Medical Image Analysis 73 (2021) 102153

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Computer-aided diagnosis of prostate cancer using multiparametric MRI and clinical features: A patient-level classification framework

Pritesh Mehta[a,*], Michela Antonelli[b], Hashim U. Ahmed[c], Mark Emberton[d], Shonit Punwani[e], Sébastien Ourselin[b]

[a] Department of Medical Physics and Biomedical Engineering, University College London, UK
[b] Biomedical Engineering & Imaging Sciences School, King's College London, UK
[c] Imperial Prostate, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, UK
[d] Division of Surgery and Interventional Science, University College London, UK
[e] Centre for Medical Imaging, University College London, UK

ARTICLE INFO

Article history:
Received 24 July 2020
Revised 3 April 2021
Accepted 28 June 2021
Available online 29 June 2021

Keywords:
Computer-aided diagnosis
Convolutional neural network
Multiparametric magnetic resonance
Imaging
Prostate cancer
Prostate-specific antigen density
Support vector machine

ABSTRACT

Computer-aided diagnosis (CAD) of prostate cancer (PCa) using multiparametric magnetic resonance imaging (mpMRI) is actively being investigated as a means to provide clinical decision support to radiologists. Typically, these systems are trained using lesion annotations. However, lesion annotations are expensive to obtain and inadequate for characterizing certain tumor types e.g. diffuse tumors and MRI invisible tumors. In this work, we introduce a novel patient-level classification framework, denoted PCF, that is trained using patient-level labels only. In PCF, features are extracted from three-dimensional mpMRI and derived parameter maps using convolutional neural networks and subsequently, combined with clinical features by a multi-classifier support vector machine scheme. The output of PCF is a probability value that indicates whether a patient is harboring clinically significant PCa (Gleason score $\geq 3+4$) or not. PCF achieved mean area under the receiver operating characteristic curves of 0.79 and 0.86 on the PICTURE and PROSTATEx datasets respectively, using five-fold cross-validation. Clinical evaluation over a temporally separated PICTURE dataset cohort demonstrated comparable sensitivity and specificity to an experienced radiologist. We envision PCF finding most utility as a second reader during routine diagnosis or as a triage tool to identify low-risk patients who do not require a clinical read.

1. Introduction

Prostate cancer (PCa) is the second most frequently diagnosed cancer in men worldwide and the fifth leading cause of cancer death in men (Bray et al., 2018). Despite considerable research, the etiology of PCa is not yet well understood (Bray et al., 2018). While age, race, and family history have shown the strongest associations to PCa risk, no conclusive preventable risk factors have been identified. Therefore, work continues to identify methods for accurate early diagnosis and effective treatment of PCa.

The prostate-specific antigen (PSA) blood test is an established tool used by clinicians to determine which patients require confirmatory transrectal ultrasound-guided (TRUS) biopsy (Ahmed et al., 2017). However, PSA alone has a low specificity for PCa due to elevation under benign circumstances and TRUS biopsy misses clinically significant PCa (CSPCa). Typically, CSPCa refers to histopathologically defined Gleason score (GS) $\geq 3+4$ disease, which poses an increased mortality risk (Woo et al., 2018). On a large cohort of 576 men with raised PSA and no previous biopsy, the "Prostate MRI Imaging Study" (PROMIS) (Ahmed et al., 2017) found that TRUS biopsy missed CSPCa in 52% of disease positive patients. The deficiencies of the PSA blood test and TRUS biopsy can lead to incorrect diagnoses, inaccurate risk assessments, and suboptimal therapy choices (Hoeks et al., 2011).

Multiparametric magnetic resonance imaging (mpMRI) is increasingly being incorporated into the PCa diagnostic pathway to enable non-invasive cancer detection, targeted biopsy and targeted treatment planning (Wang et al., 2014). The most commonly collected sequences are those advocated by the revised "Prostate Imaging Reporting and Data System" (PI-RADS v2) (American College of Radiology, 2015) and the Likert assessment

* Corresponding author.
E-mail address: pritesh.mehta.17@ucl.ac.uk (P. Mehta).

https://doi.org/10.1016/j.media.2021.102153

system (Brizmohun Appayya et al., 2018); namely, T2-weighted imaging (T2WI), diffusion-weighted imaging (DWI), including apparent diffusion coefficient (ADC) maps, and dynamic contrast-enhanced imaging (DCEI). The PROMIS study found that mpMRI prior to biopsy can identify one quarter of men presenting with elevated PSA who might safely avoid biopsy, and that mpMRI followed by targeted biopsy can reduce the over-diagnosis of clinically insignificant cancer and improve the detection of clinically significant cancer. However, reading mpMRI requires a high level of expertise and is a time-consuming task (Wang et al., 2014).

Computer-aided diagnosis (CAD) systems that use mpMRI for PCa diagnosis are actively being investigated. Many PCa CAD works present systems that perform lesion classification by extracting features from radiologist delineated lesion contours or from radiologist defined lesion-centred patches (Antonelli et al., 2019; Bonekamp et al., 2018; Dinh et al., 2018; Woźnicki et al., 2020; Zhong et al., 2019). Antonelli et al. (2019) trained machine learning classifiers to classify prostate tumors into those with/without Gleason pattern 4, which is a difficult task for radiologists; they showed that machine learning classifiers trained using MRI radiomic features and clinical features outperformed radiologists on this task. Similarly, Woźnicki et al. (2020) showed that an ensemble machine learning classifier that combined radiomics, PI-RADS scores, and clinical features performed comparably to radiologists for differentiating CSPCa lesions from other lesion types. In Zhong et al. (2019), a convolutional neural network (CNN) classifier was shown to achieve a similar area under the receiver operating characteristic curve (ROC AUC) to PI-RADS v2 scoring, for classification of lesion-centered patches as CSPCa or not CSPCa. However, in clinical practice, the success of all aforementioned methods would depend heavily on the experience of the radiologist(s) who performs the initial lesion candidate selection. Other studies explored CAD assistance during the initial detection stage by way of producing voxel probability maps that can help radiologists detect CSPCa tumors (Greer et al., 2018; Giannini et al., 2017; Gaur et al., 2018; Thon et al., 2017; Zhu et al., 2020). Greer et al. (2018) showed that CAD-assisted mpMRI interpretation increased detection sensitivity and agreement between nine radiologists with varying levels of experience, while Gaur et al. (2018) found that CAD-assisted mpMRI interpretation improved the specificity of both moderately and highly experienced radiologists on an external multi-center validation cohort. Fewer works present fully automated solutions that both detect and classify lesions. The CAD system presented by Litjens et al. (2014a) extracted a combination of intensity, texture, shape, anatomy, and pharmacokinetic features from T2WI, DWI, and DCEI to generate a voxel probability map for each patient, followed by candidate selection, candidate feature extraction, and classification of each candidate using a random forest classifier. Patient-level ROC AUCs of 0.81 and 0.83 were obtained for PCa vs. normal/benign and CSPCa vs. normal/benign respectively, on a large cohort of 347 patients; their dataset is available for download from the PROSTATEx Challenges database (Litjens et al., 2017b). In Schelb et al. (2019), the authors compared the performance of experienced radiologists to a U-Net CNN optimized for detection and segmentation of CSPCa tumors using T2WI and DWI. Notably, they thresholded the probabilistic output of their system by picking an operating point that most closely matched PI-RADS v2 performance in the training set, in terms of both sensitivity and specificity. On a held-out testing set, they reported a per-patient sensitivity of 92% on 26 patients with CSPCa and specificity of 47% on 36 patients without CSPCa, which was similar to the sensitivity and specificity of PI-RADS v2 scoring for threshold ≥ 4.

The PCa CAD system works described above required lesion contour or lesion centroid ground-truth for training their systems. Producing either type of lesion annotation on mpMRI can be challenging and/or time-consuming for a number of reasons. First and foremost, producing lesions annotations on mpMRI, following prostatectomy or a MRI-blinded biopsy technique such as systematic biopsy, saturation biopsy, or transperineal template prostate mapping (TTPM) biopsy is not clinical routine and therefore, must be performed retrospectively (Cao et al., 2019). While lesion centroid annotations may be made prospectively ahead of targeted biopsy, targeted biopsy alone is not recommended as a reference standard (Simmons et al., 2014). Second, cognitive matching of biopsy or prostatectomy findings to mpMRI may be required, which is not trivial. Third, once location on mpMRI is determined, if a contour is sought, it will typically be drawn on T2WI (due to its high spatial resolution and superior tissue contrast), in-plane and on all other slices containing the lesion. Should registration issues arise between mpMRI modalities, lesion contours or lesion centroids may be required on the other modalities also. A further challenge is posed by diffuse non-focal tumors and MRI invisible tumors (Borofsky et al., 2017); it is unclear how these tumors should be annotated. Fourth and finally, to account for inter-observer variability (Steenbergen et al., 2015), lesion annotations should be made by more than one radiologist, which can increase the overall time taken to perform annotations multiplicatively. Due to the annotation difficulties described, CAD systems for PCa are typically trained on small, carefully prepared datasets.

In this work, we introduce a novel patient-level classification framework, denoted PCF, that is trained using patient-level labels only, therefore avoiding the need for lesion annotations. In PCF, feature vectors are extracted from three-dimensional T2WI, ADC map, computed high b-value DWI, and four semi-quantitative parameter maps extracted from DCEI using CNNs, where each CNN is a modified 3D ResNet architecture, proposed in this work. During the training phase of PCF, feature selection is applied to select the optimal subset of CNN feature vectors and available clinical features for patient classification. Subsequently, selected CNN feature vectors and clinical features are combined for classification using a two-level multi-classifier support vector machine (SVM) scheme. The output of PCF is a patient probability associated to the presence of CSPCa in the patient's prostate; here, CSPCa is defined as GS $\geq 3+4$ disease. Utilizing features extracted from the full-breadth of mpMRI and available clinical features in combination to enhance classification performance is in line with the guidance provided by the Likert assessment system for radiologists. The primary contribution of this work is the proposal of PCF for patient-level CSPCa classification, while a secondary contribution is our proposed method for DCEI analysis i.e. extraction of CNN features from constructed semi-quantitative volumetric DCEI parameter maps. We envision PCF being applied as a second reader during routine diagnosis or as a triage tool which can identify low risk patients that do not require a clinical read; both applications could help alleviate the workload of radiologists who are an increasingly stretched resource (The Royal College of Radiologists, 2018).

