European Radiology (2023) 33:5087–5096
https://doi.org/10.1007/s00330-023-09410-9

European Radiology
EUROPEAN SOCIETY OF RADIOLOGY

IMAGING INFORMATICS AND ARTIFICIAL INTELLIGENCE

Automated prostate multi-regional segmentation in magnetic resonance using fully convolutional neural networks

Ana Jimenez-Pastor[1] · Rafael Lopez-Gonzalez[1,2] · Belén Fos-Guarinos[1] · Fabio Garcia-Castro[1] · Mark Wittenberg[1] · Asunción Torregrosa-Andrés[3] · Luis Marti-Bonmati[3] · Margarita Garcia-Fontes[4] · Pablo Duarte[4] · Juan Pablo Gambini[4] · Leonardo Kayat Bittencourt[5,6] · Felipe Campos Kitamura[6] · Vasantha Kumar Venugopal[7] · Vidur Mahajan[7] · Pablo Ros[8] · Emilio Soria-Olivas[2] · Angel Alberich-Bayarri[1]

Received: 1 August 2022 / Revised: 6 November 2022 / Accepted: 27 December 2022 / Published online: 24 January 2023

Abstract

Objective Automatic MR imaging segmentation of the prostate provides relevant clinical benefits for prostate cancer evaluation such as calculation of automated PSA density and other critical imaging biomarkers. Further, automated T2-weighted image segmentation of central-transition zone (CZ-TZ), peripheral zone (PZ), and seminal vesicle (SV) can help to evaluate clinically significant cancer following the PI-RADS v2.1 guidelines. Therefore, the main objective of this work was to develop a robust and reproducible CNN-based automatic prostate multi-regional segmentation model using an intercontinental cohort of prostate MRI.

Methods A heterogeneous database of 243 T2-weighted prostate studies from 7 countries and 10 machines of 3 different vendors, with the CZ-TZ, PZ, and SV regions manually delineated by two experienced radiologists (ground truth), was used to train ($n = 123$) and test ($n = 120$) a U-Net-based model with deep supervision using a cyclical learning rate. The performance of the model was evaluated by means of dice similarity coefficient (DSC), among others. Segmentation results with a DSC above 0.7 were considered accurate.

Results The proposed method obtained a DSC of 0.88 ± 0.01, 0.85 ± 0.02, 0.72 ± 0.02, and 0.72 ± 0.02 for the prostate gland, CZ-TZ, PZ, and SV respectively in the 120 studies of the test set when comparing the predicted segmentations with the ground truth. No statistically significant differences were found in the results obtained between manufacturers or continents.

Conclusion Prostate multi-regional T2-weighted MR images automatic segmentation can be accurately achieved by U-Net like CNN, generalizable in a highly variable clinical environment with different equipment, acquisition configurations, and population.

Key Points

- *Deep learning techniques allows the accurate segmentation of the prostate in three different regions on MR T2w images.*
- *Multi-centric database proved the generalization of the CNN model on different institutions across different continents.*
- *CNN models can be used to aid on the diagnosis and follow-up of patients with prostate cancer.*

Keywords Prostate · Magnetic resonance imaging · Deep learning · Artificial intelligence · Diagnosis computer-assisted

✉ Ana Jimenez-Pastor
anajimenez@quibim.com

[1] Quantitative Imaging Biomarkers in Medicine (Quibim S.L.), Aragon Avenue, 30, 13th floor, Office I-J, 46021 Valencia, Spain

[2] Intelligent Data Analysis Laboratory (IDAL), ETSE, Universidad de Valencia, Valencia, Spain

[3] Department of Radiology and GIBI230 Research Group On Biomedical Imaging, Hospital Universitario Y Politécnico de La Fe and Instituto de Investigación Sanitaria La Fe, Valencia, Spain

[4] Centro Uruguayo de Imagenologia Molecular (CUDIM), Montevideo, Uruguay

[5] University Hospitals Cleveland Medical Center, Cleveland, USA

[6] DasaInova, Dasa, Unifesp, São Paulo, Brazil

[7] Mahajan Imaging, New Delhi, India

[8] Stony Brook University, Stony Brook, NY, USA

Abbreviations

ANOVA	Analysis of variance
CADe	Computer-assisted detection
CLR	Cyclical learning rate
CNN	Convolutional neural network
CZ	Central zone
DL	Deep learning
DRE	Digital rectal examination
DSC	Dice score coefficient
MAD	Mean absolute distance
PCa	Prostate cancer
PG	Prostate gland
PSA	Prostate-specitic antigen
PSAD	Prostate-specitic antigen density
PZ	Peripheral zone
SV	Seminal vesicles
TRUS	Transrectal ultrasound
TZ	Transition zone
ΔV	Volume difference

Introduction

Magnetic resonance imaging (MRI) is widely used for prostate cancer (PCa) detection, localization, and diagnosis [1], as digital rectal examination (DRE), prostate-specific antigen (PSA) testing, and blind biopsy have some limitations [2]. PSA protein is produced by both cancerous and noncancerous prostate tissue [3], being non-specific and usually not suitable for rendering a final diagnosis. Transrectal ultrasound (TRUS) blind biopsy might miss clinically significant tumors [4]. In recent years, MRI-guided and MRI/US-fusion-targeted biopsies have been developed, with a growing number of commercial devices allowing the multiparametric capability of MRI to target lesions for sampling. MRI/US fusion biopsies blend the advantages of real-time guidance by US with the benefits of specific targeting enabled by MRI [5].

