AI-based diagnosis of prostate cancer from multiparametric MRI

Dimitri Hamzaoui

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

Thèse de Doctorat

Aide au diagnostic du cancer de la prostate depuis l'IRM multiparamétrique grâce à l'apprentissage profond.

Dimitri Hamzaoui

Inria, Équipe Epione

Thèse dirigée par Hervé Delingette, co-dirigée par Nicholas Ayache et co-encadrée par Raphaële Renard-Penna

Soutenue le 26 juin 2023

Présentée en vue de l'obtention du grade de Docteur en Automatique, Traitement du Signal et des Images de l'Université Côte d'Azur.

Devant le jury composé de :

Carole Lartizien	Directrice de Recherche à l'INSA Lyon	Rapporteur
Bjoern Menze	Professeur à l'Université de Zurich	Rapporteur
Marc-Michel Rohé	AI Tech Lead à Guerbet Research	Examinateur
Olivier Rouvière	PU-PH à l'Univ. Claude Bernard Lyon 1 & HCL	Examinateur
Hervé Delingette	Directeur de recherche au Centre INRIA d'Univ. Côte d'Azur	Directeur de thèse
Nicholas Ayache	Directeur de recherche au Centre INRIA d'Univ. Côte d'Azur	Co-directeur de thèse
Raphaële Renard-Penna	PU-PH à Sorbonne Université & AP-HP	Co-encadrante
Sarah Montagne	Praticienne hospitalière à l'AP-HP	Invitée

3iA Côte d'Azur Institut interdisciplinaire d'intelligence artificielle

ASSISTANCE PUBLIQUE HÔPITAUX DE PARIS

Inria

HEALTH DATA HUB

DIGITAL SYSTEMS FOR HUMANS

La vie m'a appris qu'il y a deux choses dont on peut parfaitement se passer, la prostate et la présidence de la République.

- George Clémenceau

Abstract

The objective of our work is the development of a method for the detection of prostate cancer from multiparametric MRI sequences. In this thesis, we detail the main sources of difficulties in the development of such a method as well as ways to overcome them.

Chapter 2 deals with the inter-rater variability of volume estimates and zonal segmentations of the prostate, two important factors for the establishment of the diagnosis and the construction of the databases necessary for the training of automatic methods. We exploit a database of 40 cases for which 7 radiologists of various levels have provided zonal segmentations as well as volume estimates. We evaluate their variations depending on the experience of the clinicians, the estimation methods used and some characteristics of the considered prostates. For the generation of segmentation masks, we show that variability is the highest at the apex and base of the prostate, and that it is independent of radiologists' experience. Furthermore, we show that the most robust volume estimation method for a prostate is to compute it directly from its segmentation.

In chapter 3, we introduce a new method to merge binary segmentation masks provided by several raters into a single consensus segmentation. The introduced MACCHIATO algorithm is based on the combination of local Fréchet means for well-chosen distances. It differs from the two main existing consensus determination methods (Averaging and STAPLE) on two points: contrary to averaging it is not computed at the voxel-level, and contrary to STAPLE it is independent of the background size. We exhibit the differences between the consensus produced by the three methods and show that our method can be placed between the two other methods with regards to consensus size. In addition, we make an in-depth analysis of the STAPLE algorithm and show its limitations, especially in case of large background size.

Chapter 4 presents a method based on deep neural network and attention mechanisms for the zonal segmentation of the prostate from 2D and/or 3D T2 MRI sequences. We evaluate our method on two databases and show that our method was on par with state-of-the-art methods for automatic segmentation and with the 7 available radiologists, being in the middle of the pack. Finally, we measure the impact that our method has on the determination of tumor location, both at the zonal and sector levels, with promising results on their localization accuracy.

In chapter 5 we study the influence of annotation quality and dataset size on a prostate cancer detection method. To this end, we develop two pseudo-labeling methods based on weak annotations of the lesion position extracted from radiological information: the

former based on prostate sector only, and the latter also including intensity and size information. The lesion detection method is a deep learning network taking as inputs biparametric MRI and zonal segmentation. This network was trained using each pseudo-labelling method on a large weakly annotated dataset, with or without the inclusion of a small amount of fully annotated cases. We compare those configurations at both patient and lesion levels to a network trained only on a fully annotated dataset.

Finally, we discuss areas of potential improvement and remaining challenges.

Keywords: medical imaging, segmentation, inter-rater variability, artificial intelligence, machine learning, consensus, prostate.

Résumé

L'objectif de notre travail est le développement d'une méthode de détection du cancer de la prostate à partir de séquences IRM multiparamétriques. Dans cette thèse, nous détaillons les principales sources de difficultés dans le développement d'une telle méthode ainsi que les moyens de les surmonter.

Le chapitre 2 traite de la variabilité inter-experts des estimations de volume et des segmentations zonales de la prostate, deux facteurs importants pour l'établissement du diagnostic et la construction des bases de données nécessaires à l'entraînement des méthodes automatiques. Nous exploitons une base de données de 40 cas pour lesquels 7 radiologues de différents niveaux ont fourni des segmentations zonales ainsi que des estimations de volume. Nous évaluons leurs différences en fonction de l'expérience des cliniciens, des méthodes d'estimation utilisées et de certaines caractéristiques des prostates considérées. Pour la génération des masques de segmentation, nous montrons que la variabilité est la plus élevée à l'apex et à la base de la prostate, et qu'elle est indépendante de l'expérience des radiologues. En outre, nous montrons que la méthode la plus robuste d'estimation du volume d'une prostate consiste à le calculer directement à partir de sa segmentation.

Dans le chapitre 3, nous présentons une nouvelle méthode pour fusionner les masques de segmentation binaires fournis par plusieurs annotateurs en une seule segmentation consensuelle. L'algorithme MACCHIATO repose sur la combinaison de moyennes de Fréchet locales pour des distances bien choisies. Il diffère des deux principales méthodes existantes de détermination du consensus (moyenne et STAPLE) sur deux points : contrairement à la moyenne, il n'est pas calculé au niveau du voxel, et contrairement à STAPLE, il est indépendant de la taille du fond. Nous présentons les différences entre les consensus produits par les trois méthodes et montrons que notre méthode peut être placée entre les deux autres méthodes en ce qui concerne la taille du consensus. En outre, nous effectuons une analyse approfondie de l'algorithme STAPLE et montrons ses limites, en particulier lorsque la taille du fond est importante.

Le chapitre 4 présente une méthode basée sur un réseau neuronal profond et des mécanismes d'attention pour la segmentation zonale de la prostate à partir de séquences d'IRM T2 2D et/ou 3D. Nous évaluons notre méthode sur deux bases de données et montrons qu'elle se situe au même niveau que les méthodes de l'état de l'art pour la segmentation automatique et que les 7 radiologues disponibles, avec des perfomances similaires à ceux-ci sans les surpasser. Enfin, nous mesurons l'impact de notre méthode

sur la détermination de la localisation des tumeurs, tant au niveau zonal que sectoriel, avec des résultats prometteurs sur leur précision de localisation.

Dans le chapitre 5, nous étudions l'influence de la qualité des annotations et de la taille de l'ensemble de données sur une méthode de détection du cancer de la prostate. À cette fin, nous développons deux méthodes de pseudo-étiquetage basées sur des annotations faibles de la position des lésions extraites des informations radiologiques : la première basée sur le secteur de la prostate uniquement, et la seconde comprenant également des informations sur l'intensité et la taille. La méthode de détection des lésions est un réseau d'apprentissage profond prenant comme entrées l'IRM biparamétrique et la segmentation zonale. Ce réseau a été entraîné à l'aide de chaque méthode de pseudo-étiquetage sur un vaste ensemble de données faiblement annotées, avec ou sans l'inclusion d'un petit nombre de cas entièrement annotés. Nous comparons ces configurations au niveau du patient et de la lésion à un réseau entraîné uniquement sur un ensemble de données entièrement annotées.

Enfin, nous discutons des domaines d'amélioration possibles et des défis qui restent à relever.

Mots-clés: imagerie médicale, segmentation, variabilité inter-expert, intelligence artificielle, apprentissage profond, consensus, prostate.

Funding

This work has been supported by the French government, through the 3IA Côte d'Azur and the Université Côte d'Azur DS4H Investments in the Future project managed by the National Research Agency (ANR) with the reference numbers ANR-19-P3IA-0002 and ANR-17-EURE-0004, and by the Clinical Data Warehouse of the AP-HP (Assistance Publique-Hôpitaux de Paris). The authors are grateful to the OPAL infrastructure from the Université Côte d'Azur for providing resources and support. We thank the French Health Data Hub for providing resources and support

Acknowledgements

I would like to thank all the people involved in the creation of this manuscript and the associated researches and experiments. I will not be able to name everyone, but even if your name is not explicitly cited, I do not forget you. First of all, I would like to thank my supervisors, Hervé Delingette and Nicholas Ayache, for their assistance during those 3 and a half years, for the constant exchanges of ideas and possibilities, for the proofreading of the manuscript and finally for giving me the possibility to work here on this projec with them. On the same level I also have to thank Raphaële Renard-Penna, who became an "unofficial" third supervisor who really accompanies me in my discovery of uro-radiology. I doubt I could have dreamed a better clinician to work with. Then, I would like to thank my reviewers, Carole Lartizien and Bjoern Menze, to have accepted this role. I would also like to thank Olivier Rouvière and Marc-Michel Rohé for begin part of my jury

In addition, I would like to thank Sarah Montagne, Julien Castelneau and Sébastien Molière for working with me on the PAIMRI project. I wish we had met in person more often than what we did, but it was already a pleasure to work with you. I would also like to thank Aurélien Maire et Yannick Jacob from the *Entrepôt des Données de Santé de l'AP-HP* for their assistance in the access to the MRI data and for the allowances of computational resources.

I would like to thank other permanents members of the Epione Team - Irene, Marco, Maxime and Xavier - for their advices during those years. I would also like to thank the other members of the team for the good times spend together: Elodie, Francisco, Hari, Jairo, Lucia, Luis, Riccardo, Tom, Victoria... with special thanks to Etrit, Hind, Paul, Santi, Yann and my office mate Zhijie.

I would also like to thanks the Wednesday's team - Rémi, Kevin, Corentin and Timothé, those evenings have been real breathing during this period. I thank my friends from Télécom Paris - especially Iann, Louison, Louis and Bernardo but I don't forget the others - for keeping hanging around with me despite the distance. I would also like to thank my friend from the LSTL team and also Anna and Anas for those good moments during my (too short) comebacks in Paris. I also thank Linda for her support during this PhD.

And finally, I would like to thank my family. They were always my first supporters, even in the difficult times and despite the distance, and I would not be there here without them.

Contents

1 Introduction

Contents

This thesis deals with the computer-aided diagnosis of prostate cancer through multi-parametric MRI via the use of deep learning methods.

1.1 Context of the thesis

1.1.1 Prostate cancer: clinical aspect and diagnosis methods

Prostate is a gland of the masculine reproductive system, playing a key role in the production of seminal fluid. Located under the urinary bladder and around the urethra, it can be decomposed into several zones: the central zone (CZ), the transition zone (TZ), the peripheral zone (PZ) and the anterior fibromuscular stroma (AFMS). An example of the prostate anatomy is provided in Fig. 1.1.
Prostate cancer (PCa) is one of the most frequent cancers in the world, especially in developed countries. Between 2015 and 2019, the incidence rate in the USA was estimated at 109.9 per 100,000 population, the second highest among all types of cancers behind breast cancer [Sie+23]. Therefore, it has been estimated that 1 American man over 8 will develop it during his lifetime. Similarly, it has been estimated that one quarter of cancer cases on French men was prostate cancer [Ins22]. This high prevalence makes it an important public health concern with major economic impacts [RB11], despite its high 5-year survival rate (97%) compared to other frequent cancers such as breast (91%), colorectal (65%) and lung (23%) cancers.

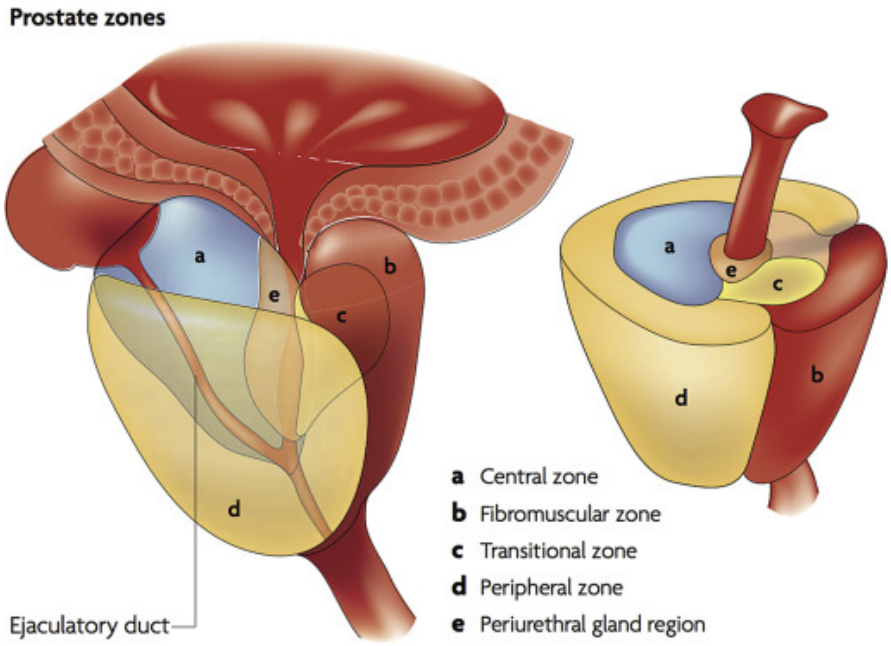


Fig. 1.1.: Prostate zonal anatomy. Reproduced with permission from Elsevier [Ree+16].

For years, the standard protocol for PCa detection has consisted of three steps. First, clinicians check family history and perform non-invasive, low-cost tests such as digital rectal exam (DRE) and measurement of the level of prostate-specific antigen (PSA) through blood tests. Then, if a cancer is suspected, imaging-based screening is performed to find possible abnormalities. And finally, if such an anomaly is detected, a biopsy is performed to confirm the diagnosis and characterize the cancer if found. This biopsy can either be targeted, i.e. the location of the extracted tissue sample is determined from the mpMRI sequence, or systematic, in which case several samples from different locations are extracted.
Originally, the recommended imaging technique for screening was transrectal ultrasonography (TRUS). However in the last decade, TRUS has been gradually replaced by multiparametric MRI (mpMRI), a combination of T2-weighted (T2W), diffusion-weighted (DWI and ADC), and dynamic contrast-enhanced (DCE) sequences. While more expensive, this new imaging protocol allows for targeted biopsies (less invasive for the patient) and reduces the number of benign lesions wrongly estimated as clinically significant (i.e. the number of false positives) [Rou+19; Kas+18; Ahm+17]. This advantage of mpMRI over TRUS made it the recommended system of screening by the European Association of Urology [Mot+21]. An example of the different sequences used in mpMRI can be seen in Fig. 1.3.

To standardize the acquisition, interpretation, and reporting of prostate mpMRI examinations, radiologists conceived in 2012 a set of radiological guidelines named as Prostate Imaging-Reporting and Data System (PI-RADS v1 [Bar+12]) , which was then updated in 2015 (PI-RADS v2 [Wei+16]) and in 2019 (PI-RADS v2.1 [Tur+19]. It first provides a standardized sector map of the prostate for the location of lesions, represented in Fig. 1.2.

In addition, it defines a five-point scale to assess the suspected severity of the lesion only based on observations of the mpMRI. The main innovation brought by PI-RADS v2 is the differentiated weight of each sequence according to the location of the suspected lesion. Lesions located in the TZ are mainly assessed using T2W sequences, whereas diffusion sequences will be prevalent on the PZ. DCE is only used to refine the grade for some specific PZ lesions. PI-RADS is only defined on those two zones since PZ and TZ represent the location of approximately 70% and 25% of prostatic cancerous lesions [McN+88; Wei+16] and due to of the absence of clear delineations between PZ and CZ. The impact of each sequence on the PI-RADS score is shown on Tab. 1.1.

When the mpMRI indicates a suspicion of cancerous lesion (corresponding to a PI-RADS score ≥ 3), the current method to confirm the diagnosis is to perform a biopsy. The diagnosis is then given as a score, either the Gleason score [Hum04] or the ISUP score [Eps+16]. The Gleason score ranges from 2 to 10 and is based on the two dominant cell patterns observed on the biopsy sample under the microscope. A score>6 is the usual threshold for malign lesions. The cell patterns used to determine the score are represented in Fig. 1.4a. This score is generally given as a sum, since the order of prevalence has an impact on the estimated severity of the lesion. For example, a cancerous lesion with a score of 4+3 is usually more aggressive than one with a score of 3+4. The ISUP score [Eps+16], equivalent to the Gleason Grade Group, is a 5-point scale derived from the Gleason score to simplify the interpretation of the diagnosis. The correspondence between ISUP and Gleason scores is available in Tab. 1.4b. For more information, the whole process of PCa assessment is detailed in Appendix A.

1.1.2 Computer-aided diagnosis methods for Prostate cancer and associated challenges

Computer-aided detection/diagnosis (CAD) solutions to detect PCa are currently investigated by several teams around the world in order to help radiologists in their diagnosis tasks. Several of those methods have already been approved by the competent authorities on their internal markets and are commercially available, a list of them being available in [TH22]. Currently their level of performance are not equivalent to those of human radiologists, but they can already provide them with assistance [Gig+23; Cac+23; Rou+22]. The global framework of those CAD methods is often similar [Lem+15; Cuo+19]: first, the mpMRI undergoes a preprocessing phase including normalization and registration of the different sequences as well as segmentation of the prostate gland from the MRI (generally using the T2W sequence). Then the detection/segmentation of tumoral lesions is performed and for the most advanced methods, a characterization of those lesions such as a prediction of their PI-RADS/Gleason score is also determined. Those methods correspond to a Computed-Aided Detection (CADe). Another family of CAD methods are methods processing lesions given to them by radiologists (Computer-aided diagnosis, CADx).

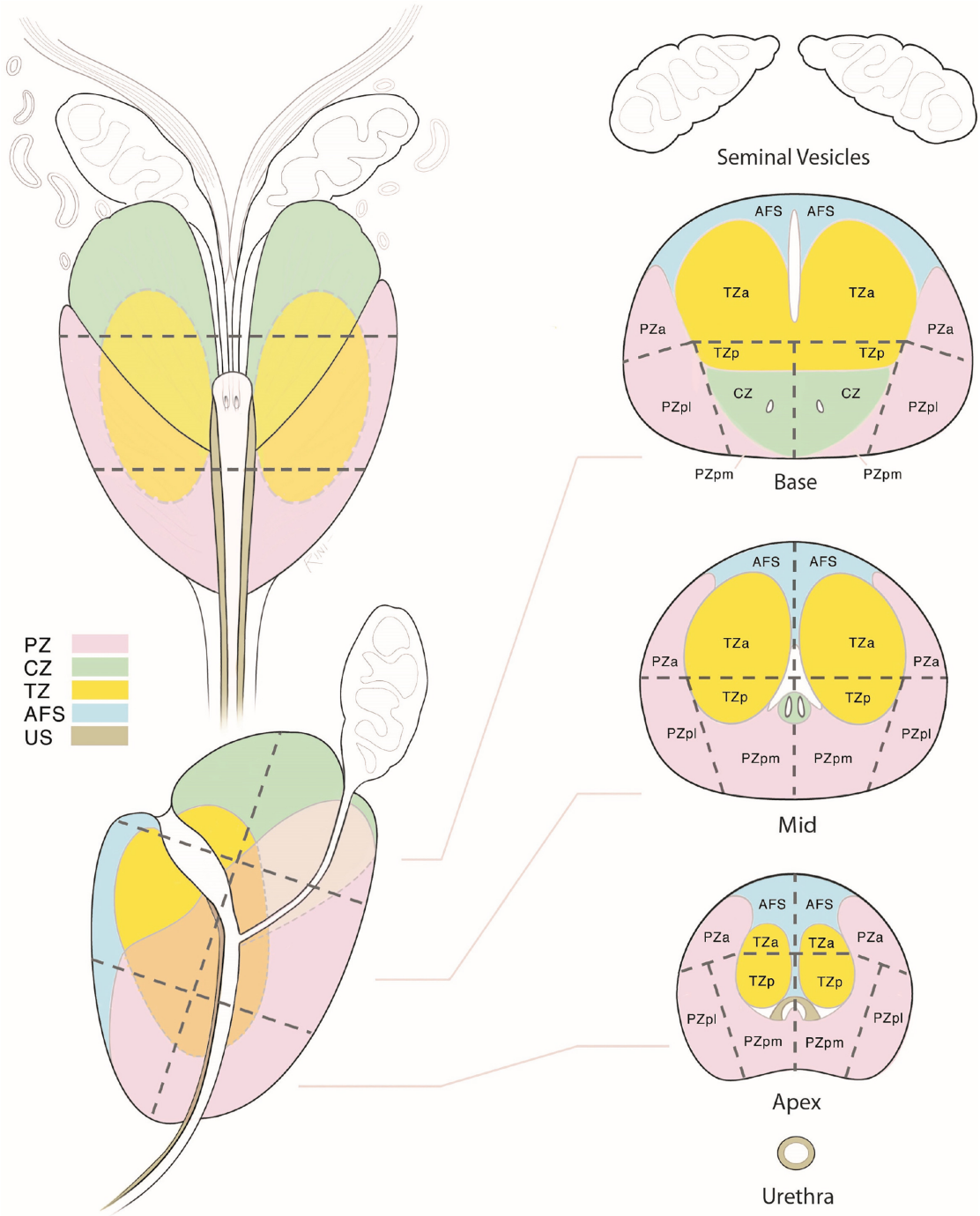
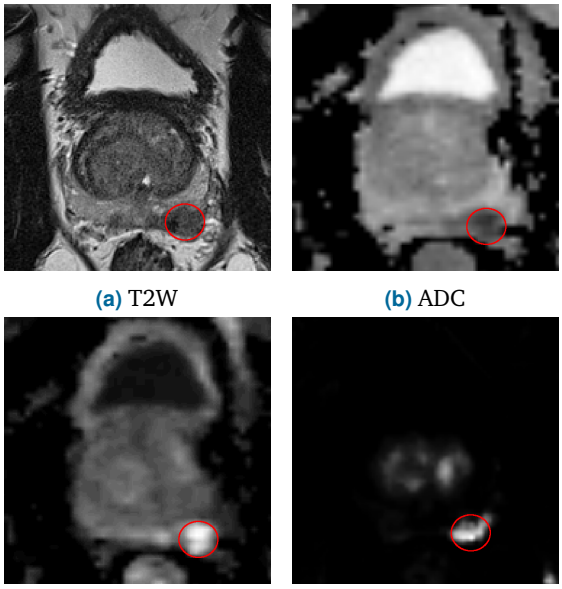


Fig. 1.2.: Prostate sector map as defined by PI-RADS v2.1. Prostate zones: PZ - Peripheral Zone; CZ - Central Zone; TZ - Transition Zone; AFS- Anterior Fibromuscular Stroma; US - Urethra.
Zones Subdivisions: a-anterior; p-posterior; pm-posteromedial; pl-posterolateral.
Reproduced with permission from Elsevier [Tur+19].



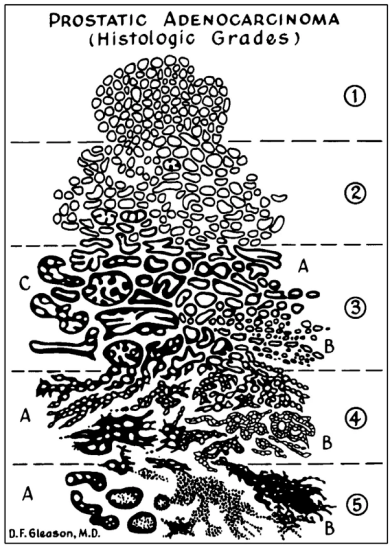
(a) T2W (b) ADC

(c) DWI (b-value = 800) (d) DCE

Fig. 1.3.: Example of mpMRI, with a lesion surrounded in red. Figure (d) represents Ktrans, a parameter computed from DCE sequences. Compared to the surrounding areas, the lesion is hypointense on the T2W and ADC sequences, and hyperintense on the other sequences.

Peripheral zone			
T2W	**DWI**	DCEI	PI-RADS v2
Any	1	Any	1
Any	2	Any	2
Any	3	-	3
		+	4
Any	4	Any	4
Any	5	Any	5
Transition zone			
T2W	DWI	DCEI	PI-RADS v2
1	Any	Any	1
2	Any	Any	2
3	≤ 4	Any	3
	5		4
4	Any	Any	4
5	Any	Any	5

Tab. 1.1.: Method of determination of PI-RADS v2 grade according to the zonal location as described in [Wei+16]. The grade is ranging from 1 (very low probability of PCa) to 5 (very high probability of PCa) and is computed on each sequence based on specific radiological criteria before merging them on a final grade.



(a) Cell patterns used to determine the Gleason score. Reproduced with permission from Springer Nature [Hum04].

ISUP 2014 score	Gleason Grade	Glandular aspects
1	≤ 6	Only individual discrete well-formed glands
2	3+4=7	Predominantly well-formed glands with lesser component of poorly-formed/fused/cribriform glands
3	4+3=7	Predominantly poorly formed/fused/cribriform glands with lesser component of well-formed glands
4	8	Only poorly formed/fused/cribriform glands *or* predominantly well-formed glands and lesser component lacking glands *or* predominantly lacking glands and lesser component of well-formed glands
5	9-10	Lack of gland formation (or with necrosis) with or without poorly formed/fused/cribriform glands

(b) Correspondence between the Gleason scores and the ISUP scores. Adapted from [Eps+16].

