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

© Quantitative Imaging in Medicine and Surgery. All rights reserved.
https://dx.doi.org/10.21037/qims-23-791

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 S7*).

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 S7*. 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.

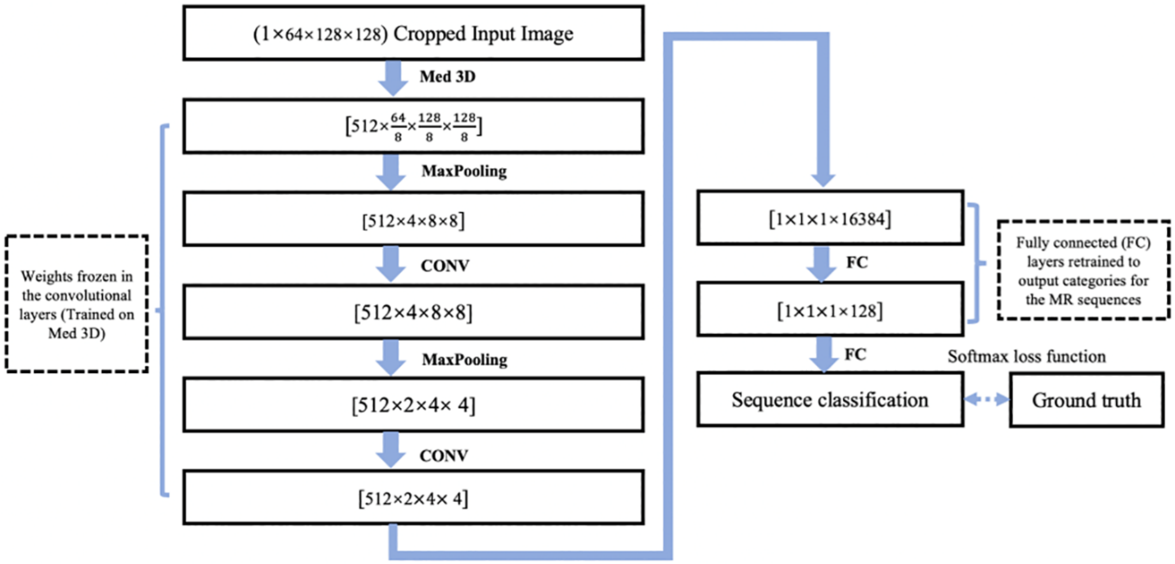
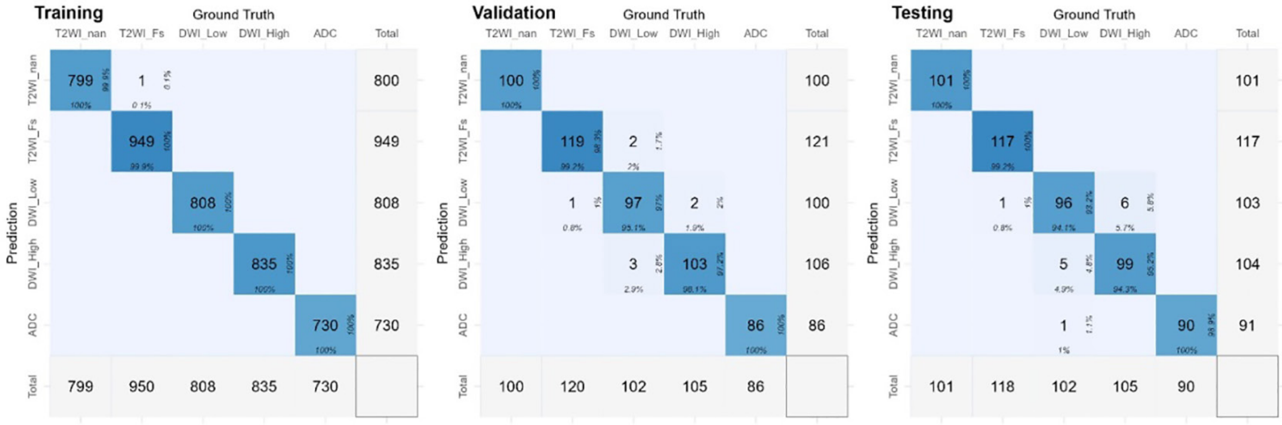


Figure S1 The modified Med3D network. 3D, three-dimensional; CONV, convolution; FC, fully connected.



Training	Ground Truth					
Prediction	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	Ground Truth					
Prediction	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	Ground Truth					
Prediction	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.