MRI entails a significant diagnostic workload for radiologists. Prostate MRI reporting will benefit of the automated analysis of the gland to quantify volumes for indicators such as PSA density and tumor/prostate ratio with advanced computer-assisted detection (CADe) applications to help the radiologist in identification, location, and scoring of lesions. A robust and widely validated automated segmentation algorithm would provide efficiency in prostate and tumor volume calculations, streamlining of MRI/US fusion procedures and radiotherapy contours for dose painting [6, 7].

The differentiation between the central-transition zone (CZ-TZ), peripheral zone (PZ), and seminal vesicles (SVs) enables further development of clinical assessment tools following PI-RADS v2.1 [8], since PIRADS considers lesion location in scoring of suspected aggressive cancer. Additionally, prostate segmentation can be used for the automated extraction of imaging biomarkers such as apparent diffusion coefficient from diffusion-weighted imaging (DWI) and K^{trans} from dynamic contrast-enhanced (DCE) sequences, used to evaluate potential tumoral regions [9, 10].

Automating the segmentation of the prostate presents a challenge due to the intra-patient, inter-patient, and inter-equipment variability present in MRI acquisitions. Recent developments in the field of deep learning (DL) can significantly improve the performance of automatic prostate segmentation using convolutional neural networks (CNN) [11], as well as in the differentiation of the CZ-TZ and PZ [12, 13]. To date, there has been no approach for validating CZ-TZ and PZ segmentation across multiple institutions, countries, and continents to support an accurate generalization. Furthermore, the inclusion of SV can help radiologists evaluate PCa invasiveness beyond the prostate gland. The validation of segmentation of CZ-TZ, PZ, and SV would add important insights into the use of CNN-based automated prostate segmentation for accelerating radiation therapy dose planning.

We propose a deeply supervised CNN-based automatic prostate multi-regional segmentation method to identify and differentiate CZ-TZ, PZ, and SV, validated across different institutions, countries, and continents.

Materials and methods

This highly heterogeneous datasets were standard-of-care prostate MRI examinations collected from 7 different countries across 3 different continents to address data variability (minimize overfitting) and external validation challenges. The total number of MRI studies gathered was 243. The dataset was collected retrospectively through an observational study approved by the Ethics Committee at every institution and waived from informed consent collection. The age of the subject cohort was within the range of 25 to 92 years old. A balance (50/50%) between non-clinically significant (PIRADS < 3) and clinically (PIRADS ≥ 3) significant prostate cancer patients was found in the dataset (121 non-clinically significant vs. 122 clinically significant, respectively). The latter was built by 18% PIRADS 3 (22 cases), 25% PIRADS 4 (30 cases), and 57% PIRADS 5 (70 cases). The studies were acquired from 10 different MRI scanner models from 3 vendors using a wide variety of acquisition parameters (Table 1). The cases were split into training set, validation set, and testing set (Fig. 1), and the volumes of the regions segmented for each dataset partition were measured (Table 2). An analysis of variance was conducted to check for differences across datasets. No statistically significant differences were found between the training, validation, and test datasets in terms of volume for both CZ-TZ and PZ and for the whole prostate gland (PG), which considers both CZ-TZ and PZ. However, statistically significant differences were found for the SV.

Table 1 Ranges of the technical parameters of the T2-weighted sequence included in the database

TECHNICAL PARAMETERS T2-WEIGHTED SEQUENCE	
SEQUENCE	Fast-spin echo (FSE)/turbo-spin-echo (TSE)
MAGNETIC FIELD	1.5–3 T
PIXEL SIZE	[0.28–0.94] mm
ACQUISITION MATRIX	[220–512] × [160×512] mm
SPACING BETWEEN SLICES	[2.1–6.5] mm
GAP	[0–2] mm
SLICE THICKNESS	[2.1–6] mm
TR	[2020–13,634] ms
TE	[80–170] ms
FLIP ANGLE	[90–160]◦

Training and tuning datasets

The training set included 123 transverse T2-weighted MR prostate examinations acquired from different MRI scanners with multiple acquisition protocols received and segmented between April 2017 and October 2019. The origin of the training database is described in Table 3. This dataset was split, leaving 88% (108 studies) for training purposes and the remaining (15 studies) for the tuning of the training hyperparameters. The proposed solution is based on 2D images; therefore, each single slice from each subset was used for training (2468 images) or tuning (363 images).

Manual delineations of the CZ-TZ, PZ, and SV (Fig. 2) were performed by a team of two radiologists, each with at least 20 years' experience on pelvic MR, from the same center (University and Polytechnic Hospital La Fe in Valencia, Spain). When there were significant differences between the delineations, both radiologists agreed the consensus prostate segmentation mask considered the ground truth to train the model. All annotations were performed on a slice-by-slice basis using ITK-SNAP tool [15].

Testing dataset

The testing set included 120 T2-weighted MR images (3168 2D images) received and segmented between November 2019 and April 2020. No patients overlapped across the training and testing sets. The delineations of CZ-TZ, PZ, and SV were performed by the same 2 radiologists following the identical

