


## ORIGINAL RESEARCH

# Segmentation versus detection: Development and evaluation of deep learning models for prostate imaging reporting and data system lesions localisation on Bi-parametric prostate magnetic resonance imaging

Zhe Min<sup>1,2</sup>  | Fernando J. Bianco<sup>3</sup> | Qianye Yang<sup>2</sup> | Wen Yan<sup>2,4</sup> | Ziyi Shen<sup>2</sup> |  
David Cohen<sup>3</sup> | Rachael Rodell<sup>2</sup> | Dean C. Barratt<sup>2</sup> | Yipeng Hu<sup>2</sup>

<sup>1</sup>School of Control Science and Engineering, Shandong University, Jinan, China

<sup>2</sup>Centre for Medical Image Computing and Wellcome/EPSRC Centre for Interventional & Surgical Sciences, University College London, London, UK

<sup>3</sup>Urological Research Network, Miami Lakes, Florida, USA

<sup>4</sup>City University of Hong Kong, Hong Kong, China

**Correspondence**

Zhe Min.

Email: [minzhe@sdu.edu.cn](mailto:minzhe@sdu.edu.cn), [z.min@ucl.ac.uk](mailto:z.min@ucl.ac.uk)

**Funding information**

National Natural Science Foundation of China, Grant/Award Number: 62303275; International Alliance for Cancer Early Detection, Grant/Award Numbers: C28070/A30912, C73666/A31378; Wellcome / EPSRC Centre for Interventional and Surgical Sciences, Grant/Award Number: 203145Z/16/Z

**Abstract**

Automated prostate cancer detection in magnetic resonance imaging (MRI) scans is of significant importance for cancer patient management. Most existing computer-aided diagnosis systems adopt segmentation methods while object detection approaches recently show promising results. The authors have (1) carefully compared performances of most-developed segmentation and object detection methods in localising prostate imaging reporting and data system (PIRADS)-labelled prostate lesions on MRI scans; (2) proposed an additional customised set of lesion-level localisation sensitivity and precision; (3) proposed efficient ways to ensemble the segmentation and object detection methods for improved performances. The ground-truth (GT) perspective lesion-level sensitivity and prediction-perspective lesion-level precision are reported, to quantify the ratios of true positive voxels being detected by algorithms over the number of voxels in the GT labelled regions and predicted regions. The two networks are trained independently on 549 clinical patients data with PIRADS-V2 as GT labels, and tested on 161 internal and 100 external MRI scans. At the lesion level, nnDetection outperforms nnUNet for detecting both PIRADS  $\geq 3$  and PIRADS  $\geq 4$  lesions in majority cases. For example, at the average false positive prediction per patient being 3, nnDetection achieves a greater Intersection-of-Union (IoU)-based sensitivity than nnUNet for detecting PIRADS  $\geq 3$  lesions, being  $80.78\% \pm 1.50\%$  versus  $60.40\% \pm 1.64\%$  ( $p < 0.01$ ). At the voxel level, nnUNet is in general superior or comparable to nnDetection. The proposed ensemble methods achieve improved or comparable lesion-level accuracy, in all tested clinical scenarios. For example, at 3 false positives, the lesion-wise ensemble method achieves  $82.24\% \pm 1.43\%$  sensitivity versus  $80.78\% \pm 1.50\%$  (nnDetection) and  $60.40\% \pm 1.64\%$  (nnUNet) for detecting PIRADS  $\geq 3$  lesions. Consistent conclusions are also drawn from results on the external data set.

**KEYWORDS**

artificial intelligence, medical image processing, robotics

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Authors. *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

## 1 | INTRODUCTION

Prostate cancer is one of the leading causes for males' death worldwide, and is the third most common newly diagnosed cancers worldwide with 1.41 million incidences in 2020 [1]. Early prostate cancer detection plays an important role in cancer treatment [2, 3]. However, accurate prostate cancer diagnosis with both high sensitivity and specificity remains a challenging task [4]. For example, conventional diagnostic standard such as prostate-specific antigen (PSA) test together with transrectal ultrasound (TRUS) can potentially either overestimate non-clinically significant cancer or miss clinically significant prostate cancer [5].

The magnetic resonance imaging (MRI) scans have great potentials for assisting the screening and diagnostic tasks of prostate cancer in a non-invasive way [6–8], possibly compensating the limitations of PSA and TRUS [5]. Specifically, MR imaging could help avoid unnecessary biopsies [8, 9] and reduce chances of missing clinically significant prostate cancer [8, 9]. For example, the PROMIS study on 740 men has shown that (a) using mp-MRI to triage men will allow 27% of patients avoid unnecessary biopsy and 5% fewer clinically insignificant cancers; (b) if subsequent TRUS-biopsies are directed by mp-MRI findings, 18% more clinically significant prostate cancer cases will be detected [9]. The latest version of the European Association of Urology guidelines now recommend performing MRI scans prior to biopsies [10]. Nevertheless, following factors hinder the wide adoption of MRIs for prostate cancer diagnosis (1) reading prostate MRI scans requires substantial expertise and is time-consuming for radiologists; (2) findings from MRI scans are quite radiologist-dependent [11, 12].

Automated especially deep-learning-based prostate cancer detection approaches can alleviate the above-mentioned limitations of radiologists [13–15]. However, there are also several challenges for accurate automated prostate cancer segmentation or detection in MRI. First, multifocal prostate cancer lesions may appear with different shapes and sizes, making it rather difficult for one algorithm to exhibit high sensitivity and specificity [2, 4]. Second, clinically significant prostate cancer may resemble much the non-malignant benign regions in MRI, which may cause algorithms to generate inevitable false positives (FPs) [2]. Prostate imaging-reporting and data system: version2 (PI-RADS v2) is a widely-adopted reporting scheme for prostate MRI, aiming to standardise the early diagnostic process of prostate cancer in MRI [16], where each suspicious lesion is assigned to a score from 1 to 5 indicating its likelihood of being clinically significant [16, 17].

Computer-aided diagnosis (CAD) of prostate cancer in MRI can be either regarded as a segmentation or detection task, where the later one is under-utilised [2, 3, 18, 19]. This study aims to answer the following two research questions. (1) Considering that the lesion-level and voxel-level accuracies of prostate cancer localisation are important clinically relevant metrics to directly assess the usefulness of a CAD system, we would like to ask which method achieves better performances with those metrics at same levels of FPs per patient, segmentation or object detection? (2) Based on observations that

segmentation and object detection methods exhibit different sensitivities and specificities (as will be shown in the Results Section) which indicates potentials of achieved improved performances by combining the two, we'd like to ask what the performances of ensemble the segmentation and object detection algorithms at the lesion and voxel levels are?

To answer the above two questions, (1) two segmentation and object detection networks (i.e. nnUNet (<https://github.com/MIC-DKFZ/nnUNet>) and nnDetection (<https://github.com/MIC-DKFZ/nnDetection>) respectively) have been independently trained on a sizeable data set with bpMRI scans and corresponding radiologists-annotated prostate imaging reporting and data system (PI-RADS)  $\geq 3$  and PI-RADS  $\geq 4$  lesions' masks. (2) two effective ensemble methods for combining nnDetection and nnUNet have been proposed. (3) two limitations have been identified for the Intersection of Union (IoU) measure that is utilised to determine true positive (TP) lesion-level detection: (a) the IoU measure is defined in a symmetric way; (b) neither predicted nor ground-truth (GT) lesions are permitted to be counted more than once [20]. To address the above-mentioned limitations, two customised lesion-level metrics have been thus developed and utilised for evaluations.

### 1.1 | Contributions

This study demonstrates the first essential step towards systematically comparing and combining deep-learning-based segmentation versus detection methods for localising PI-RADS-labelled lesions on prostate MRI scans in a clinically relevant manner. The main contributions of this study are summarised as follows.

1. The performances of two typical self-configuring semantic segmentation and object detection networks (i.e. nnUNet [21] and nnDetection [22] respectively) have been comprehensively trained and evaluated on sizeable labelled clinical data, for localising PI-RADS-labelled prostate lesions in bp-MRI scans.
2. Two simple yet effective ensemble methods of nnUNet and nnDetection have been proposed and carefully validated, which have demonstrated superior *or* on-par performances against individual approaches.
3. Two customised lesion-level metrics, that is, GT-perspective sensitivity and the prediction-perspective precision, have been developed and utilised to evaluate segmentation, object detection and ensemble prostate cancer methods.

## 2 | RELATED WORK

The related work on automatic deep-learning-based prostate cancer detection methods can be classified into two categories: segmentation and object detection methods. The segmentation methods will output voxel-wise classifications of being cancerous or not over original medical images while detection

methods generally predict bounding boxes (BBs). Another difference is that the detection algorithm has an explicit concept of individual object (as a group of pixels or area of BBs), while segmentation treats the problem as a conditionally independent pixel classification problem.

## 2.1 | Segmentation networks

The segmentation methods have been used and demonstrate the potential of detecting prostate cancer in bp-MRI scans [5]. Most recent segmentation approaches resemble encoder-decoder Convolutional Neural Networks (CNNs) network structure such as 2D [23] and 3D UNet [24]. Other UNet variants include nested structures 2D UNet++ [25], 3D V-Net [26], 3D Attention UNet that utilises attention gates to suppress irrelevant regions [27–30].

**2D or 2.5D Networks** 2D and 2.5D segmentation networks can be identified according to whether neighbouring slices are utilised [27]. Cao et al. has presented a novel multi-class 2D CNN for detecting and classifying Gleason-Score-graded lesions with the focal loss in training [28]. Yu et al. has proposed a 2.5D UNet-based PIRADS lesion detection network with a FP reduction module [29]. **3D Networks** Saha et al. has proposed an attention-based 3D-CNN prostate cancer CAD system with a FP reduction module [5]. ProsAttention-Net is a multi-class 3D UNet-based network that jointly segments the prostate gland and Gleason Score (GS)-graded lesions [3, 18].

**nnUNet** In above-mentioned segmentation methods, the network structures need to be changed adaptively and requires additional hyper-parameter (e.g. learning rate) tuning when they are deployed on new datasets with different image sizes and resolutions, which is time-consuming and cumbersome. In contrast, nnUNet is a self-configuring UNet-based medical image segmentation method, which eliminates the need for extra manual hyper-parameter tuning [21, 31]. nnUNet has been demonstrated to surpass most existing approaches, including highly specialised solutions on 23 public datasets and 53 segmentation tasks used in international biomedical segmentation competitions [21]. nnUNet has also been validated in prostate gland segmentation [21, 32], peripheral zone (PZ) and transition zone (TZ) segmentation tasks [21, 33].