The majority of recent CAD methods are based on deep learning. Deep learning consists in training artificial neural networks on a large dataset. These neural networks are

composed of multiple layers of interconnected nodes, which are designed to process and analyze complex patterns in data. Each layer in a neural network transforms the input data in a way that makes it more suitable for the next layer to process, until the final layer produces the desired output. The process of training a neural network involves adjusting the weights and biases of the connections through stochastic gradient error back-propagation. One of the main families of neural networks are convolutional neural networks (CNNs), which are based on convolution operations to extract features from images. Deep learning-based techniques are being actively developed to address various stages of PCa detection, encompassing tasks from registering mpMRI with other imaging modalities (ultrasound and histopathology) to detecting tumors on histopathology images, including tumors on mpMRI scans [Bha+22].

1.2 Thesis overview

If the development of CAD methods for PCa detection is a rapidly expanding field, several limitations still exist restraining them from being widely exploited by clinicians in hospitals.

1.2.1 Challenges tackled in this thesis

Evaluation of rater variability for prostate cancer management Accuracy of the labels used in the training of such networks is important as poor labels will lead to ineffective methods. But on medical data, determining the ground truth can be hard [Ren+20], especially for a organ such as the prostate which has an important inter-subject variability and can have different intensity and morphological properties according to the considered zone of the organ. Thus, variations of prostate segmentations among raters can be observed [Mon+21; Bec+19] which in turn can impact lesion detection, especially for low-grade (PI-RADS=3) lesions [Smi+19; Gre+19; Mus+19] as well as their estimated location [Gre+18]. This leads to the following challenge: *what is the extent of inter-rater variability of the prostate delineation in mpMRI* and *what is its potential impact on label generation and clinical decision?*

Construction of consensus prostate segmentation masks A solution to mitigate this inter-rater variability effect is to combine the information provided by several experts into one consensus segmentation. For binary segmentation masks, Majority Voting and STAPLE [WZW04] are among the most common methods to create such a label. However, both of them have limitations : the former is computed at the voxel-level, the latter depends on the image size. From there, we can wonder *how to generate a consensus segmentation computed at the lesion-level and independent of the background size?*

Use of weak annotations in large training sets In most cases, deep learning methods are trained on datasets in a supervised way, i.e. which there is a reliable label for each sample. Those conditions can be easily met on natural images [Lin+15; Eve+15], but are harder to obtain on medical datasets for several reasons.
On the one hand, the creation and sharing of medical data is strictly regulated by GDPR [Cou16b] and national health authorities. This is particularly true for MRI data as the DICOM format used to store them can contain identifying metadata. In addition, collecting data coming from different hospitals requires to have the authorization from each of them separately, taking more time and with more administrative burden. This tends to restrict the possibility of models trained on large datasets. Moreover, even if such large datasets are available, manually segmenting prostate and/or lesions and, in the latter case, assessing their PI-RADS score can take several minutes, even for experienced radiologists [Ros+17], and thus doing so on a large scale is not realist. So, *can we use train PCa detection methods on weakly-annotated data, and what are the performance of such methods?*

1.2.2 Organization of the Thesis

This manuscript is organized as follows:
In Chapter 2, we study the inter-rater variability of prostate segmentation and prostate volume measurements. To do so, we used a database containing 40 instances for which three expert, two senior and two junior radiologists have submitted zonal segmentations and volume estimates. We assess these variabilities in light of the doctors' experience, the estimating techniques employed, and specific features of the prostates under consideration. Comparisons on segmentations were done using either pairwise metrics or metrics with respect to a consensus segmentation. We demonstrate that variability of segmentations is maximal at the apex and base of the prostate, and that it is unrelated to radiologists' level of experience. Those results have been obtained whatever which type of metrics were used. In addition, we demonstrate that computing a prostate's volume directly from its segmentation is the most reliable volume estimation technique. We published on this subject in several clinical journals [Mon+21; Ham+22a]. Appendix B extends this study by exploring the impact of the number of raters on the estimation of segmentation variability [Mol+23].

The chapter 3 focuses on the computation of binary masks consensuses by studying some classical methods and introducing a new one. First, we thoroughly examine the STAPLE method, a state-of-the-art method to compute segmentation consensus from binary masks, and demonstrate its drawbacks, particularly when the processed images have large background sizes. Second, we introduce a new method to compute a segmentation consensus that is independent of the background size while taking into account local context. This method, coined MACCHIatO, is based on the local Fréchet means

for Jaccard-based or Dice-based distances and thus is not impacted by background size. In addition, by design this method produces consensuses computed at a larger level than the voxel, contrary to averaging. We analyze how the three approaches' produced masks differ and show that, in terms of consensus size, our method can be positioned between the other two methods. This work has first been presented at MICCAI-UNSURE 2022 [Ham+22c] and then extended and submitted to MELBA - Journal of Machine Learning for Biomedical Imaging [Ham+23b].

In Chapter 4, we present a deep-learning based method exploiting attention mechanisms to produce zonal segmentation of prostate using 2D and/or 3D T2 MRI. We test our method on two databases and demonstrate that its level of performance was similar to those of state-of-the-art methods and not far from those of radiologists. Finally, we assess the influence of our approach on the localization of tumors at the zonal and sector levels, showing that the high zonal accuracy was encouraging in terms of clinical exploitation of those algorithms. This work as been published in Journal of Medical Imaging [Ham+22b]. An associated review of methods for automatic segmentation of the prostate, published in a peer-reviewed journal [Wu+22] is available in Appendix C

In Chapter 5, we introduce a new method to generate pseudo-masks of prostate lesions from radiological information and mpMRIs. We use the radiological information on the location and the size of the lesion to guide a intensity-based method to create a pseudo-mask of this lesion. We compare it with another pseudo-mask generation method only based on the location information provided by radiologists. We show that between the two pseudo-mask generation strategies, the intensity-based one produces results closest to ground truth lesion segmentations and can be used to train deep learning methods on large weakly-annotated datasets. Moreover, the obtained levels of performance are at least similar to those obtained while training on a smaller fully-annotated dataset. Finally, we show the benefit of mixed-supervision to improve generalization abilities [Ham+23c].

In Chapter 6, the main contributions of this thesis are summarized. Finally, potential future work and perspectives are drawn.

This thesis was conducted in partnership with the Radiology Department of the Pitié-Salpêtrière Hospital (Paris, France) and in collaboration with the APHP's *Entrepôt des Données de Santé* (EDS) and the *Health Data Hub* (HDH), France's national health data platform.

1.3 Datasets used during this PhD

During this thesis work we used several mpMRI datasets. In the following table we sum up the characteristics of those different datasets. For each of them, we provide:

- Their name
- #Cases: Number of MRIs
- #csPCa cases: Number of MRIs with at least one clinically significant lesion, as well as the criteria used to define the clinical significance: PI-RADS (radiological) or ISUP (histological)
- Available labels: Zonal segmentation of the prostate into TZ and PZ, precise segmentations of the lesions, ...
- The chapters in which they were used
- The public availability of the dataset
- Other information detailing the specificities of each dataset: number of involved raters, how the segmentations have been obtained...

Name	#Cases	#csPCa cases	Labels	Chapters	Accessibility	Other information
PAIMRI-RV*	40	17§	Zonal & Tumors segmentations	2, 3, 4, 5, B	Private	Number of raters: 7
PAIMRI-FA	160	63§	Zonal & Tumors segmentations	4, 5	Private	-
PAIMRI-WA	5290	2850§	Weak labels	4, 5	Private	Available labels: PI-RADS, sector, lesion diameter, PSAd Automatic zonal segmentations
ProstateX†[Arm+18]	204	76★	Zonal & Tumors segmentations	4, 5	Public	Segmentations by [Cuo+21b]
PI-CAI[Sah+22]	1500	425★	Zonal & Tumors segmentations	5	Public	Automatic zonal segmentations. Half of lesions automatically segmented

★: ISUP≥2; §: PI-RADS≥3
*: Included into PAIMRI-FA; †: Included into PI-CAI

Tab. 1.2.: Description of all the mpMRI prostate datasets used in this PhD thesis.

PAIMRI-RV was originally designed to study inter-rater variability and was then included into PAIMRI-FA for evaluation of automatic zonal segmentation. PAIMRI-FA was exploited for both its zonal segmentation and its segmented lesions. ProstateX was designed for PCa detection but first we used it for automatic zonal segmentation, before exploiting it for its original objective through its inclusion into the PI-CAI dataset. PAIMRI-WA was used in 5 for its large size despite providing only weak annotations (sector and size of each lesion). A more precise description for each of those datasets is present in their corresponding chapters.

1.4 Publications

The contributions listed above led to the following list of peer-reviewed publications.

Journal Articles

- [Mon+21] Montagne, S.*, **Hamzaoui, D.***, Allera, A., Ezziane, M., Luzurier, A., Quint, R., Kalai, M., Ayache, N., Delingette, H., & Renard-Penna, R. (2021). Challenge of prostate MRI segmentation on T2W-weighted images: Inter-observer variability and impact of prostate morphology. Insights into Imaging, 12(1), 71.

- [Ham+22a] **Hamzaoui, D.***, Montagne, S.*, Granger, B., Allera, A., Ezziane, M., Luzurier, A., Quint, R., Kalai, M., Ayache, N., Delingette, H., & Renard-Penna, R. (2022). Prostate volume prediction on MRI: Tools, accuracy and variability. European Radiology, 32(7), 4931–4941.

- [Ham+23b] **Hamzaoui, D.**, Montagne, S., Renard-Penna, R., Ayache, N., & Delingette, H. (2023). Morphologically-Aware Consensus Computation via Heuristics-based IterATive Optimization (MACCHIatO). Submitted to MELBA - Journal of Biomedical Imaging.

- [Wu+22] Wu, C., Montagne, S., **Hamzaoui, D.**, Ayache, N., Delingette, H., & Renard-Penna, R. (2022). Automatic segmentation of prostate zonal anatomy on MRI: A systematic review of the literature. Insights into Imaging, 13(1), 202.

- [Ham+22b] **Hamzaoui, D.**, Montagne, S., Renard-Penna, R., Ayache, N., & Delingette, H. (2022). Automatic zonal segmentation of the prostate from 2D and 3D T2W-weighted MRI and evaluation for clinical use. Journal of Medical Imaging, 9(2), 024001.

- [Ham+23c] **Hamzaoui, D.**, Renard-Penna, R., Montagne, S., Molière, S., Ayache, N., & Delingette, H. (2023). Weak and Mixed supervision for prostate cancer detection through radiological annotations. In preparation

Conference Papers

- [Ham+22c] **Hamzaoui, D.**, Montagne, S., Renard-Penna, R., Ayache, N., & Delingette, H. (2022, September 18). MOrphologically-aware Jaccard-based ITerative Optimization (MOJITO) for Consensus Segmentation. In Uncertainty for Safe Utilization of Machine Learning in Medical Imaging: 4th International Workshop, UNSURE 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings (pp. 3-13).

- [Mol+23] Moliere, S.*, **Hamzaoui, D.***, Montagne, S., Allera, A., Ezziane, M., Luzurier, A., Quint, R., Kalai, M., Ayache, N., Delingette, H., & Renard-Penna, R. (2023). Reference standard for evaluation of automatic segmentation algorithms: quantification of inter observer variability of manual delineation of prostate contour on MRI. Abstract accepted to RSNA 2023.

* indicates that both authors contributed equally to the work.

We also contributed to the redaction of the following papers. As they are not directly related to the subject of PCa detection, they will not be discussed in his manuscript and are only evoked here to show the other collaborations that occured during this PhD:

- [Aud+22] Audelan, B., **Hamzaoui**, **D**., Montagne, S., Renard-Penna, R., & Delingette, H. (2022) Robust Bayesian fusion of continuous segmentation maps, Medical Image Analysis, 78, 102398
- [Aud+20] Audelan, B., **Hamzaoui**, **D**., Montagne, S., Renard-Penna, R., & Delingette, H. (2020). Robust fusion of probability maps. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23 (pp. 259-268).
- [Bri+23] Brillat-Savarin, N., Wu, C., Aupin, L.,, Thoumin, C., **Hamzaoui**, **D**., & Renard-Penna, R. (2023). 3.0 T Prostate MRI : Visual assessment of T2-weighted imaging 2D and 3D from the PI-QUAL score. European Journal of Radiology, 166, 110974.

2 Variability of prostate segmentations and volume measurements between raters

Contents

Abstract Reliable estimation of prostate volume (PV) and of prostate zonal anatomy is essential for prostate cancer management. To improve them, in this chapter, we focus on how those measurements can vary across radiologists. In more details, first, we study the variability of manual prostate zonal segmentation by radiologists on T2W sequences. Second, we determine intra and inter-rater variability of PV from manual planimetry and ellipsoid formulas. In both cases, we analyze the impact of factors such as radiologist' experience and prostates' morphological properties. Forty treatment-naive patients who underwent prostate 3D T2-weighted MRI were selected from a local database, and whole prostate gland (WG) and transition zone (TZ) were segmented by 7 independent radiologists. In addition, they estimated PV and corresponding PSA density (PSAd) using the traditional ellipsoid formula

(TEF), the newer biproximate ellipsoid formula (BPEF), and the manual planimetry method (MPM) used as ground truth. Segmentation variabilities were evaluated based on: anatomical and morphological variation of the prostate, variation in image acquisition, and reader's experience, based on several classic metrics. Volume intra and inter-rater variability was calculated using the mixed model based intraclass correlation coefficient (ICC) and relative standard deviation (rSTD). We showed that segmentation inter-rater variability is higher in the extreme parts of the gland, is influenced by changes in prostate morphology (volume, zone intensity ratio), and is relatively unaffected by the radiologist's level of expertise. All three methods of volume measurements are highly reproducible, however MPM is the one with the lowest variability. TEF showed a high degree of concordance with MPM but a slight overestimation of PV. Precise anatomical landmarks as defined with the BPEF led to a more accurate PV estimation, but also to a higher variability.

This chapter is adapted from two articles published in Insights Into Imaging [Mon+21] and European Radiology [Ham+22a].

2.1 Introduction

Volume estimation and segmentation of prostate MRI play a crucial role in many existing and developing clinical applications, including prostate cancer staging and treatment planning. PSA density (PSAd), one of the strongest predictors of PCa in risk models [Ben+92; Sea+93; Dis+17], is obtained by dividing the prostate-specific antigen (PSA) level by PV. Hence, it is highly dependent on accurate PV measurement. MRI has become the new standard imaging method for prostate volume estimation and segmentation as its higher spatial resolution and better soft tissue contrast compared to previously existing technologies (TRUS) makes it easier for the radiologist to select outer boundaries and provide more accurate and more reproducible volume estimations [Rah+92; LC07; Pat+16]. The PI-RADS V2.1 stipulates that PV should always be reported on MRI and should be determined using either manual or automated segmentation, or calculated using the formula for a conventional prolate ellipse [Tur+19]. Manual segmentation is considered to give the closest volume estimation to pathological specimen volume [Gar+14; Bul+12; Jeo+08]. However, this segmentation is usually performed by contouring the prostate in a slice-by-slice manner using either the axial, sagittal, or coronal views, or a combination of different views. Hence, it is extremely time-consuming, tedious, and prone to inter and intra-observer variation due to the large variability in prostate anatomy across patients [Kor+15], and prostate gland intensity heterogeneity.

The traditional ellipsoid formula (TEF) for volume estimation is very easy and quick, for clinical situations (only a few minutes). However, it relies on geometric models that "approximate" the prostatic contour by considering the prostate as a regular ellipse-like

shape, whereas in reality it is usually irregular and often has an eccentrically enlarged median lobe. To enhance measurement consistency and reduce intra- and inter-rater variability in PV approximation, Wasserman et al [WNS20] recently proposed a new ellipsoid formula, called the biproximate method (BPEF). BPEF is based on very well-defined anatomic landmarks and includes measurement of intravesical prostatic protrusion (IPP), to locate prostate boundaries with more precision. Few studies have examined the precision and accuracy of ellipsoid and planimetry volumetrics measurements. None of them evaluated the recently published BPEF [WNS20], and most of them used one single reader segmentation as ground truth [Bez+18; Sos+03; Gha+21; Tur+13; Maz+15].

In addition, if volume estimation only requires whole gland segmentation of the prostate, PI-RADS is based on the internal structure of the prostate, divided into four histological zones called the peripheral (PZ), transitional (TZ), central (CZ) zones and the anterior fibromuscular stroma (AFMS) [McN68]. Thus, the focus for automatic prostate segmentation went from whole gland segmentation to zonal segmentation of the gland [Mey+19; Ald+20], which is now necessary for the development of AI algorithms for prostate cancer detection.

The quality of a segmentation is evaluated by comparing it to a reference segmentation, often designated as ground truth. Manual delineation of the prostate gland performed by human experts (radiologists or radiation oncologists) is the main approach to generate ground truth. Several teams [MNA16; Wan+19a; Ise+21] have trained their models on prostate MRIs and the relative manual ground truth annotation available from the PROMISE12 challenge [Lit+14b], based on the final segmentation of a single expert reader. Very few studies have systematically investigated inter-reader variability in zonal segmentation due to reader expertise [Bec+19], anatomical or disease-induced variations in the prostate aspect, or technically-induced variability in the image acquisition. There are no current guidelines for prostate zonal segmentation.

In this chapter, we investigate the inter-reader variability when delineating prostate zonal anatomy on T2W sequences with a 3-T MRI without endorectal coil, and the impact of reader expertise, variations in prostate anatomy, cancer-induced modifications, and, for a subgroup of patients, technical differences in image acquisition. Moreover, we evaluate intra- and inter-rater variability in PV estimation, when using manual planimetry measurement (MPM), TEF, and BPEF.

2.2 Dataset

This work was supported by the Clinical Data Warehouse of the AP-HP (Assistance Publique-Hôpitaux de Paris) and was approved by our joint institutional review boards. Data were extracted from the Clinical Data Warehouse of the Greater Paris University Hospitals. We compiled a cohort of 40 patients from a larger cohort/dataset (in house, n = 150) of treatment-naive patients who underwent a prostate MRI before the first

round of biopsy for clinical suspicion of PCa between October 2013 and July 2019. This dataset included patients fulfilling the inclusion criterion for clinical indication of prostate MRI for suspicion of PCa (elevated prostate-specific antigen (PSA), positive Digital Rectal Evaluation, genetic susceptibility) with a standardized PI-RADS V2 score. In the compiled cohort, patients were randomly selected in order to have a large distribution of PI-RADS scores and prostate volumes.

2.2.1 MRI protocol

MRI exams were performed using a 3T clinical system (SIGNA ^{TM} Architect, GE Healthcare, Chicago, IL and MAGNETOM ^{TM} Skyra, Siemens Healthcare, Erlangen, Germany) using a 32-channel phased-array torso coil. Patients were advised to perform bowel preparation before the exam and to empty their bladder; 1mg glucagon was administered intramuscularly to reduce peristaltic motion. All MRI protocols included 3D T2W images (characteristics of the acquisition are presented in Additional file 2.9), and for a subgroup of 12 patients, a supplementary axial 2D T2W sequences acquisition.

2.2.2 MRI manual segmentation

Seven radiologist readers performed manual segmentation: 3 expert (>1000 prostate MRI interpreted, G1), 2 senior (500 prostate MRI, G2) and 2 junior (< 100 prostate MRI, G3)) radiologists. A training meeting with the 7 readers was organized before the beginning of the study in order to reach an agreement on segmentation criteria. The basic zonal anatomy of the prostate was reviewed (especially base and apex limits, and the distinction between the TZ and PZ at the base). The readers were instructed to segment the whole gland (WG) and then the transition zone (TZ) first on the axial plane of the 3D T2W sequence (n = 40) and then for a sub group of patients on the axial 2D T2W sequences (n = 12). The PZ was obtained by subtracting the WG and the TZ. The CZ and AFMS were not segmented separately for two reasons. The first was that PCa originating in the CZ is uncommon, and because there are no guidelines regarding delineation of the CZ, which is mostly posterior to the TZ, we chose to include it in the PZ. Second, PCa does not originate from the AFMS which is an entirely non glandular zone. Most suspicious lesions in the AFMS arise in the TZ, therefore we considered the AFMS to be part of the TZ. Example of anatomic zonal segmentation is provided in Fig. 2.1.

Segmentation was performed using MedInria, an open-source software developed by the Inria Research Institute. Polygons were delineated on the axial plane of the 3D (n = 40) and 2D (n=12) T2W sequences, from the lowest part of the apex to the extreme base: approximately one in every six slices on the 3D T2Ws sequences (between 35 and 75 polygons per prostate) and one in every three slices on 2D T2W sequences. The software performed an interpolation between these polygons to create the whole segmentation. All contours were then carefully checked using MedInria's capability for visualization in

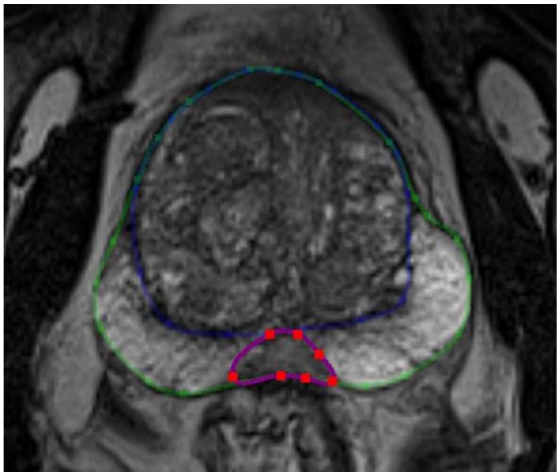


Fig. 2.1.: Example of anatomic zonal segmentation. The central zone (purple) is included in PZ (green contour minus blue contour), and not in TZ (blue) on this slice

three dimensions (axial, sagittal and coronal) and modified if necessary with a repulsor tool or by directly moving one vertex of the polygon (2.13).

Signal intensity Two readers placed similar sized ROIs in the TZ and in the PZ to evaluate TZ (SI_{TZ}) and PZ (SI_{PZ}) signal intensity, and then the squared contrast between both was calculated as $\frac{(SI_{TZ}-SI_{PZ})^2}{(SI_{TZ}+SI_{PZ})^2}$.

2.2.3 Prostate volume evaluation

We tested three evaluation methods: whole gland segmentation with computation of the extracted volume (MPM), and two estimations by ellipsoid formulas (TEF and BPEF).

1. Manual Planimetry (MPM) We computed the whole gland segmentation's volume estimations ($n = 280, 40 \times 7$) using the toolbox SimpleITK [Low+13; Yan+18]. MPM was considered as the ground truth [Gar+14; Bul+12; Jeo+08; Bez+18].

2. Traditional ellipsoid formula (TEF) According to the PI-RADS V2.1 recommendations [Tur+19], the ellipsoid formula was based on the following: (maximum antero-posterior (AP) dimension) × (maximum longitudinal dimension) [both placed on the mid-sagittal T2W sequence] × (maximum transverse dimension) [placed on an axial T2W sequence] ×0.52(Fig. 2.2a, b). Ellipsoid volumes were used as the reference volume for tests on segmentation variability, as the method is the most usual in clinical settings.

3. Biproximate ellipsoid formula (BPEF) This method was described by Wasserman et al [WNS20] and is based on the same formula as TEF but with differences on axes measurements. Length measurement was made on the mid-sagittal plane. Transverse and AP measurements were made on the axial plane showing maximal diameter, and were drawn from the inside border of the external prostatic capsule (Fig. 2.2c, d).

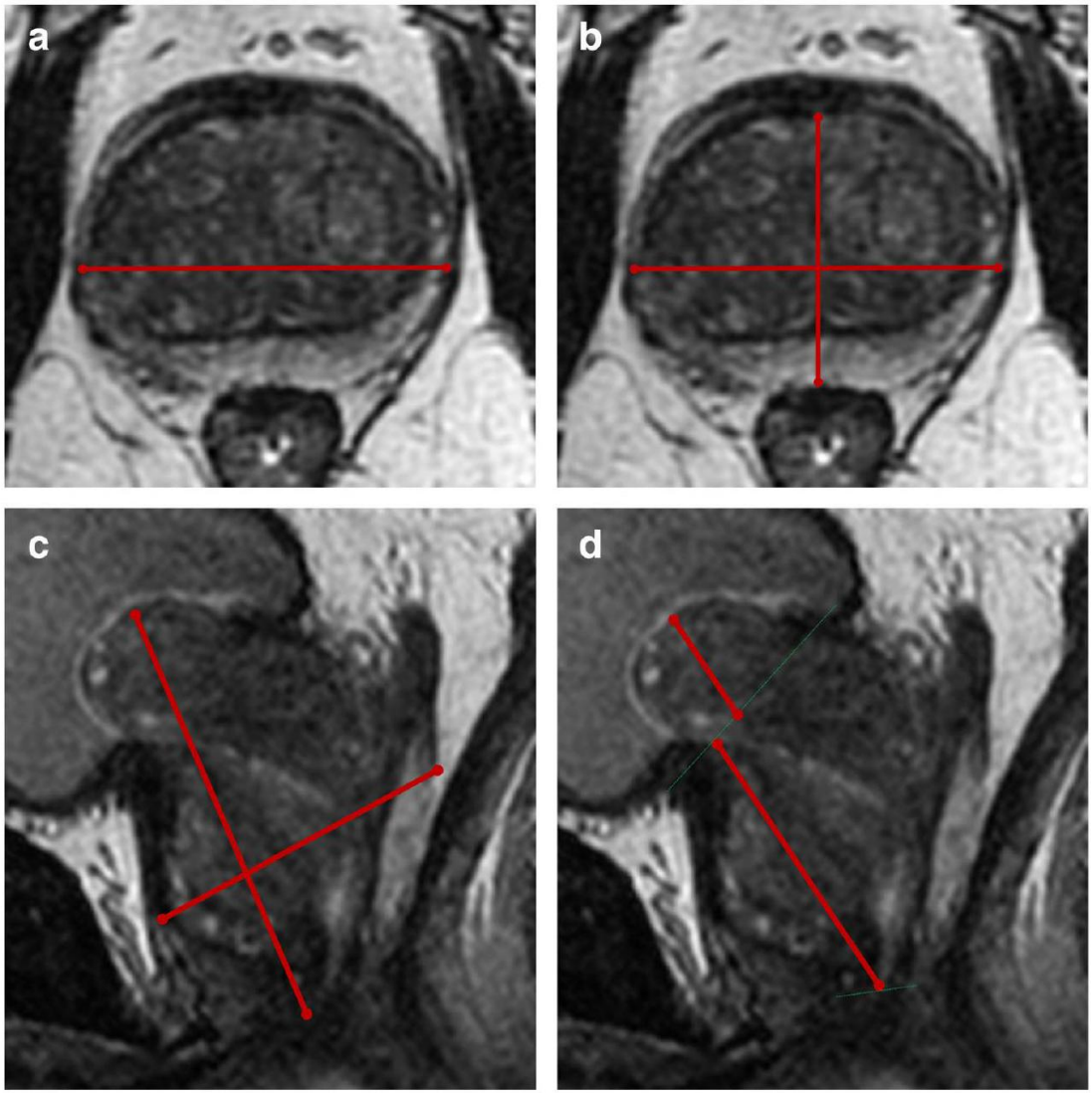


Fig. 2.2.: Example of 3D T2W MRI showing manual prostate measurement: measures are made in the axial plane showing the biggest prostate width (Fig. 1a, c) and the midsagittal plane (Fig. 1b, d). **a** and **b** show the 3 axes used to determine prostate volume by the TEF, and **c** and **d** are the ones used for the BPEF. In d, the line joining the vesicoprostatic angles and the apical line are shown as green dotted lines. Prostate length is calculated by summing both red lines (gland length + median lobe length)

2.3 Metrics and methods

2.3.1 Classical statistical tests

The paired Wilcoxon signed-rank test and Mann-Whitney-U test were used respectively for related samples and independent samples comparisons. The Spearman correlation ρ was used for the correlation calculations. p-values from multiple tests were corrected with the Holm-Bonferroni method. All statistical tests were two-sided. A p-value < 0.05 after correction was considered indicative of a statistically significant difference. We used the Python modules statsmodels, and Pingouin to compute those statistical tests.