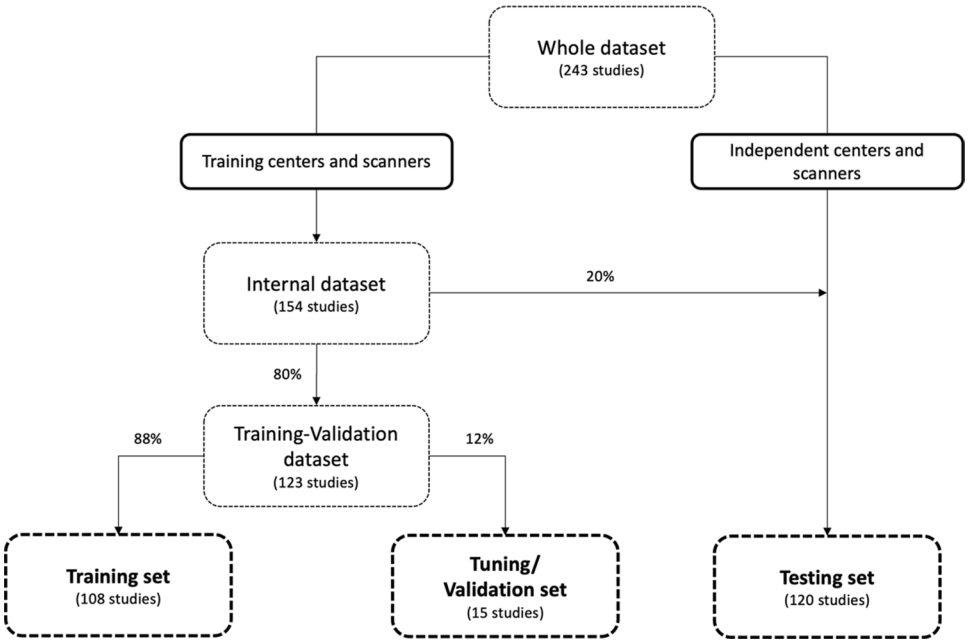

Fig. 1 Dataset for training, validation and testing of the proposed model

Table 2 Volumes of the regions in each database partition. * statistical significant differences across datasets

REGIONAL VOLUMES (mL)				
	CZ-TZ	PZ	SV	PG
Train	37.72 ± 85.64	17.74 ± 13.35	15.27 ± 14.97	55.46 ± 84.77
Validation	25.68 ± 20.93	18.35 ± 9.70	15.01 ± 15.06	44.03 ± 25.24
Test	39.06 ± 67.11	17.02 ± 18.37	11.73 ± 10.71	56.08 ± 73.91
p-value	0.44	0.72	0.00*	0.53

Table 3 Description of the sources used for the training-validation and test of the proposed model. For each source the manufacturer and magnetic fields are specified. Number of cases (*n*) refers to the number of cases in the training-validation/test datasets. *Promise12 public dataset [14]

ID	*n*	Origin	*n*	Manufacturer	Magnetic field
#1	26/6	Valencia, Spain	4/2	GE (Signa HDxt, Optima MR360)	1.5
			21/3	GE (Signa HDxt)	3
			1/1	Siemens (MAGNETOM Symphony)	3
#2	20/9	Montevideo, Uruguay	20/9	GE (DISCOVERY MR750w)	3
#3	17/5	Brazil	3/1	GE (Signa HDxt)	1.5
			4/0	GE (Signa Pioneer)	3
			1/0	Philips (Achieva)	1.5
			3/0	Siemens (MAGNETOM Avanto)	1.5
			6/4	Siemens (MAGNETOM Verio)	3
#4	18/10	New Delhi, India	6/2	GE (Signa HDxt)	1.5
			4/2	GE (DISCOVERY MR750w)	3
			7/6	Philips (Ingenia)	3
			1/0	Siemens (MAGNETOM Symphony)	1.5
#5*	1/0	Boston, Massachusetts, US	1/0	GE	3
#6*	41/1	Amsterdam, The Netherlands	41/1	Siemens	3
#7	0/89	New York, US	0/42	GE (Signa Pioneer)	3
			0/47	Siemens (MAGNETOM Prisma_fit)	3

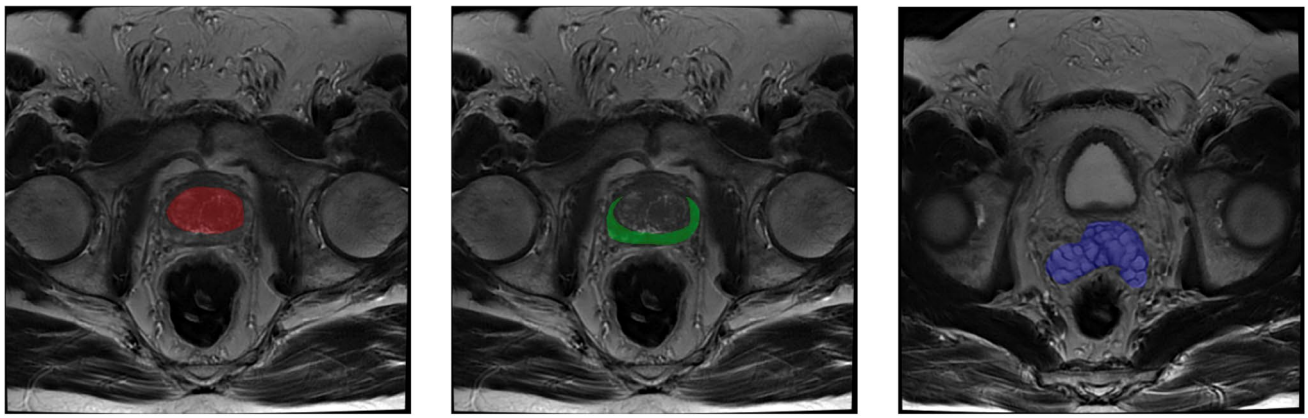

Fig. 2 Ground truth labels manually segmented by an expert. Three regions were segmented in the images: CZ-TZ (red), PZ (green), and SV (blue)

procedure described for the training and tuning sets. The origin of the test database is described in Table 3. A New York medical institution was added for external validation; cases from this institution were not included in the training dataset.

Network architecture

An architecture based in U-Net [16] was used in this experiment. To improve the performance of the model, a deep supervision stage [17] was added which applied a 1×1 convolutional filter to the outputs of the decoder blocks to combine them with the output of the network to force earlier layers to provide activation maps closer to the desired segmentation (Fig. 3). To train this network, the inverse of the Dice score coefficient (DSC) was used as cost function and Adam as optimization algorithm [18] along 300 epochs dividing the data in batches of 100 2D input slices of 256×256 and their respective segmentation masks of $256 \times 256 \times 4$ (one channel per class).

