Journal Pre-proof

End-to-end prostate cancer detection in bpMRI via 3D CNNs: effects of attention mechanisms, clinical priori and decoupled false positive reduction

Anindo Saha, Matin Hosseinzadeh, Henkjan Huisman

PII: S1361-8415(21)00201-2
DOI: https://doi.org/10.1016/j.media.2021.102155
Reference: MEDIMA 102155

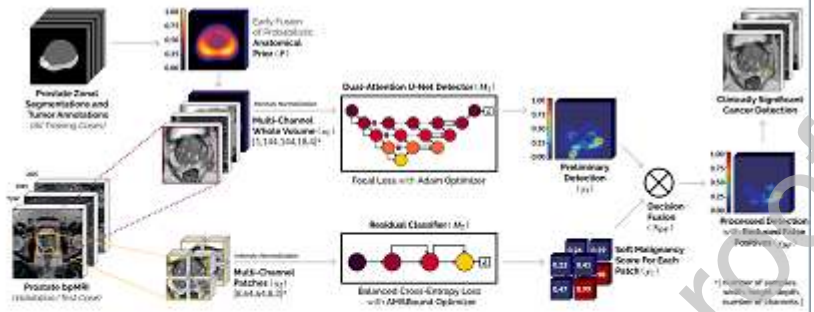
To appear in: *Medical Image Analysis*

Received date: 2 January 2021
Revised date: 30 May 2021
Accepted date: 8 June 2021

Please cite this article as: Anindo Saha, Matin Hosseinzadeh, Henkjan Huisman, End-to-end prostate cancer detection in bpMRI via 3D CNNs: effects of attention mechanisms, clinical priori and decoupled false positive reduction, *Medical Image Analysis* (2021), doi: https://doi.org/10.1016/j.media.2021.102155

Graphical Abstract

Proposed end-to-end 3D computer-aided detection and diagnosis framework for automated localization of clinically significant prostate cancer in bi-parametric MR imaging. Two ROIs are center-cropped from the multi-channel concatenation of the patient's T2W, DWI and ADC scans as the input for the model's 3D CNN sub-models (M_1, M_2). M_1 is a dual-attention detection network, that adaptively targets the most discriminate feature dimensions and spatially salient prostatic structures in the image volume. Additionally, it leverages an anatomical prior P in its input x_1 to synthesize domain-specific clinical priori and generate a preliminary detection map y_1. M_2 is a decoupled residual classifier that infers on a set of overlapping patches x_2 and maps them to a corresponding set of probabilistic malignancy scores y_2. Decision fusion node N_{DF} aggregates y_1, y_2 to produce the model output y_{DF}, i.e. a post-processed detection map with high sensitivity and reduced false positives.



HIGHLIGHTS

- Novel 3D CAD architecture for automated localization of cancer in prostate MRI.
- Dual-attention mechanisms to discriminate clinically significant prostate cancer.
- Decoupled false positive reduction, while retaining high detection sensitivity.
- Anatomical prior to infuse domain-specific clinical knowledge into learning cycle.
- Trained to detect histologically-confirmed cancer using large cohort of 1950 bpMRI scans with radiologically-estimated PI-RADS v2 annotations.
- Moderate agreement with expert radiologists and independent pathologists in patient-based diagnosis.

Medical Image Analysis (2021)

Contents lists available at ScienceDirect

Medical Image Analysis

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

End-to-end prostate cancer detection in bpMRI via 3D CNNs: effects of attention mechanisms, clinical priori and decoupled false positive reduction

Anindo Saha[a,*], Matin Hosseinzadeh[a,*], Henkjan Huisman[a]

[a]*Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen 6525 GA, The Netherlands*

ARTICLE INFO

Article history:

Keywords: prostate cancer, magnetic resonance imaging, convolutional neural network, computer-aided detection and diagnosis, anatomical prior, deep attention

ABSTRACT

We present a multi-stage 3D computer-aided detection and diagnosis (CAD) model[1] for automated localization of clinically significant prostate cancer (csPCa) in bi-parametric MR imaging (bpMRI). Deep attention mechanisms drive its detection network, targeting salient structures and highly discriminative feature dimensions across multiple resolutions. Its goal is to accurately identify csPCa lesions from indolent cancer and the wide range of benign pathology that can afflict the prostate gland. Simultaneously, a decoupled residual classifier is used to achieve consistent false positive reduction, without sacrificing high sensitivity or computational efficiency. In order to guide model generalization with domain-specific clinical knowledge, a probabilistic anatomical prior is used to encode the spatial prevalence and zonal distinction of csPCa. Using a large dataset of 1950 prostate bpMRI paired with radiologically-estimated annotations, we hypothesize that such CNN-based models can be trained to detect biopsy-confirmed malignancies in an independent cohort.

For 486 institutional testing scans, the 3D CAD system achieves 83.69±5.22% and 93.19±2.96% detection sensitivity at 0.50 and 1.46 false positive(s) per patient, respectively, with 0.882±0.030 AUROC in patient-based diagnosis –significantly outperforming four state-of-the-art baseline architectures (U-SEResNet, UNet++, nnU-Net, Attention U-Net) from recent literature. For 296 external biopsy-confirmed testing scans, the ensembled CAD system shares moderate agreement with a consensus of expert radiologists (76.69%; *kappa* = 0.51±0.04) and independent pathologists (81.08%; *kappa* = 0.56±0.06); demonstrating strong generalization to histologically-confirmed csPCa diagnosis.

1. Introduction

Prostate cancer (PCa) is one of the most prevalent cancers in men worldwide. It is estimated that as of January, 2019, over 45% of all men living with a history of cancer in the United States had suffered from PCa (Miller et al., 2019). One of the main challenges surrounding the accurate diagnosis of PCa is its broad spectrum of clinical behavior. PCa lesions can range from low-grade, benign tumors that never progress into clinically significant disease to highly aggressive, invasive malignancies, i.e. clinically significant PCa (csPCa), that can rapidly advance towards metastasis and death (Johnson et al., 2014). In clinical practice, prostate biopsies are used to histologically assign a Gleason Score (GS) and Gleason Grade Group (GGG)

*Authors with equal contribution to this research.
e-mail: anindya.shaha@radboudumc.nl (Anindo Saha)

[1]Source code and algorithm have been made publicly available at:
https://github.com/DIAGNijmegen/prostateMR_3D-CAD-csPCa
https://grand-challenge.org/algorithms/prostate-mri-cad-cspca

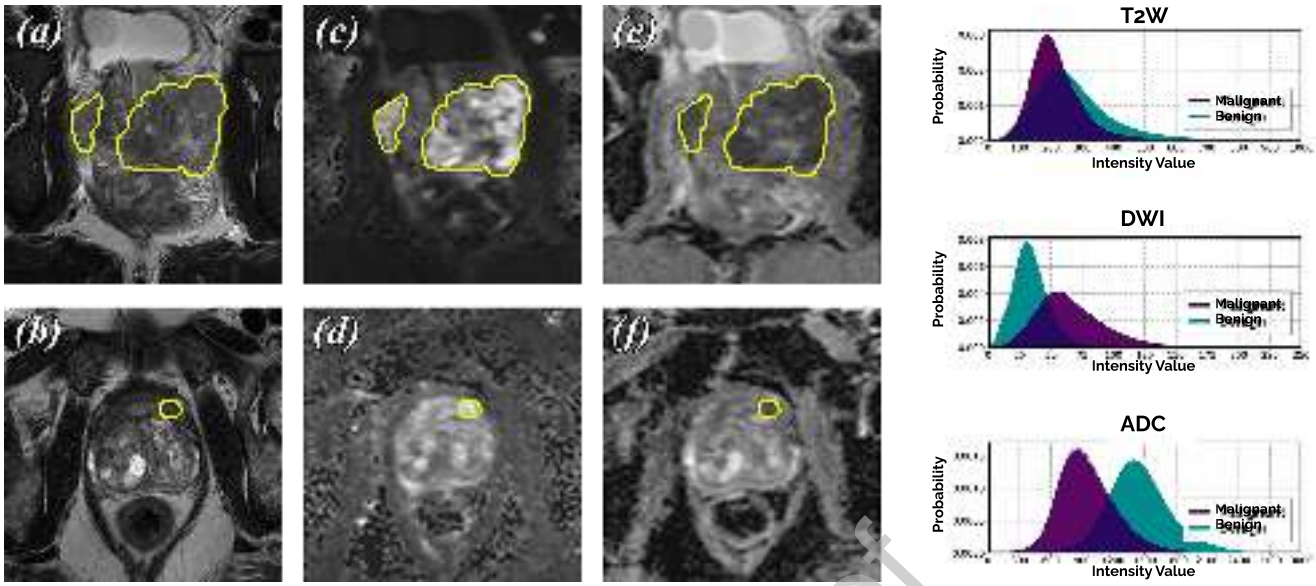


Fig. 1. The challenge of discriminating csPCa due to its morphological heterogeneity. *(a-b)* T2-weighted imaging (T2W), *(c-d)* diffusion-weighted imaging (DWI) and *(e-f)* apparent diffusion coefficient (ADC) maps constituting the prostate bpMRI scans for two different patients are shown above, where yellow contours indicate csPCa lesions. While one of the patients had large, severe csPCa developing from both ends *(top row)*, the other was afflicted by a single, relatively focal csPCa lesion surrounded by perceptually similar nodules of benign prostatic hyperplasia (BPH) *(bottom row)*. Furthermore, probability density functions *(right)* compiled from all 2732 scans used in this study, revealed a large overlap between the distributions of csPCa and non-malignant prostatic tissue across all three MRI channels.

to each lesion as a measure of cancer aggressiveness (Epstein et al., 2016). Non-targeted transrectal ultrasound (TRUS) is generally employed to guide biopsy extractions, but it remains severely prone to an underdetection of csPCa and overdiagnosis of indolent PCa (Verma et al., 2017). Prostate MR imaging can compensate for these limitations of TRUS (Johnson et al., 2014; Isral et al., 2020; Engels et al., 2020), where negative MRI can rule out unnecessary biopsies by 23–45% (Kasivisvanathan et al., 2018; van der Leest et al., 2019; Elwenspoek et al., 2019; Rouvire et al., 2019). Prostate Imaging Reporting and Data System: Version 2 (PI-RADS v2) (Weinreb et al., 2016) is the standard guideline for reading and acquiring prostate MRI, but it follows a qualitative and semi-quantitative assessment that mandates substantial expertise for proper usage. Meanwhile, csPCa can manifest as multifocal lesions of different shapes and sizes, bearing a strong resemblance to numerous non-malignant conditions (as seen in Fig. 1). In the absence of experienced radiologists, these factors can lead to low inter-reader agreement (<50%) and sub-optimal interpretation (Garcia-Reyes et al., 2015; Rosenkrantz et al., 2016; Smith et al., 2019; Westphalen et al., 2020). The development of proficient and reliable csPCa detection algorithms has therefore become an important research focus in medical image computing.

1.1. Related Work

The advent of deep convolutional neural networks (CNN) has paved the way for powerful computer-aided detection and diagnosis (CAD) systems that rival human performance (Esteva et al., 2017; McKinney et al., 2020). Machine learning models are increasingly applied for PCa detection, leveraging the high soft-tissue contrast and rich blend of anatomical and functional information present in prostate MRI.

Multiple studies have explored architectural enhancements to extend functionality. Cao et al. (2019a) proposed a hybrid 2D network titled *FocalNet* for joint csPCa detection and GS prediction. Over 5-fold cross-validation using 417 prostatectomy patient scans, *FocalNet* achieved 0.81 AUROC in csPCa diagnosis and 87.9% sensitivity at 1.0 false positive per patient, exclusively for MRI slices containing lesions. Yu et al. (2020a) proposed a two-stage 2D U-Net for csPCa detection, where the auxiliary second-stage module reduces false positives with contextual data from neighboring slices. Meanwhile, Sanyal et al. (2020) proposed a two-stage CNN model, where the first-stage segments and registers the prostate gland, while the second-stage 2D U-Net performs binary classification of csPCa (0.86 pixel-level AUROC on 20 unseen histologically-confirmed cases). Alternatively, Bhattacharya et al. (2020) proposed a two-stage CNN model, which extracts correlated cancer signatures from prostate MRI combined with its whole-mount histopathology mapping (0.86 pixel-level AUROC on 20 unseen prostatectomy patients).

