2260, BIOQIC) and the Berlin Institute of Health within the Clinician Scientist Program (TP).

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Prof. Dr. Marc Dewey.

Conflict of interest The authors of this manuscript declare relationships with the following companies: Nader Aldoj is a PhD student supported by a graduate program of the German Research Foundation (GRK2260, BIOQIC). Prof. Dewey (a PI of BIOQIC funded by the German Research Foundation, GRK2260) has also received grant support from the Heisenberg Program of the DFG for a professorship (DE 1361/14-1) and the FP7 Program of the European Commission for the randomized multi-center DISCHARGE trial (603266-2, HEALTH-2012.2.4.-2). Prof. Dewey is a cardiac section editor of *European Radiology*. Prof. Dewey has received lecture fees from Toshiba Medical Systems, Guerbet, Cardiac MR Academy Berlin, and Bayer (Schering-Berlex). Tobias Penzkofer received research support from Siemens Healthcare and Philips Healthcare. Tobias Penzkofer received grant support from the Berlin Institute of Health within the Clinician Scientist Program. Outside of the current work, Tobias Penzkofer is involved in clinical trials with AGO, Aprea AB, Astellas Pharma Global Inc., AstraZeneca, Celgene, Genmab A/S, Incyte Corporation, Lion Biotechnologies, Inc., Millennium Pharmaceuticals, Inc., Morphotec Inc., MSD Tesaro Inc., and Roche. Institutional master research agreements exist with Siemens Medical Solutions, Philips Medical Systems, and Toshiba Medical Systems. The terms of these arrangements are managed by the legal department of Charité – Universitätsmedizin Berlin. The other authors declared no conflicts of interest.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent No consent was necessary. The dataset is publicly available.

Ethical approval No ethical approval was necessary.

Methodology
- Retrospective
- Performed at one institution

References

1. Siegel R, Naishadham D, Jemal A (2013) Cancer statistics. CA Cancer J Clin 63(1):11–30

2. Johnson LM, Turkbey B, Figg WD, Choyke PL (2014) Multiparametric MRI in prostate cancer management. Nat Rev Clin Oncol 11(6):346–353
3. Chou R, Croswell JM, Dana T et al (2011) Screening for prostate cancer: a review of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med 155(11):762–771
4. Schröder FH, Hugosson J, Roobol MJ et al (2009) Screening and prostate-cancer mortality in a randomized European study. N Engl J Med 360(13):1320–1328
5. Sidhu HS, Benigno S, Ganeshan B et al (2017) Textural analysis of multiparametric MRI detects transition zone prostate cancer. Eur Radiol 27(6):2348–2358
6. Fehr D, Veeraraghavan H, Wibmer A et al (2015) Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images. Proc Natl Acad Sci U S A 112(46): E6265–E6273
7. Peng Y, Jiang Y, Yang C et al (2013) Quantitative analysis of multiparametric prostate MR images: differentiation between prostate cancer and normal tissue and correlation with Gleason score–a computer-aided diagnosis development study. Radiology. 267(3): 787–796
8. Valerio M, Donaldson I, Emberton M et al (2015) Detection of clinically significant prostate cancer using magnetic resonance imaging-ultrasound fusion targeted biopsy: a systematic review. Eur Urol 68(1):8–19
9. Lemaitre G, Martí R, Freixenet J, Vilanova JC, Walker PM, Meriaudeau F (2015) Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review. Comput Biol Med 60:8–31
10. Tiwari P, Kurhanewicz J, Madabhushi A (2013) Multi-kernel graph embedding for detection, Gleason grading of prostate cancer via MRI/MRS. Med Image Anal 17(2):219–235
11. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. Conference paper at ICLR
12. Szegedy C, Liu W, Jia Y et al (2015) Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA
13. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Computer Vision and Pattern Recognition
14. Huang G, Liu Z, Van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
15. Shin HC, Roth HR, Gao M et al (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 35(5):1285–1298
16. Tajbakhsh N, Shin JY, Gurudu SR et al (2016) Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans Med Imaging 35(5):1299–1312
17. Kasel-Seibert M, Lehmann T, Aschenbach R et al (2016) Assessment of PI-RADS v2 for the detection of prostate cancer. Eur J Radiol 85:726–731
18. Le MH, Chen J, Wang L et al (2017) Automated diagnosis of prostate cancer in multi-parametric MRI based on multimodal convolutional neural networks. Phys Med Biol 62(16):6497
19. Yang X, Liu C, Wang Z et al (2017) Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric MRI. Med Image Anal 42:212–227
20. Mehrtash A, Sedghi A, Ghafoorian A et al (2017) Classification of clinical significance of MRI prostate findings using 3D convolutional neural networks. Proc SPIE Int Soc Opt Eng 10134. pii: 101342A
21. Fawcett T (2006) An introduction to ROC analysis. Pattern Recogn Lett 27(8):861–874
22. Litjens G, Debats O, Barentsz J, Karssemeijer N, Huisman H (2017) ProstateX Challenge data. The Cancer Imaging Archive DOI: 10.7937/K9TCIA.2017.MURS5CL
23. Litjens G, Debats O, Barentsz J, Karssemeijer N, Huisman H (2014) Computer-aided detection of prostate cancer in MRI. IEEE Trans Med Imaging 33:1083–1092
24. Clark K, Vendt B, Smith K et al (2013) The cancer imaging archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 26(6):1045–1057
25. Kingma DP, Lei Ba J (2017) Adam: a method for stochastic optimization arXiv:1412.6980v9 [cs.LG] 30 Jan.
26. Lui S, Zheng H, Feng Y, Wi L (2017) Prostate cancer diagnosis using deep learning with 3D multiparametric MRI. In: Medical imaging 2017: computer-aided diagnosis, vol 10134. International Society for Optics and Photonics, p 1013428

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.