Prostate cancer detection from multi-institution multiparametric MRIs using deep convolutional neural networks

Yohan Sumathipala
Nathan Lay
Baris Turkbey
Clayton Smith
Peter L. Choyke
Ronald M. Summers

Prostate cancer detection from multi-institution multiparametric MRIs using deep convolutional neural networks

Yohan Sumathipala,^a Nathan Lay,^a Baris Turkbey,^b Clayton Smith,^b Peter L. Choyke,^b and Ronald M. Summers^{a,*}

^aNational Institutes of Health Clinical Center, Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, Bethesda, Maryland, United States

^bNational Institutes of Health, National Cancer Institute, Molecular Imaging Program, Bethesda, Maryland, United States

Abstract. Multiparametric magnetic resonance imaging (mpMRI) of the prostate aids in early diagnosis of prostate cancer, but is difficult to interpret and subject to interreader variability. Our objective is to generate probability maps, overlaid on original mpMRI images to help radiologists identify where a cancer is suspected as a computer-aided diagnostic (CAD). We optimized the holistically nested edge detection (HED) deep convolutional neural network. Our dataset contains T2, apparent diffusion coefficient, and high b -value images from 186 patients across six institutions worldwide: 92 with an endorectal coil (ERC) and 94 without. Ground-truth was based on tumor segmentations manually drawn by expert radiologists based on histologic evidence of cancer. The training set consisted of 120 patients and the validation set and test set included 19 and 47, respectively. Slice-level probability maps are evaluated at the lesion level of analysis. The best model: HED using 5×5 convolutional kernels, batch normalization, and optimized using Adam. This CAD performed significantly better ($p < 0.001$) in the peripheral zone ($AUC = 0.94 \pm 0.01$) than the transition zone. It outperforms a previous CAD from our group in a head-to-head comparison on the same ERC-only test cases ($AUC = 0.97 \pm 0.01$; $p < 0.001$). Our CAD establishes a state-of-the-art performance for predicting prostate cancer lesions on mpMRIs. © 2018 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: [10.1117/1.JMI.5.4.044507](https://doi.org/10.1117/1.JMI.5.4.044507)]

Keywords: computer-aided diagnosis; prostate cancer; deep learning; neural networks; image processing; magnetic resonance imaging.

Paper 18161R received Jul. 26, 2018; accepted for publication Nov. 5, 2018; published online Dec. 15, 2018; corrected Sep. 11, 2019.

1 Introduction

With more than 200,000 new cases diagnosed each year in the United States, prostate cancer is the most common form of nonskin cancer among men.¹ Over 27,000 men die from this disease each year.^{1,2} A key contributor to the disease burden and mortality rate is the underdiagnoses of aggressive cancers through the traditional transrectal ultrasound (TRUS) biopsy which randomly samples the prostate.^{3–5}

Multiparametric magnetic resonance imaging (mpMRI) is increasingly used in the early diagnosis of prostate cancer because of its ability to identify lesions for image-guided biopsy. Furthermore, this imaging modality can stratify patients into risk categories owing to development of the prostate imaging reporting and data system (PI-RADS) scale.^{6,7} Conducted with a T2-weighted image (T2WI) and diffusion-weighted sequences, mpMRI has an 80% sensitivity in both the peripheral and transition zones; its negative predictive value is nearly 98%.⁸ As such, mpMRI-guided biopsies can significantly improve the sensitivity and specificity of cancer detection over the random TRUS biopsy: mpMRI biopsies are more likely to detect aggressive cancers and less likely to detect indolent cancers.^{5,9,10}

Despite the several clear clinical advantages of mpMRI, there are several barriers to its full implementation. MR scans can be of varied quality interfering with diagnosis. Interpreting mpMRI requires expertise that the general radiologist may not possess.

Furthermore, interpretations of even good quality scans can vary considerably, depending on mpMRI acquisition parameters, tumor visibility, and the experience and opinions of radiologists.¹¹ Although guidelines such as PI-RADS v2 attempt to standardize many of these variables, they are inherently subjective and are associated with a high degree of interreader variation.^{7,12–15} The result is that value of the mpMRI is compromised in many cases.

One step toward capturing the full value of mpMRI in prostate cancer is computer-aided detection (CAD). Applied to prostate cancer detection and other medical imaging tasks, CAD tools have been shown to reduce interobserver variation.^{16–20} To achieve this success, CAD systems employ machine learning methods. At a high level, these methods use a myriad of mathematical functions to create a quantitative feature set from which predictions about tumor location and malignancy are made using a classification algorithm that is trained against a specific subset of training examples. In this research, we developed a CAD system that automatically detects suspicious lesions in mpMRIs of the prostate that readers can use as a reference to reduce interpretation time and interreader variability.

Most of the successful CAD systems developed for prostate cancer use a handcrafted set of several numerical features that were historically developed specifically for natural or otherwise nonmedical images; classification is typically performed using a random forest or similar method.^{16,17,21–23} However, such

*Address all correspondence to: Ronald M. Summers, E-mail: rms@nih.gov

methods are limited by the features they use. Thus, they perform differently on mpMRI scans from different institutions and even from different machines of the same magnetic field strength within a single institution.^{24–26} While such methods tend to work well on most problems, even with default parameters, their performance tends to saturate as the training set size grows.^{21,27}

An alternative approach that has enjoyed much recent success is deep learning.²⁸ These systems directly learn from the training images a set of discriminative features for a given learning task, such as classification. Due to the success of deep learning methods, we investigated a deep convolutional neural network (DCNN) to find a solution to prostate cancer detection on mpMRI. Using fully supervised learning, our proposed CAD system aims to determine precisely where there is likely to be cancer at the pixel level. As such, we train our model against a database of pixelwise annotations or segmentations for known prostate cancer. We generate probability maps that can be overlaid on the original T2WI, apparent diffusion coefficient (ADC) map, and/or high *b*-value images to help radiologists identify regions where a cancer is suspected. We chose to create such a CAD because a recent international, multireader study by Greer et al.²⁹ showed that a probability map can significantly improve the sensitivity of readers to more subtle lesions.

Greer et al.²⁹ also found that the CAD substantially reduced radiologists' specificity, which may be related to its performance. It is also possible that using probability maps to present lesion proposals to radiologists is responsible. There is yet no published comparison of presenting lesions using probability maps versus other methods. As the purpose of our research is to assess whether deep learning can improve the detection of individual prostate cancers (lesion level of analysis), we consider the problem of how to present probability maps to radiologists out-of-scope. Thus, we aim to develop a new CAD system that outperforms the Ref. 21 CAD on the probability map task.

