Original Article

Using an artificial intelligence model to detect and localize visible clinically significant prostate cancer in prostate magnetic resonance imaging: a multicenter external validation study

Zhaonan Sun[1], Kexin Wang[2], Chenchao Wu[3], Yuntian Chen[4], Zixuan Kong[5], Lilan She[3], Bin Song[4], Ning Luo[5], Pengsheng Wu[6], Xiangpeng Wang[6], Xiaodong Zhang[1], Xiaoying Wang[1]

[1]Department of Radiology, Peking University First Hospital, Beijing, China; [2]School of Basic Medical Sciences, Capital Medical University, Beijing, China; [3]Department of Radiology, Fujian Medical University Union Hospital, Fuzhou, China; [4]Department of Radiology, West China Hospital, Sichuan University, Chengdu, China; [5]Department of Radiology, The Second Affiliated Hospital of Dalian Medical University, Dalian, China; [6]Beijing Smart Tree Medical Technology Co. Ltd., Beijing, China

Contributions: (I) Conception and design: Xiaoying Wang; (II) Administrative support: Z Sun, X Zhang; (III) Provision of study materials or patients: C Wu, Y Chen, Z Kong, L She, B Song, N Luo; (IV) Collection and assembly of data: Z Sun, K Wang; (V) Data analysis and interpretation: Z Sun, K Wang, P Wu, Xiangpeng Wang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Xiaoying Wang, MD, PhD. Department of Radiology, Peking University First Hospital, 8 Xishiku Street, Xicheng District, Beijing 100034, China. Email: wangxiaoying@bjmu.edu.cn.

Background: An increasing number of patients with suspected clinically significant prostate cancer (csPCa) are undergoing prostate multiparametric magnetic resonance imaging (mpMRI). The role of artificial intelligence (AI) algorithms in interpreting prostate mpMRI needs to be tested with multicenter external data. This study aimed to investigate the diagnostic efficacy of an AI model in detecting and localizing visible csPCa on mpMRI a multicenter external data set.

Methods: The data of 2,105 patients suspected of having prostate cancer from four hospitals were retrospectively collected to develop an AI model to detect and localize suspicious csPCa. The lesions were annotated based on pathology records by two radiologists. Diffusion-weighted imaging (DWI) and apparent diffusion coefficient (ADC) values were used as the input for the three-dimensional U-Net framework. Subsequently, the model was validated using an external data set comprising the data of 557 patients from three hospitals. Sensitivity, specificity, and accuracy were employed to evaluate the diagnostic efficacy of the model.

Results: At the lesion level, the model had a sensitivity of 0.654. At the overall sextant level, the model had a sensitivity, specificity, and accuracy of 0.846, 0.884, and 0.874, respectively. At the patient level, the model had a sensitivity, specificity, and accuracy of 0.943, 0.776, and 0.849, respectively. The AI-predicted accuracy for the csPCa patients (231/245, 0.943) was significantly higher than that for the non-csPCa patients (242/312, 0.776) ($P<0.001$). The lesion number and tumor volume were greater in the correctly diagnosed patients than the incorrectly diagnosed patients (both $P<0.001$). Among the positive patients, those with lower average ADC values had a higher rate of correct diagnosis than those with higher average ADC values ($P=0.01$).

Conclusions: The AI model exhibited acceptable accuracy in detecting and localizing visible csPCa at the patient and sextant levels. However, further improvements need to be made to enhance the sensitivity of the model at the lesion level.

Keywords: Clinically significant prostate cancer (csPCa); artificial intelligence (AI); magnetic resonance imaging (MRI)

Submitted Jun 02, 2023. Accepted for publication Oct 07, 2023. Published online Oct 29, 2023.
doi: 10.21037/qims-23-791
View this article at: https://dx.doi.org/10.21037/qims-23-791

Introduction

Prostate multiparametric magnetic resonance imaging (mpMRI) is a non-invasive examination used for the detection and localization of clinically significant prostate cancer (csPCa). It aims to increase the positive rate of biopsies and reduce unnecessary biopsies (1). Multiple studies (1-3) have advocated for the use of mpMRI prior to biopsy to identify high-risk patients and pinpoint target areas for subsequent biopsy. However, work efficiency needs to be enhanced and variation in prostate mpMRI interpretations needs to be minimized to meet the escalating demand for MRI-based diagnosis (4).

The Prostate Imaging Reporting and Data System (PI-RADS) (5-8) is continuously updated to standardize prostate mpMRI interpretation. However, it still demonstrates some degree of variability (9,10). Intra-reader agreement ranges from 60% to 74%, while inter-reader agreement falls below 50% (9). Such variation has resulted in discrepancies in the diagnostic efficacy of image-guided targeted biopsies, with csPCa detection rates varying by up to 40% for PI-RADS 5 lesions (11). Further, the steep learning curve poses challenges for practitioners, and mpMRI reporting necessitates a high level of expertise (12,13).

The integration of deep-learning models and mpMRI has shown promise in providing automated and scalable assistance in identifying biopsy candidates and guiding biopsy sampling (14). Previous studies have reported that the performance of prostate artificial intelligence (AI) models has been encouraging in specific data sets; however, the generalization and clinical applicability of these models have not been extensively investigated. AI models need to be validated in external cohorts before any consideration is given to their clinical deployment (15).

In this multicenter study, we developed an AI model and validated it in an external data set to evaluate its diagnostic efficacy in detecting and localizing csPCa. We present this article in accordance with the TRIPOD reporting checklist (available at https://qims.amegroups.com/article/view/10.21037/qims-23-791/rc).

Methods

This retrospective study was approved by the Institutional Review Board (IRB) of Peking University First Hospital (IRB number: 2021060). The requirement of individual consent for this retrospective analysis was waived. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Data inclusion

Figure 1 shows the data enrollment process. In total, 2,221 consecutive prostate MR images acquired between January 2014 and December 2019 were retrospectively collected from 16 MR scanners in four hospitals (Hospital 1, Hospital 2, Hospital 3, and Hospital 4) to develop an AI model for csPCa segmentation. Further, 580 consecutive prostate MR images acquired between January 2020 and November 2021 were retrospectively collected from 14 MR scanners in three hospitals (Hospital 1, Hospital 2, and Hospital 3) to establish an external validation data set. Importantly, the images used in the development data set for the AI model and the external validation data set were mutually exclusive and adhered to the same inclusion and exclusion criteria.

To be eligible for inclusion in the study, patients had to meet the following inclusion criteria: (I) have undergone mpMRI prior to biopsy with a clinical suspicion of prostate cancer (PCa) due to an elevated serum prostate-specific antigen (PSA) level, abnormal findings during a digital rectal examination, and/or abnormal transrectal ultrasound results; (II) have undergone an image-guided biopsy, transurethral prostatectomy, or radical prostatectomy within one month of the MRI examination and received pathological confirmation; (III) have not undergone any PCa-related treatment prior to the examination; and (IV) have tested negative for PCa during biopsy and showed no potential signs of PCa during clinical follow-up for over 1 year. Clinical information, including age, total PSA levels in serum, and pathology results, was collected for all patients.

Patients were excluded from the study if they met any

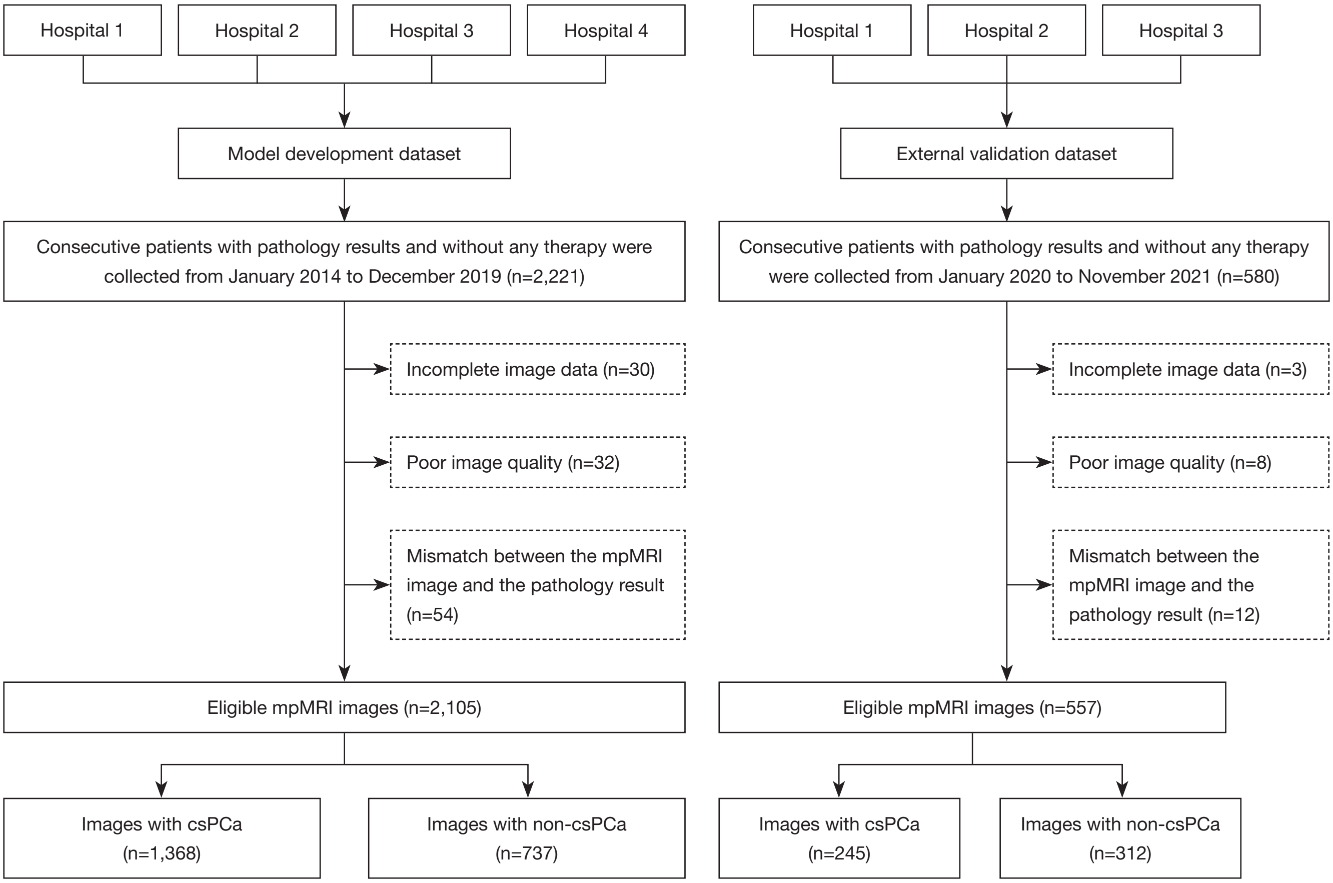

Hospital 1
Hospital 2
Hospital 3
Hospital 4
Model development dataset
Consecutive patients with pathology results and without any therapy were collected from January 2014 to December 2019 (n=2,221)
Incomplete image data (n=30)
Poor image quality (n=32)
Mismatch between the mpMRI image and the pathology result (n=54)
Eligible mpMRI images (n=2,105)
Images with csPCa (n=1,368)
Images with non-csPCa (n=737)
Hospital 1
Hospital 2
Hospital 3
External validation dataset
Consecutive patients with pathology results and without any therapy were collected from January 2020 to November 2021 (n=580)
Incomplete image data (n=3)
Poor image quality (n=8)
Mismatch between the mpMRI image and the pathology result (n=12)
Eligible mpMRI images (n=557)
Images with csPCa (n=245)
Images with non-csPCa (n=312)

Figure 1 Diagram showing the inclusion of patients in the study. mpMRI, multiparametric magnetic resonance imaging; csPCa, clinically significant prostate cancer.

of the following exclusion criteria: (I) they had incomplete image data; (II) the quality of the image was poor; and/or (III) there were inconsistencies between the MRI image and the pathology results, such as variations in tumor location, MRI-invisible csPCa, and imaging that indicated csPCa but pathology results that confirmed its absence.

MR scanning protocols

Prostate MRI images were acquired using 16 different MR scanners for the model development data set and 14 different MR scanners for the external validation data set. The imaging setup involved the use of body coils as transmit coils and phased array coils as receiver coils, but no use of endorectal coils. The MRI sequences included T1-weighted imaging, T2-weighted imaging (T2WI), diffusion-weighted imaging (DWI), and apparent diffusion coefficient (ADC) maps. The acquisition of a dynamic contrast-enhanced sequence was not mandatory. For the DWI, a diffusion-weighted single-shot gradient echo planar imaging sequence was used. For the T2WI, a T2-weighted fast spin echo sequence at both 3.0 T and 1.5 T was used. The ADC maps were generated from the DWI sequence using high and low b-values. Detailed information regarding the MR scanning protocols for both the model development data set and the external validation data set is presented in *Table 1*.

Patient classification and localization of csPCa

All the patients included in this study underwent a transrectal ultrasonography-guided systematic biopsy using either 12- or 6-core needles, as well as a cognitive-targeted biopsy. The process of the systematic and cognitive-targeted biopsy is described in detail in Appendix 1. The pathology results from both biopsy approaches were combined (16). Sextants showing the highest grade assigned by either

Table 1 Information of MR scanners and protocols

Parameters	Model development data set				External validation data set
	Training (N=1,681)	Validation (N=212)	Test (N=212)	P	Overall (N=557)
DWI/ADC					
Model name				0.87	
Achieva	52 (3.1)	5 (2.4)	4 (1.9)		13 (2.3)
Aera	217 (12.9)	32 (15.1)	22 (10.4)		22 (3.9)
Amira	1 (0.1)	1 (0.5)	0 (0.0)		0 (0.0)
DISCOVERY MR750	844 (50.2)	112 (52.8)	109 (51.4)		243 (43.6)
DISCOVERY MR750w	75 (4.5)	13 (6.1)	7 (3.3)		44 (7.9)
Ingenia	107 (6.4)	8 (3.8)	19 (9.0)		45 (8.1)
Ingenia CX	1 (0.1)	0 (0.0)	0 (0.0)		1 (0.2)
Multiva	24 (1.4)	2 (0.9)	4 (1.9)		3 (0.5)
Prisma	29 (1.7)	2 (0.9)	4 (1.9)		20 (3.6)
SIGNA EXCITE	36 (2.1)	2 (0.9)	4 (1.9)		0 (0.0)
Signa HDxt	3 (0.2)	0 (0.0)	0 (0.0)		1 (0.2)
SIGNA Premier	1 (0.1)	0 (0.0)	0 (0.0)		0 (0.0)
Skyra	46 (2.7)	6 (2.8)	8 (3.8)		25 (4.5)
TrioTim	114 (6.8)	17 (8.0)	18 (8.5)		108 (19.4)
uMR 790	87 (5.2)	8 (3.8)	7 (3.3)		6 (1.1)
Verio	44 (2.6)	4 (1.9)	6 (2.8)		25 (4.5)
MAGNETOM_ESSENZA	0 (0.0)	0 (0.0)	0 (0.0)		1 (0.2)
Magnetic field				0.38	
1.5 T	254 (15.1)	36 (17.0)	26 (12.3)		27 (4.8)
3.0 T	1,427 (84.9)	176 (83.0)	186 (87.7)		530 (95.2)
B value (s/mm^2)	1,400 [1,400, 1,400]	1,400 [1,400, 1,400]	1,400 [1,400, 1,400]	0.70	1,400 [1,400, 1,400]
Slice thickness (mm)	4.00 [4.00, 4.00]	4.00 [4.00, 4.00]	4.00 [4.00, 4.00]	0.96	4.0 [3.0, 5.0]
Slice spacing (mm)	4.00 [4.00, 4.50]	4.00 [4.00, 4.50]	4.00 [4.00, 4.50]	0.87	4.0 [3.0, 6.5]
Repetition time (ms)	3,000 [2,640, 4,380]	2,930 [2,640, 4,110]	2,910 [2,640, 4,130]	0.76	2,671 [2,000, 6,759]
Echo time (ms)	61.3 [59.7, 63.8]	61.2 [60.0, 63.7]	61.3 [59.7, 63.5]	0.97	61 [51, 93]
Field of view (mm)	240 [220, 250]	240 [220, 250]	240 [220, 250]	0.34	240 [220, 250]
Flip angle (°)	90 [90, 90]	90 [90, 90]	90 [90, 90]	0.81	90 [90, 90]
T2WI					
Slice thickness (mm)	–	–	–	–	4.0 [3.5, 4.0]
Slice spacing (mm)	–	–	–	–	4.0 [4.0, 4.0]
Repetition time (ms)	–	–	–	–	3,340 [3,000, 4,000]
Echo time (ms)	–	–	–	–	95 [87, 106]
Field of view (mm)	–	–	–	–	240 [200, 240]
Flip angle (°)	–	–	–	–	90 [90, 90]

The quantitative variables are presented as the median [Q1, Q3] for the non-normalized data. The categorical variables are presented as n (%). MR, magnetic resonance; DWI, diffusion-weighted imaging; ADC, apparent diffusion coefficient; T2WI, T2-weighted imaging.

the systematic biopsy or the cognitive-targeted biopsy were categorized based on the International Society of Urological Pathology (ISUP) grade, while all other sextants were graded based on the systematic biopsy results. Some patients in this study subsequently underwent transurethral prostatectomy or radical prostatectomy. When obtaining the final comprehensive pathological diagnosis, priority was given to the pathology results from radical prostatectomy over those from transurethral prostatectomy, with the combined biopsy results ranked last. The ISUP grading system was used to classify patients as non-csPCa (the no-cancer or ISUP group 1) or csPCa (ISUP group 2). Patients with negative PCa biopsies were classified as negative if they had an additional follow-up period of at least one year and no evidence of underlying PCa.