The paper is organized as follows: In Section 2, we describe the technical details of PCF. In Section 3, we introduce the two patient datasets used to evaluate the performance of PCF, the classification tasks performed, the evaluation measures used, and the experimental settings employed. Section 4 presents results for patient classification. In Section 5, we conclude by discussing the implications of our results and future work.

2. Methods

PCF is visualized schematically in Fig. 1. First, mpMRI and clinical features are pre-processed. This involves automated prostate region segmentation, calculation of high b-value DWI and semi-quantitative DCEI parameter maps, and finally, normalization/standardization of images, parameter maps, and clinical features. Second, CNN encoders are employed to extract feature

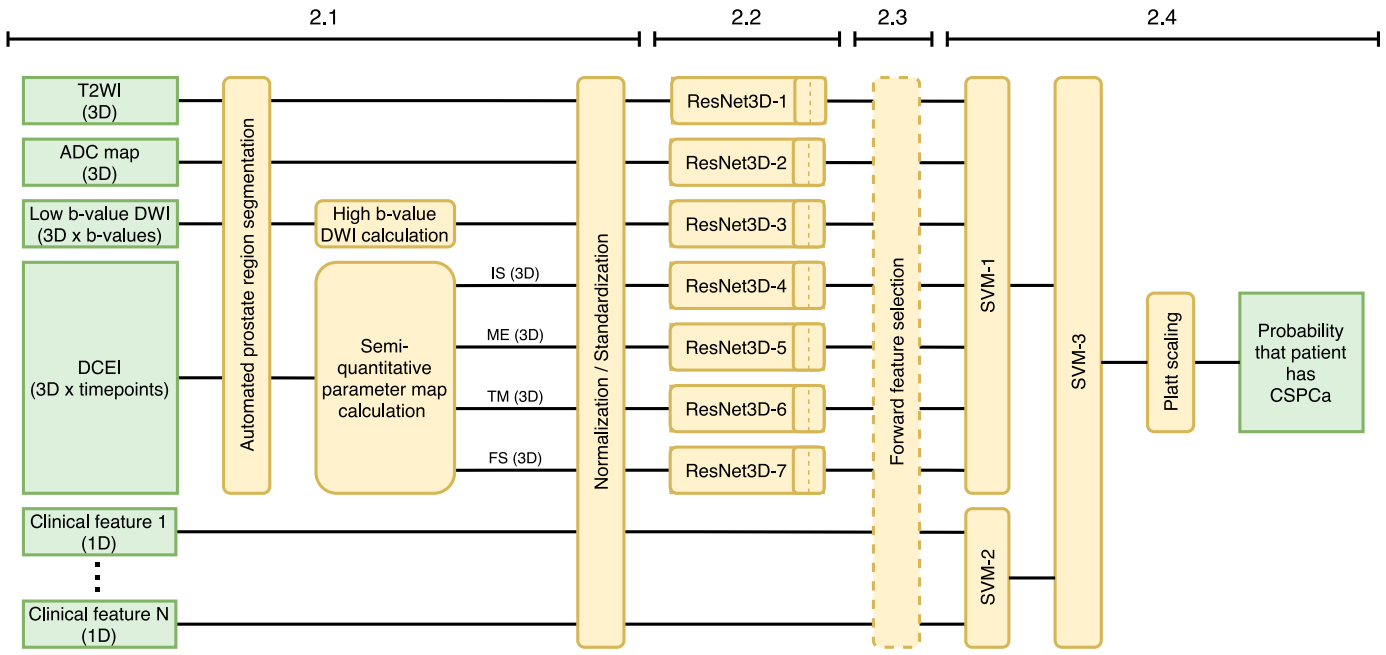


Fig. 1. Workflow of the proposed patient-level classification framework, PCF. Green rectangles indicate original scans or clinical features or the probabilistic classification outcome; yellow rectangles indicate processes applied to the data. In experiments, we refer to PCF with forward feature selection disabled as PCF-ALL and PCF with forward feature selection enabled as PCF-SEL.

vectors from three-dimensional MR images and parameter maps. Third, forward feature selection is used to select the CNN feature vectors and clinical features that are most pertinent for classification. Fourth and finally, a two-level SVM scheme is used to output patient probability of CSPCa. Each stage is described in more detail in the sections to follow, while experimental parameters used to collect results are described in Section 3.3.

2.1. Pre-processing

2.1.1. Automated prostate region segmentation

As a first step, the prostate is segmented on T2WI. Segmentation of the prostate creates a simpler classification task, unsullied by excess background information. In PCF, we use HighRes3DNet (Li et al., 2017) to segment the prostate on T2WI. HighRes3DNet is a high-resolution compact CNN for volumetric image segmentation. Given a three-dimensional T2WI, I_{T2WI}, HighRes3DNet outputs a prostate mask, S_{T2WI}, with the same spatial dimensions as I_{T2WI}. The prostate mask for DWI, S_{DWI}, is obtained by transforming S_{T2WI} from T2WI space into DWI space using a registration-driven transformation T such that $T : S_{T2WI} \rightarrow S_{DWI}$, which accounts for resolution differences between T2WI and DWI, as well as voluntary/involuntary patient movement between acquisitions, and distortions on DWI caused by air in the rectum (De Luca et al., 2011). Here, $T = T_{nrd} \circ T_{aff}$, where T_{aff} is the transformation given by the affine registration of I_{T2WI} to the three-dimensional ADC map, I_{ADC}, using the symmetric block-matching algorithm (Modat et al., 2014) and T_{nrd} is given by the subsequent non-rigid registration using the free-form deformation (FFD) algorithm (Modat et al., 2010). The convolution-based fast local normalized correlation coefficient (LNCC) (Cachier et al., 2003) is used as similarity measure for FFD to enable robustness to bias field inhomogeneity. The same approach is used to obtain the prostate mask in DCEI space, S_{DCEI}; in this case driven by the registration of T2WI to the first DCEI timepoint. The prostate masks, S_{T2WI}, S_{DWI}, and S_{DCEI} obtained for each patient are used to crop a sub-volume containing the prostate in corresponding imaging.

2.1.2. Computed high b-value DWI

High b-value images with b-value ≥ 1400 are a key component of mpMRI (American College of Radiology, 2015). Computed high b-value DWI has been shown to achieve superior image quality and lesion conspicuity than acquired high b-value DWI (Verma et al., 2016). In PCF, we compute high b-value DWI using a monoexponential model (Blackledge et al., 2011) for the per-voxel observed signal:

$$s(b) = s(0)\exp(-b \cdot \mathrm{ADC}). \tag{1}$$

Non-linear least squares is used to fit Eq. (1) to the observed points given by low b-value DWI intensities, giving per-voxel estimates of $s(0)$ and ADC: $s^*(0)$ and ADC^*. High b-value images are then computed using the equation:

$$s(b_c) = s^*(0)\exp(-b \cdot \mathrm{ADC}^*), \tag{2}$$

where b_c is the high b-value being extrapolated to.

2.1.3. Semi-quantitative DCEI parameter maps

Semi-quantitative analysis of DCEI has been shown to provide good discrimination between benign and malignant lesions (Zelhof et al., 2009), while avoiding the challenging estimation of the arterial input function needed for computing pharmacokinetic parameters (Haq et al., 2015). In PCF, semi-quantitative analysis of DCEI is used to compute parameter maps in an automated manner.

First, per-voxel signal intensity-versus-time curves are normalized to a standard pre-contrast level using a mean baseline computed from the first three signal values from T timepoints:

$$\hat{s}_t = \frac{s_t}{k}, \quad t = 1, \ldots, T, \quad k = \sum_{t=1}^{3} \frac{s_t}{3}. \tag{3}$$

Then, four voxel-wise variables are extracted: initial slope of enhancement (IS), maximum enhancement (ME), time to maximum enhancement (TM), and final slope (FS); originally defined in Zelhof et al. (2009) and Kubassova et al. (2007), but we are the first to use them to construct three-dimensional parameter maps

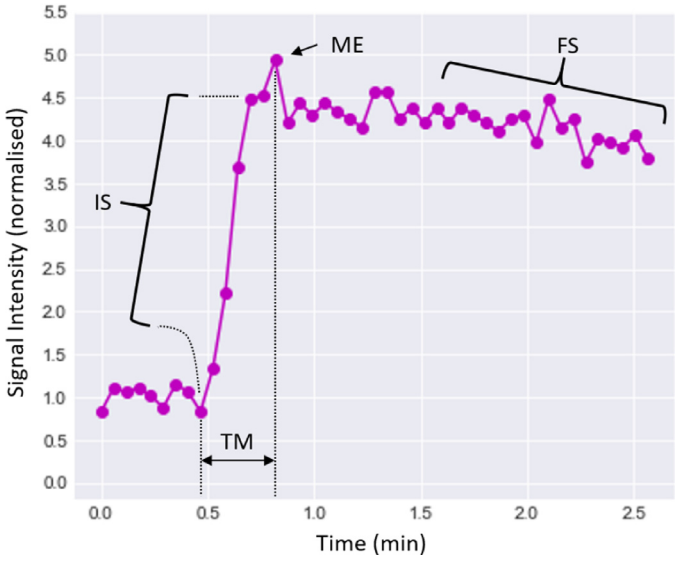


Fig. 2. Normalized signal intensity-versus-time curve corresponding to voxel at the center of a GS 3+4 lesion of a patient from the PROSTATEx dataset.

and in addition, to later extract spatial features from the parameter maps using CNNs. An illustration of the variables is given in Fig. 2.

IS is assumed to be the gradient of the steepest portion of the normalized signal intensity-versus-time curve. First, an averaging window of length l_{IS} is passed over the normalized signal $\{\hat{s}_t\}_{t=1}^{T}$ in steps of one timepoint as in Kubassova et al. (2007). The gradient of the linear best fit in each window is computed, giving the set of gradients $\{g_t\}_{t=1}^{T-l_{IS}+1}$. Subsequently, IS is computed by taking the maximum of the gradients:

$$IS = \max(\{g_t\}_{t=1}^{T-l_{IS}+1}), \tag{4}$$

where l_{IS} is determined empirically based on the temporal resolution of the DCEI.

ME is calculated as the maximum value of the normalized signal:

$$ME = \max(\{\hat{s}_t\}_{t=1}^{T}). \tag{5}$$

TM is calculated as the difference between onset time, denoted t_{OS}, and the time of ME, denoted t_{ME}, in minutes, where t_{OS} is defined as the first time point in the averaging window to which IS corresponds:

$$TM = t_{ME} - t_{OS}. \tag{6}$$

FS is defined as the gradient of the normalized signal over the wash-out phase of the contrast agent. Here, we compute it as the the gradient g_{FS} of the linear best fit over the final m_{FS} minutes of the normalized signal, where m_{FS} is determined empirically based on the length of the wash-out phase of the contrast agent.

