SCIENTIFIC REPORTS
nature research

OPEN

Prostate Cancer Detection using Deep Convolutional Neural Networks

Sunghwan Yoo[1], Isha Gujrathi[1], Masoom A. Haider[1,2,3,4] & Farzad Khalvati[1,2,3,5*]

Prostate cancer is one of the most common forms of cancer and the third leading cause of cancer death in North America. As an integrated part of computer-aided detection (CAD) tools, diffusion-weighted magnetic resonance imaging (DWI) has been intensively studied for accurate detection of prostate cancer. With deep convolutional neural networks (CNNs) significant success in computer vision tasks such as object detection and segmentation, different CNN architectures are increasingly investigated in medical imaging research community as promising solutions for designing more accurate CAD tools for cancer detection. In this work, we developed and implemented an automated CNN-based pipeline for detection of clinically significant prostate cancer (PCa) for a given axial DWI image and for each patient. DWI images of 427 patients were used as the dataset, which contained 175 patients with PCa and 252 patients without PCa. To measure the performance of the proposed pipeline, a test set of 108 (out of 427) patients were set aside and not used in the training phase. The proposed pipeline achieved area under the receiver operating characteristic curve (AUC) of 0.87 (95% Confidence Interval (CI): 0.84–0.90) and 0.84 (95% CI: 0.76–0.91) at slice level and patient level, respectively.

Prostate cancer is the most common form of cancer among males in the United States. In 2017, it was the third leading cause of death from cancer in men in the United States, with around 161,360 new cases which represented 19% of all new cancer cases and 26,730 deaths, which represented 8% of all cancer deaths[1]. Despite the fact that prostate cancer is the most common form of cancer, if detected in the early stages, the survival rates are high due to slow progression of the disease[1]. Therefore, effective monitoring and early detection are the key for improved patients' survival.

Currently, accepted clinical methods to diagnose clinically significant prostate cancer (PCa) are a combination of the prostate-specific antigen (PSA) test, digital rectal exam, trans-rectal ultrasound (TRUS), and magnetic resonance imaging (MRI). However, PSA screening leads to over-diagnosis, which leads to unnecessary expensive and painful needle biopsies and potential over-treatment[2]. Multiparametric MRI which relies heavily on diffusion-weighted imaging (DWI) has been increasingly becoming the standard of care for prostate cancer diagnosis in radiology settings where the area under the receiver operating characteristic curve (ROC) varies from 0.69 to 0.81 for radiologists detecting PCa[3]. A standardized approach to image interpretation called PI-RADS v2[4] has been developed for radiologists, however, there remain issues with inter-observer variability in the use of the PI-RADS scheme[5].

Machine learning (ML) is a branch of artificial intelligence (AI) that is based on the idea of the system learning a pattern from a large scale database by using probabilistic and statistical tools and making decisions or predictions on the new data[6–8]. In medical imaging field, computer-aided detection and diagnosis (CAD), which is a combination of imaging feature engineering and ML classification, has shown potential in assisting radiologists for accurate diagnosis, decreasing the diagnosis time and the cost of diagnosis. Traditional feature engineering methods are based on extracting quantitative imaging features such as texture, shape, volume, intensity, and various statistical features from imaging data followed by a ML classifier such as Support Vector Machines (SVM), Adaboost, and Decision Trees[9–14].

Deep learning methods have shown promising results in a variety of computer vision tasks such as segmentation, classification, and object-detection[15–17]. These methods consist of convolution layers that are able to

[1]Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON, Canada. [2]Institute of Medical Science, University of Toronto, Toronto, ON, Canada. [3]Department of Medical Imaging, University of Toronto, Toronto, ON, Canada. [4]Sunnybrook Research Institute, Toronto, ON, Canada. [5]Department of Mechanical and Industrial Engineering, University of Toronto, Toronto, ON, Canada. *email: farzad.khalvati@utoronto.ca

extract different features from low-level local features to high-level global features from input images. A fully connected layer at the end of the convolutional neural layers converts convoluted features into the probabilities of certain labels[15]. Different types of layers, such as batch normalization layer[18], which normalizes the input of a layer with a zero mean and a unit variant, and dropout layer[19], which is one of regularization techniques that ignores randomly selected nodes, have been shown to improve the performance of deep learning-based methods. Nevertheless, to achieve convincing performance, an optimal combinations and structures of the layers as well as precise fine-tuning of hyper-parameters are required[15,17,20]. This remains as one of the main challenges of deep learning-based methods when applied to different fields such as medical imaging.

With CNNs' promising results in computer vision field[15,21], the medical imaging research community has shifted their interest toward deep learning-based methods for designing CAD tools for cancer detection. As a widely used approach, most of proposed algorithms require user-drawn regions of interest (ROI) to classify these user-annotated ROIs to PCa lesions and non PCa lesions. Tsehay *et al.*[22] conducted a 3×3 pixel level analysis by 5 convolution layers deep VGGNet[20] inspired CNN with 196 patients. They fine-tuned their classifier by cross-validation method within the training set with 144 patients and achieved area under ROC curve (AUC) of 0.90 AUC on a separated test set of 52 patients. The result was based on 3×3 windows of pixels extracted from MRI slices of DWI, T2-weighted images (T2w), and b-value images of 2000s mm^{-2}.