All the cases included in this study underwent a retrospective review and were delineated based on the comprehensive pathological diagnosis results of each lesion by two uro-radiologists (Zhaonan Sun and Xiaoying Wang, who had 5 and 30 years of experience in prostate MRI interpretation, respectively) in consensus. The open-source software ITK-SNAP (version 3.8.0; available at http://www.itksnap.org) was used to annotate the tumor foci.

Sextant areas were automatically generated using the prostate sextant location model (Appendix 2). Subsequently, these sextant areas were classified as cancerous or non-cancerous based on the presence or absence of cancer within each sextant. For instance, if the ground-truth segmentation overlapped with a sextant, it was categorized as a cancerous area.

AI model development

In this study, we employed an end-to-end AI model comprising the following four components: (I) MRI sequence classification; (II) prostate gland segmentation and measurement (17); (III) prostate zonal anatomy segmentation; and (IV) csPCa foci segmentation and measurement (18). These models were executed automatically in sequence. Based on the identified suspicious areas, various parameters, including the number of suspicious lesions, three-dimensional (3D) diameter, volume, sextant location, and average ADC value, were calculated and automatically incorporated into the PI-RADS structured report (19). The primary focus of this study was the csPCa foci segmentation model, which provided a binary output. Detailed information about the other AI models can be found in Appendix 2.

Preprocessing

The DWI, ADC maps, and T2WI were registered through rigid transformation using the coordinate information stored in the Digital Imaging and Communications in Medicine (DICOM) image headers. For the automatic pre-segmentation of the prostate gland region, a model (17) previously developed at our institution was employed; detailed information about the latest updated model can be found in Appendix 2. All the prostate areas were standardized and cropped to a size of 64×64×64 (x, y, z), with pixel intensity normalized to the range of [0, 1]. To augment the training set, random rotations (within a range of 10 degrees), random noise, and parallel translations within the range of [(–0.1, 0.1); (–0.1, 0.1)] pixels were applied.

Deep learning

The collected data were divided randomly into the following three data sets: a training data set (comprising 80% of the data); a validation data set (comprising 10% of the data); and an internal test data set (comprising10% of the data). The AI model used a combination of DWI and ADC maps as input for the PCa segmentation (18). A cascade 3D U-Net framework (20) was employed for the segmentation process. The training processes were conducted using the NVIDIA Tesla P100 16G GPU. The algorithm was implemented using Python 3.6, PyTorch 0.4.1, OpenCV 3.4.0.12, Numpy 1.16.2, and SimpleITK 1.2.0. Taking into consideration aspects such as computational efficiency, generalization, training stability, and data set size, a batch size of 60 was employed during training, and 1,000 epochs were conducted to train the networks. The optimization algorithm employed was the Adaptive Moment Estimation (Adam) gradient descent, which had a learning rate of 0.0001, and the binary cross-entropy loss function was minimized.

Post-processing

Based on the current capabilities of mpMRI, csPCa volumes greater than or equal to 0.5 cc can feasibly be detected (7). However, the current use of AI models for csPCa diagnosis is often associated with a high rate of false positive (FP) results (21-23). To address the potential impact of the very small, predicted tumor foci, an output threshold of 0.5 cc was adopted. It is important to emphasize that this post-processing step was only performed on the test and external validation data sets and was not performed on the training and validation sets. The rationale for adopting this approach during the training phase was to enhance the training efficiency and enable the model to learn from a diverse range of cancerous voxels in which a volume threshold was deliberately omitted.

Evaluation of the AI prediction results

At the lesion level, each connected domain predicted by the AI model was regarded as a predicted lesion. For each patient, we focused on studying the four largest lesions. To assess the performance of the AI model, we compared the predicted lesions with the reference standard. The spatial overlap between the predicted lesion and the reference standard was measured using the mean dice similarity coefficient (DSC). If a lesion overlapped with the reference standard, it was classified as a true positive (TP) lesion. Conversely, if an AI-predicted lesion did not overlap with the reference standard, it was classified as an FP lesion. Similarly, if a reference lesion did not overlap with any AI-predicted lesion, it was classified as a false negative (FN) lesion.

At the sextant level, the right-upper, right-middle, right-lower, left-upper, left-middle, and left-lower regions of the prostate gland were independently examined. The presence of lesion overlap within a sextant was defined as a positive finding. Thus, if a sextant overlapped with both the reference lesion and AI-predicted lesion, it was classified as a TP sextant. If a sextant only overlapped with the reference lesion and not the AI-predicted lesion, it was classified as a FN sextant. Conversely, if a sextant only overlapped with the AI-predicted lesion and not the reference lesion, it was classified as a FP sextant. Finally, if a sextant did not overlap with either the reference lesion or the AI-predicted lesion, it was classified as a true negative (TN) sextant.

At the patient level, a patient with any TP sextant was classified as a TP case. A patient with all TN sextants was classified as a TN case. A patient with TN and FP sextants, without any TP sextants, was classified as a FP case. Similarly, a patient with TN and FN sextants without any TP sextants was classified as a FN case.

To assess the ability of the AI model to detect lesions, we calculated the sensitivity at the lesion level. At the sextant level, we evaluated sensitivity, specificity, and accuracy to assess the ability of the AI model to localize lesions, comparing these metrics across the six sextants. Additionally, sensitivity, specificity, and accuracy were calculated at the patient level to evaluate the ability of the AI model to identify biopsy candidates. The accuracy of the AI model at the patient level was also examined; the tumor number, volume, and average ADC value were considered contributing factors.

Statistical analysis

The statistical analyses were conducted using the R software (version 4.2.0). For the normalized data, the quantitative variables are presented as the mean (standard deviation). For the non-normalized data, the quantitative variables are presented as the median (Q1, Q3). The categorical variables are reported as numbers (percentages). The Wilcoxon test was used to compare the quantitative variables, and the chi-square test was used to compare the categorical variables. Segmentation metrics, including the DSC, Jaccard Coefficient (JACARD), volume similarity (VS), Hausdorff distance (HD), and average distance (AD), were calculated and compared by an analysis of variance.

A free-response receiver operating characteristic (FROC) analysis was conducted to assess the lesion detection accuracy of the AI model. A Bland-Altman analysis was used to evaluate the measured values of PCa. All the statistical tests were two-tailed, and a significance level of 5% was applied.

Results

Clinical characteristics

The model development data set comprised 2,105 patients, of whom 1,368 had a confirmed diagnosis of PCa and 737 did not have PCa. For the external validation, the data set comprised 557 patients, of whom 245 had a confirmed diagnosis of PCa and 312 did not have PCa. The clinical characteristics of these patients are presented in *Table 2*.

Segmentation metrics and quantitative evaluation

In the test set of the model development data set, the model exhibited proficient performance in lesion segmentation, achieving a DSC value of 0.80 (0.61, 0.86). The DSC, JACRD, VS, HD, and AD values of the different data sets are displayed in *Table 3*. Further, *Table 4* and *Figure 2* present the results of the Bland-Altman analysis for the measured values of PCa foci, including the right-left (RL) diameter, anteroposterior (AP) diameter, superoinferior (SI) diameter, volume, and ADC value. The Bland-Altman analysis used radiologist-annotated results as the reference

Table 2 Clinical characteristics of the patients in the study

Parameters	Model development data set		External validation data set		P value
	Negative (N=737)	Positive (N=1,368)	Negative (N=312)	Positive (N=245)	
Age (years)	64.5 (60, 69)	69.0 (63.0, 77.0)	65.5 (60.0, 71.0)	70.0 (64.0, 76.0)	0.23
tPSA value	7.5 (6.2, 8.6)	16.2 (10.6, 32.5)	8.22 (5.07, 18.8)	12.1 (6.87, 33.2)	<0.001
Final ISUP group					<0.001
Biopsy-NoPCa	698 (94.7)	0 (0.0)	299 (95.8)	0 (0.0)	
1	39 (5.3)	0 (0.0)	13 (4.2)	0 (0.0)	
2	0 (0.0)	628 (45.9)	0 (0.0)	101 (41.2)	
3	0 (0.0)	208 (15.2)	0 (0.0)	40 (16.3)	
4	0 (0.0)	164 (12.0)	0 (0.0)	37 (15.1)	
5	0 (0.0)	232 (17.0)	0 (0.0)	46 (18.8)	
Positive	0 (0.0)	136 (9.9)	0 (0.0)	21 (8.6)	
ISUP source					<0.001
Biopsy	701 (95.1)	808 (59.1)	307 (98.4)	137 (55.9)	
TURP	36 (48.9)	54 (3.9)	5 (1.6)	9 (3.7)	
RP	0 (0.0)	506 (37.0)	0 (0.0)	99 (40.4)	
Per-patient PI-RADS category					<0.001
PI-RADS 1–2	492 (66.8)	112 (8.2)	232 (74.3)	17 (6.9)	
PI-RADS 3	194 (26.3)	364 (26.6)	67 (21.5)	61 (24.9)	
PI-RADS 4	26 (3.5)	263 (19.2)	9 (2.9)	43 (17.6)	
PI-RADS 5	25 (3.4)	629 (46.0)	4 (1.3)	124 (50.6)	
Number of lesions					<0.001
0	737 (100.0)	0 (0.0)	312 (100.0)	0 (0.0)	
1	0 (0.0)	596 (43.6)	0 (0.0)	140 (57.1)	
2	0 (0.0)	340 (24.9)	0 (0.0)	47 (19.2)	
3	0 (0.0)	168 (12.3)	0 (0.0)	23 (9.4)	
4	0 (0.0)	98 (7.2)	0 (0.0)	35 (14.3)	
>4	0 (0.0)	166 (12.1)	0 (0.0)	0 (0.0)	
Hospital					<0.001
Hospital 1	641 (87.0)	930 (68.0)	263 (84.3)	69 (28.2)	
Hospital 2	74 (10.0)	229 (16.7)	21 (6.7)	67 (27.3)	
Hospital 3	9 (1.2)	186 (13.6)	28 (9.0)	109 (44.5)	
Hospital 4	13 (1.8)	23 (1.7)	0 (0.0)	0 (0.0)	

The quantitative variables are presented as the median (Q1, Q3) for the non-normalized data. The categorical variables are presented as n (%). tPSA, total prostate-specific antigen; ISUP, Pathological International Society of Urological Pathology; TURP, transurethral resection of the prostate; RP, radical prostatectomy; PI-RADS, Prostate Imaging Reporting and Data System.

Table 3 Segmentation metrics of the model in the model development data set

Parameters	Training (N=1,681)	Validation (N=212)	Test (N=212)	P value
DSC	0.93 (0.91, 0.95)	0.81 (0.70, 0.87)	0.80 (0.61, 0.86)	<0.001
JACARD	0.87 (0.83, 0.90)	0.68 (0.53, 0.77)	0.67 (0.44, 0.76)	<0.001
VS	0.99 (0.98, 1.00)	0.90 (0.80, 0.95)	0.87 (0.73, 0.95)	<0.001
HD	5.63 (2.50, 12.4)	9.94 (6.76, 18.0)	12.5 (7.91, 20.4)	<0.001
AD	0.08 (0.05, 0.18)	0.40 (0.18, 1.11)	0.49 (0.23, 1.67)	<0.001

The quantitative variables are presented as the median (Q1, Q3) for the non-normalized data. DSC, dice similarity coefficient; JACARD, Jaccard Coefficient; VS, volumetric similarity; HD, Hausdorff distance; AD, average distance.

Table 4 Bland-Altman analysis of the measured values of prostate cancer

Parameters	RL diameter	AP diameter	SI diameter	Volume	ADC value
Bias	–0.464	–0.431	–0.526	–1.306	–0.464
BiasUpperCI	–0.394	–0.37	–0.464	–1.053	–0.394
BiasLowerCI	–0.533	–0.492	–0.588	–1.56	–0.533
BiasStdDev	1.63	1.425	1.453	5.926	1.63
BiasSEM	0.036	0.031	0.032	0.129	0.036
LOA_SEM	0.061	0.053	0.054	0.221	0.061
UpperLOA	2.73	2.361	2.322	10.309	2.73
UpperLOA_upperCI	2.85	2.465	2.428	10.742	2.85
UpperLOA_lowerCI	2.611	2.257	2.216	9.876	2.611
LowerLOA	–3.657	–3.223	–3.374	–12.922	–3.657
LowerLOA_upperCI	–3.538	–3.119	–3.267	–12.489	–3.538
LowerLOA_lowerCI	–3.777	–3.327	–3.48	–13.355	–3.777
Regression.fixed.slope	0.15	0.12	0.12	–0.02	0.15
Regression.fixed.intercept	–0.83	–0.71	–0.82	–1.1	–0.83

RL, right-left; AP, anteroposterior; SI, superoinferior; ADC, apparent diffusion coefficient; CI, confidence interval; LOA, limits of agreement; SEM, standard error of mean.

standard to assess the consistency of the AI measurements with these reference standards. These results demonstrated a high level of agreement in the predicted volumes, 3D diameters, and ADC values of the PCa lesions compared to the reference standard. The differences observed between the majority of the AI results and the reference standard were minimal and fell within the 95% limit of agreement.

Lesion-level performance in the external validation data set

In total, 434 locations of PCa lesions were annotated by the radiologists and used as the reference standard for evaluating the AI model. Additionally, the AI model identified an additional 96 locations, resulting in a total of 530 locations for analysis. The AI model had a sensitivity of 0.654 and a positive predictive value of 0.747 at the lesion level. The AI-predicted results at the lesion level are depicted in line 3 of *Figure 3*. The AI model correctly identified 284 cancer foci in 231 patients, yielding a lesion-level sensitivity of 65.4%. The prediction results at the lesion level on the MR images are illustrated in *Figure 4*. The results of the FROC analysis of the lesion detection ability of the AI model are shown in *Figure 5*. By illustrating the trade-off between the sensitivity and FP results per patient, this curve visually presents the

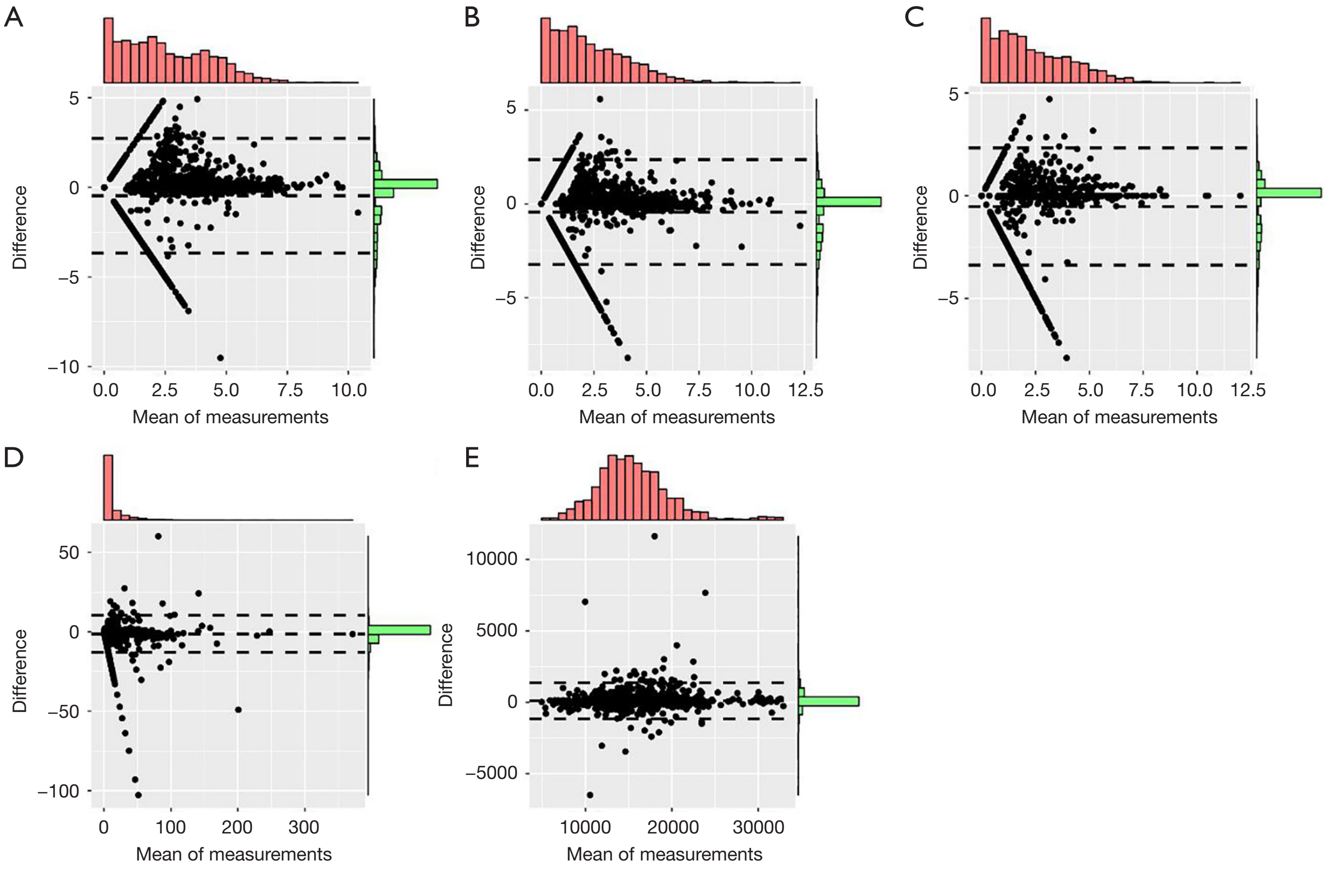


Figure 2 Bland-Altman analysis of the values of the RL diameter (A), AP diameter (B), SI diameter (C), volume (D), and ADC value (E) of the manual label and the predicted label of prostate cancer. RL, right-left; AP, anteroposterior; SI, superoinferior; ADC, apparent diffusion coefficient.

performance characteristics of the algorithm.