## 2.2 | Object detection

Object detection methods first determine whether there are instances of interest in an image, and if present estimate each instance's spatial location and extent usually defined with a BB [4, 34]. Object detection methods are generally classified into two categories: (1) the two-stage methods where proposals are utilised [35, 36]; (2) the one-stage region-proposal-free method that directly outputs classification probabilities and box coordinates [37, 38]. Object detection methods have recently been explored for detecting prostate cancer in MRI [2, 19].

**nnDetection** nnDetection [22] is a self-configuring one-stage Retina-UNet [38] based 3D object detection approach. nnDetection has set a new benchmark on the LUNG Nodule Analysis [39] and achieves competitive performances of Aneurysm Detection And segmentation Challenge [22]. nnDetection's effectiveness were also validated on another 11 data sets, including ProstateX, Kits19 etc. [22]. In ref. [22], nnDetection and nnUNet were compared with the metric mean average precision commonly used in the computer vision community, which is neither directly clinically relevant nor intuitive to radiologists or urologists.

## 3 | METHODS

nnUNet [21] is chosen as the segmentation baseline, where GT labels are the radiologists-annotated lesions' masks and the network's raw outputs are the voxel-wise softmax probabilities. nnDetection [22] is chosen as the detection baseline, where GT labels are the lesions' instance segmentations while the network's raw outputs are the predicted BBs with probability/confidence scores. It is noted that the probability/confidence score is associated with an individual predicted BB.

### 3.1 | Lesion-wise ensemble

In lesion-wise ensemble, predicted BBs from nnUNet and nnDetection are used where BBs can be readily extracted from predicted individual lesions' masks in nnUNet. At the lesion-level, ensemble the predicted BBs from both networks is done with the weighted box clustering [22]. Suppose that there are  $L \in \mathbb{N}^+$  predicted BBs whose IoUs with the highest scoring and unclustered BB are larger than a threshold, the confidence score  $o_s \in (0, 1]$  of the ensemble BB at that location is computed as follows:

$$o_s = \frac{\sum_{i=1}^L s_i w_i}{\sum_{i=1}^L w_i + n_{\text{missing}} \frac{1}{L} \sum_{i=1}^L w_i}, \quad (1)$$

where  $w_i \in \mathbb{R}$  is the weight associated with the  $i$ th BB,  $s_i \in (0, 1]$  is the corresponding confidence score,  $n_{\text{missing}} = \max(0, n_{\text{prediction}} - L)$  where  $n_{\text{prediction}} = 2$  is the number of expected predictions per location since at a specific location one prediction from either independent model are expected. The weighted coordinates of the clustered BB are computed as follows

$$o_c = \frac{\sum_{i=1}^L c_i w_i}{\sum_{i=1}^L w_i},$$

where  $c_i \in \mathbb{R}$  denotes the  $i$ th BB's coordinates taking value from  $\{x_1, y_1, x_2, y_2, z_1, z_2\}$  recursively. In order to get the lesion-level sensitivities, the two networks' predicted BBs whose confidence scores are larger than a varying *lesion-level cutoff* value, are ensemble with the weighted box clustering technique and the results are kept in the final predictions. At the lesion-level, ensemble the nnDetection with *nnUNet* and

*nnUNet Argmax* is referred to as *Lesion-wise Ensemble* and *Lesion-wise Ensemble Argmax* respectively.

### 3.2 | Voxel-wise ensemble

To ensemble nnUNet and nnDetection voxel-wisely, the softmax probabilities from both networks are to be utilised, which requires us to generate the *pseudo* softmax probabilities from the nnDetection predicted BBs. In nnDetection, two different scenarios are considered at one specific voxel as follows. Let  $O \in \mathbb{N}^+$  denote the number of predicted instances by the object detection network. (a) There is  $O = 1$  predicted instance with probability  $p_{object} \in (0, 1]$  at that voxel, the probability of that voxel being foreground is thus equal to that of the predicted instance  $p_{voxel}^{detection} = p_{object}$ ; (b) there are  $O > 1$  predicted instances at that voxel, then

$$p_{voxel}^{detection} = \max/\text{mean} \left\{ p_{object}^i \right\}_{i=1}^O. \quad (2)$$

It is noted that the case (a) can be unified into Equation (2). The corresponding probability of that voxel being background is  $p_{voxel}^{background} = 1 - p_{voxel}^{foreground} = p_{voxel}^{detection}$ , and the softmax probability is  $\left( 1 - \max \left\{ p_{object}^i \right\}_{i=1}^O, \max \left\{ p_{object}^i \right\}_{i=1}^O \right)$ . We should determine the follows before voxel-wise ensemble of the two networks. (a) Should we utilise all or a subset of raw predicted BBs in nnDetection to generate the *pseudo* softmax, which are denoted as ‘*all*’ or ‘*not all*’ respectively. The subset is acquired by applying a *lesion-level cutoff* to the BBs’ scores to retain only above-threshold ones. (b) It can be either max or average operations in Equation (2), which are denoted as ‘*max*’ or ‘*average*’ respectively. (c) The foreground probability  $p_{voxel}^{ensemble} \in (0, 1]$  in the final ensemble softmax is acquired with the max or average operations as

$$p_{voxel}^{ensemble} = \max/\text{mean} \left\{ p_{voxel}^{detection}, p_{voxel}^{segmentation} \right\}, \quad (3)$$

whose corresponding scenarios are denoted as ‘*max*’ or ‘*average*’ respectively,  $p_{voxel}^{segmentation} \in (0, 1]$  denotes the nnUNet foreground probability at that voxel. In total, there are 8 different combinations of above choices, which were evaluated on the 50 split validation data and the best combination was identified.

At the voxel level, ensemble the nnDetection with *nnUNet* and *nnUNet Argmax* are referred to as *Voxel-wise Ensemble* and *Voxel-wise Ensemble Argmax*, respectively.

### 3.3 | Lesion-level evaluation metrics

For one patient scan, suppose there are  $M \in \mathbb{N}^+$  GT and  $N \in \mathbb{N}^+$  predicted lesions,  $i \in \mathbb{N}^+$  and  $j \in \mathbb{N}^+$  respectively denote the index of GT and predicted lesions.

*Intersection-of-Union (IoU)-based Lesion-Level Sensitivity and Precision.* The lesion-level IoU between the  $i$ th GT lesion and the  $j$ th predicted lesion is defined as follows:

$$\text{IoU}_{\text{lesion-level}}^{ij} = \frac{S_{\text{overlap}}^{ij}}{S_{\text{GT}}^i + S_{\text{Predicted}}^j - S_{\text{overlap}}^{ij}}, \quad (4)$$

where  $S_{\text{GT}}^i \in \mathbb{N}^+$  and  $S_{\text{Predicted}}^j \in \mathbb{N}^+$  are the numbers of foreground voxels within the  $i$ th GT and the  $j$ th predicted lesions respectively,  $S_{\text{overlap}}^{ij} \in \mathbb{N}^+$  counts the number of foreground voxels belonging to both the  $i$ th GT lesion and the  $j$ th predicted lesion.

The matching process between GT and predicted lesions will start with the predicted lesion that has the largest object score, and iterate over all predicted lesions. The  $j$ th predicted lesion is matched to the  $i$ th GT lesion if all follows hold (1) the  $i$ th GT lesion has the largest  $\text{IoU}_{\text{Lesion-level}}^{ij}$  among unmatched GT lesions; (2)  $\text{IoU}_{\text{Lesion-level}}^{ij}$  is larger than a pre-set threshold; (3) the  $i$ th GT lesion is unmatched. The number of matched GT or predicted lesions is counted as TP ( $\text{TP}^{\text{IoU}_{\text{Lesion-level}}}$ ), while the numbers of unmatched GT and predicted lesions are counted as false negative (FN $^{\text{IoU}_{\text{Lesion-level}}}$ ) and FPs (FP $^{\text{IoU}_{\text{Lesion-level}}}$ ) respectively. The IoU-based lesion-level sensitivity and precision are computed as  $\text{Sensitivity}_{\text{Lesion-level}}^{\text{IoU-based}} = \frac{\text{TP}^{\text{IoU}_{\text{Lesion-level}}}}{\text{TP}^{\text{IoU}_{\text{Lesion-level}}} + \text{FN}^{\text{IoU}_{\text{Lesion-level}}}}$  and  $\text{Precision}_{\text{Lesion-level}}^{\text{IoU-based}} = \frac{\text{TP}^{\text{IoU}_{\text{Lesion-level}}}}{\text{TP}^{\text{IoU}_{\text{Lesion-level}}} + \text{FP}^{\text{IoU}_{\text{Lesion-level}}}}$ .

*IoU-based Box-Level Sensitivity* The evaluations based on Equation (4) are biased towards the nnUNet whose outputs are masks, which motivates us to also utilise the box-level IoU is used to further determine TP/FP predictions, and false negative (FN) GT lesions. Formally, the box-level IoU between the  $i$ th GT lesion’s BB and the  $j$ th predicted lesion’s BB is defined as follows:

$$\text{IoU}_{\text{box-level}}^{ij} = \frac{V_{\text{overlap}}^{ij}}{V_{\text{GT}}^i + V_{\text{Predicted}}^j - V_{\text{overlap}}^{ij}} \quad (5)$$

where  $V_{\text{GT}}^i \in \mathbb{R}$  is the volume of the BB enclosing the  $i$ th GT lesion mask,  $V_{\text{Predicted}}^j \in \mathbb{R}$  is the volume of  $j$ th predicted BB in nnDetection or the smallest BB enclosing the  $j$ th predicted lesion mask in nnUNet,  $V_{\text{overlap}}^{ij} \in \mathbb{R}$  is the volume of intersected BB between the  $i$ th GT BB and  $j$ th predicted BB. The matching process is similar to those in computing  $\text{IoU}_{\text{lesion-level}}^{ij}$ . Afterwards, the IoU-based box-level sensitivity  $\text{Sensitivity}_{\text{Box-Level}}^{\text{IoU-based}}$  and precisions  $\text{Precision}_{\text{Box-Level}}^{\text{IoU-based}}$  are computed.