Le *et al.*[23] conducted two dimensional (2D) ROI classification with combination of fused multimodal Residual Network (ResNet)[17] and the traditional handcrafted feature extraction method. They augmented the training dataset and used the test set for fine-tuning and evaluating their classifier. They achieved ROI-level (lesion-level) AUC of 0.91. Liu *et al.*[24] used VGGNet inspired 2D CNN classifier to classify each sample corresponding to a 32 $\times$ 32 ROI (lesion) centered around biopsy location using a dataset, which was part of ProstateX challenge competition ("SPIE-AAPM-NCI Prostate MR Classification Challenge")[25]. They separated the dataset of 341 patients into 3 sets, the training set with 199 patients for training, validation set with 30 patients for fine-tuning, and test set with 107 patients for evaluation, and applied data augmentation to all 3 sets. They used 4 different types of input images which were generated with different combinations of DWI, apparent diffusion coefficient map (ADC), K_{trans} from dynamic contrast enhanced magnetic resonance imaging (DCE-MRI), and T2w for their study. They achieved AUC of 0.84 with the augmented test test.

Mehrtash *et al.*[26] also used VGGNet inspired 9 convolution layers deep three dimensional (3D) CNN classifier to classify 3D PCa lesions vs. non PCa lesions with $32 \times 32 \times 12$ ROI using ADC, high b-value images, and K_{trans} (DCE-MRI) of ProstateX challenge dataset[25]. They separated the data set with 341 patients into training set with 201 patients and test set with 140 patients, and achieved lesion-level performance of 0.80 AUC on their test set. They applied cross-validation method within the augmented training set during training. As it will be discussed in Discussion section, the proposed method in this paper is superior compared to these ROI-based solutions in terms of robustness and applicability in clinical usage since it foregoes the need for manually or automatically generating ROIs.

Slice-level detection algorithms classify each MRI slice into with or without PCa tumors. Ishioka *et al.*[27] performed the slice-level analysis with 316 patients by U-Net[28] combined with ResNet. They created non-augmented training, validation, and test sets and achieved AUC of 0.79 on the test set, which included only 17 individual slices. The proposed algorithm in this paper performs slice-level detection as well using a much larger sample size with superior performance compared to that proposed by Ishioka *et al.*[27].

Patient-level algorithms classify patients into with and without PCa. It is generally a challenging task to merge ROI-based or slice-level results into patient-level results[22–24,26,27]. Wang *et al.*[29] compared the performance of deep learning-based methods to non-deep learning-based methods on the classification of PCa MRI slices vs non PCa MRI slices with 172 patients. They evaluated their VGGNet inspired 7 layers (5 convolution layers and 2 inner product layers) CNN classifier's performance based on cross-validation. First, they classified each slice of a given patient and then converted the slice-level results into patient-level results by a simple voting strategy and achieved the patient-level AUC of 0.84, positive prediction value (PPV) of 79%, and negative prediction value (NPV) of 77%. In this work, we achieved similar results with an independent test set and larger sample size.

In this paper, we propose an automated pipeline for two levels of PCa classification: slice level and patient level. For slice-level classification, we have proposed a stack of individually trained modified ResNet[17] CNNs. We have also proposed a novel approach to convert slice-level classification results into patient level using first-order statistical features extractor, a decision tree-based feature selector, and a Random Forest classifier[30,31]. For the robustness of the performance, we divided the dataset into three separate sets, the training, validation, and test sets, and ensured that the test set was never seen by the classifier during training and fine-tuning[6]. We also included all slices that contain prostate and did not limit the pipeline to slices that have been selected for biopsy. Our proposed pipeline's performance on the independent test set was superior and more robust compared to similar studies that proposed CAD tools for PCa detection using CNNs.

Methods

Data. A cohort of 427 consecutive patients with a PI-RADS score of 3 or higher who underwent biopsy were included. Out of 427 patients, 175 patients had clinically significant prostate cancer and 252 patients did not. A total of 5,832 2D slices of each DWI sequence (e.g., b0) which contained prostate gland were used as our dataset. We set the patient with Gleason score higher than or equal to 7 (International Society of Uropatholgists grade group (GG $>= 2$) as the patient with a clinically significant prostate cancer and patient with Gleason score lower than or equal to 6 (GG $= 1$) or with no cancer (GG $= 0$) as the patient without a clinically significant prostate cancer.

MRI Acquisition. The DWI data was acquired between January 2014 to July 2017 using a Philips Achieva 3T whole body unit MR imaging scanner. The transverse plane of DWI sequences was obtained using a single-slot spin-echo echo-planar imaging sequence with four b values (0, 100, 400, and 1000s mm^{-2}), repetition time (TR)

Data Set	Patients with PCa	Patients without PCa	Slices with PCa tumors	Slices without PCa tumors
Training Set	105	166	439	3,253
Validation Set	18	30	66	588
Test Set	52	56	226	1,260

Table 1. Number of patients and slices with and without PCa for training, validation, and test sets.

5000~7000 ms, echo time (TE) 61ms, slice thickness 3mm, field of view (FOV) 240 mm × 240 mm and matrix of 140 × 140.

DWI is an MRI sequence which measures the sensitivity of tissue to Brownian motion and it has been found to be a promising imaging technique for PCa detection[32]. The DWI image is usually generated with different b values (0, 100, 400, and 1000s mm^{-2}) which generates various signal intensities representing the amount of water diffusion in the tissue and can be used to estimate ADC and compute high b-value images (b1600)[33].

In order to use DWI images as input to our deep learning network, we resized all of the DWI slices into 144 × 144 pixels, and center cropped them with 66 × 66 pixels such that the prostate was covered. The CNNs were modified to feed DWI data with 6 channels (ADC, b0, b100, b400, b1000, and b1600) instead of images with 3 channels (red, green and blue.)