Sextant-level performance in the external validation data set

Of the 3,342 sextant areas evaluated in the 557 patients, the AI model diagnosed 745 TP, 136 FN, 2,175 TN, and 286 FP sextants. As a result, the overall sensitivity, specificity, and accuracy of the AI model at the sextant level were 0.846, 0.884, and 0.874, respectively (*Table 5*). In relation to each specific type of sextant (i.e., right-superior, right-middle, right-inferior, left-middle, and left-inferior), the sensitivity ranged from 0.772 to 0.902 (P=0.01), the specificity ranged from 0.851 to 0.915 (P=0.02), and the accuracy ranged from 0.858 to 0.894 (P=0.34) (*Figure 6*).

Patient-level performance in the external validation data set

At the patient level, the AI model correctly detected that 231 of 245 patients had csPCa, and 242 of 312 patients did not have csPCa (*Figure 3*). Thus, the sensitivity, specificity, and accuracy for the detection of csPCa patients were 0.943, 0.776, and 0.849, respectively (*Table 5*).

Table 6 and *Figure 7* present the factors that influenced the accuracy of the AI model at the patient level. The AI model accurately predicted the presence of csPCa in a higher proportion of csPCa patients (231/245, 0.943) than non-csPCa patients (242/312, 0.776) (P<0.001) (*Figure 7A*). The lesion number and tumor volume (24) were greater in the correctly diagnosed patients than the incorrectly diagnosed patients [0 (0, 1) *vs.* 0 (0, 0); 0.00 (0.00, 0.00) *vs.* 0.00 (0.00, 2.36) cm^3, both P<0.001] (*Figure 7B,7C*). Additionally, among the positive patients, those with lower average ADC values were more accurately diagnosed than those with higher average ADC values [0.750 (0.643, 0.866) ×10^{-3} mm^2/s *vs.* 0.884 (0.765, 0.966)×10^{-3} mm^2/s, P=0.011] (*Figure 7D*). The statistical analysis indicated that

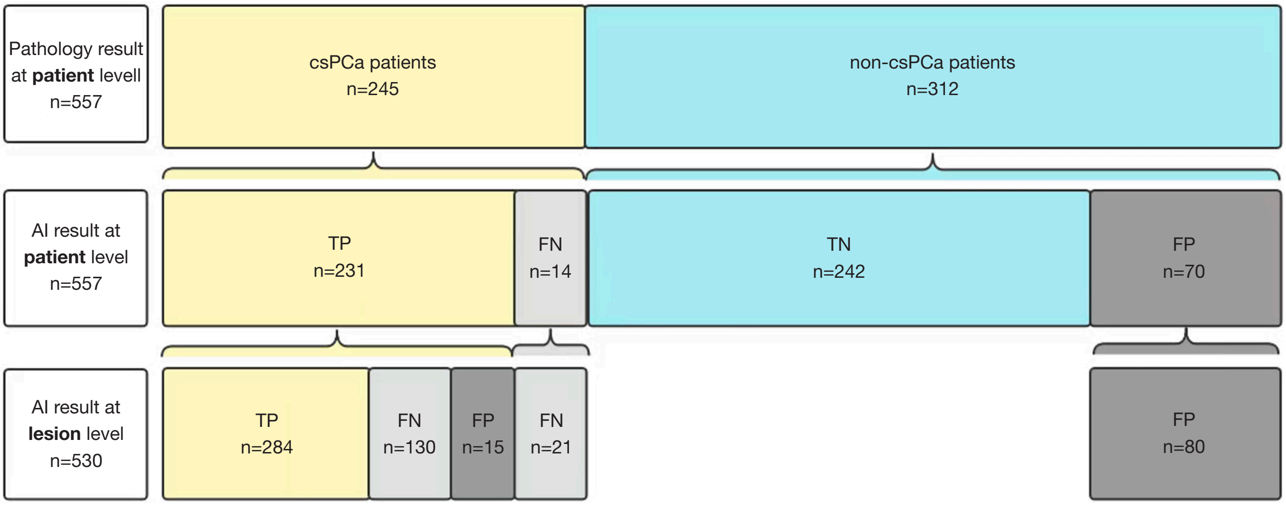


Figure 3 AI-predicted results at the lesion level. csPCa, clinically significant prostate cancer; TP, true positive; FN, false negative; TN, true negative; FP, false positive; AI, artificial intelligence.

with the exception of the hospital variable (P=0.04), there were no significant differences in the accuracy of the AI model among the different magnetic fields, vendors of MR scanners, or PCa locations (all P>0.05).

Discussion

Before any AI model can be used to assist radiologists to manage the increase in imaging volumes, it must undergo external evaluation. Thus, this study sought to demonstrate that an AI model could accurately detect and localize suspicious findings on prostate MRI. A diverse multicenter data set was employed for the external validation of our PCa AI model. The correlation between model efficacy and disease characteristics was investigated to promote the transparent validation and facilitate the clinical translation of our findings.

The meta-analysis (25) showed that the model had a pooled sensitivity of 0.89 and a specificity of 0.73 in diagnosing patients with csPCa using PI-RADS Version 2. In relation to sextant location, a previous study reported that radiologists had a sensitivity of 0.55 to 0.67 and a specificity of 0.68 to 0.80 (26). In our study, the sensitivity and specificity of the AI model were 0.943 and 0.776 at the patient level, and 0.846 and 0.884 at the sextant level, respectively. It is crucial to compare the results of the AI model with those of the radiologists. The AI model provides a binary classification of positive and negative results. Radiologists use the PI-RADS scoring system to assess the likelihood of PCa during image interpretation. Since the PI-RADS scores provided by radiologists are subjective evaluations, there may be inconsistencies between different observers and even within the same observer. We have previously conducted research on this very issue, comparing the diagnostic accuracy, diagnostic confidence, and diagnostic time of radiologists when using AI versus not using AI. As the results of this previous study have already been published (18), we elected not to include this content in the current manuscript to avoid duplicate publication. It is crucial to note that our AI model performed well at both the patient and sextant levels, effectively fulfilling the two main objectives of identifying biopsy candidates and providing accurate guidance for targeted biopsies. Our AI model had an average sensitivity of 94.3% in detecting index lesions but only 65.4% in detecting all lesions; thus, the lesion-level sensitivity of the model needs to be further enhanced. The use of PIRADS V2 enabled radiologists to achieve a sensitivity of 91% in identifying index PCa lesions and a sensitivity of 63% in detecting all lesions (27). The sensitivity of the AI model at the lesion level is comparatively lower; however, it remains consistent with the diagnostic performance currently achieved by medical professionals using PIRADS V2. The per-sextant sensitivity was much higher than the per-lesion sensitivity. The observed difference between the per-sextant sensitivity and per-lesion sensitivity can be attributed to the inherent complexity of lesion detection and the manner in which lesions are distributed within the sextants. The lower per-

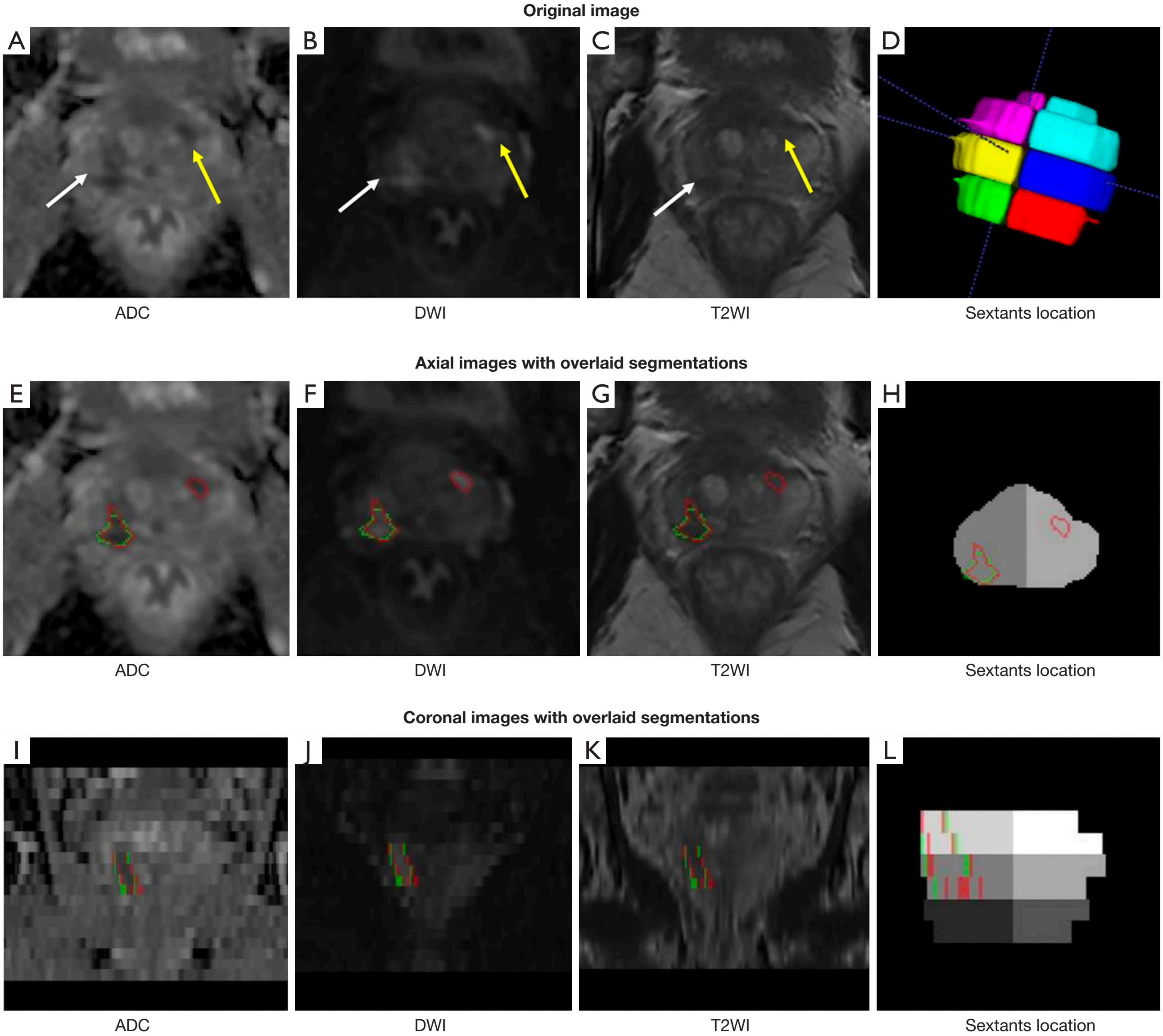


Figure 4 An example of algorithm segmentation and location. ADC map (A), DWI (B), and T2WI (C) showed two prostate cancer lesions in the right peripheral zone (white arrow) and the left transition zone (yellow arrow). (D) Sextant location image. (E-L) Axial and coronal images with overlaid segmentations (the red polygon outlines the ground truth of clinically significant prostate cancer, and the green polygon outlines the predicted area by the AI algorithm). The lesion located in the right peripheral zone is a TP lesion and occupies the right-upper and right-middle sextants; thus, the two sextants are TP sextants. The predicted area located in the left transitional zone is a FN lesion; thus, the left-middle sextant is an FN sextant, and the remaining sextants are TN zones. At the patient level, the patient was diagnosed with TP. ADC, apparent diffusion coefficient; DWI, diffusion-weighted imaging; T2WI, T2-weighted imaging; AI, artificial intelligence; TP, true positive; FN, false negative; TN, true negative.

lesion sensitivity arises from the potential occurrence of missed diagnoses at the lesion level, which led to a decrease in the model's overall sensitivity for detecting individual lesions. Lesions with larger volumes tend to have a lower rate of missed diagnosis and higher detection accuracy, as their size facilitates easier identification. Consequently, larger lesions often span multiple sextants, resulting in the correct diagnosis of more sextants.

Recently, numerous studies have presented compelling evidence in this domain (28). The commercially available software provided at https://fuse-ai.de/prostatecarcinoma-ai has a comparable level of reliability to that of radiologists in detecting carcinomas, with a patient-level sensitivity of 0.86, which, in contrast, is inferior to our model's sensitivity value of 0.94. Several researchers (18,26,29) have sought to evaluate the efficacy of the U-Net network for PCa detection at various levels, including the lesion, sextant, and patient levels. However, these studies did not examine the effect of different cancer foci on the diagnostic efficiency of the models. Zhong *et al.* (30) employed a training set (comprising data from 110 prostate patients) and a test set (comprising data from 30 patients) to assess the effectiveness of their model, and found that the transfer learning-based model exhibited superior efficacy in diagnosing PCa than the deep-learning model without transfer learning. However, it should be noted that their analysis was limited to the lesion level. Further, Cao *et al.* (31) trained a FocalNet model to autonomously identify prostate lesions. Their model achieved a confidence score that yielded outcomes on par with those achieved by experienced radiologists but was specifically limited to high-sensitivity or high-specificity scenarios. These models were trained and validated on private data sets, which are often homogeneous and lack external validation with data from other institutions and scanners produced by different manufacturers. Table S8 provides an overview of the key differences between our study and recent previous studies using deep learning in prostate MRI.

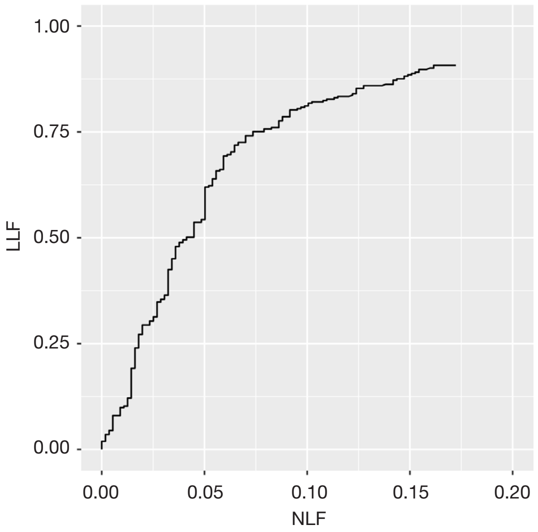


Figure 5 AI-predicted results at the lesion level as shown by the FROC curve. The y-axis represents the LLF, which corresponds to the sensitivity of lesion detection. The x-axis represents the NLF, indicating the average number of FPs per case. LLF, lesion localization fraction; NLF, negative localization fraction; AI, artificial intelligence; FROC, free-response receiver operating characteristic; FP, false positive.