2.1.4. Normalization/standardization

Histogram-based standardization is applied to the prostate region in each patient's T2WI to homogenize tissue intensities across patients in line with the work by Toivonen et al. (2019). Images are transformed by matching their intensity histograms to a mean histogram calculated using training data. The algorithm, including pseudo-code, is presented in Nyúl et al. (2000). A simple per-patient z-score normalization is then applied to each patient's standardized T2WI, computed b2000 DWI, and DCEI parameter maps in line with the work by Isensee et al., 2018, who showed this to be an effective strategy for MRI as CNN input. ADC maps are not normalized since ADC is a quantitative measurement.

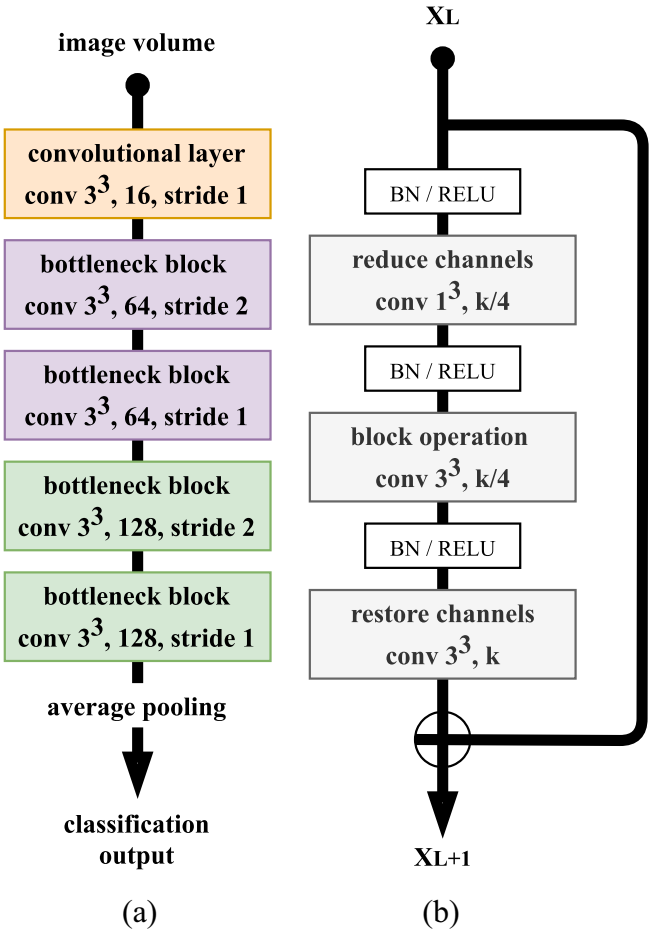


Fig. 3. (a) Proposed ResNet3D CNN used to extract features from volumetric images. (b) A bottleneck block, where k = #kernels.

Each clinical feature included in PCF is standardized to have zero mean and unit standard deviation, using mean and standard deviation computed from training data.

2.2. Convolutional neural network feature extraction

ResNet CNN architectures have demonstrated superior performance in several image classification tasks (He et al., 2015; 2016). In PCF, seven identical 3D ResNet CNN architectures, denoted $\{\text{ResNet3D-i}\}_{i=1}^{7}$, are employed to extract features from T2WI, ADC map, high b-value DWI, and each of the four DCEI parameter maps. ResNet3D is a modified 3D implementation of the standard 2D ResNet. Our implementation is composed of a convolutional layer C_1, followed by four bottleneck blocks B_1, B_2, B_3 and B_4, and a fully-connected layer FC. A network diagram is shown in Fig. 3a. Bottleneck blocks reduce the computational load of 3D convolutional layers by performing a channel reduction and restoration operation either side of the core convolution operation as shown in Fig. 3b. Preactivation (He et al., 2016) (batch normalization and rectified linear unit activation prior to weight layer computation) is used to ease optimization and regularize the networks. The last bottleneck block, B_4, outputs a set of feature maps to which global average pooling is applied to transform each feature map f_j into a feature value v_j.

During the training phase of PCF, each ResNet3D-i is trained end-to-end. The feature values v_j are linearly combined in the FC layer, followed by softmax to produce a classification output, followed by loss computation, backpropagation, and weight updates.

During inference, the feature values v_j, corresponding to ResNet3D-i, are grouped into a feature vector $V_i = \{v_j\}_{j=1}^{128}$. Subsequently, each feature vector V_i corresponding to each ResNet3D-i is passed to a two-level SVM scheme where the final patient classification is made.

2.3. Forward feature selection

The optimal subset of ResNet3D feature vectors $V = \{V_i\}_{i=1}^{7}$ and normalized clinical features $F = \{F_j\}_{j=1}^{N}$, for some quantity of clinical features N, is found during the training phase of PCF using forward feature selection (FFS) (Efroymson, 1966). FFS is used to remove features that are acting as noise; removing noise is especially important when training classification algorithms using small datasets. In our implementation of FFS, each ResNet3D feature vector V_i is considered a feature. We denote the total feature set $ALL = V \cup F$. We begin by assuming the null set of selected features $SEL = \varnothing$. At each iteration we induct the feature into SEL which maximizes an evaluation metric M computed over SEL. The FFS procedure is summarized as follows:

1. Initialize $SEL = \{\varnothing\}$;
2. For each feature $X_k \in ALL$, $k = 1, \ldots, N+7$, compute $M(X_k)$ and select the feature $\hat{X}_k$ that maximizes M;
3. Remove $\hat{X}_k$ from the set ALL and add $\hat{X}_k$ to the set SEL; thus $ALL := ALL \backslash \{\hat{X}_k\}$ and $SEL := \{\hat{X}_k\}$;
4. Repeat until a decrease in M is observed:
   (a) For each X_k in ALL, compute $M(SEL \cup \{X_k\})$ and select the feature $\hat{X}_k$ that maximizes M;
   (b) Remove $\hat{X}_k$ from the set ALL and add $\hat{X}_k$ to the set SEL; thus $ALL := ALL \backslash \{\hat{X}_k\}$ and $SEL := SEL \cup \{\hat{X}_k\}$;

2.4. Support vector machine classification

A two-level multi-classifier SVM scheme is used to combine the selected ResNet3D feature vectors and normalized clinical features to produce a final patient classification. Two SVMs, denoted SVM-1 and SVM-2, are included in the first layer and a third SVM, denoted SVM-3, is included in the second layer. First, SVM-1 takes the ResNet3D feature vectors $V_i \in SEL$ as input and outputs a patient classification probability $\hat{y}_1$. Concurrently, SVM-2 takes the normalized clinical features $F_j \in SEL$ as input and outputs a patient classification probability $\hat{y}_2$. Since SVMs do not naturally output a probability, Platt scaling (Platt, 1999) is used to obtain probability estimates. Then, SVM-3 accepts $\hat{y}_1$ and $\hat{y}_2$ as input to output a final classification probability $\hat{y}$ associated to the positive class i.e. probability that the patient has CSPCa. It should be noted that if clinical features are either not available or not selected by FFS, the final classification is made by SVM-1, and SVM-2 and SVM-3 will be omitted.

3. Experimental setup

In this section we describe the patient datasets used in this work, the classification tasks completed, the validation measures used to evaluate PCF, and the methodological settings employed for conducting experiments.

3.1. Patient data

The performance of PCF was evaluated using two datasets. The first is a dataset collected during the Prostate Imaging Compared to Transperineal Ultrasound-guided biopsy for significant prostate cancer Risk Evaluation (PICTURE) study (Simmons et al., 2014) and the second is the publicly available PROSTATEx dataset (Litjens et al., 2017a).

Table 1
PICTURE dataset patient characteristics. Interquartile range shown in brackets for age, PSA, TPV, and PSA density (PSAd).

Total patients following exclusions	210
Median age (years)	62 (58–67)
Median PSA (ng)	7 (5–10)
Median TPV (ml)	40 (28–51)
Median PSAd (ng/ml)	0.18 (0.13–0.28)
Breakdown by max GS	**# of patients**
Normal/Benign	30
GS 3+3	50
GS 3+4	96
GS 4+3	30
GS > 8	4

3.1.1. PICTURE dataset

Full details of the PICTURE study have been previously reported (Simmons et al., 2014). The PICTURE study recruited men who had undergone an initial standard transrectal ultrasound-guided (TRUS) biopsy, but concern remained over the accuracy of the subsequent diagnosis. As part of the study, patients were offered an ultrasound-guided TTPM biopsy with a 5 mm sampling frame and MRI-targeted biopsy, which were used as the reference standard against which the diagnostic accuracy of mpMRI could be determined. MpMRI was acquired using a 3 Tesla magnetic field scanner (Achieva, Philips Healthcare) and a pelvic-phased array coil. Sequences collected included T2WI, DWI with high b-value (2000), ADC map computed from DWI at multiple b-values (0, 150, 500, 1000), and DCEI with a temporal resolution of 13s. 249 men completed mpMRI, TTPM biopsy, and targeted biopsy. A 5-point Likert impression scale based on the outputs of a consensus group (Dickinson et al., 2011) was used by a radiologist with 10 years of experience in reading mpMRI to score at the lesion, sector, and patient level. Three definitions of clinical significance were considered during scoring: “Any cancer”, “Definition 2” ($\geq$ 0.2cc and/or $\geq$ GS 3+4) and “Definition 1” ($\geq$ 0.5cc and/or $\geq$ GS 4+3). Clinical information, including the referral PSA (ng) and estimated total prostate volume (TPV) (ml), was available to the radiologist during scoring to reflect real clinical practice.

In our work, the patient-level ground truth used for training and evaluating PCF was established as follows: a patient was allocated to the CSPCa class if any core sampled during TTPM biopsy or targeted biopsy was positive for GS $\geq 3+4$. Five patient studies were removed due to one or more missing MRI sequences and 34 patient studies were removed due to severe magnetic susceptibility artifacts on DWI. Characteristics of the included patients are shown in Table 1.

3.1.2. PROSTATEx dataset

Limited details of the PROSTATEx dataset have been previously reported (Litjens et al., 2014a). MpMRI and histopathological findings for 346 consecutive studies were downloaded from the PROSTATEx Challenges database (Litjens et al., 2017a). MpMRI was acquired using two 3 Tesla magnetic field scanners (Magnetom Trio and Skyra, Siemens) and a pelvic-phased array coil. Sequences collected included T2WI, ADC map computed from DWI at multiple b-values (50, 400, 800), and DCEI with a temporal resolution of 3.5s. All mpMRI studies were reported on by a radiologist with over 20 years of experience in reading prostate mpMRI (Barentsz), who indicated areas of suspicion per modality with a point marker. A PI-RADS v1 score was assigned to each lesion, though these were not available in the released dataset. MR-guided biopsies were carried out for PI-RADS v1 $\geq$ 3 lesions and biopsy samples were graded by a histopathologist.