Only a few prostate CAD systems utilize deep learning to detect lesions in mpMRI images. Three studies published in 2017 to 2018 used deep learning to characterize whether a specific image patch contains a cancerous lesion or is noncancerous tissue,^{30–32} an entirely different problem than the one our research aims to address. Specifically, the DCNN models proposed by Le et al.³⁰ and Liu et al.³¹ used modified VGG16 architectures that do not generate probability maps or any another predictive tool that could be immediately used to help radiologists interpret mpMRI images in the manner that Greer et al.²⁹ have shown to be successful. The only deep learning system we could find in the published literature that produces probability maps was developed by Tsehay et al.³³ to predict biopsy points. Although their scope is different from ours, their research demonstrated that directly applying the holistically nested edge detector (HED) architecture, initially developed for analyzing natural images, to the problem of prostate cancer detection can yield good predictions.³³ The HED DCNN³⁴ is a neural network architecture that has been implemented to segment a variety of anatomic structures from CT images including the pancreas, lymph node clusters, and the lungs.^{35–37}

We aim to improve on the HED-based architecture to produce probability maps that are clinically meaningful. To this end, we compare the performance of our system to the random forest-based CAD system of Lay et al.²¹ that was used in the Greer et al. multireader study²⁹ for two primary reasons: (1) Lay et al. evaluated their CAD in the exact same way as our proposed system, although the training and testing data

sets used by our study differ from theirs and (2) because the Lay et al. CAD demonstrably helps radiologists, matching (or besting) their performance by their same metrics with a deep learning CAD would indicate that this new CAD would be similarly impactful in the real world.

To unearth insights that may help overcome the many challenges present in developing a robust system to automatically detect prostate cancers from mpMRI, we take an iterative, incremental approach toward optimizing the HED-based architecture. This research implements three specific methodologies: (1) Rician normalization, instead of histogram normalization, of the raw mpMRI images of the prostate, as Ref. 38 indicates that this improves the performance of a random forest prostate CAD system; (2) batch normalization during training which is widely shown to improve the generalization of DCNNs,³⁹ and (3) larger convolutional kernels, which effectively increases the receptive field (amount of local image context) our DCNN uses to make predictions about where there is or is not prostatic cancer.^{40,41}

2 Methods and Materials

2.1 Patient Cohort

The patient cohort consists of 186 patients, each of whom underwent mpMRI scanning wherein T2WI, diffusion weighted imaging with ADC, and high *b*-value axial image volumes of the prostate were obtained. For all mpMRIs, an expert radiologist and a dedicated prostate MRI fellow provided reference standard segmentations of the whole prostate, the transition zone, and tumors. The tumor segmentations for institute 6 were based on TRUS/MRI fusion guided transrectal biopsy histopathology, whereas tumor segmentations of all the other institutes were based on radical prostatectomy derived histopathology. The radiologist used this validation information to manually draw tumor contours on axial T2WI sequences. The mpMRI scans of these patients were performed at six institutions across three continents and multiple scanner manufacturers (Table 1). All mpMRI images were acquired using 3T MR scanners using the anterior half of a 32-channel SENSE cardiac coil when applicable. When an endorectal coil (ERC) was utilized, the

Table 1 Distribution of patient cases in training, testing, and validation sets by institute and average lesion size at the slice level.

	Train	Test	Validation
Institute 1	9	2	0
Institute 2	5	0	3
Institute 3	6	1	1
Institute 4	8	3	2
Institute 5	9	1	1
Institute 6	83	40	12
Total patient cases	120	47	19
Slices with tumor	691	221	87
Average lesion size (in mm ³)	485	330	684

balloon of each coil was distended with ~ 45 mL of perfluorocarbon or an equally suitable fluid to reduce air-induced artifacts. This follows the imaging protocol cited by Ref. 21.

This data set includes mpMRI images captured with ($n = 92$) and without ($n = 94$) the insertion of an ERC. The high b -value diffusion weighted image was captured with $b = 1500$ for patients imaged without the insertion of an ERC, and $b = 2000$ with ERC insertion. If either high- b value sequence was not acquired for the patient, the appropriate sequence was calculated from the other images.

The CAD by Lay et al. was trained exclusively on mpMRI images taken with the insertion of an ERC. These systems are unlikely to perform well on images without an inserted ERC, because the coil enhances the resolution of the prostate and changes the noise distribution but also distorts the shape of the normal prostate.

2.2 Training, Validation, and Testing Splits

The 186-patient cohort was split randomly into training ($n = 120$, 65%), testing ($n = 47$, 25%), and validation ($n = 19$, 10%) sets at the patient level (Table 1).

2.3 Image Preprocessing

It is well known that prostate images are heterogeneous, with absolute intensity values and MRI noise distributions varying significantly between institutions, scanners, and the insertion of an ERC.^{24,42} As such the interpretation of absolute intensity values and differences vary among patients. This phenomenon can adversely affect machine learning programs which require absolute differences to have a consistent meaning.²⁷ To overcome this problem, we first segment the prostate, thereby removing all nonprostate anatomy from the MRI sequence. We then convert the raw intensity values to a Z-score using Rice-normalization. We invoke this normalization method because theoretical and experimental evidence suggests that an MRI scan's noise distribution can be described by a Rician distribution.⁴³ This ensures that a relative difference of, for example, three units in the Rice-normalized image has the same meaning across all patients. Furthermore, this normalization technique can be thought of as analogous to a Z-score from a normal distribution, in the sense that it does not alter

substantially the underlying distribution of intensity values. This is specifically because 3 T mpMRI images have a relatively high signal-noise ratio, a condition under which the Rician distribution approximates a Gaussian distribution. We calculate the Rician Z-scores at the patient level to the intensity of all pixels in the prostate across all the slices, ignoring the intensities of all pixels outside the prostate. We apply this normalization to the T2W, ADC, and high b -value sequences separately. For each patient, the T2W, ADC, and high b -value images are registered using the spatial coordinates of the scanner to ensure the (X, Y, Z) coordinates of any pixel in one sequence map to the same anatomic location in another sequence.

After subtracting nonprostate background and applying a Rician normalization, the images are cropped around the prostate in the same way as Ref. 33. We crop in three-dimensions (3-D) around the center of the prostate. This results in prostate-only volume that perfectly encloses the prostate, where at least one pixel touches a face of the bounding cube and no prostate pixels exist outside the cube. Naturally, the XY extent of this resulting volume will be different for each patient—a fact that will pose problems for the HED deep learning system under investigation. To standardize the XY extent of each patient's cropped prostate volume, we apply a bicubic upsampling to create a volume that is 382×382 pixels in dimension. We pad 0s around the border of this upsampled volume to give a final dimension of 384×384 . This process of cropping and resampling is applied to the T2W, ADC, and high b -value sequences. We crop the ground-truth segmentations using the same bounding cube and apply a linear upsampling to rescale the tumor segmentation in the same way as the acquired MRI sequences. Just as was done in Ref. 33 and is now common in the field of medical computer vision, we consider the T2W, ADC, and high b -value sequences as different channels of a 3-D volume. This preprocessing protocol is applied to all the images and reference tumor segmentations in the training, validation, and test images.

3 Neural Networks

All the neural networks described are built using the PyTorch 0.30 deep learning framework in Python 2.7 on an Ubuntu 14.7 computer with a NVIDIA Tesla K40 graphics card. They are all initialized using the He initialization, which is the default in PyTorch.⁴⁴

Fig. 1 Illustration showing the HED network producing outputs of reduced scale at each layer and the final fused image.³⁴ The fused image is generated by up sampling the side-outputs and using 1×1 convolutions.