As we enter a new phase in the application of AI to prostate mpMRI, our goal is to prioritize transparent validation and clinical translation. Previous studies have primarily focused on reporting the performance metrics of AI models, such as sensitivity, specificity, and areas under the curve, but have often failed to offer interpretations of these values. Such interpretations are crucial in enhancing radiologists' confidence in the results and facilitating the clinical implementation of AI models. In our research, we examined the effects of inherent characteristics of PCa on the performance of the AI model. In our analysis, we observed a substantial discrepancy in the AI-predicted accuracy between patients with csPCa and those without csPCa. Specifically, our model's accuracy for csPCa patients

Table 5 Evaluation metrics for the detection of prostate cancer at different levels

Parameters	Patient level	Sextant level	Lesion level
Accuracy	0.849 (0.849, 0.850)	0.874 (0.874, 0.874)	–
Sensitivity	0.943 (0.914, 0.972)	0.846 (0.822, 0.869)	0.654 (0.608, 0.699)
Specificity	0.776 (0.729, 0.822)	0.884 (0.871, 0.896)	*
Positive predictive value	0.767 (0.720, 0.815)	0.723 (0.695, 0.750)	0.747 (0.701, 0.790)
Negative predictive value	0.945 (0.917, 0.973)	0.941 (0.932, 0.951)	*

*, due to the absence of TN findings for lesion detection, calculating the specificity and negative predictive value was infeasible at the lesion level. The data presented in the table represent various metrics with their corresponding values and their respective 95% confidence intervals. TN, true negative.

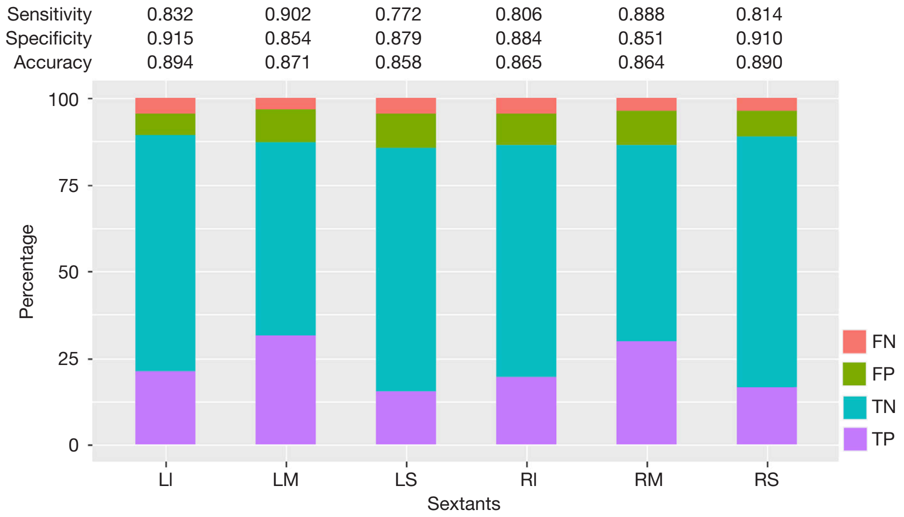


Figure 6 AI-predicted results at the sextant level. LI, left-inferior; LM, left-middle; LS, left-superior; RI, right-inferior; RM, right-middle; RS, right-superior; FN, false negative; FP, false positive; TN, true negative; TP, true positive; AI, artificial intelligence.

Table 6 Factors affecting AI accuracy at the patient level

Parameters	Overall (n=557)	Error (n=84)	Right (n=473)	P value
Patient diagnosis				<0.001
Negative	312 (56.0)	70 (83.3)	242 (51.2)	
Positive	245 (44.0)	14 (16.7)	231 (48.8)	
Lesion number	0.00 (0.00, 1.00)	0.00 (0.00, 0.00)	0.00 (0.00, 1.00)	<0.001
Tumor volume (cm^3)	0.00 (0.00, 1.87)	0.00 (0.00, 0.00)	0.00 (0.00, 23.55)	<0.001
Average ADC value (10^{-3} mm^2/s)	0.753 (0.643, 0.875)	0.884 (0.765, 0.966)	0.750 (0.643, 0.866)	0.011
Hospital				0.042
Hospital 1	332 (59.6)	52 (61.9)	280 (59.2)	
Hospital 2	88 (15.8)	19 (22.6)	69 (14.6)	
Hospital 3	137 (24.6)	13 (15.5)	124 (26.2)	
Magnetic field				0.999
1.5 T	27 (4.8)	4 (4.8)	23 (4.9)	
3.0 T	530 (95.2)	80 (95.2)	450 (95.1)	
Vendor				0.364
GE	288 (51.7)	50 (59.5)	238 (50.3)	
PHILIPS	62 (11.1)	8 (9.5)	54 (11.4)	
SIEMENS	201 (36.1)	26 (31.0)	175 (37.0)	
UIH	6 (1.1)	0 (0.0)	6 (1.3)	
PCa location				0.512
PZ	32 (5.7)	3 (21.4)	29 (12.6)	
TZ	188 (33.8)	2 (14.3)	23 (10.0)	
PZ + TZ	25 (4.5)	9 (64.3)	179 (77.4)	

The quantitative variables are presented as the median (Q1, Q3) for the non-normalized data. The categorical variables are presented as n (%). AI, artificial intelligence; ADC, apparent diffusion coefficient; PCa, prostate cancer; PZ, peripheral zone; TZ, transition zone.

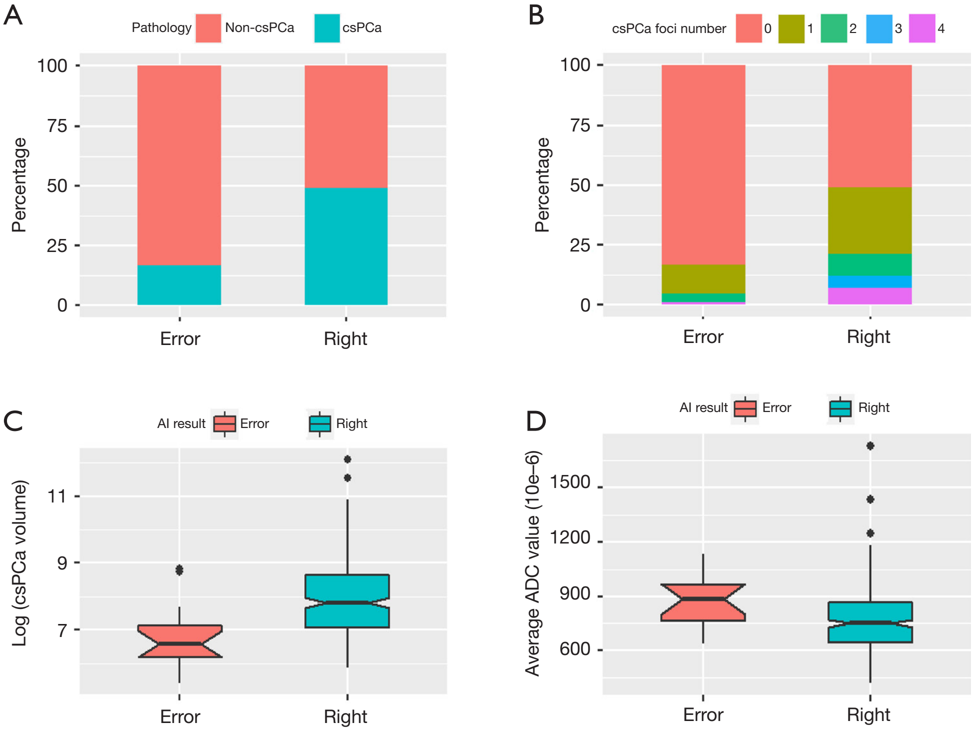


Figure 7 Factors affecting the accuracy of the AI model at the patient level. (A) The AI model was significantly more accurate in predicting csPCa patients (231/245, 0.943) than non-csPCa patients (242/312, 0.776) (P<0.001). (B,C) The lesion number and tumor volume were greater in the correctly diagnosed patients than the incorrectly diagnosed [0 (0, 1) *vs.* 0 (0, 0); 0.00 (0.00, 0.00) *vs.* 0.00 (0.00, 2.36) cm^3, both P<0.001]. (D) Among the positive patients, those with lower average ADC values had a higher rate of correct diagnosis than those with higher average ADC values [0.750 (0.643, 0.866) *vs.* 0.884 (0.765, 0.966)×10^{-3} mm^2/s, P=0.011]. csPCa, clinically significant prostate cancer; AI, artificial intelligence; ADC, apparent diffusion coefficient.

was significantly higher than that for non-csPCa patients. Thus, the AI model's enhanced ability to detect csPCa could be leveraged to mitigate FN results and unnecessary biopsies. However, it is important to emphasize that further validation at a higher level is required to establish the association between the AI models' diagnostic results and their clinical value. Thus, future studies with prospective cohorts of patients with long-term follow-up periods need to be conducted to validate these results. Other findings indicate that patients with a greater number of lesions, a larger tumor volume, and lower average ADC values tend to receive a more accurate diagnosis from the AI model. These trends align with the performance of radiologists. Using the PI-RADS Version 2 system, radiologists detected PCa in 12–33%, 22–70%, and 72–91% of lesions with PI-RADS scores of 3, 4, and 5, respectively (32,33).

Recent studies have shown the favorable performance of architectures that integrate two cascaded networks: a first model that segments or performs a crop around the prostate area, and a second binary model that segments the PCa lesion. In the study conducted by Yang *et al.* (34), convolutional neural networks (CNNs) were employed to crop square regions encompassing the entirety of the prostate area. Subsequently, their proposed co-trained CNNs were fed with pairs of aligned ADC and T2WI squares. Expanding on this work, Wang *et al.* (35) and Zhu *et al.* (29) adapted the workflow to create an end-to-end trainable deep neural network comprising two sub-networks. The first sub-network was responsible for prostate detection and ADC-T2WI registration, while the second sub-network was a dual-path multimodal CNN that generated a classification score for csPCa and non-csPCa. In the study conducted by Saha *et al.* (36), a preparatory network called anisotropic 3D U-Net was employed to generate deterministic or probabilistic zonal segmentation maps. These maps were subsequently fused in a second CNN that produced a probability map for csPCa. Despite the additional weights and more intensive training process introduced by the two-step workflow, it has proven to be an effective method. Certain attention mechanisms have demonstrated improved segmentation performance in prostate imaging tasks. Rundo *et al.* (37) employed a "squeeze & excitation" (SE) module for prostate zonal segmentation on diverse MRI data sets, and reported a 1.4–2.9% increase in the DSC compared to the U-Net baseline, specifically for peripheral zone segmentation, when evaluating their multi-source model across different data sets. Zhang *et al.* (38) introduced combined channel-attention (inspired by the SE module) and position-attention layers, and achieved a DSC that surpassed their baseline by 1.8% for prostate lesion segmentation on T2WI. Additionally, Saha *et al.* (36) improved the sensitivity by 4.34% by incorporating the aforementioned SE model and grid-attention gates mechanisms into a 3D UNet++ (39) backbone architecture for binary PCa segmentation.

At the lesion level, 530 lesions were analyzed, of which 53.6% (284/530) lesions were correctly diagnosed (TP lesions) by the AI model, 28.3% (150/530) lesions were missed (FN lesions), and 18.1% (96/530) additional lesions were over detected (FP lesions). Despite some misdiagnosed lesions, the overall accuracy of the AI model at the lesion level was similar to the results of the imaging interpretation by radiologists. In previous studies, the FP rates of radiologists for lesions with PI-RADS scores ≥3 have varied from 32% to 50% (40). Additionally, the FN rate of radiologists may reach 12% for high-grade cancers during screening, and may be as high as 34% in men undergoing radical prostatectomy (1,41). There might be other reasons for the high prevalence of FN and FP lesions in our study. In terms of the FN lesions, our study methodology may have falsely increased the FN rate. One limitation of our study is that when multiple lesions were detected by the AI model in a case, only the four largest lesions detected by the AI model were studied. Because PI-RADS recommends reporting no more than four lesions in structured reporting, we followed this rule to simulate a real clinical scenario. Thus, if AI detects more than four TP lesions in a patient, the extra lesions would be considered FN lesions. In terms of the FP lesions, the imperfect match between the reference standard annotation and the real pathology might have increased the FP rate. In this study, the diagnoses of 79.7% of the patients were pathologically proven by image-guided biopsy. In comparison to radical prostatectomy specimens, image-guided biopsy pathology results may miss some lesions (16). When radiologists outline the reference area of csPCa foci on MR images according to the biopsy pathology results, they may miss lesions or underestimate the extent of the lesions (42). Conversely, if the missed lesions were detected by the AI model and considered FPs, the lesions may be correctly detected.

In addition to the above-mentioned methodological limitations, our study had a number of other limitations. First, the retrospective study design and unbalanced data prevented a robust assessment of the clinical impact of the model. Ideally, AI models should be deployed in prospective randomized studies to test their performance. Second, if there were mismatches between the MR images and the pathology data, the data were not analyzed. The reference standard was based on image-guided biopsy, and the result obtained by using whole-mount step section pathology as the reference standard was more credible than image-guided biopsy pathology. Third, poor image quality data were excluded; however, we intend to address this in future clinical applications. Finally, there is a lack of research on radiologists interacting with AI models. In the future, we plan to invite radiologists with varying levels of experience to interpret mpMRI reports with the assistance of AI models to determine whether AI models add value in real clinical settings.

Conclusions

In the external validation, the AI model achieved acceptable accuracy for the detection and localization of csPCa at the patient level and the sextant level. However, the sensitivity at the lesion level should be improved for future clinical application.

Acknowledgments

Funding: This work was supported by the Capital's Fund for Health Improvement and Research (No. 2020-2-40710).

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://qims.amegroups.com/article/view/10.21037/qims-23-791/rc

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-791/coif). P.W. and Xiangpeng Wang were employees of Beijing Smart Tree Medical Technology Co., Ltd., and had no financial or other conflicts with respect to this study. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the institutional review board (IRB number: 2021060), and the requirement of individual consent for this retrospective analysis was waived.

References

1. Ahmed HU, El-Shater Bosaily A, Brown LC, Gabe R, Kaplan R, Parmar MK, Collaco-Moraes Y, Ward K, Hindley RG, Freeman A, Kirkham AP, Oldroyd R, Parker C, Emberton M; PROMIS study group. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. Lancet 2017;389:815-22.
2. Ahdoot M, Wilbur AR, Reese SE, Lebastchi AH, Mehralivand S, Gomella PT, Bloom J, Gurram S, Siddiqui M, Pinsky P, Parnes H, Linehan WM, Merino M, Choyke PL, Shih JH, Turkbey B, Wood BJ, Pinto PA. MRI-Targeted, Systematic, and Combined Biopsy for Prostate Cancer Diagnosis. N Engl J Med 2020;382:917-28.
3. Kasivisvanathan V, Rannikko AS, Borghi M, Panebianco V, Mynderse LA, Vaarala MH, et al. MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis. N Engl J Med 2018;378:1767-77.
4. van der Leest M, Cornel E, Israël B, Hendriks R, Padhani AR, Hoogenboom M, Zamecnik P, Bakker D, Setiasti AY, Veltman J, van den Hout H, van der Lelij H, van Oort I, Klaver S, Debruyne F, Sedelaar M, Hannink G, Rovers M, Hulsbergen-van de Kaa C, Barentsz JO. Head-to-head Comparison of Transrectal Ultrasound-guided Prostate Biopsy Versus Multiparametric Prostate Resonance Imaging with Subsequent Magnetic Resonance-guided Biopsy in Biopsy-naïve Men with Elevated Prostate-specific Antigen: A Large Prospective Multicenter Clinical Study. Eur Urol 2019;75:570-8.
5. Barentsz JO, Richenberg J, Clements R, Choyke P, Verma S, Villeirs G, Rouviere O, Logager V, Fütterer JJ; European Society of Urogenital Radiology. ESUR prostate MR guidelines 2012. Eur Radiol 2012;22:746-57.
6. Vargas HA, Hötker AM, Goldman DA, Moskowitz CS, Gondo T, Matsumoto K, Ehdaie B, Woo S, Fine SW, Reuter VE, Sala E, Hricak H. Updated prostate imaging reporting and data system (PIRADS v2) recommendations for the detection of clinically significant prostate cancer using multiparametric MRI: critical evaluation using whole-mount pathology as standard of reference. Eur Radiol 2016;26:1606-12.
7. Weinreb JC, Barentsz JO, Choyke PL, Cornud F, Haider MA, Macura KJ, Margolis D, Schnall MD, Shtern F, Tempany CM, Thoeny HC, Verma S. PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. Eur Urol 2016;69:16-40.
8. Turkbey B, Rosenkrantz AB, Haider MA, Padhani AR, Villeirs G, Macura KJ, Tempany CM, Choyke PL, Cornud F, Margolis DJ, Thoeny HC, Verma S, Barentsz J, Weinreb JC. Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2. Eur Urol 2019;76:340-51.
9. Smith CP, Harmon SA, Barrett T, Bittencourt LK, Law YM, Shebel H, An JY, Czarniecki M, Mehralivand S, Coskun M, Wood BJ, Pinto PA, Shih JH, Choyke PL, Turkbey B. Intra- and interreader reproducibility of PI-RADSv2: A multireader study. J Magn Reson Imaging 2019;49:1694-703.
10. Girometti R, Giannarini G, Greco F, Isola M, Cereser L, Como G, Sioletic S, Pizzolitto S, Crestani A, Ficarra V, Zuiani C. Interreader agreement of PI-RADS v. 2 in assessing prostate cancer with multiparametric MRI: A study using whole-mount histology as the standard of reference. J Magn Reson Imaging 2019;49:546-55.
11. Sonn GA, Fan RE, Ghanouni P, Wang NN, Brooks JD, Loening AM, Daniel BL, To'o KJ, Thong AE, Leppert JT. Prostate Magnetic Resonance Imaging Interpretation Varies Substantially Across Radiologists. Eur Urol Focus 2019;5:592-9.
12. Stabile A, Giganti F, Kasivisvanathan V, Giannarini G, Moore CM, Padhani AR, Panebianco V, Rosenkrantz AB, Salomon G, Turkbey B, Villeirs G, Barentsz JO.