2.3.2 Metrics and methods for segmentation comparisons

We used the open-source software VISCERAL Evaluate Segmentation (Apache License v2) for computation of the metrics used for the comparisons and SimpleITK [Low+13; Yan+18]. Two methods were used to evaluate the similarity of the segmentations: classic pairwise calculation (by comparing each mask one by one, and then considering the mean and the standard deviation of the metrics to compare both readers) and, inspired by Shahedi et al. [Sha+17], consensus comparison based on STAPLE algorithm [WZW04](computation of a consensus between the seven raters' segmentations and calculation of the metrics comparing the masks and the consensus mask generated with SimpleITK [Low+13; Yan+18]). Because of correlations existing between those metrics, we only performed statistical tests on some of the most commonly used in the literature: The Dice Score (DSC), the Hausdorff Distance (HD), and the Average Hausdorff Distance (AHD) [TH15]. All metrics are in 3D unless stated otherwise. To investigate the segmentation variability along the cranio-caudal axis we computed HD and DSC for each third of the prostate: apex, mid-gland and base, taking as limits the upper and lower slices of the masks for the pairwise comparison, and the limits of the consensus mask for the STAPLE comparison.

2.3.3 Metrics and methods for volume comparisons

To assess the inter-rater and the intra-rater variability on volume measurements, we used the relative standard deviation (rSTD), defined for an element with multiple measures $X_1, X_2, \ldots, Xn$ as $\frac{\sigma(X)}{\mu(X)}$, and the intraclass correlation coefficient (ICC) derived from a two-way mixed, average measures, absolute agreement model. Empirical statistical power was computed with the R package MKpower (version 0.5), using the mclust package [Scr+16] to estimate distributions with a Gaussian mixture model.

We also used the R package Lme4 (version 1.1-26) to fit a linear mixed-effect model, considering the rater and method effects as fixed effects and the subject impact as a random effect. Inspired by McGraw and al [MW96], we defined $\text{ICC}_{\text{rater}}$ as $\frac{\sigma_{\text{method}} + \sigma_{\text{sujpet}}}{\sigma_{\text{residual}} + \sigma_{\text{method}} + \sigma_{\text{sujet}} + \sigma_{\text{rater}}}$

and $\text{ICC}_{\text{method}}$ as $\frac{\sigma_{\text{rater}} + \sigma_{\text{sujet}}}{\sigma_{\text{residual}} + \sigma_{\text{mechod}} + \sigma_{\text{sujet}} + \sigma_{\text{rater}}}$ to estimate the overall impact of raters and methods on the variability.

No statistical tests were made for experience level impact on volume estimation.
To assess the impact of PV variability on PSAd, we estimated the number of cases that would lead to a clinical disagreement between the PSAd scores computed by raters using the same PV method. Specifically, we identify cases in which there was no unanimous consensus on whether the PSAd fell above or below the classical threshold of 0.15ng/mL for PCa suspicion [Mot+21; Roz+20]. We also computed specificity, sensitivity, and area under the curve (AUC), taking as PV for a given patient and a given method the mean of the volumes obtained by the seven raters [Eri+02].

2.4 Results

The demographic, biologic and morphological data for our population are summarized in Table 2.1. Median age at MRI was 64 years [range 45 – 76 years], mean PSA level was 8.4 ± 5.6ng/mL, and median prostate volume was 57.8 cm^3 [range 15-199]. Among the 40 patients, 17 (42.5%) were classified with a PI-RADS ≥ 3. ($n = 40$)

Variable	Value
Age (years)[a]	64 [45–76]
PSA (ng/mL)[b]	8.4 (± 5.6)
MRI equipment	
3 T SIGNA™ Architect, General Electrics	11 (27%)
3 T MAGNETOM™ Skyra, Siemens Healthcare	29 (73%)
Prostate Volume (cm^3)[a,c]	57.8 [15–199]
PI-RADS	
PI-RADS 1–2	23 (57.5%)
PI-RADS 3	4 (10%)
PI-RADS 4	6 (15%)
PI-RADS 5	7 (17.5%)
Tumor location evaluated on MRI	
Peripheral zone (PZ)	10 (25%)
Transitional zone (TZ)	7 (17%)

Tab. 2.1.: Demographic and clinical characteristics of study participants.
[a] Median [range]; [b] Mean (± STD).
[c] Median volume was estimated by the median of all the volumes the readers estimated from the MRIs, using the ellipsoid formula

2.4.1 Segmentation

Inter-reader variability of prostate segmentation: WG versus TZ

Pairwise comparison When evaluating the WG, we obtained a mean DSC of 0.92(±SD = 0.02), a mean HD of 9.8(±3.8) voxels, and a mean AHD of 0.17(±0.08) voxels.

Concerning the TZ we found a higher variability with a mean DSC of 0.88(±SD = 0.05), an increase of the mean HD to 12.0 (±4.9) voxels, and an increase of the mean AHD to 0.31 (±0.19) voxels. An example of segmentation variability between the different readers groups is shown in Fig. 2.3, and the global results are illustrated in Fig. 2.4.

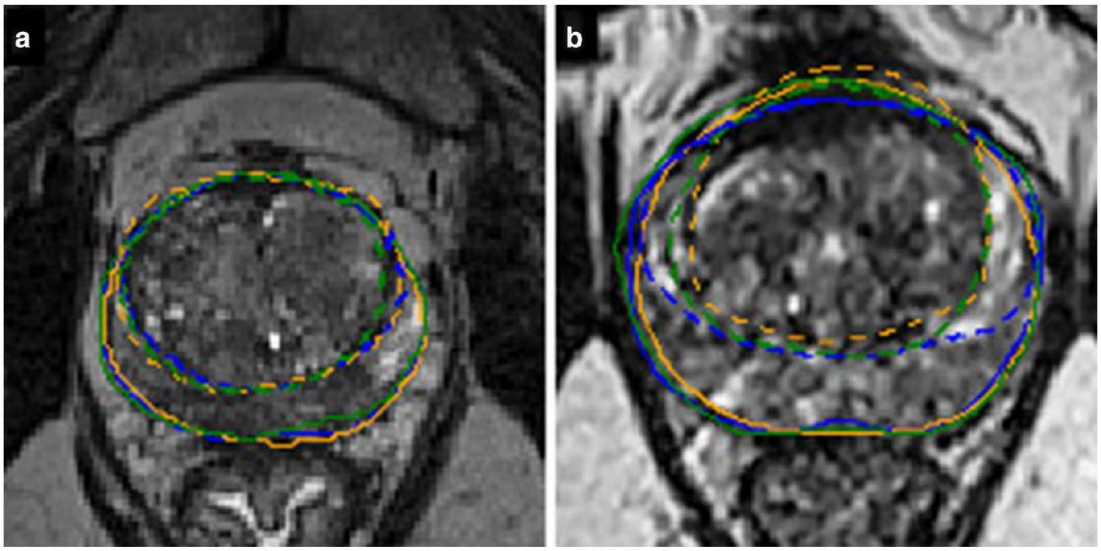


Fig. 2.3.: Examples of low (a) and high (b) segmentation variabilities for WG (full line) and TZ (dashed line) on a transverse slice for one rater of each group of experience (blue for expert, orange for senior, green for resident)

Consensus comparison (STAPLE method) Results (summarized in Table 2.3) were similar for the WG with a mean DSC of 0.94(±SD = 0.03), a mean HD of 8.15 (±3.33) voxels and a mean AHD of 0.11 (±0.07) voxels, and a higher variability for TZ with a mean DSC of 0.91(±SD = 0.05), a mean HD of 10.0 (±4.2) voxels, and a mean AHD of 0.21 (±0.16) voxels.

Inter-reader variability of prostate segmentation: regions/cranio-caudal axis

With the pairwise method the lowest similarity was found at the base with a mean DSC and HD respectively of 0.87(±SD = 0.06) and 9.66(±4.61) voxels, compared to the apex (mean DSC and HD respectively of 0.90 (±0.06), and 7.12 (±3.72) voxels), and to the mid-gland (mean DSC and HD respectively of 0.95(±0.02) and 7.51 (±3.63) voxels). All comparisons between the base and other regions were found to be significant for

Factor	Metric	Pair-wise comparison			STAPLE reference		
		ρ^a	CI 95%	p-value	ρ^a	CI 95%	p-value
Volume[a]	DSC	0.84	[0.72; 0.91]	< 0.001	0.86	[0.76; 0.93]	< 0.001
	HD (voxels)	0.28	[− 0.04; 0.54]	0.5	0.20	[− 0.12; 0.48]	1.0
	AHD (voxels)	− 0.42	[− 0.65; − 0.13]	0.06	− 0.63	[− 0.78; − 0.39]	< 0.001
Squared TZ to PZ contrast[a]	DSC	0.50	[0.23; 0.71]	0.01	0.45	[0.17; 0.67]	0.03
	HD (voxels)	− 0.03	[− 0.34; 0.29]	1.0	− 0.04	[− 0.34; 0.28]	1.0
	AHD (voxels)	− 0.49	[− 0.7; − 0.22]	0.01	− 0.45	[− 0.67; − 0.17]	0.03
Presence of a PI-RADS ≥ 3 lesion[b]	DSC			0.53			0.71
	HD (voxels)			0.3			0.47
	AHD (voxels)			1.0			1.0
Large median lobe[b]	DSC			0.08			0.11
	HD (voxels)			0.61			1.0
	AHD (voxels)			1.0			1.0

Tab. 2.2.: Impact of various factors on segmentation variability with 2 methods (pair-wise comparison and consensus comparison (STAPLE reference)).
[a] Test: Spearman correlation;
[b] Test: Mann-u-Whitney

both metrics. Similar results were obtained with the STAPLE method. These results are summarized in Table 2.3 and illustrated in Fig. 2.4.

Inter-reader variability of prostate segmentation: impact of prostate morphological differences

We found that the smaller the prostate was, the higher the variability was (using DSC for both methods), $\rho > 0.8$ (p-value < 0.001).

A low squared TZ to PZ contrast was significantly associated with a higher segmentation variability ($\rho = 0.5$(CI 95% = [0.23; 0.7], p-value = 0.01 and 0.45 (CI 95% = [0.17; 0.67], p-value = 0.03) for the pairwise method and the consensus comparison (STAPLE method).
No significant difference was found when considering the impact of the presence of tumor (p-value = 0.53 for the mean DSC on the WG). Finally, a retro-urethral lobe protruding into the bladder showed no significant influence on segmentation variability (p-value = 0.08 for the mean DSC on the WG). These results are detailed in Table 2.2 and illustrated in Figs. 2.5, 2.6 and Additional figures 2.14 and 2.15.

Inter-reader variability of prostate segmentation: impact of reader expertise

Masks from the 3 different groups of radiologists (expert, senior, and junior) were compared to the consensus (STAPLE reference).

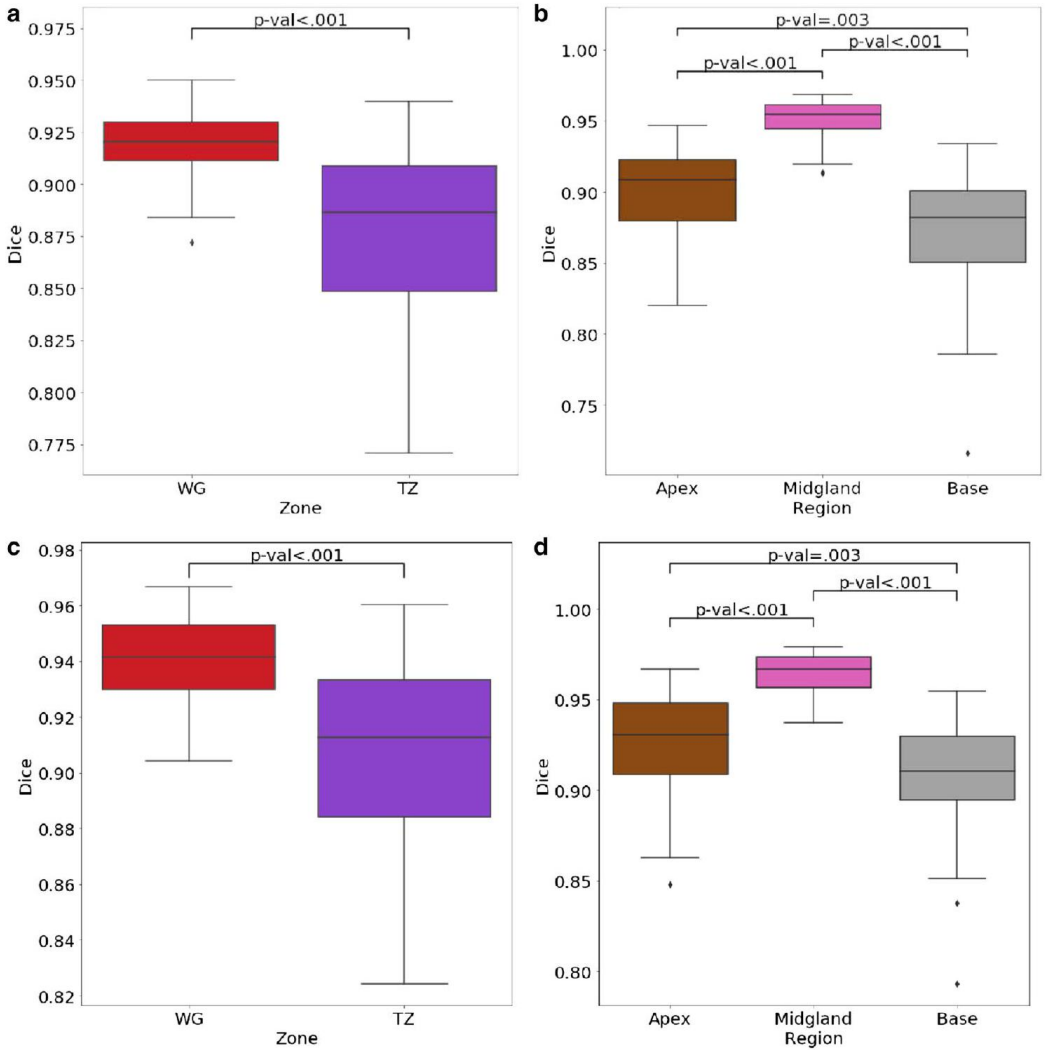


Fig. 2.4.: Comparison of DSC for segmentations of WG and TZ (a, c), and for WG segmentation when the prostate is divided along the cranio-caudal axis in the base/mid-gland/apex (b, d), using a pairwise comparison (**a**, **b**) and a consensus comparison (**c**, **d**)

Structure	Method	DSC	HD (voxels)	AHD (voxels)
WG	Pairwise	0.92 ± 0.02	9.77 ± 3.78	0.17 ± 0.08
TZ		0.88 ± 0.05	11.98 ± 4.92	0.31 ± 0.19
WG	STAPLE	0.94 ± 0.03	8.15 ± 3.33	0.11 ± 0.07
TZ		0.91 ± 0.05	10.03 ± 4.25	0.21 ± 0.16
Base	Pairwise	0.87 ± 0.06	9.66 ± 4.61	
Mid-gland		0.95 ± 0.02	7.51 ± 3.63	
Apex		0.90 ± 0.06	7.12 ± 3.73	
Base	STAPLE	0.91 ± 0.06	7.87 ± 3.69	
Mid-gland		0.96 ± 0.02	6.10 ± 3.05	
Apex		0.93 ± 0.05	5.88 ± 3.05	

Tab. 2.3.: Summarized similarity metrics for all radiologists and all structures (WG vs. TZ, and WG divided along cranio-caudal axis in base, mid-gland and apex), with 2 methods (pair-wise comparison and consensus comparison (STAPLE reference))

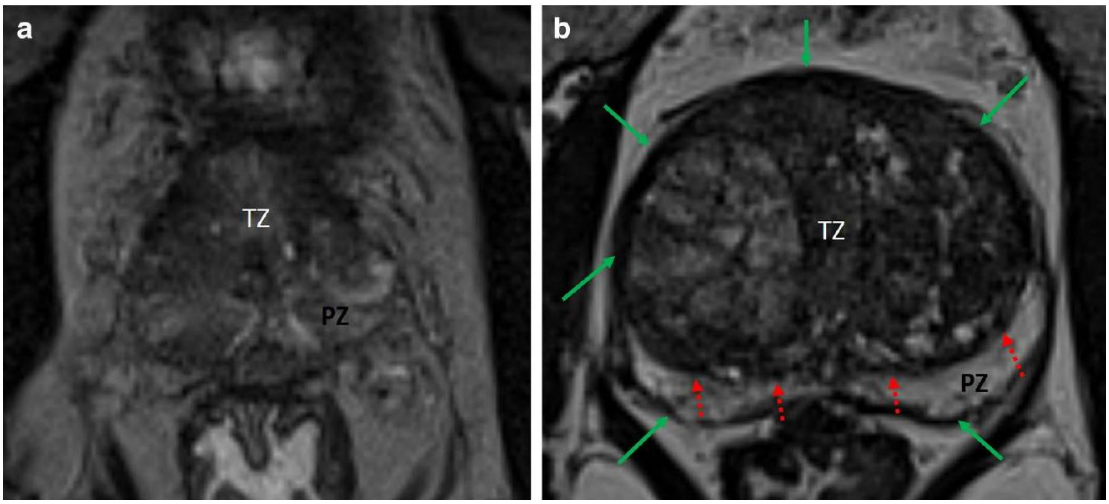


Fig. 2.5.: Influence of variation in prostate volume on zonal differentiation.
a Poor zonal differentiation in a small prostate volume (20 cm^3).
b Clear zonal anatomy differentiation in a larger prostate volume 120 cm^3) : pseudo-capsule (green arrows) and TZ delimitation (red dotted arrows) are clearly individualizable.

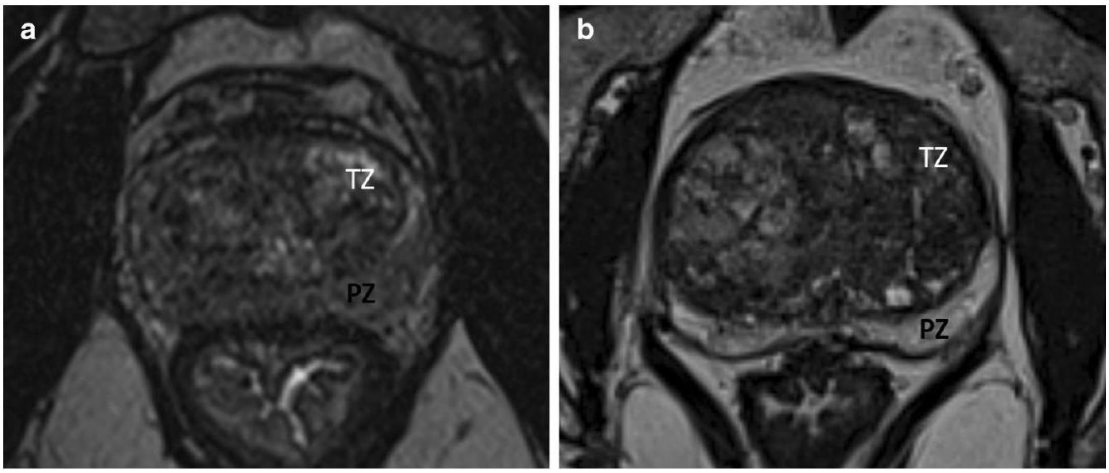


Fig. 2.6.: Influence of intensity signal ratio between the TZ and the PZ on zonal differentiation.
a Moderate signal difference between zones (signal ratio = 0.98).
b Marked difference in signal intensity, facilitating zonal differentiation (signal ratio = 0.37)

For WG, G1, G2, and G3 had respectively a mean DSC of 0.944(±0.023), 0.936(±0.031), and 0.938(±0.025). G1 was the closest to the consensus (p-value = 0.009 and 0.03 for G1/G2 and G1/G3 comparison) (Fig. 2.7). Similar results were obtained using HD and AHD.

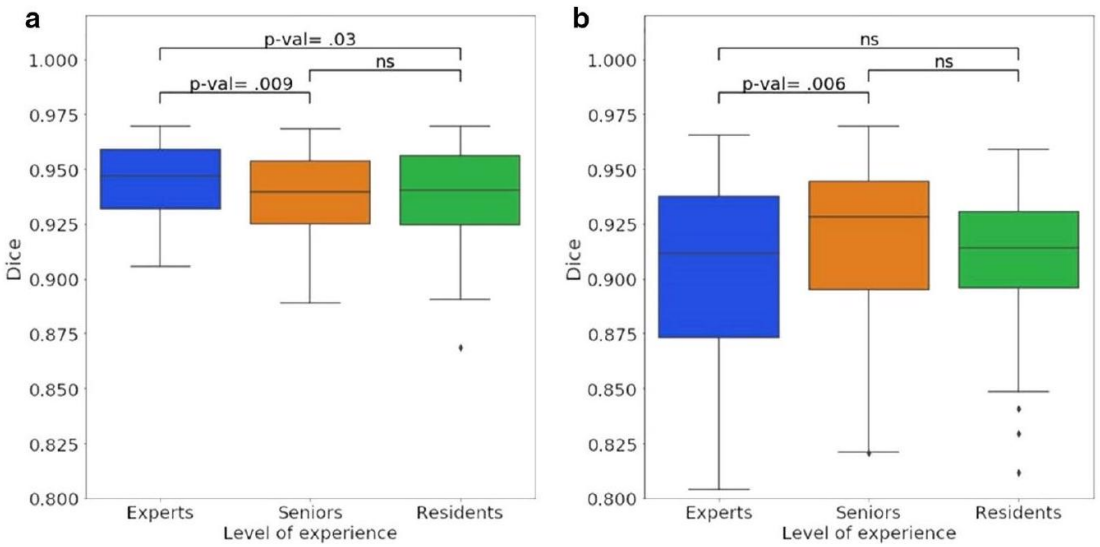


Fig. 2.7.: Impact of the readers'level of expertise (expert/senior/resident) on segmentation variability evaluated by DSC, for WG (a) and TZ (b) segmentation (ns = not significant)

On TZ, G1, G2, and G3 had respectively a mean DSC of 0.903 (±0.061), 0.916(±0.055), and 0.907(±0.04). G2 was the closest to the consensus but was not significantly closer than G3 (p-value = 0.27). The results are summarized in Table 2.4.

Structure	Group	DSC	p-value versus seniors	p-value versus residents	HD (voxels)	p-value versus seniors	p-value versus residents	AHD (voxels)	p-value versus seniors	p-value versus residents
WG	Experts	0.944±0.023	0.09	0.03	7.74±3.15	1.15	0.14	0.10±0.05	0.03	0.009
	Seniors	0.936±0.031	–	0.91	8.49±3.47	–	1.0	0.12±0.08	–	1.0
	Residents	0.938±0.025	0.91	–	8.42±3.41	1.0	–	0.13±0.08	1.0	–
TZ	Experts	0.903±0.061	0.01	1.0	10.50±4.60	0.007	1.0	0.22±0.18	0.002	1.0
	Seniors	0.916±0.055	–	0.27	8.93±3.66	–	0.01	0.17±0.16	–	0.09
	Residents	0.907±0.040	0.27	–	10.42±4.12	0.01	–	0.22±0.13	0.09	–

Tab. 2.4.: Segmentation variability according to the reader's level of expertise (3 experts/2 seniors/2 residents), with their comparison' associated p-values

Inter-reader variability: 2D versus 3D segmentation

WG versus TZ No significant difference was shown when comparing segmentation on 3D T2W MRI versus segmentation on 2D T2W MRI, neither with DSC for the TZ with 0.860 versus 0.861 (p-value = 0.8), nor with HD on WG and TZ (p-value = 0.24 and 0.44 respectively). The only exception was the mean DSC for WG with 0.91 vs 0.90 (p-value = 0.006).

Cranio-caudal axis We found higher mean slicewise DSC and HD for 3D versus 2D MRI segmentations, but these differences were statistically significant only for mid-gland DSC and base HD(p-value = 0.01 and 0.03).

All those results are summarized in Table 2.5.