Training, validation, and test sets. We separated 427 patients DWI images into three different sets, the training set with 271 patients (3,692 slices), the validation set with 48 patients (654 slices), and the test set with 108 patients (1,486 slices) where the training/validation/test ratio was 64%, 11%, 25%. The separation procedure of the dataset was as follows. First, we separated the dataset into two sets, the training/validation set as 75% and the test set as 25% to maintain a reasonable sample size for the test set. Second, we separated the training/validation set into two sets with training set as 85% of training/validation set and the validation set as 15% of training/validation set (Table 1). The ratios between the PCa patients and non PCa patients were kept roughly similar throughout the data sets.

Data preprocessing. All of DWI images in the dataset were normalized across the entire dataset using the following function.

$$X_{i\_normalized} = \frac{X_i - \mu}{std} \tag{1}$$

where X_i is the pixels in an individual MRI slice, μ is the mean of the dataset, std is the standard deviation of the dataset, and $X_{i\_normalized}$ is the normalized individual MRI slice.

Pipeline. The proposed pipeline consists of three stages. In the first stage, each DWI slice is classified using five individually trained CNNs models. In the second stage, first-order statistical features (e.g., mean, standard deviation, median, etc.) are extracted from the probability sets of CNNs outputs, and important features are selected through a decision tree-based feature selector. In the last stage, a Random Forest classifier is used to classify patients into groups with and without PCa using these first order statistical features. The Random Forest classifier was trained and fine-tuned by the features extracted from the validation set with 10 fold cross-validation method. Figure 1 shows the block diagram of the proposed pipeline.

ResNet. Since ResNet architecture has shown promising performance in multiple computer vision tasks[17], we chose it as our base architecture for this research. Each Residual Block consists of convolutional layers[21] and identity shortcut connection[17] that skips those layers, and their outcomes are added at the end, as shown in Figure 2-a. When input and output dimensions are the same, the identity shortcuts, denoted by x, can be directly applied. The following formula shows the identity mapping process.

$$y = F(x, \{W_i\}) + x \tag{2}$$

where $F(x, W_i)$ is the output from convolutional layers and x is the input. When the dimension of input is not the same as that of the output (e.g., at the end of the Residual Block), the linear projection W_s changes the dimension of the input to be same as that of the output which is defined as:

$$Y = F(x, \{W_i\}) + W_s x. \tag{3}$$

To improve the performance of the architecture, we implemented a fully pre-activated residual network[34]. In the original ResNet, batch normalization and ReLU activation layers were followed after the convolution layer, but in pre-activation ResNet, batch normalization and ReLU activation layers comes before the convolution layers. The advantage of this structure is that the gradient of a layer does not vanish even when the weights are arbitrarily small[34]. Instead of 2-layer deep ResNet block, we implemented a 3-layer deep "bottleneck" building block since it significantly reduces training time without sacrificing the performance[17] (Figure 2-b).

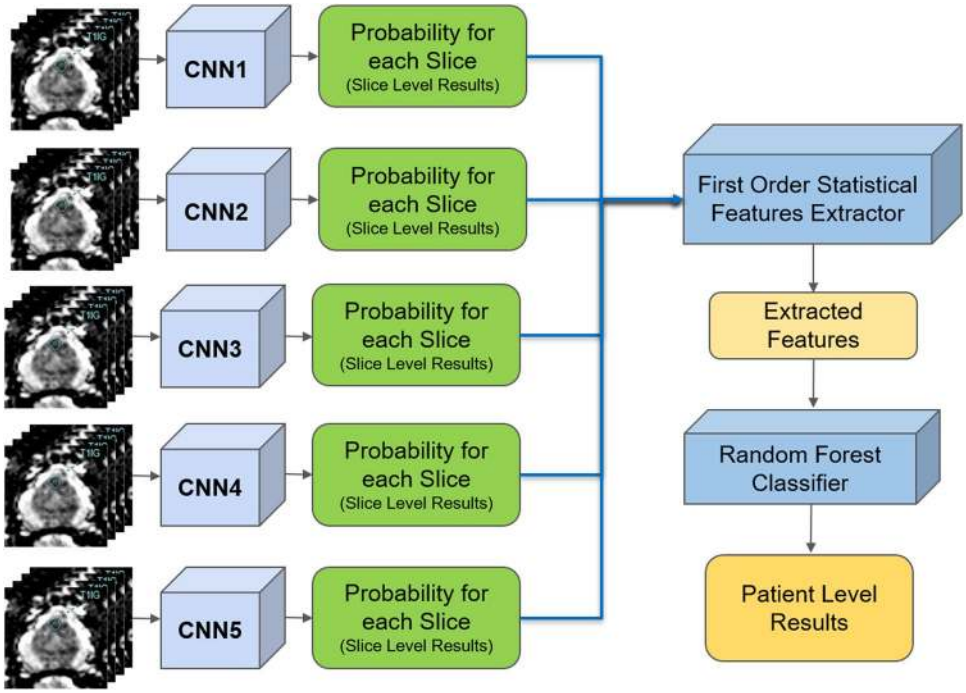

Figure 1. Block diagram of the proposed pipeline for prostate cancer detection. The inputs to each CNN are $66 \times 66 \times 6$ (ADC, b0, b100, b400, b1000, b1600) MRI slices. The output is the slice level and patient level results.