4 Holistically Nested Edge Detector

The holistically nested edge detector is a neural network architecture that was initially developed for detecting edges in natural images³⁴ and has since grown in popularity for use in anatomy segmentations from CT scans.^{35–37} The original network has five main layers, each of which produces a segmentation map, but at reduced scales, as shown in Fig. 1. These so-called side-outputs are fused together in a separate layer to give rise to the output image. Training is performed in a deeply supervised fashion wherein an error against the ground-truth is calculated and back-propagated from all the side-outputs and the fused image. In other words, each layer has its own loss function. At test and validation time, only the final fused image is used to render a segmentation map. In the original network, each layer is comprised of two (layers 1 and 2) or three (layers 3, 4, and 5) 3×3 convolutions followed by ReLU activation. Between each layer, a twofold downsampling operation is performed. Bilinear interpolation is applied to rescale each side-output to the original input dimensions as Xie and Tu found this gave better results than other learnable methods.³⁴

The HED network is end-to-end, taking a full three channel image as an input and outputting the corresponding probability map. As the network is designed to output two-dimensional (2-D) probability maps, we train the network and generate outputs at the slice level.

5 Modifications to the HED Network

5.1 Gradient Descent Methods

The HED network is trained using stochastic gradient descent (SGD) with learning rate = $1e-8$, momentum = 0.9, and weight decay = 0.0002, which mirrors the parameters used in Ref. 33. We compared this native HED trained using SGD with the same network architecture trained using the Adam gradient descent method with learning rate = $2e-4$. Finding the Adam method optimized the HED network 50 times faster than the SGD method, we implemented it in all subsequent experiments.

5.2 Batch Normalization

Batch normalization is a technique to normalize the inputs of each layer in a neural network so that the activations follow a standard normal distribution. This procedure has been shown to improve the generalization of neural networks. For prostate mpMRI images, where the input data in the ADC and high b -value images may be quite noisy, this procedure is likely to be of significant help. We implemented batch normalization before every ReLU activation in every layer of the HED network. Finding that adding batch normalization improved performance, we used it in all subsequent experiments.

5.3 Kernel Size

Increasing kernel size increases the effective receptive field of each layer in the network, which is shown to improve performance because each convolution effectively uses more local context to determine if a pixel is or is not a part of a cancerous tumor.^{40,41} The original HED architecture uses convolutions with 3×3 kernels everywhere; we call this the HED 3×3 network. We constructed an HED network with 5×5 kernels in every convolution and compared its performance to the original

network to increase local context at all levels and call this HED 5×5 .

We do not consider an HED network with 7×7 kernels in every convolution. Because each 7×7 kernel has 49 parameters (compared to 25 for HED 5×5 and 9 for HED 3×3), such a network has a substantially increased complexity. This increases the putative HED 7×7 model's propensity to over-fit the training data and impedes its generalization. Moreover, training such a network is too computationally expensive (over 20 min to train each epoch), as there are twice as many arithmetic operations using 7×7 convolutions than 5×5 .

Instead, we created a network that emphasizes increased local context at the top of the network, where the receptive field is initially the smallest, and retain the 3×3 convolutions at the bottom of the network, where the receptive field is largest due to natural downsampling from the HED architecture. This network uses 7×7 kernels in the first layer, 5×5 kernels in the second and third layers, and 3×3 convolutions in the last two layers; this is our HED 7×7 , 5×5 , 3×3 network which we abbreviate as HED 7,5,3.

6 Training

We train the HED network on a binary classification task of identifying whether a given pixel in the image belongs to a prostate cancer or noncancerous tissue—a similar task as the network was initially designed for. As such, the HED network uses a sigmoid as its output layer and is trained using binary cross-entropy loss. The loss function is weighted so that the weight for cancerous pixels is the same as the weight for non-cancerous pixels. These weights are determined at the batch level of six 2-D slices per batch. We train our system only on image slices that have prostate cancer and ignore slices without prostate cancer. Table 1 shows the number of slices used to train, validate, and test the DCNNs. We perform data augmentation by flipping slices about the vertical axis. This training paradigm is identical to that used by Ref. 33.

7 Statistical Analysis

7.1 Direct Evaluation

We evaluate the trained HED model at the patient level against all mpMRI slices that contain the prostate following preprocessing. Receiver operating curve (ROC) analysis is computed on the output probability maps in 3-D: if (1) the predicted probability scores from the CAD lie within the volume of a ground-truth tumor contour segmentation and (2) the 90th percentile of the probability scores are larger than a threshold value, then the lesion is considered detected. To determine false positives, the CAD probability map is divided into 3×3 voxel cells; we define a false positive if (1) the 90th percentile of probability scores within a cell exceeds the threshold and (2) the cell is more than 3 mm away from the ground-truth segmentation volume. We primarily report lesion detection in the entire prostate. In addition, we report lesion detection in the transition zone and peripheral zones separately, using segmentations of these zones that were drawn by an expert radiologist. The evaluation methodology and reporting of CAD performance in the entire prostate plus its two major anatomic zones are identical to that used by Ref. 33.

Only one model state for each network is evaluated at test time. We select the model state that attains the best performance on the validation set. To estimate the variance of CAD

performance on the test set, we use 10× bootstrapping. We create 10 lists of 47 patients each by randomly sampling cases in the original test list, and evaluate performance on each list. Then, we compute the variance of the performance on these lists for each HED model. This allows us to directly assess the variance of performance on our test list. To determine whether two AUCs (from different neural network architectures or performance on different cases by the same architecture) are statistically significantly different, we use the Student's *t*-test. We compute the two-sample *t*-test statistic using the aforementioned variance estimate and use the performance on the original test list (all 47 distinct cases for testing) as the mean performance on the test dataset.²⁷ This evaluation method has precedent in the biomedical imaging literature, including prostate cancer applications (Ref. 45).^{27,45,46}

7.2 Head-to-Head Comparison with Lay et al. Random Forest CAD

The random forest CAD developed by Lay et al. was trained and tested exclusively against mpMRIs acquired with an ERC.^{42,47,48} Our HED models are trained and evaluated on cases with and without an ERC. Thus, a fair head-to-head comparison of our deep-learning CAD system with the random forest is only possible on cases with an ERC. To perform this evaluation, we construct a new test set that only contains the 26 cases used by Lay et al. to evaluate their CAD, and obtain their performance on these cases as described in Ref. 21. Of these head-to-head evaluation cases, 19 were in our original training set, 1 was in the validation set, and 6 were in the test set. This necessitated retraining the HED 5 × 5 architecture. To retrain the neural network, we first created a test list containing all 26 cases from Ref. 21. We substituted the overlapping case in our validation set with a randomly chosen case (i.e., a case not in the new test set). This helped ensure that we picked the optimal model state posttraining in the same way as before. All the remaining 142 patient cases were used to train the HED 5 × 5 model. We reported performance on the head-to-head comparison using the ROC evaluation described above.