Structure	DSC			HD (mm)		
	2D	3D	p-value	2D	3D	p-value
Zone[a]						
WG	0.90 ± 0.03	0.91 ± 0.03	0.006	6.97 ± 2.54	6.68 ± 2.06	0.24
TZ	0.86 ± 0.06	0.86 ± 0.06	0.8	7.92 ± 3.07	7.53 ± 2.23	0.44
Region[b]						
Base	0.70 ± 0.16	0.71 ± 0.15	0.32	6.89 ± 2.79	6.31 ± 2.2	0.03
Mid-gland	0.94 ± 0.02	0.95 ± 0.02	0.01	4.14 ± 1.11	4.31 ± 1.59	0.33
Apex	0.76 ± 0.16	0.79 ± 0.12	0.11	4.54 ± 1.69	4.23 ± 1.56	0.05*

Tab. 2.5.: 2D versus 3D T2W MRI segmentation variability (n = 12).
[a] Computations with 3D metrics.
[b] Computations with slicewise metrics.
* Not significant

2.4.2 Volumes

Prostate volume measurements, MPM, TEF, BPEF

Mean volumes were 67.00(±36.61), 66.07(±35.03), and 64.77(±38.27)cm^3 with TEF, BPEF, and MPM, respectively. Median PV measurements for each technique and each rater are given in Table 2.6 and illustrated in Fig. 2.9. Figure 2.8 shows detailed results for each rater and each method.

Group		TEF [a]	BPEF[a]	MPM [a]
Group 1 (experts)	Reader 3	49.95/62.33/104.68	57.82/69.52/97.73	36.31/55.26/85.98
	Reader 6	32.48/49.86/68.70	32.78/50.51/72.55	35.56/57.92/83.58
	Reader 7	43.75/58.5/85.75	38.32/59.41/81.20	36.85/56.46/87.50
Group 2 (seniors)	Reader 1	39.90/64.53/81.76	35.65/59.18/77.41	30.08/51.14/79.35
	Reader 2	41.00/58.98/85.55	46.36/66.74/87.01	38.11/60.20/91.52
Group 3 (juniors)	Reader 4	33.50/51.4/76.63	32.94/52.02/79.68	34.03/54.34/81.27
	Reader 5	35.25/61.35/87.43	36.19/52.16/69.78	36.38/56.11/84.86

Tab. 2.6.: Distribution of prostate volume estimations for each method and rater.
[a]Q1/median/Q3

While considering the volume distribution, taking as PV for a given patient and a given method the mean of the volumes obtained by the seven raters, we observed that the median difference of calculated volumes compared to the reference (MPM) was significant for TEF with a slight overestimation of PV of 1.91 cm^3 (IQ = $[-0.33\ \text{cm}^3, 5.07\ \text{cm}^3]$, p val = 0.03, power = 0.71) but not for BPEF (1.45 cm^3, IQ = $[-1.07\ \text{cm}^3, 5.63\ \text{cm}^3]$, p-val = 0.43, power = 0.28)(Table 2.7). No statistical difference was found between BPEF

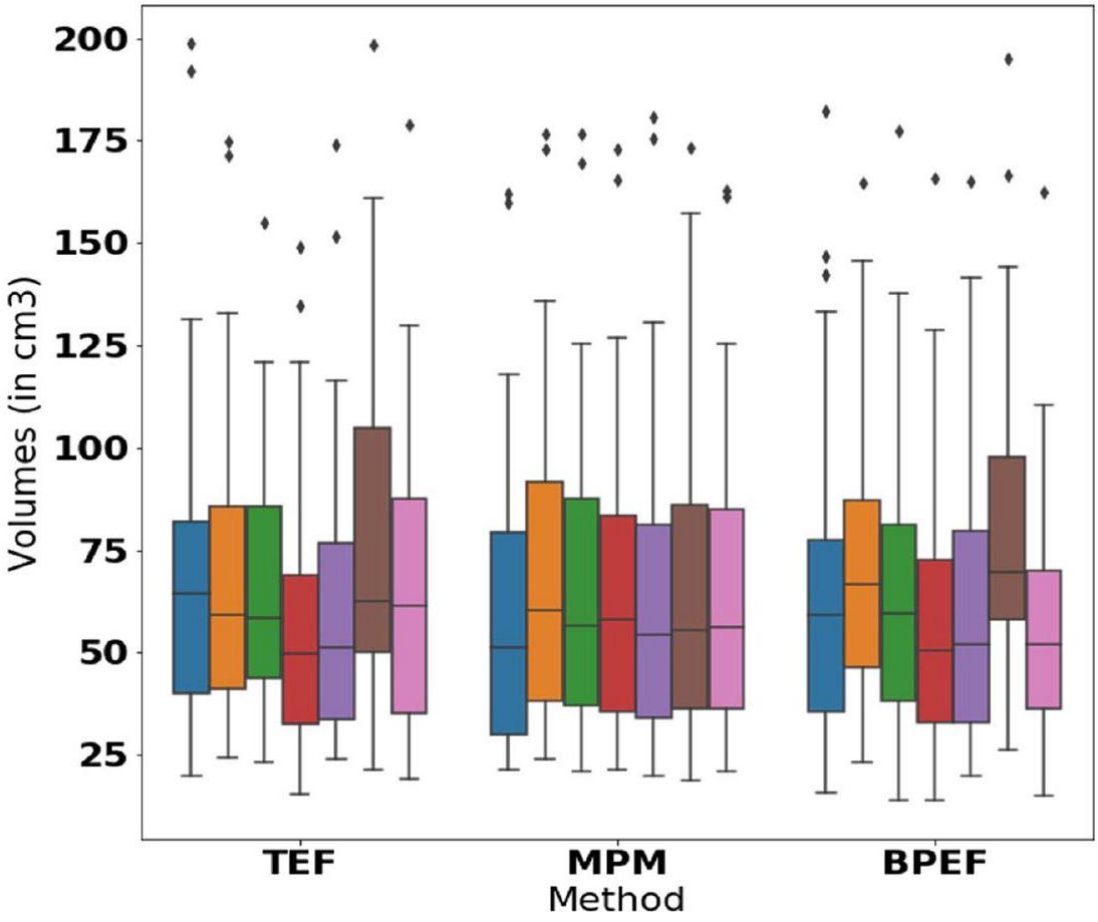


Fig. 2.8.: Volume estimations for the 7 raters and the 3 methods. Each color corresponds to one rater.

and TEF (median difference = $-0.58\ \text{cm}^3$, IQ = $[-3.32\ \text{cm}^3, 2.56\ \text{cm}^3]$, p-val = 0.15, power = 0.27)

Group	Rater	TEF versus MPM	p-val	BPEF versus MPM	p-val
Group 1 (experts)	Reader 3	13.317 ± 12.226	< 0.001	16.517 ± 15.270	< 0.001
	Reader 6	−7.966 ± 9.985	< 0.001	−6.545 ± 12.584	0.02
	Reader 7	0.846 ± 11.935	0.62	1.493 ± 12.728	0.62
Group 2 (seniors)	Reader 1	9.111 ± 9.932	< 0.001	4.637 ± 13.235	0.22
	Reader 2	0.473 ± 7.503	0.69	2.374 ± 14.218	0.15
Group 3 (juniors)	Reader 4	−3.195 ± 9.350	0.62	−2.664 ± 10.156	0.62
	Reader 5	3.000 ± 9.420	0.0498	−6.720 ± 12.869	0.004

Tab. 2.7.: Mean difference between estimated volumes for each rater when using MPM versus ellipsoid methods (TEF or BPEF). More detailed results are available in Additional Table 2.10

PSAd measurements

The median values were 0.117 (IQ = [0.079, 0.193]), 0.127 (IQ = [0.082, 0.202]), and 0.119 (IQ = [0.078, 0.197]) for tefPSAd, mpmPSAd, and bpefPSAd (Fig. 2.9b). As seen for PV, there was a significant difference between mpmPSAd and tefPSAd (p-val = 0.01).

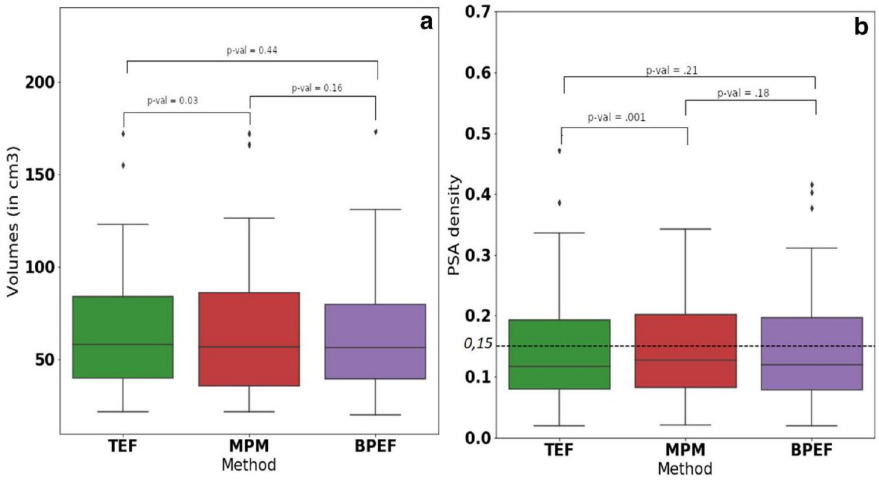


Fig. 2.9.: Subject-wise mean prostate volumes (a) and PSAd (b) for each method. The dotted line in b represents the 0.15ng/mL clinical threshold

Volume Intra-reader variability

For each reader, the ICC between the three PV methods was above 0.90. Detailed results are presented in Table 2.8. No substantial differences were observed according to experience.

Group	Rater	ICC	CI 95%
Group 1 (experts)	Reader 3	0.960	[0.885, 0.982]
	Reader 6	0.981	[0.960, 0.990]
	Reader 7	0.978	[0.963, 0.988]
Group 2 (seniors)	Reader 1	0.981	[0.961, 0.990]
	Reader 2	0.982	[0.969, 0.990]
Group 3 (juniors)	Reader 4	0.987	[0.978, 0.993]
	Reader 5	0.977	[0.948, 0.999]

Tab. 2.8.: Intra-rater reproducibility of volume estimation (evaluated by ICC) for each rater

Volume Inter-rater variability

Using ICC to assess the inter-rater variability, the highest ICC was obtained by MPM (ICC = 0.999, CI 95% = [0.997, 0.9995]), followed by TEF (ICC = 0.988, CI 95% = [0.978, 0.994]) and BPEF (ICC = 0.984, CI 95% = [0.968, 0.992]). MPM's ICC is significantly higher than other methods' ICC, while rSTD for MPM was significantly lower than those of the ellipsoid methods (Fig. 2.10a, b). In addition, the 1^{st} quartile, median, and 3^{rd} quartile of PV measurements are more consistent between raters using MPM than with the ellipsoid methods, as illustrated in Fig. 2.8.

Similar results were obtained on PSAd (Fig. 2.10c, d).

Inter-rater agreement for axes measurement with BPEF

Length measurements are the main source of variability (compared to width or AP measurements), with a mean rSTD and ICC of 0.13 ± 0.04 and 0.943 for length against 0.04 ± 0.03 and 0.986 for transverse diameter and 0.08 ± 0.05 and 0.969 for AP diameter. The differences on rSTD are significant when comparing all three axes (Fig. 2.11).

Evaluation of variability using a linear mixed-effect model

The linear mixed-effect models comparing both ellipsoid methods with MPM returned an ICC of 0.956 (CI 95%: [0.923 – 0.969]) for TEF and 0.932 (CI 95% : [0.8880.953]) for BPEF. Once again, we found a statistically significant difference between MPM and TEF ($p = 0.01$), but not between MPM and BPEF ($p = 0.116$). ICC_{rater} was 0.956 for TEF and 0.911 for BPEF, and ICC_{method} was 0.955 for TEF and 0.910 for BPEF.

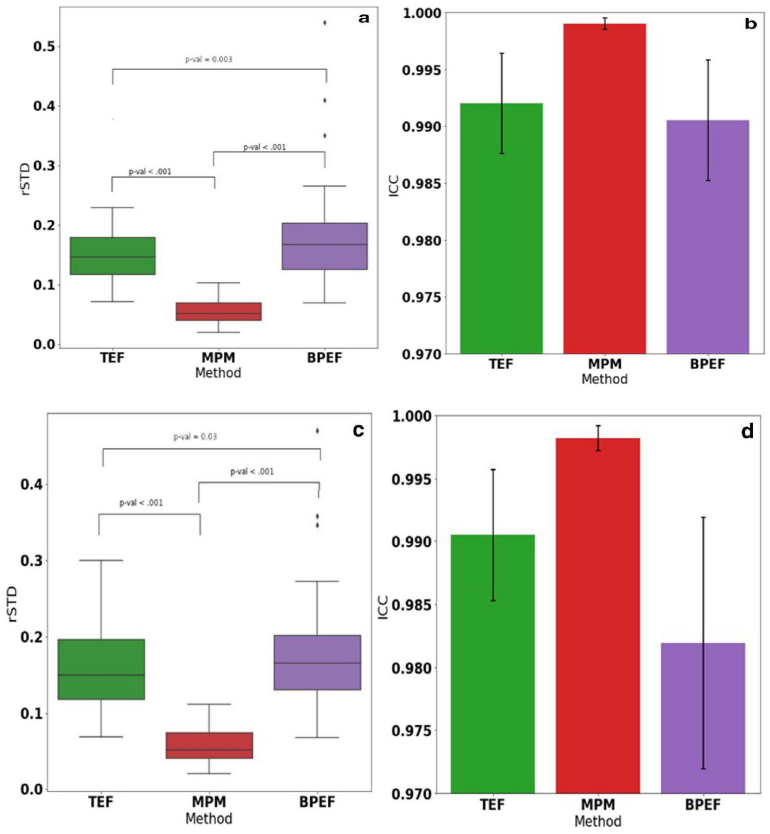


Fig. 2.10.: Inter-rater variability for prostate volume measurement (**a**, **b**) and PSAd (**c**, **d**), depending on the estimation method (p-value < 0.05 for all 3 distributions). a and **c** show relative standard deviation (rSTD); **b** and **d** show intraclass correlation (ICC)

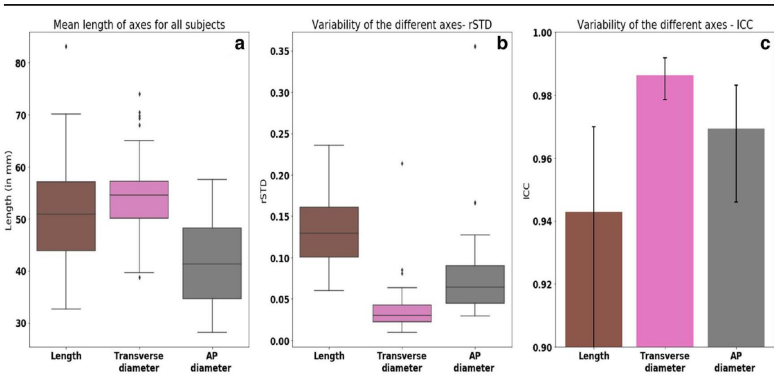


Fig. 2.11.: BPEF axis measure variability. a shows mean measures for e axis (length in brown, width in pink, and antero-posterior in gray). **b** shows the rSTD variability distribution for each axis. rSTD distribution is significantly different for each axis (p-val < 0.001). c shows the ICC distribution for each axis

Impact of volume methods on PSAd measurement and linked diagnosis

In addition to variability estimation using ICC and rSTD which gives similar results to those obtained with volumes, we also estimated the number of disagreements arising from PSAd differences. There were two patients (5%), seven patients (17.5%), and nine patients (22.5%) with disagreements when using MP, TEF, and BPEF, respectively.

At the clinical threshold of 0.15ng/mL, bpefPSAd had a sensitivity of 59% and a specificity of 83% against 65% and 78% for tefPSAd, and 65% and 73% for mpmPSAd. AUC from the three PSAd were similar (Fig. 2.12).

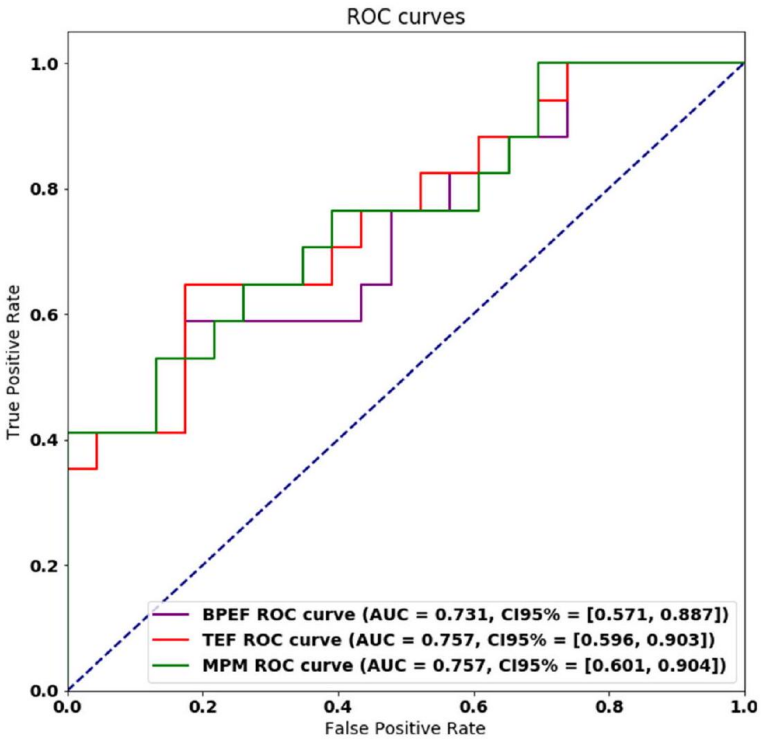


Fig. 2.12.: ROC curves and AUC for PSAd determination when prostate volume is estimated by the three methods (TEF in red, BPEF in purple, MPM in green).

2.5 Discussion

2.5.1 Prostate segmentation

Manual delineation of the internal structure of the prostate performed by human experts is the main approach for generating the ground truth in order to develop automated PCa diagnosis algorithms. Very few studies have investigated the variability of the manual zonal prostate delineation, and for automated segmentation tools under development, a quality and well-described ground truth is rarely available.

To identify sources of variability that may influence the quality of the ground truth for the development of automatic zonal segmentation of the prostate gland, we evaluated in this study the influence of reader expertise, variation of prostate morphology, and in a subgroup of patients variability due to images acquisition differences.

We found a low variability when evaluating the WG (DSC of 0.92 and 0.94 with pairwise and STAPLE method respectively) and slightly higher variability for the TZ segmentation (DSC of 0.88 and 0.91). In the cranio-caudal axis we found a lower similarity at the base (DSC 0.87) and the apex (DSC 0.90) of the prostate.

To our knowledge, two studies have evaluated the inter-reader variability of the zonal anatomy [Bec+19; Pad+19]; Becker et al. [Bec+19; Pad+19] found in a multi-reader study (2 expert radiologists, 2 residents, and 2 computer vision scientists), in a cohort of 80 patients using a 3T MRI and endorectal coil, a DSC of 0.733 for the WG and a higher variability for the TZ (DSC 0.738), in the apex (2D DSC 0.85) and basal part of the gland (2D DSC 0.87). Padgett et al. [Pad+19] in a multi-reader study (n = 2) of zonal segmentation on 2D T2W sequences obtained on 3T of 30 consecutive patients found for the WG a DSC of 0.88 ± 0.04 and 0.81 ± 0.1 for the TZ.

Our results are partly in line with these previously published studies and highlight the difficulty of zonal segmentation especially at the ends of the gland: (a) the apex that has an intensity profile similar to surrounding structures, fuzzy borders, and poor image contrast at the boundary, and (b) at the base with the tricky challenge of partial volume effect between the PZ and the TZ.

Unlike those two previous studies we have chosen to include the CZ in the PZ segmentation. There are no current guidelines in particular regarding whether the CZ should be delineated separately or included in the PZ or in the TZ, and figures provided in the different studies do not clearly indicate this specific point. The CZ, which appears as a symmetric band of tissue between the peripheral and the transition zones at the base of the prostate, extending from below the seminal vesicles to the verumontanum, is extremely difficult to delineate, because it is usually compressed and displaced. Very few cancers arise from this area, (around 7%) [RT14], and even in the PI-RADS score [Tur+19], there is no guidance on how to derive the PI-RADS assessment category for such lesions involving the CZ. It is suggested that CZ lesions should receive PI-RADS score as if they were located in the zone from which they are most likely to be coming from (the PZ or TZ) [Tur+19; Pur+21]. This highlights the need to work on guidelines for prostate delineation for the development of automatic tools.

We evaluated the influence of expertise with 7 readers of varying experiences divided into 3 groups. We did not find any substantial difference on TZ and WG segmentation. Results were statistically significant but numerical DSC values were very close (0.94 for the 3 groups) and showed no substantial difference. Our overall segmentation variability scored higher than those previously published whatever the region analyzed and the level

of expertise. However, all readers in our study were radiologists and have benefited from a training meeting before the start of the study, in order to precisely define the segmentation criteria. This is concordant with the results of Becker et al. [Bec+19], who only found significant differences between non-radiologists and radiologists and concluded that inter-reader baseline of non-radiologists may not be sufficient for meaningful comparison to new segmentation algorithms. Previous studies emphasize the challenge of automated segmentation because of variation in prostate size and shape but there is not description of such variability in the databases, and no evaluation of the influence of anatomical variations such as prostate volume, intensity contrast ratio between TZ and PZ, or the presence of visible lesions.

The prostate gland is a complex organ with varied size, shape and appearance. Morphological differences may contribute to segmentation variability. We found that the smaller the prostate, the higher the segmentation variability was ($p < 0.001$). Hyperplasia of the TZ leading to prostate hypertrophy is the most common change attributed to aging [McN83]. We hypothesized that the increase in size of TZ was associated with sharper contours (surgical capsule) which are then easier to draw, whereas in small volume prostate without prostatic hyperplasia, the glandular tissues of the transition and peripheral zone are histologically identical [McN83] and therefore more difficult to differentiate. In our cohort, 42.5% of MRIs had a PI-RADS score > 2 with lesions in both the PZ and the TZ that may alter the appearance of anatomical structure under segmentation. The presence or absence of a PI-RADS score > 2 lesion did not translate into an increase in segmentation variability ($p = 0.53$). However, the variability increased with a lower PZ to TZ contrast ratio (p-value = 0.01) which can be explain by poor contrast at boundary between zones.

Variation in image acquisition such as 3D versus 2D T2W sequences could translate into variability of segmentation. Unlike 2D T2W sequences, the 3D T2W sequences are acquired with sub-millimeter resolution, to allow the acquisition of a volume that can be reconstructed into any plane with an improvement of anatomic delineation. Although we are aware of the limited number of patients, we didn't find any substantial differences in the subgroup ($n = 12$) who benefited from both types of acquisition.

Zonal prostate segmentation is a fundamental step in the development of automated PCa diagnosis algorithms. In the PROMISE12 challenge [Lit+14b] reference segmentations of the WG were provided in each center by an experienced reader, and were checked by a second expert (with more than 1000 prostate MRIs analyses) who was asked to correct the potential WG segmentation inconsistencies. The resulting segmentation was used as the reference standard and served as a training set for the development for multiple AI algorithms. However, the PROMISE12 database does not provide any zonal information of the prostate besides the WG and furthermore relies only on a single reference standard. Yet, the estimation of inter-observer variability is very important to assess the practical performance of an algorithm with respect to human experts. Indeed,

this variability reflects the intrinsic ambiguity of the segmentation task, and an algorithm performance can be properly assessed by testing whether its output falls within the range of inter-observer variability. Knowledge of the factors influencing the quality of prostate zonal segmentation may also contribute to producing high-quality labeled training data essential for PCa detection and PI-RADS score application. Well-defined guidelines to ensure consistency and accuracy of manual delineation of the prostate are currently not available and should be developed and followed to generate ground truth segmentations. To account for the anatomical and disease-related variability among different patients, as well as the variability in image acquisition, image databases should include representative clinical samples with anatomical variation and patients with different tumors according to their localization.

2.5.2 Volume estimation

Despite the difficulty of delineating the prostate gland, in particular at its extremities, we found MPM to be the most reproducible method (ICC = 0.999, CI 95% = [0.997, 0.9995]). Compared to planimetry, we found a slight overestimation of PV with both ellipsoid formulas, significant with TEF (p-val = 0.03) but not with BPEF (p-val = 0.43) with a median difference of 1.91 cm^3 and 1.485 cm^3. Empirical powers are coherent with those results. Nevertheless, supplementary tests with more subjects are necessary to confirm them with a better statistical power.

Several studies have looked into the accuracy of PV estimation with ellipsoid formulas on 3T MRI with a different type of PV estimation method, in particular by measuring the AP dimension in the axial vs sagittal plane. They found high levels of concordance between ellipsoid formulas and reference (manual planimetry or prostatectomy specimen), with either a slight overestimation [Sos+03; Gha+21] for an underestimation [Bul+12; Maz+15], probably due to variations in the measurements and image interpretation.

Sosna et al [Sos+03] compared values from the ellipsoid formula among 6 different datasets and found that the best estimate was obtained using two diameters from the sagittal plane multiplied by the right-left diameter of the axial plane as recommended in PI-RADS V2.1 [Tur+19].

These authors argued that measuring the AP dimension from sagittal rather than axial images could result in a more precise estimate of PV since the shape of the prostate is more oval or ellipsoid in the sagittal plane. However, measurement of the AP dimension in the sagittal plane may lead to an overestimation because of the inclusion of peri-capsular veins and/or thick anterior fibromuscular stroma. This has been shown by Ghafoor et al [Gha+21], who compared the gland volume measurement between the TEF as defined in the PI-RADS V2.0 and V2.1 [Wei+16; Tur+19] (AP measurement in the sagittal vs axial plane). They found a slight but significant overestimation ($p <$

0.001) of gland volume with AP measurement in the sagittal plane by 2.6 mL compared to the reference.