CNNs architecture and training. A 41 layers deep ResNet was created for the slice-level classification. The architecture is composed of 2D convolutional layers with a 7×7 filter followed by a 3×3 Max pooling layer and residual blocks (Res Block). The depth of 41 layers were found to be optimal through hyper-parameter fine-tuning procedure using the validation set. Since the input images were small (66×66 pixels) and the tumorous regions were even smaller (e.g., 4×3 pixels), additional ResNet blocks or deeper networks were needed. The first ResNet Block (ResNet Block1 in Table 2) is 3-layer bottleneck blocks with 2D CNN layers with filter sizes 64, 64 and 256 which is stacked 4 times. The second ResNet Block (ResNet Block2 in Table 2) is 3-layer bottleneck blocks with 2D CNN layers with filter sizes 128, 128, and 512 which is stacked 9 times. 2×2 2D Average Pooling, Dropout layer, and 2D Fully connected Layer with 1000 nodes for two probabilistic outputs are followed by the end of Res Blocks. Table 2 shows the overview of the proposed CNNs architecture.

Stochastic Gradient Decent[35] was used as the optimizer with the initial learning rate of 0.001, and it was reduced by a factor of 10 when the model stopped improving after iterations. The model was trained with the batch size set to 8. Dropout rate was set to 0.90. We used a weight decay of 0.000001 and a momentum of 0.90. Since the dataset is extremely unbalanced, binary cross entropy[36] was used as the loss function.

Stacked generalization. Due to the randomness in training CNNs (for instance, at the beginning of training CNNs, weights are set to arbitrary random numbers), each CNN may be different despite identical set of hyper-parameters and input datasets. This means each CNN may capture different features for the patient-level classification. Stacked generalization[37] is an ensemble technique that trains multiple classifiers with the same dataset and makes a final prediction using a combination of individual classifiers' predictions. Stacked generalization typically yields better classification performance compared to a single classifier[37]. We implemented a simple stacked generalization method using five CNNs. The number of stacked CNNs was selected based on the best performance and increasing the number of CNNs did not show improvement on the patient-level performance. Since there is a limited sample size for patient level (48 patients for validation, which was used to train Random Forest classifier for patient-level detection), increasing the number of CNNs, which leads to an increased number of patient-level features (as discussed in the next section), increases the likelihood of overfitting and hence, decreases the model's robustness[38]. All the slice-level probabilities generated by the five CNNs were fed into a first-order statistical features extractor to generate one set of features for each patient. In the proposed pipeline, the patient-level performance significantly improved (2-tailed $P = 0.048$) using five CNNs compared to a single CNN (AUC: 0.84, CI: 0.76–0.91, vs. AUC: 0.71, CI: 0.61–0.81).

First order statistical feature extraction. Let p_{ij} and n_{ij} be the probabilities of a MRI slice associated with PCa and non PCa, respectively, where i represents one of five individually trained CNNs and j represents each MRI slice of a patient. Each CNN produces two probability sets, $P_i = \{p_{i1}, \ldots, p_{iN}\}$ and $N_i = \{n_{i1}, \ldots, n_{iN}\}$ where N is the total number of MRI slices for each patient. Within the probability sets, top five probabilities which are higher than 0.74 were selected ($\acute{P}_i$ and $\acute{N}_i$). This was done to ensure less relevant probabilities at slice level were not used for

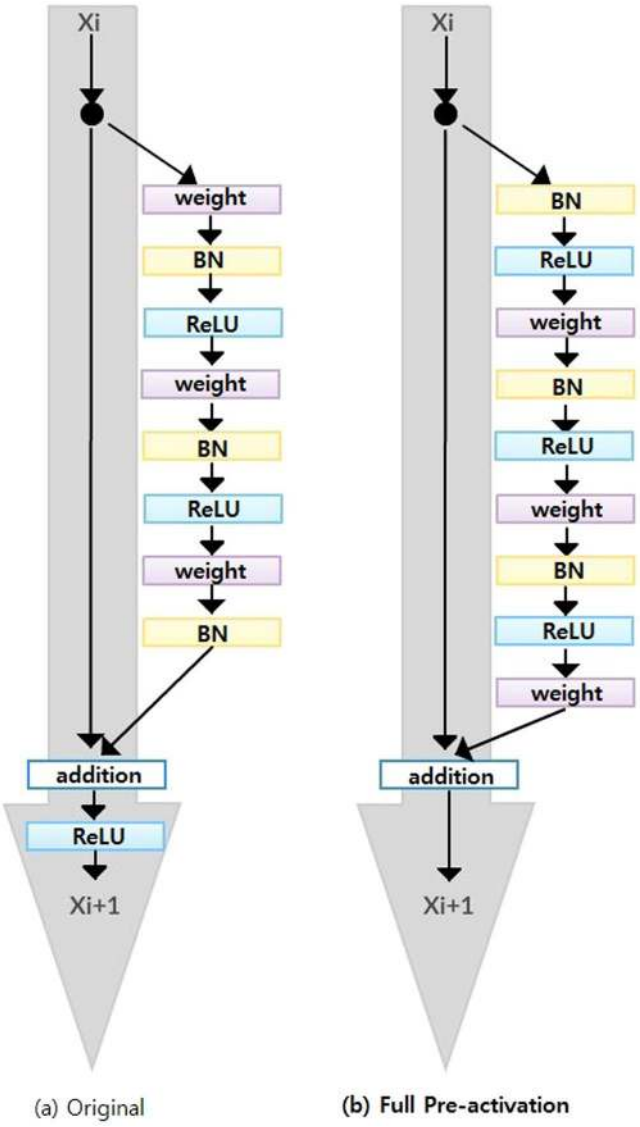

Figure 2. The structural difference between original residual network and fully pre-activated residual network.