Cancerous lesions stemming from the prostatic peripheral zone (PZ) exhibit different morphology and pathology than those developing from the transitional zone (TZ) (Chen et al., 2000; Weinreb et al., 2016; Isral et al., 2020). Hosseinzadeh et al. (2019) highlights the merits of utilizing this priori through an early fusion of probabilistic zonal segmentations inside a 2D CAD system –demonstrating how the inclusion of PZ and TZ segmentations can introduce an average increase of 5.3% detection sensitivity, between 0.5–2.0 false positives per patient. In a separate study, Cao et al. (2019b) constructed a probabilistic 2D prevalence map from 1055 MRI slices. Depicting the typical sizes, shapes and locations of malignancy across the prostate anatomy, this map was used to weakly supervise a 2D

U-Net for PCa detection. Both methods underline the value of clinical priori and anatomical features –factors known to play an equally important role in radiomics-based solutions (Litjens et al., 2014; Lematre et al., 2017).

The vast majority of CAD systems for csPCa operate solely on a 2D-basis, citing computational limitations and the non-isotropic imaging protocol of prostate MRI as their primary rationale. Yoo et al. (2019) tackled this challenge by employing dedicated 2D ResNets for each slice in a patient scan and aggregating all slice-level predictions with a Random Forest classifier (0.84 patient-level AUROC on 108 unseen biopsy-confirmed cases). Aldoj et al. (2020) proposed a patch-based approach, passing highly-localized regions of interest (ROI) through a standard 3D CNN (0.90 AUROC over 8-fold cross-validation using 200 biopsy-confirmed patients). Alkadi et al. (2019) followed a 2.5D approach as a compromise solution, sacrificing the ability to harness multiple MRI channels for an additional pseudo-spatial dimension.

In recent years, a number of retrospective studies have investigated the growing potential of CAD systems relative to radiologists. Sanford et al. (2020) compared the PI-RADS classification performance of a four-class 2D ResNet with expert radiologists, reaching 56% agreement on 68 testing scans. Schelb et al. (2019) used an ensemble of 2D U-Nets to achieve statistically similar csPCa detection performance as a cohort of trained radiologists on 62 biopsy-confirmed testing scans. Seetharaman et al. (2021) developed a 2.5D HED architecture, reaching 0.75 and 0.80 lesion-level AUROC on 293 biopsy-confirmed and 23 prostatectomy patient cases, respectively –approaching radiologists performance. Studies have also investigated the robustness of CAD systems across multi-institutional, multi-vendor testing sets (Castillo et al., 2021), reaching 0.93 AUROC on 40 biopsy-confirmed and 7 prostatectomy patient cases collected across five institutions (Sumathipala et al., 2018).

1.2. Contributions

In this research, we harmonize several state-of-the-art techniques from recent literature to present a novel end-to-end 3D CAD system that generates voxel-level detections of csPCa in prostate MRI. Key contributions of our study are, as follows:

- We examine a detection network with dual-attention mechanisms, which can adaptively target highly discriminative feature dimensions and salient prostatic structures in bpMRI, across multiple resolutions, to reach peak detection sensitivity at lower false positive rates.
- We study the effect of employing a residual patch-wise 3D classifier for decoupled false positive reduction and we investigate its utility in improving baseline specificity, without sacrificing high detection sensitivity.
- We develop a probabilistic anatomical prior, capturing the spatial prevalence and zonal distinction of csPCa from a large training dataset of 1584 MRI scans. We investigate the impact of encoding the computed prior into our CNN architecture and we evaluate its ability to guide model generalization with domain-specific clinical knowledge.
- We hypothesize that our model can train on radiologically-estimated annotations, begin to generalize beyond, and in turn, accurately detect histologically-confirmed malignancies, given a large training cohort with rich information. We evaluate performance across large, multi-institutional testing datasets: 486 institutional and 296 independent patient scans annotated using PI-RADS v2 and histological grades, respectively. Our benchmark includes a consensus score of expert radiologists to assess clinical viability.

2. Material and Methods

2.1. Dataset

The primary dataset was a cohort of 2436 consecutive prostate MRI studies (2317 patients) from Radboud University Medical Center (RUMC), acquired over the period January, 2016–January, 2018. All cases were paired with radiologically-estimated annotations of csPCa derived via PI-RADS v2. From here, 1584 (65%), 366 (15%) and 486 (20%) scans were split into training, validation and testing (TS1) sets, respectively, via double-stratified sampling –preserving the same class balance (*benign* or *malignant*) while ensuring non-overlapping patients, between each subset of data. Additionally, 296 prostate MRI scans (296 patients) from Ziekenhuisgroep Twente (ZGT), acquired over the period March, 2015–January, 2017, were used to curate an external testing set (TS2). TS2 annotations included biopsy-confirmed histological grades (GS, GGG).

2.1.1. Bi-Parametric MRI Scans

Patients were biopsy-naive men (RUMC: {median age: 66 yrs, IQR: 61–70}, ZGT: {median age: 65 yrs, IQR: 59–68}) with elevated levels of PSA (RUMC: {median level: 8 ng/mL, IQR: 5–11}, ZGT: {median level: 6.6 ng/mL, IQR: 5.1–8.7}). Imaging was performed on 3T MR scanners with surface coils (RUMC: {89.9% on Magnetom Trio/Skyra, 10.1% on Prisma}, ZGT: {100% on Skyra}; Siemens Healthineers, Erlangen). Acquisitions were obtained following standard mpMRI protocols in compliance with PI-RADS v2 (Engels et al., 2020). Given the limited role of dynamic contrast-enhanced (DCE) imaging in mpMRI, in recent years, bpMRI has emerged as a practical alternative –achieving similar performance, while saving time and the use of contrast agents (Turkbey et al., 2019; Brancato et al., 2020; Bass et al., 2020). Similarly, in this study, we used bpMRI sequences only, which included T2-weighted (T2W) and diffusion-weighted imaging (DWI). Apparent diffusion coefficient (ADC) maps and high b-value DWI (b≥1400 s/mm^2) were computed from the raw DWI scans. Due to the standardized precautionary measures (e.g. minimal temporal difference between acquisitions, administration of antispasmodic agents to reduce bowel motility, use of rectal catheter to minimize distension, etc.) (Engels et al., 2020) taken in the imaging protocol, we observed negligible patient motion across the different sequences. Thus, no additional registration techniques were applied, in agreement with clinical recommendations (Epstein et al., 2016) and recent studies (Cao et al., 2019a). Further details on patient demographics, study inclusion/exclusion criteria and acquisition parameters can be found in the *Supplementary Materials*.

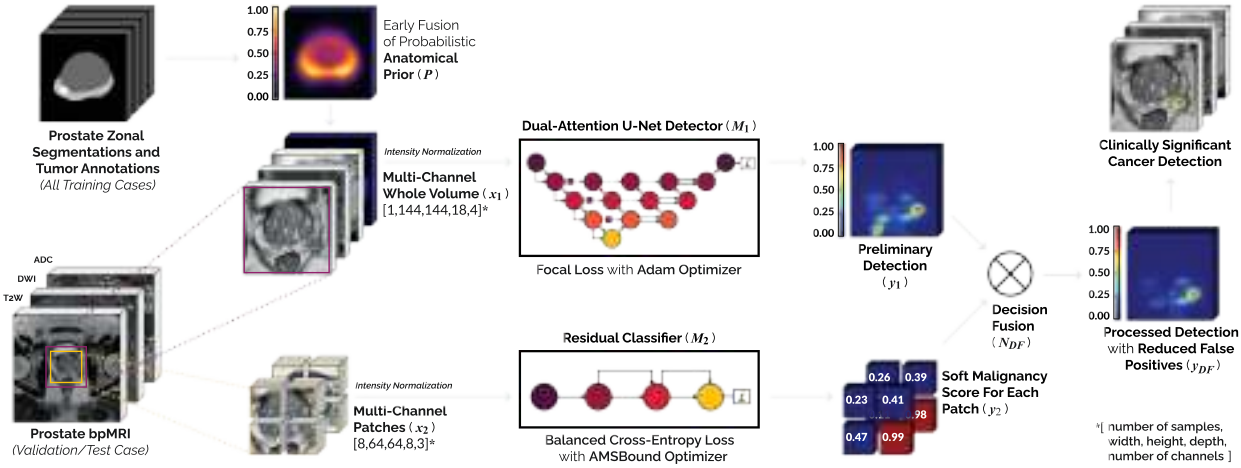


Fig. 2. Proposed end-to-end framework for computing voxel-level detections of csPCa in validation/test samples of prostate bpMRI. The model center-crops two ROIs from the multi-channel concatenation of the patient's T2W, DWI and ADC scans for the input of its detection and classification 3D CNN sub-models (M_1, M_2). M_1 leverages an anatomical prior P in its input x_1 to synthesize spatial priori and generate a preliminary detection y_1. M_2 infers on a set of overlapping patches x_2 and maps them to a set of probabilistic malignancy scores y_2. Decision fusion node N_{DF} aggregates y_1, y_2 to produce the model output y_{DF} in the form of a post-processed csPCa detection map with high sensitivity and reduced false positives.

2.1.2. Clinical Annotations

All patient cases were read during regular clinical routine via PI-RADS v2. At RUMC, all cases were read by at least one of six radiologists (4–25 years of experience) and difficult cases were jointly examined with an expert radiologist (25 years of experience with prostate MRI). At ZGT, all cases were read by two radiologists (6, 24 years of experience) and independently reviewed by two expert radiologists (5, 25 years of experience with prostate MRI) in consensus. In this study, we flagged any detected lesions marked PI-RADS 4 or 5 as csPCa$^{(PR)}$. All patients at ZGT underwent TRUS-guided biopsies performed by a urologist, blinded to the imaging results. In the presence of any suspicious lesions (PI-RADS> 2), patients also underwent in-bore MRI-guided biopsies. All tissue samples were graded by general pathologists (as detailed in van der Leest et al., 2019) and independently reviewed by an experienced uropathologist (25 years of experience), where cores containing cancer were assigned GS and GGG (in compliance with Epstein et al., 2016). Any lesion graded GS > 3+3 or GGG > 1 was marked as csPCa$^{(GS)}$. All instances of csPCa$^{(PR)}$ and csPCa$^{(GS)}$ were carefully delineated on a voxel-level basis by trained students (6–18 months of expertise), under the supervision of an experienced radiologist (7 years of experience).

Upon complete annotation, the RUMC and ZGT datasets contained 1527 and 210 *benign* cases, along with 909 and 86 *malignant* cases (≥ 1 csPCa lesion), respectively. Moreover, on a lesion-level basis, the RUMC dataset contained 1092 csPCa$^{(PR)}$ lesions (mean frequency: 1.21 lesions per *malignant* scan; median size: 1.05 cm^3, range: 0.01–61.49 cm^3), while the ZGT dataset contained 97 csPCa$^{(GS)}$ lesions (mean frequency: 1.05 lesions per *malignant* scan; median size: 1.69 cm^3, range: 0.23–22.61 cm^3). Further details on the distribution of PI-RADS and GGG scores across the study population can be found in the *Supplementary Materials*.

2.1.3. Prostate Zonal Segmentations

Multi-class segmentations of prostatic TZ and PZ were generated for each scan using a multi-planar, anisotropic 3D U-Net from a separate study (Riepe et al., 2020). Trained using a subset of 40 patient scans from the RUMC training cohort, the network achieved an average Dice Similarity Coefficient (DSC) of 0.90±0.01, 0.85±0.02 and 0.63±0.03 for whole-gland, TZ and PZ segmentation, respectively, over 5×5 nested cross-validation (Cawley and Talbot, 2010). We used these zonal segmentations to construct and apply the anatomical prior (as detailed in Section 2.2.3). For this study, the goal of the zonal segmentations was to establish object-level, prior-to-image correspondence, rather than voxel-level matching. Thus, high quality segmentations with precise contour definitions were not mandatory.