*Ground-truth-perspective Lesion-Level Sensitivity.* The overlap ratio of  $i$ th GT lesion, from the GT perspective is defined as follows:

$$\text{Overlap}_{\text{GT,Perspective}}^i = \frac{S_{\text{overlap}}^i}{S_{\text{GT}}^i} \quad (6)$$

where  $S_{\text{overlap}}^i = \sum_{j=1}^N S_{\text{overlap}}^{ij}$  is the number of voxels within both the  $i$ th GT lesion and any predicted lesion. The inherent rationale of the formula is to compute the percentage of 1 GT lesion being detected by all predictions.

One GT lesion is considered as a TP if  $\text{Overlap}_{\text{GT,Perspective}}^i$  in Equation (6) is larger than a pre-set threshold, a FN otherwise. FP prediction is not defined within this metric, and the definition of lesion-level FP using  $\text{IoU}_{\text{lesion-level}}^{ij}$  is used when we plot sensitivities with respect to FPs per patient. With the numbers being  $\text{TP}^{\text{GT}} \in \mathbb{N}^+$  and  $\text{FN}^{\text{GT}} \in \mathbb{N}^+$ , the GT-perspective lesion-level sensitivity is  $\text{Sensitivity}_{\text{Lesion-Level}}^{\text{GT-perspective}} = \frac{\text{TP}^{\text{GT}}}{\text{TP}^{\text{GT}} + \text{FN}^{\text{GT}}}$  while the corresponding precision is not defined within this metric.

**Prediction-Perspective Lesion-level Precision.** The overlap of  $j$ th prediction lesion with all GT lesions is defined as follows:

$$\text{Overlap}_{\text{Prediction-Perspective}}^j = \frac{S_{\text{overlap, prediction-perspective}}^j}{S_{\text{Prediction}}^j}, \quad (7)$$

where  $S_{\text{overlap, prediction-perspective}}^j = \sum_{i=1}^M S_{\text{overlap}}^{ij}$  represents the number of voxels within both the  $j$ th predicted lesion and any GT lesion. The inherent rationale behind this formula is to compute the percentage of one predicted lesion's voxels being any GT lesion. One prediction is a  $\text{TP}^{\text{Pred}}$  if  $\text{Overlap}_{\text{Prediction-Perspective}}^j$  is larger than a pre-set threshold, and a  $\text{FP}^{\text{Pred}}$  otherwise. True negative is not defined within this metric. With  $\text{TP}^{\text{Pred}} \in \mathbb{N}^+$  and  $\text{FP}^{\text{Pred}} \in \mathbb{N}^+$ , the prediction-perspective lesion-level precision is  $\text{Precision}_{\text{Lesion-Level}}^{\text{Prediction-perspective}} = \frac{\text{TP}^{\text{Pred}}}{\text{TP}^{\text{Pred}} + \text{FP}^{\text{Pred}}}$  while the corresponding sensitivity is not defined within this metric.

### 3.4 | Voxel-level evaluation metric

At the voxel-level, the Dice Similarity Coefficient (DSC) is utilised to evaluate the predictions from different models with the GT lesions' masks, and is computed as follows:

$$\text{DSC} = \frac{2 \times \text{TP}^{\text{voxel}}}{2 \times \text{TP}^{\text{voxel}} + \text{FP}^{\text{voxel}} + \text{FN}^{\text{voxel}}}, \quad (8)$$

where here  $\text{TP}^{\text{voxel}}$ ,  $\text{FP}^{\text{voxel}}$ , and  $\text{FN}^{\text{voxel}}$  are the numbers of corresponding voxels. One voxel is considered as TP if that voxel is within both the GT and predicted lesions at the same time. The reported overall DSC for the test data set is computed by averaging the DSCs with all patients.

## 4 | EXPERIMENTS

### 4.1 | Dataset

All cases were labelled with radiologists' annotations of suspicious cancerous regions via PIRADS-V2 [16]. This work

involved patient imaging data that were acquired from multiple studies, approved by the local research ethic committees, including SmartTarget Biopsy (14/LO/0830, 22-08-2014) [40], PROMIS (11/LO/0185, 21-07-2021) [40], INDEX (NCT01194648, 02-09-2010) [41], PICTURE (11/LO/1657, 06-11-2011) [42], and SmartTarget Therapy (14/LO/1375, 01-08-2015) [40]. In all, 760 patients were selected for the study with all three image types, which include T2-weighted (T2-W), diffusion-weighted imaging (DWI), and apparent diffusion coefficient (ADC), following the local bpMR imaging protocols. All scanners vendor of the utilised data set are SIEMENS. The mean and median lesion volume are 2.55 and 1.72  $\text{cm}^3$  (range, 0.18–27.02  $\text{cm}^3$ ). There are multiple radiologists that delineate the lesions. The GT segmentation is the consensus results of one or multiple radiologists. In the utilised data set, 23.57% lesions are in the PZ while 76.43% lesions are in the TZ. The cohort is further partitioned into the training, validation and test sets, each with 549 (72.24%), 50 (6.58%) and 161 (21.18%) cases respectively. In this study, the lesions whose PIRADS score is larger than 3 and those whose PIRADS score is larger than 4 are segmented, which are denoted as  $\text{PIRADS} \geq 3$  and  $\text{PIRADS} \geq 4$  lesions, respectively.

A detailed summary of the data cohort is given in Table 1. In the training cohort, the numbers of  $\text{PIRADS} \geq 3$  and  $\text{PIRADS} \geq 4$  lesions are 1478 (83.74%) and 287 (16.26%) and the numbers of cases whose index lesions being  $\text{PIRADS} \geq 3$  and  $\text{PIRADS} \geq 4$  are 337 (61.38%) and 212 (38.62%), respectively. In the validation cohort, the numbers of the numbers of  $\text{PIRADS} \geq 3$  and  $\text{PIRADS} \geq 4$  lesions are 98 (82.35%) and 21 (17.65%) and the numbers of cases whose index lesions being  $\text{PIRADS} \geq 3$  and  $\text{PIRADS} \geq 4$  are 33 (66%) and 1734%, respectively. In the test cohort, the numbers of the numbers of  $\text{PIRADS} \geq 3$  and  $\text{PIRADS} \geq 4$  lesions are 300(76.53%) and 92(23.47%) and the numbers of cases whose index lesions being  $\text{PIRADS} \geq 3$  and  $\text{PIRADS} \geq 4$  are 86 (53.42%) and 75 (46.58%), respectively. In the training, validation and test cohorts, the ratios of numbers of  $\text{PIRADS} \geq 3$  and  $\text{PIRADS} \geq 4$  lesions are 5.15 1, 4.67 1 and 3.26:1 respectively while the ratios of number of cases whose index lesions are  $\text{PIRADS} \geq 3$  and  $\text{PIRADS} \geq 4$  are 1.59:1, 1.94:1 and 1.15:1.

All three modalities are resampled to image size of  $128 \times 128 \times 32$  and spatial resolutions being  $0.625 \times 0.625 \times 3 \text{ mm}^3$  with higher resolution in the x-y plane. For each patient, the DWI and ADC images are spatially resampled once (using the rigid image-to-scanner transformations) to their corresponding T2 image. All modalities are also normalised to [0,1]. For both detection and segmentation networks, T2W, ADC and DWI are concatenated together into  $128 \times 128 \times 32 \times 3$  as the network inputs.

### 4.2 | nnUNet

An appropriate strategy is needed to aggregate lesions from raw voxel-wise softmax probabilities of nnUNet [22]. First, each voxel in the MRI volume is classified as foreground or

**TABLE 1** A detailed summary of the utilised internal data set.

Properties	Training cohort	Validation cohort	Test cohort
Number of Patients	549	50	161
No. of MRI-detected lesions per patient			
One lesion	69	12	33
Two lesions	126	18	57
Three lesions	144	13	45
Four lesions	104	6	22
Five lesions	61	0	2
Six lesions	29	0	2
Seven lesions	11	0	0
Eight lesions	5	1	0
Nine lesions	0	0	0
MRI index lesion per patient			
PI-RADS 3	337	33	86
PIRADS 45	212	17	75
MRI assessment per lesion			
Total	1765	119	392
PIRADS 3	1478	98	300
PIRADS 45	287	21	92

background, which can be achieved either (a) by applying a minimum threshold value (*voxel-level cutoff*) to foreground probability in the softmax. In the binary-class segmentation, the voxel is considered as foreground if the foreground softmax probability is greater or equal to the *voxel-level cutoff*; or (b) with the *argmax* operation, which can also be understood as applying certain *voxel-level cutoffs* to the softmax probability scores. In the binary-class segmentation, the *argmax* operation is equivalent to applying a *voxel-level cutoff* being 0.5 to the foreground softmax probability. Second, one predicted lesion's confidence score has to be determined from the aggregated voxels's foreground probabilities, which could be max (used in our study), mean, median, 95% percentile operations [22]. Third, the aggregated lesions in the final diagnostic prediction is determined by keeping those lesions whose confidence scores are not less than a minimum threshold (*lesion-level cutoff*).

Given the above introduced strategy, two nnUNet variants are utilised in this study. (a) The *argmax* operation is used to determine foreground voxels, which are then aggregated into individual lesions with the connected component analysis (CCA) technique [22]. After that, the lesions whose confidence scores are higher than varying *lesion-level cut-offs* are retained in the final diagnostic prediction. This variant of nnUNet is referred to as *nnUNet Argmax*. (b) The foreground voxels are determined by identifying voxels whose foreground probabilities are higher than varying *voxel-level cutoffs*, and foreground voxels are then aggregated into individual lesions with the CCA technique [22]. This variant of nnUNet is referred to as

*nnUNet*. The BB is determined as the smallest cuboid that encloses the predicted lesion mask in 3D. Two experiments were conducted with PIRADS  $\geq 3$  and PIRADS  $\geq 4$  lesions. In both cases, nnUNet was trained in a five fold cross validation manner for 1000 epochs with 250 minimatches per epoch [21].