Table 2
PROSTATEx dataset patient characteristics.

Total patients following exclusions	282
Breakdown by max GS	# of patients
No CSPCa*	212
GS 3+4	38
GS 4+3	19
GS > 8	13

*Either GS $\leq$ 6, benign or PI-RADS = 2. PI-RADS = 2 lesions were not biopsied; assumed not CSPCa, CSPCa occurrence in PI-RADS = 2 lesions at Radboud Medical Center less than 5%.

In our work, the patient-level ground truth used for training and evaluating PCF was established as follows: a patient was allocated to the CSPCa class if the patient's prostate contained any lesion with GS $\geq 3+4$.

64 patient studies were removed due to missing ground-truth labels; of these, two patients belonged to the PROSTATEx Challenges training set and 62 patients belonged to the PROSTATEx Challenges test set. Characteristics of the remaining 282 patient studies are shown in Table 2.

3.2. Experiments

PCF was trained to classify patients into those with CSPCa and those without CSPCa, where CSPCa refers to the presence of max GS $\geq 3+4$ tissue, as determined through histopathological analysis. In the PICTURE dataset a total of 130 patients with CSPCa and 80 patients without CSPCa were available for analysis, while in the PROSTATEx dataset a total of 70 patients with CSPCa and 212 patients without CSPCa were available for analysis. The following experiments were conducted:

- Intra-dataset evaluation: The following classifiers were trained: (i) ResNet3D with individual MRI modalities or parameter maps (ResNet3D-x, where $x \in X = \{$T2WI, ADC, Cb2000, IS, ME, TM, FS$\}$); (ii) SVM with individual clinical features (SVM-y, where $y \in Y = \{$PSA, TPV, PSAd$\}$); (iii) PCF with the set of available MRI modalities and parameter maps (PCF-ALL-MR); (iv) PCF with the set of MRI modalities and parameter maps selected by FFS (PCF-SEL-MR); (v) PCF with the set of available MRI modalities, parameter maps, and clinical features (PCF-ALL); and (vi) PCF with the set of MRI modalities, parameter maps. and clinical features selected by FFS (PCF-SEL). The performance of classifiers was evaluated using a five-fold cross-validation on the PICTURE and PROSTATEx datasets separately. The mean receiver operating characteristic (ROC) curve, precision-recall (PR) curve, and respective areas under the curve (AUC) were calculated in each instance. The Wilcoxon-signed rank test for pairwise comparison (Wilcoxon, 1945) was applied to statistically validate the comparison between different classifiers.
- Inter-dataset evaluation: ResNet3D classifiers, PCF-ALL-MR classifiers, and PCF-SEL-MR classifiers, obtained from the PICTURE dataset intra-dataset five-fold cross-validation were used to perform inference on the PROSTATEx dataset and vice versa. The mean and standard deviation of the ROC and PR AUCs of the five cross-validation models is presented.
- Clinical evaluation: The PICTURE dataset alone was used for clinical evaluation as radiologist PI-RADS v1 scores associated to the PROSTATEx dataset have not been released publicly. The PICTURE dataset was divided temporally into 170 patients for training (scan dates: 11/01/2012 to 25/06/2013) and 40 patients for testing (scan dates: 26/06/2013 to 29/01/2014). The test set comprised of 20 patients with CSPCa and 20 patients without CSPCa. PROSTATEx scans were used to augment the training set. PCF-SEL was used in the clinical evaluation. The probabilistic output of PCF-SEL was binarized by selecting a cutoff that matched the sensitivity of the radiologist on the PICTURE training set at two cutoffs: Likert $\geq$ 3 and Likert $\geq$ 4. The sensitivity, specificity, precision, and negative predictive value (NPV) were computed; 95% confidence intervals (CI) were calculated using bootstrapping. McNemar's test (McNemar, 1947) was used to statistically compare the sensitivity and specificity of the radiologist and PCF-SEL, while the weighted generalized score (WGS) test statistic (Kosinski, 2013) was used to compare the precision and NPV of the radiologist and PCF-SEL.

3.3. Experimental settings

In this section we describe the methodological settings used for conducting experiments with PCF.

3.3.1. Pre-processing settings

HighRes3DNet was trained using the T2WI of 82 patients from the PICTURE dataset for which manual contours of the whole prostate were available, and 50 training cases from the publicly available PROMISE12 dataset (Litjens et al., 2014b). All images were whitened and resampled to isotropic 1mm resolution as pre-processing, and resampled to original voxel resolution as post-processing. During training subvolumes of size 64^3 were sampled to maintain a 50:50 ratio of foreground to background voxels. Flip and rotation augmentations were applied on-the-fly. Training was conducted using Dice loss (Milletari et al., 2016), Adam optimisation (Kingma and Ba, 2015), learning rate equal to 0.001, and batch size 4. The network was trained until we observed a plateau in performance on the validation set. The trained network was used to segment the remainder of the PICTURE dataset and the entirety of the PROSTATEx dataset. A mean Dice score of 0.90 was achieved on a ten-fold cross-validation of the 82 PICTURE dataset patients.

Registration of T2WI to ADC map and first timepoint of DCEI, used to obtain the transformation of prostate masks into DWI and DCEI space, used default parameters for affine registration via symmetric block-matching (Modat et al., 2014). The subsequent non-rigid FFD registration used a Gaussian kernel with standard deviation equal to 5mm for LNCC calculation, control point spacing equal to 10mm, and bending energy constraint equal to 0.1.

A high b-value, b_c=2000, was selected for computing high b-value DWI as in Verma et al. (2016).

The DCEI parameter IS was calculated using an averaging window of length $l_{IS} = 3$ for the PICTURE dataset and $l_{IS} = 5$ for the PROSTATEx dataset. DCEI parameter FS was calculated over the final $m_{FS} = 2$ minutes of the normalized signal for the PICTURE dataset and $m_{FS} = 1$ minutes of the normalized signal for the PROSTATEx dataset.

As recommended in Nyúl et al. (2000), deciles were used as landmarks for histogram standardization of T2WI.

3.3.2. Training settings

For each experiment, training data was further subdivided 80:20 into training and validation sets. The training set was used for training constituent ResNet3D and SVM classifiers, while the validation set was used for selecting feature vectors and normalized clinical features during FFS.

All images were resized to a common size of $65 \times 65 \times 45$ prior to ResNet3D training. Each ResNet3D in PCF was trained using cross-entropy loss, Adam optimisation, learning rate equal to 0.00001, and batch size 8. In-plane flip and random deformation augmentations were applied to the training set to balance classes and reduce overfitting.

Table 3
Intra-dataset evaluation. Mean ROC AUC and PR AUC ± one standard deviation, for ResNet3D, SVM and PCF classifiers, averaged over five-fold cross validation, for the PICTURE and PROSTATEx datasets. Highest value in each column shown in bold.

	PICTURE dataset		PROSTATEx dataset	
Classifier	Mean ROC AUC	Mean PR AUC	Mean ROC AUC	Mean PR AUC
ResNet3D-T2WI	0.70 ± 0.06	0.79 ± 0.05	0.78 ± 0.07	0.60 ± 0.08
ResNet3D-ADC	0.74 ± 0.09	0.83 ± 0.08	0.80 ± 0.05	0.64 ± 0.04
ResNet3D-Cb2000	0.67 ± 0.10	0.79 ± 0.08	0.82 ± 0.05	0.66 ± 0.07
ResNet3D-IS	0.65 ± 0.03	0.75 ± 0.04	0.70 ± 0.08	0.47 ± 0.07
ResNet3D-ME	0.67 ± 0.06	0.76 ± 0.06	0.79 ± 0.05	0.55 ± 0.07
ResNet3D-TM	0.68 ± 0.13	0.77 ± 0.09	0.70 ± 0.08	0.44 ± 0.11
ResNet3D-FS	0.65 ± 0.08	0.75 ± 0.05	0.72 ± 0.03	0.44 ± 0.03
PCF-ALL-MR	0.72 ± 0.09	0.80 ± 0.07	0.82 ± 0.04	0.63 ± 0.05
PCF-SEL-MR	0.77 ± 0.11	0.84 ± 0.09	$\mathbf{0.86 \pm 0.04}$	$\mathbf{0.72 \pm 0.03}$
SVM-PSA	0.54 ± 0.06	0.64 ± 0.03	n/a	n/a
SVM-TPV	0.70 ± 0.12	0.80 ± 0.10	n/a	n/a
SVM-PSAd	0.73 ± 0.07	0.82 ± 0.06	n/a	n/a
PCF-ALL	0.74 ± 0.10	0.81 ± 0.09	0.82 ± 0.04	0.63 ± 0.05
PCF-SEL	$\mathbf{0.79 \pm 0.09}$	$\mathbf{0.86 \pm 0.07}$	$\mathbf{0.86 \pm 0.04}$	$\mathbf{0.72 \pm 0.03}$

The following metric M is proposed for observation during FFS:

$$M = \frac{ROC\,AUC + PR\,AUC}{2}, \tag{7}$$

as it maximizes both model evaluation metrics of interest.

A radial basis kernel was used in SVM-1, SVM-2, and SVM-3 as there existed no reason to assume linear separability of data. The misclassification penalty was set to $C = 0.1$ for SVM-1 and SVM-2, and $C = 1$ for SVM-3, in all experiments.

4. Results

In this section we present the results obtained from the intra-dataset and inter-dataset evaluations of PCF, as well as the clinical evaluation of PCF using a temporally separated patient cohort from the PICTURE dataset.

4.1. Intra-dataset model evaluation

The mean ROC and PR AUCs averaged over five-fold cross-validation for ResNet3D, SVM, and PCF classifiers are shown in Table 3 for both the PICTURE and PROSTATEx datasets. Fig. 4a–d show the mean ROC and PR curves calculated for PCF-SEL for both datasets. Reliability diagrams for PCF-SEL are shown in Fig. 4e for both datasets. An additional comparison of PCF to an end-to-end CNN is shown in Appendix A and an analysis of the relationship between the probabilistic output of PCF with radiologist Likert score and biopsy maximum cancer core length is shown in Appendix B.