Training process

Pre-processing

To achieve a balance between resolution and computational load, all the 2D images were resized to common size of 256×256. Then, all the images were intensity-normalized to the interval [0, 1] to reduce intensity range, which is required by CNNs. During training, random data augmentation transformations were

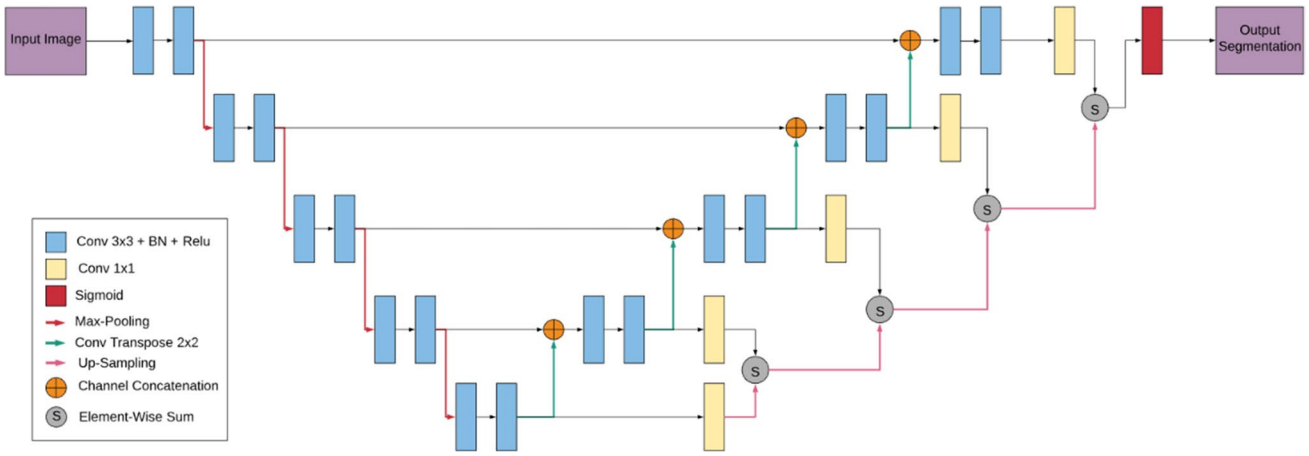

Fig. 3 Deeply supervised U-Net architecture

used to produce slightly different batches of data to improve the model's performance and generalization [19]. The transformations applied to the images consisted of Gaussian noise addition of standard deviation varying in the interval [0, 0.08], rotation in the range of [−15, 15] degrees and left–right flip. The probability that a specific transformation was applied to a given image was 30% in each data augmentation iteration.

Cyclical learning rate

To avoid the time-consuming task of learning rate fine-tuning, a cyclical learning rate [20] methodology was used. To find the reasonable bounds of the cycle, the model was trained with a few epochs using incremental learning rates. The lower bound was considered the learning rate where the loss function started to decrease and the upper bound the learning rate that produced the minimum value of the loss function [20].

Post-processing

To avoid randomly mislabeled pixels and background-labeled pixels, morphological operators [21] were applied, consisting of a per-class opening and a black hat to the combination of the CZ-TZ and PZ masks, that was used to find the background-labeled pixels, which were assigned as CZ-TZ pixels afterwards (Fig. 4).

Performance metrics evaluation

To evaluate the performance of the CNN, complementary boundary-based, regional overlap-based, and volume-based metrics were evaluated using the 120 studies that composed the test set. The mean absolute distance was used as the boundary-based error metric; the 3D DSC, recall rate, and precision rate as regional overlap-based error metrics; and the volume difference (ΔV) in cm^3 as volume-based metric. These metrics were used to compare the automated multi-regional segmentation against the manual segmentation. The metrics were extracted independently for the CZ-TZ, PZ, SV, and PG. To evaluate the generalization of the model, the metrics were obtained for studies of the test set as whole and, independently, for the studies of the test set acquired in the USA (not used for model training) and for the studies of the test set acquired with scanners from the different manufacturers (GE, Siemens, and Philips). To analyze for statistically significant differences between the USA cohort and the rest, the Mann–Whitney U test was used based after evaluating the non-Gaussian distribution of the obtained metrics with the Shapiro–Wilk test. Finally, to compare results between the three manufacturers, the Kruskal Wallis test was conducted.

Experiments and results

Cyclic learning rate

The obtained learning rate cycle bounds were 10^{-4} and 10^{-2}. The method chosen to vary the learning rate cyclically was a triangular function that scaled its peak with a gamma factor of 0.99994 each time 8 batches of training data were passed through the network (Fig. 5).

Metrics

The metrics obtained for the whole test set can be seen in Table 4. The PG was segmented with an average DSC of 0.88 ± 0.01 (DSC of 0.85 ± 0.02 for the CZ-TZ and of 0.72 ± 0.02 for the PZ), and the SV was segmented with a DSC of 0.72 ± 0.02. For the results presented in Table 4, similar results can be observed between the ones obtained over the whole testing dataset and those from the subset of the USA (0.88 ± 0.01 DSC in PG segmentation). No statistically significant differences ($p > 0.05$) were found between the metrics in the USA cases and the ones obtained in the rest of the test dataset.