Turkbey et al [Tur+13] compared the accuracy of fully automated segmentation, manual segmentation, and ellipsoid volumetric measurement using post-operative prostate specimens as "ground truth." Authors found a strong positive correlation between true PV, PV derived from the ellipsoid formula ($R = 0.86 - 0.90, p < 0.0001$), and manual segmentations ($R = 0.89\ 0.91, p < 0.0001$). The strongest correlation was between true PV and manual segmentations.

Bezinque et al [Bez+18] found an excellent correlation between the ellipsoid formula and MRI R3D (automatic prostate segmentation with manual adjustments by an experienced radiologist) measurement (ICC = 0.90), showing that MRI using the ellipsoid formula provides accurate estimates of PV for most patients.

Wasserman et al [WNS20] found an excellent inter- and intra-rater reliability (precision of 0.95 and 0.98, respectively) of their updated method (BPEF), with the length measurement being the most common cause of variation between readers. Our results are concordant with an excellent correlation between BPEF and the reference (median difference of PV of 1.45 cm^3, IQ = $\left[-1.07\ cm^3, 5.63\ cm^3\right]$, p val = 0.43). However, inter-rater variability was the highest with this method (ICC = 0.984, CI 95% = $[0.968, 0.992]$), mainly due to length measurements with a mean rSTD of 0.13 ± 0.04 against 0.04 ± 0.03 and 0.08 ± 0.05 for width and AP, respectively. This underlines the difficulty of delineating precise lower and upper landmarks in the mid-sagittal plane while measuring length.

Very few studies have examined precision, accuracy, and agreements of ellipsoid and planimetry volumetric measurements with MRI, and most of them are limited by not taking into account the level of rater experience in the analysis, and the absence of inter- and intra-rater variability evaluation.

Ghafoor et al [Gha+21] found an excellent inter-rater agreement between four readers for TEF (ICC > 0.90); however, only one reader provided the reference (whole gland manual segmentation).

Bulman et al [Bul+12] compared results from pathological standard, planimetry by two different readers, ellipsoid formula by two other readers, and results obtained by an automated method, with overall good to excellent agreement between the different methods and readers.

Other works only considered one reference segmentation with no analysis on either inter- or intra-rater variability [Bez+18; Sos+03; Tur+13; Maz+15].

Evaluating them and determining their sources are essential to provide accurate gland measurements, in order to obtain the lowest variability impact on PSAd calculation. In our study, we found very low intra- and inter-rater variability between planimetry and the two ellipsoid derived formulas with an ICC > 0.90 for all readers with no impact

of level of experience, although no statistical tests confirmed this last result. Ellipsoid methods may be a time-saver for expert radiologists and allow young radiologists without experience in prostate MRI to safely perform accurate prostate volumetry, an essential step in learning prostate MRI.

However, the MPM method was significantly more reproducible than the ellipsoid-based methods, but also the most time consuming. Until fast, reliable, automated or manually adjusted MRI software is available, ellipsoid formula methods are appropriate for routine clinical work with a high degree of concordance. Although both TEF and BPEF differed from the reference, overestimation was higher with TEF due to less-defined anatomical boundaries, but BPEF was less reproducible probably because of a new definition of these landmarks.

The observed high level of concordance between the measurements translates into a high level of concordance for PSAd risk classifications. However, considering 0.15ng/mL as threshold, disagreement of volumetry-based PSAd levels was lower with MPM (only 5% of disagreements) compared to TEF and BPEF (17.5% and 22.5% disagreement). These individual case errors were not apparent in statistical analysis, but highlighted the relevance of accurate volume estimation for PSAd measurements as it may affect biopsy strategy whether for tumor detection or as part of active surveillance.

2.5.3 Limitations

Our study has several limitations. First, due to the prohibitive time cost, the sample size was small. However, this is because manual segmentation is an extremely time consuming process and this limit was partially offset by the number of readers (seven radiologists) who each provided a planimetry for all forty 3D MRIs of the dataset. We found only one study [Bec+19] with more cases segmented (80 vs. 40) but fewer readers (6 vs. our 7), and with technical differences such as the use of 2D T2W sequences (vs 3D in our case) and an endo-rectal coil (instead of the pelvic coil we used). Second, we chose manual planimetry as the "ground truth" measure for total prostatic volume rather than the pathological specimen, which may be considered as a limitation. However, it has been shown that the mean PV was significantly smaller ex vivo than in vivo with an average change in volume of 19.5% because of loss of vascularity. In addition, tissue shrinkage during specimen processing is one of the factors that may significantly affect the accuracy of PV measurement [Jon+06; Orc+14]. MRI volume imaging eliminates these variables, and it has been argued by multiple authors that MRI volumetric measurements in the living patient should replace postmortem measurement as the "gold standard" [Rah+92; WNS20; Haa+17]. We accepted in this study that manual planimetry could be considered to have the highest level of accuracy and should be considered as ground truth. Finally, we should also point out the lack of non-radiologist readers, which would have been interesting as it was discussed for example by Becker et al. [Bec+19] to evaluate the impact of expertise.

2.6 Conclusion

Identifying sources of variability of prostate zonal segmentation that may influence the quality of the ground truth is a prerequisite for the development of automated PCa detection algorithms. In this study we found that segmentation variability was higher in the extreme parts of the gland, influenced by change in prostate morphology such volume and intensity ratio between zones and was not substantially influenced by radiologist's expertise. Despite those variations, manual planimetry is a robust and reproducible method for PV measurement and PSAd calculation, with the lowest variability between readers. Volumes computed with the traditional ellipsoid formula showed a high degree of agreement with those estimated by planimetry but with a slight overestimation of PV. Delineation of clear anatomical boundaries as defined in the biproximate ellipsoid method leads to a more accurate assessment of PV but with a slight decrease in reproducibility. This highlights the need to include representative clinical samples with morphological variation in image databases to help developing efficient, reproducible and robust automatic segmentation tools in prostate MRI in the future.

2.7 Appendices

	3T SIGNA™ Architect, GE Healthcare, Chicago, IL*		3T MAGNETOM™ Skyra, Siemens Healthcare, Erlangen, Germany**	
Parameters	Axial T2W	3D T2W	Axial T2W	3D T2W
Sequence type	FSE	Echo de Spin Cube	TSE	SPACE
Field of view (mm)	200	280	250	230
Acquisition matrix	512×512	512×512	296×334	230×320
Repetition time (ms)	9861	1602	3050	1550
Echo time (ms)	153.14	102.87	84	173
Flip angle (degrees)	170	-	133	115
Slice thickness (mm)	2.5	1	2.5	0.85
Image reconstruction matrix (pixels)	0.7×0.9×2.5	0.8×0.8×1	0.7×0.7×2.5	0.4×0.4×0.85
Time for acquisition (min:s)	3min34	5min11	3min14	5min35

* Receiver frequency coils: 16-channel phased array body small coil and 32-channel spine coil.
** Receiver frequency coils: 18-channel phased array body coil and 32-channel spine coil.
T2W = T2-weighted imaging, FSE = Fast Spin Echo, TSE = turbo spin-echo,
SPACE = Sampling Perfection with Application optimized Contrasts using different flip angle Evolution

Tab. 2.9.: MRI acquisition specificities

	TEF versus MPM[1]	BPEF versus MPM[1]	BPEF versus TEF[1]
Reader 1	3.43/6.95/13.12	-3.77/4.62/9.99	-8.62/-3.52/2.53
Reader 2	-3.97/0.09/7.49	-1.50/4.78/10.25	-1.15/2.35/9.95
Reader 3	5.63/13.09/19.43	9.21/13.68/21.14	-3.04/2.34/9.21
Reader 4	-8.92/-0.91/2.06	-7.48/-2.04/3.10	-4.95/0.61/5.46
Reader 5	-1.68/2.74/8.80	-9.86/-4.32/0.51	-16.78/-9.01/-0.89
Reader 6	-12.58/-7.45/-2.60	-11.43/-5.03/-0.96	-2.94/2.01/6.14
Reader 7	-2.23/1.68/7.29	-4.95/2.10/10.61	-9.42/0.29/9.76
Mean	-0.33/1.91/5.07	-1.04/1.45/5.63	-3.32/-0.58/2.66

[1] Q1/Median/Q3

Tab. 2.10.: Distribution of estimated volumes difference for each rater

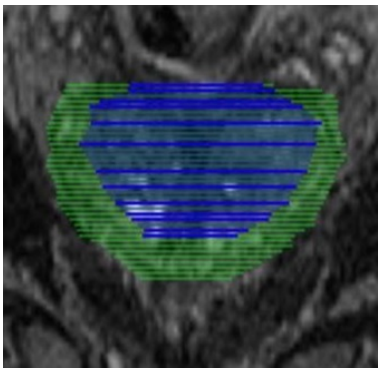


Fig. 2.13.: Example of a segmentation with manually drawn polygons (thick lines visible on TZ) and result of the interpolation between them (thin lines), reformat in a coronal plan. TZ is drawn in blue WG is drawn in green.

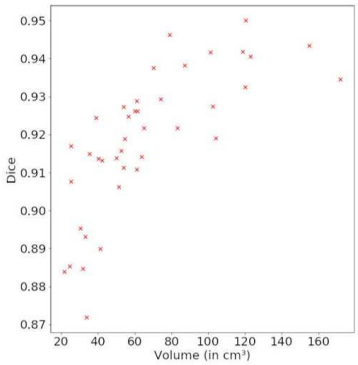


Fig. 2.14.: Relationship between variability (evaluated by DSC) and prostate volume for WG segmentation, using pairwise comparison.

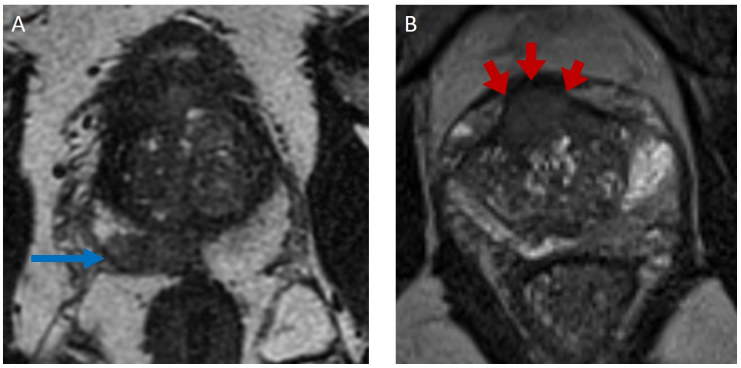


Fig. 2.15.: Example of prostate tumor modifying zones contours. a: PI-RADS 5 tumor in the PZ (blue arrow); b: PI-RADS 5 tumor in the TZ, with contour deformation (red arrows).

3 Morphologically-Aware Consensus Computation via Heuristics-based IterATive Optimization (MACCHIatO)

Contents

Abstract The extraction of consensus segmentations from several binary or probabilistic masks is important to solve various tasks such as the analysis of inter-rater variability or the fusion of several neural network outputs. One of the most widely used method to obtain such a consensus segmentation is the STAPLE algorithm. In

this chapter, we first demonstrate that the output of that algorithm is heavily impacted by the background size of images and the choice of the prior. We then propose a new method to construct a binary or a probabilistic consensus segmentation based on the Fréchet means of carefully chosen distances which makes it totally independent of the image background size. We provide a heuristic approach to optimize this criterion such that a voxel's class is fully determined by its morphological distance, the connected component it belongs to and the group of raters who segmented it. We compared extensively our method on several datasets with the STAPLE method and the naive segmentation averaging method, showing that it leads to binary consensus masks of intermediate size between Majority Voting and STAPLE and to different posterior probabilities than Mask Averaging and STAPLE methods.

This chapter has been submitted to MELBA - Journal of Biomedical Imaging [Ham+23b] as an extension to a paper previously accepted into the MICCAI 2022-UNSURE workshop [Ham+22c].

3.1 Introduction

The fusion of several segmentations into a single consensus segmentation is a classical problem in the field of medical image analysis related to the need to merge multiple segmentations provided by several clinicians into a single "consensus" segmentation. This problem has been recently revived by the development of deep learning and the multiplication of ensemble methods based on neural networks [Ise+21]. One of the most well-known methods to obtain a consensus segmentation is the STAPLE algorithm [WZW04], where an Expectation-Maximization algorithm is used to jointly construct a consensus segmentation and to estimate the raters' performances posed in terms of sensitivities and specificities. The seminal STAPLE method [WZW04] creating a probabilistic consensus from a set of binary segmentations was followed by several follow-up works. For instance, [AL12] replaced global indices of performance by spatially dependent performance fields and [CAAW12] combined STAPLE with a sliding window approach, in order to allow spatial variations of rater performances. Another improvement consisted in introducing the original image intensity information [AL13]. Several alternatives to STAPLE were proposed, with a large diversity of approaches. Some of them decided to use a generative model but with different properties. For example, [Aud+20] modeled raters' input maps by heavy-tailed distributions which parameters are estimated by variational calculus, and [Sab+10] presented a model using a random field learnt on the whole set to model the interaction between the intensity maps and the corresponding label maps. Methods based on deep learning were also conceived, as in [Zha+20] where two CNNs are trained together to estimate simultaneously the consensus segmentation and each rater's performance via an estimation of their spatial confusion matrices. In addition to those complex methods, several studies [RM07; Alj+09] show that simple

majority voting (MV) could remain a suitable pick. However STAPLE and its simple yet robust probabilistic model remains the go-to method for consensus segmentation estimation [WZW04; DV+15] despite suffering from several limitations, some of them already addressed in the literature [AL12; CAAW12; AL13] and some, to the best of our knowledge, never raised before.

In this article, we first analytically characterize the dependence of the STAPLE algorithm on the size of the background image and the choice of prior consensus probability. We then introduce an alternative consensus segmentation method, coined MACCHIatO, which is based on the minimization of the squared distance between each binary segmentation and the consensus. After choosing a distance between binary or probabilistic shapes, the consensus is thus posed as the estimation of the Fréchet mean of this distance, which is independent of the size of the background image for a well-chosen distance. We show that the adoption of specific heuristics based on morphological distances during the optimization allows to provide a novel binary or probabilistic globally consistent consensus method which creates masks of intermediate size between Majority Voting and the STAPLE methods.

3.2 Estimation of a soft or hard consensus from binary segmentations

In the remainder, we consider the problem of generating a consensus segmentation T_n, $1 \leq n \leq N$ given K binary segmentations $\mathcal{S} = \{S^1, ..., S^K\}$, $S_n^k \in \{0, 1\}$ of size N provided by each rater k. The consensus segmentation may be either a *hard* binary segmentation $T_n \in \{0, 1\}$ or a *soft* probabilistic segmentation $\tilde{T}_n \in [0, 1]$, the tilde sign indicating that we are dealing with a continuous probabilistic consensus value, rather than a binary one. Given a soft consensus, one can easily generate a hard consensus by thresholding the soft consensus voxels at the 0.5 limit. Yet, this raises the issue of dealing with voxels that are exactly at the 0.5 value which can be either set arbitrarily to one of the 2 classes, or set aside to a third class.

In terms of probabilistic framework, the main approach is to consider that each observed binary segmentation S^k results from a random process applied on a consensus segmentation T which is captured by the likelihood distribution $p(S^k|T, \theta_k)$ also involving some parameters θ_k specific to each rater k. A prior probability on the consensus $p(T)$ is also defined related to the general *a priori* knowledge about the consensus segmentation. Then a hard consensus can be obtained as a maximum likelihood $T = \arg\max_M p(\mathcal{S}|M)$ or maximum *a posteriori* estimate $U = \arg\max_U p(\mathcal{S}|U)p(U)$ whereas a soft consensus is obtained as the posterior probability $p(\tilde{T}|\mathcal{S}) = p(\mathcal{S}|\tilde{T})p(\tilde{T})/p(\mathcal{S})$. The parameters θ_k are also estimated by maximum likelihood for hard consensuses or maximum marginal likelihood for soft ones.

We make use of the following notations : FP_k, TP_k, FN_k, and TN_k are respectively the number of false positives, true positives, false negatives, and true negatives between observed mask S^k and consensus T, i.e. $FP_k = \sum_{n=1}^{N} S_n^k \wedge T_n$.

We consider as baseline methods to create a hard consensus the majority voting (MV) and the ML STAPLE algorithms whereas mask averaging (MA) and STAPLE algorithm are baseline approaches for the soft consensus estimation. We describe below the hypotheses in terms of probability distribution associated with those baseline models and discuss their limitations.

3.2.1 Majority Voting and Mask Averaging Models

We first make the hypothesis of voxel independence, i.e. that the binary value of each voxel of an observed segmentation mask S^k is independent of the values of other voxels : $p(S^k|T) = \prod_{n=1}^{N} p(S_n^k|T_n)$. Furthermore, we consider that the prior and likelihood probability are simple Bernouilli distribution of the same parameter $b_n \in [0,1]$: $p(S_n^k = 1|b_n) = p(T_n = 1|b_n) = b_n$. This means that the probability parameter b_n is potentially different for all voxels, but the same for all raters : $\theta_k = \theta = \{b_n\}$. Also, the observed masks $\mathcal{S}$ do not directly depend from the consensus but share the same distribution.

Therefore the likelihood of observing the whole segmentation data is then

$$p(\mathcal{S}|\theta) = \prod_{k=1}^{K}\prod_{n=1}^{N} b_n^{S_n^k}(1-b_n)^{1-S_n^k} = \prod_{n=1}^{N} b_n^{S_n^+}(1-b_n)^{S_n^-}$$

where S_n^+ (resp. $S_n^- = K - S_n^+$) is the number of times voxel n is equal to 1 (resp. 0) in the observed segmentation masks S^k,$1 \leq k \leq K$. After maximizing the likelihood, one trivially gets the Bernouilli parameter as $p(S_n^k = 1|b_n) = p(T_n = 1|b_n) = \frac{S_n^+}{K} = b_n$, leading to the Mask Averaging consensus formula where the probability of having a foreground voxel is the frequency of positive voxels in the observed masks S^k. To estimate the hard consensus, one needs to maximize $p(T_n|b_n)$ thus leading to majority voting : $T_n = 1$ if $S_n^+ > S_n^-$ and $T_n = 0$ if $S_n^+ < S_n^-$.

Limitations Majority voting and mask averaging are simple and easy to understand mechanisms to choose a consensus. Yet they suffer from the fact that this decision is purely local without any influence from the neighboring pixels. This can lead to situations where the hard consensus includes some isolated voxels or has very irregular boundaries. Another limitation of majority voting is the case where the number of raters K is even and therefore many decisions are ambiguous with as many foreground than background voxels. Finally, those simple models assume that all raters contributions to the consensus are equal which may not be the case. In particular, an underperforming rater will bias the soft consensus with mask averaging.

3.2.2 STAPLE model

In the STAPLE algorithm [WZW04], all voxels are also assumed independent but the probability that S_n^k is equal to T_n depends on whether T_n is a background or foreground voxel, and on the rater k. More precisely, $p(S_n^k = T_n|T_n = 1) = p_k$ and $p(S_n^k = T_n|T_n = 0) = q_k$ where p_k is the sensitivity of rater k and q_k its specificity.

Prior Consensus The consensus prior probability is here supposed to factorize as the product of voxel priors w_n values $p(T) = \prod_{n=1}^{N} P(T_n) = \prod_{n=1}^{N} w_n$. The original STAPLE paper [WZW04] also introduced an Ising Markov random field model as a prior consensus probability to enforce that a voxel prior value depends on that of its neighbors. However this approach leads to solving iteratively graph cuts problems and is not available in most widely used STAPLE implementations. Instead, the original paper assumes simple independent priors that lead to closed form updates. Choosing $w_n = w = \frac{1}{2}$ is a non-informative prior but another common choice is to have a spatially uniform value $w_n = w = \frac{1}{NK} \sum_{n,k} S_n^k$ which is the average relative size of the foreground object in the observed segmentation masks. We further consider more general priors of the form $w = \frac{A}{N^\alpha}$, with A a constant independent of the image size, and $\alpha \in \mathbb{N}$ an exponent. The non-informative case $w_n = 0.5$ corresponds to $\alpha = 0$ while the average object size to $\alpha = 1$.

Maximum likelihood (ML STAPLE) The likelihood of the observed data simply writes as $\mathcal{L}(T, \theta) = \prod_{k=1}^{K} p_k^{\mathrm{TP}_k}(1 - p_k)^{\mathrm{FN}_k} q_k^{\mathrm{TN}_k}(1 - q_k)^{\mathrm{FP}_k}$ and does not involved the prior on the consensus. There is no closed form expression for the estimation of the rater parameters (p_k, q_k) and the hard consensus (T) maximizing the likelihood. But an iterative maximization of the likelihood is possible by setting its derivatives to zero which leads to the update equation :

$$p_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k} \qquad q_k = \frac{\mathrm{TN}_k}{\mathrm{TN}_k + \mathrm{FP}_k} \tag{3.1}$$

$$s_n^+ = \prod_{k=1}^{K} p_k^{S_n^k}(1 - p_k)^{1-S_n^k} \qquad s_n^- = \prod_{k=1}^{K} q_k^{1-S_n^k}(1 - q_k)^{S_n^k} \tag{3.2}$$

$$T_n = 1 \text{ if } s_n^+ > s_n^- \qquad T_n = 0 \text{ if } s_n^+ < s_n^-$$

Maximum marginal likelihood (MML STAPLE) The *marginal likelihood* or *evidence* writes as $p(\mathcal{S}|\theta) = \prod_{n=1}^{N}(w_n \prod_k p_k^{S_n^k}(1 - p_k)^{1-S_n^k} + (1 - w_n) \prod_k q_k^{1-S_n^k}(1 - q_k)^{S_n^k})$ and is only a function of the rater parameters θ_k. Its maximisation is not tractable in closed form but the expectation-maximisation algorithm provides a way to estimate some local maxima.

The E-step consists in evaluating the posterior probability from Bayes law with the current estimated sensitivities and specificities :

$$u_n = p(\tilde{T}|\theta, \mathcal{S}) = \frac{w_n \prod_k p_k^{S_n^k}(1-p_k)^{1-S_n^k}}{w_n \prod_k p_k^{S_n^k}(1-p_k)^{1-S_n^k} + (1-w_n)\prod_k q_k^{1-S_n^k}(1-q_k)^{S_n^k}} \tag{3.3}$$

The M-step updates the the parameters p_k and q_k as follows:

$$p_k = \frac{\sum_{n,S_n^k=1} u_n}{\sum_n u_n} = \frac{\text{sTP}_k}{\text{sFN}_k + \text{sTP}_k} \qquad q_k = \frac{\sum_{n,S_n^k=0}(1-u_n)}{\sum_n (1-u_n)} = \frac{\text{sTN}_k}{\text{sTN}_k + \text{sFP}_k} \tag{3.4}$$

where sTP_k, sTN_k, sFP_k, sFN_k are the "soft extension" of the number of true positive, true negative, false positive and false negative voxels from rater k.

Influence of the prior term

We can better understand the influence of the prior when estimating the probability to belong to a consensus by writing its logit $\text{logit}(u_n) = \ln\left(\frac{u_n}{1-u_n}\right)$ from Eq.3.3 :

$$\text{logit}\,(u_n) = \text{logit}(w_n) + \sum_{k,S_n^k=1} \log\left(\frac{p_k}{1-q_k}\right) + \sum_{k,S_n^k=0} \log\left(\frac{1-p_k}{q_k}\right) \tag{3.5}$$

Thus, we see that to estimate u_n each foreground voxel of rater k "votes" with a (usually) positive quantity $\log\left(\frac{p_k}{1-q_k}\right)$ whereas each background voxel "votes" with a (usually) negative quantity $\log\left(\frac{1-p_k}{q_k}\right)$. The prior term $\text{logit}(w_n)$ then biases this votes depending whether w_n is greater or smaller than $\frac{1}{2}$.

Influence of the background size

In many cases, the size N of images that contain the objects delineated by the raters is arbitrary since it can be the size of the original image (with large value of N) or the size of a restricted region of interest (with small value of N). It is therefore important to estimate the influence of the background size, i.e the number of true negative voxels TN_k, in the estimation of the hard and soft consensus. This notion of background size can be extended to the whole set $\mathcal{S}$ as the number of voxels segmented by no rater (i.e. $|\{n|\forall k, S_n^k = 0\}|$).

Influence on hard consensus Based on Eqs.3.1 and 3.2, the sensitivity and coefficient s_n^+ are not influenced by TN_k, but the specificities are. More precisely, we have $q_k = 1 - \frac{\text{FP}_k}{\text{TN}_k} + O((\text{TN}_k)^{-2})$, and therefore the quantity s_n^- tends towards 0 when TN_k reaches

large values. This implies that the hard consensus converges towards the union of all observed segmentation masks when the background size becomes large.

Influence on soft consensus The posterior probability u_n and specificities q_k are mainly impacted by the increase of the background size, while the sensitivities are more marginally influenced. The nature of the soft consensus depends on the α exponent of the prior expression $w_n = \frac{A}{N^\alpha}$, and in particular we have :

$$\text{logit}\,(u_n) = (\sum_{k=1}^{K} S_n^k - \alpha)\log N + \log A + \ln\,(\frac{p_k}{\text{sFP}_k})) + \sum_{k, S_n^k=0} \ln\,(1-p_k) + O(N^{-2})$$

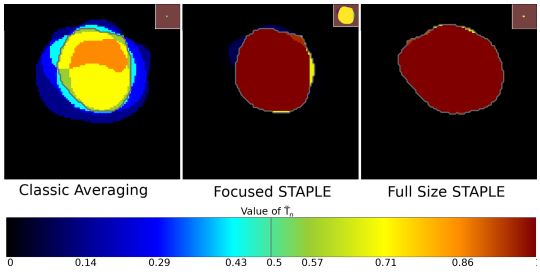
As seen in Fig.3.1b, the soft consensus when having a large background size depends on the value of α, with larger α corresponding to smaller consensus. The detailed proof is presented in Appendix 3.6.1.