For the PICTURE dataset, ResNet3D-ADC had the best performance among ResNet3D and SVM classifiers that were trained using a single MRI modality, parameter map, or clinical feature. PCF-ALL did not improve the result. However, PCF-SEL did improve upon the result of ResNet3D-ADC, with an increase in ROC AUC from 0.74 to 0.79 ($p < 0.05$) and an increase in PR AUC from 0.83 to 0.86 ($p = 0.08$); during the five-fold cross-validation of PCF-SEL, FFS selected ADC map, PSAd, Cb2000 DWI, and TM map in the majority of fold experiments run.

For the PROSTATEx dataset, ResNet3D-Cb2000 had the best performance among ResNet3D classifiers that were trained using a single MRI modality of parameter map. PCF-ALL did not improve the result. However, PCF-SEL did improve upon the result of ResNet3D-Cb2000, with an increase in ROC AUC from 0.82 to 0.86 ($p = 0.07$) and an increase in PR AUC from 0.66 to 0.72 ($p = 0.07$); during the five-fold cross-validation of PCF-SEL, FFS selected Cb2000 DWI, ADC map, and ME map in the majority of fold experiments run.

In addition to the ability to discriminate between classes, it is desirable for models to produce well-calibrated probability estimates. For output probability $\hat{P}$, perfect calibration is defined as:

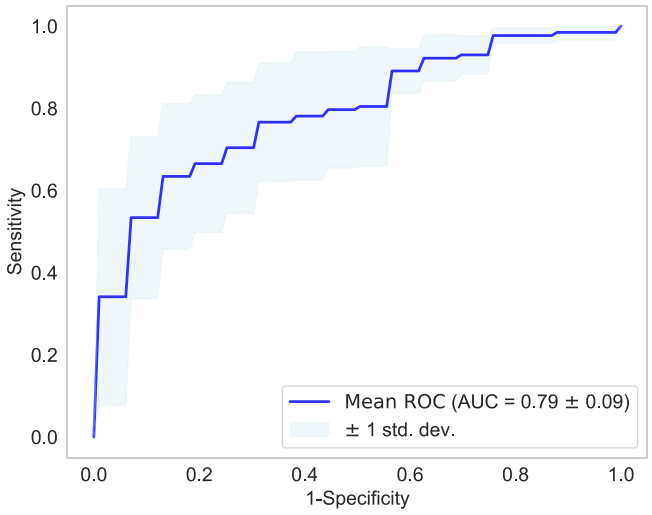
$$P(CSPCa \mid \hat{P} = p) = p, \quad \forall p \in [0, 1], \tag{8}$$

i.e. $\hat{P}$ should represent a true probability (Guo et al., 2017). Fig. 4a shows reliability diagrams for PICTURE and PROSTATEx dataset patient probabilities output by PCF-SEL. Perfect calibration is represented by the identity diagonal. As observed for both datasets, the identity diagonal is broadly tracked indicating reasonable calibration. For the PICTURE dataset we observe better calibration at the higher probability end, while for the PROSTATEx dataset we observe better calibration at the lower probability end. This may be explained by the higher prevalence of CSPCa patients in the PICTURE dataset and the higher prevalence of patients with benign conditions or low-grade PCa in the PROSTATEx dataset.

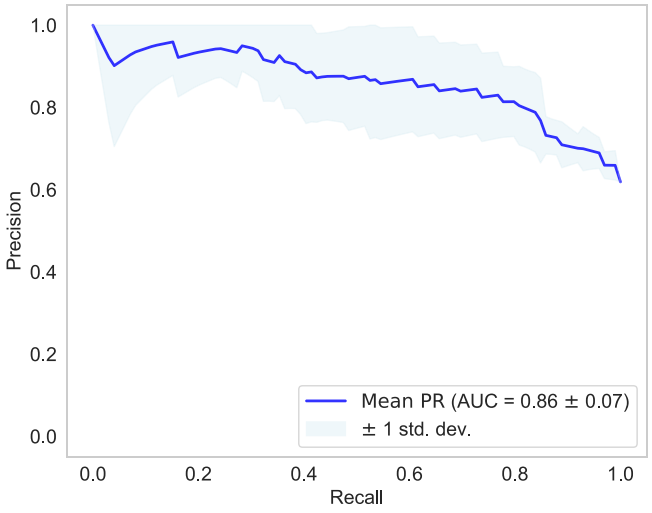
4.2. Inter-dataset model evaluation

ResNet3D classifiers, PCF-ALL-MR classifiers, and PCF-SEL-MR classifiers, obtained from the PICTURE dataset intra-dataset five-fold cross-validation, were used to perform inference on the PROSTATEx dataset and vice versa. The mean and standard deviation of the ROC and PR AUCs of the five cross-validation models is presented in Table 4. Clinical features were not considered since they were not available for the PROSTATEx dataset.

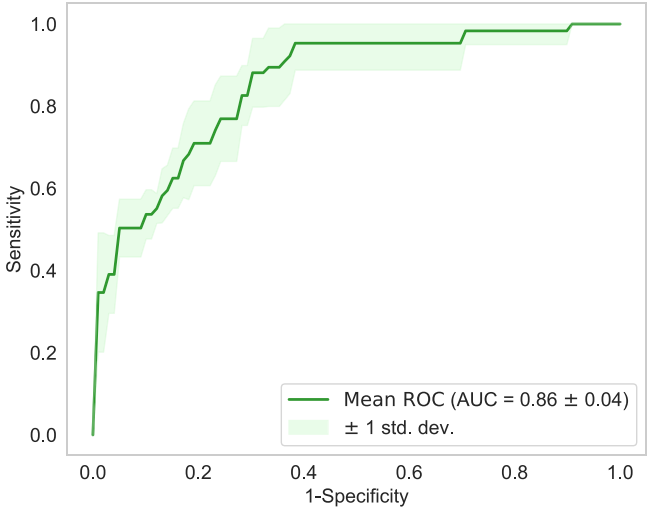
ResNet3D-ADC trained using the PROSTATEx dataset and applied to the PICTURE dataset maintained a similar performance level to ResNet3D-ADC trained with the PICTURE dataset. Similarly, ResNet3D-Cb2000 trained using the PICTURE dataset and applied to the PROSTATEx dataset maintained a similar performance level to ResNet3D-Cb2000 trained with the PROSTATEx dataset. For both datasets we observed a decrease in the performance of ResNet3D classifiers trained using DCEI parameter maps likely due to the differences in temporal resolution of the DCEI between the PICTURE and PROSTATEx datasets (13s vs. 3.5s). Notably, for both datasets we observed a drop in the performance of PCF-SEL-MR as compared to its performance in the intra-dataset evaluation, primarily as the datasets do not share the same optimal modalities and due to the reduction in performance of constituent ResNet3D classifiers trained using DCEI parameter maps.



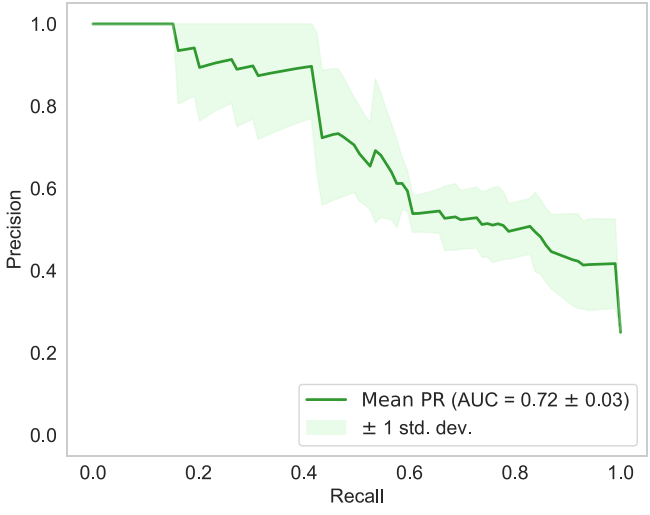
(a)



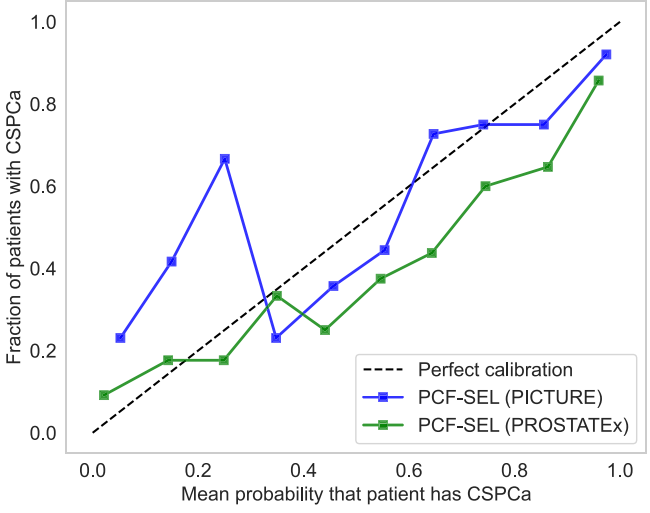
(b)



(c)



(d)



(e)

Fig. 4. Intra-dataset evaluation. Graphs (a, b) show the mean ROC and PR curves, averaged over five-fold cross validation, for PCF-SEL, for the PICTURE dataset, while graphs (c, d) correspond to the PROSTATEx dataset. Reliability diagrams for PCF-SEL are shown in (e), for both datasets.

Table 4
Inter-dataset evaluation. Mean ROC AUC and PR AUC ± one standard deviation for ResNet3D and PCF classifiers obtained during the intra-dataset evaluation and subsequently applied to the dataset that was not used to train the classifier. Highest value in each column shown in bold.

	PROSTATEx train, PICTURE test		PICTURE train, PROSTATEx test	
Classifier	Mean ROC AUC	Mean PR AUC	Mean ROC AUC	Mean PR AUC
ResNet3D-T2WI	0.70 ± 0.01	0.79 ± 0.01	0.75 ± 0.01	0.51 ± 0.03
ResNet3D-ADC	$\mathbf{0.73 \pm 0.03}$	$\mathbf{0.82 \pm 0.02}$	0.73 ± 0.02	0.47 ± 0.01
ResNet3D-Cb2000	0.68 ± 0.00	0.79 ± 0.00	$\mathbf{0.81 \pm 0.02}$	$\mathbf{0.65 \pm 0.03}$
ResNet3D-IS	0.63 ± 0.01	0.72 ± 0.01	0.51 ± 0.08	0.30 ± 0.06
ResNet3D-ME	0.62 ± 0.02	0.72 ± 0.01	0.59 ± 0.06	0.33 ± 0.05
ResNet3D-TM	0.65 ± 0.03	0.74 ± 0.02	0.58 ± 0.07	0.33 ± 0.06
ResNet3D-FS	0.60 ± 0.04	0.69 ± 0.02	0.62 ± 0.04	0.33 ± 0.03
PCF-ALL-MR	$\mathbf{0.73 \pm 0.01}$	0.80 ± 0.01	0.75 ± 0.03	0.50 ± 0.04
PCF-SEL-MR	0.72 ± 0.03	0.81 ± 0.03	0.77 ± 0.07	0.56 ± 0.11

Table 5
Clinical comparison of the patient-level diagnostic performance of radiologist Likert scoring and PCF-SEL, on temporally separated training and test cohorts from the PICTURE dataset.