2.2. Model Architecture

The architecture of our proposed CAD solution comprises of two parallel 3D CNNs (M_1, M_2) followed by a decision fusion node N_{DF}, as shown in Fig. 2. Based on our observations in previous work (Hosseinzadeh et al., 2019; Riepe et al., 2020), we opted for anisotropically-strided 3D convolutions in both M_1 and M_2 to process the bpMRI data, which resemble multi-channel stacks of 2D images rather than full 3D volumes. Prior to usage, all acquisitions were spatially resampled to a common axial in-plane resolution of 0.5 mm^2 and slice thickness of 3.6 mm via B-spline interpolation. T2W and DWI channels were normalized to zero mean and unit standard deviation, while ADC channels were linearly normalized from [0,3000] to [0,1] in order to retain their clinically relevant numerical significance (Isral et al., 2020). Anatomical prior P, constructed using the prostate zonal segmentations and csPCa$^{(PR)}$ annotations in the training dataset, is encoded in M_1 to infuse spatial priori. At train-time, M_1 and M_2 are independently optimized using different loss functions and target labels. At test-time, N_{DF} is used to aggregate their predictions (y_1, y_2) into a single output detection map y_{DF}.

2.2.1. Detection Network

The principal component of our proposed model is the dual-attention detection network or M_1, as shown in Fig. 2, 3. It is used to generate the preliminary voxel-level detection of csPCa in prostate bpMRI scans with high sensitivity. Typically, a prostate gland occupies 45–50 cm^3, but it can be significantly enlarged in older males and patients afflicted by BPH (Basillote et al., 2003). The input ROI of M_1, measuring 144×144×18 voxels per channel or nearly 336 cm^3, includes and extends well beyond this window to utilize surrounding peripheral and global anatomical information. M_1 trains on whole-image volumes equivalent to its total ROI, paired with fully delineated annotations of csPCa$^{(PR)}$ as target labels. Since the larger ROI and voxel-level labels contribute to a severe class imbalance (1:153) at train-time, we use the focal loss function to train M_1. Focal loss addresses extreme class imbalance in one-stage dense detectors by weighting the contribution of easy to hard examples, alongside conventional class-weighting (Lin et al., 2017). In a similar study for joint csPCa detection in prostate MRI, the authors credited focal loss as one of the pivotal enhancements that enabled their CNN solution, titled *FocalNet* (Cao et al., 2019a).

For an input volume, $x_1 = ({x_1}^1, {x_1}^2, ..., {x_1}^n)$ derived from a given scan, let us define its target label $Y_1 = ({Y_1}^1, {Y_1}^2, ..., {Y_1}^n) \in \{0, 1\}$, where n represents the total number of voxels in x_1. We can formulate the focal loss function of M_1 for a single voxel in each scan, as follows:

$$FL({x_1}^i, {Y_1}^i) = -\alpha(1-{y_1}^i)^\gamma {Y_1}^i \log {y_1}^i - (1-\alpha)({y_1}^i)^\gamma (1-{Y_1}^i)\log(1-{y_1}^i) \quad i \in [1, n]$$

Here, ${y_1}^i = p(O{=}1|{x_1}^i) \in [0, 1]$, represents the probability of ${x_1}^i$ being a *malignant* tissue voxel as predicted by M_1, while α and γ represent weighting hyperparameters of the focal loss. At test-time, $y_1 = ({y_1}^1, {y_1}^2, ..., {y_1}^n) \in [0, 1]$, i.e. a voxel-level, probabilistic csPCa detection map for x_1, serves as the final output of M_1 for each scan.

We choose 3D U-Net (Ronneberger et al., 2015; Çiçek et al., 2016) as the base architecture of M_1, for its ability to summarize multi-resolution, global anatomical features (Dalca et al., 2018; Isensee et al., 2020) and generate an output detection map with voxel-level precision. Pre-activation residual blocks (He et al., 2016) are used at each scale of M_1 for deep feature extraction. Architecture of the decoder stage is adapted into that of a modified UNet++ (Zhou et al., 2020) for improved feature aggregation. UNet++ uses redesigned encoder-decoder skip connections that implicitly enable a nested ensemble configuration. In our adaptation, its characteristic property of feature fusion from multiple semantic scales is used to achieve similar performance, while dense blocks and deep supervision from the original design are forgone to remain computationally lightweight.

Two types of differentiable, soft attention mechanisms are employed in M_1 to highlight salient information throughout the training process, without any additional supervision. Channel-wise *Squeeze-and-Excitation* (SE) attention (Hu et al., 2019; Rundo et al., 2019) is used to amplify the most discriminative feature dimensions at each resolution. Grid-attention gates (Schlemper et al., 2019) are used to automatically learn spatially important prostatic structures of varying shapes and sizes. While the former is integrated into every residual block to guide feature extraction, the latter is placed at the start of skip-connections to filter the semantic features being passed onto the decoder. During backpropagation, both attention mechanisms work collectively to suppress gradients originating from background voxels and inessential feature maps. Similar combinations of dual-attention mechanisms have reached state-of-the-art performance in semantic segmentation challenges (Fu et al., 2019) and PCa diagnosis (Yu et al., 2020b), sharing an ability to integrate local features with their global dependencies.

2.2.2. Classifier for False Positive Reduction

The goal of the classification network, M_2, is to improve overall model specificity via independent, binary classification of each scan and its constituent segments. It is effectuated by N_{DF}, which factors in these predictions from M_2 to locate and penalize potential false positives in the output of M_1. M_2 has an input ROI of 112×112×12 voxels per channel or nearly 136 cm^3, tightly centered around the prostate. While training on the full ROI volume has the advantage of exploiting extensive spatial context, it results in limited supervision by the usage of a single coarse, binary label per scan. Thus, we propose patch-

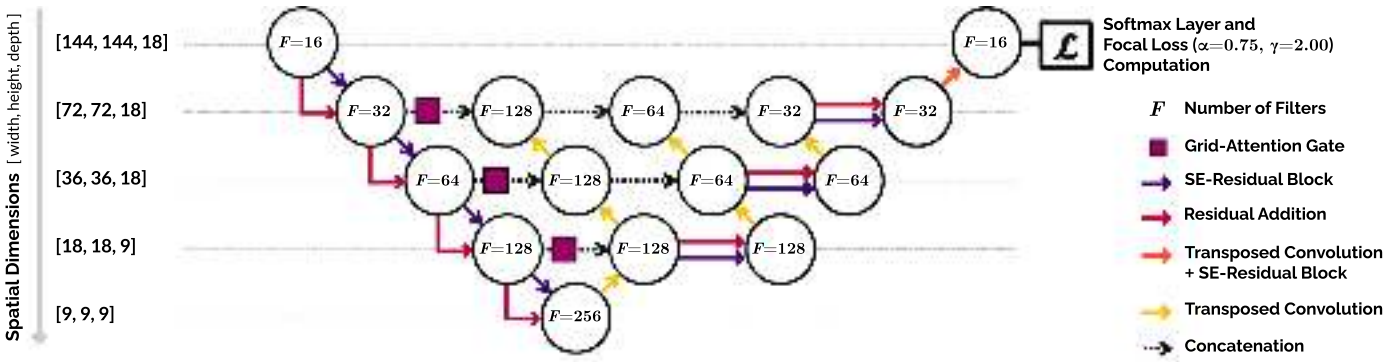


Fig. 3. Architecture schematic for the Dual-Attention U-Net (M_1). M_1 is a modified adaptation of the UNet++ architecture (Zhou et al., 2020), utilizing a pre-activation residual backbone (He et al., 2016) with *Squeeze-and-Excitation* (SE) channel-wise attention mechanism (Hu et al., 2019) and grid-attention gates (Schlemper et al., 2019). All convolutional layers in the encoder and decoder stages are activated by ReLU and LeakyReLU, respectively, and use kernels of size $3 \times 3 \times 3$ with L_2 regularization ($\beta = 0.001$). Both downsampling and upsampling operations throughout the network are performed via anisotropic strides. Dropout nodes ($rate = 0.50$) are connected at each scale of the decoder to alleviate train-time overfitting.

wise training using multiple, localized labels, to enforce fully supervised learning. We define an effective patch extraction policy as one that samples regularly across the ROI to densely cover all spatial positions. Sampled patches must also be large enough to include a sufficient amount of context for subsequent feature extraction. Random sampling within a small window, using the aforementioned criteria, poses the risk of generating highly overlapping, redundant training samples. However, a minimum level of overlap can be crucial, benefiting regions that are harder to predict by correlating semantic features from different surrounding context (Xiao et al., 2018). To facilitate these conditions, the ROI is uniformly divided into a set of eight octant training samples x_2, measuring $64 \times 64 \times 8$ voxels each, with an overlap stride of 16 and 4 voxels across their in-plane and through-plane dimensions, respectively. As a result, training samples consistently and deterministically cover the total ROI, while 71.85% of each patch overlaps with a portion of its seven neighboring patches.

For input patches, $x_2 = ({x_2}^1, {x_2}^2, ..., {x_2}^8)$ derived from a given scan, let us define its set of target labels $Y_2 = ({Y_2}^1, {Y_2}^2, ..., {Y_2}^8) \in \{0, 1\}$. Using a pair of complementary class weights to adjust for the patch-level class imbalance (1:4), we formulate the balanced cross-entropy loss function of M_2 for a single patch in each scan, as follows:

$$BCE({x_2}^i, {Y_2}^i) = -\beta {Y_2}^i \log {y_2}^i - (1-\beta)(1-{Y_2}^i)\log(1-{y_2}^i) \quad i \in [1, 8]$$

Here, ${y_2}^i = p(O{=}1|{x_2}^i) \in [0, 1]$, represents the probability of ${x_2}^i$ being a *malignant* patch as predicted by M_2. At test-time, $y_2 = ({y_2}^1, {y_2}^2, ..., {y_2}^8) \in [0, 1]$, i.e. a set of probabilistic malignancy scores for x_2, serves as the final output of M_2 for each scan.

Transforming voxel-level annotations into patch-wise labels can introduce additional noise in the target labels used at train-time. For instance, a single octant patch contains $64 \times 64 \times 8$ or 32768 voxels per channel. In a naive patch extraction system, if the fully delineated ground-truth for this sample includes even a single voxel of *malignant* tissue, then the patch-wise label would be inaccurately set as *malignant*, despite a voxel-level imbalance of 1:32767 supporting the alternate class. Such a training pair carries high label noise and proves detrimental to the learning cycle, where the network associates semantic features to the wrong target class. Therefore, we define a constraint τ, representing the minimum percentage of *malignant* tissue voxels required to assign the label *malignant*.

For M_2, we consider CNN architectures based on residual learning for feature extraction, due to their modularity and continued success in supporting state-of-the-art segmentation and detection performance in the medical domain (Yoo et al., 2019; McKinney et al., 2020; Jiang et al., 2020).

2.2.3. Decision Fusion

The goal of the decision fusion node N_{DF} is to aggregate M_1 and M_2 predictions (y_1, y_2) into a single output y_{DF}, which retains the same sensitivity as y_1, but improves specificity by reducing false positives. False positives in y_1 are fundamentally clusters of positive values located in the *benign* regions of the scan. N_{DF} employs y_2 as a means of identifying these regions. We set a threshold T_P on $(1 - {y_2}^i)$ to classify each patch ${x_2}^i$, where $i \in$[1,8]. T_P represents the minimum probability required to classify ${x_2}^i$ as a *benign* patch. A high value of T_P adapts M_2 as a highly sensitive classifier that yields very few false negatives, if any at all. Once all *benign* regions have been identified, any false positives within these patches are suppressed by multiplying their corresponding regions in y_1 with a penalty factor λ. Resultant detection map y_{DF}, i.e. essentially a post-processed y_1, serves as the final output of our proposed CAD system. Limiting N_{DF} to a simple framework of two hyperparameters alleviates the risk of overfitting. While T_P addresses which patches should be considered for false positive reduction, λ addresses what factor candidate false positives should be suppressed by, within these selected patches. An appropriate combination of T_P and λ can either suppress clear false positives or facilitate an aggressive reduction scheme at the expense of fewer true positives in y_{DF}. In this research, we opted for the former policy to retain maximum csPCa detection sensitivity. Optimal values of T_P and λ were determined to be 0.98 and 0.90, respectively, via a coarse-to-fine hyperparameter grid search (detailed in the *Supplementary Materials*).