Factors Influencing Variability in the Performance of Multiparametric Magnetic Resonance Imaging in Detecting Clinically Significant Prostate Cancer: A Systematic Literature Review. Eur Urol Oncol 2020;3:145-67.

13. Ruprecht O, Weisser P, Bodelle B, Ackermann H, Vogl TJ. MRI of the prostate: interobserver agreement compared with histopathologic outcome after radical prostatectomy. Eur J Radiol 2012;81:456-60.
14. Wildeboer RR, van Sloun RJG, Wijkstra H, Mischi M. Artificial intelligence in multiparametric prostate cancer imaging with focus on deep-learning methods. Comput Methods Programs Biomed 2020;189:105316.
15. Suarez-Ibarrola R, Sigle A, Eklund M, Eberli D, Miernik A, Benndorf M, Bamberg F, Gratzke C. Artificial Intelligence in Magnetic Resonance Imaging-based Prostate Cancer Diagnosis: Where Do We Stand in 2021? Eur Urol Focus 2022;8:409-17.
16. Radtke JP, Schwab C, Wolf MB, Freitag MT, Alt CD, Kesch C, Popeneciu IV, Huettenbrink C, Gasch C, Klein T, Bonekamp D, Duensing S, Roth W, Schueler S, Stock C, Schlemmer HP, Roethke M, Hohenfellner M, Hadaschik BA. Multiparametric Magnetic Resonance Imaging (MRI) and MRI-Transrectal Ultrasound Fusion Biopsy for Index Tumor Detection: Correlation with Radical Prostatectomy Specimen. Eur Urol 2016;70:846-53.
17. Zhu Y, Wei R, Gao G, Ding L, Zhang X, Wang X, Zhang J. Fully automatic segmentation on prostate MR images based on cascaded fully convolution network. J Magn Reson Imaging 2019;49:1149-56.
18. Sun Z, Wu P, Cui Y, Liu X, Wang K, Gao G, Wang H, Zhang X, Wang X. Deep-Learning Models for Detection and Localization of Visible Clinically Significant Prostate Cancer on Multi-Parametric MRI. J Magn Reson Imaging 2023;58:1067-81.
19. Zhu L, Gao G, Liu Y, Han C, Liu J, Zhang X, Wang X. Feasibility of integrating computer-aided diagnosis with structured reports of prostate multiparametric MRI. Clin Imaging 2020;60:123-30.
20. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. editors. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Cham: Springer International Publishing; 2016.
21. Padhani AR, Turkbey B. Detecting Prostate Cancer with Deep Learning for MRI: A Small Step Forward. Radiology 2019;293:618-9.
22. Gaur S, Lay N, Harmon SA, Doddakashi S, Mehralivand S, Argun B, et al. Can computer-aided diagnosis assist in the identification of prostate cancer on prostate MRI? a multi-center, multi-reader investigation. Oncotarget 2018;9:33804-17.
23. Giannini V, Mazzetti S, Armando E, Carabalona S, Russo F, Giacobbe A, Muto G, Regge D. Multiparametric magnetic resonance imaging of the prostate with computer-aided detection: experienced observer performance study. Eur Radiol 2017;27:4200-8.
24. Mayer R, Simone CB 2nd, Turkbey B, Choyke P. Algorithms applied to spatially registered multi-parametric MRI for prostate tumor volume measurement. Quant Imaging Med Surg 2021;11:119-32.
25. Woo S, Suh CH, Kim SY, Cho JY, Kim SH. Diagnostic Performance of Prostate Imaging Reporting and Data System Version 2 for Detection of Prostate Cancer: A Systematic Review and Diagnostic Meta-analysis. Eur Urol 2017;72:177-88.
26. Schelb P, Kohl S, Radtke JP, Wiesenfarth M, Kickingereder P, Bickelhaupt S, Kuder TA, Stenzinger A, Hohenfellner M, Schlemmer HP, Maier-Hein KH, Bonekamp D. Classification of Cancer at Prostate MRI: Deep Learning versus Clinical PI-RADS Assessment. Radiology 2019;293:607-17.
27. Greer MD, Brown AM, Shih JH, Summers RM, Marko J, Law YM, Sankineni S, George AK, Merino MJ, Pinto PA, Choyke PL, Turkbey B. Accuracy and agreement of PIRADSv2 for prostate cancer mpMRI: A multireader study. J Magn Reson Imaging 2017;45:579-85.
28. Wang Y, Wang W, Yi N, Jiang L, Yin X, Zhou W, Wang L. Detection of intermediate- and high-risk prostate cancer with biparametric magnetic resonance imaging: a systematic review and meta-analysis. Quant Imaging Med Surg 2023;13:2791-806.
29. Zhu L, Gao G, Zhu Y, Han C, Liu X, Li D, Liu W, Wang X, Zhang J, Zhang X, Wang X. Fully automated detection and localization of clinically significant prostate cancer on MR images using a cascaded convolutional neural network. Front Oncol 2022;12:958065.
30. Zhong X, Cao R, Shakeri S, Scalzo F, Lee Y, Enzmann DR, Wu HH, Raman SS, Sung K. Deep transfer learning-based prostate cancer classification using 3 Tesla multi-parametric MRI. Abdom Radiol (NY) 2019;44:2030-9.
31. Cao R, Zhong X, Afshari S, Felker E, Suvannarerg V, Tubtawee T, Vangala S, Scalzo F, Raman S, Sung K. Performance of Deep Learning and Genitourinary Radiologists in Detection of Prostate Cancer Using 3-T

Multiparametric Magnetic Resonance Imaging. J Magn Reson Imaging 2021;54:474-83.

32. Mehralivand S, Bednarova S, Shih JH, Mertan FV, Gaur S, Merino MJ, Wood BJ, Pinto PA, Choyke PL, Turkbey B. Prospective Evaluation of PI-RADS™ Version 2 Using the International Society of Urological Pathology Prostate Cancer Grade Group System. J Urol 2017;198:583-90.
33. Greer MD, Shih JH, Lay N, Barrett T, Kayat Bittencourt L, Borofsky S, Kabakus IM, Law YM, Marko J, Shebel H, Mertan FV, Merino MJ, Wood BJ, Pinto PA, Summers RM, Choyke PL, Turkbey B. Validation of the Dominant Sequence Paradigm and Role of Dynamic Contrast-enhanced Imaging in PI-RADS Version 2. Radiology 2017;285:859-69.
34. Yang X, Liu C, Wang Z, Yang J, Min HL, Wang L, Cheng KT. Co-trained convolutional neural networks for automated detection of prostate cancer in multi-parametric MRI. Med Image Anal 2017;42:212-27.
35. Wang Z, Liu C, Cheng D, Wang L, Yang X, Cheng KT. Automated Detection of Clinically Significant Prostate Cancer in mp-MRI Images Based on an End-to-End Deep Neural Network. IEEE Trans Med Imaging 2018;37:1127-39.
36. Saha A, Hosseinzadeh M, Huisman H. End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effects of attention mechanisms, clinical priori and decoupled false positive reduction. Med Image Anal 2021;73:102155.
37. Rundo L, Han C, Nagano Y, Zhang J, Hataya R, Militello C, Tangherloni A, Nobile MS, Ferretti C, Besozzi D. USE-Net: Incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI data sets. Neurocomputing 2019;365:31-43.
38. Zhang G, Wang W, Yang D, Luo J, He P, Wang Y, Luo Y, Zhao B, Lu J. A bi-attention adversarial network for prostate cancer segmentation. IEEE Access 2019;7:131448-58.
39. Zhou Z, Siddiquee MMR, Tajbakhsh N, Liang J. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. IEEE Trans Med Imaging 2020;39:1856-67.
40. Stolk TT, de Jong IJ, Kwee TC, Luiting HB, Mahesh SVK, Doornweerd BHJ, Willemse PM, Yakar D. False positives in PIRADS (V2) 3, 4, and 5 lesions: relationship with reader experience and zonal location. Abdom Radiol (NY) 2019;44:1044-51.
41. Johnson DC, Raman SS, Mirak SA, Kwan L, Bajgiran AM, Hsu W, Maehara CK, Ahuja P, Faiena I, Pooli A, Salmasi A, Sisk A, Felker ER, Lu DSK, Reiter RE. Detection of Individual Prostate Cancer Foci via Multiparametric Magnetic Resonance Imaging. Eur Urol 2019;75:712-20.
42. Priester A, Natarajan S, Khoshnoodi P, Margolis DJ, Raman SS, Reiter RE, Huang J, Grundfest W, Marks LS. Magnetic Resonance Imaging Underestimation of Prostate Cancer Geometry: Use of Patient Specific Molds to Correlate Images with Whole Mount Pathology. J Urol 2017;197:320-6.

Cite this article as: Sun Z, Wang K, Wu C, Chen Y, Kong Z, She L, Song B, Luo N, Wu P, Wang X, Zhang X, Wang X. Using an artificial intelligence model to detect and localize visible clinically significant prostate cancer in prostate magnetic resonance imaging: a multicenter external validation study. Quant Imaging Med Surg 2024;14(1):43-60. doi: 10.21037/qims-23-791

Appendix 1 The process of systematic and cognitive-targeted biopsy

A total of 12 dedicated urologists from three hospitals performed the prostate biopsies using the same biopsy techniques with their own hardware; that is, double-plane B-ultrasounds (LOGIQ E9, GE; EPIQ 7, Philips; Hivision Ascendus, Hitachi; RS80A, Samsung), transrectal probes, and corresponding puncture needle guns. For the system biopsy, 12- or 6-core needle biopsies were performed. For the targeted biopsy, based on structured reports prepared by dedicated urogenital radiologists during during routine clinical procedure, lesions suspected of malignancy were marked on a prostate sector map for the targeted biopsy. At least one urologist and one urogenital radiologist would review the MR images before biopsy in a multidisciplinary meeting to ensure the accurate localization of suspicious lesions. When performing the biopsies, the urologists examined each suspicious lesion with an additional needle core (a 2- to 5-core needle). The dedicated genitourinary pathologists analyzed and recorded the histopathology on each specimen.

Appendix 2 The components of the end-to-end AI model

In our processing pipeline, we trained distinct models to conduct MRI sequence classification, prostate gland segmentation and measurement, and prostate zonal anatomy segmentation. The training data for these models was acquired from 2009 to 2021, with varying volumes for each task. The training process for each model was concluded upon achieving a notable level of effectiveness. It should be noted that in the entire end-to-end AI model, the data of the external validation data set were not only mutually exclusive with the fourth model, but also were not used in the first three models.

Conversely, the csPCa foci segmentation and measurement model was trained using data from 2014 to 2019, and subsequently tested with data from 2020 to 2021. This approach ensured that the images used in the model's development data set and the external validation data set remained mutually exclusive.

It is worth noting that clinicopathological information on prostate mpMRI prior to 2014 was unavailable and thus data from 2014 onwards were exclusively used for the training process. Notably, the training process for the first three models relied solely on image data (and was not restricted by pathological information), which enabled us to incorporate data from 2009 to 2021.

MRI sequence classification

Data enrollment

The mpMRI images were retrospectively collected from 1,086 patients (1,153 mpMRI examinations) studied from July 28, 2009, to November 26, 2021. After importing the anonymized data, the DICOM data were converted to Nifty format using dicom2nii.py (Python 3.5) to obtain the image data. First, the DICOM data were split into multiple scan sequences for one MR examination. Individual sequences with more than 15 slices were included in the study. Then, each sequence was further split into an image group. The images with the same acquisition parameters and the same spatial location were split into one image group. The diffusion weighted imaging (DWI) sequence was grouped by b-value, for example, a DWI sequence with three b-values was split into three independent image groups, with each image group having only one unique b-value. In total, 5,151 images from five image types were ultimately classified, including (I) DWI_High (b value ≥500 s/mm^2, N=1,045); (II) DWI_Low (b value ≤100 s/mm^2, N=1,012); (III) apparent diffusion coefficient (ADC) map (N=906); (IV) T2-weighted imaging_nan (T2WI_nan) (non-fat-sat T2WI, N=1,000); and (V) T2WI_fs (fat-sat T2WI, N=1,188). The T1-weighted imaging (T1WI) and dynamic enhancement (DCE) images were scanned but excluded from the study.

MR scanners and imaging protocols

The mpMRI images were obtained from 15 MR scanners from four vendors. The transmit coils were body coils, and the receiver coils were phased array coils. No endorectal coils were used. Information on the MR scanners and image types is provided in *Table S1*.

Development of deep learning model

The input image was set to the automatic window width window level. Histogram equalization was performed. Each image was resized to 64×128×128 pixels. The training and validation data sets were augmented by some image transformations: rotation by –10° to 10°, random noise addition, perspective transformation, and translation of 0.01 pixels in cardinal or ordinal directions.

In total, 5,151 images were randomly split into 80% training, 10% validation, and 10% test sets. A modified Med3D network (*Figure S1*) was retrained to classify the sequences of prostate mpMRI. Using the method of transfer learning, we adopted the weight of the encoder to extract the image features. The encoder part was retained, and the decoder part (deconvolution part) of the network was replaced with the convolution layer and full connection layer of the classical classification network structure. The convolution layer used for classification had the following four layers: (I) the max-pooling layer (stride: 2); (II) the convolution layer (kernel: 3); (III) the max-pooling layer (stride: 2); (IV) the convolution layer (kernel: 3). The full-connection layer of the classification network was composed of 128 neurons, and the image features were combined and classified. The result was calculated, and output the classification array by the softmax function.

All the training processes were performed using the GPU NVIDIA Tesla P100 16G. The algorithm was coded by Python 3.6, PyTorch 0.4.1, OpenCV 3.4.0.12, Numpy 1.16.2, and SimpleITK 1.2.0. The parameters of the training options were set as follows: initial learning rate: 0.0001; mini-batch size: 4; maximum epochs: 400. The classification efficiency was evaluated by the confusion matrix.

Results

The confusion matrixes of the prediction results in different data sets are shown in *Figure S2*. The corresponding prediction efficacies of the image classification model in different data sets are shown in *Table S2*. The prediction accuracies of the training, validation, and test data sets were 0.992–1.000, 0.989–1.000, and 0.995–1.000, respectively.

Prostate gland segmentation and measurement

Data enrollment

The mpMRI images were retrospectively collected from 2,673 patients (2,849 mpMRI examinations) studied from July 28, 2009, to November 26, 2021.

After importing the anonymized data, the DICOM data were converted to nifty format using dicom2nii.py (Python 3.5). The ADC maps (N=2,320) were calculated from the DWI sequence with high and low b-values. Conventional T2WI and fat saturation T2WI (fat-sat T2WI) (N=3,654) were selected.

MR scanners and imaging protocols

The mpMRI images were obtained from 19 MR scanners from four vendors. The transmit coils were body coils, and the receiver coils were phased array coils. No endorectal coils were used. Information on the MR scanners and image types is shown in *Table S3*.