Metric	Value	95% CI	Value	95% CI	P-value
Training set ($\mathbf{n = 170}$)					
	Likert ≥ 3		PCF-SEL ≥ 0.17		
Sensitivity / Recall (%)	95 (104/110)	(90-98)	95 (104/110)	(90-98)	1.00
Specificity (%)	33 (20/60)	(22-45)	65 (39/60)	(53-77)	< 0.01
Precision / PPV (%)	72 (104/144)	(65-79)	83 (104/125)	(76-90)	< 0.01
NPV (%)	77 (20/26)	(59-92)	87 (39/45)	(76-96)	0.27
	Likert ≥ 4		PCF-SEL ≥ 0.75		
Sensitivity / Recall (%)	69 (76/110)	(60-78)	69 (76/110)	(60-78)	1.00
Specificity (%)	77 (46/60)	(66-87)	87 (52/60)	(78-95)	0.21
Precision / PPV (%)	84 (76/90)	(77-92)	90 (76/84)	(84-96)	0.14
NPV (%)	58 (46/80)	(47-68)	60 (52/86)	(50-71)	0.54
Test set ($\mathbf{n = 40}$)					
	Likert ≥ 3		PCF-SEL ≥ 0.17		
Sensitivity / Recall (%)	100 (20/20)	(100-100)	95 (19/20)	(83-100)	1.00
Specificity (%)	20 (4/20)	(5-39)	35 (7/20)	(14-57)	0.51
Precision / PPV (%)	56 (20/36)	(39-71)	59 (19/32)	(42-76)	0.47
NPV (%)	100 (4/4)	(100-100)	87 (7/8)	(60-100)	0.46
	Likert ≥ 4		PCF-SEL ≥ 0.75		
Sensitivity / Recall (%)	75 (15/20)	(55-94)	75 (15/20)	(55-94)	1.00
Specificity (%)	75 (15/20)	(55-93)	55 (11/20)	(33-77)	0.23
Precision / PPV (%)	75 (15/20)	(55-93)	63 (15/24)	(43-82)	0.34
NPV (%)	75 (15/20)	(55-94)	69 (11/16)	(44-91)	0.57

4.3. Clinical evaluation

In this section we present the results of the clinical evaluation of PCF-SEL. To simulate prospective use we temporally split the PICTURE dataset into 170 patients for training and 40 patients for testing (20 patients with CSPCa and 20 patients without CSPCa). The performance of PCF-SEL is compared to the performance of an experienced radiologist (10 years of experience in reading prostate mpMRI) who assigned a Likert score to each patient. To enable calculation of sensitivity, specificity, precision, and NPV for PCF-SEL, the probabilistic output of PCF-SEL was thresholded to match the sensitivity of the radiologist on the training set. The results of the clinical evaluation are shown in Table 5. Fig. 5 shows the training and test set ROC and PR curves calculated for PCF-SEL, as well as the performance of the radiologist and PCF-SEL at two operating thresholds.

FFS selected SEL = {T2WI, ADC map, Cb2000 DWI, PSAd}. Using the training cohort a probability threshold equal to 0.17 was selected for PCF-SEL to match the sensitivity of the radiologist at Likert threshold ≥ 3, while a probability threshold equal to 0.75 was selected to match the sensitivity of the radiologist at Likert threshold ≥ 4. On the test cohort, PCF-SEL achieved sensitivities of 95% and 75%, compared to the radiologist who achieved sensitivities of 100% and 75% and PCF-SEL achieved specificities of 35% and 55%, compared to the radiologist who achieved specificities of 20% and 75%.

While differences in specificity can be observed in favour of PCF-SEL at the higher sensitivity setting and in favour of the radiologist at the lower sensitivity setting, McNemar's test did not find statistically significant differences between PCF-SEL and the radiologist on the test cohort.

5. Discussion

In this work we proposed a patient-level classification framework, denoted PCF, that uses volumetric mpMRI, derived parameter maps, and clinical features, jointly, to classify patients into those with and without CSPCa. PCF is trained using patient-level labels only, thus avoiding the need for lesion annotations, which can be challenging and time-consuming to obtain. The performance of PCF was evaluated using the PICTURE and PROSTATEx datasets. We performed an intra-dataset five-fold cross-validation, an inter-dataset generalization experiment, and a clinical evaluation of PCF on a temporally separated patient cohort from the PICTURE dataset.

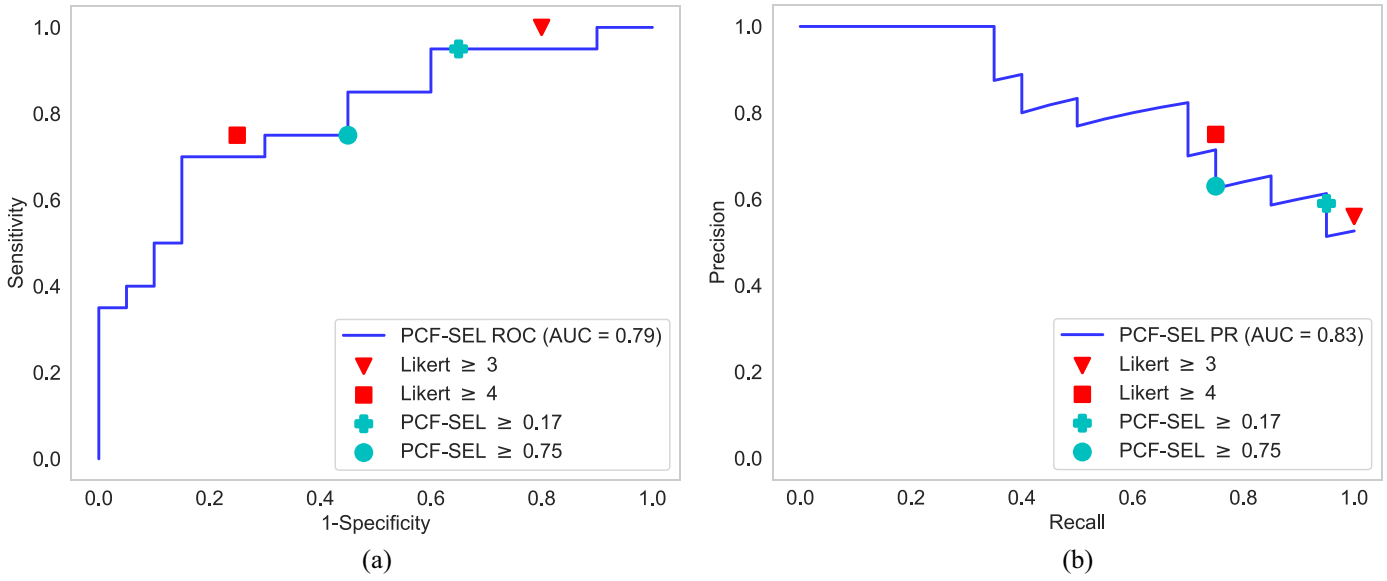


Fig. 5. Graph (a) shows the ROC curve and graph (b) shows the PR curve for PCF-SEL on the temporally separated PICTURE dataset test cohort. Radiologist performance at Likert cutoffs (≥ 3 and ≥ 4) and PCF-SEL performance at probability cutoffs (≥ 0.17 and ≥ 0.75) are shown, where probability cutoffs for PCF-SEL were selected using the PICTURE dataset training cohort.

In the intra-dataset five-fold cross-validation, the performance of PCF with feature selection enabled (PCF-SEL) was compared to the performance of PCF with feature selection disabled (PCF-ALL), to assess whether feature selection in PCF has a performance benefit. Further comparison was made to ResNet3D and SVM classifiers trained using individual MRI modalities, parameter maps, and clinical features. On both the PICTURE and PROSTATEx datasets, PCF-SEL outperformed all other classifiers. On the PICTURE dataset, PCF-SEL achieved a mean ROC AUC of 0.79 and mean PR AUC of 0.86; ADC map, PSAd, Cb2000 DWI and TM map were selected for inference over at least three out of five folds during the five-fold cross-validation. On the PROSTATEx dataset, PCF-SEL achieved a mean ROC AUC of 0.86 and mean PR AUC of 0.72; for this dataset, Cb2000 DWI, ADC map, and ME map were selected for inference over at least three out of five folds during the five-fold cross-validation. Three observations are made based on the results of the intra-dataset evaluation. First, we observe that the inclusion of feature selection during the training stage of PCF yields performance benefit, as shown by the superior performance of PCF-SEL as compared to PCF-ALL. The feature selection step improves generalizability to unseen data by removing MRI modalities, parameter maps, and clinical features that are acting as noise; removing sources of noise is especially important when training classification algorithms with small datasets which are common in PCa CAD works primarily due to the need for a consistent and accurate reference standard. Second, we observe that PCF-SEL successfully uses clinical features alongside MRI to improve patient classification performance. Our method uses a stacked ensemble of SVMs, where MRI features and clinical features are processed by separate dedicated SVMs, whose outputs are combined by a third SVM, to produce to final patient classification. Using both clinical features and MRI features for improved classification performance is in line with works by Antonelli et al. (2019) and Woźnicki et al. (2020) who showed the utility of PSAd in lesion classification tasks. Third, we observed a performance benefit from using DCEI parameter maps. The semi-quantitative DCEI parameters calculated in this work avoid the challenging estimation of the arterial input function needed for computing pharmacokinetic parameters (Haq et al., 2015). However, prior to clinical adoption it would be important to consider whether the gain in performance from using DCEI justifies the additional costs and risks associated to gadolinium injection; this is beyond the scope of this paper.