### 4.3 | nnDetection

In nnDetection, lesions whose confidence scores are higher than varying *lesion-level cutoffs* are retained in the final predictions, in order to achieve different lesion-level sensitivities, precisions, and voxel-level DSCs. The predicted lesion mask is acquired from the BB noting that all voxels inside the BBs are considered as foreground.

nnDetection was trained independently with PIRADS  $\geq 3$  and PIRADS  $\geq 4$  lesions' labels, similar to the scenario of nnUNet, in a five-fold cross validation manner for 60 epochs with 2500 mini batches per epoch [22].

### 4.4 | Ensemble

*Lesion-wise Ensemble Hyperparameter Tuning.* The hyperparameters that need to be tuned include the weights with the nnunet predictions  $w_i$  in Equation (1) and the IoU threshold  $\text{IoU}_{ensemble}$  to determine whether the two boxes are to be clustered, which are chosen from

$$w_i \in [0.1, 0.3, 0.5, 0.7, 0.9], \quad (9)$$

and

$$\text{IoU}_{\text{ensemble}} \in [0.1, 0.3, 0.5, 0.7, 0.9]. \quad (10)$$

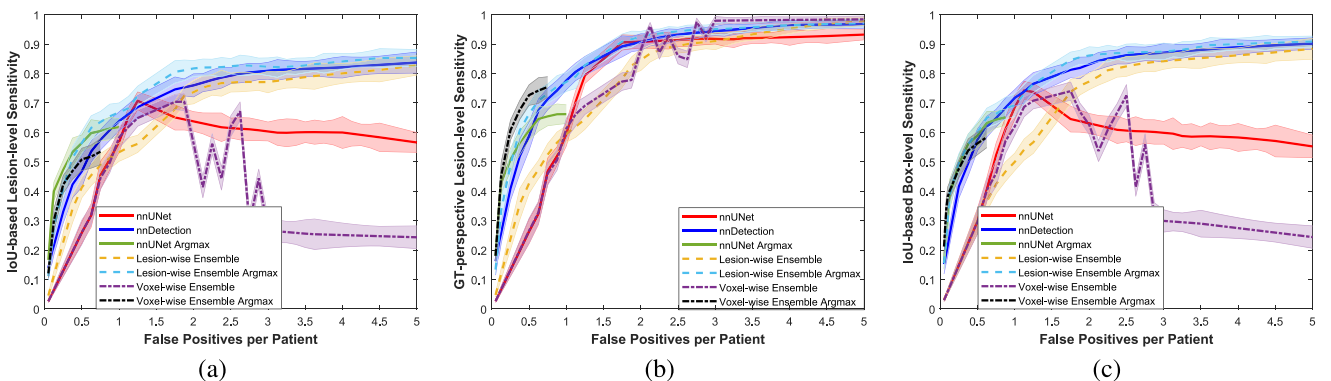
Totally, there are 25 different cases that contain different parameter combinations, which have been optimised on the split validation set.

**Voxel-Wise Ensemble Hyperparameter Tuning.** The softmax probabilities from both individual methods at the inference stage are ensemble with the approach in Voxel-Wise Ensemble of Methods Section. The combinations in Voxel-Wise Ensemble of Methods Section were chosen according to the results on the validation data set.

## 5 | RESULTS

In this section, we compare the results of nnDetection, nnUNet, nnUNet Argmax, *Lesion-wise Ensemble*, *Lesion-wise Ensemble Argmax*, *Voxel-wise Ensemble*, and *Voxel-wise Ensemble Argmax* at both the lesion and voxel levels, where nnUNet, nnDetection and nnUNet Argmax are considered as baselines. To compare methods, multi-way ANOVA tests have been conducted and statistical significance results of paired *t*-test are reported after multiple testing correction. More specifically, we have done the ANOVA test using 7 (i.e. nnUNet, nnDetection, nnUNet Argmax, *Lesion-wise Ensemble*, *Lesion-wise Ensemble Argmax*, *Voxel-wise Ensemble*, *Voxel-wise Ensemble Argmax*) or 5 methods (i.e. nnUNet, nnDetection, *Lesion-wise Ensemble*, *Lesion-wise Ensemble Argmax*, *Voxel-wise Ensemble*) with 3 overlap measures (i.e.  $\text{IoU}_{\text{lesion-level}}$ ,  $\text{Overlap}_{\text{GT, Perspective}}$  and  $\text{IoU}_{\text{box-level}}$ ) at the 4 different levels of FP (i.e. 0.5FP, 1FP, 2FPs, 3 FPs) for the two classification problems (i.e. PIRADS  $\geq 3$  or not and PIRADS  $\geq 4$  or not).

Figure 1 includes IoU-based lesion-level sensitivity (a), GT-perspective lesion-level sensitivity (b) and IoU-based box-level sensitivity (c) at different FPs per patient



**FIGURE 1** The performance comparison in detecting PIRADS  $\geq 3$  lesions with IoU-based lesion-level sensitivity (a), GT-perspective lesion-level sensitivity (b) and IoU-based box-level sensitivity (c). The transparent areas denote 95% confidence intervals. PIRADS, prostate imaging reporting and data system.

respectively, in detecting the PIRADS  $\geq 3$  lesions. Figure 2 includes similar results in detecting PIRADS  $\geq 4$  lesions, respectively.

Table 2 and Table 3 include the mean values of three different sensitivities and DSC in detecting PIRADS  $\geq 3$  and PIRADS  $\geq 4$  lesions, respectively. Table 4 and Table 5 include the mean values of two precisions and DSC in detecting PIRADS  $\geq 3$  and PIRADS  $\geq 4$  lesions, respectively. In all the tables, the above-mentioned evaluation metrics are presented at four specified FPs: 0.5,1,2,3.

### 5.1 | Comparison between nnDetection and nnUnet at the lesion-level

*nnDetection outperforms nnUnet, when considerable lesion-level sensitivity is required for specific clinical application such as MRI-targeted surgical biopsies, in detecting PIRADS  $\geq 3$  and PIRADS  $\geq 4$  lesions.*

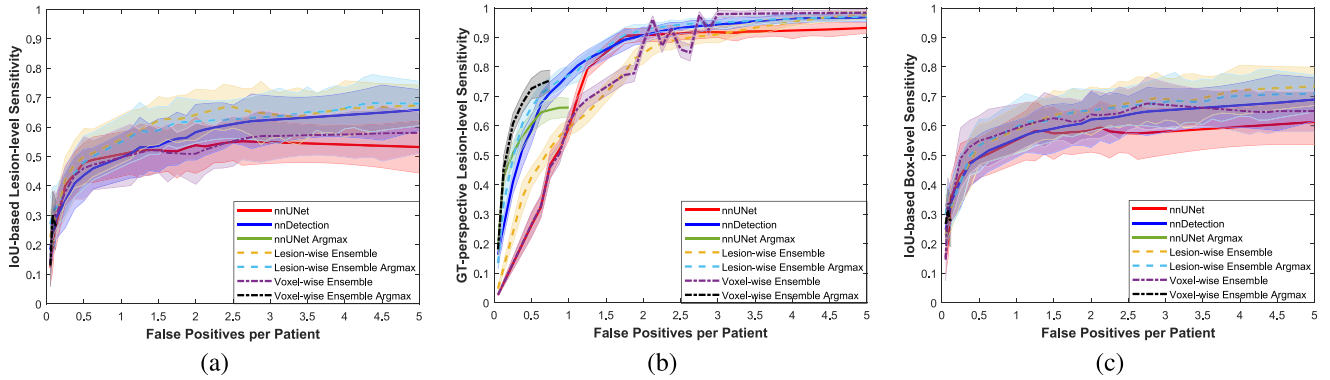
**PIRADS  $\geq 3$ . Sensitivities** As it is shown in Figure 1 and Table 2, in most cases, nnDetection achieves significantly larger sensitivities than nnUNet at four specific levels of FPs (11 out of 12 different cases). For example, the IoU-based lesion-level sensitivities at 3FPs are  $80.78\% \pm 1.50\%$  versus  $60.40\% \pm 1.64\%$  for nnDetection and nnUNet respectively.

**Precisions** nnDetection is better than nnUNet for five of the eight different FPs significantly (Table 4). For example, as shown in Table 4, the two precisions at 3FPs are  $39.70\% \pm 1.24\%$  versus  $33.00\%$  and nnUNet respectively.

**PIRADS  $\geq 4$ . Sensitivities**  $4\% \pm 1.95\%$ ,  $65.70\% \pm 1.83\%$  versus  $40.43\% \pm 2.27\%$  for nnDetection and **Precisions** As shown in Tables 3 and 5, in most cases, nnDetection has significantly higher sensitivities (10 out of 12 different cases).

### 5.2 | Comparison between nnDetection and nnUnet at the voxel-level

*nnUNet achieves higher DSC than nnDetection in detecting the PIRADS  $\geq 3$  lesions when more than 1 FP is allowed,*



**FIGURE 2** The performance comparison in detecting PIRADS  $\geq 4$  lesions with IoU-based lesion-level sensitivity (a), GT-perspective lesion-level sensitivity (b), IoU-based box-level sensitivity (c). The transparent areas are 95% confidence intervals. PIRADS, prostate imaging reporting and data system.

**TABLE 2** Mean sensitivities with corresponding average numbers of false positives (FPs) per patient, in detecting PIRADS  $\geq 3$  Lesions.

Model	Sensitivity <sup>IoU-based</sup> <sub>Lesion-Level</sub>				Sensitivity <sup>overlap-based</sup> <sub>GT-perspective</sub>				Sensitivity <sup>IoU-based</sup> <sub>Box-Level</sub>			
	0.5FP	1FP	2FPs	3FPs	0.5FP	1FP	2FPs	3FPs	0.5FP	1FP	2FPs	3FPs
nnUNet	0.26	0.57	0.64	0.60	0.27	0.60	0.91 <sup>o</sup>	0.92	0.30	0.68	0.63	0.60
nnDetection	0.47 <sup>o</sup>	0.64 <sup>o</sup>	0.76 <sup>o</sup>	0.81 <sup>o</sup>	0.59 <sup>o</sup>	<b>0.78<sup>o</sup></b>	0.90	0.94 <sup>o</sup>	0.57 <sup>o</sup>	<b>0.72<sup>o</sup></b>	0.83 <sup>o</sup>	0.87 <sup>o</sup>
nnUNet Argmax	<b>0.57</b>	0.61	N/A	N/A	0.60	0.66	N/A	N/A	0.59	N/A	N/A	N/A
Lesion-wise Ensemble	0.41	0.53	0.74	0.77	0.43	0.59	0.84	0.91	0.29	0.50	0.77	0.84
Lesion-wise Ensemble Argmax	0.56	<b>0.67*</b>	<b>0.82*</b>	<b>0.82*</b>	0.66*	0.77	<b>0.93*</b>	<b>0.95*</b>	<b>0.60*</b>	0.69	<b>0.86*</b>	<b>0.88*</b>
Voxel-wise Ensemble	0.26	0.58	0.56	0.26	0.27	0.61	0.88	<b>0.98*</b>	0.29	0.63	0.63	0.30
Voixe-wise Ensemble Argmax	0.51	N/A	N/A	N/A	<b>0.73*</b>	N/A	N/A	N/A	0.56	N/A	N/A	N/A

Note: \*: significantly better than all baselines that are nnUNet, nnDetection and nnUNet Argmax ( $p < 0.01$ ). <sup>o</sup>: the differences between nnUNet and nnDetection are significant ( $p < 0.01$ ). N/A: not applicable.