2.2.4. Anatomical Prior

Parallel to recent studies in medical image computing (Gibson et al., 2018; Dalca et al., 2018; Wachinger et al., 2018; Cao et al., 2019b; Faryna et al., 2021) on infusing clinical priori into CNN architectures, we hypothesize that M_1 can benefit from an explicit anatomical prior for csPCa detection in bpMRI. To this end, we construct a probabilistic population prior $P \in [0, 1]$, measuring $144 \times 144 \times 18$ voxels, as introduced in our previous work (Saha et al., 2020). P captures the spatial prevalence and zonal distinction of csPCa using 1584 radiologically-estimated csPCa$^{(PR)}$ annotations and CNN-generated prostate zonal segmentations, respectively, from the training dataset. We opt for an early fusion technique to encode the priori (Hosseinzadeh et al., 2019), where P is concatenated as an additional channel to every input scan passed through M_1, thereby guiding its learning cycle as a spatial weight map embedded with domain-specific clinical knowledge (refer to Fig. 2). Prior-to-image correspondence is established at both train-time and inference by using case-specific prostate segmentations to translate, orient and scale P (via matching object centroids and maximizing volumetric overlap), accordingly, for each bpMRI scan.

2.3. Experimental Design

Several experiments were conducted to statistically evaluate performance and analyze the design choices throughout the end-to-end model. We facilitated a fair comparison by maintaining an identical preprocessing, augmentation, tuning and train-validation pipeline for each candidate system in a given experiment. Patient-based diagnosis performance was evaluated using the Receiver Operating Characteristic (ROC), where the area under ROC (AUROC) was estimated from the normalized Wilcoxon/Mann-Whitney U statistic (Hanley and McNeil, 1982). Lesion-level performance was evaluated using the Free-Response Receiver Operating Characteristic (FROC) to address PCa multifocality. Detections sharing a minimum DSC with

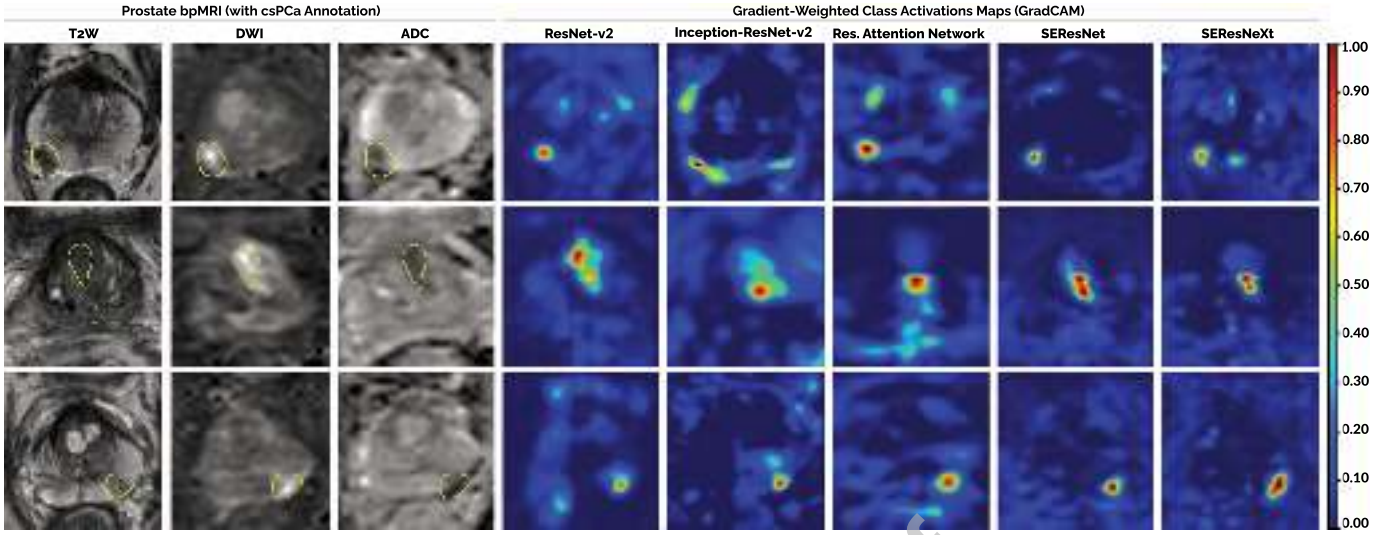


Fig. 4. Model interpretability of the candidate CNN architectures for classifier M_2 at τ =0.1%. Gradient-weighted class activation maps (GradCAM) and their corresponding T2W, DWI and ADC scans for three patient cases from the validation set are shown above. Each case included a single instance of csPCa$^{(PR)}$ located in the prostatic TZ (*center row*) or PZ (*top, bottom rows*), as indicated by the yellow contours. Whole-image GradCAMs were generated by restitching and normalizing (*min-max*) the eight patch-level GradCAMs generated per case. Maximum voxel-level activation was observed in close proximity of csPCa$^{(PR)}$, despite training each network using patch-level binary labels only.

Table 1. Patient-based diagnosis performance of the candidate CNN architectures and training schemes (whole-image versus patch-wise training with four different values of τ to regulate label noise) for classifier M_2. Performance scores indicate mean of 5-fold cross-validation, followed by 95% confidence intervals estimated as twice the standard deviation.

Model	Params	AUROC (Whole-Image)	AUROC (Patches)			
			τ = 0.0%	τ = 0.1%	τ = 0.5%	τ = 1.0%
ResNet-v2 (He et al., 2016)	0.089 M	0.819±0.018	0.830±0.010	0.844±0.011	0.868±0.013	0.897±0.008
Inception-ResNet-v2 (Szegedy et al., 2017)	6.121 M	0.823±0.017	0.822±0.014	0.860±0.015	0.883±0.009	0.905±0.008
Res. Attention Network (Wang et al., 2017)	1.233 M	0.826±0.024	0.837±0.012	0.850±0.007	0.876±0.008	0.901±0.008
SEResNet (Hu et al., 2019)	0.095 M	0.836±0.014	0.842±0.019	0.861±0.005	0.886±0.008	0.912±0.008
SEResNeXt (Hu et al., 2019)	0.128 M	0.820±0.022	0.833±0.013	0.843±0.005	0.875±0.009	0.896±0.012

their ground-truth delineations were considered true positives. Given that the vast majority of csPCa lesions are small (median volume < 2 cm^3 across both centers), have indistinct margins and share large inter-reader variability in their interpretation, we set this threshold as 0.10 DSC, in agreement with McKinney et al. (2020). All metrics were computed in 3D, across complete image volumes. Confidence intervals were estimated as twice the standard deviation from the mean of 5-fold cross-validation (applicable to validation sets) or 1000 replications of bootstrapping (applicable to testing sets) –as per standard practice (Cao et al., 2019a; McKinney et al., 2020). Statistically significant improvements were verified with a p-value on the difference in case-level AUROC and lesion-level sensitivity at clinically relevant false positive rates (0.5, 1.0) using 1000 replications of bootstrapping (Chihara et al., 2014). Bonferroni correction was used to adjust the significance level for multiple comparisons.

3. Results and Analysis

3.1. Effect of Architecture and Label Noise on Classification

To determine the effect of the classification architecture for M_2, five different 3D CNNs (ResNet-v2, Inception-ResNet-v2, Residual Attention Network, SEResNet, SEResNeXt) were implemented and tuned across their respective hyperparameters to maximize patient-based AUROC over 5-fold cross-validation. Furthermore, each candidate CNN was trained using whole-images and patches, in separate turns, to draw out a comparative analysis surrounding the merits of spatial context versus localized labels. In the latter case, we studied the effect of τ on patch-wise label assignment (refer to Section 2.2.2). We investigated four different values of τ: 0.0%, 0.1%, 0.5%, 1.0%; which correspond to minimum csPCa volumes of 9, 297, 594 and 1188 mm^3, respectively. Each classifier was assessed qualitatively via 3D GradCAMs (Selvaraju et al., 2017) to ensure adequate interpretability for clinical usage.

From the results noted in Table 1, we observed that the SEResNet architecture consistently scored the highest AUROC across every training scheme. However, in each case, its performance remained statistically similar ($p \geq 0.01$) to the other candidate models. We observed that a higher degree of supervision from patch-wise training proved more useful than the near 8× additional spatial context provided per sample during whole-image training. Increasing the value of τ consistently improved performance for all candidate classifiers (upto 10% in patch-

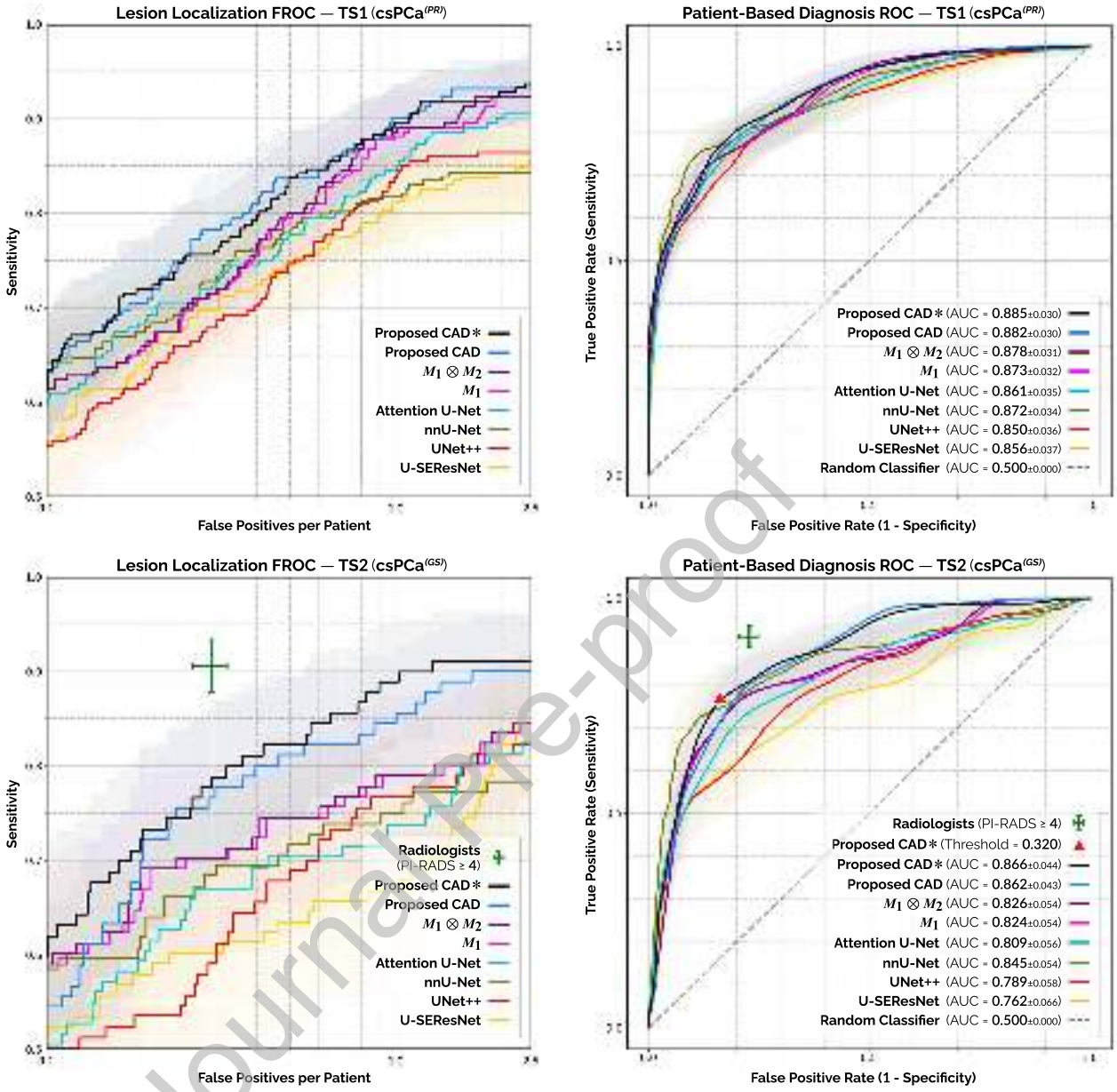


Fig. 5. Lesion-level FROC (*left*) and patient-based ROC (*right*) analyses of csPCa$^{(PR)}$ (*top row*) / csPCa$^{(GS)}$ (*bottom row*) detection sensitivity against the number of false positives generated per patient scan using the baseline, ablated and proposed detection models on the institutional testing set TS1 (*top row*) and the external testing set TS2 (*bottom row*). Transparent areas indicate the 95% confidence intervals. Mean performance for the consensus of expert radiologists and their 95% confidence intervals are indicated by the centerpoint and length of the green markers, respectively, where all observations marked PI-RADS 4 or 5 are considered positive detections (as detailed in Section 2.3).

level AUROC). While we attribute this improvement to lower label noise at train-time, it is important to note that the total csPCa volume per patient is typically small (refer to Section 2.1.2). If τ is set too large, not only are patch labels regulated, as intended, but multiple patch-level label swaps can compound to the point where entire patient cases can swap labels –resulting in an inaccurate evaluation. Such a phenomenon was observed for 9 patient cases across the 1950 training-validation scans at τ ={0.5, 1.0}%. At τ =0.1%, there were no cases of patient-level label swaps (as seen at τ ={0.5, 1.0}%), while patch-level AUROC still improved by nearly 2% relative to a naive patch extraction system (τ =0.0%). Thus, for the 3D CAD system, we chose the SEResNet patch-wise classifier trained at τ =0.1% as M_2. GradCAMs confirm that M_2 accurately targets csPCa$^{(PR)}$ lesions (if any) on a voxel-level basis, despite being trained on patch-level binary labels (as highlighted in Fig. 4). Further details regarding the network and training configurations of M_2 are listed in the *Supplementary Materials*.