In the intra-dataset evaluation we considered the ability of PCF to generalize to unseen patient data from the same distribution as the training patient data. However, it is also of interest to consider the ability of CAD systems to generalize to external patient cohorts, since this type of generalizability if observed would allow for wider deployment of a trained system. However, our inter-dataset evaluation revealed a generalization gap. More precisely, for the PICTURE dataset we observed a drop in the performance of PCF-SEL-MR as compared to its performance in the intra-dataset evaluation, from a ROC AUC of 0.77 to 0.72. As the feature selection step uses validation data from the same distribution as the training data, it does not guarantee selection of the optimal modalities in the external dataset. However, a small increase in ROC AUC was observed for PCF-ALL-MR from 0.72 to 0.73. On the PROSTATEx dataset, both PCF-SEL-MR and PCF-ALL-MR had diminished performance, again, as the datasets do not share the same optimal modalities and additionally due to the reduction in the performance of constituent ResNet3D classifiers. Our findings suggest that training CAD systems with data from the institution in which deployment is intended is the optimal strategy and should be sought where possible.

It is important to clinically evaluate prostate CAD systems. Central to this is the need to compare CAD system performance to the performance of radiologists who are the current clinical standard. Moreover, clinical evaluations should consider how CAD systems may perform prospectively which can be simulated using a temporally separated patient cohort or an external patient cohort. Furthermore, an effective clinical evaluation requires the probabilistic output of the CAD system to be thresholded, allowing measures such as sensitivity, specificity, precision, and NPV to be reported as opposed to ROC AUC or PR AUC, which allow model comparison, but are less useful measures clinically. We compared the performance of PCF-SEL to the performance of a radiologist with 10 years of experience in reading prostate mpMRI, who gave a Likert score to each patient's prostate indicating the likelihood

of CSPCa. On a temporally separated cohort of 40 patients from the PICTURE dataset, the radiologist achieved a sensitivity of 100% and a specificity of 20% at Likert threshold $\geq$ 3, while PCF-SEL achieved a sensitivity of 95% and a specificity of 35% at a probability threshold equal to 0.17. At Likert threshold $\geq$ 4, the radiologist achieved a sensitivity of 75% and a specificity of 75%, whereas PCF-SEL achieved a sensitivity of 75% and specificity of 55% at a probability threshold equal to 0.75. The differences in performance between the radiologist and PCF-SEL were not found to be statistically significant, providing initial evidence for PCF-SEL to be evaluated in a larger clinical trial.

Our future work will introduce uncertainty estimation into our framework. Several works have investigated epistemic (model-based) uncertainty and aleatoric (data-based) uncertainty for classification and regression tasks (Gal and Ghahramani, 2016; Kendall and Gal, 2017; Eaton-Rosen et al., 2018; Wang et al., 2019). These are especially important to consider for medical systems to indicate cases where human intervention may be required or to facilitate active learning for continuous optimisation of systems. From a clinical viewpoint, we are planning clinical trials to examine the performance of PCF as a second reader and as a triage tool which can identify low-risk patients that do not require a clinical read; both applications can help alleviate the ever-increasing workload of radiologists.

CRediT authorship contribution statement

Pritesh Mehta: Conceptualization, Methodology, Software, Writing - original draft. **Michela Antonelli:** Conceptualization, Supervision, Writing - review & editing. **Hashim U. Ahmed:** Resources. **Mark Emberton:** Resources. **Shonit Punwani:** Supervision. **Sébastien Ourselin:** Supervision.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [EP/R512400/1]. This work was additionally supported by the EPSRC-funded UCL Centre for Doctoral Training in Intelligent, Integrated Imaging in Healthcare (i4health) [EP/S021930/1].

Appendix A. Performance comparison of proposed patient classification framework, PCF, with an end-to-end convolutional neural network

An end-to-end seven-stream CNN (E2E-Net) was trained using volumetric T2WI, ADC map, computed b2000 DWI, and four DCEI parameter maps, to output the patient-level probability of CSPCa. Each stream of E2E-Net follows the architecture of the 3D ResNet CNN shown in Fig. 3a, with the output node removed. In E2E-Net, convolutional streams are joined using a 128-way fully connected layer. Table A.1 compares the performance of E2E-Net to the performance of PCF-ALL-MR and PCF-SEL-MR using an intra-dataset five-fold cross-validation, where PCF-ALL-MR is trained using all seven of the MRI modalities and parameter maps, while PCF-SEL-MR applies the forward feature selection step described in Section 2.3, during training.

Table A.1
Mean ROC AUC $\pm$ one standard deviation for E2E-Net, PCF-ALL-MR, and PCF-SEL-MR, averaged over five-fold cross validation, for the PICTURE and PROSTATEx datasets. Highest value in each column shown in bold.

Classifier	PICTURE dataset	PROSTATEx dataset
E2E-Net	0.69 ± 0.07	0.80 ± 0.06
PCF-ALL-MR	0.72 ± 0.09	0.82 ± 0.04
PCF-SEL-MR	$\mathbf{0.77 \pm 0.11}$	$\mathbf{0.86 \pm 0.04}$

For both the PICTURE and PROSTATEx datasets we observe a large increase in ROC AUC from PCF-SEL-MR. The forward feature selection step improves generalizability to unseen data by removing MRI modalities, parameter maps, and clinical features that are acting as noise; removing sources of noise is especially important when training classification algorithms with small datasets, which are common in PCa CAD works, primarily due to the need for a consistent and accurate reference standard.

Appendix B. Mean probability of clinically significant prostate cancer, grouped by radiologist Likert score and maximum cancer core length

During the PICTURE study an experienced radiologist (>10 years of experience in reading prostate mpMRI) assigned a Likert score to a maximum of six lesions per patient, based on T2WI alone, DWI alone, DCEI alone, and a combination of all modalities. It would be reasonable to expect PCF to output the highest probabilities of CSPCa for patients who have one or more large Likert 4 or 5 lesions. Table B.1 shows the averaged probabilities of CSPCa associated to 210 patients from the PICTURE dataset who either had no scorable lesion, a max Likert 2 or 3 lesion, a max Likert 4 or 5 lesion with maximum cancer core length (MCCL) $<$ 6mm, or a max Likert 4 or 5 lesion with MCCL $\geq$ 6mm. The probabilities were output by ResNet3D-T2WI, ResNet3D-ADC, and PCF-SEL during the intra-dataset five-fold cross validation described in Section 4.1.

We observe the highest average probability of CSPCa for patients with max Likert 4 or 5 lesion with MCCL $\geq$ 6mm, followed by patients with max Likert 4 or 5 lesion with MCCL $<$ 6mm, followed by patients with max Likert 2 or 3 lesion, and the lowest average probability of CSPCa for patients with no scorable lesion.

Table B.1
Mean probability of CSPCa, grouped by Likert score and MCCL, for the PICTURE dataset.

Patient grouping	ResNet3D-T2WI	ResNet3D-ADC	PCF-SEL
No scorable lesion	0.45	0.39	0.39
Max Likert 2 or 3	0.52	0.43	0.47
Max Likert 4/5, MCCL $<$ 6mm	0.53	0.49	0.51
Max Likert 4/5, MCCL $\geq$ 6mm	0.57	0.61	0.76

References

Ahmed, H.U., El-Shater Bosaily, A., Brown, L.C., Gabe, R., Kaplan, R., Parmar, M.K., Collaco-Moraes, Y., Ward, K., Hindley, R.G., Freeman, A., Kirkham, A.P., Oldroyd, R., Parker, C., Emberton, M., 2017. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. Lancet 389 (10071), 815–822.

American College of Radiology, 2015. PI-RADS version 2. ACR.

Antonelli, M., Johnston, E.W., Dikaios, N., Cheung, K.K., Sidhu, H.S., Appayya, M.B., Giganti, F., Simmons, L.A.M., Freeman, A., Allen, C., Ahmed, H.U., Atkinson, D., Ourselin, S., Punwani, S., 2019. Machine learning classifiers can predict Gleason pattern 4 prostate cancer with greater accuracy than experienced radiologists. Eur. Radiol. 29, 4754–4764.

Blackledge, M.D., Leach, M.O., Collins, D.J., Koh, D.-M., May, I., Tumor, I., Blackledge, M.D., Leach, M.O., Collins, D.J., 2011. Computed diffusion-weighted MR imaging may improve tumor detection. Radiology 261 (2), 573–581.

Bonekamp, D., Kohl, S., Wiesenfarth, M., Schelb, P., Radtke, J.P., Gotz, M., Kickingereder, P., Yaqubi, K., Hitthaler, B., Gahlert, N., Kuder, T.A., Deister, F., Freitag, M., Hohenfellner, M., Hadaschik, B.A., Schlemmer, H.P., Maier-Hein, K.H., 2018. Radiomic machine learning for characterization of prostate lesions with MRI: comparison to ADC values. Radiology 289, 128–137.

Borofsky, S., George, A.K., Gaur, S., Bernardo, M., Greer, M.D., Mertan, F.V., Taffel, M., Moreno, V., Merino, M.J., Wood, B.J., Pinto, P.A., Choyke, P.L., Turkbey, B., 2017. What are we missing? False-negative cancers at multiparametric MR imaging of the prostate. Radiology 286 (1), 186–195.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A., 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 68, 394–424.

Brizmohun Appayya, M., Adshead, J., Ahmed, H.U., Allen, C., Bainbridge, A., Barrett, T., Giganti, F., Graham, J., Haslam, P., Johnston, E.W., Kastner, C., Kirkham, A.P.S., Lipton, A., McNeill, A., Moniz, L., Moore, C.M., Nabi, G., Padhani, A.R., Parker, C., Patel, A., Pursey, J., Richenberg, J., Staffurth, J., van der Meulen, J., Walls, D., Punwani, S., 2018. National implementation of multi-parametric magnetic resonance imaging for prostate cancer detection recommendations from a UK consensus meeting. BJU Int. 122, 13–25.

Cachier, P., Bardinet, E., Dormont, D., Pennec, X., Ayache, N., 2003. Iconic feature based nonrigid registration: the PASHA algorithm. Comput. Vis. Image Underst. 89 (3), 272–298.

Cao, R., Mohammadian Bajgiran, A., Afshari Mirak, S., Shakeri, S., Zhong, X., Enzmann, D., Raman, S., Sung, K., 2019. Joint prostate cancer detection and Gleason score prediction in mp-MRI via FocalNet. IEEE Trans. Med. Imaging 38, 2496–2506.

De Luca, M., Giannini, V., Vignati, A., Mazzetti, S., Bracco, C., Stasi, M., Armando, E., Russo, F., Bollito, E., Porpiglia, F., Regge, D., 2011. A fully automatic method to register the prostate gland on T2-weighted and EPI-DWI images. In: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS. IEEE, pp. 8029–8032.