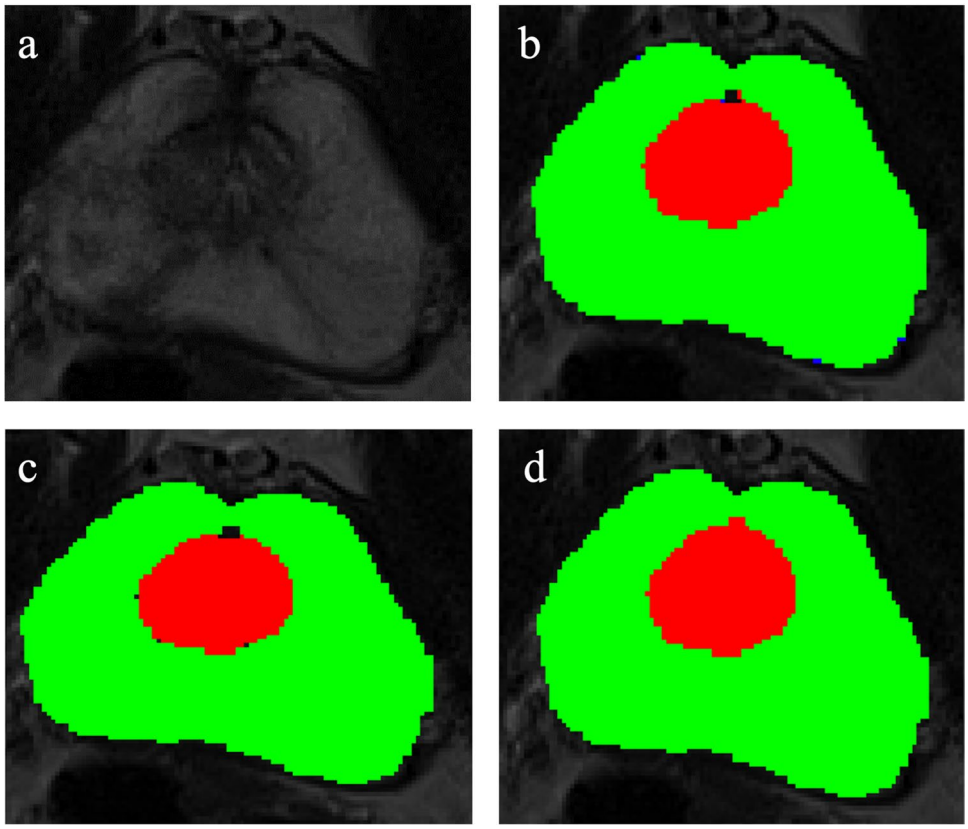

Fig. 4 Segmentation mask prost-processing process. **a** Prostate T2-weighted MRI without segmentation mask. **b** Output segmentation of the proposed model, some noisy segmented voxels can be found in the border areas. **c** Output segmentation after removing the noisy segmented regions. **d** Output segmentation with closed borders (final segmentation)

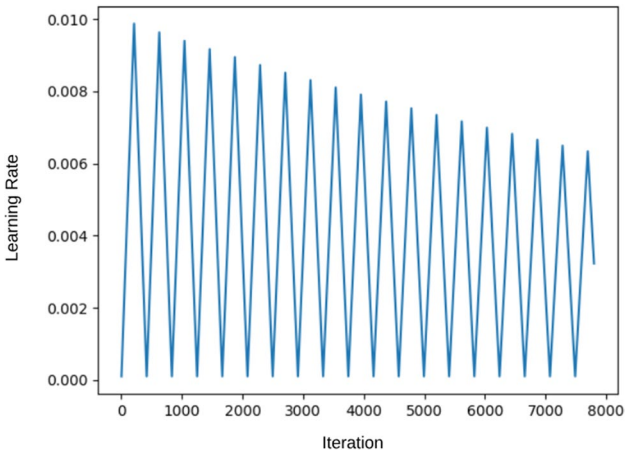

Fig. 5 Cyclical learning rate evolution over model training, each iteration is a forward–backward evaluation of a training set batch

When evaluating the results over scans from different manufacturers, the best results were obtained for images acquired with Siemens scanners (0.88 ± 0.02 DSC in PG segmentation), obtaining slightly lower results with Philips scanners (0.86 ± 0.07 DSC in PG segmentation). No statistically significant differences ($p > 0.05$) were found between the results obtained among the three manufacturers.

Figure 6 shows the results obtained in 2 test patients, one showing the case where the worst DC for the PG segmentation was obtained and another one where a good result was obtained for all the labels.

Execution time

The experiments were executed on a high computing performance node with 2 processors Intel SkyLake 6132 Gold 2.6 GHz with 14 cores, 2 GPUs NVIDIA QUADRO GP100 of 16 GB, and 192 GB of RAM memory running CentOS 7. TensorFlow was the deep learning framework used to run the experiments, and Keras was used as interface to speed up the code development. During the training process, the average epoch time was 44 s.

The inference time was measured on a MacBook Pro of 2018 with an Intel Core i7 2.6 GHz with 6 cores, a GPU Radeon Pro 560X of 4 GB, and 16 GB of RAM memory running macOS Big Sur. The average inference time over the test set was 0.46 s/slice including the pre-processing and post-processing.