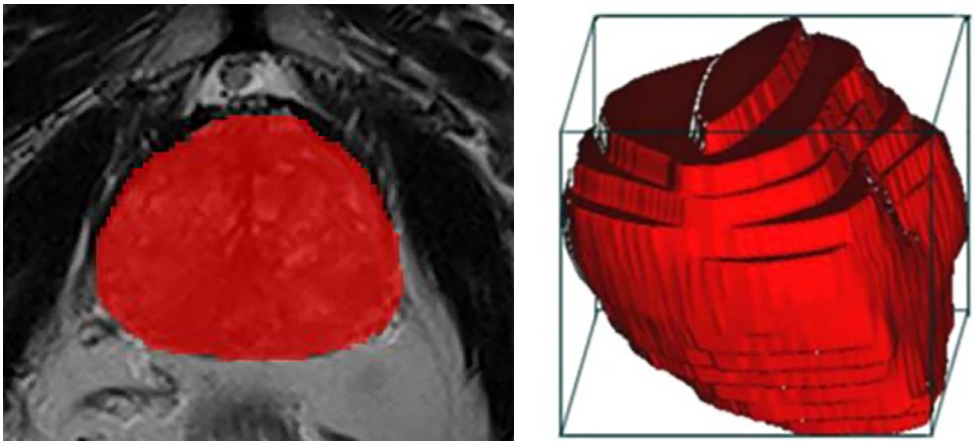


Figure S3 Whole prostate segmentation and the algorithm rule of the minimum volume bounding box.

© Quantitative Imaging in Medicine and Surgery. All rights reserved. https://dx.doi.org/10.21037/qims-23-791