Dickinson, L., Ahmed, H.U., Allen, C., Barentsz, J.O., Carey, B., Futterer, J.J., Heijmink, S.W., Hoskin, P.J., Kirkham, A., Padhani, A.R., Persad, R., Puech, P., Punwani, S., Sohaib, A.S., Tombal, B., Villers, A., Meulen, J.V.D., Emberton, M., 2011. Magnetic resonance imaging for the detection, localisation, and characterisation of prostate cancer : recommendations from a European consensus meeting. Eur. Urol. 59 (4), 477–494.

Dinh, A.H., Melodelima, C., Souchon, R., Moldovan, P.C., Bratan, F., Pagnoux, G., Mege-Lechevallier, F., Ruffion, A., Crouzet, S., Colombel, M., Rouviere, O., 2018. Characterization of prostate cancer with Gleason score of at least 7 by using quantitative multiparametric MR imaging: validation of a computer-aided diagnosis system in patients referred for prostate biopsy. Radiology 287, 525–533.

Eaton-Rosen, Z., Bragman, F., Bisdas, S., Ourselin, S., Cardoso, M.J., 2018. Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions. Lect. Notes Comput. Sci. 11070 LNCS, 691–699.

Efroymson, M., 1966. Stepwise regressiona backward and forward look. In: Proceedings of the Eastern Regional Meetings of the Institute of Mathematical Statistics.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML 2016), 3, pp. 1651–1660.

Gaur, S., Lay, N., Harmon, S.A., Doddakashi, S., Mehralivand, S., Argun, B., Barrett, T., Bednarova, S., Girometti, R., Karaarslan, E., Kural, A.R., Oto, A., Purysko, A.S., Antic, T., Magi-Galluzzi, C., Saglican, Y., Sioletic, S., Warren, A.Y., Bittencourt, L., Futterer, J.J., Gupta, R.T., Kabakus, I., Law, Y.M., Margolis, D.J., Shebel, H., Westphalen, A.C., Wood, B.J., Pinto, P.A., Shih, J.H., Choyke, P.L., Summers, R.M., Turkbey, B., 2018. Can computer-aided diagnosis assist in the identification of prostate cancer on prostate MRI? A multi-center, multi-reader investigation. Oncotarget 9, 33804–33817.

Giannini, V., Mazzetti, S., Armando, E., Carabalona, S., Russo, F., Giacobbe, A., Muto, G., Regge, D., 2017. Multiparametric magnetic resonance imaging of the prostate with computer-aided detection: experienced observer performance study. Eur. Radiol. 27, 4200–4208.

Greer, M.D., Lay, N., Shih, J.H., Barrett, T., Bittencourt, L.K., Borofsky, S., Kabakus, I., Law, Y.M., Marko, J., Shebel, H., Mertan, F.V., Merino, M.J., Wood, B.J., Pinto, P.A., Summers, R.M., Choyke, P.L., Turkbey, B., 2018. Computer-aided diagnosis prior to conventional interpretation of prostate mpMRI: an international multi-reader study. Eur. Radiol. 28, 4407–4417.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: Proceedings of the Machine Learning Research.

Haq, N.F., Kozlowski, P., Jones, E.C., Chang, S.D., Goldenberg, S.L., Moradi, M., 2015. A data-driven approach to prostate cancer detection from dynamic contrast enhanced MRI. Comput. Med. Imaging Graph. 41, 37–45.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. CoRR abs/1512.03385.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. CoRR abs/1603.05027.

Hoeks, C., Barentsz, J., Hambrock, T., Yakar, D., Somford, D., Heijmink, S., Scheenen, T., Vos, P., Huisman, H., van Oort, I., Witjes, J.A., Heerschap, A., Fütterer, J., 2011. Prostate cancer: multiparametric MR imaging for detection, localization, and staging. Radiology 261, 46–66.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., Maier-Hein, K.H., 2018. nnU-net: self-adapting framework for U-Net-based medical image segmentation. CoRR abs/1809.10486.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in Bayesian deep learning for computer vision? In: Proceedings of the Advances in Neural Information Processing Systems, NIPS 2017, pp. 5575–5585.

Kingma, D.P., Ba, J., 2015. Adam: a method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations.

Kosinski, A.S., 2013. A weighted generalized score statistic for comparison of predictive values of diagnostic tests. Stat. Med. 32 (6).

Kubassova, O.A., Boyle, R.D., Radjenovic, A., 2007. Quantitative analysis of dynamic contrast-enhanced MRI datasets of the metacarpophalangeal joints. Acad. Radiol. 14, 1189–1200.

Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M.J., Vercauteren, T., 2017. On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task. CoRR abs/1707.01992.

Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., Huisman, H., 2014a. Computer-aided detection of prostate cancer in MRI. IEEE Trans. Med. Imaging 33, 1083–1092.

Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., Huisman, H., 2017a. Prostatex challenge data. Data retrieved from: The Cancer Imaging Archive, DOI: 10.1109/TMI.2014.2303821.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I., 2017b. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Litjens, G., Toth, R., van de Ven, W., Hoeks, C., Kerkstra, S., van Ginneken, B., Vincent, G., Guillard, G., Birbeck, N., Zhang, J., Strand, R., Malmberg, F., Ou, Y., Davatzikos, C., Kirschner, M., Jung, F., Yuan, J., Qiu, W., Gao, Q., Edwards, P.E., Maan, B., van der Heijden, F., Ghose, S., Mitra, J., Dowling, J., Barratt, D., Huisman, H., Madabhushi, A., 2014b. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Med. Image Anal. 18 (2), 359–373.

McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12 (2), 153–157.

Milletari, F., Navab, N., Ahmadi, S.-a., 2016. V-Net : fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the 3DV.

Modat, M., Cash, D.M., Daga, P., Winston, G.P., Duncan, J.S., Ourselin, S., 2014. Global image registration using a symmetric block-matching approach. J. Med. Imaging 1.

Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S., 2010. Fast free-form deformation using graphics processing units. Comput. Methods Progr. Biomed. 98, 278–284.

Nyúl, L.G., Udupa, J.K., Zhang, X., 2000. New variants of a method of MRI scale standardization. IEEE Trans. Med. Imaging 19, 143–150.

Platt, J., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv. Large Margin Classif. 10, 61–74.

Schelb, P., Kohl, S., Radtke, J.P., Wiesenfarth, M., Kickingereder, P., Bickelhaupt, S., Kuder, T.A., Stenzinger, A., Hohenfellner, M., Schlemmer, H.-P., Maier-Hein, K.H., Bonekamp, D., 2019. Classification of cancer at prostate MRI: deep learning versus clinical PI-RADS assessment. Radiology 293 (3), 607–617.

Simmons, L.A.M., Ahmed, H.U., Moore, C.M., Punwani, S., Freeman, A., Hu, Y., Barratt, D., Charman, S.C., Van der Meulen, J., Emberton, M., 2014. The PICTURE study - prostate imaging (multi-parametric MRI and Prostate HistoScanning compared to transperineal ultrasound guided biopsy for significant prostate cancer risk evaluation. Contemp. Clin. Trials 37, 69–83.

Steenbergen, P., Haustermans, K., Lerut, E., Oyen, R., De Wever, L., Van Den Bergh, L., Kerkmeijer, L.G.W., Pameijer, F.A., Veldhuis, W.B., Van Der Voort Van Zyp, J.R.N., Pos, F.J., Heijmink, S.W., Kalisvaart, R., Teertstra, H.J., Dinh, C.V., Ghobadi, G., Van Der Heide, U.A., 2015. Prostate tumor delineation using multiparametric magnetic resonance imaging: inter-observer variability and pathology validation. Radiother. Oncol. 115 (2), 186–190.

The Royal College of Radiologists, 2018. UK workforce census report 2018. Clinical radiology.

Thon, A., Teichgräber, U., Tennstedt-Schenk, C., Hadjidemetriou, S., Winzler, S., Malich, A., Papageorgiou, I., 2017. Computer aided detection in prostate cancer diagnostics: a promising alternative to biopsy? A retrospective study from 104 lesions with histological ground truth. PLoS ONE 12, 1–21.

Toivonen, J., Perez, I.M., Movahedi, P., Merisaari, H., Pesola, M., Taimen, P., Boström, P.J., Pohjankukka, J., Kiviniemi, A., Pahikkala, T., Aronen, H.J., Jambor, I., 2019. Radiomics and machine learning of multisequence multiparametric prostate MRI: towards improved non-invasive prostate cancer characterization. PLoS ONE 14 (7), 1–23.

Verma, S., Sarkar, S., Young, J., Venkataraman, R., Yang, X., Bhavsar, A., Patil, N., Donovan, J., Gaitonde, K., 2016. Evaluation of the impact of computed high b–value diffusion-weighted imaging on prostate cancer detection. Abdom. Radiol. 41 (5), 934–945.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 338, 34–45.

Wang, S., Burtt, K., Turkbey, B., Choyke, P., Summers, R., 2014. Computer aided-diagnosis of prostate cancer on multiparametric MRI: a technical review of current research. BioMed. Res. Int. 2014.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biom. Bull. 1 (6), 80–83.

Woo, S., Suh, C.H., Kim, S.Y., Cho, J.Y., Kim, S.H., Moon, M.H., 2018. Head-to-head comparison between biparametric and multiparametric MRI for the diagnosis of prostate cancer: a systematic review and meta-analysis. Am. J. Roentgenol. 211, 226–241.

Woźnicki, P., Westhoff, N., Huber, T., Riffel, P., Froelich, M.F., Gresser, E., von Hardenberg, J., Mühlberg, A., Michel, M.S., Schoenberg, S.O., Nörenberg, D., 2020. Multiparametric MRI for prostate cancer characterization: combined use of radiomics model with PI-RADS and clinical parameters. Cancers 12 (7), 1–14.

Zelhof, B., Lowry, M., Rodrigues, G., Kraus, S., Turnbull, L., 2009. Description of magnetic resonance imaging-derived enhancement variables in pathologically confirmed prostate cancer and normal peripheral zone regions. BJU Int. 104, 621–627.

Zhong, X., Cao, R., Shakeri, S., Scalzo, F., Lee, Y., Enzmann, D.R., Wu, H.H., Raman, S.S., Sung, K., 2019. Deep transfer learning-based prostate cancer classification using 3 Tesla multi-parametric MRI. Abdom. Radiol. 44, 2030–2039.

Zhu, L., Gao, G., Liu, Y., Han, C., Liu, J., Zhang, X., Wang, X., 2020. Feasibility of integrating computer-aided diagnosis with structured reports of prostate multiparametric MRI. Clin. Imaging 60, 123–130.