3.2. Effect of Architecture and Clinical Priori on Detection

We analyzed the effect of the M_1 architecture, in comparison to the four baseline 3D CNNs (U-SEResNet, UNet++, nnU-Net, Attention U-Net) that inspire its design. We evaluated the

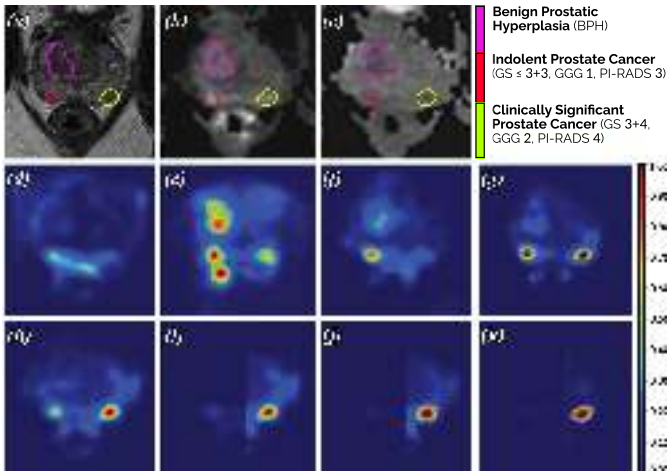


Fig. 6. (*a*) T2W, (*b*) DWI, (*c*) ADC scans for a patient case in the external testing set TS2, followed by its csPCa detection map as predicted by each candidate system: (*d*) U-SEResNet, (*e*) UNet++, (*f*) Attention U-Net, (*g*) nnU-Net, (*h*) M_1, (*i*) $M_1 \otimes M_2$, (*j*) proposed CAD, (*k*) proposed CAD∗. Three stand-alone detection networks (UNet++, nnU-Net, M_1) successfully identified the csPCa$^{(GS)}$ lesion, albeit with additional false positive(s). In the case of the proposed CAD/CAD∗ system, while the classifier in $M_1 \otimes M_2$ was able to suppresses these false positive(s) from M_1, inclusion of prior P further strengthened the confidence and boundaries of the true positive. At the same time, it should be noted that small lesions carry higher uncertainty in their grading –i.e. indolent GGG 1 lesions may create suspicious findings in MRI, while clinically significant GGG 2 lesions risk being undersampled during biopsy (Verma et al., 2017; Srivastava et al., 2019).

end-to-end 3D CAD system, along with the individual contributions of its constituent components (M_1, M_2, P), to examine the effects of false positive reduction and clinical priori. Additionally, we applied the ensembling heuristic of the nnU-Net framework (Isensee et al., 2020) to create CAD∗, i.e. an ensemble model comprising of multiple CAD instances, and we studied its impact on overall performance. Each candidate setup was tuned over 5-fold cross-validation and benchmarked on the testing datasets (TS1, TS2).

3.2.1. Generalization to Radiologically-Estimated csPCa

Lesion Localization: From the FROC analysis on the institutional testing set TS1 (refer to Fig. 5), we observed that M_1 reached 88.15±4.19% detection sensitivity at 1.0 false positive per patient, significantly ($p \leq 0.01$) outperforming the baseline U-SEResNet (81.18±4.99%), UNet++ (83.81±4.80%), nnU-Net (81.67±4.64%) and Attention U-Net (84.76±4.64%). With the addition of classifier M_2 to M_1 ($M_1 \otimes M_2$), upto 12.89% ($p \leq 0.001$) less false positives were generated per patient, while retaining the same maximum detection sensitivity (92.29%) as before. The working principle of $M_1 \otimes M_2$ is illustrated in Fig. 6 through a particularly challenging patient case, where the prostate gland is afflicted by multiple, simultaneously occurring conditions. With the inclusion of anatomical prior P in $M_1 \otimes M_2$, the 3D CAD system benefited from a further 3.14% increase in partial area under FROC (pAUC) between 0.10–2.50 false positives per patient, reaching 1.676±0.078 pAUC. At 0.5 false positive per patient, the 3D CAD system reached 83.69±5.22% detection sensitivity, surpassing the best baseline (nnU-Net) by 5.59% ($p \leq 0.001$), while detecting 4.10% ($p \leq 0.01$) and 3.63% ($p \leq 0.01$) more csPCa$^{(PR)}$ lesions than its component systems M_1 and $M_1 \otimes M_2$, respectively. It reached a maximum detection sensitivity of 93.19±2.96% at 1.46 false positives per patient, identifying a higher percentage of csPCa occurrences than all other candidate systems.

Patient-Based Diagnosis: From ROC analysis on the institutional testing set TS1 (refer to Fig. 5), we observed that the 3D CAD system reached 0.882±0.03 AUROC in case-level diagnosis, ahead of all other candidate systems by a margin of 0.4–3.2%. While it performed significantly better than the baseline U-SEResNet ($p \leq 0.01$), UNet++ ($p \leq 0.001$) and Attention U-Net ($p \leq 0.01$), its ability to discriminate between *benign* and *malignant* patient cases was statistically similar ($p \geq 0.01$) to the nnU-Net, M_1 and $M_1 \otimes M_2$.

3.2.2. Generalization to Histologically-Confirmed csPCa

Both the FROC and ROC analyses on the external testing set TS2 (refer to Fig. 5) indicate similar patterns emerging as those observed in Section 3.2.1, but with an overall decrease in performance. Given the near-identical MRI scanners and acquisition conditions employed between both institutions (refer to Section 2.1.1), we primarily attribute this decline to the disparity between the imperfect radiologically-estimated training annotations (csPCa$^{(PR)}$) and the histologically-confirmed testing annotations (csPCa$^{(GS)}$) in TS2 (refer to Section 3.3 for ra-

Table 2. Computational requirements (in terms of the number of trainable parameters, VRAM usage and the average time taken per patient scan during inference on a single NVIDIA RTX 2080 Ti) against the localization performance (in terms of the maximum csPCa detection sensitivity achieved and its corresponding false positive (FP) rate across both testing datasets) for each candidate detection system.

Model	Params	VRAM	Inference	Maximum Sensitivity {FP/Patient}	
				TS1 – csPCa$^{(PR)}$	TS2 – csPCa$^{(GS)}$
U-SEResNet (Hu et al., 2019)	1.615 M	0.94 GB	1.77±0.20 s	85.63%±4.70 {2.44}	84.42%±7.36 {2.26}
UNet++ (Zhou et al., 2020)	14.933 M	2.97 GB	1.79±0.19 s	86.41%±4.54 {1.74}	82.28%±7.62 {2.25}
nnU-Net (Isensee et al., 2020)	30.599 M	4.69 GB	2.09±0.03 s	84.34%±4.40 {1.44}	77.23%±8.14 {1.12}
Attention U-Net (Schlemper et al., 2019)	2.235 M	1.96 GB	1.77±0.19 s	90.46%±3.63 {2.07}	82.43%±7.79 {2.32}
Dual-Attention U-Net – M_1	15.250 M	3.01 GB	1.79±0.19 s	92.29%±3.24 {1.94}	84.60%±7.45 {2.31}
M_1 with *False Positive Reduction* – $M_1 \otimes M_2$	15.335 M	3.75 GB	1.89±0.23 s	92.29%±3.24 {1.69}	84.60%±7.45 {2.22}
$M_1 \otimes M_2$ with *Prior* – Proposed CAD	15.335 M	3.98 GB	1.90±0.23 s	93.19%±2.96 {1.46}	90.03%±5.80 {1.67}
Ensemble of CAD – Proposed CAD∗	40.069 M	9.85 GB	2.41±0.42 s	93.69%±3.13 {2.36}	91.05%±5.24 {1.29}

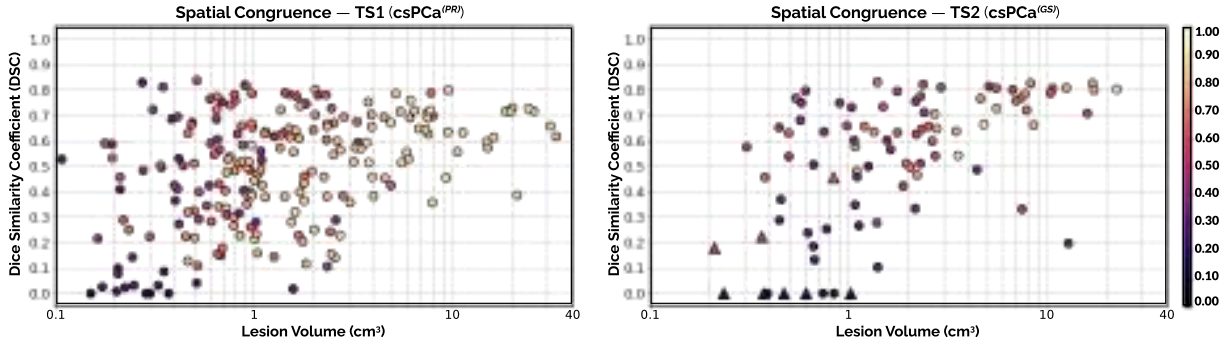


Fig. 7. Distribution of per-lesion Dice Similarity Coefficient (DSC) (relative to csPCa lesion volume) for CAD∗ detections against the ground-truth annotations of csPCa$^{(PR)}$ in the institutional testing TS1 *(left)* and csPCa$^{(GS)}$ in the external testing set TS2 *(right)*. All DSC values were computed in 3D for the model-specific operating point with maximum detection sensitivity (91.05±5.24%). Encoded color for each marker indicates its corresponding likelihood of malignancy, as predicted by CAD∗. Triangular markers for TS2 *(right)* indicate csPCa$^{(GS)}$ lesions missed by the radiologists.

diologists' performance). By comparing the relative drop in performance for each candidate model, we can effectively estimate their generalization and latent understanding of csPCa, beyond our provided training samples.

Lesion Localization: At 1.0 false positive per patient, the 3D CAD system achieved 85.55±7.04% detection sensitivity on TS2 (refer to Fig. 5), performing significantly better ($p \leq 0.001$) than the baseline U-SEResNet (66.74±9.65%), UNet++ (76.66±9.05%), nnU-Net (74.73±7.72%) and Attention U-Net (73.64±8.97%). It also detected 6.56% ($p \leq 0.005$) more csPCa$^{(GS)}$ lesions than its ablated counterparts M_1 and $M_1 \otimes M_2$, respectively. The 3D CAD system reached a maximum detection sensitivity of 90.03±5.80% at 1.67 false positives per patient, scoring higher than all other candidate systems. On average, all baseline models underwent 7–13% drops in detection sensitivity at 1.0 false positive per patient, relative to their performance on TS1. Similarly, the average detection sensitivities of M_1 and $M_1 \otimes M_2$ fell by nearly 10%. From the inclusion of P in $M_1 \otimes M_2$, this decline came down to only 3% for the 3D CAD system at the same false positive rate. Furthermore, an overall 11.54% increase in pAUC was observed between 0.10–2.50 false positives per patient, relative to $M_1 \otimes M_2$.