Removing the Influence of the background size We explore under which conditions the STAPLE model leads to consensus estimations that are independent of the background size. A first simplification of the model is to assume that all raters perform equally $p_k = p$, $q_k = q$. In this case, the global specificity maximizing the likelihood is $q = \frac{\sum_{k=1}^{K} \text{TN}_k}{\sum_{k=1}^{K} \text{TN}_k + \text{FP}_k}$ which is still dependent on the size of the background through TN_k.

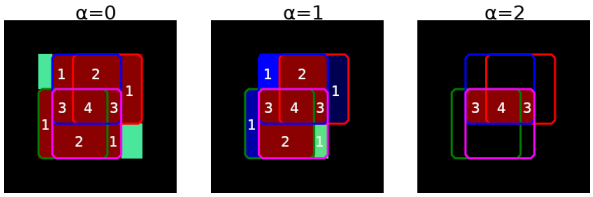
A second simplification is to consider that each rater sensitivity and specificity are equal, i.e. $p_k = q_k = \gamma_k$. This implies that the rater performance is independent of the fact the consensus voxel is in the background or foreground. In this case, the parameter $p_k = q_k = \gamma_k$ can be interpreted as the accuracy parameter and its optimization leads to $\gamma_k = \frac{\text{TP}_k + \text{TN}_k}{N}$. It is easy to see that in that case, $\frac{s_n^+}{s_n^-} = \left(\frac{\gamma_k}{1-\gamma_k}\right)^{S_n^+ - S_n^-}$, and therefore the maximum likelihood is equivalent to majority voting when $\gamma_k > \frac{1}{2}$ which is independent of background size. With this simplification, and from Eq.3.5, the soft consensus obtained by maximizing the marginal likelihood with a non-informative prior $w_n = \frac{1}{2}$, is such that $\text{logit}(u_n) = (S_n^+ - S_n^-)\,\text{logit}(\gamma_k)$. The value of γ_k depends on the background size, but whether a voxel is more likely to be a background pixel $u_n > \frac{1}{2}$ does not depend on the background size.

Limitations

The STAPLE algorithm addresses the problem of taking into account the performance of raters when building a consensus segmentation. However, this approach has the drawback of being dependent on the choice of the prior, and the background size. This dependence of the STAPLE consensus can be explained by the fact that it is a generative



(a) Impact of background size on a simple 2D case with 7 raters (the 7th is an empty map), with STAPLE computed on a 67×61 image (middle), and 640×640 (right) with $w = (\sum_{n,k} S_n^k)/NK$. The relative size of the structure can be seen at the top right corner.



(b) Limits of STAPLE algorithm for 3 different values of α on a toy example with 4 raters providing the red, blue, green and magenta contours. The figures are the number of raters who segmented this zone, and the colors are the probability of the soft consensus with colormap borrowed from Fig.3.1a.

Fig. 3.1.: Impact of STAPLE hyperparameters and background size on the soft consensus

model which should explain the foreground and the background voxels separately. When assuming that the rater performance is the same in both background and foreground, then the model becomes equivalent to majority voting.

The use of local sliding windows in STAPLE as in [CAAW12] can somewhat mitigate the background size effect, but smallest structures in images can still be impacted and the window size remains a hyperparameter which is difficult to set.

3.3 MACCHIatO framework

3.3.1 Main approach description

In the previous section, we have seen that only the majority voting and mask averaging algorithms lead to a consensus which is independent of the background size. Yet, those algorithms are purely local at the voxel level and can lead to irregular boundaries or isolated voxels.

In this section we introduce a new framework to compute soft and hard consensus that are i) invariant from the background size and ii) dependent on the global morphology of each binary object. This approach is coined MACCHIatO for Morphologically-Aware Consensus Computation via Heuristics-based IterATive Optimization.

Distance-based approach We formulate the estimation of a hard consensus T as the minimization of the sum of the square distance between the consensus T and each observed binary mask S^k :

$$T = \arg \min_{M \in \{0,1\}^N} \sum_{k=1}^{K} d(M, S^k)^2 \tag{3.6}$$

where $d(T, S^k)$ is a distance as defined in [DD16] between the two masks S^k and T. This is equivalent to estimating the consensus as a maximum likelihood where the likelihood can be written as $p(S^k|T) \propto \exp(-\lambda d(T, S^k)^2)$. Note that the square sum $\sum_{k=1}^{K} d(M, S^k)^2$ can be seen as the Fréchet variance, and T as the Fréchet mean of the set of binary masks $\mathcal{S}$.

Link with baseline models In section 3.2.2, we have seen that when the sensitivity and specificity are equal, the maximization of the STAPLE model leads to the majority voting algorithm. In this case, we can write the likelihood $p(S^k|T) = \gamma_k^{\mathrm{TP}_k+\mathrm{TN}_k}(1-\gamma_k)^{\mathrm{FP}_k+\mathrm{FN}_k}$ (where γ_k is the accuracy parameter) which is a product of N independent Bernouilli distributions. Since the Bernouilli distribution is a member of the exponential family [DDW13], it can be also written as $p(S^k|T) \propto \exp(-\lambda_k(\mathrm{FP}_k + \mathrm{FN}_k))$ where $\lambda_k = \mathrm{logit}(\gamma_k)$. The number of false positives or false negatives $\mathrm{FP}_k + \mathrm{FN}_k$ is the number of elements of symmetric difference between the two sets S^k and T: $\mathrm{FP}_k + \mathrm{FN}_k = |T \Delta S^k| = |(T \cup S^k) \setminus (T \cap S^k)|$ and is also called the *Hamming distance* in information theory. Thus, for instance by choosing $d(T, S^k) = \sqrt{|T \Delta S^k|}$, the maximum likelihood leads to majority voting consensus (as detailed in Appendix 3.6.2).

Soft consensus framework On the baseline models, soft consensuses were obtained as posterior probability of having a consensus from the observed binary masks. However, from the likelihoods $p(S^k|\tilde{T}) \propto \exp(-\lambda d(\tilde{T}, S^k)^2)$, the computation of the posterior $p(\tilde{T}|\mathcal{S})$ may not be tractable due to the difficulty of computing the normalization constant. Instead, we propose to approximate $p(\tilde{T}_n|\mathcal{S})$ by the quantity $\tilde{U}_n \in [0, 1]$ such that $\tilde{U} \in [0, 1]^N$ minimizes the quantity :

$$\tilde{U} = \arg \min_{\tilde{X} \in [0,1]^N} \sum_{k=1}^{K} d^s(\tilde{X}, S^k)^2 \tag{3.7}$$

where $d^s(\tilde{X}, S^k)$ is a distance between the probabilistic array $\tilde{X}$ and the binary mask S^k. More precisely, the distances $d^s(\tilde{X}, S^k)$ considered are *soft surrogate* of the distance between binary sets $d(\tilde{X}, S^k)$ such that $d^s(\tilde{X}, S^k)^2 = d(\tilde{X}, S^k)^2$ when $\tilde{X} \in \{0, 1\}^N$. For instance, the distance $d(\tilde{X}, S^k) = \|\tilde{X} - S^k\|$ is a soft surrogate of the Hamming distance since $|\tilde{X} \Delta S^k| = \|\tilde{X} - S^k\|^2$. Besides it is clear that the mask averaging (MA) method is a soft consensus minimizing the following square sum $\sum_{k=1}^{K} \|\tilde{U} - S^k\|^2$.

Optimization approach The estimation of the soft and hard consensus is independent of the background size if the distance $d(T, S^k)$ is invariant to the number of true negatives. Besides, unlike the MV and MA algorithms, the optimization cannot be performed at the voxel level when the distance cannot be split voxelwise. Instead of optimizing the whole foreground object, we chose to consider each connected component separately from each

other as to obtain more coherent results. Finally, we further split the optimization on sub-crowns with various heuristics to speed-up the computation.

3.3.2 Distances between binary masks

We detail below the selected distances between binary sets that are considered and their associated soft surrogates. We mainly focus on distances based on two widely used methods to measure the overlap between binary segmentations : the Jaccard and Dice coefficients.

Jaccard distance The Jaccard coefficient (aka IoU) between binary masks A and $B \in \{0,1\}^N$ is defined as: $\text{Jac}(A,B) = \frac{|A \cap B|}{|A \cup B|}$. In [Kos19], it is shown that its complementary to 1 $\text{dist}_J(A,B) = 1 - \text{Jac}(A,B) = \frac{|A \Delta B|}{|A \cup B|}$ is a metric between binary sets following the triangular inequality. Several formulations of soft surrogates exist that extend the Jaccard distance. We focused specifically on two of them: the Soergel metric [Spä81; DD16] $d_{\text{Sg}}(x,y) = \frac{\sum_i \max(x_i,y_i) - \min(x_i,y_i)}{\sum_i \max(x_i,y_i)}$ which follows the triangular inequality but is not differentiable, and the widely-used Tanimoto distance [WBD98; DD16; LG07] $d_{\text{Tan}}(x,y) = 1 - \frac{\sum_i x_i y_i}{\sum_i x_i^2 + y_i^2 - x_i y_i} = \frac{||x-y||^2}{||x-y||^2 + <x,y>}$.

Dice coefficient It is defined as $\text{DSC}(A,B) = \frac{2|A \cap B|}{|A| + |B|}$ and is widely used in image segmentation as a performance index. Indeed, the Dice index is equal to the F1-score and corresponds to the harmonic mean of the sensitivity and positive predictive value. It is closely related to the Jaccard coefficient as $\text{DSC}(A,B) = \frac{2\text{Jac}(A,B)}{1+\text{Jac}(A,B)}$. The Dice distance $\text{dist}_D(A,B) = 1 - \text{DSC}(A,B)$ is a near-metric i.e. it respects a relaxed form of the triangular inequality [GS18]. Soft surrogates of the Dice distance have been developed especially as a loss function in deep learning. We consider in the remainder two main extensions of the Dice distance [Ma+21] on non-binary sets defined as $d_{\text{pSD}}(x,y) = 1 - \frac{2\sum_i x_i y_i}{\sum_i x_i^p + \sum_i y_i^p}$ where $p \in \{1,2\}$.

By construction, all those distances only depend on segmented pixels and are independent of the background size. Note that both distances are extended to get a null distance between two empty sets. The different formulations of the MACCHIatO framework are summarized in the table 3.1.

3.3.3 Heuristic computation based on morphological distance and crowns

Domain of optimization Since the distances listed in the previous section are independent of the number of true negatives, their computations can be restricted to the union of all rater masks : $\mathcal{E}_S = \{n | \sum_{k=1}^K S_n^k > 0\}$. Furthermore, we consider that to decide whether a voxel belongs to the consensus, one should only take into account the regional

Hard Consensus Method	Soft Consensus Method	Distance	Soft Surrogate	Computation level
Majority Voting	Mask Averaging	$\|A \Delta B\|$	$\|\|x - y\|\|$	Voxel-level
ML STAPLE	MML STAPLE	NA	NA	Image-level
MACCHIatO-J	MACCHIatO-TJ	Jaccard d_J	Tanimoto d_{Tan}	Connected component level
	MACCHIatO-SJ		Soergel d_{Sg}	
MACCHIatO-D	MACCHIatO-1SD	Dice d_D	$d_{1\text{SD}}$	
	MACCHIatO-2SD		$d_{2\text{SD}}$	

Tab. 3.1.: Distances between binary sets and their soft surrogate considered to compute hard and soft consensus with the MACCHIatO framework

context associated with the connected components surrounding that voxel, since far away components may not be relevant. Therefore, we choose to minimize separately the Fréchet variances of Eqs. 3.6 and 3.7 for each connected component St of the masks union $\mathcal{E}_{\mathcal{S}}$. Therefore, in practice, we minimize the *Local Mean Squared Distance* between $\mathcal{S}$ and the consensus: $\text{LMSD}_d(\mathcal{S}, M) = \sum_{St \subset \mathcal{E}_{\mathcal{S}}} \frac{1}{K} \sum_k d(S^k_{\|St}, M_{\|St})^2$ where $S^k_{\|St}$ (resp. $M_{\|St}$) are the restriction of the binary masks S^k (resp. M) to the connected component St. To lighten notations, we drop in the remainder the St index which is equivalent to consider that $\mathcal{E}_{\mathcal{S}}$ has only one single connected component.

Sub-crown based optimization The minimization of the Fréchet variance is a combinatorial problem with a complexity of $2^{|\mathcal{E}_{\mathcal{S}}|}$ for the naive approach. Furthermore, it may lead to several global minima when the number of raters K is small. This is why we propose instead to seek a local minimum of the Fréchet variance by introducing some heuristics in the optimization. With this approach, the local minimum has a lower complexity to compute and is by construction maximally connected to avoid isolated voxels. More precisely, instead of a computationally expensive per voxel minimization of the Fréchet variance, we decompose the set $\mathcal{E}_{\mathcal{S}}$ into a set of *sub-crowns* that take into account the global morphological relationships between each rater mask. The formal definition of sub-crowns requires the specification of distance maps $Dm_{\mathcal{N}}(S^k)$ to each binary mask S^k on $\mathcal{E}_{\mathcal{S}}$ according to a chosen neighborhood $\mathcal{N}$. This one can be either the 4 or 8 (resp. 6 or 26) connectivity in 2D (resp. 3D) and the distance $Dm_{\mathcal{N}}(S^k)$ is set to 0 for all voxels inside the object S^k. The global morphological distance map is the sum of those distance maps $D^{\mathcal{N}}_{\mathcal{S}} = \sum_{S^k \in \mathcal{S}} Dm_{\mathcal{N}}(S^k)$ for all raters on $\mathcal{E}_{\mathcal{S}}$. A *crown* $C^{\mathcal{N}}_{td}$ is then defined as the set of voxels having a global morphological distance td. Those crowns realize a partition of $\mathcal{E}_{\mathcal{S}}$ ($\mathcal{E}_{\mathcal{S}} = \coprod_{td} C^{\mathcal{N}}_{td}$), and the 0-crown corresponds by construction to the intersection of all masks in $\mathcal{S}$. We further split each crown as a set of *sub-crowns* by grouping the voxels that have been produced by the same set of raters. In other words, a sub-crown corresponds to a set of voxels located at the same morphological distance from the intersection of all rater masks and which have been segmented by exactly the same group of raters, as seen in Fig. 3.2a. Formally, a sub-crown is noted $(C^{\mathcal{N}}_{td})^g$ where

the superscript g corresponds to a group of raters and sub-crowns realize a partition of a crown :

$$C_{td}^{\mathcal{N}} = \bigsqcup_{g \in \mathcal{P}([\![1,K]\!])} (C_{td}^{\mathcal{N}})^g, \text{with } (C_{td}^{\mathcal{N}})^g = \{n | n \in C_{td}^{\mathcal{N}} \ \& \ \forall k \ S_n^k = (k \in g)\} \tag{3.8}$$

where $\mathcal{P}([\![1, K]\!])$ is the power set (i.e. the set of all subsets) of the first K integers.

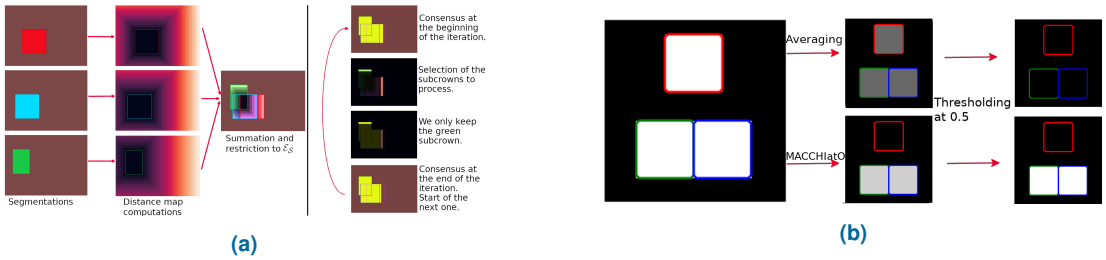


Fig. 3.2.: (a) Left: Preprocessing step of the MACCHIatO algorithm, with the construction of the crowns. Right: An iteration of the shrinking approach with selection of sub-crowns and the evaluation of their contribution to the LMSD_d. (b) Application of averaging and soft MACCHIatO on a toy example with three segmentations (red, green and blue contours). After thresholding, averaging gives an empty segmentation whereas the soft MACCHIatO method is more inclusive and outputs one connected component.

3.3.4 Hard consensus algorithm

The optimization proceeds in a greedy fashion by iteratively removing or adding sub-crowns to the current estimate of the consensus until the LMSD_d criterion stops decreasing. In Alg. 1, we use two concurrent strategies: either we start from the union of all masks (as seen in Fig. 3.2a) and then remove sub-crowns with decreasing distances or we start with the crown with the minimum distance and then add sub-crowns of increasing distances. Both growing and shrinking strategies are applied in order to mitigate the risk of falling into a local minimum and the consensus associated with the minimum LMSD_d of both strategies and the null set is kept. The empty consensus is also tested in a last stage, since there are often a lot of local minima when dealing with large disagreement among raters.

Examples of consensus obtained with this strategy can be seen in Fig. 3.3. Thus, the resulting consensus leads to a consistent grouping since all voxels belonging to the same connected component, having the same morphological distance, and being generated by the same group of raters will end up in the same class. Alternative optimization approaches could have been based on the removal or addition of single voxels (smaller than sub-crowns) or crowns (larger than sub-crowns). While voxel-based minimization would be very time consuming especially in 3D, conversely crown-based would lead to suboptimal results as crowns can be fairly large. Thus, the Morphologically-Aware Consensus Computation via Heuristics-based IterATive Optimization (MACCHIatO) algorithm is designed be a good compromise between computational efficiency and consistency, with a number of iterations exponentially depending on K but which is lower than the

naive $2^{|\mathcal{E}_\mathcal{S}|}$ complexity.

Input: $\mathcal{S}$ segmentation maps, $\mathcal{N}$ neighborhood, d distance
Result: T
Initialization: Computation of $D_\mathcal{S}^\mathcal{N}$, $td_u = \max(D_\mathcal{S}^\mathcal{N})$, $td_i = \min(D_\mathcal{S}^\mathcal{N})$;
$T^u = \bigcup_k S^k$; $T^i = \{n | (D_\mathcal{S}^\mathcal{N})_n = td_i\}$
while $\mathrm{LMSD}_d(T^u, \mathcal{S})$ *decreases* **do** `// Shrinking strategy`
  for $g \in \mathcal{P}(\llbracket 1, K \rrbracket)$ **do**
    if $\mathrm{LMSD}_d((T^u / (C_{td_u}^\mathcal{N})^g), \mathcal{S}) < \mathrm{LMSD}_d(T^u, \mathcal{S})$ **then**
      $T^u \leftarrow T^u / (C_{td_u}^\mathcal{N})^g$
    end
  end
  $td_u \leftarrow \max(\{x \in D_\mathcal{S}^\mathcal{N} | x < td_u\})$
end
while $\mathrm{LMSD}_d(T^i, \mathcal{S})$ *decreases* **do** `// Growing strategy`
  for $g \in \mathcal{P}(\llbracket 1, K \rrbracket)$ **do**
    if $\mathrm{LMSD}_d((T^i \cup (C_{td_i}^\mathcal{N})^g), \mathcal{S}) < \mathrm{LMSD}_d(T^i, \mathcal{S})$ **then**
      $T^i \leftarrow T^i \cup (C_{td_i}^\mathcal{N})^g$
    end
  end
  $td_i \leftarrow \min(\{x \in D_\mathcal{S}^\mathcal{N} | x > td_i\})$
end
$T \leftarrow \arg \min_{T \in \{T^u, T^i, \emptyset\}} \mathrm{LMSD}_d(T, \mathcal{S})$

Algorithm 1: Hard consensus algorithm.

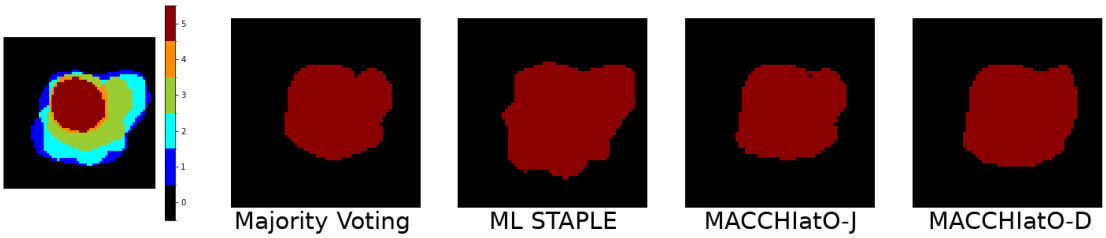


Fig. 3.3.: Comparison of several hard consensus methods on a 2D slice with 5 raters using MV, ML STAPLE and both hard MACCHIatO. On the left is indicated the number of raters who segmented each pixel.

3.3.5 Soft consensus algorithm

The estimation of a probabilistic or soft consensus is based on the minimization of the sum of square surrogate distances as displayed in Eq. 3.7 and the optimization is split for each connected component of the mask union $\mathcal{E}_\mathcal{S}$.

The *soft MACCHIatO* algorithm extends the previous approach to minimize the criterion $\mathrm{LMSD}_{d^s}(\tilde{T}, \mathcal{S})$. A brute force approach would lead to the optimization of a sum of K rational polynomials over a set of $|\mathcal{E}_\mathcal{S}|$ scalars. Instead, we proceed in a greedy manner, separately on each connected component of $\mathcal{E}_\mathcal{S}$, by starting with the mean consensus and

optimizing successively sub-crowns of increasing distances. All sub-crowns of increasing distances are iteratively considered until $\text{LMSD}_d(\tilde{T}, \mathcal{S})$ stops decreasing. For each sub-crown $r = (C_{td}^{\mathcal{N}})^g$, we seek the scalar value $p_r \in [0, 1]$ such that it minimizes

$$p_r = \arg\min_{x\in[0,1]} (d(\tilde{T}_{(td,g),x}, \mathcal{S})), \text{ with } \tilde{T}_{(td,g),x} = \begin{cases} x \text{ if } n \in r \\ \tilde{T}_n \text{ otherwise} \end{cases}.$$

The algorithm is described in Alg.2 and iteratively optimizes each sub-crown from the inside to the outside of the $\mathcal{E}_{\mathcal{S}}$ set. We have observed no gain to combine a growing and a shrinking exploration of sub-crowns unlike Alg. 1. For the optimization process of Eq. 3.3.5, we use the SLSQP algorithm [Kra88] implemented in Scipy v1.7.3 [Vir+20]. Resulting consensus can be seen in Figs. 3.4, 3.6 and 3.7.

Input: $\mathcal{S}$ segmentation maps, $\mathcal{N}$ neighborhood, d^s distance
Result: $\tilde{T}$
Initialization: Computation of $D_{\mathcal{S}}^{\mathcal{N}}$; $\tilde{T} = \frac{1}{K}\sum_{k=1}^{K} S^k$
while $\text{LMSD}_{d^s}(\tilde{T}, \mathcal{S})$ *decreases* **do**
  for $td \in D_{\mathcal{S}}^{\mathcal{N}}$ *in increasing order* **do**
    for $g \in \mathcal{P}(\llbracket 1, K \rrbracket)$ **do**
      $p = \arg\min_{x\in[0,1]} (\text{LMSD}_{d^s}(\tilde{T}_{(td,g),x}, \mathcal{S}))$ with $\tilde{T}_{(td,g),x} = \begin{cases} x \text{ on } (C_{td}^{\mathcal{N}})^g \\ \tilde{T} \text{ elsewhere} \end{cases}$
      $\tilde{T} \leftarrow \tilde{T}_{(td,g),p}$
    end
  end
end

Algorithm 2: Soft consensus algorithm

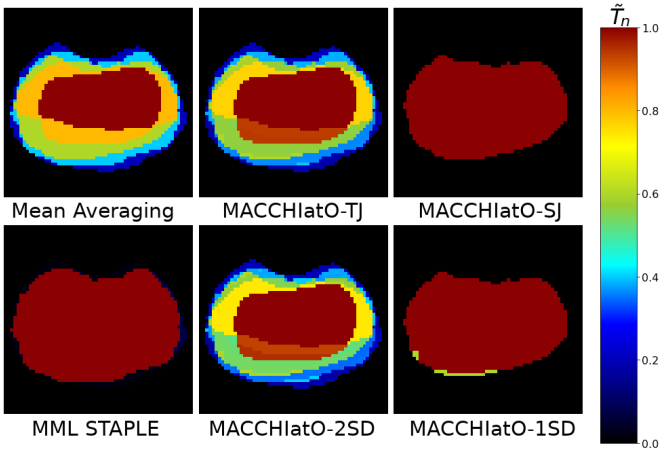


Fig. 3.4.: Comparison of several soft consensus methods on a 2D case with 5 raters using MA, STAPLE and MACCHIatO with different distances.

3.4 Results

3.4.1 Datasets and Implementation Details

We applied our method on 3 datasets:

- A private database of transition zones of prostate T2W sequences, composed of 40 cases segmented by 5 raters.
- The publicly available MICCAI MSSEG 2016 dataset of Multiple Sclerosis lesions segmentations [Com+18] segmented from Brain MR images, with 15 subjects segmented by 7 raters
- The publicly available SCGM dataset [Pra+17], with 40 spinal cords and their grey matter segmented by 4 raters. We used the whole spinal cord segmentation (SCGM-SC) and the grey matter segmentation (SCGM-GM).

Images from the private dataset (resp. MSSEG dataset, SCGM dataset) have a size of [80-288]×[320-640]×[320-640] voxels (resp. [144-261]×[224-512]×[224-512] voxels and [3-28]×[100-655]×[100-776] voxels). It was possible to extract from the private dataset bounding boxes of size [58-227]×[53-184]×[62-180] voxels. Similarly, we were able to extract from SCGM-SC (resp. SCGM-GM) bounding boxes of size [3-20]×[15-90]×[24-131] voxels(resp.) From the 3D private dataset, we created a 2D subset by extracting a single slice for each patient located at the base of the prostate since this region is subject to a high inter-rater variability [Bec+19; Mon+21].
Examples for each dataset of segmentations by the different raters of the same case is available in Appendix 3.6.3 (Fig. 3.8).