Abbreviation: PIRADS, prostate imaging reporting and data system.

**TABLE 3** Mean sensitivities with corresponding average numbers of false positives (FPs) in detecting PIRADS  $\geq 4$  Lesions.

Model	Sensitivity <sup>IoU-based</sup> <sub>Lesion-Level</sub>				Sensitivity <sup>overlap-based</sup> <sub>GT-perspective</sub>				Sensitivity <sup>IoU-based</sup> <sub>Box-Level</sub>			
	0.5FP	1FP	2FPs	3FPs	0.5FP	1FP	2FPs	3FPs	0.5FP	1FP	2FPs	3FPs
nnUNet	0.47 <sup>o</sup>	0.51 <sup>o</sup>	0.53	0.54	0.48	0.52	0.55	0.56	0.49	0.54	0.59	0.58
nnDetection	0.44	0.50	0.58 <sup>o</sup>	0.62 <sup>o</sup>	0.55 <sup>o</sup>	0.61 <sup>o</sup>	0.67 <sup>o</sup>	0.72 <sup>o</sup>	0.50 <sup>o</sup>	0.55 <sup>o</sup>	0.62 <sup>o</sup>	0.65 <sup>o</sup>
nnUNet Argmax	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Lesion-wise Ensemble	<b>0.49*</b>	0.56*	<b>0.64*</b>	<b>0.65*</b>	<b>0.59*</b>	0.62*	0.70*	<b>0.74*</b>	0.54*	0.58*	<b>0.65*</b>	<b>0.69*</b>
Lesion-wise Ensemble Argmax	0.48*	<b>0.56*</b>	0.62*	0.64*	0.56*	<b>0.64*</b>	<b>0.70*</b>	0.72	0.53*	<b>0.61*</b>	0.65*	0.68*
Voxel-wise Ensemble	0.47	0.51	0.51	0.57	0.57*	0.61	0.66	0.72	<b>0.55*</b>	0.59*	0.64*	0.66*
Voixe-wise Ensemble Argmax	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Note: \*: significantly better than all baselines that are nnUNet, nnDetection ( $p < 0.01$ ). <sup>o</sup>: the difference between the nnUNet and nnDetection are significant ( $p < 0.01$ ).

Abbreviation: PIRADS, prostate imaging reporting and data system.

while outperforms nnDetection at nearly all test levels of FPs in detecting PIRADS  $\geq 4$  lesions.

**PIRADS  $\geq 3$ .** As shown in Table 4, the DSC with nnUNet is greater than nnDetection at 1FP, 2FPs and 3FPs. At 0.5FP, as shown in Table 4, the DSC with nnUNet

are larger than nnDetection being  $0.16 \pm 0.01$  versus  $0.31 \pm 0.01$ .

**PIRADS  $\geq 4$ .** As shown in Table 3, nnUNet significantly outperforms nnDetection at the four sampled levels of FPs (i.e. 0.5,1,2,3) ( $p < 0.01$ ).



**TABLE 4** Mean precisions and Dice Similarity Coefficient (DSC) with corresponding average numbers of false positives (FPs) in detecting PIRADS  $\geq 3$  Lesions.

Model	Precision <sub>IoU-based Lesion-Level</sub>				Precision <sub>overlap-based Prediction-perspective</sub>				DSC			
	0.5FP	1FP	2FPs	3FPs	0.5FP	1FP	2FPs	3FPs	0.5FP	1FP	2FPs	3FPs
nnUNet	N/A	0.58	0.44	0.33	N/A	0.74	0.53	0.40	0.16	0.38°	<b>0.45°</b>	<b>0.45°</b>
nnDetection	0.69	0.61°	0.48°	<b>0.40°</b>	0.76	0.74	0.69°	0.66°	0.31°	0.35	0.36	0.35
nnUNet Argmax	<b>0.73</b>	0.60	N/A	N/A	0.81	0.74	N/A	N/A	<b>0.40</b>	<b>0.41</b>	N/A	N/A
Lesion-wise Ensemble	0.67	0.56	0.48	0.39	<b>0.84*</b>	<b>0.77*</b>	<b>0.71*</b>	<b>0.67*</b>	0.31	0.35	0.36	0.34
Lesion-wise Ensemble Argmax	<b>0.73</b>	<b>0.62*</b>	<b>0.50*</b>	<b>0.40*</b>	0.78	0.73	0.65	0.58	0.35	0.37	0.38	0.35
Voxel-wise Ensemble	N/A	0.59	0.40	0.18	N/A	0.73	0.49	0.25	0.16	0.38	0.40	0.29
Voxel-wise Ensemble Argmax	0.71	N/A	N/A	N/A	0.78	N/A	N/A	N/A	0.35	N/A	N/A	N/A

Note: \*: significantly better than all baselines that are nnUNet, nnDetection and nnUNet Argmax ( $p < 0.01$ ). °: the differences between nnUNet and nnDetection are significant ( $p < 0.01$ ).

Abbreviation: PIRADS, prostate imaging reporting and data system.

**TABLE 5** Mean precisions and dice similarity coefficient (DSC) with corresponding average numbers of false positives (FPs) in detecting PIRADS  $\geq 4$  Lesions.

Model	Precision <sub>IoU-based Lesion-Level</sub>				Precision <sub>overlap-based Prediction-perspective</sub>				DSC			
	0.5FP	1FP	2FPs	3FPs	0.5FP	1FP	2FPs	3FPs	0.5FP	1FP	2FPs	3FPs
nnUNet	0.35°	0.23°	0.13	0.10	0.39	0.28	0.18	0.14	<b>0.21°</b>	<b>0.19°</b>	<b>0.16°</b>	<b>0.16°</b>
nnDetection	0.33	0.22	0.14°	0.11°	<b>0.43°</b>	<b>0.35°</b>	<b>0.28°</b>	<b>0.23°</b>	0.19	0.18	0.15	0.13
nnUNet Argmax	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Lesion-wise Ensemble	<b>0.36*</b>	0.24*	0.16*	0.11	0.41	0.28	0.18	0.14	0.19	0.15	0.14	0.13
Lesion-wise Ensemble Argmax	0.35	<b>0.24*</b>	<b>0.16*</b>	<b>0.11*</b>	0.41	0.29	0.20	0.17	0.19	0.16	0.13	0.11
Voxel-wise Ensemble	0.35	0.23	0.13	0.10	0.36	0.26	0.16	0.12	0.19	0.16	0.14	0.13
Voxel-wise Ensemble Argmax	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Note: \*: significantly better than all baselines that are nnUNet, nnDetection ( $p < 0.01$ ). °: the differences between the nnUNet and nnDetection are significant ( $p < 0.01$ ).

Abbreviation: PIRADS, prostate imaging reporting and data system.

### 5.3 | Performance of lesion-wise ensemble methods

Lesion-wise ensemble methods can significantly improve the lesion-level sensitivities and precisions in detecting PIRADS  $\geq 3$  and PIRADS  $\geq 4$  lesions.

**PIRADS  $\geq 3$ . Sensitivities** Improved performances with respect to baselines are found with the *Lesion-wise Ensemble Argmax* (refer to Figure 1 and Table 2). According to results of paired *t*-tests, *Lesion-wise Ensemble Argmax* achieves significantly greater values than baselines (a) in terms of IoU-based lesion-level sensitivity at 1FP, 2FPs and 3 FPs; (b) in terms of IoU-based box-level sensitivity at 0.5FP, 2FPs, and 3FPs. For example, at 2 FPs, *Lesion-wise Ensemble Argmax* achieves Sensitivity<sub>IoU-based Lesion-Level</sub> of  $81.81\% \pm 1.41\%$ , outperforming  $63.81\% \pm 1.67\%$  (nnUNet) and  $76.05\% \pm 2.0\%$  (nnDetection). It is noted that nnUNet Argmax cannot achieve 2 FPs. *Lesion-wise Ensemble Argmax* also outperforms baselines at 0.5 FP, 2FPs and 3FPs for Sensitivity<sub>IoU-based Box-Level</sub> with statistical significance ( $p < 0.01$ ).

**Precisions** *Lesion-wise Ensemble* and *Lesion-wise Ensemble Argmax* outperform baselines at 0.5FP, 1FP, 2FPs

and 3FPs for Precision<sub>IoU-based Lesion-Level</sub> and Precision<sub>IoU-based Lesion-Level</sub>, respectively (refer to Table 4).

**PIRADS  $\geq 4$ . Sensitivities** Both ensemble methods achieve significantly greater sensitivities than baselines in almost all cases at 0.5FP, 1FP, 2FPs and 3FPs (refer to Table 3). The only exception is that the difference between *Lesion-wise Ensemble Argmax* ( $72.31\% \pm 4.26\%$ ) and *nnDetection* ( $72.15\% \pm 3.87\%$ ) at 3FP are not statistically significant for Sensitivity<sub>overlap-based GT-perspective</sub> ( $p$ -value = 0.3946).