7.3 Performance on Images With and Without an ERC

Because prostate mpMRIs are often acquired with or without an ERC—there is no clinical best-practice for the routine use of an ERC^{47,48}—it is important that our new CAD system perform

equally well on both image types. In general, images taken without an ERC are of lower quality than images with one. To ensure we have a sufficiently large number of cases ($n > 30$) acquired with and without an ERC, we pool our test and validation cases together ($n = 66$ cases). We then evaluate our CAD against cases in this dataset that were acquired with an ERC ($n = 35$) and cases acquired without an ERC ($n = 31$). We use the bootstrapping method and a Student's *t*-test as described above to assess statistical significance.

7.4 Postprocessing

No postprocessing is performed, as the goal of this study is to evaluate the raw efficacy of the proposed neural networks in the task of producing probability maps to detect prostate cancer. We do not apply a threshold when rendering probability maps nor do we suppress predictions made outside of the prostate. Predictions outside the prostate are likely an artifact from upsampling and convolving the lower-resolution images from the lower layers of the HED neural network.

8 Results

8.1 Gradient Descent

In every case, the SGD trained model (following the approach of Tsehay et al.) predicted tumors with a probability 1.0 across most of the prostate and much of the background outside the prostate, while the Adam trained model gave predictions in only a few specific regions. Furthermore, the SGD model failed to converge after 150 epochs (~30 h of training time) whereas the Adam model converged after just three epochs (~30 min of training time). Thus, we used the Adam method.

8.2 Batch Normalization

Both gradient optimization methods used to train the HED neural network give “probability map” outputs that are almost exclusively binary: predictions are either exactly 1 or 0. Introducing batch normalization solves this problem, as there is a far better distribution of values across the [0, 1] interval, which is an observed limitation of the CAD proposed in Ref. 33. This also significantly improves the AUC performance (Table 2). Qualitatively, the introduction of batch normalization gives probability maps that use a wider range of colors, allowing a reader to clearly distinguish areas that the network has strongly

Table 2 AUC performance of the all the HED models trained.

Model	Prostate AUC		Transition zone AUC		Peripheral zone AUC	
	Val	Test	Val	Test	Val	Test
Original HED 3 × 3, no batch normalization	0.60 ± 0.01	0.63 ± 0.01	0.54 ± 0.01	0.54 ± 0.01	0.65 ± 0.01	0.67 ± 0.01
HED 3 × 3	0.88 ± 0.01	0.91 ± 0.01	0.88 ± 0.01	0.90 ± 0.01	0.89 ± 0.01	0.92 ± 0.01
HED 5 × 5	0.86 ± 0.01	0.93 ± 0.01	0.85 ± 0.01	0.92 ± 0.01	0.86 ± 0.01	0.94 ± 0.01
HED 7 × 7, 5 × 5, 3 × 3	0.84 ± 0.01	0.88 ± 0.02	0.81 ± 0.01	0.86 ± 0.02	0.86 ± 0.01	0.90 ± 0.01
Lay et al. (21)	0.93					

Unless specified, HED was trained with batch normalization. All mean test AUCs are significantly different from each other at the $\alpha = 0.001$ level using the Student's *t*-test. Values for test AUC are given as mean ± one standard deviation. Bold indicates highest performance in each column.

Fig. 2 True-positive example on a scan acquired with an ERC from institute 6. All images show the same axial slice. (a)–(d) Original mpMRI images windowed to show full range of intensity, as this is what HED sees: (a) T2WI, (b) ADC, (c) B2000, and (d) T2WI with ground-truth tumor segmentation (red). (e)–(h) T2WI with probability map overlays with consistent windowing (blue is 0, dark red is 1): (e) original HED with 3×3 kernels and no batch normalization, (f) HED 3×3 , (g) HED 5×5 , and (h) HED 7,5,3. The lesion is clearly hypointense in ADC and TWI and hyperintense in B2000. The probability maps generated by the original HED are a binary mask of either ones or zeroes with many false-positive predictions. With batch normalization, the HED 3×3 , 5×5 , and HED 7,5,3 models correctly identify the lesion in the anterior midline lesion with minimal false positives.

predicts are cancerous from areas that it predicts as less likely. Without batch normalization, this information is lost (Figs. 1 and 2, e and f). Because it improves the performance of the base HED CAD system both quantitatively and qualitatively, we modified the base HED architecture, implementing batch normalization before every ReLU activations when training subsequent models.

8.3 Kernel Size

By increasing the size of the kernels used in every convolutional layer, we sought to reduce the aforementioned type of false positives (Figs. 2 and 3). We show that increasing kernel size reduces false-positive predictions. However, the general qualitative locations where lesions are predicted within the prostate

Fig. 3 Mix of true-positive and false-positive detections on scan acquired without an ERC from institute 2. All images show the same axial slice. (a)–(d) Original mpMRI images windowed to show full range of intensity. (a) T2WI, (b) ADC, (c) B2000, and (d) T2WI with ground-truth tumor segmentation (red). (e)–(h) T2WI with probability map overlays with consistent windowing (blue is 0, dark red is 1): (e) HED 3×3 no batch norm, (f) HED 3×3 predictions, (g) HED 5×5 predictions, and (h) HED 7,5,3. Probability maps generated without batch normalization appear as a binary mask. Increasing kernel size reduces false-positive predictions but does not change the general region where tumors are proposed. The false-positive predictions in the posterior prostate appear because the ADC image is strongly hypointense at the left and right apex of the posterior prostate; the B2000 sequence in this region is patchy with some hyperintense regions.

Fig. 4 False-positive example on a scan from institute 6 acquired with an ERC. All images show the same axial slice. (a)–(c) Original mpMRI images windowed to show full range of intensity, as this is what HED sees: (a) T2WI with ground-truth tumor segmentation (red), (b) ADC, and (c) calculated B2000. (d)–(f) T2WI with probability map overlays with consistent windowing (blue is 0, dark red is 1): (d) HED 3×3 predictions, (e) HED 5×5 predictions, and (f) HED 7,5,3. The ground truth lesion is weakly detected by HED 3×3 and 5×5 . The large false-positive region in the anterior region of the image is patchy and only intermediately hypointense in ADC and hyperintense in B2000. This case highlights a very rare shortcoming of this CAD.

do not change, as shown in the false-positive case (Fig. 4) and the failure cases (Figs. 5 and 6). Quantitatively, a kernel size of 5×5 gives the best performance, while the HED 7,5,3 gives the worst (Table 2; Fig. 7). Our HED 5×5 system significantly outperforms all our other systems ($p < 0.001$).

8.4 Performance in Prostate Zones

Every HED model performed significantly better in the peripheral zone than the transition zone at the ($p < 0.05$ all) using a t -test (Table 2). This difference was most pronounced in the

Fig. 5 False negative on an image acquired with an ERC from institution 4. All images show the same axial slice. (a)–(c) Original mpMRI images windowed to show full range of intensity, as this is what HED sees: (a) T2WI with ground-truth tumor segmentation (red), (b) ADC, and (c) b-1500. (d)–(f) T2WI with probability map overlays with consistent windowing (blue is 0, dark red is 1): (d) HED 3×3 predictions, (e) HED 5×5 predictions, and (f) HED 7,5,3. The ground-truth lesion in the right anterior transition zone is only marginally hypointense in the T2 and ADC sequences and ambiguous in the b-1500 sequence. All three HED systems miss the lesion, although the HED 3×3 system gives a low probability prediction near the lesion.