Development of the deep-learning model

The ground truth of the prostate gland was manually outlined by two experts, both of whom had more than five years of experience. The ADC and T2WI images were resized to 64×256×224 (z, y, x) pixels and were taken as the input of the network. We augmented the data in the training set by random rotation (rotation angle within 10°), adding random noise, and parallel translation at a range of [(–0.1; 0.1); (–0.1; 0.1)] pixels.

The results of the preliminary experiment have been published (17). We used the classic U-Net (20) framework, which enables accurate pixelwise prediction by combining spatial and contextual information in a network architecture comprising convolutional layers. All the training and experiments were conducted on a personal computer equipped with an Intel Core i5 3.2 GHz CPU with 16 GB main memory and an NVIDIA GTX1060 GPU. The proposed deep-learning network was implemented using the Keras open-source deep-learning library, and TensorFlow was chosen as a backend deep-learning engine. The learning rate was set as 0.0001, and the U-Net models were trained for up to 400 iterations.

The T2WI images were resized to 64×256×224 (z, y, x) pixels and were taken as the input of the network. We augmented the data in the training set by random rotation (rotation angle within 10°), adding random noise, and parallel translation at a range of [(–0.1; 0.1); (–0.1; 0.1)] pixels. In total, 1,225 images were randomly split into 80% training, 10% validation, and 10% test sets. A 3D U-Net segmentation framework (20) was used for the prostate anatomic segmentation. The model took the T2 weight image as input. All the training processes were performed using the GPU NVIDIA Tesla P100 16G. The algorithm was coded by Python 3.6, PyTorch 0.4.1, OpenCV 3.4.0.12, Numpy 1.16.2, and SimpleITK 1.2.0. The batch size was set as 10. The networks were trained for a total of 300 epochs. Adam was employed as an optimizer to minimize loss with a learning rate of 0.0001 and a binary cross-entropy loss function.

Results

Dice similarity coefficient (DSC), Jacard index, volumetric similarity (VS), Hausdorf distance (HD), and average distance (AD) values were used to compare the model and manual segmentation results. The right and left (RL) diameter, anterior and posterior (AP) diameter, and superior and inferior (SI) diameter of the prostate gland were automatically measured using the algorithm rule of the minimum volume bounding box (*Figure S3*).

The DSC, Jacard index, VS, HD, and AD in different data sets are shown in *Table S4* and *Figure S4*. The segmentation metrics of the T2WI were superior to those of the ADC map in all data sets (all P<0.001). The Bland-Altman analysis of the measured values of the prostate gland, including RL diameter, AP diameter, SI diameter, volume, and signal intensity, are shown in *Table S5* and *Figure S5*. The differences between the manual label and the predicted label to their means were –2.058% to 4.257%.

Prostate zonal anatomy segmentation

Prostate sextant locations model

First, the prostate gland was segmented by the established model (refer to part II). For the sextant location, the prostatic gland was then trisected to obtain the base, mid-gland, and apex in the longitudinal axis direction. It was bisected to divide the prostate gland into left and right parts in the horizontal axis direction. Thus, the sextants were automatically generated (*Figure S6*). When one sextant overlapped with a lesion, it was considered a cancer sextant; otherwise, it was considered a non-cancer sextant.

Prostate zonal anatomy segmentation

Second, for the anatomic zone locations, we developed an anatomic regional model to segment the peripheral zone (PZ), transition zone (TZ), central zone (CZ), anterior fibromuscular stroma (AFS), urethra (URE), left seminal vesicle (LS), and right seminal vesicle (RS) (*Figure S*7).

Data enrollment

The mpMRI images were retrospectively collected from 1,225 patients from August 29, 2012, to November 26, 2021. After importing the anonymized data, the DICOM data were converted to nifty format using dicom2nii.py (Python 3.5). T2WI images were used to develop the prostate zonal anatomy segmentation model.

MR scanners and imaging protocols

The T2WI images were obtained from 17 MR scanners from four vendors. The transmit coils were body coils, and the receiver coils were phased array coils. No endorectal coils were used. Information on the MR scanning protocols is provided in *Table S6*.

Development of deep-learning model

The T2WI images were resized to 64×256×224 (z, y, x) pixels and were taken as the input of the network. We augmented the data in the training set by random rotation (rotation angle within 10°), adding random noise, and parallel translation at a range of [(–0.1; 0.1); (–0.1; 0.1)] pixels. In total, 1,225 images were randomly split into 80% training, 10% validation, and 10% test sets. A 3D U-Net segmentation framework (20) was used for the prostate anatomic segmentation. The model took the T2 weight image as input. All the training processes were performed using the GPU NVIDIA Tesla P100 16G. The algorithm was coded by Python 3.6, PyTorch 0.4.1, OpenCV 3.4.0.12, Numpy 1.16.2, and SimpleITK 1.2.0. The batch size was set as 10. The networks were trained for a total of 300 epochs. Adam was employed as an optimizer to minimize loss with a learning rate of 0.0001 and a binary cross-entropy loss function.

Results

The DSC, JACRD, volume similarity, Hausdorff distance, and average distance in different data sets are shown in *Table S*7. The median metrics in the training, validation, and test data set showed statistically significant differences (P<0.001). When one zone overlapped with a lesion, it was considered a cancer zone; otherwise, it was considered a non-cancer zone.

Table S1 Information on the MR scanners and image types

Parameters	Overall (N=5,151)	Training (N=4,122)	Validation (N=513)	Test (N=516)	P value
Age (years)					
Median [Q1, Q3]	71.0 [65.0, 76.0]	71.0 [65.0, 76.0]	71.0 [66.0, 77.0]	71.0 [65.0, 76.0]	0.46
Image type					
ADC	906 (17.6%)	730 (17.7%)	86 (16.8%)	90 (17.4%)	>0.99
DWI_High	1,045 (20.3%)	835 (20.3%)	105 (20.5%)	105 (20.3%)	
DWI_Low	1,012 (19.6%)	808 (19.6%)	102 (19.9%)	102 (19.8%)	
T2WI_Fs	1,188 (23.1%)	950 (23.0%)	120 (23.4%)	118 (22.9%)	
T2WI_nan	1,000 (19.4%)	799 (19.4%)	100 (19.5%)	101 (19.6%)	
Magnetic field					
1.5 T	657 (12.8%)	523 (12.7%)	59 (11.5%)	75 (14.5%)	0.33
3.0 T	4494 (87.2%)	3599 (87.3%)	454 (88.5%)	441 (85.5%)	
Manufacture					
GE Medical Systems	2,635 (51.2%)	2100 (50.9%)	253 (49.3%)	282 (54.7%)	0.50
Philips Medical Systems	491 (9.5%)	397 (9.6%)	50 (9.7%)	44 (8.5%)	
SIEMENS	2,025 (39.3%)	1625 (39.4%)	210 (40.9%)	190 (36.8%)	
Station name					
AWP145938	597 (11.6%)	468 (11.4%)	73 (14.2%)	56 (10.9%)	0.17
AWP152194	119 (2.3%)	96 (2.3%)	12 (2.3%)	11 (2.1%)	
AWP166059	194 (3.8%)	164 (4.0%)	17 (3.3%)	13 (2.5%)	
AWP174090	8 (0.2%)	5 (0.1%)	0 (0.0%)	3 (0.6%)	
AWP39300	6 (0.1%)	5 (0.1%)	0 (0.0%)	1 (0.2%)	
DVMRDVMR	1,172 (22.8%)	939 (22.8%)	124 (24.2%)	109 (21.1%)	
GEHC	1,023 (19.9%)	821 (19.9%)	87 (17.0%)	115 (22.3%)	
GEHCGEHC	440 (8.5%)	340 (8.2%)	42 (8.2%)	58 (11.2%)	
MRC35207	696 (13.5%)	567 (13.8%)	69 (13.5%)	60 (11.6%)	
MRC40764	387 (7.5%)	306 (7.4%)	37 (7.2%)	44 (8.5%)	
MRSUZTB03A	57 (1.1%)	49 (1.2%)	4 (0.8%)	4 (0.8%)	
PHILIPS-8FA1B4E	72 (1.4%)	62 (1.5%)	5 (1.0%)	5 (1.0%)	
PHILIPS-CB0GKAC	12 (0.2%)	9 (0.2%)	0 (0.0%)	3 (0.6%)	
PHILIPS-DSALI1J	156 (3.0%)	124 (3.0%)	17 (3.3%)	15 (2.9%)	
PHILIPS-NK6RG9A	194 (3.8%)	153 (3.7%)	24 (4.7%)	17 (3.3%)	

The quantitative variables are presented as the median [Q1, Q3] for the non-normalized data. Fs, fat saturation; T2WI, T2-weighted imaging; ADC, apparent diffusion coefficient; DWI, diffusion-weighted imaging.

Table S2 Prediction efficacies of the image classification model in different data sets

Image type	Image number	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	Kappa	Prevalence	Detection rate	Detection prevalence
Training											
ADC	718	1.000	1.000	1.000	0.999	1.000	0.999	0.999	0.174	0.174	0.174
DWI_High	849	0.996	0.994	0.998	0.991	0.998	0.992	0.990	0.206	0.205	0.207
DWI_Low	815	0.992	0.987	0.998	0.991	0.997	0.989	0.986	0.198	0.195	0.197
T2WI_Fs	957	0.998	0.997	0.999	0.998	0.999	0.997	0.997	0.232	0.231	0.232
T2WI_nan	783	1.000	1.000	1.000	0.999	1.000	0.999	0.999	0.190	0.190	0.190
Validation											
ADC	96	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.187	0.187	0.187
DWI_High	93	0.989	0.978	1.000	1.000	0.995	0.989	0.987	0.181	0.177	0.177
DWI_Low	101	0.998	1.000	0.995	0.981	1.000	0.990	0.988	0.197	0.197	0.201
T2WI_Fs	107	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.209	0.209	0.209
T2WI_nan	116	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.226	0.226	0.226
Test											
ADC	92	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.178	0.178	0.178
DWI_High	103	0.995	0.990	1.000	1.000	0.998	0.995	0.994	0.200	0.198	0.198
DWI_Low	96	0.999	1.000	0.998	0.990	1.000	0.995	0.994	0.186	0.186	0.188
T2WI_Fs	124	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.240	0.240	0.240
T2WI_nan	101	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.196	0.196	0.196

ADC, apparent diffusion coefficient; T2WI, T2-weighted imaging; DWI, diffusion weighted imaging; Fs, fat saturation; PPV, positive predictive value; NPV, negative predictive value.

Table S3 Information on the MR scanners and image types

Parameters	Overall (N=5,974)	Training (N=4,780)	Validation (N=601)	Test (N=593)	P value
Age (years), median [Q1, Q3]	70.0 [64.0, 76.0]	70.0 [64.0, 76.0]	70.0 [64.0, 76.0]	70.0 [63.0, 75.0]	0.29
Magnetic field					0.50
1.5 T	1,034 (17.3%)	841 (17.6%)	96 (16.0%)	97 (16.4%)	
3.0 T	4,940 (82.7%)	3,939 (82.4%)	505 (84.0%)	496 (83.6%)	
Image type					0.35
ADC	2,320 (38.8%)	1,873 (39.2%)	217 (36.1%)	230 (38.8%)	
T2WI	3,654 (61.2%)	2,907 (60.8%)	384 (63.9%)	363 (61.2%)	
Manufacture					0.79
GE Medical Systems	3,243 (54.3%)	2,599 (54.4%)	316 (52.6%)	328 (55.3%)	
Philips Medical Systems	810 (13.6%)	637 (13.3%)	91 (15.1%)	82 (13.8%)	
SIEMENS	1,695 (28.4%)	1,368 (28.6%)	168 (28.0%)	159 (26.8%)	
UIH	226 (3.8%)	176 (3.7%)	26 (4.3%)	24 (4.0%)	
Model name					0.87
Achieva	150 (2.5%)	123 (2.6%)	17 (2.8%)	10 (1.7%)	
Ingenia	583 (9.8%)	456 (9.5%)	64 (10.6%)	63 (10.6%)	
Ingenia CX	3 (0.1%)	2 (0.0%)	1 (0.2%)	0 (0.0%)	
Discovery MR750	2,753 (46.1%)	2,204 (46.1%)	276 (45.9%)	273 (46.0%)	
Discovery MR750w	304 (5.1%)	250 (5.2%)	21 (3.5%)	33 (5.6%)	
Signa EXCITE	173 (2.9%)	135 (2.8%)	16 (2.7%)	22 (3.7%)	
Signa HDxt	11 (0.2%)	9 (0.2%)	2 (0.3%)	0 (0.0%)	
Signa Premier	2 (0.0%)	1 (0.0%)	1 (0.2%)	0 (0.0%)	
Aera	889 (14.9%)	724 (15.1%)	81 (13.5%)	84 (14.2%)	
Amira	4 (0.1%)	3 (0.1%)	1 (0.2%)	0 (0.0%)	
Essenza	2 (0.0%)	2 (0.0%)	0 (0.0%)	0 (0.0%)	
Multiva	74 (1.2%)	56 (1.2%)	9 (1.5%)	9 (1.5%)	
Prisma	127 (2.1%)	99 (2.1%)	16 (2.7%)	12 (2.0%)	
Skyra	188 (3.1%)	156 (3.3%)	18 (3.0%)	14 (2.4%)	
TrioTim	348 (5.8%)	273 (5.7%)	39 (6.5%)	36 (6.1%)	
Verio	137 (2.3%)	111 (2.3%)	13 (2.2%)	13 (2.2%)	
uMR 790	226 (3.8%)	176 (3.7%)	26 (4.3%)	24 (4.0%)	

The quantitative variables are presented as the median [Q1, Q3] for the non-normalized data. ADC, apparent diffusion coefficient; T2WI, T2-weighted imaging.

Table S4 Segmentation metrics in different data sets

Param-eters	Overall		Training		Validation		Test	
	ADC (N=2,320)	T2WI (N=3,654)	ADC (N=1,873)	T2WI (N=2,907)	ADC (N=217)	T2WI (N=384)	ADC (N=230)	T2WI (N=363)
DSC	0.921 (0.0337)	0.937 (0.0322)	0.925 (0.0279)	0.940 (0.0270)	0.902 (0.0491)	0.922 (0.0445)	0.900 (0.0446)	0.923 (0.0449)
JACRD	0.854 (0.0549)	0.883 (0.0537)	0.862 (0.0467)	0.889 (0.0458)	0.825 (0.0755)	0.857 (0.0717)	0.822 (0.0704)	0.861 (0.0726)
VS	0.970 (0.0303)	0.979 (0.0245)	0.974 (0.0236)	0.981 (0.0200)	0.953 (0.0478)	0.967 (0.0370)	0.955 (0.0449)	0.971 (0.0339)
HD	6.460 (3.150)	5.880 (2.830)	6.20 (2.720)	5.640 (2.490)	7.450 (4.700)	6.780 (3.730)	7.700 (3.990)	6.840 (3.730)
AD	0.140 (0.149)	0.168 (4.01)	0.122 (0.0962)	0.172 (4.49)	0.213 (0.293)	0.152 (0.196)	0.214 (0.233)	0.149 (0.186)

Data conforming to a normal distribution are presented as the mean (standard deviation). DSC, dice similarity coefficient; VS, volumetric similarity; HD, Hausdorff distance; AD, average distance; T2WI, T2-weighted imaging; ADC, apparent diffusion coefficient.

Table S5 Bland-Altman analysis of the measured values of the prostate gland

Parameters	RL diameter (mm)	AP diameter (mm)	SI diameter (mm)	Volume (cm^3)	Signal intensity
Means of label and plabel	56.025	59.670	63.655	108.623	56.025
Differences	2.230	2.540	–1.310	–0.924	2.230
Means/differences proportion	3.980	4.257	–2.058	–0.851	3.980
Means of label	57.140	60.940	63.000	108.160	57.140
Means of plabel	54.910	58.400	64.310	109.085	54.910
Bias of the label and plabel	0.733	1.094	–1.180	–0.623	0.733
Bias upper CI	0.800	1.172	–1.063	–0.477	0.800
Bias lower CI	0.665	1.017	–1.298	–0.769	0.665
Bias std dev	2.679	3.057	4.643	5.756	2.679
Bias standard error	0.035	0.040	0.060	0.745	0.035
LOA standard error	0.059	0.068	0.103	0.127	0.059
Upper LOA	5.984	7.085	7.919	10.659	5.984
Upper LOA_upperCI	6.100	7.218	8.120	10.908	6.100
Upper LOA_lowerCI	5.868	6.953	7.718	10.409	5.868
Lower LOA	–4.519	–4.897	–10.280	–11.904	–4.519
Lower LOA_upperCI	–4.403	–4.764	–10.078	–11.655	–4.403
Lower LOA_lowerCI	–4.635	–5.029	–10.481	–12.154	–4.635
Regression fixed slope	0.076	0.071	0.032	0.023	0.076
Regression fixed intercept	–3.100	–2.100	–2.700	–1.900	–3.100

LOA, limits of agreement; CI, confidence interval; RL, right-left; AP, anteroposterior; SI, superoinferior.