Patient-Based Diagnosis: 3D CAD reached 0.862±0.04 AUROC on TS2 (refer to Fig. 5), ahead of the baseline U-SEResNet, UNet++, nnU-Net and Attention U-Net by 10.0% ($p \leq 0.001$), 7.3% ($p \leq 0.001$), 1.7% ($p > 0.1$) and 5.3% ($p \leq 0.05$), respectively. Compared to TS1, the 3D CAD model underwent 2% decrease in AUROC, while all other candidate systems underwent an average reduction of 5–6%. Once again, the anatomical prior proved vital, enabling 3D CAD to outperform its immediate counterpart $M_1 \otimes M_2$ by 3.6% ($p \leq 0.05$).

3.2.3. Effect of Ensembling

The ensembled prediction of CAD∗ is the weighted-average output of three member models: 2D, 3D and two-stage cascaded 3D variants of the proposed CAD system (refer to *Supplementary Materials* for detailed implementation). In comparison to the standard 3D CAD system, CAD∗ carries 2.6× trainable parameters, occupies 2.5× VRAM for hardware acceleration and requires 1.3× inference time per patient scan (as noted in Table 2). In terms of its performance, CAD∗ demonstrated 0.3–0.4% improvement in patient-based AUROC across both testing datasets and shared statistically similar lesion localization on TS1. It generated a considerably large improvement in lesion detection on TS2, amounting to 4.01% increase in pAUC between 0.10–2.50 false positives per patient (refer to Fig 5), as well as a higher maximum detection sensitivity (91.05±5.24%) at a lower false positive rate (1.29) (as noted in Table 2).

3.3. Relative Performance to Consensus of Radiologists

To evaluate CAD∗ in comparison to the consensus of expert radiologists, we analyzed their relative performance on the external testing set TS2. Agreements in patient-based diagnosis were computed with Cohen's *kappa*.

Radiologists achieved 90.72±2.78% detection sensitivity at 0.30 false positives per patient and 91.11±2.67% sensitivity at 77.18±2.37% specificity in lesion localization and patient-based diagnosis, respectively (refer to Fig. 5). After binarizing its case-level detections using a probability threshold (>0.32) (as demonstrated by Schelb et al., 2019), the CAD∗ system reached 75.31±3.64% sensitivity at 85.83±2.22% specificity in patient-based diagnosis, where it shared 76.69% (227/296 cases; *kappa* = 0.51 ± 0.04) and 81.08% (240/296 cases; *kappa* = 0.56 ± 0.06) agreement with expert radiologists and independent pathologists, respectively. In comparison, radiologists shared 81.42% (241/296 cases; *kappa* = 0.61±0.05) agreement with pathologists. Pathologists often struggle to accurately differentiate GGG 1 from GGG 2 patterns in tissue specimens, resulting in ≥ 30% inter-reader variability (Egevad et al., 2013; Ozkan et al., 2016). As observed in TS2, this challenge persists across prostate MRI, where radiologists' performance dropped to 88.24±2.67% sensitivity and 51.67±3.45% specificity, while discriminating patients with GGG 1 lesions (64/296 cases) from those carrying GGG 2 lesions (42/296 cases), exclusively. At the same specificity, CAD∗ performed statistically similar ($p \geq 0.05$), reaching 81.76±10.72% sensitivity.

3.4. Spatial Congruence Analysis

On average, CAD∗ shared 0.49 ± 0.22 DSC and 0.58 ± 0.21 DSC on TS1 (csPCa$^{(PR)}$) and TS2 (csPCa$^{(GS)}$), respectively (refer to Fig. 7). In comparison, similar studies on csPCa$^{(GS)}$ detection from MRI reported 0.35–0.46 DSC (Artan and Yetik,

2012; Chung et al., 2015; Kohl et al., 2017; Schelb et al., 2019; d. Vente et al., 2021). Increasing model confidence and higher DSC values were observed consistently for larger lesions. However, DSC remains limited as an evaluation metric for detection –demonstrating high variance with minute changes in predicted contours and/or volume for smaller lesions (Carass et al., 2020). High DSC variance among smaller lesions could also be indicative of misaligned MRI sequences (potentially due to physical distortions from the low bandwidth of DWI) (Donato Jr et al., 2014). Unlike hard, binary masks; training CNNs with probabilistic target labels and output detections could potentially capture such uncertainty along lesion contours (Gros et al., 2021).

4. Discussion and Conclusion

We conclude that a detection network (M_1), harmonizing state-of-the-art attention mechanisms, can accurately discriminate more malignancies at the same false positive rate, than one without (refer to Section 3.2.1). Among four other recent adaptations of the 3D U-Net that are popularly used for biomedical segmentation, M_1 detected significantly more csPCa lesions at 1.00 false positive per patient and consistently reached the highest detection sensitivity on the testing datasets between 0.10–2.50 false positives per patient, closely followed by the Attention U-Net (refer to Fig. 5). As soft attention mechanisms continue to evolve, supporting ease of optimization, sharing equivariance over permutations (Goyal and Bengio, 2020) and suppressing gradient updates from inaccurate annotations (Wang et al., 2017; Min et al., 2019), deep attentive models, such as M_1, become increasingly more applicable for csPCa detection in bpMRI (Duran et al., 2020; Yu et al., 2020b).

We conclude that a residual patch-wise 3D classifier (M_2) can significantly reduce false positives, without sacrificing high sensitivity. In stark contrast to ensembling, which scaled up the number of trainable parameters nearly 3× for limited improvements in performance (refer to Section 3.2.3), M_2 produced flat increases in specificity (upto 12.89% less false positives per patient) across both testing datasets, while requiring less than 1% of the total parameters in the 3D CAD system (as noted in Table 2). Furthermore, as a decoupled classifier, M_2 shares two major advantages. Firstly, unlike the jointly-trained, cascaded approach proposed by Yu et al. (2020a), where the second-stage classifier was able to reduce false positives at the expense of nearly an 8% decrease in detection sensitivity, in our case, the effect of M_2 on the overall 3D CAD system could be controlled via the decision fusion node N_{DF}, such that the maximum detection sensitivity of the system was completely retained (refer to Table 2). Secondly, due to its independent training scheme, M_2 remains highly modular, i.e. it can be easily tuned, upgraded or swapped out entirely upon future advancements, without retraining or affecting the stand-alone performance of M_1.

We conclude that encoding an anatomical prior (P) into the CNN architecture can guide model generalization with domain-specific clinical knowledge. Results indicated that P played the most important role in the generalization of the 3D CAD system (via M_1) and in retaining its performance across the multi-institutional testing datasets (refer to Section 3.2.2). Remarkably, its contribution was substantially more than any other architectural enhancement proposed in recent literature, while introducing negligible changes in the number of trainable parameters (refer to Table 2). However, it is worth noting that similar experiments with classifier M_2, yielded no statistical improvements. Parallel to the methods proposed by Cheng et al. (2018) and Tang et al. (2019), M_2 was designed to learn a different set of feature representations for csPCa than M_1, using its smaller receptive field size, patch-wise approach and decoupled optimization strategy. Thus, while M_1 was trained to learn translation covariant features for localization, M_2 was trained to learn translation invariant features for classification, i.e. patch-wise

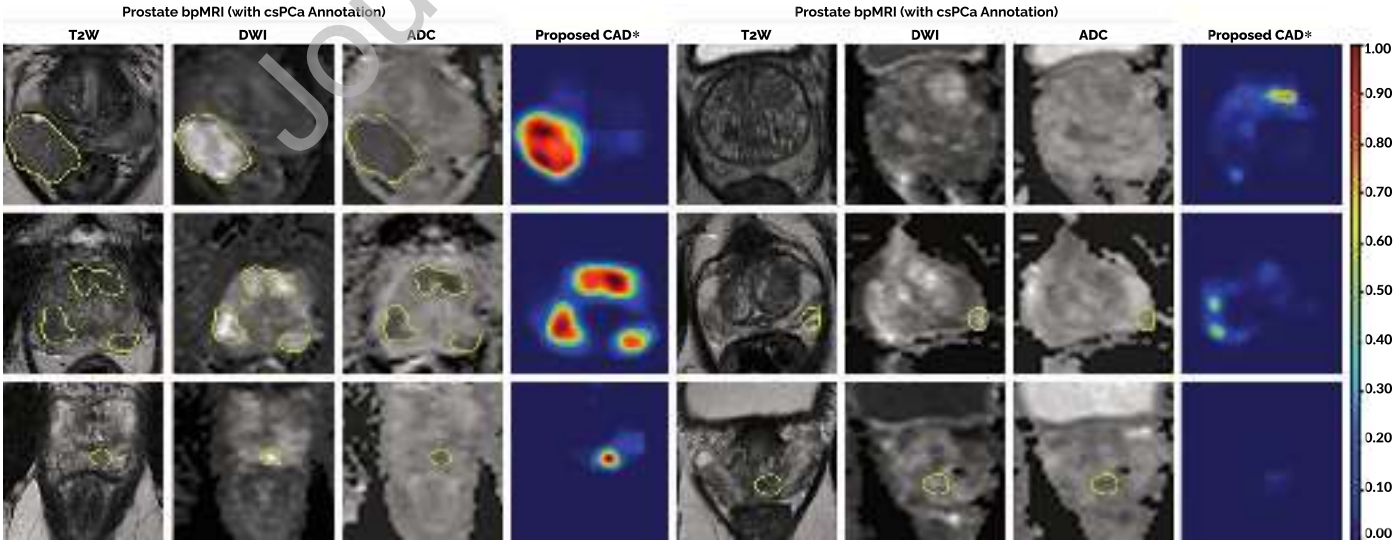


Fig. 8. Six patient cases from the external testing set TS2 and their corresponding csPCa detection maps, as predicted by the proposed CAD∗ system. Yellow contours indicate csPCa$^{(GS)}$ lesions, if present. While CAD∗ was able to successfully localize large, multifocal and apical/basal instances of csPCa$^{(GS)}$ *(left)*, in the presence of severe inflammation/fibrosis induced by other non-malignant conditions (eg. BPH, prostatitis), CAD∗ misidentified smaller lesions, resulting in false positive/negative predictions *(right)*.

prediction of the presence/absence of csPCa, irregardless of its spatial context in the prostate gland. We presume this key difference to be the primary reason why M_2 was effective at independent false positive reduction, yet unable to leverage the spatial priori embedded in P. Nonetheless, our study confirmed that powerful anatomical priors, such as P, can substitute additional training data for deep learning-based CAD systems and improve model generalization, by relaying the inductive biases of csPCa in bpMRI (Goyal and Bengio, 2020).