Fig. 6 False-negative example on a scan from institution 4 acquired without an ERC. All images show the same axial slice. (a)–(d) Original mpMRI images windowed to show full range of intensity, as this is what HED sees (a) T2WI, (b) ADC, (c) b-1500, and (d) T2WI with ground-truth tumor segmentation (red). (e) The peripheral zone (white) surrounding the transition zone (black area enclosed by the white) and the pathology confirmed lesion (red). (f)–(h) The T2WI with probability map overlays with consistent windowing (blue is 0, dark red is 1): (f) HED 3×3 predictions, (g) HED 5×5 predictions, and (g) HED 7,5,3. The tumor itself is intermediately hypointensity on T2WI, ambiguous in b-1500, and largely ambiguous on ADC. All three CAD systems make a strong false-positive prediction in the left lateral center region that is driven by (b) the strong ADC hypointensity and (c) b-1500 hyperintensity. Only the HED 7,5,3 network makes weak true-positive prediction. Both the HED 5×5 and 3×3 systems miss the lesion.

Fig. 7 ROC curves showing HED models with different kernel sizes evaluated at the patient level against the (a) validation and (b) test datasets. Performance evaluated on the entire prostate (both the transition zone and peripheral zone).

original HED and the HED 7,5,3 models. Our best model state, the HED 5×5 , achieves optimal performance in both the peripheral and transition zones at the $\alpha = 0.001$ level. Higher performance in the peripheral zone is preferred because cancer is more prevalent in this region of the prostate.⁴⁹

8.5 Performance on Cases Acquired With and Without an Endorectal Coil

The original HED network performs significantly better ($\alpha = 0.05$) on the cases acquired with an ERC, likely because

these images are higher quality. Curiously the HED 7,5,3 network performs better on the lower-quality images acquired without an ERC. HED 3×3 and HED 5×5 perform do not perform differently on images with and without an ERC at the $\alpha = 0.05$ confidence level (Table 3). Both the HED 7,5,3 and the original HED networks are significantly outperformed by the HED 5×5 and HED 3×3 on both types of images at the $\alpha = 0.01$ confidence level. Our best system, the HED 5×5 is the highest performer on both methods for acquiring images (AUC = 0.91 for both scans with and without an ERC).

Table 3 AUC performance of all the CAD systems against cases acquired with and without the insertion of an ERC.

Model	ERC coil	No ERC coil
Original HED 3×3	0.62 ± 0.01	0.60 ± 0.02
HED 3×3	0.90 ± 0.01	0.91 ± 0.02
HED 5×5	0.91 ± 0.02	0.91 ± 0.02
HED 7×7 , 5×5 , 3×3	0.86 ± 0.02	0.89 ± 0.02

Performance on scans acquired with and without the insertion of an ERC is significantly different at the $\alpha = 0.05$ level for only the original HED system and the HED 7,5,3 system—our two lowest performing systems. The HED 3×3 and HED 5×5 systems perform similarly on scans acquired with or without a coil, demonstrating the robustness of this CAD system.

8.6 Successful Qualitative Performance

The lesion in the patient shown in Fig. 2 is clearly hypointense in ADC and TWI and hyperintense in B2000. The probability maps generated by the original HED are a binary mask of either ones or zeroes and have numerous false-positive predictions. With batch normalization, the HED 3×3 , 5×5 , and HED 7,5,3 models correctly identify the lesion in the anterior midline lesion. The larger kernels used by the HED 5×5 and 7,5,3 models remove the false positives on the patient's left. Increasing kernel size generally improves the shape of the prediction, as the HED 7,5,3 prediction quite closely resembles the shape of the ground-truth tumor annotation.

8.7 False-Positive Qualitative Performance

In many cases, the CAD will correctly detect the real lesion and make false predictions. The patient in Fig. 3 is one such case, where the false-positive predictions in the posterior prostate appear because the ADC image is strongly hypointense at the left and right apex of the posterior prostate; the B2000 sequence in this region is patchy with some hyperintense regions. The HED 3×3 and 5×5 models make false-positive predictions in this region. The large false-positive prediction in the right posterior prostate from the HED 5×5 model is in a less patchy hypointense ADC and hyperintense B2000 region. The medial posterior prostate shows another region that is hypointense in ADC and hyperintense in B2000, but this is a false-positive lesion. The HED 3×3 and HED 7,5,3 model detect this region, with the latter detecting this false positive strongly and ignoring the apical false-positive regions.

Another example is shown in Fig. 4, where the CAD systems weakly propose the ground-truth lesion. In this patient, the HED 3×3 model predicts part of the ground-truth lesion in the left peripheral zone while the HED 5×5 predicts the lesion's shape more accurately. In addition, the HED 5×5 probability map creates better separation between predictions—the large anterior false positive and posterior true positive are more disconnected in (4E) than in (4F). Curiously, the HED 7,5,3 model hardly predicts the lesion at all, suggesting that local context is not the only key factor to perfectly accurate probability maps. The large false-positive region in the anterior region of the image is patchy and only intermediately hypointense in ADC and hyperintense in B2000. Performance on this case, with a large false positive and weakly predicted ground-truth tumor,

highlights a very rare, but impactful shortcoming of this CAD system.

8.8 False-Negative Qualitative Performance

The lesions the CAD fails to detect are difficult to observe in mpMRI (Figs. 5 and 6). For the patient in Fig. 5, the ground-truth lesion in the right anterior transition zone is only marginally hypointense in the T2 and ADC sequences and ambiguous in the b-1500 sequence. All three HED systems miss the lesion, although the HED 3×3 system gives a low probability prediction near the lesion. The large false-positive prediction in the medial transition zone is clearly hypointense in the both T2WI and ADC sequences, and ambiguous in the b-1500 sequence. For the patient in Fig. 6, the tumor itself is intermediately hypointense on T2WI, ambiguous in b-1500, and largely ambiguous on ADC. All three CAD systems make a strong false-positive prediction in the left lateral center region that is driven by the strong ADC hypointensity (6B) and b-1500 hyperintensity (6C). The HED 7,5,3 network makes weak predictions in a region that overlaps with the tumor that is much larger and of higher probability than the HED 3×3 , which only weakly detects a small part. The HED 5×5 system misses the lesion entirely. It is possible the increased context provided to the HED 7,5,3 network improves its ability to recognize this lesion. In both cases, the expert radiologists among us (BT, PLC, and RMS) unanimously agreed that a clinician reading these two cases prospectively would have great difficulty interpreting them correctly. Together, they highlight an uncommon occurrence that some tumors confirmed by histopathology are not visible in mpMRI. Clinically, it is more acceptable for the CAD to miss these lesions, as they are generally associated with low-risk cancer.^{50,51}

8.9 Comparison to Random Forest

Due to its superior qualitative performance and similar, but statistically significantly different quantitative performance, we propose HED 5×5 as the best deep learning model tested here. We compare the performance of a retrained HED 5×5 architecture (trained on more cases, validated on 19), against the performance of the random forest by Lay et al. against the exact same cases (Fig. 8). Our model gives an AUC = 0.97 ± 0.01 ,

Fig. 8 ROC curves comparing the HED models to the random forest CAD developed by Lay et al. The HED 5×5 model outperforms the random forest at every threshold, but the original HED does not.

which is significantly higher than the $AUC = 0.93$ reported by Ref. 21 ($p < 0.001$). The original HED system without batch normalization yields the worst performance ($AUC = 0.79 \pm 0.01$).