Layer Name	Details about the layer
Conv layer	2D Convolutional Layer (7 × 7, 64, stride 2)
Max Pool	3 × 3 max pool, stride
ResNet Block 1	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 4$
ResNet Block 2	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 9$
Ave Pool	2D Average Pooling (7 × 7)
FC	Fully Connected Layer (2D, softmax)

Table 2. The Architecutre of the proposed CNNs.

patient-level classification. The probability cutoff of 0.74 was selected by grid-search using the validation set. Next, from the new probability sets, $\acute{P}_i$ and $\acute{N}_i$, the first-order statistical features set, $F_i = \{f_{i1}, \ldots f_{iK}\}$ where K represents the total number of statistical features, were extracted for each patient. Next, the important features, $\acute{F}_i$ were selected by a decision tree-based feature selector[39]. The final feature set was constructed by combining important features, $\acute{F}_i$, for all five CNNs where $F = \{\acute{F}_1, \ldots \acute{F}_5\}$.

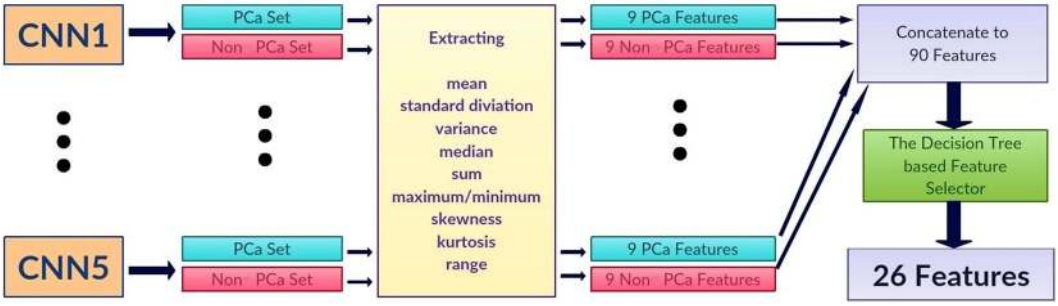

Figure 3. Block diagram of the proposed first-order statistical feature extractor. PCa Set: probabilistic output set from each CNN which is associated with PCa class. Non PCa Set: probabilistic output set from each CNN which is associated with non PCa class.

Architecture	Test AUC (95 % CI)
CNN1	**0.87** (0.84–0.90)
CNN2	0.87 (0.84–0.90)
CNN3	0.86 (0.83–0.89)
CNN4	0.85 (0.82–0.88)
CNN5	0.85 (0.82–0.88)

Table 3. Slice-level performances of five individually trained CNNs.

We extracted nine first-order features which are the mean, standard deviation, variance, median, sum, minimum (only from non PCa class), maximum (only from PCa class), skewness[40], kurtosis[40], and range from the minimum to maximum from each probability set. This produced 90 features for each patient (9 features for PCa and 9 features for non PCa class for each CNN). We selected 26 best features using the decision tree-based feature selector[39]. The decision tree based-feature selector was fine-tuned and trained with 10 fold cross-validation method using the validation set (Fig. 3).

Once first-order statistical features were extracted for each patient, a Random Forest classifier[30,31] was trained using the validation set and tested on the test set for patient-level classification.

Computational time. The CNNs were trained using one Nvidia Titan X GPU, 8 cores Intel i7 CPU and 32 GB memory. It took 6 hours to train all five CNNs with up to 100 iterations, less than 10 seconds to train the Random Forest classifier, and less than 1 minute to test all 108 patients.

Ethics approval and consent to participate. The Sunnybrook Health Sciences Centre Research Ethics Boards approved this retrospective single institution study and waived the requirement for informed consent.

Results

The AUC and ROC curve[41] were used to evaluate the performance of the proposed pipeline. A ROC curve is a commonly used method to visualize the performance of a binary classifier by plotting true positive rates and false positive rates with different thresholds, and an AUC summarizes its performance with a single number. The great advantage of AUC is its validity in an unbalanced dataset. Since only a small number of DWI slices have PCa tumor (e.g., average of 1 to 3 slices per patient where the total number of slices are an average of 14), AUC is the best way to evaluate the performance of the pipeline. In addition, ROC curve allows us to pick desired specificity and/or sensitivity of the classifier through the threshold. This evaluation method is applied to slice-level and patient-level classifications using the test set with 108 patients (1,486 slices).

Slice-level performance. Since the pipeline contains five individually trained CNNs, there are five different test results at slice level. Table 3 shows individual performance on the test set for each CNN. Our best CNN (CNN1) achieved the DWI slice-level AUC of 0.87 (95% Confidence Interval (CI): 0.84–0.90). Figure 4 shows the ROC curve of CNN1 performance.

Patient-level performance. The patient-level AUC by our Random Forest classifier with the features extracted through CNNs was 0.84 (95% CI: 0.76–0.91) (Fig. 5).

Discussion

In the literature, several PCa classification methods for MRI images have been developed to address the inherent challenges of CAD tools for cancer detection, which can be categorized into two classes: radiomics-driven feature-based methods[9–12,42,43] and deep learning-based methods[22–24,26,27,29].