Discussion

An automated prostate multi-regional segmentation algorithm based in a fully convolutional CNN with T2-weighted MRI as the input was proposed and externally validated in the present

Table 4 Metrics obtained with the automatic prostate multi-regional segmentation CNN in the studies of the whole test set, the studies acquired in the USA, the test studies acquired with GE equipment, the test studies acquired with Siemens equipment, and the test studies acquired with Philips equipment. For each metric the mean ± standard deviation values are provided

Region	DSC	Precision	Recall	$\lvert\Delta V\rvert$ [cm^3]	MAD [in mm]
Whole test dataset ($n = 120$)					
CZ-TZ	0.85 ± 0.02	0.87 ± 0.02	0.85 ± 0.02	3.33 ± 0.67	2.01 ± 0.17
PZ	0.72 ± 0.02	0.75 ± 0.02	0.72 ± 0.03	4.02 ± 0.94	2.23 ± 0.30
SV	0.72 ± 0.02	0.68 ± 0.03	0.81 ± 0.03	3.65 ± 0.49	2.75 ± 0.40
PG	0.88 ± 0.01	0.89 ± 0.01	0.87 ± 0.02	4.98 ± 0.91	1.92 ± 0.14
US subset ($n = 89$)					
CZ-TZ	0.85 ± 0.02	0.87 ± 0.02	0.85 ± 0.02	3.21 ± 0.72	1.97 ± 0.17
PZ	0.73 ± 0.02	0.73 ± 0.03	0.76 ± 0.03	3.04 ± 0.69	2.44 ± 0.39
SV	0.70 ± 0.03	0.65 ± 0.04	0.82 ± 0.03	3.89 ± 0.56	3.08 ± 0.51
PG	0.88 ± 0.01	0.88 ± 0.01	0.88 ± 0.02	4.80 ± 0.96	2.03 ± 0.17
GE subset ($n = 61$)					
CZ-TZ	0.85 ± 0.02	0.86 ± 0.02	0.84 ± 0.03	3.24 ± 0.86	1.93 ± 0.25
PZ	0.70 ± 0.04	0.76 ± 0.03	0.70 ± 0.05	4.24 ± 1.19	2.06 ± 0.45
SV	0.70 ± 0.03	0.68 ± 0.04	0.80 ± 0.04	3.48 ± 0.62	2.48 ± 0.46
PG	0.87 ± 0.01	0.90 ± 0.01	0.86 ± 0.02	4.68 ± 1.16	1.72 ± 0.14
Siemens subset ($n = 53$)					
CZ-TZ	0.86 ± 0.02	0.87 ± 0.03	0.87 ± 0.03	3.28 ± 1.05	2.22 ± 0.25
PZ	0.73 ± 0.03	0.73 ± 0.03	0.76 ± 0.05	3.26 ± 1.05	2.58 ± 0.47
SV	0.72 ± 0.04	0.66 ± 0.04	0.84 ± 0.05	3.98 ± 0.88	3.38 ± 0.76
PG	0.88 ± 0.02	0.88 ± 0.02	0.89 ± 0.02	5.26 ± 1.54	2.30 ± 0.25
Philips subset ($n = 6$)					
CZ-TZ	0.79 ± 0.18	0.91 ± 0.05	0.76 ± 0.23	3.84 ± 2.58	1.65 ± 0.46
PZ	0.67 ± 0.19	0.76 ± 0.09	0.65 ± 0.22	4.29 ± 4.56	1.87 ± 0.59
SV	0.77 ± 0.11	0.83 ± 0.04	0.75 ± 0.16	1.84 ± 1.77	1.38 ± 0.15
PG	0.86 ± 0.07	0.93 ± 0.03	0.82 ± 0.10	5.62 ± 4.14	1.42 ± 0.29

manuscript. The model was built from studies across 10 different MRI systems from 7 different countries and 3 vendors.

It is broadly accepted in training DL robust models that datasets must be carefully built to represent variability within the domain of the problem. MRI properties are affected by multiple factors such as scanner model and technical acquisition parameters. Due to these factors and the accessibility of fast spin-echo T2-weighted MRI, this sequence was the preferred one for the development of the segmentation model over other sequences such as DWI and DCE. This variability has a significant impact in the performance of DL models if diversity is not represented in the training dataset due to the difficulties of gathering medical imaging data from multiple sources [11]. Additionally, to properly test DL models in the medical domain, an independent database must be collected [22]. The use of a non-exposed independent database to validate the algorithms ensures that their biases based on the acquisition machine manufacturer or acquisition parameters are minimized. Following these, highly heterogeneous intercontinental training and testing standard-of-care prostate MRI examination databases were collected to undertake this development.

As derived from the results obtained in the testing database, the prostate segmentation resulted in higher DSC, most likely due to its larger size, homogeneity, and roundness, whereas the SV provided the lower DSC due to its smaller size and larger variability across individuals, factors that have a significant impact in the DSC calculation. The performance of the proposed method for automated gland segmentation is comparable with manual performance by expert observers [23, 24]. We obtained DSC of 0.88 ± 0.01 in PG segmentation, whereas [23] obtained a DSC ranging from 0.74 to 0.90 (mean and confidence range were not provided) when comparing manual prostate segmentations in T2-weighted MR images among three expert observers, and [24] obtained a DSC of 0.74 ± 0.14 when comparing the segmentations between six readers. Therefore, the results obtained in this work are within the ranges to be used in clinical practice, since they are comparable with those obtained by experts. According to these previous references, we can consider, therefore, a threshold of DSC over 0.7 as an accurate segmentation result.

The performance of our prostate segmentation method was compared (Table 5) with the results of recent similar methods [6, 7, 12, 13, 25–27]. The best performing algorithm [12] had a DSC of 0.94, 0.91, and 0.77 for PG, CZ-TZ, and PZ, respectively; in comparison, we obtained a DSC of 0.88 ± 0.01, 0.85 ± 0.02, and 0.72 ± 0.01.