Implementation details In the remainder, STAPLE results were produced by using the algorithm implemented in SimpleITK v2.0.2 [Low+13]. All MACCHIatO methods used the 8 or 26-connectivity neighborhood for 2D or 3D cases. MACCHIatO code is available at https://gitlab.inria.fr/dhamzaou/jaccardmap

3.4.2 Heuristics relevance

In Section 3.3.3, we have presented the sub-crown based heuristics which drives the optimization of the local mean square distance criteria. Indeed, those sub-crown group voxels based on three properties: their morphological distance, the connected component they belong to and the raters who segmented them. To check if this heuristics is appropriate, we compared it with two alternatives:

- One iteratively minimizing the LMSD_d at the crown level, without any rater-related property.
- One iteratively processing each voxel separately.

Heuristics	LMSD_d	Time
Subcrown-based heuristics	0.159	0.26s
Crown-based heuristics	0.176	0.07s
Voxel-based approach	0.159	0.92s
Estimated True minimizer	0.159	0.55s

Tab. 3.2.: Computed LMSD_{d^s} and computation time for the soft consensus with Tanimoto distance on the toy example of Fig.3.5 using three different heuristics and the true minimizer.

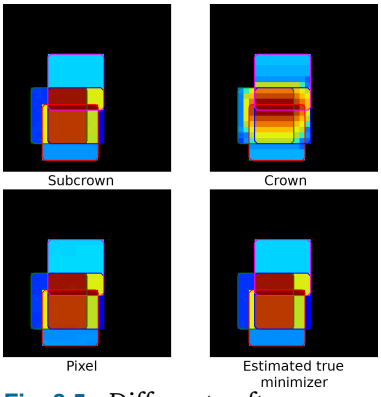


Fig. 3.5.: Different soft consensus obtained on a toy example. Each contour corresponds to one of the raters' segmentation and colors indicate the probability using the same colormap as Fig 3.4.

We compared the 3 heuristics by computing a soft consensus (with the Tanimoto distance) on the toy example of Fig. 3.5 and we display their optimized value of LMSD_{d^s} and their computation time in Table 3.2. Furthermore, since the size of $\mathcal{E}_\mathcal{S}$ is small, we could estimate the true minimizer of LMSD_{d^s} which involves the optimization of $|\mathcal{E}_\mathcal{S}|$ parameters.

Unlike the crown-based heuristics, the subcrown-based and voxel-based heuristics appears to compute a consensus close to the true LMSD_{d^s} minimizer. In addition, the sub-crown method is significantly faster than the voxel-based approach.

We have also compared the three heuristics on two datasets in Table 3.3. The crown-based heuristics is the fastest method to compute but with the highest criteria LMSD_{d^s}, whereas the voxel-based method requires far more time to compute than the subcrown-based heuristics, and even several hours for some Prostate 3D cases. Surprisingly, the subcrown-based heuristics reaches in average a lower LMSD_{d^s} criteria than the voxel-based method, although the difference may hardly be seen on the produced consensus. In those datasets, we were not able to estimate the true minimizer of LMSD_{d^s}, due to the important memory resources those computations would require.

Dataset	Sub-crown	Crown	Voxel
MSSEG	16.36 (57.48s)	16.50 (23.41s)	16.36 (20min30s)
Prostate 3D	1.24e-2 (31.5s)	1.26e-2 (5.46s)	NA
Prostate 2D	5.98e-3 (0.29s)	6.22e-3 (0.07s)	6.10e-3 (5.30s)

Tab. 3.3.: Mean LMSD_{d^s} and computation time for three different heuristics on some datasets

3.4.3 Comparison with baseline methods

Comparison of inter-rater variabilities A first set of experiments consist in measuring the impact of the choice of the consensus method when computing a measure of inter-rater variability. More precisely, we compute the average precision, recall and F1-score between the hard consensus (considered as ground truth) and each rater segmentation. Those metrics have been computed on the MSSEG dataset where there are potentially large disagreements between raters. Table 3.4 reports those metrics averaged among all lesions of all images, a lesion corresponding to a connected component of the mask union $\mathcal{E}_S$. The MV consensus has the highest recall and lowest precision which can be interpreted by a MV consensus smaller than other methods. Conversely, the STAPLE consensus has largest precision and lowest recall, thus corresponding to a consensus of larger size. In terms of F1-score, MV and MACCIHIatO methods are close to each other, but it is highest for MACCHIatO-D (0.449).

Measure \ Method	ML STAPLE	MV	MACCHIatO-J	MACCHIatO-D
Precision	0.976	0.497	0.562	0.570
Recall	0.273	0.817	0.769	0.758
F1-score	0.297	0.437	0.448	0.449

Tab. 3.4.: Averaged lesion-wise measures on the MSSEG dataset for all hard consensus methods

In addition, we also compared the methods on the number of connected components. To do so, we defined each consensus as a potential ground truth and from there computed the average precision, recall and F1-score of each rater for lesion detection (considering the existence of a non-null intersection with the rater's segmentation as a sufficient threshold to detect). We made this experiment on the MSSEG dataset, as it is our only dataset with several connected components per case. Table 3.5 reports those metrics averaged among all patients. The MV consensus has the highest detection recall and lowest detection precision which can be interpreted by a MV consensus not segmenting some lesions conserved by the other methods. Conversely, the STAPLE consensus has largest precision and lowest recall, thus corresponding to the presence of lesions rarely segmented by the raters. In terms of F1-score, MV and MACCHIatO methods are close to each other, but it is highest for MACCHIatO-D (0.894).

Measure \ Method	ML STAPLE	MV	MACCHIatO-J	MACCHIatO-D
Precision	0.994	0.887	0.914	0.931
Recall	0.643	0.967	0.931	0.930
F1-score	0.746	0.892	0.888	0.894

Tab. 3.5.: Measures of lesion detection on the MSSEG dataset for all hard consensus methods

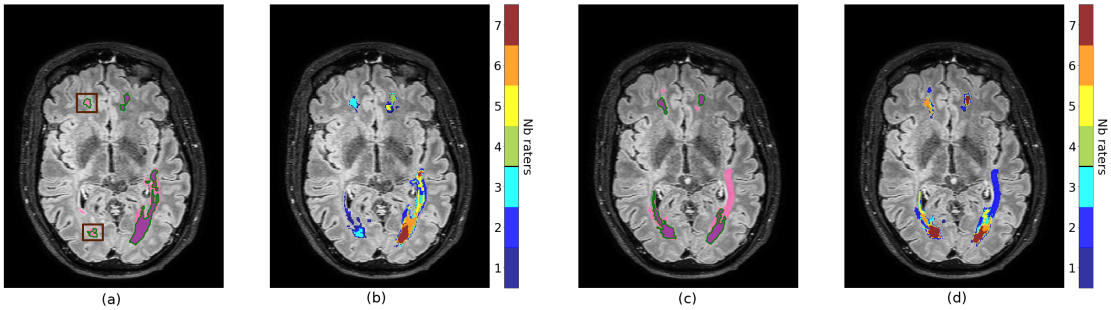


Fig. 3.6.: Two consecutive slices of a MSSEG sample on which we applied STAPLE (pink), Majority Voting (purple) and MACCHIatO-TJ (green contour) (a, c), and for each voxel of those slices the number of raters who segmented them (b, d). We can note that some zones (highlighted by brown squares) were selected by soft MACCHIatO-TJ whereas less than the majority of raters segmented them.

Comparison of consensus areas or volumes In Table 3.6, we compare the relative size of hard consensus on 2 datasets, taking the MV consensus as reference. In average, all methods lead to consensus of larger size than MV. For the MACCHIatO methods, the difference with MV consensus is modest on a massive organ (prostate) but significant for small lesions (>16%). The ML STAPLE method generates much larger consensus than MV, especially when dealing with small lesions. Note that for the MSSEG dataset, ML STAPLE is computed on the whole image, thus with a large background size. Finally, the MACCHIatO-D and MACCHIatO-J methods lead to consensus of similar size, without any clear order. Table 3.7 compares the soft area or volumes of the soft consensus given by $\sum_{n=1}^{N} \tilde{U}_n$ generated by all methods, taking the mask averaging as reference. Fig. 3.6 illustrates those soft consensuses on the MSSEG dataset. The variation of volumes is smaller for soft consensus than for hard consensus. In general, the MA method produces the smallest volumes, and STAPLE the largest ones. The methods using surrogate Dice or Jaccard distances give similar volumes, although the Soergel and $1SD$ are more diverging on the MSSEG dataset. We also compare the size of the thresholded maps $\tilde{U}_n > 0.5$ which provide similar trends than their soft maps.

For both hard and soft consensuses, the largest differences between the different methods are observed on the MSSEG dataset, followed by SCGM-GM.

	Avg. size variation w.r.t MV			Frequencies of size > |MV|		
Method / Dataset	Jaccard	Dice	ML STAPLE	Jaccard	Dice	ML STAPLE
Prostate 3D	+0.4%	+0.6 %	+22%	87.5%	85%	100%
MSSEG	+19%	+16%	+151%	100%	93%	100%
SCGM-SC	+2.36%	+2.30%	+11%	97.5%	97.5%	100%
SCGM-GM	+17%	+15%	+47%	100%	100%	100%

Tab. 3.6.: Left: Average size variation on 3D datasets for hard consensus, with the Majority Voting serving as the reference size. Right: percentage of cases where the computed consensus is strictly larger than the MV consensus. Red color indicates that for this setting, all cases are at least of equal size.

Dataset \ Method	Avg. soft volume variation w.r.t MA				
	TJ	SJ	2SD	1SD	STAPLE
Prostate 3D	+0.4%	+0.1%	+0.1%	+0.7%	+10%
Thresholded	+0.1%	+0.07%	+0.09%	+0.03%	+11%
MSSEG	+4%	+16%	+2%	-3%	+43%
Thresholded	+8%	+37%	+4%	+11%	+68%
SCGM-SC	-0.4%	+0.5%	-0.5%	+0.3%	+4%
Thresholded	+1%	+1.3%	+0.9%	+0.9%	+5.7%
SCGM-GM	+1.2%	+4.4%	+1%	+2.9%	+8.6%
Thresholded	+13%	+16%	+11%	+14%	+19%

Dataset \ Method	Frequencies of soft volume > |MA|			
	TJ	SJ	2SD	1SD
Prostate 3D	80%	65%	60%	80%
Thresholded	22.5%	12.5%	7.5%	7.5%
MSSEG	87%	100%	73%	33%
Thresholded	93%	100%	80%	93%
SCGM-SC	10%	52.5%	5%	37.5%
Thresholded	35%	67.5%	25%	27.5%
SCGM-GM	92.5%	95%	92.5%	82.5%
Thresholded	100%	100%	100%	100%

Tab. 3.7.: Top: Average soft volume variation on 3D datasets for soft consensus, with the MA serving as the reference. Bottom: Percentage of cases where the obtained consensus has a higher volume than the MA consensus. Red color indicates for the thresholded case that for this setting, all cases are at least of equal size.

We recorded the cumulative running time for STAPLE and soft MACCHIatO methods to generate a consensus for all structures of our datasets in Table 3.8. We did not consider MA as it requires far less computation than the other methods. Among the considered algorithms STAPLE is in general the fastest method, being approximately 2-3 times faster than MACCHIatO methods. The exception here being the computation time on SCGM, which always involve small structure sizes and large image sizes.

Method / Dataset	TJ	SJ	2SD	1SD	STAPLE
Prostate 2D	11.1s	14.6s	7.4s	9.8s	2.3s
Prostate 3D	15m02s	12m52s	9m19s	9m48s	4m17s
MSSEG	14m29s	11m31s	11m42s	11m13s	3m38s
SCGM-SC	16.7s	15.1s	14s	14.3	40.6s
SCGM-GM	14.1s	12.8s	12.4s	13.3s	34.7s

Tab. 3.8.: Computation time of continuous methods on all datasets

3.4.4 Entropy of soft consensus

In Figs. 3.3 and 3.7 we show examples of soft consensus on the prostate and lesions datasets. It appears that MACCHIatO-SJ and MACCHIatO-1SD methods often assign to sub-crowns probability values very close to 0 or 1 despite being soft consensus methods. To confirm this behaviour, we compared on all 3D datasets the Shannon entropy $-\sum_n \tilde{U}_n \log \tilde{U}_n - (1-\tilde{U}_n)\log(1-\tilde{U}_n)$ obtained by MA and by the four soft MACCHIatO methods. Table 3.9 confirms the strong binary behavior of MACCHIatO-SJ and MACCHIatO-1SD methods while MACCHIatO-TJ and MACCHIatO-2SD have a similar spread than mask averaging. Thus, we classify the surrogate distances between two families: the ones associated with low-entropy consensus (Soergel, d_{1SD}), and the ones generating high-entropy consensuses (Tanimoto, d_{2SD}).

Dataset	MA	TJ	SJ	2SD	1SD
Prostate 3D	63850	63658	6928	63799	19361
MSSEG	41295	37377	3805	37720	6107
SCGM-SC	2401	2467	259	2483	305
SCGM-GM	757	736	97	736	118

Tab. 3.9.: Mean entropy on 3D datasets for soft MACCHIatO methods. MA entropy is given as a reference.

3.4.5 Discussion

Experiments confirmed the dependence on background size of the STAPLE method, as shown on Fig. 3.1a and in Appendix 3.6.1 (Tab. 3.10). We also observed that hard consensuses obtained by MACCHIatO were generally slightly larger than those obtained by MV, particularly with MACCHIatO-J which never produces consensuses smaller than

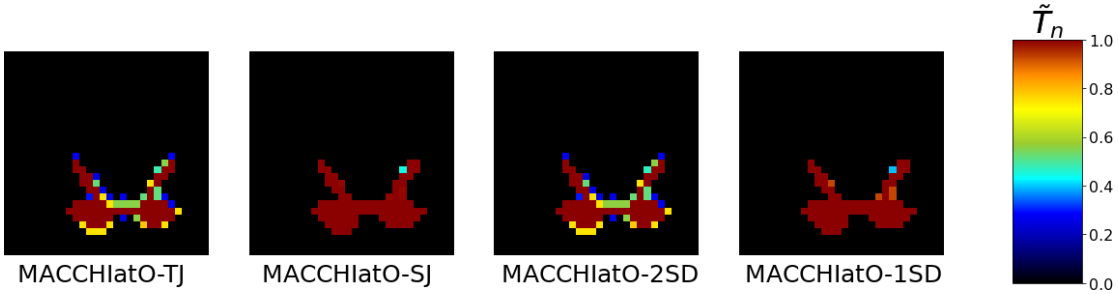


Fig. 3.7.: Impact of the choice of the distance on the computed soft MACCHIatO consensus on a SCGM-GM example

MV's. This can be explained by the fact that the MACCHIatO consensus may include voxels segmented by less than half of the raters (as seen in Figs. 3.3 and 3.6). Finally, STAPLE consensuses always have a larger size than both MACCHIatO and MV. Similar observations can be made on soft consensus but with a smaller difference between methods on soft volumes compared to hard volumes. The MACCHIatO methods by construction create consensuses, independent from the background size, that maximizes the local average (soft) Dice or Jaccard coefficients between the consensus and rater masks for each connected component. Furthermore, they produce masks that are different from the MV and STAPLE methods and having in general larger volumes than MV consensus and smaller volumes than STAPLE ones. Finally, the MACCHIatO algorithms are in general more computationally expensive than MV or STAPLE algorithms but only to a reasonable extent (about 2 or 3 times more).

It can also be noted that the size variation observed on a dataset seems to be correlated with its inter-rater variability, the observed differences being more important on the MSSEG and SCGM-GM dataset than on the others.

In this article, we always considered 8-connexity in 2D cases and 26-connexity in 3D cases, as it performed better on preliminary experiments. However, use of other neighborhoods (such as the 4-neighborhood in 2D, or the 6 and 18-neighborhood in 3D) could be envisaged. Moreover, we did not consider the case of highly anisotropic images, like in the SCGM dataset where a ratio of anisotropy greater than 10 in the voxel size is encountered. For those cases, it could be considered to apply a 2.5D approach consisting in applying our method to each slice independently. Comparisons between 2.5D and 3D neighborhood on SCGM is available in Appendix 3.6.4.

The proposed method has several limitations. First, we only considered a binary segmentation problem. Extension to multiclass segmentation could be foreseen using for instance the generalization method presented in [CCH06] and [Sud+17]. Second, the considered distances between binary sets are based on region overlap measures (Dice, Jaccard indices) and discard distances between boundaries such as Hausdorff Distance (HD). Our experiments based on HD were not conclusive. The reasons for this may be similar to the ones described in [KS19]: instability of the methods to minimize a distance

only defined from the largest error, HD sensitivity to outliers, difficulties to optimize it from an optimization point of view. To mitigate those effects, we made some tests using two of the Hausdorff alternatives defined in [KS19] and based respectively on distance maps and erosion, to no avail.
Third, the proposed criteria LMSD_d, weights all raters equally for all connected components unlike the STAPLE algorithm. It is possible to extend the MACCHIatO framework by attributing a weight to each rater based on their precision and recall (as those measures are independent of background size), either at the local or at the global level. Yet, this extension would require additional optimization steps, since the weights depend on the current estimate of the consensus.

Extending the MACCHIatO method to generate consensuses from K (soft) probability maps instead of binary segmentations is not straightforward. Indeed, while minimizing the Fréchet variance of Eq. 3.7 is well-posed, we can no longer restrict its computation to the set $\mathcal{E}_S$ and define sub-crowns as optimization blocks. An alternative method that we have explored in our prior work [Aud+20], is to map probabilities to real values through a link function (e.g. a logit function) and then use robust parametric models (t-distributions) to fuse the probability maps.

3.5 Conclusion

In this chapter, we have shown that the STAPLE method is impacted by the image background size and the choice of prior law. We have also introduced a new background-size independent method to generate a consensus based on Jaccard and Dice-based distances, thus extending the Majority Voting and mean consensus methods. More precisely, the generated masks minimize the average Jaccard or Dice distance between the consensus and each rater segmentation. The MACCHIatO algorithms are efficient and provide consistent masks by taking into account local morphological configurations between rater masks. The consensus masks are usually of larger size than those generated by the majority voting or mask averaging methods but smaller than those issued by STAPLE. Therefore, we believe based on the experiments performed on two datasets, that the hard and soft MACCHIatO algorithms are good alternatives to MV-based and STAPLE-based methods to define consensus segmentation.

3.6 Appendices

3.6.1 Influence of background size in STAPLE

We can see that by definition u_n is impacted by the value of w_n and, through TN_k, by the background size $BS = |\{n|\forall k, S_n^k = 0\}|$ (i.e. the number of voxels that no rater segmented). In the following subsections we will characterize the dependence of the produced consensus to those parameters.

STAPLE dependence on background size at fixed foreground

By definition, when the background size increases TN_k also increases whereas TP_k, FP_k and FN_k remain constants. So, $q_k \to 1$ when $BS \to \infty$ and we can write

$$\begin{aligned}
\text{logit}\,(u_n) &\sim \text{logit}\,(w_n) + \sum_{k,S_n^k=1} (\ln(p_k) - \ln(1 - \frac{TN_k}{TN_k + FP_k})) + \sum_{k,S_n^k=0} \ln(1-p_k) \\
&\sim \text{logit}\,(w_n) + \sum_{k,S_n^k=1} (\ln(p_k) - \ln(\frac{FP_k}{N - B_k})) + \sum_{k,S_n^k=0} \ln(1-p_k) \\
&\sim \text{logit}\,(w_n) + \sum_{k,S_n^k=1} (\ln(N - B_k) + \ln(\frac{p_k}{FP_k})) + \sum_{k,S_n^k=0} \ln(1-p_k)
\end{aligned}$$

with $B_k = TP_k + FN_k$.

Impact of the consensus prior w_n on the limit

In [WZW04], they proposed to set w_n as a spatially uniform value $w_n = w$ where w is either a constant (typically $w = 0.5$) or defined as the average occurrence ratio ($w = \frac{1}{NK}\sum_{n,k} S_n^k$). We further consider more general priors of the form $w = \frac{A}{N^\alpha}$, with A a constant independent of the image size BS, thus having $\text{logit}(w_n) = -\ln(\frac{N^\alpha - A}{A})$.

From there, we can write

$$\begin{aligned}
\lim_{BS\to\infty} \text{logit}\,(u_n) &= -\ln(\frac{N^\alpha - A}{A}) + \sum_{k,S_n^k=1} \ln(N - B_k) + \sum_{k,S_n^k=1} \ln(\frac{p_k}{FP_k}) + \sum_{k,S_n^k=0} \ln(1-p_k) \\
&= \sum_{k,S_n^k=1} \ln(N - B_k) - \ln(N^\alpha - A) + \ln(A) + \sum_{k,S_n^k=1} \ln(\frac{p_k}{FP_k}) + \sum_{k,S_n^k=0} \ln(1-p_k) \\
&\sim \sum_{k,S_n^k=1} \ln(N) - \alpha\ln(N) + \ln(A) + \sum_{k,S_n^k=1} \ln(\frac{p_k}{FP_k}) + \sum_{k,S_n^k=0} \ln(1-p_k)
\end{aligned}$$

And

$$\lim_{BS\longrightarrow\infty} u_n = \frac{1}{1 + \left(\frac{1}{A}\prod_k \frac{FP_k^{S_n^k}}{p_k^{S_n^k}(1-p_k)^{1-S_n^k}}\right)N^{\alpha-\sum_k S_n^k}}$$

Dataset	Measure	Full size STAPLE	Focused STAPLE
Prostate 3D	Entropy	2019	10992
	Size	300534	285329
SCGM-SC	Entropy	74	269
	Size	11406	11275
SCGM-GM	Entropy	71	118
	Size	1854	1838

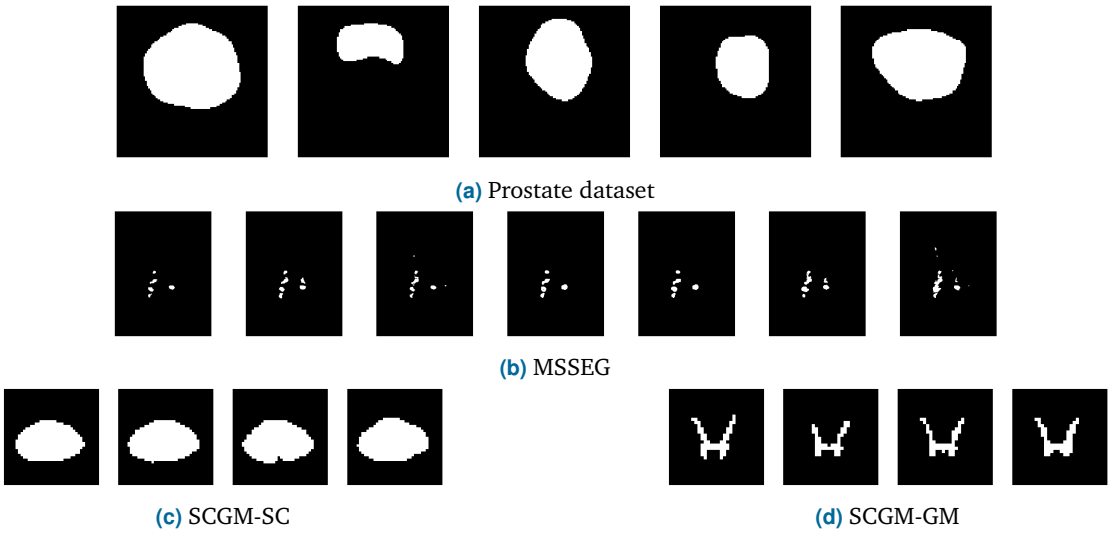
Tab. 3.10.: Mean soft consensus entropy and volume comparisons on Prostate 3D between STAPLE on the full image and on a bounded box.

3.6.2 Proof of Majority Voting as a Fréchet Mean

With $S^1, S^2, ..., S^K \in \{0,1\}^N$ binary segmentation maps and T their Fréchet mean with regards to the function $\sqrt{A \triangle B} = \sqrt{|(A \cup B) \setminus (A \cap B)|}$, we have

$$\begin{aligned}
T &= \arg\min_{M\in\{0,1\}^N} \sum_k \left(\sqrt{|(S^k \cup M) \setminus (S^k \cap M)|}\right)^2 = \arg\min_{M\in\{0,1\}^N} \sum_k \left(\sqrt{|(S^k \cup M) \setminus (S^k \cap M)|}\right)^2 \\
&= \arg\min_{M\in\{0,1\}^N} \sum_k (\sum_n (S_n^k + M_n - S_n^k M_n) - S_n^k M_n) = \arg\min_{M\in\{0,1\}^N} \sum_{k,n} {S_n^k}^2 + {M_n}^2 - 2S_n^k M_n \\
&= \arg\min_{M\in\{0,1\}^N} \sum_n \left(\sum_k (S_n^k - M_n)^2\right) = (\delta(\sum_k S_n^k > \frac{K}{2}))_n \text{(the Majority Voting consensus).}
\end{aligned}$$

3.6.3 Inter-rater variability



(a) Prostate dataset

(b) MSSEG

(c) SCGM-SC

(d) SCGM-GM

Fig. 3.8.: Example of the inter-rater variability between the raters for the different datasets.