Radiomics-driven feature-based methods consist of two stages: extraction of hand crafted features and classification based on these features. These methods require a comprehensive set of radiomic features, which include

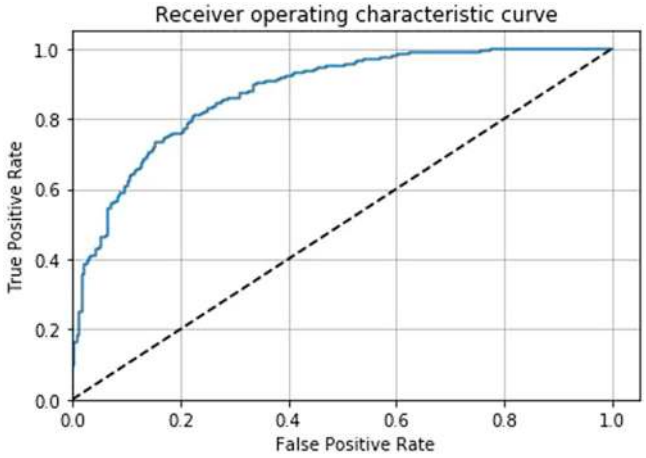

Figure 4. Slice-level ROC curve of the proposed ResNet inspired deep learning architecture (AUC: 0.87, CI: 0.84–0.90).

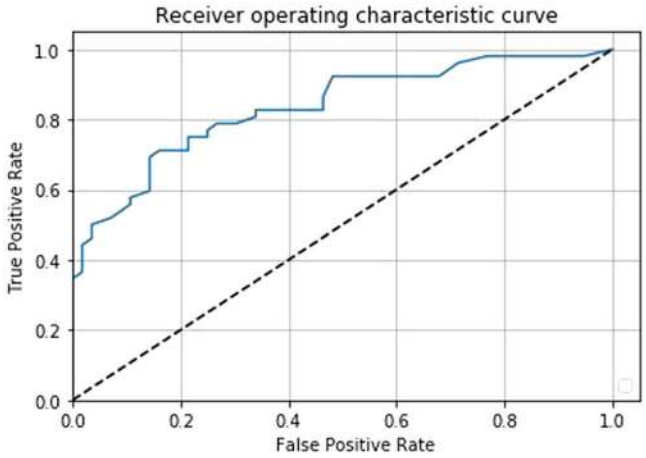

Figure 5. Patient-level ROC curve of the proposed pipeline: Random Forest classifier trained on the features extracted by the CNNs (AUC: 0.84, CI: 0.76–0.91).

first- and second-order statistical features, high-level features such as morphological features[9], and voxel-level features[10]. For the classification using radiomic features, several approaches have been proposed. Different machine learning classifiers[6] such as naive Bayesian classifier[9], SVM[42,43], and Random Forest classifier[10] have been used. However, it has been shown that deep learning methods are superior to radiomics-driven feature-based methods in classification of PCa[29].

ROI is one of commonly used data structures in medical image analysis. Usually delineated by the user, ROIs are samples within medical images identified for a particular purpose[44], which often contain cancer tumors. ROI-based methods directly compare and only classify regions or bounding boxes that contains tumors over healthy tissues. ROI-based methods have been used in both radiomics-driven and CNN-based methods for PCa CAD tool design. In CNN-based methods, Liu *et al*[24], Tsehay *et al*[22], and Le *et al*.[23] used 2D ROIs of cancer tumors and Mehrtash *et al*[26] and used 3D ROIs of cancer tumors as their data structures (eg. $32 \times 32 \times 12$ ROI).

ROI-based CAD algorithms have several limitations. First, ROI-based algorithms require a time consuming manually generated (by expert reader) or automatically generated segmentation of ROI as a part of the pipeline to generate ROI-based dataset. If it is a manually generated segmentation, the application for clinical use is limited because it ultimately relies on the clinician's review and expertise and hence, it is not fully automated. If, on the other hand, it is an automatically generated segmentation, the result of classification depends on the performance of the segmentation algorithm, and inaccuracies from ROI segmentation algorithm can lead to poor PCa detection performance. Moreover, most of ROI-based methods use sliding windows of pixels as data structures to feed the CNNs, which makes it a challenging task to achieve an acceptable performance on classification of PCa at patient level due to the fact that each patient's MRI data constitutes several thousands of windows of pixels. Therefore, ROI-based methods[22,24,26] struggle to merge individual ROI-based results into patient-level classification and they usually rely on basic merging methods such as simple voting[29], which makes it a challenging task to achieve acceptable performance at patient level.

In this work, instead of feeding ROIs into CNNs, we used automatically center-cropped DWI images with the only user intervention being to indicate the first and last slice that contained prostate gland. This is advantageous because it does not require generation of ROIs by either hand or segmentation algorithms. Thus, the proposed pipeline performance is independent of ROI generation method. In other words, our pipeline is able to perform PCa diagnosis on patients without the aid of expert readers. A similar approach was taken by Liu *et al.*[24] where 32 $\times$ 32 ROIs were constructed around biopsy locations. The main difference between this approach and ours is that in the former, only slices with biopsy were used and the remaining of slices, which are the majority of them, were excluded from the model. In our pipeline, we built the ROI around the prostate with no a priori knowledge on the biopsy locations, which makes our approach independent of radiologists. Although the result for Liu *et al.* approach[24] was reported for augmented test set and only for slices with biopsy, our pipeline AUC was superior (AUC of 0.87 vs. 0.84).