We hypothesized that modern CNN models, such as the CAD system presented in this study, should be able to learn and generalize beyond their training annotations of csPCa$^{(PR)}$ to reliably detect csPCa$^{(GS)}$ in an independent dataset (Cheplygina et al., 2019). To test this hypothesis, we acquired a large cohort of 1950 patient scans and produced voxel-level delineations of all PI-RADS>3 lesions. In practice, as every prostate bpMRI scan shares a corresponding radiology report compliant with PI-RADS, our training cohort was able to effectively capture a broad range of highly diverse studies –representing the complete distribution of patients encountered in the clinical workflow, and not only those who underwent biopsies. We benchmarked the CAD system against a consensus of radiologists on the ZGT cohort –an external testing set (TS2) graded by independent pathologists. Radiologists operated with high sensitivity (refer to Section 3.3), detecting 265 PI-RADS $\geq$ 3 lesion(s) among 152/296 patients (as noted in the *Supplementary Materials*), who subsequently underwent both TRUS and MRI-guided biopsies. When jointly performed, these techniques have been reported to reach upto 97% sensitivity for csPCa$^{(GS)}$ detection, relative to radical prostatectomy specimens (Radtke et al., 2016). While cohorts of radical prostatectomy patients share potentially stronger ground-truth, they typically carry much fewer samples (Sumathipala et al., 2018; Cao et al., 2019a; Bhattacharya et al., 2020; Seetharaman et al., 2021) and mostly, if not only, cases with severe malignancies –thereby deviating from the general distribution of patients encountered during clinical routine. Hence, an independent analysis of prostatectomy patients, exclusively, was not performed in this study (in agreement with Schelb et al., 2019). For the ZGT cohort, we observed that CAD$*$ demonstrated higher agreement with pathologists (81.08%; $kappa = 0.56 \pm 0.06$) than it did with radiologists (76.69%; $kappa = 0.51 \pm 0.04$). Moreover, CAD$*$ detected csPCa$^{(GS)}$ lesions in three patient cases, which were missed by four experienced radiologists operating at high sensitivity. Notably, this verified that CNNs can train effectively on csPCa$^{(PR)}$ annotations and achieve competitive test performance for csPCa$^{(GS)}$ detection, in comparison to state-of-the-art solutions that exclusively utilize biopsy-confirmed training annotations (Sumathipala et al., 2018; Cao et al., 2019a; Schelb et al., 2019; Yoo et al., 2019; Sanyal et al., 2020; Bhattacharya et al., 2020; Aldoj et al., 2020; Seetharaman et al., 2021) (refer to Section 1.1). Although, CNNs remain inadequate as stand-alone solutions (refer to Fig. 5, 8), the moderate agreement of CAD$*$ with both clinical experts, while inferring predictions relatively dissimilar to radiologists, highlights its potential to improve diagnostic certainty as a viable second reader in a screening setting (Sanford et al., 2020; Schelb et al., 2020).

The study is limited in a few aspects. Within the scope of this research, all prostate scans were acquired using MRI scanners developed by the same vendor. Thus, generalizing our proposed solution to a vendor-neutral model requires special measures, such as domain adaptation (Chiou et al., 2020), to account for heterogeneous acquisition conditions. Inclusion of PI-RADS>3 lesions for constructing the anatomical prior may introduce bias at train-time, particularly if the corresponding radiologist(s) lack(s) substantial expertise. Radiologists utilize additional clinical variables (e.g. prior exams, DCE scans, PSA density levels, etc.) to inform their diagnosis for each patient case –limiting the equity of any direct comparisons against the 3D CNNs developed in this study.

In summary, an automated novel end-to-end 3D CAD system, harmonizing several state-of-the-art methods from recent literature, was developed to diagnose and localize csPCa in bpMRI. To the best of our knowledge, this was the first demonstration of a deep learning-based 3D detection and diagnosis system for csPCa trained using radiologically-estimated annotations only and evaluated on large, multi-institutional testing datasets. The promising results of this research motivate the ongoing development of new techniques, particularly those which factor in the breadth of clinical knowledge established in the field beyond limited training datasets, to create comprehensive CAD solutions for the clinical workflow of prostate cancer management.

Acknowledgements

The authors would like to acknowledge Maarten de Rooij and Ilse Slootweg from Radboud University Medical Center for the annotation of fully delineated masks of prostate cancer for every bpMRI scan used in this study. This research is supported in parts by the European Union H2020: ProCAncer-I project (EU grant 952159) and Siemens Healthineers (CID: C00225450). Anindo Saha is also supported by an European Union: EACEA grant in the Erasmus+: Medical Imaging and Applications (MaIA) program.

References

Aldoj, N., Lukas, S., Dewey, M., Penzkofer, T., 2020. Semi-Automatic Classification of Prostate Cancer on Multi-parametric MR Imaging using a Multi-Channel 3D Convolutional Neural Network. European Radiology 30, 1243–1253. doi:10.1007/s00330-019-06417-z.

Alkadi, R., El-Baz, A., Taher, F., Werghi, N., 2019. A 2.5D Deep Learning-Based Approach for Prostate Cancer Detection on T2-Weighted Magnetic Resonance Imaging, in: Computer Vision – ECCV 2018 Workshops, Springer International Publishing. pp. 734–739.

Artan, Y., Yetik, I.S., 2012. Prostate Cancer Localization Using Multiparametric MRI based on Semisupervised Techniques With Automated Seed Initialization. IEEE Transactions on Information Technology in Biomedicine 16, 1313–1323. doi:10.1109/TITB.2012.2201731.

Basillote, J.B., Armenakas, N.A., Hochberg, D.A., Fracchia, J.A., 2003. Influence of Prostate Volume in the Detection of Prostate Cancer. Urology 61, 167–171. doi:10.1016/S0090-4295(02)02103-9.

Bass, E., Pantovic, A., Connor, M., Gabe, R., , Ahmed, H., 2020. A Systematic Review and Meta-Analysis of the Diagnostic Accuracy of Biparametric Prostate MRI for Prostate Cancer in Men at Risk. Prostate Cancer and Prostatic Diseases , 1–16.

Bhattacharya, I., Seetharaman, A., Shao, W., Sood, R., Rusu, M., 2020. CorrSigNet: Learning CORRelated Prostate Cancer SIGnatures from Radiology and Pathology Images for Improved Computer Aided Diagnosis, in: Medical Image Computing and Computer Assisted Intervention (MICCAI 2020), Springer International Publishing. pp. 315–325. doi:10.1007/978-3-030-59713-9_31.

Brancato, V., Di Costanzo, G., Basso, L., Tramontano, L., Puglia, M., Ragozzino, A., Cavaliere, A.C., 2020. Assessment of DCE Utility for PCa Diagnosis Using PI-RADS v2.1: Effects on Diagnostic Accuracy and Reproducibility. Diagnostics (Basel) 10.

Cao, R., Mohammadian Bajgiran, A., Afshari Mirak, S., Shakeri, S., Zhong, X., Enzmann, D., Raman, S., Sung, K., 2019a. Joint Prostate Cancer Detection and Gleason Score Prediction in mp-MRI via FocalNet. IEEE Transactions on Medical Imaging 38, 2496–2506.

Cao, R., Zhong, X., Scalzo, F., Raman, S., Sung, K., 2019b. Prostate Cancer Inference via Weakly-Supervised Learning using a Large Collection of Negative MRI, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), pp. 434–439.

Carass, A., Roy, S., Gherman, A., Reinhold, J.C., Shinohara, R.T., Oguz, I., 2020. Evaluating White Matter Lesion Segmentations with Refined Sørensen-Dice Analysis. Scientific Reports 10, 8242. doi:10.1038/s41598-020-64803-w.

Castillo, J.M., Starmans, M.P.A., Arif, M., Niessen, W.J., Klein, S., Bangma, C.H., Schoots, I.G., Veenland, J.F., 2021. A Multi-Center, Multi-Vendor Study to Evaluate the Generalizability of aRadiomics Model for Classifying Prostate cancer: High Grade vs. Low Grade. Diagnostics (Basel) 11.

Cawley, G.C., Talbot, N.L., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research 11, 20792107.

Chen, M.E., Johnston, D.A., Tang, K., Babaian, R.J., Troncoso, P., 2000. Detailed mapping of prostate carcinoma foci: biopsy strategy implications. Cancer 89, 1800–1809.

Cheng, B., Wei, Y., Shi, H., Feris, R., Xiong, J., Huang, T., 2018. Revisiting RCNN: On Awakening the Classification Power of Faster RCNN, in: Proceedings of the European Conference on Computer Vision (ECCV).

Cheplygina, V., de Bruijne, M., Pluim, J.P., 2019. Not-So-Supervised: A Survey of Semi-supervised, Multi-Instance, and Transfer Learning in Medical Image Analysis. Medical Image Analysis 54, 280–296. doi:10.1016/j.media.2019.03.009.

Chihara, L.M., Hesterberg, T.C., Dobrow, R.P., 2014. Mathematical Statistics with Resampling and R & Probability: With Applicati. John Wiley & Sons. OCLC: 941516595.

Chiou, E., Giganti, F., Punwani, S., Kokkinos, I., Joskowicz, L., 2020. Harnessing Uncertainty in Domain Adaptation for MRI Prostate Lesion Segmentation, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Springer International Publishing. pp. 510–520.

Chung, A.G., Khalvati, F., Shafiee, M.J., Haider, M.A., Wong, A., 2015. Prostate Cancer Detection via a Quantitative Radiomics-Driven Conditional Random Field Framework. IEEE Access 3, 2531–2541. doi:10.1109/ACCESS.2015.2502220.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, Springer International Publishing. pp. 424–432.

Dalca, A.V., Guttag, J., Sabuncu, M.R., 2018. Anatomical Priors in Convolutional Networks for Unsupervised Biomedical Segmentation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9290–9299.

Donato Jr, F., Costa, D.N., Yuan, Q., Rofsky, N.M., Lenkinski, R.E., Pedrosa, I., 2014. Geometric Distortion in Diffusion-Weighted MR Imaging of the Prostate Contributing Factors and Strategies for Improvement. Academic Radiology 21, 817–823.

Duran, A., Jodoin, P.M., Lartizien, C., 2020. Prostate Cancer Semantic Segmentation by Gleason Score Group in Bi-parametric MRI with Self Attention Model on the Peripheral Zone, in: International Conference on Medical Imaging with Deep Learning (MIDL) – Full Paper Track, Montreal, QC, Canada. pp. 193–204.

Egevad, L., Ahmad, A.S., Algaba, F., Berney, D.M., Varma, M., Camparo, P., 2013. Standardization of Gleason grading among 337 European pathologists. Histopathology 62, 247–256. doi:10.1111/his.12008.

Elwenspoek, M.M.C., Sheppard, A.L., McInnes, M.D.F., Whiting, P., 2019. Comparison of Multiparametric Magnetic Resonance Imaging and Targeted Biopsy With Systematic Biopsy Alone for the Diagnosis of Prostate Cancer: A Systematic Review and Meta-analysis. JAMA Network Open 2, e198427–e198427. doi:10.1001/jamanetworkopen.2019.8427.

Engels, R.R., Isral, B., Padhani, A.R., Barentsz, J.O., 2020. Multiparametric Magnetic Resonance Imaging for the Detection of Clinically Significant Prostate Cancer: What Urologists Need to Know. Part 1: Acquisition. European Urology 77, 457 – 468. doi:10.1016/j.eururo.2019.09.021.

Epstein, J.I., Egevad, L., Amin, M.B., Delahunt, B., 2016. The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System. Am. J. Surg. Pathol. 40, 244–252.

Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., 2017. Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. Nature 542, 115–118. doi:10.1038/nature21056.

Faryna, K., van der Laak, J., Litjens, G., 2021. Tailoring Automated Data Augmentation to H&E-Stained Histopathology, in: International Conference on Medical Imaging with Deep Learning (MIDL) – Full Paper Track, Lbeck, Germany. URL: https://openreview.net/pdf?id=JrBfXaoxbA2.

Fu, J., Liu, J., Tian, H., Lu, H., 2019. Dual Attention Network for Scene Segmentation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3141–3149.

Garcia-Reyes, K., Passoni, N.M., Palmeri, M.L., Kauffman, C.R., 2015. Detection of Prostate Cancer with Multiparametric MRI (mpMRI): Effect of Dedicated Reader Education on Accuracy and Confidence of Index and Anterior Cancer Diagnosis. Abdominal Imaging 40, 134–142. doi:10.1007/s00261-014-0197-7.

Gibson, E., Giganti, F., Hu, Y., Bonmati, E., Bandula, S., Gurusamy, K., Davidson, B., Pereira, S.P., Clarkson, M.J., Barratt, D.C., 2018. Automatic Multi-Organ Segmentation on Abdominal CT With Dense V-Networks. IEEE Transactions on Medical Imaging 37, 1822–1834. doi:10.1109/TMI.2018.2806309.

Goyal, A., Bengio, Y., 2020. Inductive Biases for Deep Learning of Higher-Level Cognition. arXiv:2011.15091.

Gros, C., Lemay, A., Cohen-Adad, J., 2021. SoftSeg: Advantages of Soft versus Binary Training for Image Segmentation. Medical Image Analysis 71, 102038. doi:10.1016/j.media.2021.102038.

Hanley, J.A., McNeil, B.J., 1982. The Meaning and Use of The Area Under A Receiver Operating Characteristic (ROC) Curve. Radiology 143, 29–36. doi:10.1148/radiology.143.1.7063747. pMID: 7063747.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity Mappings in Deep Residual Networks, in: Leibe, B., Matas, J., Sebe, N., Welling, M. (Eds.), Computer Vision – ECCV 2016, Springer International Publishing. pp. 630–645.