**Precisions** *Lesion-wise Ensemble* and *Lesion-wise Ensemble Argmax* respectively achieve significantly higher values of Precision<sub>IoU-based Lesion-Level</sub> than baselines at 0.5FP, 1FP, 2FPs and 1FP, 2FPs, 3FPs (refer to Table 5). For Precision<sub>IoU-based Lesion-Level</sub>, (a) the difference between *Lesion-wise Ensemble* ( $11.06\% \pm 1.11\%$ ) and *nnDetection* ( $10.98\% \pm 1.23\%$ ) is not significant at 3FP ( $p$ -value = 0.3290); (b) the difference between *Lesion-wise Ensemble Argmax* ( $34.86\% \pm 4.07\%$ ) and *nnUNet* ( $35.33\% \pm 3.19\%$ ) is not statistically significant at 0.5FP ( $p$ -value = 0.1348). It is observed that both ensemble variants are inferior to *nnDetection* for Precision<sub>overlap-based Prediction-perspective</sub> (Table 5).

## 5.4 | Performance of voxel-wise ensemble methods

Voxel-wise ensemble methods can significantly improve the GT-perspective sensitivities for detecting both PIRADS  $\geq 3$  and PIRADS  $\geq 4$  lesions, and IoU-based box-level sensitivity for detecting PIRADS  $\geq 4$  lesions, both at certain levels of FPs.

**PIRADS  $\geq 3$ .** *Sensitivities Voxel-wise Ensemble Argmax* and *Voxel-wise Ensemble* has achieved significantly greater Sensitivity<sub>GT-perspective</sub><sup>overlap-based</sup> than the best of baselines at 0.5FP and 3FPs, with  $73.00\% \pm 1.94$  versus  $60.20\% \pm 1.56\%$  and  $97.99\% \pm 0.69\%$  versus  $94.34\% \pm 0.82\%$  respectively (refer to Table 2). Both ensemble variants perform worse than best of baselines at other sampled FPs and/or sensitivities.

*Precisions* Both voxel-wise ensemble variants are inferior to the best of baselines at 0.5FP, 1FP, 2FP and 3FP (Table 4).

**PIRADS  $\geq 4$ .** *Sensitivities* It is observed that *Voxel-wise Ensemble* outperform the baselines for IoU-based box-level sensitivity significantly at 0.5FP, 1FP, 2FPs, 3FPs (refer to Table 3). For Sensitivity<sub>Lesion-Level</sub><sup>IoU-based</sup> and Sensitivity<sub>GT-perspective</sub><sup>overlap-based</sup>, *Voxel-wise Ensemble* is worse than the best of baselines (refer to Table 3).

*Precisions* Similar to the case of PIRADS  $\geq 3$ , *Voxel-wise Ensemble* is worse than the best of baselines (Table 5).

## 6 | EXPERIMENTS AND RESULTS ON EXTERNAL DATA SET

In this section, we show the results of trained models being tested on the external data set with 100 patient scans, which were acquired from different institutions from those of the internal data set. In the external data, radiologist contours were obtained for all lesions with Likert-scores  $\geq 3$  and served as GT labels. The models used is the one that have been trained with the PIRADS  $\geq 3$  lesions. The mean and median lesion volume are 2.05 and 1.18 cm<sup>3</sup> (range, 0.11–18.34 cm<sup>3</sup>). In the external test data set, 81.48% lesions are in the PZ while 18.52% lesions are in the TZ, which is quite different from the lesions' spatial distribution in the internal data set as described in the Dataset of the Experiments Section. All the data were centre cropped and resampled to  $128 \times 128 \times 32$  with the spatial resolutions being  $0.625 \times 0.625 \times 3$  mm.

For brevity, we only show the two lesion-level sensitivities. Table 6 shows Sensitivity<sub>Lesion-Level</sub><sup>IoU-based</sup> and Sensitivity<sub>GT-perspective</sub><sup>overlap-based</sup> at the average numbers of FPs being 0.5, 1, 2, 3, 5, 10 FPs. In terms of Sensitivity<sub>Lesion-Level</sub><sup>IoU-based</sup>, either lesion-wise ensemble method achieve greater values ( $p < 0.01$ ) than nnDetection and nnUNet at five out of six FP levels (i.e. 1,2,3,5,10 FPs). In terms of Sensitivity<sub>GT-perspective</sub><sup>overlap-based</sup>, either lesion-wise or voxel-wise ensemble (or Argmax) methods achieve greater values ( $p < 0.01$ ) than

**TABLE 6** Mean Sensitivity<sub>Lesion-Level</sub><sup>IoU-based</sup> and Sensitivity<sub>GT-perspective</sub><sup>overlap-based</sup> on the external data set with corresponding average numbers of false positives (FPs) per patient, in detecting Likert-scores  $\geq 3$  Lesions.

Model	Sensitivity <sub>Lesion-Level</sub> <sup>IoU-based</sup>					
	0.5FP	1FP	2FPs	3FPs	5FPs	10FPs
nnUNet	0.08 $\pm$ 0.02	0.20 $\pm$ 0.03	0.32 $\pm$ 0.03	0.34 $\pm$ 0.03	0.35 $\pm$ 0.03	0.36 $\pm$ 0.03
nnDetection	<b>0.21° <math>\pm</math> 0.03°</b>	0.27° $\pm$ 0.03°	0.33° $\pm$ 0.04°	0.38° $\pm$ 0.04°	0.44° $\pm$ 0.04°	0.50° $\pm$ 0.04°
nnUNet Argmax	0.18 $\pm$ 0.03	0.26 $\pm$ 0.03	N/A	N/A	N/A	N/A
Lesion-wise Ensemble	0.13 $\pm$ 0.02	0.23 $\pm$ 0.03	0.34 $\pm$ 0.04	0.35 $\pm$ 0.04	0.42 $\pm$ 0.04	0.49 $\pm$ 0.04
Lesion-wise Ensemble Argmax	0.17 $\pm$ 0.03	<b>0.28* <math>\pm</math> 0.04*</b>	<b>0.36* <math>\pm</math> 0.04*</b>	<b>0.40* <math>\pm</math> 0.04*</b>	<b>0.45* <math>\pm</math> 0.04*</b>	<b>0.52* <math>\pm</math> 0.04*</b>
Voxel-wise Ensemble	0.08 $\pm$ 0.02	0.20 $\pm$ 0.03	0.32 $\pm$ 0.04	0.32 $\pm$ 0.04	0.29 $\pm$ 0.03	0.21 $\pm$ 0.03
Voxel-wise Ensemble Argmax	0.19 $\pm$ 0.03	0.28 $\pm$ 0.03	N/A	N/A	N/A	N/A
Model	Sensitivity <sub>GT-perspective</sub> <sup>overlap-based</sup>					
	0.5FP	1FP	2FPs	3FPs	5FPs	10FPs
nnUNet	0.09 $\pm$ 0.02	0.20 $\pm$ 0.03	0.47 $\pm$ 0.04	0.58 $\pm$ 0.04	0.63 $\pm$ 0.04	0.68 $\pm$ 0.04
nnDetection	0.33° $\pm$ 0.04°	0.45° $\pm$ 0.04°	0.61° $\pm$ 0.04°	<b>0.70° <math>\pm</math> 0.04°</b>	0.76° $\pm$ 0.03°	0.84° $\pm$ 0.03°
nnUNet Argmax	0.23 $\pm$ 0.03	0.32 $\pm$ 0.04	N/A	N/A	N/A	N/A
Lesion-wise Ensemble	0.15 $\pm$ 0.03	0.32 $\pm$ 0.04	0.56 $\pm$ 0.04	0.68 $\pm$ 0.04	<b>0.80* <math>\pm</math> 0.03*</b>	<b>0.88* <math>\pm</math> 0.03*</b>
Lesion-wise Ensemble Argmax	0.31 $\pm$ 0.04	<b>0.50* <math>\pm</math> 0.04*</b>	<b>0.64* <math>\pm</math> 0.04*</b>	0.68 $\pm$ 0.04	0.74 $\pm$ 0.04	0.87 $\pm$ 0.03
Voxel-wise Ensemble	0.09 $\pm$ 0.02	0.20 $\pm$ 0.03	0.43 $\pm$ 0.04	0.54 $\pm$ 0.04	0.70 $\pm$ 0.04	0.81 $\pm$ 0.03
Voxel-wise Ensemble Argmax	<b>0.33 <math>\pm</math> 0.04</b>	0.46 $\pm$ 0.04	N/A	N/A	N/A	N/A

Note: \*: significantly better than all baselines including nnUNet, nnDetection and nnUNet Argmax ( $p < 0.01$ ). °: the differences between nnUNet and nnDetection are significant ( $p < 0.01$ ). N/A: not applicable.

nnDetection and nnUNet at four out of six FP levels (i.e. 1,2,5,10 FPs). Although lower sensitivity values have been achieved on the external data set than those tested on the internal data set. Two similar trends/observations/conclusions from both the internal and external data sets: nnDetection outperforms nnUNet in terms of lesion-level sensitivities and ensemble methods can significantly improve the lesion-level metrics.

## 7 | DISCUSSIONS

Results reveal that (1) nnDetection generally achieves significant greater lesion-level sensitivities and precisions than nnUNet; (2) nnUNet (or its variants nnUNet Argmax) reaches higher DSC than nnDetection at most levels of FPs; (3) ensemble nnUNet and nnDetection significantly improves or is comparable with the individual, regarding of the localisation ability of detecting PIRADS  $> 3$  and PIRADS  $\geq 4$  lesions.

Results reported in this study also indicate the strong impact of the evaluation metric for comparing algorithms in prostate cancer detection, where slight different observations and thus following conclusions may be made. For example, it is observed that under different definitions of sensitivities both *Voxel-wise Ensemble* (dashed purple line) and *nnUNet* (solid red line) methods exhibit different trends with increasing FPs (refer to Figure 1, more specifically comparing Figures 1(a,c) and (b)). It is also noticed that all methods generally reach higher values with the definition of prediction-perspective lesion-level precision (b) than with that of IoU-based lesion-level precision (a). For example, at 3FP, nnDetection achieves  $\text{Precision}_{\text{Lesion-Level}}^{\text{Prediction-perspective}}$  of  $39.70\% \pm 1.24\%$  versus  $\text{Precision}_{\text{Lesion-Level}}^{\text{IoU-based}}$  of  $65.70\% \pm 1.83\%$  and  $23.36\%$  (0.0273) versus  $10.98\% \pm 1.23\%$  in detecting PIRADS  $\geq 3$  and PIRADS  $\geq 4$  lesions respectively. The difference can be explained by the fact that one prediction is considered as TP if its ‘overlap’ with all GT lesions exceeds the preset threshold (set to be 0.1 in this study) using the definition of  $\text{Precision}_{\text{Lesion-Level}}^{\text{Prediction-perspective}}$  in Lesion-Level Evaluation Metrics of the Methods Section whilst it has to ‘overlap’ with one specific GT lesion more than that preset threshold using the definition of  $\text{Precision}_{\text{Lesion-Level}}^{\text{IoU-based}}$  in Lesion-Level Evaluation Metrics of the Methods Section.