There are other studies in the literature that proposed slice-based analysis[27,29], but our slice-level performance (AUC: 0.87) and the sample size of the test set (108 patients or 1,486 slices), were significantly superior to their performance and sample size. For example, Ishioka *et al.*[27] proposed a slice-level algorithm using 316 patient data for training and validation. The test set was only 17 slices withe AUC of 0.79. Furthermore, we used the results generated by CNNs as features for classifying PCa at patient level, which was not the case with these previous works on slice-level algorithms.

Completely isolating test data from training and validation is crucial to measure true performance of a (deep) machine learning-based classifier. Cross-validation is a well-known method to evaluate the performance of the classifier[29]. However, it is only relevant for optimizing or fine-tuning the model because there is a possibility that cross-validation leads to a model that overfits. Fine-tuning the classifier based on the performance of the test set (e.g., adopted in[23]) makes the test set not independent from the trained and optimized classifier, and hence, the performance achieved is optimistic and not realistic. The fine-tuning and optimization of the model must be done through a validation set, which is separate than both training and test sets as adopted in our work in this paper and those of[22,26,27]. Moreover, the test set should not be augmented (e.g., adopted in[24]) to keep the robustness of the results. Due to test data cross-contamination with training or validation sets via cross-validation or data augmentation, the performance of some of the proposed models in the literature is rather optimistic.

In this work, we divided the entire dataset into three different sets, training, validation, and test set. In the slice-level analysis, the training set was used to train the model, and the validation set was used to fine-tune and optimize our CNNs architecture, and the test set was used to evaluate the performance of the CNNs. In the patient-level analysis, cross validation was used within the validation set to fine-tune and optimize our decision tree-based feature selector and Random Forest classifier, and tested on the test set. As a result, our classifier's results were more robust compared to studies that used cross-validation as a measure of performance[29] or training deep learning classifier without the validation set[23]. For the studies that used independent test set[24,26,27], our results are superior. For example, Liu *et al.*[24] conducted 2D ROI slice-level analysis and achieved 0.84 AUC for ROC-based (centered around biopsy location) classification only compared to 0.87 AUC for our proposed pipeline for slice-level classification.

Turning ROI-level results or the slice-level results of MRI data into patient-level result has been a major challenge in PCa classification via deep learning[22–24,26,27]. This is due to the fact that the 3D MRI volume of each patient may have hundreds or thousands of ROIs. Wang *et al.*[29] converted their slice-level result into patient-level by averaging all of the slice-level probabilities for patient and thresholding the average probabilities to classify PCa at patient level. Although this method achieved patient-level performance similar to our proposed pipeline's results (AUC: 0.84), it is based on cross validation, which makes it an optimistic result. In contrast, the results presented in this paper is based on a test set which is completely separate than the training and validation sets. Moreover, our test data contained 108 patients, which is significantly larger than the dataset with 17 patients for each fold[29].

The main limitation of this work is the fact that similar to CAD papers, the data is inherently biased; those patients are sent to MRI who have an indication of prostate cancer (e.g., higher PSA). Thus, the dataset is not a true reflection of the population. In addition, the labels for the data are based on biopsy locations, which are determined by radiologists. In other words, slices with no biopsy are assumed to be negative, based on radiology reports. However, the positive slices are based on pathology (biopsy) reports. Finally, an external validation with a dataset from a different institute is required to verify the performance and robustness of the proposed pipeline across scanners and institutions.

Conclusion

In this work, we built a two step automated deep learning pipeline for slice-level and patient-level PCa diagnosis using DWI images. Instead of manual ROI annotation, automated center-cropping was used to maintain independence from expert readers' intervention. A stack of five CNNs were used to produce improved classification results at slice level. First-order statistical features were extracted from slice-level probabilities to compile slice-level classification results into patient level. The pipeline was tested on an independent test set of 108 patients and the results at both slice level and patient level was superior to the state-of-the-art. As future work, other CNN architectures such as 3D CNNs (to feed the 3D DWI) and recurrent neural networks[45] (to account for sequentiality of lesions in neighboring slices) will be used.

Data availability

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request pending the approval of the institution(s) and trial/study investigators who contributed to the dataset.

Received: 8 November 2018; Accepted: 2 December 2019;
Published online: 20 December 2019