Table S6 Scanning protocols of the T2WI

Parameters	Overall (N=1,225)	Training (N=973)	Validation (N=99)	Test (N=153)	P value
Magnetic field					
1.5 T	271 (22.1%)	213 (21.9%)	14 (14.1%)	44 (28.8%)	0.02
3.0 T	954 (77.9%)	760 (78.1%)	85 (85.9%)	109 (71.2%)	
Manufacture					
GE Medical Systems	665 (54.3%)	540 (55.5%)	65 (65.7%)	60 (39.2%)	<0.001
Philips Medical Systems	155 (12.7%)	111 (11.4%)	10 (10.1%)	34 (22.2%)	
SIEMENS	365 (29.8%)	289 (29.7%)	20 (20.2%)	56 (36.6%)	
UIH	40 (3.3%)	33 (3.4%)	4 (4.0%)	3 (2.0%)	
Model name					
Achieva	26 (2.1%)	20 (2.1%)	3 (3.0%)	3 (2.0%)	0.01
Aera	239 (19.5%)	192 (19.7%)	8 (8.1%)	39 (25.5%)	
Amira	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
DISCOVERY MR750	547 (44.7%)	446 (45.8%)	54 (54.5%)	47 (30.7%)	
DISCOVERY MR750w	78 (6.4%)	64 (6.6%)	5 (5.1%)	9 (5.9%)	
Ingenia	111 (9.1%)	79 (8.1%)	5 (5.1%)	27 (17.6%)	
Ingenia CX	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
MAGNETOM_ESSENZA	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
Multiva	17 (1.4%)	11 (1.1%)	2 (2.0%)	4 (2.6%)	
Prisma	10 (0.8%)	7 (0.7%)	1 (1.0%)	2 (1.3%)	
SIGNA EXCITE	36 (2.9%)	28 (2.9%)	5 (5.1%)	3 (2.0%)	
Signa HDxt	3 (0.2%)	1 (0.1%)	1 (1.0%)	1 (0.7%)	
SIGNA Premier	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
Skyra	34 (2.8%)	28 (2.9%)	3 (3.0%)	3 (2.0%)	
TrioTim	41 (3.3%)	35 (3.6%)	2 (2.0%)	4 (2.6%)	
uMR 790	40 (3.3%)	33 (3.4%)	4 (4.0%)	3 (2.0%)	
Verio	39 (3.2%)	25 (2.6%)	6 (6.1%)	8 (5.2%)	
FatSat					
fs	87 (7.1%)	67 (6.9%)	10 (10.1%)	10 (6.5%)	0.47
Non-fs	1,138 (92.9%)	906 (93.1%)	89 (89.9%)	143 (93.5%)	
Repetition time (ms)	3,560 [3,040, 3,880]	3,460 [3,040, 3,850]	3,560 [3,070, 3,790]	3,730 [3,000, 4,200]	0.30
Echo time (ms)	92.9 [87.5, 112]	92.2 [87.4, 110]	90.3 [87.4, 103]	99.0 [88.0, 115]	0.05
Pixel bandwidth (Hz)	163 [163, 200]	163 [163, 200]	163 [122, 188]	200 [160, 218]	<0.001
Flip angle	111 [111, 140]	111 [111, 140]	111 [111, 111]	111 [111, 150]	0.37
Reconstruction diameter (mm)	240 [200, 240]	240 [200, 240]	240 [200, 240]	220 [200, 240]	0.01
Slice thickness (mm)	4.00 [3.50, 4.00]	4.00 [3.50, 4.00]	4.00 [3.40, 4.00]	4.00 [3.50, 4.00]	0.44
Slice spacing (mm)	4.00 [4.00, 4.00]	4.00 [4.00, 4.00]	4.00 [4.00, 4.00]	4.00 [3.60, 4.00]	0.06
Pixel spacing (mm)	0.469 [0.469, 0.577]	0.469 [0.469, 0.625]	0.469[0.417, 0.469]	0.469 [0.344, 0.625]	0.05

Data are presented as n (%) or median [Q1, Q3].

Table S7 Segmentation metrics of the model

Parameters	Overall (N=1,225)	Training (N=979)	Validation (N=123)	Test (N=123)	P value
AFS					
DSC					
Median [Min, Max]	0.790 [0, 0.920]	0.800 [0, 0.920]	0.710 [0, 0.890]	0.690 [0.0300, 0.860]	<0.001
Missing	11 (0.9%)	8 (0.8%)	1 (0.8%)	2 (1.6%)	
JACRD					
Median [Min, Max]	0.650 [0, 0.850]	0.670 [0, 0.850]	0.550 [0, 0.800]	0.530 [0.020, 0.760]	<0.001
Missing	11 (0.9%)	8 (0.8%)	1 (0.8%)	2 (1.6%)	
VS					
Median [Min, Max]	0.930 [0.0300, 1.00]	0.940 [0.0300, 1.00]	0.885 [0.230, 1.00]	0.890 [0.140, 1.00]	<0.001
Missing	11 (0.9%)	8 (0.8%)	1 (0.8%)	2 (1.6%)	
HD					
Median [Min, Max]	5.05 [1.56, 50.3]	4.77 [1.56, 50.3]	6.89 [2.21, 47.0]	6.64 [2.50, 48.6]	<0.001
Missing	11 (0.9%)	8 (0.8%)	1 (0.8%)	2 (1.6%)	
AD					
Median [Min, Max]	0.260 [0.0900, 25.3]	0.230 [0.090, 25.3]	0.410 [0.100, 9.58]	0.450 [0.130, 3.73]	<0.001
Missing	11 (0.9%)	8 (0.8%)	1 (0.8%)	2 (1.6%)	
PZ					
DSC					
Median [Min, Max]	0.870 [0, 0.960]	0.88 [0, 0.96]	0.84 [0.48, 0.92]	0.840 [0.390, 0.930]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
JACRD					
Median [Min, Max]	0.770 [0, 0.920]	0.780 [0, 0.920]	0.720 [0.31, 0.86]	0.720 [0.240, 0.870]	
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
VS					
Median [Min, Max]	0.970 [0.100, 1.00]	0.970 [0.100, 1.00]	0.95 [0.61, 1.00]	0.960 [0.530, 1.00]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
HD					
Median [Min, Max]	7.55 [2.50, 50.2]	7.20 [2.50, 43.7]	8.91 [2.58, 50.2]	8.11 [3.85, 44.3]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
AD					
Median [Min, Max]	0.160 [0.0500, 19.5]	0.150 [0.0500, 19.5]	0.240 [0.080, 2.05]	0.240 [0.080, 4.43]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
CZ					
DSC					
Median [Min, Max]	0.810 [0, 0.930]	0.820 [0.410, 0.930]	0.650 [0.05, 0.87]	0.630 [0, 0.900]	<0.001
Missing	6 (0.5%)	6 (0.6%)	0 (0.0%)	0 (0.0%)	
JACRD					
Median [Min, Max]	0.680 [0, 0.880]	0.700 [0.260, 0.880]	0.480 [0.03, 0.770]	0.460 [0, 0.810]	<0.001
Missing	6 (0.5%)	6 (0.6%)	0 (0.0%)	0 (0.0%)	
VS					
Median [Min, Max]	0.920 [0.170, 1.00]	0.930 [0.490, 1.00]	0.88 [0.180, 1.00]	0.880 [0.170, 1.00]	<0.001
Missing	6 (0.5%)	6 (0.6%)	0 (0.0%)	0 (0.0%)	
AD					
Median [Min, Max]	4.60 [2.00, 45.5]	4.29 [2.00, 34.1]	6.53 [2.80, 45.5]	6.50 [2.73, 33.0]	<0.001
Missing	6 (0.5%)	6 (0.6%)	0 (0.0%)	0 (0.0%)	
HD					
Median [Min, Max]	0.240 [0.0600, 9.09]	0.220 [0.060, 3.89]	0.600 [0.140, 8.41]	0.610 [0.120, 9.09]	<0.001
Missing	6 (0.5%)	6 (0.6%)	0 (0.0%)	0 (0.0%)	
TZ					
DSC					
Median [Min, Max]	0.930 [0.610, 0.970]	0.940 [0.720, 0.970]	0.910 [0.610, 0.970]	0.920 [0.700, 0.970]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
JACRD					
Median [Min, Max]	0.870 [0.440, 0.950]	0.880 [0.560, 0.950]	0.830 [0.44, 0.940]	0.850 [0.540, 0.940]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
VS					
Median [Min, Max]	0.980 [0.770, 1.00]	0.990 [0.840, 1.00]	0.970 [0.770, 1.00]	0.970 [0.830, 1.00]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
HD					
Median [Min, Max]	4.59 [2.34, 38.3]	4.46 [2.34, 34.0]	5.33 [2.47, 38.3]	4.91 [2.72, 19.4]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
AD					
Median [Min, Max]	0.080 [0.020, 1.07]	0.080 [0.020, 0.890]	0.130 [0.03, 1.07]	0.110 [0.0300, 0.730]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
URE					
DSC					
Median [Min, Max]	0.910 [0, 0.980]	0.920 [0.520, 0.980]	0.830 [0, 0.960]	0.830 [0.490, 0.960]	<0.001
Missing	8 (0.7%)	5 (0.5%)	1 (0.8%)	2 (1.6%)	
JACRD					
Median [Min, Max]	0.830 [0, 0.960]	0.840 [0.350, 0.960]	0.700 [0, 0.930]	0.700 [0.320, 0.930]	<0.001
Missing	8 (0.7%)	5 (0.5%)	1 (0.8%)	2 (1.6%)	
VS					
Median [Min, Max]	0.940 [0.0800, 1.00]	0.950 [0.550, 1.00]	0.890 [0.080, 1.00]	0.900 [0.490, 1.00]	<0.001
Missing	8 (0.7%)	5 (0.5%)	1 (0.8%)	2 (1.6%)	
HD					
Median [Min, Max]	1.88 [0.780, 49.5]	1.75 [0.780, 49.5]	3.31 [0.940, 17.1]	3.13 [0.780, 33.8]	<0.001
Missing	8 (0.7%)	5 (0.5%)	1 (0.8%)	2 (1.6%)	
AD					
Median [Min, Max]	0.0900 [0.020, 723]	0.080 [0.020, 1.18]	0.220 [0.04, 723]	0.200 [0.03, 1.13]	<0.001
Missing	8 (0.7%)	5 (0.5%)	1 (0.8%)	2 (1.6%)	
RS					
DSC					
Median [Min, Max]	0.920 [0, 0.970]	0.930 [0, 0.970]	0.900 [0.710, 0.97]	0.900 [0, 0.970]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
JACRD					
Median [Min, Max]	0.860 [0, 0.940]	0.860 [0, 0.940]	0.82 [0.550, 0.930]	0.830 [0, 0.940]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
VS					
Median [Min, Max]	0.970 [0.760, 1.00]	0.980 [0.780, 1.00]	0.970 [0.760, 1.00]	0.960 [0.760, 1.00]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
AD					
Median [Min, Max]	4.17 [1.37, 52.9]	3.98 [1.37, 52.9]	4.94 [1.92, 39.7]	4.74 [1.88, 37.0]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
HD					
Median [Min, Max]	0.090 [0.030, 805]	0.080 [0.030, 805]	0.130 [0.03, 107]	0.120 [0.030, 15.1]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
LS					
DSC					
Median [Min, Max]	0.920 [0.080, 0.970]	0.930 [0.260, 0.970]	0.90 [0.08, 0.960]	0.900 [0.260, 0.960]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
JACRD					
Median [Min, Max]	0.860 [0.040, 0.950]	0.860 [0.150, 0.950]	0.830 [0.04, 0.920]	0.830 [0.150, 0.920]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
VS					
Median [Min, Max]	0.980 [0.120, 1.00]	0.980 [0.800, 1.00]	0.970 [0.120, 1.00]	0.960 [0.260, 1.00]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
HD					
Median [Min, Max]	3.75 [1.33, 42.3]	3.75 [1.33, 41.3]	4.26 [1.88, 35.5]	4.42 [2.08, 42.3]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	
AD					
Median [Min, Max]	0.0900 [0.030, 9.54]	0.080 [0.030, 9.54]	0.110 [0.04, 4.80]	0.130 [0.0400, 1.70]	<0.001
Missing	1 (0.1%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	

The categorical variables are presented as numbers (percentages). The quantitative variables are presented as the median [Min, Max] for the non-normalized data. DSC, dice similarity coefficient; VS, volumetric similarity; HD, Hausdorff distance; AD, average distance; AFS, anterior fibromuscular stroma; PZ, peripheral zone; CZ, central zone; TZ, transition zone; URE, urethra; LS, left seminal vesicle; RS, right seminal vesicle.

Table S8 Recent deep-learning studies in prostate cancer detection or segmentation

Study	Algorithm	Sequences	Scanner	Field strength	Cohort (patients)	Validation cohort (patients)	Ground truth	Performance						
								Lesion	Sextant			Patient		
									AUC	Sen	Spe	AUC	Sen	Spe
Sun (18)	U-Net	DWI, ADC	7	1.5, 3.0	1,628	200	WMHP, Biopsy	Sen: 0.9	0.895	0.92	0.908	0.865	0.97	0.77
Zhu (19)	Res-Unet	T2W, ADC	1	3.0	347	88	Biopsy	Sen: 0.955	–	0.956	0.915	–	0.986	0.648
Schelb (26)	U-Net	T2WI, DWI	1	3.0	312	62	Biopsy	–	–	0.59	0.66	–	0.96	0.31
Zhong (30)	ResNet	T2W, ADC	6	3.0	140	30	WMHP	AUC: 0.726, lesion pach level	–	–	–	–	–	–
Cao (31)	CNN	T2W, ADC	4	3.0	553	126	WMHP	FROC: 0.50, 0.80, and 0.90 at 0.43, 3.39, and 11.7 false-positive detections per patient	–	–		–	–	–

CNN, convolutional neural network; DWI, diffusion-weighted imaging; ADC, apparent diffusion coefficient; T2WI, T2-weighted imaging; Sen, sensitivity; Spe, specificity; AUC, area under the curve; WMHP, whole-mount histopathology; FROC, free-response receiver operating characteristic.

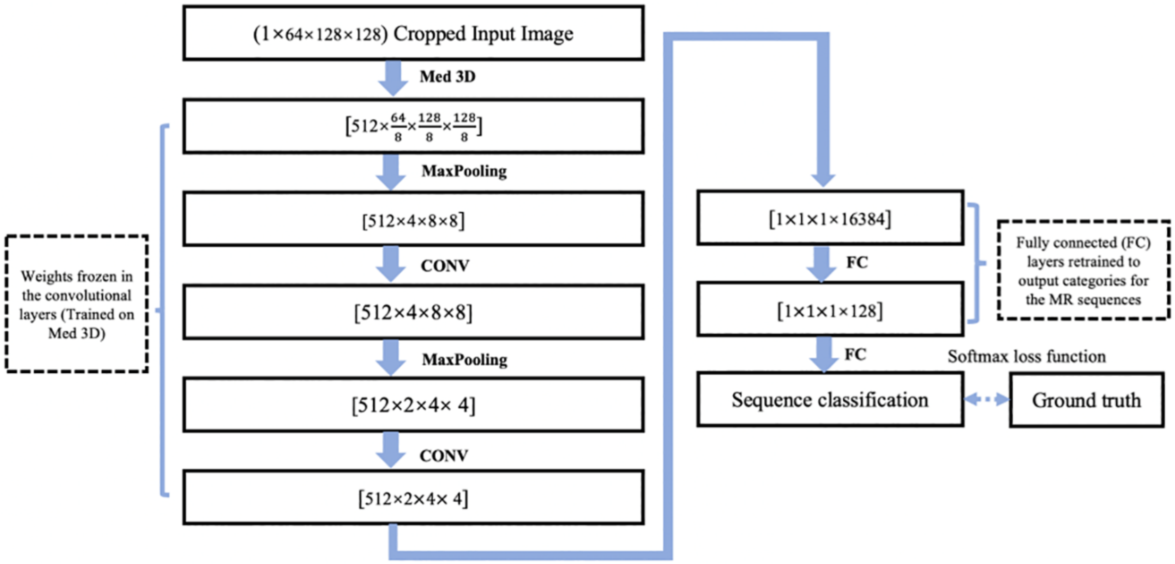
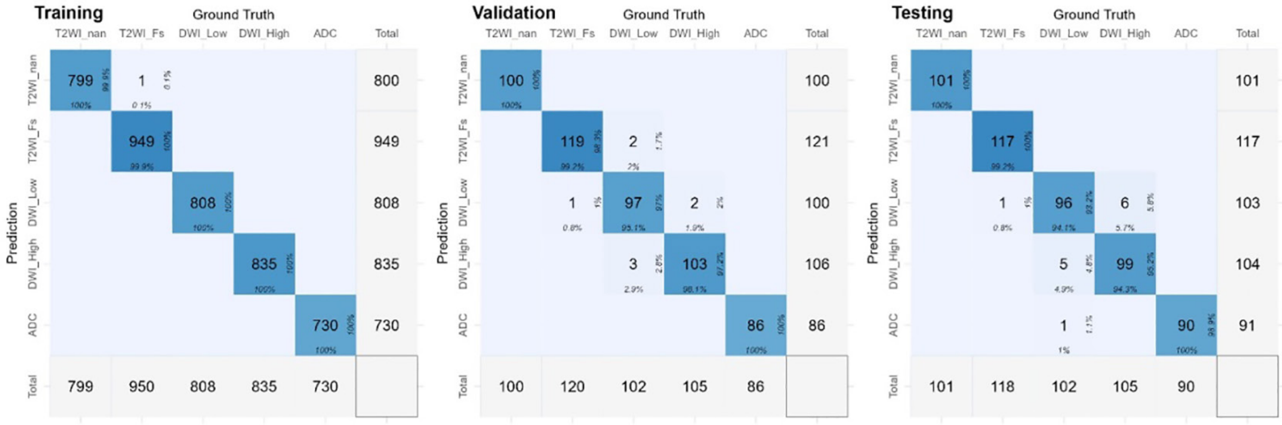


Figure S1 The modified Med3D network. 3D, three-dimensional; CONV, convolution; FC, fully connected.