Tables 7 and 8 show the maximum sensitivities and corresponding average number of FPs, in detecting PIRADS  $\geq 3$  and PIRADS  $\geq 4$  lesions respectively. For detecting PIRADS  $\geq 3$  lesions, *Lesion-wise Ensemble Argmax* achieves the maximum  $\text{Sensitivity}_{\text{Lesion-Level}}^{\text{IoU-based}}$  and  $\text{Sensitivity}_{\text{Box-Level}}^{\text{IoU-based}}$  of  $85.53\% \pm 1.28\%$  and  $91.0\% \pm 0.80\%$  at 5 FPs. *Voxel-wise Ensemble* achieves the maximum  $\text{Sensitivity}_{\text{GT-perspective}}^{\text{overlap-based}}$  of  $98.37\% \pm 0.53\%$  at 5 FPs. nnUNet achieves the maximum DSC of  $0.46 \pm 0.01$  at 1.75 FPs. Due to the fact that detecting prostate cancer in MRI scans is quite a challenging task, the DSC being less than 0.5 in the scenario of automated prostate cancer detection using deep-learning is reasonable and consistently with other reported in the literature. For example,

the study that utilises 2D UNet as the backbone and the Gleason Grade group as the GT has reported the DSC being  $0.370 \pm 0.046$  for segmenting clinically significant cancer (defined as GGG  $\geq 2$ )<sup>2</sup>. For detecting PIRADS  $\geq 4$  lesions, *Lesion-wise Ensemble Argmax* achieves the maximum  $\text{Sensitivity}_{\text{Lesion-Level}}^{\text{IoU-based}}$  of  $67.84\% \pm 3.55\%$  at 5 FPs. *Voxel-wise Ensemble* achieves the maximum  $\text{Sensitivity}_{\text{GT-perspective}}^{\text{overlap-based}}$  of  $77.27\% \pm 2.84\%$  at 5 FPs. *Lesion-wise Ensemble* achieves the maximum  $\text{Sensitivity}_{\text{Box-Level}}^{\text{IoU-based}}$  of  $73.14\% \pm 2.85\%$  at 5 FPs. nnUNet achieves the maximum DSC of  $0.22 \pm 0.02$  at 0.38 FP. It is noted that the DSC in the range of [0.17,0.22] in detecting the PIRADS  $\geq 4$  lesions are considerably smaller than the DSC lying in the range of [0.35, 0.46] in detecting the PIRADS  $\geq 3$  lesions. The reasons are two-folds: there are much fewer PIRADS  $\geq 4$  lesions than PIRADS  $\geq 3$  lesions; in the binary segmentation/detection of PIRADS  $\geq 4$  lesions, there are patients with no GT PIRADS  $\geq 4$  lesions while the DSC is computed patient-wise and averaged over all patients. As an example, if where is GT PIRADS  $\geq 4$  lesions but deep-learning models predict regions in one MRI scan, the DSC computed with (8) is expected to be low. Thus, the way of computing the DSC is pessimistic but the results are reasonable. To conclude, for both detecting PIRADS  $\geq 3$  and PIRADS  $\geq 4$  lesions (a) *Lesion-wise* ensemble methods achieves the largest IoU-based lesion-level/box-level sensitivities; (b) *Voxel-wise* ensemble method achieves the largest GT-perspective sensitivity; (c) nnUNet reaches the largest DSC.

For clinical applications like MRI-targeted biopsies that require missing as few clinically significant prostatic lesions as possible [8], ensemble methods seem better choices with its demonstrated greater sensitivities and precisions. Given that one lesion has already detected, nnUNet with its higher DSC maybe a better choice for applications like focal therapy that requires as high lesion coverage as possible.

This study has limitations in a few aspects. First, the fact that the raw nnDetection predictions are BBs may lead to bias towards nnUNet in the evaluation when the metrics are based on the lesion masks, and vice versa (i.e. it may lead to bias towards nnDetection when the box-level metrics are utilised). More specifically, the segmentation network outputs the precise segmentation (that can be of any shapes) mask of interested region while the detection network can only predict BBs (that can only be rectangles). Thus, if the evaluation GT is segmented masks of lesions (that can be of any shapes), the evaluation will be biased towards the segmentation network. Second, all networks are trained with and thus predict the radiologist-annotated PIRADS lesions, which themselves have a wide range of sensitivity and specificity due to inter-reader variability and sub-optimal analysis [28]. In the future, studies on the real-world clinical data with the histological GT will be conducted, to see the extent of detecting histological-confirmed prostate cancer of deep-learning-based segmentation and detection models. Third, although we have validated the approaches on one external data set, it should be noted that the observations and conclusions cannot be directly

**TABLE 7** Maximum sensitivities with corresponding average numbers of false positives (FPs) in detecting PIRADS  $\geq 3$  Lesions. Inside the parentheses are the numbers of FPs.

Model	Sensitivity <sup>IoU-based</sup> <sub>Lesion-Level</sub>	Sensitivity <sup>overlap-based</sup> <sub>GT-perspective</sub>	Sensitivity <sup>IoU-based</sup> <sub>Box-Level</sub>	DSC
nnUNet	70.88% $\pm$ 1.77%(1.25)	93.46% $\pm$ 0.87%(5)	73.92% $\pm$ 1.56%(1.25)	<b>0.46 <math>\pm</math> 0.01(1.75)</b>
nnDetection	83.31% $\pm$ 1.73%(5)	96.65% $\pm$ 0.83%(5)	90.05% $\pm$ 1.18%(5)	0.36 $\pm$ 0.01(1.75)
nnUnet Argmax	66.9 $\pm$ 1.64%(1)	65.99% $\pm$ 1.55%(1)	64.55% $\pm$ 1.36%(0.88)	0.41 $\pm$ 0.02 (0.63)
Lesion-wise Ensemble	83.18% $\pm$ 1.54%(5)	97.65% $\pm$ 0.64%(5)	88.58% $\pm$ 1.56%(5)	0.36 $\pm$ 0.01(2)
Lesion-wise Ensemble Argmax	<b>85.53% <math>\pm</math> 1.28%(5)</b>	97.07% $\pm$ 0.64%	<b>91.0% <math>\pm</math> 0.80%(5)</b>	0.38 $\pm$ 0.01(1.75)
Voxel-wise Ensemble	70.49% $\pm$ 1.50%(1.88)	<b>98.37% <math>\pm</math> 0.53%(5)</b>	74.17% $\pm$ 1.58%(1.75)	0.45 $\pm$ 0.01(1.88)
VoXe-wise Ensemble Argmax	53.23% $\pm$ 1.72%(0.75)	75.62% $\pm$ 1.78%(0.75)	58.39% $\pm$ 1.89%(0.63)	0.35 $\pm$ 0.01(0.75)

Abbreviation: PIRADS, prostate imaging reporting and data system.

**TABLE 8** Maximum sensitivities and corresponding average numbers of false positives (FPs) in detecting PIRADS  $\geq 4$  Lesions.

Model	Sensitivity <sup>IoU-based</sup> <sub>Lesion-Level</sub> (FPs)	Sensitivity <sup>overlap-based</sup> <sub>GT-perspective</sub> (FPs)	Sensitivity <sup>IoU-based</sup> <sub>Box-Level</sub> (FPs)	DSC (FPs)
nnUNet	54.88% $\pm$ 5.16%(2.63)	57.10% $\pm$ 4.29%(5)	60.98% $\pm$ 4.45%(5)	<b>0.22 <math>\pm</math> 0.02(0.38)</b>
nnDetection	66.42% $\pm$ 3.42%(5)	76.42% $\pm$ 3.26%(5)	69.29% $\pm$ 3.55%(5)	0.20 $\pm$ 0.02(0.5)
nnUnet Argmax	31.03% $\pm$ 4.68%(0.08)	27.33% $\pm$ 3.47%(0.08)	28.18% $\pm$ 3.44%(0.05)	0.17 $\pm$ 0.02(0.08)
Lesion-wise Ensemble	67.5% $\pm$ 3.90%(5)	76.89% $\pm$ 2.90%(5)	<b>73.14% <math>\pm</math> 2.85%(5)</b>	0.19 $\pm$ 0.02(0.25)
Lesion-wise Ensemble Argmax	<b>67.84% <math>\pm</math> 3.55%(5)</b>	74.42% $\pm$ 3.82(5)	70.55% $\pm$ 3.61%(5)	0.20 $\pm$ 0.02(0.38)
Voxel-wise Ensemble	58.04% $\pm$ 3.68%(5)	<b>77.27% <math>\pm</math> 2.84%(5)</b>	66.74% $\pm$ 4.74%(2.75)	0.19 $\pm$ 0.02(0.25)
VoXe-wise Ensemble Argmax	30.74% $\pm$ 3.96%(0.08)	31.72% $\pm$ 4.20%(0.08)	30.07% $\pm$ 4.48%(0.08)	0.21 $\pm$ 0.03(0.08)

Abbreviation: PIRADS, prostate imaging reporting and data system.

generalised to new data sets. We plan to conduct more experiments for cross-validation between more data sets [43].

ProstAttention-Net achieves 69.0%  $\pm$  14.5% sensitivity at 2.9 FP per patient, where the clinically significant lesions are defined as those with GS  $>$  6 [3]. As shown in Table 2, at 3 FPs, nnUNet, nnDetection, Lesion-level Ensemble and Lesion-level Ensemble Argmax achieve 60.40%  $\pm$  1.64%, 80.78%  $\pm$  1.50%, 77.34%  $\pm$  2.01% and 82.24%  $\pm$  1.43% in detecting the PIRADS  $\geq 3$  lesions, respectively. It should be taken into consideration that the GT labels are not the same between ProstAttention-Net (GS) and our work (PIRADS Score), and thus the direct comparisons of these numbers should be taken with care.