References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2017. *CA: a cancer journal for clinicians* **67**, 7–30 (2017).
2. Sandhu, G. S. & Andriole, G. L. Overdiagnosis of prostate cancer. *Journal of the National Cancer Institute Monographs* **2012**, 146–151 (2012).
3. Sonn, G.A.*et al.* Prostate magnetic resonance imaging interpretation varies substantially across radiologists. *Eur. urology focus* (2017).
4. Hassanzadeh, E.*et al.* Prostate Imaging Reporting and Data System Version 2 (PIRADS v2): A pictorial review. *Abdom Radiol* **42**, 278–289 https://doi.org/10.1016/j.trsl.2014.08.005. (2017).
5. Rosenkrantz, A. B. *et al.* Interobserver reproducibility of the pi-rads version 2 lexicon: a multicenter study of six experienced prostate radiologists. *Radiology* **280**, 793–804 (2016).
6. Nasrabadi, N. M. Pattern recognition and machine learning. *Journal of electronic imaging* **16**, 049901 (2007).
7. Goldberg, D. E. & Holland, J. H. Genetic algorithms and machine learning. *Machine learning* **3**, 95–99 (1988).
8. Michalski, R. S., Carbonell, J. G. & Mitchell, T. M. *Machine learning: An artificial intelligence approach* (Springer Science & Business Media, 2013).
9. Cameron, A., Khalvati, F., Haider, M. A. & Wong, A. Maps: a quantitative radiomics approach for prostate cancer detection. *IEEE Transactions on Biomed. Eng.* **63**, 1145–1156 (2016).
10. Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N. & Huisman, H. Computer-aided detection of prostate cancer in mri. *IEEE transactions on medical imaging* **33**, 1083–1092 (2014).
11. Wang, S., Burtt, K., Turkbey, B., Choyke, P. & Summers, R.M.Computer aided-diagnosis of prostate cancer on multiparametric mri: a technical review of current research. *BioMed research international* **2014** (2014).
12. Fehr, D. *et al.* Automatic classification of prostate cancer gleason scores from multiparametric magnetic resonance images. *Proc. of the Natl. Acad. of Sci.* **112**, E6265–E6273 (2015).
13. Erickson, B. J., Korfiatis, P., Akkus, Z. & Kline, T. L. Machine learning for medical imaging. *Radiographics* **37**, 505–515 (2017).
14. Orru, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G. & Mechelli, A. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci. & Biobehav. Rev.* **36**, 1140–1152 (2012).
15. Krizhevsky, A., Sutskever, I. & Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097– 1105 (2012).
16. Long, J., Shelhamer, E. & Darrell, T.Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431– 3440 (2015).
17. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770– 778 (2016).
18. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).
19. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436 (2015).
20. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
21. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436 (2015).
22. Tsehay, Y.et al. Biopsy-guided learning with deep convolutional neural networks for prostate cancer detection on multiparametric mri. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, 642– 645 (IEEE, 2017).
23. Le, M. H. *et al.* Automated diagnosis of prostate cancer in multi-parametric mri based on multimodal convolutional neural networks. *Phy. in Medicine & Bio.* **62**, 6497 (2017).
24. Liu, S., Zheng, H., Feng, Y. & Li, W. Prostate cancer diagnosis using deep learning with 3d multiparametric mri. In *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134, 1013428 (International Society for Optics and Photonics, 2017).
25. Armato, S. G., Petrick, N. A. & Drukker, K.Prostatex: Prostate mr classification challenge (conference presentation). *Proceedings of the SPIE, Volume 10134, id. 101344G 1 pp.(2017).***134** (2017).
26. Mehrtash, A.et al. Classification of clinical significance of mri prostate findings using 3d convolutional neural networks. In *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134, 101342A (International Society for Optics and Photonics, 2017).
27. Ishioka, J. *et al.* Mp20-10 deep learning with a convolutional neural network algorithm for fully automated detection of prostate cancer using pre-biopsy mri. *The Journal of Urology* **199**, e256 (2018).
28. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234– 241 (Springer, 2015).
29. Wang, X. *et al.* Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. *Sci. reports* **7**, 15415 (2017).
30. Breiman, L. Random forests. *Machine learning* **45**, 5–32 (2001).
31. Nguyen, C., Wang, Y. & Nguyen, H. N. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. *J. of Biomed. Sci. and Eng.* **6**, 551 (2013).
32. Padhani, A. R. *et al.* Diffusion-weighted magnetic resonance imaging as a cancer biomarker: consensus and recommendations. *Neoplasia* **11**, 102–125 (2009).
33. Glaister, J., Cameron, A., Wong, A. & Haider, M.A. Quantitative investigative analysis of tumour separability in the prostate gland using ultra-high b-value computed diffusion imaging. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, 420– 423 (IEEE, 2012).
34. He, K., Zhang, X., Ren, S. & Sun, J. Identity mappings in deep residual networks. In *European conference on computer vision*, 630–645 (Springer, 2016).
35. Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, 177– 186 (Springer, 2010).
36. De Boer, P.-T., Kroese, D. P., Mannor, S. & Rubinstein, R. Y. A tutorial on the cross-entropy method. *Annals of operations research* **134**, 19–67 (2005).
37. Wolpert, D. H. Stacked generalization. *Neural networks* **5**, 241–259 (1992).
38. Vapnik, V.*The nature of statistical learning theory* (Springer science and business media, 2013).
39. Saeys, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in bioinformatics. *bioinformatics* **23**, 2507–2517 (2007).
40. Kim, H.-Y. Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. *Restorative dentistry & endodontics* **38**, 52–54 (2013).
41. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* **143**, 29–36 (1982).
42. Khalvati, F., Wong, A. & Haider, M. A. Automated prostate cancer detection via comprehensive multi-parametric magnetic resonance imaging texture feature models. *BMC medical imaging* **15**, 27 (2015).
43. Khalvati, F. *et al.* Mpcad: a multi-scale radiomics-driven framework for automated prostate cancer localization and detection. *BMC medical imaging* **18**, 16 (2018).

44. Brinkmann, R.*The art and science of digital compositing: Techniques for visual effects, animation and motion graphics* (Morgan Kaufmann, 2008).
45. Mikolov, T., Karafiát, M., Burget, L., Černocky, J. & Khudanpur, S. Recurrent neural network based language model. In *Eleventh annual conference of the international speech communication association* (2010).

Acknowledgements
This study was conducted with the support of the Ontario Institute for Cancer Research (OICR) through funding provided by the Government of Ontario.

Author contributions
S.Y., M.A.H. and F.K. contributed to the design and implementation of the concept. S.Y., I.G., M.A.H. and F.K. contributed in collecting and reviewing the data. F.K. and M.A.H. are co-senior authors for this manuscript. All authors contributed to the writing and reviewing of the paper. All authors read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to F.K.

Reprints and permissions information is available at www.nature.com/reprints.

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