9 Discussion

In general, the deep learning convolutional neural networks perform similarly on images acquired with and without an ERC. The random forest CAD of Ref. 21 succeeds only on ERC cases. Our HED 5×5 architecture yielded the higher performance in the peripheral zone, transition zone, and the prostate overall. On the scans acquired with and without an ERC, the performance of this system ($AUC = 0.93$) is identical to that proposed by Ref. 21, which were found by Ref. 29 to significantly improve the sensitivity of readers to less-severe lesions and significantly reduce variability among readers of different experience levels. Furthermore, in a head-to-head comparison on the exact same cases (all acquired with an ERC), we found our HED 5×5 model substantially outperformed the Lay et al. random forest CAD ($AUC = 0.97 \pm 0.01$ and 0.93 , respectively). Taken together, these results strongly indicate our deep-learning system establishes a new state-of-the-art performance standard in the task of yielding probability maps to predict prostate cancer lesions. Our models are the first deep-learning models to achieve such performance on this task of prostate cancer detection from mpMRI images.

The original HED network proposed by Xie et al.³⁴ and used by Tsehay et al.³³ to produce probability maps from prostate mpMRI images was constructed without batch normalization and was trained using SGD gradient descent at a learning rate of 10^{-8} . We deviated from this approach, demonstrating that (1) Adam optimization method converges more rapidly without compromising generalizability and (2) batch normalization produces qualitatively and quantitatively superior probability maps (Fig. 3 and Table 2). The advantageous effects of Adam and batch normalization are well described.^{39,52} While it is possible batch normalization simply provides an insensitivity to network initialization, we consider this unlikely because all networks are initialized using the widely accepted He method.⁴⁴

Compared to the HED 3×3 model, the HED 5×5 and HED 7,5,3 models show fewer predictions in the posterior part of the prostate away from the ground-truth lesion [Fig. 3(f)–3(h)]. In the HED 7,5,3 probability maps, there are generally far fewer regions of dark red (very high probability prediction) than the HED 5×5 and HED 3×3 ; this indicates the model with the largest kernels is generally the least confident about where is a lesion. The HED 5×5 model on the other hand predicts a large, very dark region near the ground-truth [Fig. 3(g)]. Furthermore, the general spatial areas where these three HED models predict cancers are largely the same. Consider the failure case in Fig. 4, where there is a clear lesion in the left posterior peripheral zone that is hypointense in the ADC and hyperintense in the b-2000 sequences. The HED 5×5 predicts the shape of the lesion better than the HED 3×3 network but has less confidence that this region is cancerous. Together, these results suggest that increasing local context can improve the CAD's ability to predict the shape of a detected lesion, but it reduces the confidence with which a CAD system might predict a lesion, even when the lesion is rather plainly visible. This conclusion is confounded by the fact that the HED 5×5 model has twice as many parameters as HED 3×3 . Determining whether local context or a more complex model is specifically responsible for this

finding remains to be determined. The decreased performance by the HED 7,5,3 network suggests the context captured by 5×5 convolutional kernels in all layers of the HED network is at least partially responsible for the performance improvement.

In failure cases, increasing kernel size changes little. The patient shown in Fig. 5 has a lesion in the right anterior of the transition zone that all models fail to predict correctly. In this case, the HED 3×3 and HED 7,5,3 probability maps look quite similar. Both predict a large false positive in the medial transition zone that is clearly hypointense in the T2WI and ADC sequences. Both these models tend to clearly predict the shape of this false-positive region. The HED 5×5 model makes false-positive predictions with high probability in a similar region, but of distinctly different shapes. The strong similarity between the HED 3×3 and HED 7,5,3 models may be because the 3×3 kernels are used in the last two layers in the HED 7,5,3 model. In some failure cases (Fig. 6), the increased local context may improve lesion detection, and the HED 7,5,3 gives the best detection.

In both the failure cases (Figs. 5 and 6), it should be noted that the ground-truth lesion is very difficult to observe. These two images are representative of situations in which the CAD fails to detect a lesion. All “obvious” lesions that are hypointense in the T2WI and ADC sequence and hyperintense in the high b value sequence are correctly identified by the HED 3×3 and HED 5×5 systems for all patients in the test and validation datasets. This shows that the CAD is at minimum detecting the mpMRI image features that are correlated with prostate cancer. Figures 2 and 3 provide an example of what we consider an “obvious” case and how the CAD performs on them. The expert radiologists among us consider the lesion in Fig. 2 to be easier to detect than the lesion in Fig. 3. The only case where these CAD systems fail to mark an “obvious” lesion with a very strong probability is shown in Fig. 4—we present this case because it highlights a rare but potentially clinically important limitation of CAD performance.

One of the key challenges to developing a robust prostate cancer CAD system remains the heterogeneity and noisy behavior of prostate mpMRI images. Chiefly, the absolute meaning of intensity values is not consistent across mpMRI scans. We address this challenge at several junctures in the development of our CAD, starting with the application of a Rice normalization to the raw mpMRI images. Using batch normalization throughout the network to ensure that numerical values have a consistent absolute interpretation led to a substantial performance improvement that converts a CAD that generated a binary mask into a CAD that yields meaningful probability maps. We believe these two modifications are important to the successful development of future prostate CAD systems.

While the CAD is trained on a relatively large variety of cases coming from six institutions, this data set is still eclipsed by those used in other deep learning systems by several orders of magnitude.^{27,34,53} Most ($\sim 75\%$ of test and 85% of train cases) come from one institution, limiting our ability to generalize performance across all institutes globally. Furthermore, the data are inherently limited by varying quality issues such as image resolution and artifacts. This consequently limits the performance of our CAD. Privately developed CAD systems and privacy issues with medical image data also pose reproducibility and comparability challenges. Publically available datasets such as the one published for the ProstateX challenge use biopsy points as a ground-truth instead of tumor segmentations,

which is an entirely different task than the segmentation-like task considered in this research. Weakly supervised learning may be a possible solution, but we do not explore this topic in this work. The paucity of cases and corresponding tumor segmentations is further complicated by nonstandard evaluation methodology across various prostate CADs in the literature. We can only compare the performance of our CAD system with reported numbers in the literature.