Following aspects will be explored in the future. First, multi-class segmentation and object detection networks will be trained where lesions with different PIRADS or Gleason scores are distinguished and utilised. Second, ensemble methods of segmentation and detection networks at the training stage will be developed, in hope of further enhancing the performances. For example, utilising the segmented masks in the training of an object detection network has great potential for improving its voxel-level performance while maintaining high lesion-level sensitivity and precision at the same time.

## 8 | CONCLUSION

We believe that this paper provides an important set of results to advance further development in this application area, addressing a number of unanswered questions

regarding methodology choice, real-world data performance and rigorous validation. Results on real-world clinical data demonstrate that the object detection method generally achieves higher lesion-level performances while the segmentation reaches greater voxel-level DSCs, and that the proposed ensemble algorithms combining segmentation and detection methods have shown effectiveness in improving localisation accuracies. The proposed method potentially opens up the possibility for exploring new ways to fully utilise both deep-learning-based segmentation and object detection approaches for automatic prostate cancer detection in MRI (also general CAD from medical images).

## ACKNOWLEDGEMENTS

This work was supported by the International Alliance for Cancer Early Detection, an alliance between Cancer Research UK [C28070/A30912; C73666/A31378], Canary Center at Stanford University, the University of Cambridge, OHSU Knight Cancer Institute, University College London and the University of Manchester. This work was also supported by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences [203145Z/16/Z]. This work was also supported in part by the National Natural Science Foundation of China under Grant 62303275.

## CONFLICT OF INTEREST STATEMENT

None.

## DATA AVAILABILITY STATEMENT

Research data are not shared.

## ORCID

Zhe Min  <https://orcid.org/0000-0002-8903-1561>

## REFERENCES

1. Ferlay, J., et al.: Cancer statistics for the year 2020: an overview. *Int. J. Cancer* 149(4), 778–789 (2021). <https://doi.org/10.1002/ijc.33588>
2. Yu, X., et al.: Deep attentive panoptic model for prostate cancer detection using biparametric mri scans. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 594–604. Springer (2020)
3. Duran, A., et al.: Prostattention-net: a deep attention model for prostate cancer segmentation by aggressiveness in mri scans. *Med. Image Anal.* 102347 (2022)
4. Min, Z., et al.: Controlling false positive/negative rates for deep-learning-based prostate cancer detection on multiparametric mr images. In: *Annual Conference on Medical Image Understanding and Analysis*, pp. 56–70. Springer (2021)
5. Saha, A., Hosseinzadeh, M., Huisman, H.: End-to-end prostate cancer detection in bpmri via 3d cnns: effect of attention mechanisms, clinical priori and decoupled false positive reduction. *arXiv preprint arXiv:2101.03244* (2021)
6. Israël, B., et al.: Multiparametric magnetic resonance imaging for the detection of clinically significant prostate cancer: what urologists need to know. part 2: interpretation. *Eur. Urol.* 77(4), 469–480 (2020). <https://doi.org/10.1016/j.eururo.2019.10.024>
7. Kasivisvanathan, V., et al.: Mri-targeted or standard biopsy for prostate-cancer diagnosis. *N. Engl. J. Med.* 378(19), 1767–1777 (2018). <https://doi.org/10.1056/nejmoa1801993>
8. Klotz, L., et al.: Comparison of multiparametric magnetic resonance imaging–targeted biopsy with systematic transrectal ultrasonography biopsy for biopsy-naïve men at risk for prostate cancer: a phase 3 randomized clinical trial. *JAMA Oncol.* 7(4), 534–542 (2021). <https://doi.org/10.1001/jamaoncol.2020.7589>
9. Ahmed, H.U., et al.: Diagnostic accuracy of multi-parametric mri and trus biopsy in prostate cancer (promis): a paired validating confirmatory study. *Lancet* 389(10071), 815–822 (2017). [https://doi.org/10.1016/S0140-6736\(16\)32401-1](https://doi.org/10.1016/S0140-6736(16)32401-1)
10. Mottet, N., et al.: Eau-eanm-estro-esur-siog guidelines on prostate cancer—2020 update. part 1: screening, diagnosis, and local treatment with curative intent. *Eur. Urol.* 79(2), 243–262 (2021). <https://doi.org/10.1016/j.eururo.2020.09.042>
11. Litjens, G., et al.: Computer-aided detection of prostate cancer in mri. *IEEE Trans. Med. Imag.* 33(5), 1083–1092 (2014). <https://doi.org/10.1109/tmi.2014.2303821>
12. Saha, A., Hosseinzadeh, M., Huisman, H.: Encoding clinical priori in 3d convolutional neural networks for prostate cancer detection in bpmri. *arXiv preprint arXiv:2011.00263* (2020)
13. Chen, Y., et al.: Automatic intraprostatic lesion segmentation in multiparametric magnetic resonance images with proposed multiple branch unet. *Med. Phys.* 47(12), 6421–6429 (2020). <https://doi.org/10.1002/mp.14517>
14. Chiou, E., et al.: Harnessing uncertainty in domain adaptation for mri prostate lesion segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 510–520. Springer (2020)
15. Hiremath, A., et al.: Test-retest repeatability of a deep learning architecture in detecting and segmenting clinically significant prostate cancer on apparent diffusion coefficient (adc) maps. *Eur. Radiol.* 31(1), 379–391 (2021). <https://doi.org/10.1007/s00330-020-07065-4>
16. Weinreb, J.C., et al.: Pi-rads prostate imaging–reporting and data system: 2015, version 2. *Eur. Urol.* 69(1), 16–40 (2016). <https://doi.org/10.1016/j.eururo.2015.08.052>
17. Schelb, P., et al.: Classification of cancer at prostate mri: deep learning versus clinical pi-rads assessment. *Radiology* 293(3), 607–617 (2019). <https://doi.org/10.1148/radiol.2019190938>
18. Duran, A., Jodoin, P.-M., Lartizien, C.: Prostate cancer semantic segmentation by gleason score group in bi-parametric mri with self attention model on the peripheral zone. In: *Medical Imaging with Deep Learning*, pp. 193–204. PMLR (2020)
19. Dai, Z., et al.: Segmentation of the prostatic gland and the intraprostatic lesions on multiparametric magnetic resonance imaging using mask region-based convolutional neural networks. *Advances in Radiation Oncology* 5(3), 473–481 (2020). <https://doi.org/10.1016/j.adro.2020.01.005>
20. Yan, W., et al.: The Impact of Using Voxel-Level Segmentation Metrics on Evaluating Multifocal Prostate Cancer Localisation (2022). <https://doi.org/10.48550/ARXIV.2203.16415>
21. Isensee, F., et al.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* 18(2), 203–211 (2021). <https://doi.org/10.1038/s41592-020-01008-z>
22. Baumgartner, M. et al.: A self-configuring method for medical object detection. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 530–539. Springer International Publishing, Cham (2021)
23. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer (2015)
24. Çiçek, Ö., et al.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432. Springer (2016)
25. Zhou, Z., et al.: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imag.* 39(6), 1856–1867 (2019). <https://doi.org/10.1109/tmi.2019.2959609>
26. Milletari, F., Navab, N., Ahmadi, S.-A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571. IEEE (2016)
27. Saha, A., et al.: Anatomical and diagnostic bayesian segmentation in prostate mri – should different clinical objectives mandate different loss functions? *arXiv preprint arXiv:2110.12889* (2021)
28. Cao, R., et al.: Joint prostate cancer detection and gleason score prediction in mp-mri via focalnet. *IEEE Trans. Med. Imag.* 38(11), 2496–2506 (2019). <https://doi.org/10.1109/tmi.2019.2901928>
29. Yu, X., et al.: 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)*, 1355–1359. IEEE (2020)
30. Oktay, O., et al.: Attention u-net: learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999* (2018)
31. Isensee, F., et al.: No new-net. In: *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, 234–244. Springer International Publishing, Cham (2019)
32. Litjens, G., et al.: Evaluation of prostate segmentation algorithms for mri: the promise12 challenge. *Med. Image Anal.* 18(2), 359–373 (2014). <https://doi.org/10.1016/j.media.2013.12.002>
33. Simpson, A.L., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063* (2019)
34. Liu, L., et al.: Deep learning for generic object detection: a survey. *Int. J. Comput. Vis.* 128(2), 261–318 (2020). <https://doi.org/10.1007/s11263-019-01247-4>
35. Ren, S., et al.: J. Faster r-cnn: towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015)
36. He, K., et al.: Mask r-cnn. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969. Poverkhnost (2017)
37. Lin, T.-Y., et al.: Focal loss for dense object detection. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988 (2017)
38. Jaeger, P.F., et al.: Retina u-net: embarrassingly simple exploitation of segmentation supervision for medical object detection. In: *Machine Learning for Health Workshop*, pp. 171–183. PMLR (2020)
39. Setio, A.A.A., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed

- tomography images: the luna16 challenge. *Med. Image Anal.* 42, 1–13 (2017). <https://doi.org/10.1016/j.media.2017.06.015>
40. Hamid, S., et al.: The smarttarget biopsy trial: a prospective, within-person randomised, blinded trial comparing the accuracy of visual-registration and magnetic resonance imaging/ultrasound image-fusion targeted biopsies for prostate cancer risk stratification. *Eur. Urol.* 75(5), 733–740 (2019). <https://doi.org/10.1016/j.eururo.2018.08.007>
  41. Dickinson, L., et al.: A multi-centre prospective development study evaluating focal therapy using high intensity focused ultrasound for localised prostate cancer: the index study. *Contemp. Clin. Trials* 36(1), 68–80 (2013). <https://doi.org/10.1016/j.cct.2013.06.005>
  42. Simmons, L.A., et al.: Accuracy of transperineal targeted prostate biopsies, visual estimation and image fusion in men needing repeat biopsy in the picture trial. *J. Urol.* 200(6), 1227–1234 (2018). <https://doi.org/10.1016/j.juro.2018.07.001>
  43. Bi, Y., et al.: Mutual information-based us segmentation for unseen domain generalization. *arXiv preprint arXiv:2303.12649* (2023)

**How to cite this article:** Min, Z., et al.: Segmentation versus detection: Development and evaluation of deep learning models for prostate imaging reporting and data system lesions localisation on Bi-parametric prostate magnetic resonance imaging. *CAAI Trans. Intell. Technol.* 1–14 (2024). <https://doi.org/10.1049/cit2.12318>