Training

Prediction \ Ground Truth	T2WI_nan	T2WI_Fs	DWI_Low	DWI_High	ADC	Total
T2WI_nan	799 99.9% 100%	1 0.1% 0.1%				800
T2WI_Fs		949 100% 99.9%				949
DWI_Low			808 100% 100%			808
DWI_High				835 100% 100%		835
ADC					730 100% 100%	730
Total	799	950	808	835	730	

Validation

Prediction \ Ground Truth	T2WI_nan	T2WI_Fs	DWI_Low	DWI_High	ADC	Total
T2WI_nan	100 100% 100%					100
T2WI_Fs		119 98.3% 99.2%	2 1.7% 2%			121
DWI_Low		1 1% 0.8%	97 97% 95.1%	2 2% 1.9%		100
DWI_High			3 2.8% 2.9%	103 97.2% 98.1%		106
ADC					86 100% 100%	86
Total	100	120	102	105	86	

Testing

Prediction \ Ground Truth	T2WI_nan	T2WI_Fs	DWI_Low	DWI_High	ADC	Total
T2WI_nan	101 100% 100%					101
T2WI_Fs		117 100% 99.2%				117
DWI_Low		1 1% 0.8%	96 93.2% 94.1%	6 5.8% 5.7%		103
DWI_High			5 4.8% 4.9%	99 95.2% 94.3%		104
ADC			1 1.1% 1%		90 98.9% 100%	91
Total	101	118	102	105	90	

Figure S2 Confusion matrix of the prediction results in the training, validation, and test data sets. The number in the middle of each tile is the counted number of images. The percentage number at the bottom of each tile is the column percentage. The percentage number at the right side of each tile is the row percentage. The color intensity is based on the counts. T2WI, T2-weighted imaging; Fs, fat saturation; DWI, diffusion-weighted imaging; ADC, apparent diffusion coefficient.

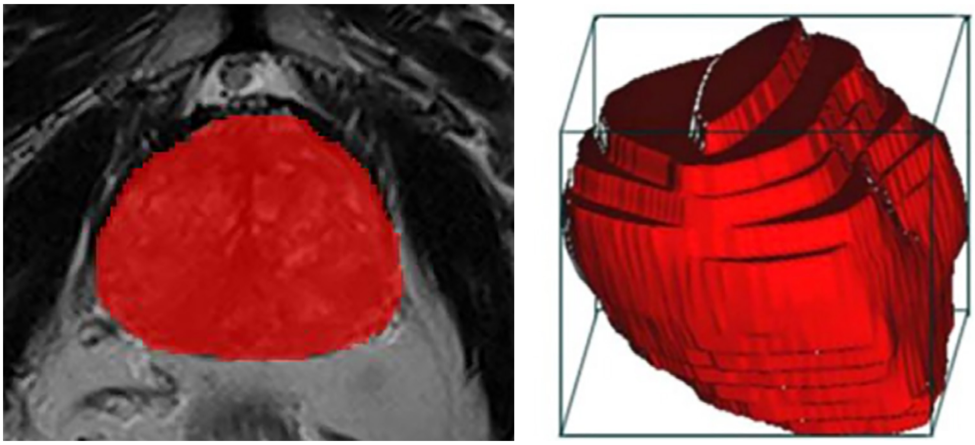


Figure S3 Whole prostate segmentation and the algorithm rule of the minimum volume bounding box.

https://dx.doi.org/10.21037/qims-23-791

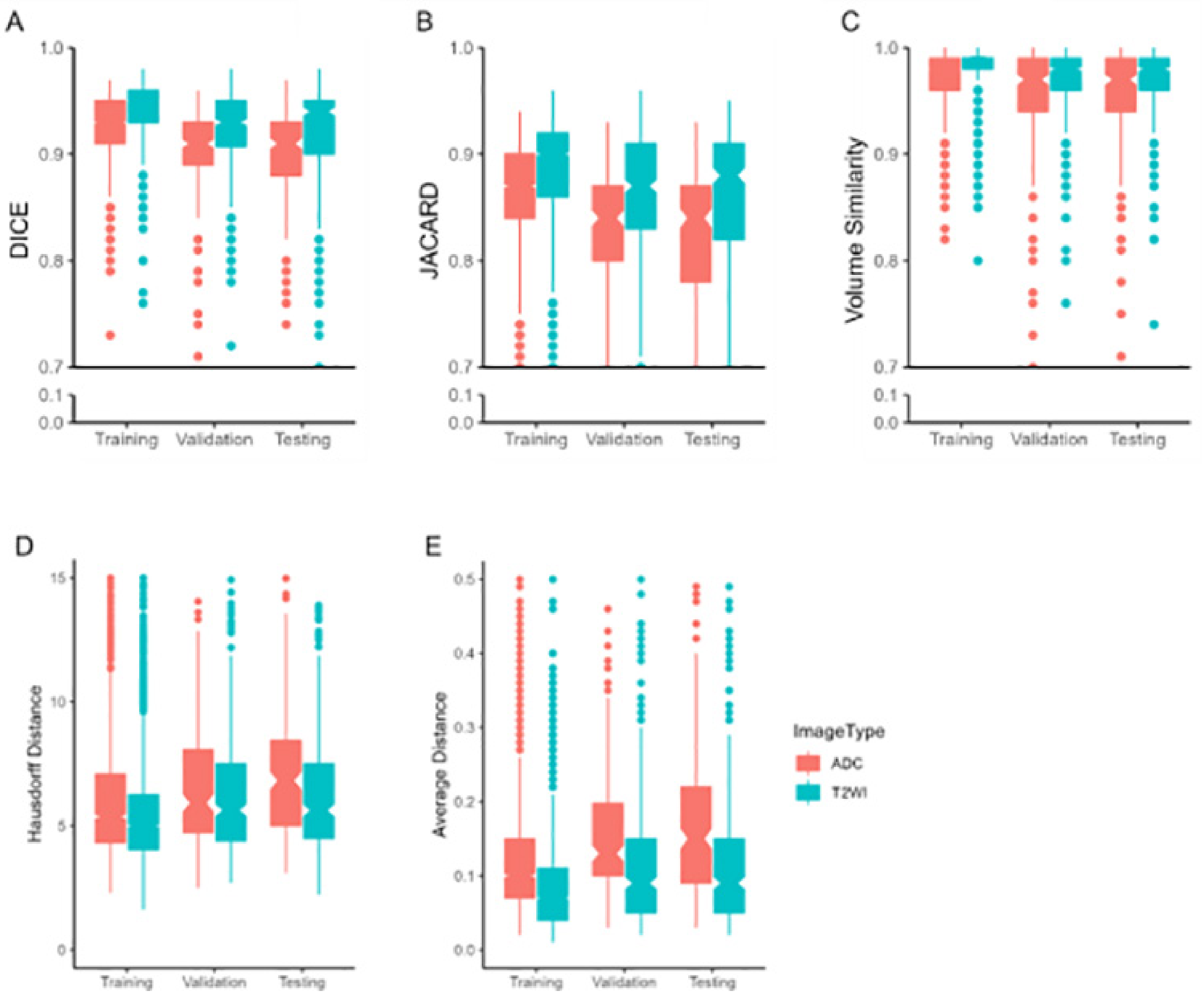


Figure S4 The DSC, Jacard index, VS, HD, and AD values in different data sets. The metrics of the T2WI were superior to those of the ADC map in all the data sets (all P<0.001). DSC, dice similarity coefficient; VS, volumetric similarity; HD, Hausdorff distance; AD, average distance; T2WI, T2-weighted imaging; ADC, apparent diffusion coefficient.

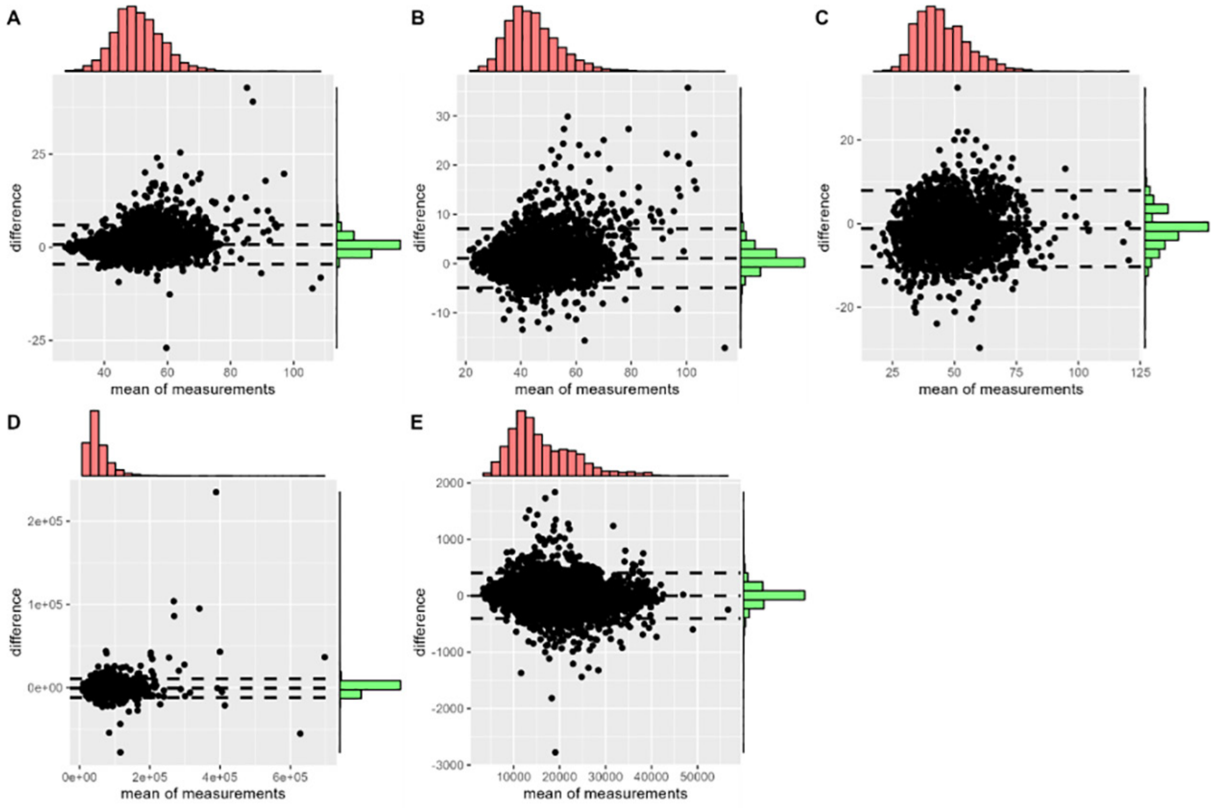


Figure S5 Bland-Altman analysis of the values of the RL diameter (A), AP diameter (B), SI diameter (C), volume (D), and signal intensity (E) of the manual label and the predicted label of the prostate gland. RL, right and left; AP, anterior and posterior; SI, superior and inferior.

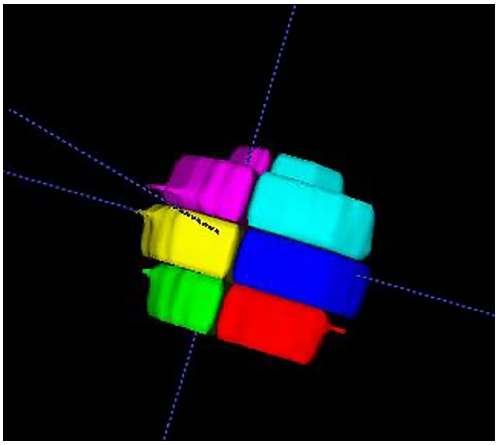


Figure S6 Sextant locations.

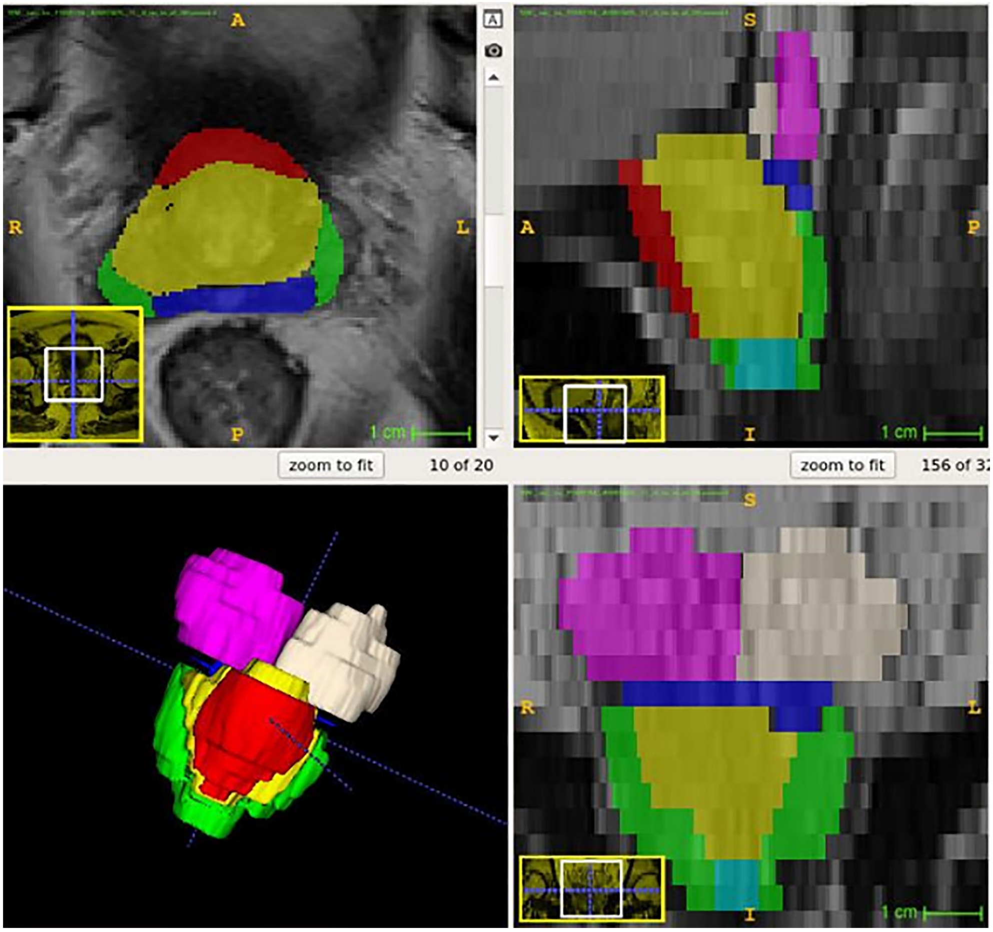


Figure S7 Anatomic zone locations.