Selecting HED 5×5 for the head-to-head comparison with Ref. 21 CAD could be biased if the reason for selecting HED 5×5 over HED 3×3 was based purely on the former's higher test set performance. This was not the case. Our reason for selecting HED 5×5 was more holistic, considering both qualitative and quantitative factors. Qualitatively, we have shown that HED 5×5 made fewer false-positives where the ADC sequence was hyperintense and the high b -value sequence was hypointense. In general, this made the probability maps from HED 5×5 easier to interpret in the opinions of the expert radiologists among us. It is because of the holistic (qualitative and quantitative) performance of the HED 5×5 network that we propose this as the best of our deep-learning networks and thus why we elected to compare it to Ref. 21 CAD. Potential bias is further reduced because the HED 5×5 network used in the head-to-head comparison was retrained and evaluated on largely new cases. In retraining, the optimal model state was chosen based on a validation set that differed from the original comparison by only one case, which helped ensure we picked the best set of learnable hyperparameters in a consistent way. Both the retraining and reasons for selecting HED 5×5 for the head-to-head comparison help to minimize the effects of bias in our conclusion that this CAD architecture establishes a new state-of-the-art performance in prostate cancer detection.

A final limitation of our system is the general lack of interpretability of deep learning systems. This is exemplified by the large false-positive prediction in the apical anterior midline transition zone seen in Fig. 4, a region that is patchy and only mildly hypointense in ADC and hyperintense in the high b -value sequence. Regrettably, this is a limitation inherent to all machine learning systems, and deep learning in particular as robust methods to interpret neural networks have not yet been developed.^{27,53}

10 Conclusions

In this research, we have presented one of the first deep learning systems to successfully produce probability maps for prostate cancer detection on mpMRI. These maps are of similar clinical utility to those evaluated by Greer et al.²⁹ Evaluated on identical cases, our best model, the HED 5×5 network, significantly outperforms the state-of-the-art random forest CAD²¹ that was clinically evaluated.²⁹ Moreover, the HED 5×5 CAD performs identically overall to the random forest on a much broader set of mpMRI images, spanning six institutes worldwide and images acquired with and without an ERC. Furthermore, it performs equally well on cases acquired with an ERC or without one, which is crucial because there is no clinical best practice for ERC usage. For these reasons, our best CAD system establishes a new state-of-the-art performance benchmark for prostate cancer prediction. The probability maps generated by our system have high clinical utility, as they outperform a clinically evaluated CAD system.

Disclosures

Dr. Summers receives royalties from iCAD, Philips, PingAn, and ScanMed and receives research support from PingAn and NVIDIA. He also reports grants from the National Institutes of Health Clinical Center under Grant No. 1Z01 CL040004. Dr. Peter Choyke reports grants from the National Cancer Institute. No other authors report conflicts of interest, financial or otherwise. This research was performed under both an institutionally approved protocol for handling patients in acquiring images and an IRB protocol for human data.

Acknowledgments

This research was funded by the Intramural Research Programs of the National Institutes of Health, Clinical Center and National Cancer Institute. We thank NVIDIA for GPU card donation.

References

1. C. A. F. Lawton, "Prostate cancer: improving the flow of research," *Radiology* **287**(1), 5–9 (2018).
2. M. S. Litwin and H.-J. Tan, "The diagnosis and treatment of prostate cancer: a review," *JAMA* **317**(24), 2532 (2017).
3. T. Graif et al., "Under diagnosis and over diagnosis of prostate cancer," *J. Urol.* **178**(1), 88–92 (2007).
4. G. Draisma et al., "Lead time and overdiagnosis in prostate-specific antigen screening: importance of methods and context," *J. Natl. Cancer Inst.* **101**(6), 374–383 (2009).
5. A. M. Brown et al., "Recent advances in image-guided targeted prostate biopsy," *Abdom. Imaging* **40**(6), 1788–1799 (2015).
6. C. J. D. Wallis, M. A. Haider, and R. K. Nam, "Role of mpMRI of the prostate in screening for prostate cancer," *Transl. Androl. Urol.* **6**(3), 464–471 (2017).
7. A. B. Rosenkrantz et al., "Proposed adjustments to PI-RADS version 2 decision rules: impact on prostate cancer detection," *Radiology* **283**(1), 119–129 (2017).
8. M. A. Bjurlin et al., "Optimization of prostate biopsy: the role of magnetic resonance imaging targeted biopsy in detection, localization and risk assessment," *J. Urol.* **192**(3), 648–658 (2014).
9. M. M. Siddiqui et al., "Comparison of MR/ultrasound fusion-guided biopsy with ultrasound-guided biopsy for the diagnosis of prostate cancer," *JAMA* **313**(4), 390–397 (2015).
10. I. G. Schoots et al., "Magnetic resonance imaging-targeted biopsy may enhance the diagnostic accuracy of significant prostate cancer detection compared to standard transrectal ultrasound-guided biopsy: a systematic review and meta-analysis," *Eur. Urol.* **68**(3), 438–450 (2015).
11. J. C. Weinreb et al., "PI-RADS prostate imaging—reporting and data system: 2015, version 2," *Eur. Urol.* **69**(1), 16–40 (2016).
12. C. Zhao et al., "The efficiency of multiparametric magnetic resonance imaging (mpMRI) using PI-RADS version 2 in the diagnosis of clinically significant prostate cancer," *Clin. Imaging* **40**(5), 885–888 (2016).
13. B. G. Muller et al., "Prostate cancer: interobserver agreement and accuracy with the revised prostate imaging reporting and data system at multiparametric MR imaging," *Radiology* **277**(3), 741–750 (2015).
14. S. Borofsky et al., "What are we missing? False-negative cancers at multiparametric MR imaging of the prostate," *Radiology* **286**(1), 186–195 (2018).
15. J. D. Le et al., "Multifocality and prostate cancer detection by multiparametric magnetic resonance imaging: correlation with whole-mount histopathology," *Eur. Urol.* **67**(3), 569–576 (2015).
16. S. Wang et al., "Computer aided-diagnosis of prostate cancer on multiparametric MRI: a technical review of current research," *Biomed. Res. Int.* **2014**, 1–11 (2014).
17. V. Giannini et al., "A fully automatic computer aided diagnosis system for peripheral zone prostate cancer detection using multi-parametric magnetic resonance imaging," *Comput. Med. Imaging Graphics* **46**(Pt 2), 219–226 (2015).
18. Y. Jiang et al., "Potential of computer-aided diagnosis to reduce variability in radiologists' interpretations of mammograms depicting microcalcifications," *Radiology* **220**(3), 787–794 (2001).