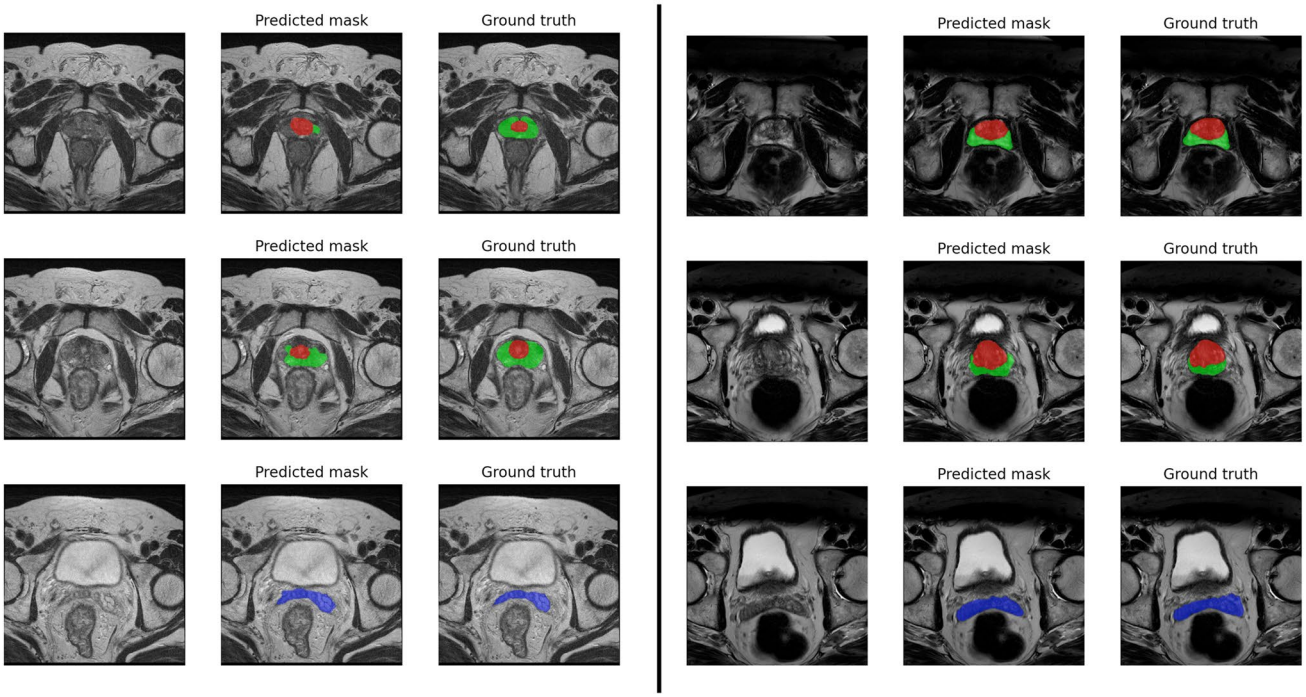

Fig. 6 Segmentation obtained in two test cases. In both cases, three different slices are shown at different prostate levels. For each slice, the original MR image (left), the predicted mask (middle), and the ground truth (right) are shown. The CZ + TZ is represented in red, the PZ in green, and the SV in blue. The left example shows the performance of the model in a case from the test set where the worst performance was observed in the PG segmentation (DSC = 0.64). The right example shows an example with good performance for all the labels

A direct comparison with these other proposed methods is difficult due to the test data variability in terms of number, quality, and size of the images. None of previous studies presented a database as large and heterogeneous as the one collected in this publication, nor did they include the segmentation of the SV. In the study [12], even though better results were obtained for all the structures, it is important to highlight that it was validated with only 48 cases coming from 2 MRI scanner models at one institution; therefore, a direct comparison between both methods should be made in future study.

Additionally, the training set of this work was acquired with the same equipment and acquisition protocols as the test set. The test set proposed in the present work is significantly larger, collected from several institutions, resulting in more robust DL models given the higher variability present. The test set used in our study represents more accurately the real-world data.

Our method presents some limitations that should be considered. As a 2D CNN has been used, the prostate regions are not evaluated as a whole 3D structure, but in a slice-by-slice manner which may have impact in the regularity of the segmentation along the axial axis. The size of the T2-weighted slices was decreased to 256 × 256 for memory and speed constraints, which may lead to information loss affecting the segmentation performance.

Table 5 Comparison of the proposed method with other prostate segmentation studies

	Test samples	PG	PZ	CZ-TZ
Proposed model	120	0.88 ± 0.01	0.72 ± 0.02	0.85 ± 0.02
Dai et al (2020) 6	12	0.86 ± 0.04	–	–
Da Silva et al (2020) 7	11	0.85	–	–
Bardis et al (2021) 12	48	0.94	0.77	0.91
Cheng et al (2019) 13	29	0.92	–	0.90
Zavala et al (2020) 25	55	0.89 ± 0.04	0.83 ± 0.11	–
Lee et al (2020) 26	70	0.87	0.76	–
Khan et al (2020) 27	51	–	0.79 ± 0.02	0.93 ± 0.00

Furthermore, the database used for the present work is composed of 243 studies coming from different sources with high variability, which is a very large size for the medical domain and one of the main values of the present work; but on the other hand, it is a moderate size for DL approaches; therefore, increasing the size of the database will lead to better results. Finally, further research and

validations should be conducted to compare our automatic method against expert manual segmentation to probe that our methods give comparable results.

In summary, prostate multi-regional T2-weighted segmentations can be properly achieved in MRI studies by U-Net like CNNs. The obtained results are generalizable in the highly variable clinical environment with different equipment, acquisition configurations, and patient's population. The results demonstrate that DL approaches are suitable to not just segment the prostate gland but to differentiate the CZ-TZ and PZ that compose it and additionally identify the SV. This automatic approach can aid radiologists in the evaluation of prostate cancer behavior following the PIRADS guidelines, and accurately evaluate clinically significant cancer in the prostate following the PIRADS v2.1 guidelines.