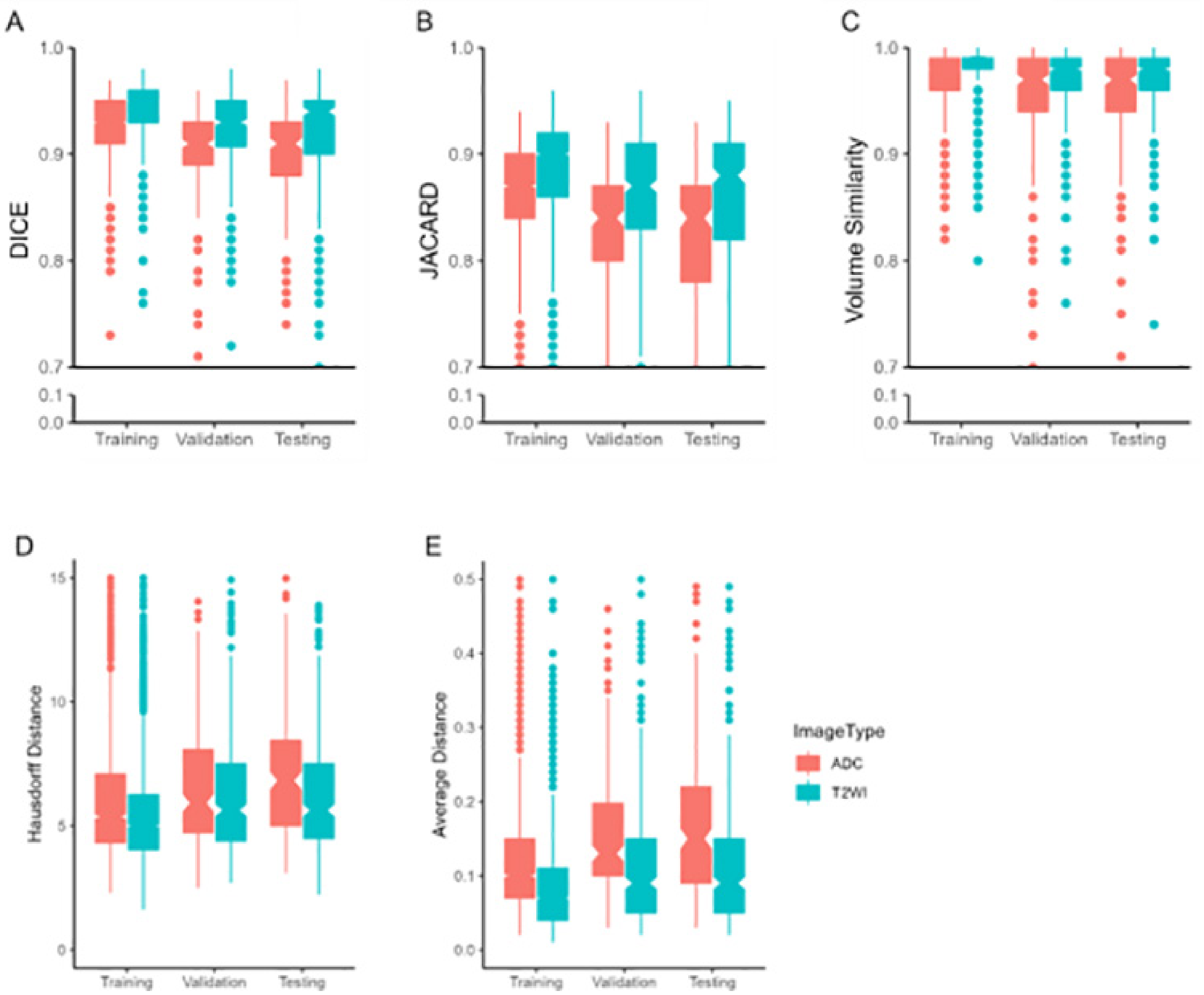


Figure S4 The DSC, Jacard index, VS, HD, and AD values in different data sets. The metrics of the T2WI were superior to those of the ADC map in all the data sets (all P<0.001). DSC, dice similarity coefficient; VS, volumetric similarity; HD, Hausdorff distance; AD, average distance; T2WI, T2-weighted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

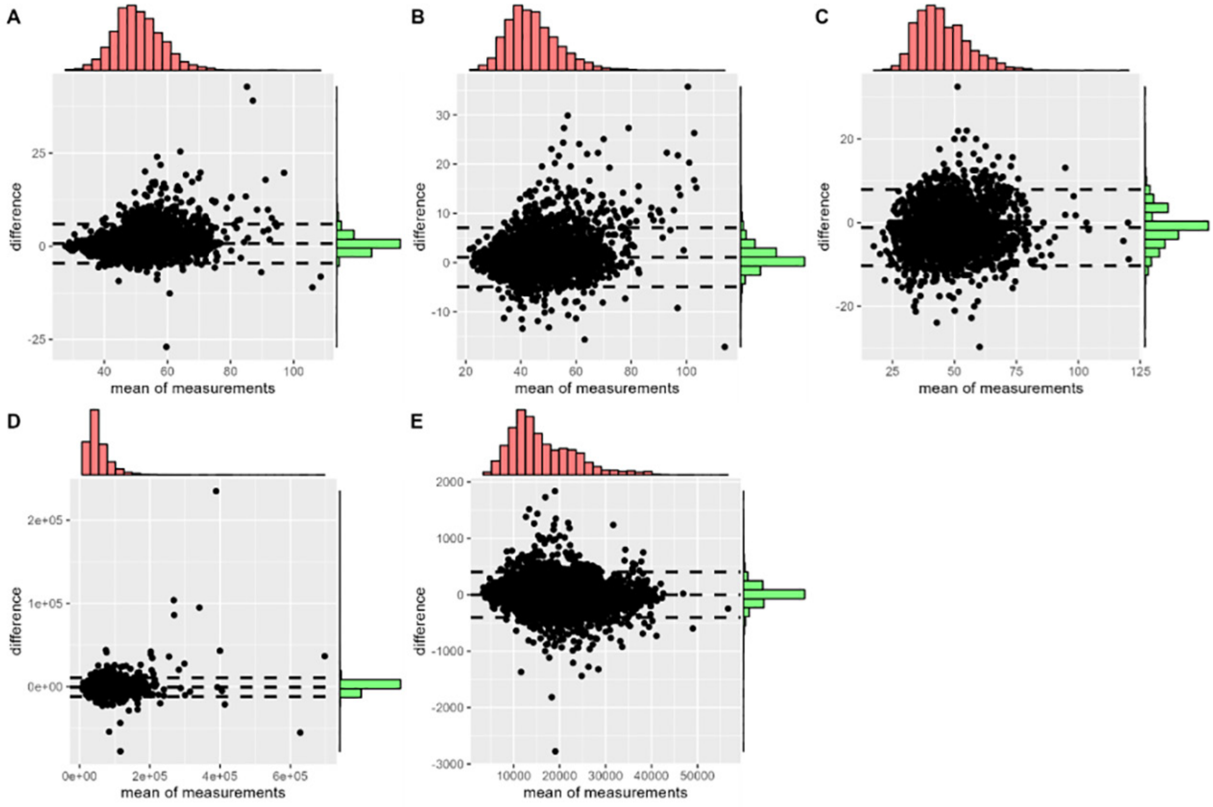


Figure S5 Bland-Altman analysis of the values of the RL diameter (A), AP diameter (B), SI diameter (C), volume (D), and signal intensity (E) of the manual label and the predicted label of the prostate gland. RL, right and left; AP, anterior and posterior; SI, superior and inferior.