19. J.-Z. Cheng et al., "Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans," *Sci. Rep.* **6**(1), 24454 (2016).
20. M. Muyoyeta et al., "The sensitivity and specificity of using a computer aided diagnosis program for automatically scoring chest x-rays of presumptive TB patients compared with Xpert MTB/RIF in Lusaka Zambia," *PLoS One* **9**(4), e93757 (2014).
21. N. Lay et al., "Detection of prostate cancer in multiparametric MRI using random forest with instance weighting," *J. Med. Imaging* **4**(2), 024506 (2017).
22. G. Litjens et al., "Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge," *Med. Image Anal.* **18**(2), 359–373 (2014).
23. A. Madabhushi et al., "Automated detection of prostatic adenocarcinoma from high-resolution ex vivo MRI," *IEEE Trans. Med. Imaging* **24**(12), 1611–1625 (2005).
24. C. M. Stonnington et al., "Interpreting scan data acquired from multiple scanners: a study with Alzheimer's disease," *Neuroimage* **39**(3), 1180–1185 (2008).
25. R. T. Shinohara et al., "Statistical normalization techniques for magnetic resonance imaging," *Neuroimage Clin.* **6**, 9–19 (2014).
26. R. Herzog et al., "Variability in diagnostic error rates of 10 MRI centers performing lumbar spine MRI examinations on the same patient within a 3-week period," *Spine J.* **17**(4), 554–561 (2017).
27. T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York (2009).
28. H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique," *IEEE Trans. Med. Imaging* **35**(5), 1153–1159 (2016).
29. M. D. Greer et al., "Computer-aided diagnosis prior to conventional interpretation of prostate mpMRI: an international multi-reader study," *Eur. Radiol.* **28**(10), 4407–4417 (2018).
30. M. H. Le et al., "Automated diagnosis of prostate cancer in multiparametric MRI based on multimodal convolutional neural networks," *Phys. Med. Biol.* **62**(16), 6497–6514 (2017).
31. S. Liu et al., "Prostate cancer diagnosis using deep learning with 3D multiparametric MRI," *Proc. SPIE* **10134**, 1013428 (2017).
32. Y. Song et al., "Computer-aided diagnosis of prostate cancer using a deep convolutional neural network from multiparametric MRI," *J. Magn. Reson. Imaging* **48**, 1570–1577 (2018).
33. Y. Tschay et al., "Biopsy-guided learning with deep convolutional neural networks for prostate cancer detection on multiparametric MRI," in *IEEE 14th Int. Symp. on Biomedical Imaging (ISBI)*, pp. 642–645 (2017).
34. S. Xie and Z. Tu, "Holistically-nested edge detection," *Int. J. Comput. Vision* **125**(1–3), 3–18 (2017).
35. A. P. Harrison et al., "Progressive and multi-path holistically nested neural networks for pathological lung segmentation from CT images," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, Springer, Cham, pp. 621–629 (2017).
36. H. R. Roth et al., "Spatial aggregation of holistically-nested networks for automated pancreas segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, Springer, Cham, pp. 451–459 (2016).
37. I. Nogues et al., "automatic lymph node cluster segmentation using holistically-nested neural networks and structured optimization in CT images," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, Springer, Cham, pp. 388–397 (2016).
38. N. Lay et al., "A decomposable model for the detection of prostate cancer in multi-parametric MRI," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI*, Granada, Spain (2018).
39. S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proc. of the 32nd Int. Conf. Machine Learning*, JMLR.org, Lille, France, Vol. 37, pp. 448–456 (2015).
40. M. B. Blaschko and C. H. Lampert, "Object localization with global and local context kernels," in *British Machine Vision Conf.*, pp. 63.1–63.11 (2009).
41. W. Luo et al., "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. of the 30th Int. Conf. on Neural Information Processing Systems*, Curran Associates Inc., Barcelona, Spain, pp. 4905–4913 (2016).
42. J. Gawlitza et al., "Impact of the use of an endorectal coil for 3 T prostate MRI on image quality and cancer detection rate," *Sci. Rep.* **7**, 40640 (2017).
43. H. Gudbjartsson and S. Patz, "The Rician distribution of noisy MRI data," *Magn. Reson. Med.* **34**(6), 910–914 (1995).
44. K. He et al., "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in *IEEE Int. Conf. on Computer Vision (ICCV)*, Santiago, Chile, pp. 1026–1034, (2015).
45. I. Cima et al., "Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer," *Proc. Natl. Acad. Sci. U. S. A.* **108**(8), 3342–3347 (2011).
46. N. Petrick et al., "CT colonography with computer-aided detection as a second reader: observer performance study," *Radiology* **246**(1), 148–156 (2008).
47. B. Turkbey et al., "Comparison of endorectal coil and nonendorectal coil T2W and diffusion-weighted MRI at 3 Tesla for localizing prostate cancer: correlation with whole-mount histopathology," *J. Magn. Reson. Imaging* **39**(6), 1443–1448 (2014).
48. A. D. J. Baur et al., "T2- and diffusion-weighted magnetic resonance imaging at 3 T for the detection of prostate cancer with and without endorectal coil: an intraindividual comparison of image quality and diagnostic performance," *Eur. J. Radiol.* **85**(6), 1075–1084 (2016).
49. Z. Shaikhibrahim et al., "The peripheral zone of the prostate is more prone to tumor development than the transitional zone: is the ETS family the key?" *Mol. Med. Rep.* **5**(2), 313–316 (2012).
50. S. H. Lee et al., "Nonvisible tumors on multiparametric magnetic resonance imaging does not predict low-risk prostate cancer," *Prostate Int.* **3**(4), 127–131 (2015).
51. S. S. Dianat et al., "Magnetic resonance-invisible versus magnetic resonance-visible prostate cancer in active surveillance: a preliminary report on disease outcomes," *Urology* **85**(1), 147–154 (2015).
52. D. P. Kingma and L. J. Ba, "Adam: a method for stochastic optimization," in *ICLR*, San Diego (2015).
53. Y. Bengio, "Practical recommendations for gradient-based training of deep architectures," *Lect. Notes Comput. Sci.* **7700**, 437–478, (2012).

Yohan Sumathipala is currently an MD candidate at the University of Wisconsin School of Medicine and Public Health. He has a master's in biotechnology from Johns Hopkins University. He has a strong interest in better understanding how artificial intelligence can improve clinical care and support medical decision making.

Baris Turkbey received his medical degree from Hacettepe University in Ankara, Turkey, in 2003. He completed his residency in diagnostic and interventional radiology at Hacettepe University. He joined the National Cancer Institute, NIH, in 2007. His main research areas are imaging of prostate cancer (multiparametric MRI, PET CT), biopsy and focal therapy of prostate cancer and image processing (artificial intelligence and decision support systems). He is a member of ACR PI-RADS Steering Committee.

Peter L. Choyke is a senior investigator and chief of the Molecular Imaging Program of the National Cancer Institute. He received his medical degree from Jefferson Medical College and trained at Yale and Penn. His research has focused on the development of novel targeted imaging and treatment agents, based on optical, MRI, and radionuclide techniques, for oncologic imaging. He has focused on cancers of the genitourinary tract including prostate, renal, bladder, and malignancies of the bone marrow.

Ronald M. Summers received his BA degree in physics and his MD and PhD degrees in medicine/anatomy and cell biology from the University of Pennsylvania. He is a tenured senior investigator and staff radiologist in the Radiology and Imaging Sciences Department at the NIH Clinical Center in Bethesda, Maryland. His awards include being named a fellow of the Society of Abdominal Radiologists, the Presidential Early Career Award for Scientists and Engineers, the NIH Director's award, and the NIH Clinical Center director's award. He is a member of the editorial boards of the *Journal of Medical Imaging*, *Radiology: Artificial Intelligence and Academic Radiology* and a past member of the editorial board of *Radiology*. He has coauthored over 400 articles and is a coinventor on 14 patents.

Biographies of the other authors are not available.