Hosseinzadeh, M., Brand, P., Huisman, H., 2019. Effect of Adding Probabilistic Zonal Prior in Deep Learning-based Prostate Cancer Detection, in: International Conference on Medical Imaging with Deep Learning (MIDL) – Extended Abstract Track, London, United Kingdom. pp. 1026–1034.

Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2019. Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence , 7132–7141.

Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2020. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods doi:10.1038/s41592-020-01008-z.

Isral, B., van der Leest, M., Sedelaar, M., Padhani, A.R., Zmecnik, P., Barentsz, J.O., 2020. Multiparametric Magnetic Resonance Imaging for the Detection of Clinically Significant Prostate Cancer: What Urologists Need to Know. Part 2: Interpretation. European Urology 77, 469 – 480. doi:10.1016/j.eururo.2019.10.024.

Jiang, Z., Ding, C., Liu, M., Tao, D., 2020. Two-Stage Cascaded U-Net: 1st Place Solution to BraTS Challenge 2019 Segmentation Task, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing. pp. 231–241.

Johnson, L.M., Turkbey, B., Figg, W.D., Choyke, P.L., 2014. Multiparametric MRI in Prostate Cancer Management. Nature Reviews Clinical Oncology 11, 346–353. doi:10.1038/nrclinonc.2014.69.

Kasivisvanathan, V., Rannikko, A.S., Borghi, M., Panebianco, V., 2018. MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis. New England Journal of Medicine 378, 1767–1777. doi:10.1056/NEJMoa1801993.

Kohl, S., Bonekamp, D., Schlemmer, H.P., Yaqubi, K., Radtke, J.P., Maier-Hein, K., 2017. Adversarial Networks for the Detection of Aggressive Prostate Cancer. arXiv:1702.08014.

Lematre, G., Mart, R., Rastgoo, M., Mriaudeau, F., 2017. Computer-Aided Detection for Prostate Cancer Detection based on Multi-parametric Magnetic Resonance Imaging, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 3138–3141.

Lin, T., Goyal, P., Girshick, R., He, K., Dollr, P., 2017. Focal Loss for Dense Object Detection, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999–3007.

Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., Huisman, H., 2014. Computer-aided detection of prostate cancer in mri. IEEE Transactions on Medical Imaging 33, 1083–1092.

McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., 2020. International Evaluation of an AI System for Breast Cancer Screening. Nature 577, 89–94. doi:10.1038/s41586-019-1799-6.

Miller, K.D., Nogueira, L., Mariotto, A.B., Rowland, J.H., Yabroff, K.R., Alfano, C.M., Jemal, A., Kramer, J.L., Siegel, R.L., 2019. Cancer Treatment and Survivorship Statistics, 2019. CA: A Cancer Journal for Clinicians 69, 363–385. doi:10.3322/caac.21565.

Min, S., Chen, X., Zha, Z.J., Wu, F., Zhang, Y., 2019. A Two-Stream Mutual Attention Network for Semi-supervised Biomedical Segmentation with Noisy Labels, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 4578–4585.

Ozkan, T.A., Eruyar, A.T., Cebeci, O.O., Memik, O., Ozcan, L., Kuskonmaz, I., 2016. Interobserver variability in Gleason histological grading of prostate cancer. Scand J Urol 50, 420–424.

Radtke, J.P., Schwab, C., Wolf, M.B., Freitag, M.T., Hohenfellner, M., Hadaschik, B.A., 2016. Multiparametric Magnetic Resonance Imaging (MRI) and MRITransrectal Ultrasound Fusion Biopsy for Index Tumor Detection: Correlation with Radical Prostatectomy Specimen. European Urology 70, 846–853.

Riepe, T., Hosseinzadeh, M., Brand, P., Huisman, H., 2020. Anisotropic Deep Learning Multi-planar Automatic Prostate Segmentation, in: Proceedings of the 28th International Society for Magnetic Resonance in Medicine Annual Meeting. URL: http://indexsmart.mirasmart.com/ISMRM2020/PDFfiles/3518.html.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing. pp. 234–241.

Rosenkrantz, A.B., Ginocchio, L.A., Cornfeld, D., Froemming, A.T., 2016. Interobserver Reproducibility of the PI-RADS Version 2 Lexicon: A Multicenter Study of Six Experienced Prostate Radiologists. Radiology 280, 793–804. doi:10.1148/radiol.2016152542.

Rouvire, O., Puech, P., Renard-Penna, R., Claudon, M., 2019. Use of Prostate Systematic and Targeted Biopsy on the Basis of Multiparametric MRI in Biopsy-Naive Patients (MRI-FIRST): A Prospective, Multicentre, Paired Diagnostic Study. The Lancet Oncology 20, 100 – 109. doi:10.1016/S1470-2045(18)30569-2.

Rundo, L., Han, C., Nagano, Y., Zhang, J., Hataya, R., Militello, C., Tangherloni, A., Nobile, M., Ferretti, C., Besozzi, D., Gilardi, M., Vitabile, S., Mauri, G., Nakayama, H., Cazzaniga, P., 2019. USE-Net: Incorporating Squeeze-and-Excitation Blocks into U-Net for Prostate Zonal Segmentation of Multi-Institutional MRI Datasets. Neurocomputing 365, 31 – 43.

Saha, A., Hosseinzadeh, M., Huisman, H., 2020. Encoding Clinical Priori in 3D Convolutional Neural Networks for Prostate Cancer Detection in bpMRI, in: Medical Imaging Meets NeurIPS Workshop–34th Conference on Neural Information Processing Systems (NeurIPS 2020). URL: https://arxiv.org/abs/2011.00263.

Sanford, T., Harmon, S.A., Turkbey, E.B., Turkbey, B., 2020. Deep-Learning-Based Artificial Intelligence for PI-RADS Classification to Assist Multiparametric Prostate MRI Interpretation: A Development Study. Journal of Magnetic Resonance Imaging n/a. doi:10.1002/jmri.27204.

Sanyal, J., Banerjee, I., Hahn, L., Rubin, D., 2020. An Automated Two-Step Pipeline for Aggressive Prostate Lesion Detection from Multi-Parametric MR Sequence. AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science 2020, 552–560.

Schelb, P., Kohl, S., Radtke, J.P., Bonekamp, D., 2019. Classification of Cancer at Prostate MRI: Deep Learning versus Clinical PI-RADS Assessment. Radiology 293, 607–617. doi:10.1148/radiol.2019190938.

Schelb, P., Wang, X., Radtke, J.P., Bonekamp, D., 2020. Simulated Clinical Deployment of Fully Automatic Deep Learning for Clinical Prostate MRI Assessment. European Radiology .

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images. Medical Image Analysis 53, 197 – 207. doi:10.1016/j.media.2019.01.012.

Seetharaman, A., Bhattacharya, I., Chen, L.C., Kunder, C.A., Rusu, M., 2021. Automated Detection of Aggressive and Indolent Prostate Cancer on Magnetic Resonance Imaging. Medical Physics doi:10.1002/mp.14855.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626.

Smith, C.P., Harmon, S.A., Barrett, T., Bittencourt, L.K., 2019. Intra- and Interreader Reproducibility of PI-RADSv2: A Multireader Study. Journal of Magnetic Resonance Imaging 49, 1694–1703. doi:10.1002/jmri.26555.

Srivastava, S., Koay, E.J., Borowsky, A.D., De Marzo, A.M., Ghosh, S., Wagner, P.D., Kramer, B.S., 2019. Cancer Overdiagnosis: A Biological Challenge and Clinical Dilemma. Nature Reviews Cancer 19, 349–358. doi:10.1038/s41568-019-0142-8.

Sumathipala, Y., Lay, N., Turkbey, B., Smith, C., Choyke, P.L., Summers, R.M., 2018. Prostate Cancer Detection from Multi-Institution Multiparametric MRIs using Deep Convolutional Neural Networks. Journal of Medical Imaging (Bellingham, Wash.) 5, 044507–044507. doi:10.1117/1.JMI.5.4.044507. pMC6294844[pmcid].

Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning, in: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI Press. p. 42784284.

Tang, H., Zhang, C., Xie, X., 2019. NoduleNet: Decoupled False Positive Reduction for Pulmonary Nodule Detection and Segmentation, in: Medical Image Computing and Computer Assisted Intervention (MICCAI 2019), pp. 266–274.

Turkbey, B., Rosenkrantz, A.B., Haider, M.A., Padhani, A.R., Margolis, D.J., 2019. Prostate Imaging Reporting and Data System version 2.1: 2019 Update of Prostate Imaging Reporting and Data System version 2. European Urology .

van der Leest, M., Cornel, E., Isral, B., Hendriks, R., 2019. Head-to-head Comparison of Transrectal Ultrasound-guided Prostate Biopsy Versus Multiparametric Prostate Resonance Imaging with Subsequent Magnetic Resonance-guided Biopsy in Biopsy-nave Men with Elevated Prostate-specific Antigen: A Large Prospective Multicenter Clinical Study. European Urology 75, 570 – 578. doi:doi.org/10.1016/j.eururo.2018.11.023.

d. Vente, C., Vos, P., Hosseinzadeh, M., Pluim, J., Veta, M., 2021. Deep Learning Regression for Prostate Cancer Detection and Grading in Bi-Parametric MRI. IEEE Transactions on Biomedical Engineering 68, 374–383. doi:10.1109/TBME.2020.2993528.

Verma, S., Choyke, P.L., Eberhardt, S.C., Oto, A., Tempany, C.M., Turkbey, B., Rosenkrantz, A.B., 2017. The Current State of MR Imagingtargeted Biopsy Techniques for Detection of Prostate Cancer. Radiology 285, 343–356. doi:10.1148/radiol.2017161684.

Wachinger, C., Reuter, M., Klein, T., 2018. DeepNAT: Deep Convolutional Neural Network for Segmenting Neuroanatomy. NeuroImage 170, 434 – 445. doi:10.1016/j.neuroimage.2017.02.035. segmenting the Brain.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang,

X., 2017. Residual Attention Network for Image Classification, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6450–6458.

Weinreb, J.C., Barentsz, J.O., Choyke, P.L., Cornud, F., 2016. PI-RADS Prostate Imaging Reporting and Data System: 2015, Version 2. European Urology 69, 16 – 40. doi:10.1016/j.eururo.2015.08.052.

Westphalen, A.C., McCulloch, C.E., Anaokar, J.M., Arora, S., Rosenkrantz, A.B., 2020. Variability of the Positive Predictive Value of PI-RADS for Prostate MRI across 26 Centers: Experience of the Society of Abdominal Radiology Prostate Cancer Disease-focused Panel. Radiology 296, 76–84. doi:10.1148/radiol.2020190646. pMID: 32315265.

Xiao, C., Deng, R., Li, B., Yu, F., Liu, M., Song, D., 2018. Characterizing Adversarial Examples Based on Spatial Consistency Information for Semantic Segmentation, in: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018, Springer International Publishing. pp. 220–237.

Yoo, S., Gujrathi, I., Haider, M.A., Khalvati, F., 2019. Prostate Cancer Detection using Deep Convolutional Neural Networks. Scientific Reports 9, 19518. doi:10.1038/s41598-019-55972-4.

Yu, X., Lou, B., Shi, B., Szolar, D., 2020a. False Positive Reduction Using Multiscale Contextual Features for Prostate Cancer Detection in Multi-Parametric MRI Scans, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), pp. 1355–1359.

Yu, X., Lou, B., Zhang, D., Winkel, D., Joskowicz, L., 2020b. Deep Attentive Panoptic Model for Prostate Cancer Detection Using Biparametric MRI Scans, in: Medical Image Computing and Computer Assisted Intervention (MICCAI 2020), Springer International Publishing. pp. 594–604.

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2020. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. IEEE Transactions on Medical Imaging 39, 1856–1867.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT Authorship Contribution Statement

Anindo Saha:
Conceptualization; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing - original draft.

Matin Hosseinzadeh:
Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Writing - review & editing.

Henkjan Huisman:
Conceptualization; Data curation; Funding acquisition; Project administration; Resources; Supervision; Writing - review & editing.