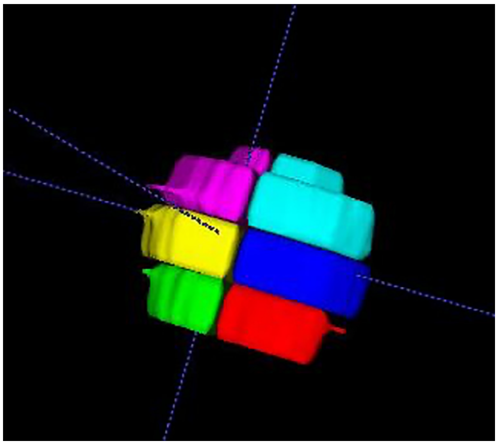


Figure S6 Sextant locations.

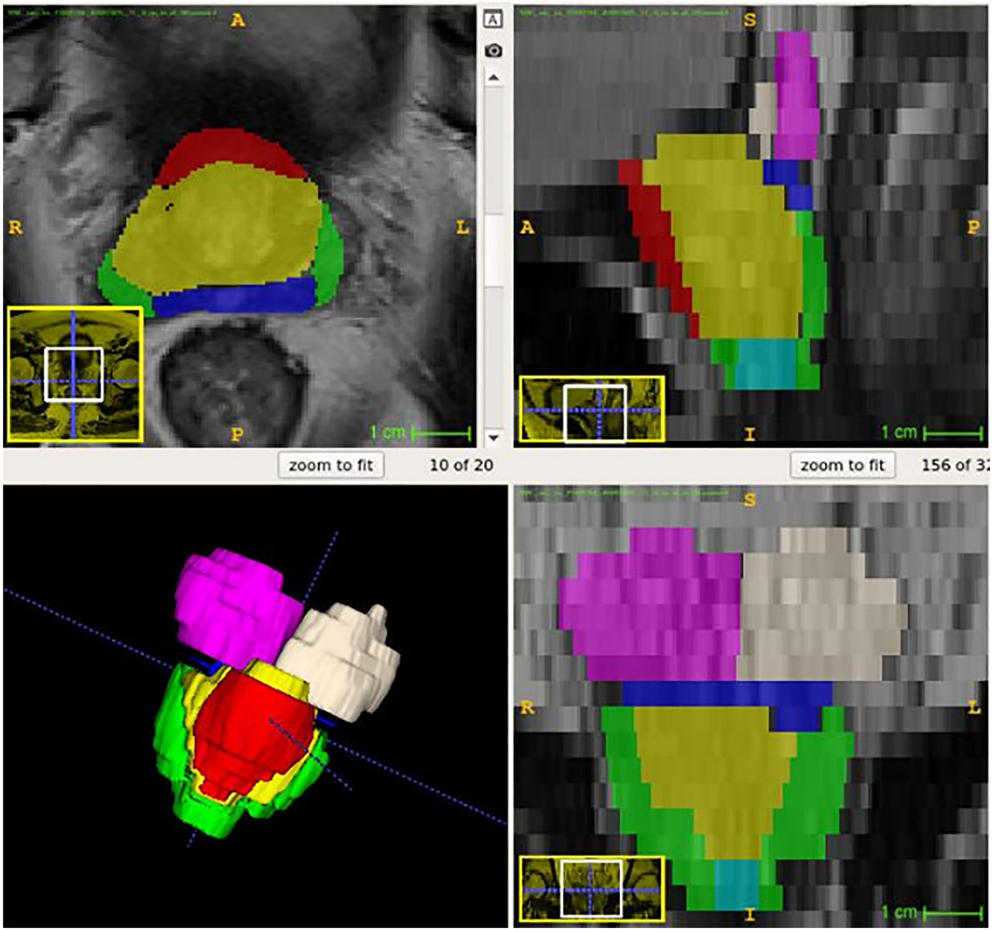


Figure S7 Anatomic zone locations.