Funding The authors declare that this work has not received any funding.

Data Availability For data access, please, contact the authors of the manuscript.

Declarations

Guarantor The scientific guarantor of this publication is Angel Alberich-Bayarri.

Conflict of interest The authors of this manuscript declare relationships with the following companies: QUIBIM SL.

Statistics and biometry No complex statistical methods were necessary for this paper.

Informed consent Written informed consent was waived by the Institutional Review Board.

Ethical approval Institutional Review Board approval was obtained.

Methodology
- Retrospective
- Observational
- Multicenter study

References

1. De Visschere P (2018) Improving the diagnosis of clinically significant prostate cancer with magnetic resonance imaging. J Belg Soc Radiol 102(1):22
2. Cui T, Kovell RC, Terlecki RP (2016) Is it time to abandon the digital rectal examination? Lessons from the PLCO Cancer Screening Trial and peer-reviewed literature. Curr Med Res Opin 32(10):1663–1669
3. Mayo Clinic (2019) PSA test. Mayo Clinic. Available via https://www.mayoclinic.org/tests-procedures/psa-test/about/pac-20384731. Accessed 19 Jan 2023
4. Brown LC, Ahmed HU, Faria R et al (2018) Multiparametric MRI to improve detection of prostate cancer compared with transrectal ultrasound-guided prostate biopsy alone: the PROMIS study. Health Technol Assess 22(39):1–176
5. Das CJ, Razik A, Netaji A, Verma S (2020) Prostate MRI-TRUS fusion biopsy: a review of the state of the art procedure. Abdom Radiol (NY) 45(7):2176–2183
6. Dai Z, Carver E, Liu C et al (2020) Segmentation of the prostatic gland and the intraprostatic lesions on multiparametic magnetic resonance imaging using mask region-based convolutional neural networks. Adv Radiat Oncol 5(3):473–481
7. da Silva GLF, Diniz PS, Ferreira JL et al (2020) Superpixel-based deep convolutional neural networks and active contour model for automatic prostate segmentation on 3D MRI scans. Med Biol Eng Comput 58(9):1947–1964
8. Barentsz JO, Weinreb JC, Verma S et al (2016) Synopsis of the PI-RADS v2 guidelines for multiparametric prostate magnetic resonance imaging and recommendations for use. Eur Urol 69(1):41–49
9. Hötker AM, Mazaheri Y, Aras Ö (2016) Assessment of prostate cancer aggressiveness by use of the combination of quantitative DWI and dynamic contrast-enhanced MRI. AJR Am J Roentgenol 206(4):756–763
10. Sanz-Requena R, Martí-Bonmatí L, Pérez-Martínez R, García-Martí G (2016) Dynamic contrast-enhanced case-control analysis in 3T MRI of prostate cancer can help to characterize tumor aggressiveness. Eur J Radiol 85(11):2119–2126
11. Gillespie D, Kendrick C, Boon I (2020) Deep learning in magnetic resonance prostate segmentation: a review and a new perspective. ArXiv: 2011.07795
12. Bardis M, Houshyar R, Chantaduly C (2021) Segmentation of the prostate transition zone and peripheral zone on MR images with deep learning. Radiol Imaging Cancer 3(3)
13. Cheng R, Lay N, Roth HR (2019) Fully automated prostate whole gland and central gland segmentation on MRI using holistically nested networks with short connections. J Med Imaging (Bellingham) 6(2):024007
14. Litjens G, Toth R, van de Ven W et al (2014) Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. Med Image Anal 18(2):359–373
15. Yushkevich PA, Piven J, Hazlett HC et al (2006) User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31(3):1116–1128
16. Ronneberger O, Fischer P, Brox T (2015) U-Net convolutional networks for biomedical image segmentation. ArXiv: 1505.04597v1
17. Bo QZ, Turkbey B, Choyke PL (2017) Deeply-supervised CNN for prostate segmentation. ArXiv: 1703.07523
18. Kingma DP, Ba J (2017) Adam: a method for stochastic optimization. ArXiv: 1412.6980
19. Perez L, Wang J (2017) The effectiveness of data augmentation in image classification using deep learning. ArXiv: 1712.04621
20. Smith LN (2017) Cyclical learning rates for training neural networks. ArXiv: 1506.01186
21. Srisha R, Khan A (2013) Morphological operations for image processing: understanding and its applications. Conference: National Conference on VLSI, Signal processing & Communications
22. Kim DW, Jang HY, Kim KW (2019) Deep learning in magnetic resonance prostate segmentation: a review and a new perspective. Korean J Radiol 20(3):405–410
23. Shahedi M, Cool DW, Romagnoli C et al (2014) Spatially varying accuracy and reproducibility of prostate segmentation in magnetic resonance images using manual and semiautomated methods. Med Phys 41(11):113503
24. Brembilla G, Dell'Oglio P, Stabile A et al (2020) Interreader variability in prostate MRI reporting using Prostate Imaging Reporting and Data System version 2.1. Eur Radiol 30(6):3383–3392

25. Zavala-Romero O, Breto AL, Xu IR et al (2020) Segmentation of prostate and prostate zones using deep learning: A multi-MRI vendor analysis. Strahlenther Onkol 196(10):932–942
26. Lee DK, Sung DJ, Kim CS et al (2020) Three-dimensional convolutional neural network for prostate MRI segmentation and comparison of prostate volume measurements by use of artificial neural network and ellipsoid formula. AJR Am J Roentgenol 214(6):1229–1238
27. Khan Z, Yahya N, Alsaih K, Ali SSA, Meriaudeau F (2020) Evaluation of deep neural networks for semantic segmentation of prostate in T2W MRI. Sensors (Basel) 20(11):3183