3.6.4 Comparison between 2.5D and 3D neighborhoods

	Avg. size variation w.r.t MV		Direct size comparisons	
Method	3D	2.5D	|3D| > |2.5D|	|3D| < |2.5D|
Jaccard	+2.37%	+1.66%	32.5%	65%
Dice	+2.3%	+1.6%	37.5%	55%

SCGM-SC

	Avg. size variation w.r.t MV		Direct size comparisons	
Method	3D	2.5D	|3D| > |2.5D|	|3D| < |2.5D|
Jaccard	+16.9%	+15.8%	77.5%	15%
Dice	+14.7%	+14.9%	67.5%	27.5%

SCGM-GM

Tab. 3.11.: Size comparisons for hard MACCHIatOs between the 2.5D and 3D neighborhood on SCGM-SC (top) and SCGM-GM (bottom)

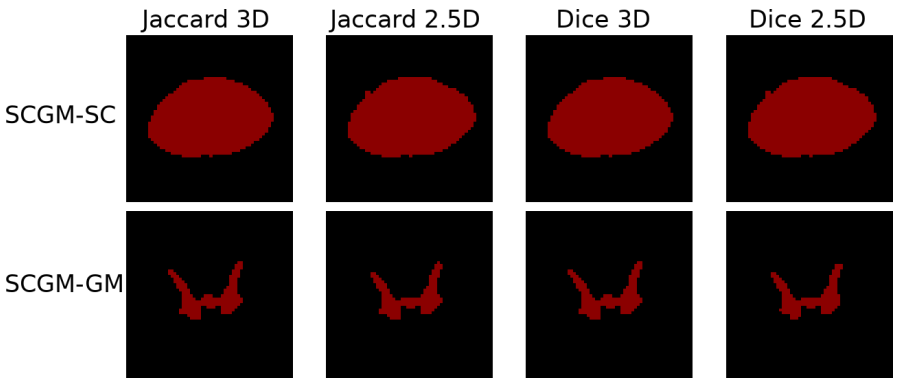


Fig. 3.9.: Examples of hard consensuses on SCGM with 2.5D and 3D neighborhoods.

4 Automatic Zonal Segmentation of the Prostate from 2D and 3D T2-weighted MRI and Evaluation for Clinical Use

Contents

Abstract An accurate zonal segmentation of the prostate is required for prostate cancer management with MRI. The aim of this work is to present UFNet, a deep learning-based method for automatic zonal segmentation of the prostate from T2W MRI. It takes into account the image anisotropy, includes both spatial and channel-wise attention mechanisms and uses loss functions to enforce prostate partition. The method was applied on a private multicentric 3D T2W MRI dataset and on the public 2D T2W MRI dataset ProstateX. To assess the model performance, the structures segmented by the algorithm on the private dataset were compared with those obtained by seven radiologists of various experience levels. On the private dataset, we obtained a Dice score (DSC) of 93.90 ± 2.85 for the whole gland (WG), 91.00 ± 4.34 for the transition zone (TZ) and 79.08 ± 7.08 for the peripheral zone

(PZ). Results were significantly better than other compared networks' (p-value<.05). On ProstateX we obtained a DSC of 90.90 ± 2.94 for WG, 86.84 ± 4.33 for TZ and 78.40 ± 7.31 for PZ. These results are similar to state-of-the art results and, on the private dataset, are coherent with those obtained by radiologists. Zonal locations and sectorial positions of lesions annotated by radiologists were also preserved. Deep learning-based methods can provide an accurate zonal segmentation of the prostate leading to a consistent zonal location and sectorial position of lesions, and therefore can be used as a helping tool for prostate cancer diagnosis.

This chapter was previously published into Journal of Medical Imaging[Ham+22b].

4.1 Introduction

Prostate cancer (PCa) is the most frequent type of cancer affecting men in Europe and North America with more than 280,000 expected cases in the USA in 2023. It is estimated that 12% of American men will develop PCa during their life [Sie+23]. For years, the standard imaging modality used to guide biopsies was transrectal ultrasound (TRUS), prone to underdetection of clinically significant PCa and to overestimation of benign lesions [Sta+19]. Now, its replacement by multiparametric MRI (mpMRI) is supported by several medical associations such as the European Association of Urology [Mot+21] and the American Urological Association [Bju+20]. Based on mpMRI, the PI-RADS score [Tur+19] is designed to improve detection, localization, characterization and risk stratification in patients with suspected PCa in treatment-naive prostate glands. It uses a 5-point scale based on the probability that a combination of mpMRI findings on T2W, DWI and DCE sequences correlates with the presence of clinically significant PCa. Then the patient will benefit from standard and targeted biopsies if a suspected lesion is detected [Sta+19; Kas+18; Rou+19; Kas+19; Elk+19]. PI-RADS defines a dominant sequence for each zone of the prostate: T2W for the Transition Zone (TZ) and DWI for the Peripheral Zone (PZ); so identification of the zonal location of a lesion is vital. Both zones are represented in Fig. 4.1.

In addition, to locate findings on MRI reports and to simplify discussions about biopsies and treatment, radiologists and urologists have defined sector maps that are based on those zones and on longitudinal, transverse and antero-posterior directions. Since both PI-RADS scores and sectorial positions are subject to a high inter-rater variability [Wes+20; Gat+19; Gre+18] there is a need for automated PCa diagnosis methods.
When required, the manual segmentation of prostate zones is commonly performed from T2W sequences. But several factors complicate this task and make it time-consuming even for a skilled physician [Sar+11]. First, boundaries of the prostatic gland and inner boundaries between TZ and PZ may be hard to detect. Second, the prostate is subject to an important inter-subject variability due to physiological differences in terms of shape, size and tissue intensities [Bec+19] as shown in Chapter 2. Finally, sequences acquired

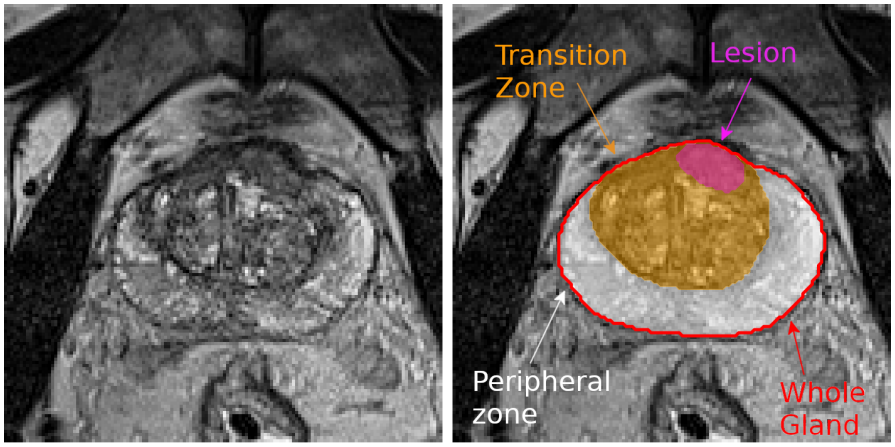


Fig. 4.1.: Left: Axial view of the T2-weighted MR image of a prostate. Right: The corresponding zonal and lesion segmentation. The whole gland is the union of transition zone and peripheral zone.

from different MRI machines increase the variability in appearance of the prostate in T2W imaging.

Related Works Several authors proposed computerized methods for the automatic segmentation of the prostate from T2W sequences. In 2012 the PROMISE12 challenge, dedicated to the segmentation of the whole prostatic gland (WG) [Lit+14b], took place and was won by Vincent et al. [VGB12] using active appearance models. Meanwhile, convolutional neural networks (CNN) began to provide promising results, especially in image classification [KSH12]. Among the different architectures, UNet [RFB15] appeared to be adapted to biomedical image segmentation. Furthermore, in 2016 Milletari et al. [MNA16] presented V-net, a 3D UNet variation with a Dice similarity coefficient (DSC) based loss function dedicated to the automatic segmentation of the prostate with consistent results on PROMISE12 (mean DSC of 86.9±3.3%) and was the first of many works on WG segmentation using deep learning. For example, in 2017, Cheng et al. [Che+17] used holistically nested networks [XT15] and coherence-enhancing diffusion filters [Wei99] to perform this task, and in 2018 Tian et al. [Tia+18] studied the use of transfer learning from large-scale datasets. Now, CNN have become the most widely used methodology for automated segmentation of WG, with best mean DSC on PROMISE12 comprised between 91.5% and 93% [Ise+21; Jia+19].

Although WG prostate segmentation is performed successfully in many cases, the zonal segmentation of the prostate is more difficult especially for the PZ. Indeed, in addition to having a croissant-like shape in axial views, this zone is subject to an important inter-subject variability. For the zonal segmentation of the prostate, many authors used 2D neural networks, as the large anisotropy of 2D T2W sequences makes them closer to stacked 2D images than to real 3D volumes. To improve the generalization on pre-

viously unseen datasets, Rundo et al. [Run+19] proposed a 2D UNet for the zonal segmentation with Squeeze-and-Excitation modules. Aldoj et al. [Ald+20] conceived a DenseNet-like network to perform a zonal segmentation, with a DSC of 92.1± 0.8% for WG and 89.5 ± 2% for the TZ. Cuocolo et al. [Cuo+21a] compared the classic 2D UNet with efficient neural network (ENet) [Pas+16] and efficient residual factorized ConvNet (ERFNet) [Rom+18] which aim to limit their number of parameters, and computation times while keeping a high level of performance. All those articles used the public dataset ProstateX [Lit+17; Lit+14a; Cla+13] consisting of 2D T2W sequences with a slice thickness of 3mm, originally dedicated to PCa diagnosis [Arm+18]. Several works also considered 3D neural networks to take into account the volumetric consistency between slices. Bardis et al. [Bar+21] used a combination of 3D UNets to respectively locate the whole prostate, segment the prostate in the image and classify each voxel of the image as TZ or PZ. Meyer et al. [Mey+19] performed zonal segmentation with a 3D neural network including anisotropic MaxPooling and deconvolutions to perform a zonal segmentation of the prostate. Zavala-Romero et al. [ZR+20] used a multiplanar 3D neural network [Mey+18] to perform this zonal segmentation, and studied in particular the impact of MRI vendors on generated segmentations, showing the importance of multicentric and multivendor datasets.

Limitations All those works obtained good results regarding the zonal segmentation of the prostate. However, they did not tackle the issue of localizing prostate lesions within zones and sectors. This localization is important for grading those lesions in the PI-RADS standard.
Moreover, prior works were based on 2D T2W sequences since they are the most widely available modality. Yet, 3D T2W prostate MR images allow a shorter acquisition time [Ros+10; Pol+17] and simplify modalities fusion among other advantages [Bat+20], while having similar performances in terms of diagnosis. For those reasons they will probably become the new radiological standard in prostate imaging in the near future. Differences between 2D and 3D T2W sequences can be seen in Fig. 4.2.
In addition, inter-rater variability in the segmentation of prostate zones and whole gland has not been widely considered [Ald+20; Mey+19; Sha+17]. The number of human raters was limited to three, and no consensus was built for the prostate zones.

Contributions To cope with the previous limitations, we introduce in this chapter the following contributions :

- To the best of our knowledge, we provide the first prostate zonal segmentation method on 3D T2W images in addition to 2D T2W images. The obtained results are similar to the state of the art.

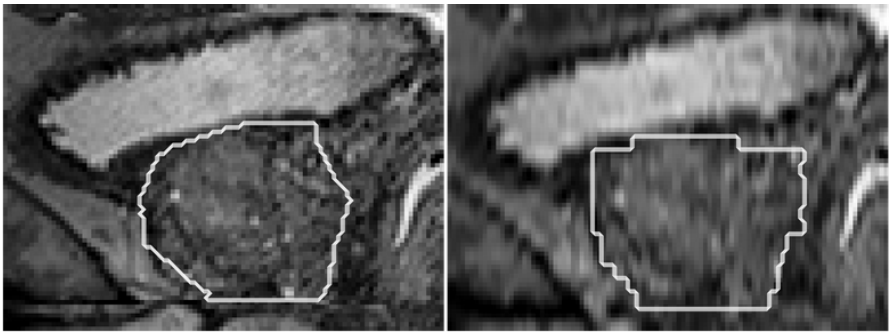


Fig. 4.2.: Left: Sagittal view of a 3D T2W MRI of the prostate and its segmentation by a radiologist (slice thickness: 1mm). Right: Sagittal view of a 2D T2W MRI of the same prostate and its segmentation from the axial views by the same radiologist, resampled to the resolution of the 3D T2W MRI (original slice thickness: 3.25mm).

- We propose a deep learning-based framework for the automatic zonal segmentation of the prostate (transition zone and peripheral zone), including a novel neural network architecture. This architecture takes into consideration the anisotropy of the data, and includes dual attention mechanisms to improve the zonal segmentation. Partition loss functions were defined to enforce the partition of the prostate.
- We compare the generated segmentations with the ones supplied by 7 radiologists of various experience levels, from which we derive a consensus segmentation. We show that our network performs similarly to the radiologists.
- Finally, we show that our method globally preserves both the zonal location and the sectorial position of lesions of the prostate, making it suitable as a helping tool for the detection and grading of lesions. Furthermore we propose the first computerized method to generate a prostate's sector map from its zonal segmentation.

4.2 Material and Methods

4.2.1 Dataset

MRI Scans

In this work approved by our joint institutional review boards, we used a private dataset of 131 3D T2W MRIs from treatment-naive patients who underwent a prostate MRI before the first round of biopsy for clinical suspicion of PCa (linked to an elevated prostate-specific antigen (PSA), a positive Digital Rectal Evaluation and a genetic susceptibility) between October 2013 and July 2019 from 3T Siemens scanners (Siemens Healthcare, Erlangen, Germany) on Pitié-Salpêtrière Hospital, Paris, France (100, 76.6%) and from 3T G.E. scanners (GE Healthcare; Chicago, IL) on Tenon Hospital, Paris, France (31, 23.4%). This dataset was built to have a diversity in terms of shapes, sizes and volumes. The voxel dimensions are [0.36-0.78, 0.36-0.78, 0.5–1.0]mm. A random split

of 91/40 (69%/31%) patients has been used between the training-validation set and the test set.In practice, the former was split into five folds in a cross-validation strategy where four folds served as a training set and the fifth one as a validation set. A minority (35.9%) had at least a clinically significant lesion, which was defined as a lesion with a PI-RADS score ≥ 3, of which 30 being on the training-validation set (33.0% of the training-validation set) and 17 on the test set (42.5%).
To assess the capacity of our network to provide a segmentation preserving not only the zonal location of the lesions but also their sectorial positions, we had access to an additional dataset of 33 3D T2W sequences of prostates with 46 clinically significant lesions that we will call private lesion dataset. These sequences have been acquired on the same scanners than the private dataset between May 2017 and December 2019, with 24 from Tenon Hospital, Paris, France (73%, voxel dimensions: [0.547, 0.547, 0.5]mm) and 9 from Pitié-Salpêtrière Hospital, Paris, France (27%, voxel dimensions: [0.36-0.78, 0.36-0.78, 0.5–1.0]mm).
In addition, we considered the public dataset ProstateX [Lit+17; Lit+14a; Cla+13] to compare our method with prior works. It consists of 204 T2W MRIs taken with 3T Siemens scanners on Radboud University Medical Center, with an in-plane dimension of [0.375-0.6]mm^2 and a slice thickness of [3-4.5]mm, the most frequent resolution being 0.5x0.5x3mm. We excluded three sequences for mismatches with their provided segmentation and randomly split the remaining data into a training-validation set of 141 sequences and a test set of 60 prostates, with a fivefold cross-validation strategy similar to the one used on the private dataset.

Zonal segmentation

The zonal segmentation of the private dataset consists of binary masks of the WG and the TZ. The segmentation of the training-validation set has been performed by a single expert radiologist, whereas on the test set 7 radiologists of various levels of experience provided each a zonal segmentation for each prostate: 3 experts (≥ 1000 prostate MRI interpreted), 2 seniors (≈ 500 prostate MRI) and 2 juniors (≤ 100 prostate MRI). This led to a total of 280 (=40×7) zonal segmentations on the test set, for a rich comparison of performance with radiological experts. The radiologists were instructed to first segment WG and then TZ on the axial plane of the 3D T2W sequence of our cohort. PZ was obtained by subtracting TZ to WG. Segmentation was performed using MedInria, an open-source software (https://med.inria.fr/).

In addition to these segmentations we also generated a consensus segmentation for WG and TZ using the STAPLE algorithm [WZW04] , which describes raters' binary segmentations by Bernoulli distributions and uses an expectation-maximization (EM) algorithm to produce a consensus of the segmentations. We binarized the obtained consensus by only keeping voxels with a probability ≥ 0.75. This threshold was chosen empirically, as its

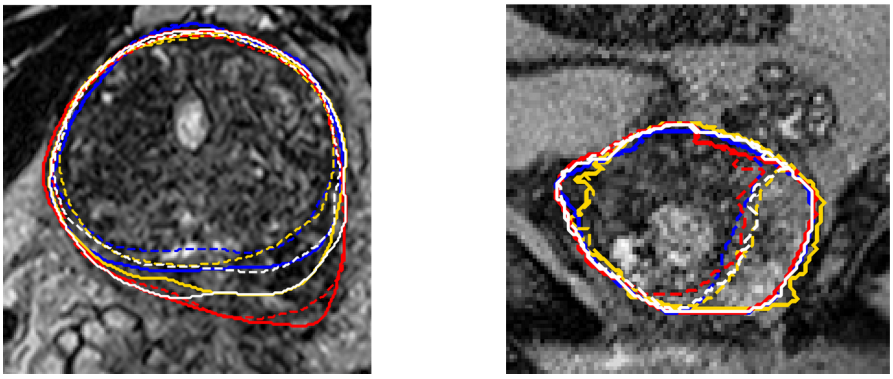


Fig. 4.3.: Inter-rater variability, with segmentations from 3 of the 7 raters (red, blue, yellow) and consensus segmentation from STAPLE using the 7 raters (white), on axial (left) and sagittal views (right). Solid line: whole gland, dashed line: transition zone.

value varies following the different authors between 0.5 and 0.95 [Sui+14; Cox+12; Pop+06]. PZ was then obtained by subtracting the consensus TZ to the consensus WG. An example with different raters segmentation and the consensus can be seen in Fig.4.3. As the segmentations supplied by the radiologists have been prone to intra-rater variability, leading to some gaps between TZ and WG border on the anterior part of the prostate which, after verification with a radiologist, do not belong to PZ, we applied on the initially determined PZ a slicewise 2D erosion, followed by a restriction to its largest connected component - or the two largest components if the second component is at most three times smaller than the largest component, then a slicewise 2D dilation. Examples of the impact of those corrections can be seen in Fig. 4.4.

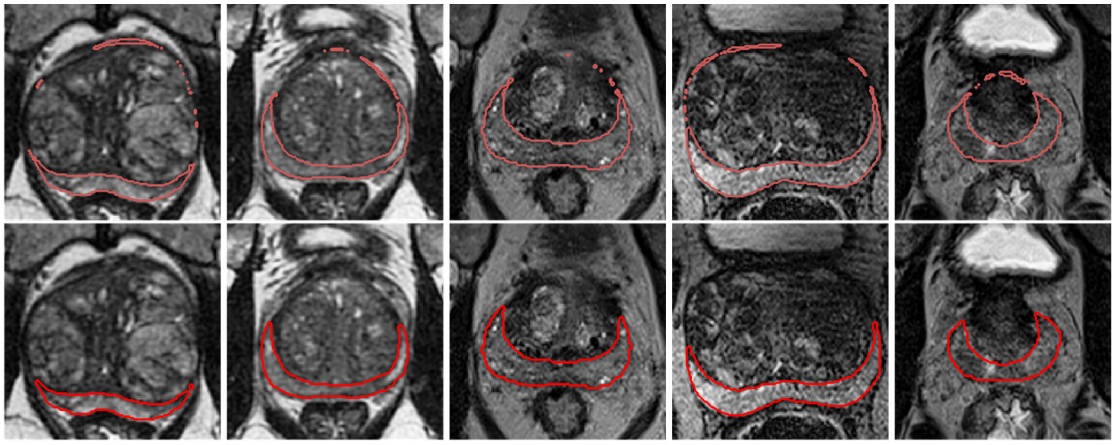


Fig. 4.4.: Top: Examples of segmentations of the peripheral zone (= whole gland - transition zone) before correction. Bottom: Same segmentations after correction.

For ProstateX, we used the zonal segmentation provided by Cuocolo et al. [Cuo+21a; Cuo+21b]. No zonal segmentation has been done on the private lesion dataset.

Lesion placement

On the private dataset, each expert or senior radiologist (5 of the 7 raters) provided a segmentation of the lesions, from which we derived a consensus using the STAPLE algorithm as explained in section 4.2.1. On the private lesion dataset, a radiologist provided for each lesion its sectorial position according to the 27 regions of interest sector map defined in Dickinson et al. [Dic+11] - with for some lesions 2 or 3 sectors indicated, as well as their size, their PI-RADS score and their Likert score [Ros+13].

4.2.2 Objectives and architecture of the networks

For the zonal segmentation of the prostate we chose a framework with two cascaded UNet-based neural networks. A sum-up of the whole framework is available in Fig.4.5.

The objective of the first network, that we will call global location network, is to roughly segment WG in order to generate a bounding box around the prostate with a fixed size of 8cm according to transverse and antero-posterior directions, and a margin of 10cm above and below that segmentation. It takes as inputs patches of size $192 \times 192 \times 32$ voxels at a resolution of 1x1x3mm, and uses as a loss function the Generalized Dice loss function [Sud+17]. For image resampling we used the python module SimpleITK [Low+13], with BSpline interpolation for images and nearest neighbors interpolation for masks.
The second network, or zonal segmentation network, operates at a higher resolution of 0.5x0.5x1mm and takes input patches of 96x96x48 voxels, which appeared to be a good compromise between the quantity of information brought to the network and the available GPU memory. Resampled and rescaled images had a size of 160x160x[57-110] voxels. The loss functions used for the training of the zonal segmentation network are defined in section 4.2.4. Detailed architecture of the networks is provided in Fig.4.6. In the final framework we used UFNet for both the global location network and the zonal segmentation network, but with fewer parameters for the former. We combined in UFNet two methods to take into account the existing anisotropy of the data. As in Meyer et al. [Mey+21], we used anisotropic MaxPooling and 3D Deconvolutions with varying kernel sizes, and we replaced the classic k×k×k kernels by a combination of k×k×1 and 1×1×k as presented in Fig.6.d inspired by Liu et al. [Liu+]. We used the activation function LeakyReLU with a parameter α=0.1 except for the last layer which uses sigmoid activation. In UFNet we used deep supervision, which consists in introducing upscaled versions of intermediate results from the decoder into the final result as a form of regularization. We also performed Dropout [Sri+14] and Instance normalization [UVL17] to fight against overfitting and improve stability of our network, and we used attention modules which are presented in section 4.2.3.

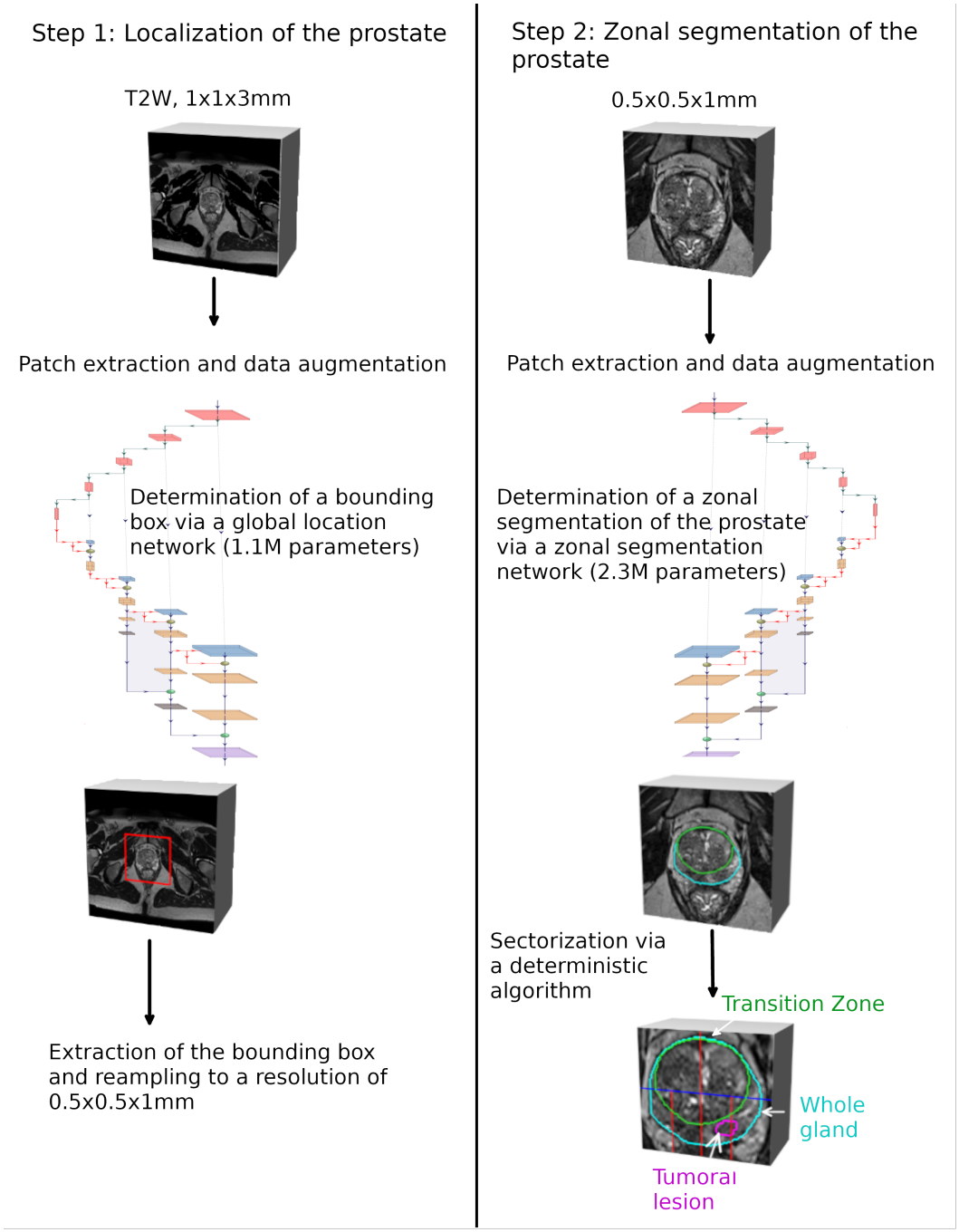


Fig. 4.5.: Framework for the zonal segmentation of the prostate. The global location network extracts from a T2W sequence a bounding box, which serves as input to the zonal segmentation network, generating the zonal segmentation of the whole gland (cyan), the transition zone (green) and the peripheral zone. Finally, a sector map is constructed from the zonal segmentation to provide information about the location of the lesion (magenta)

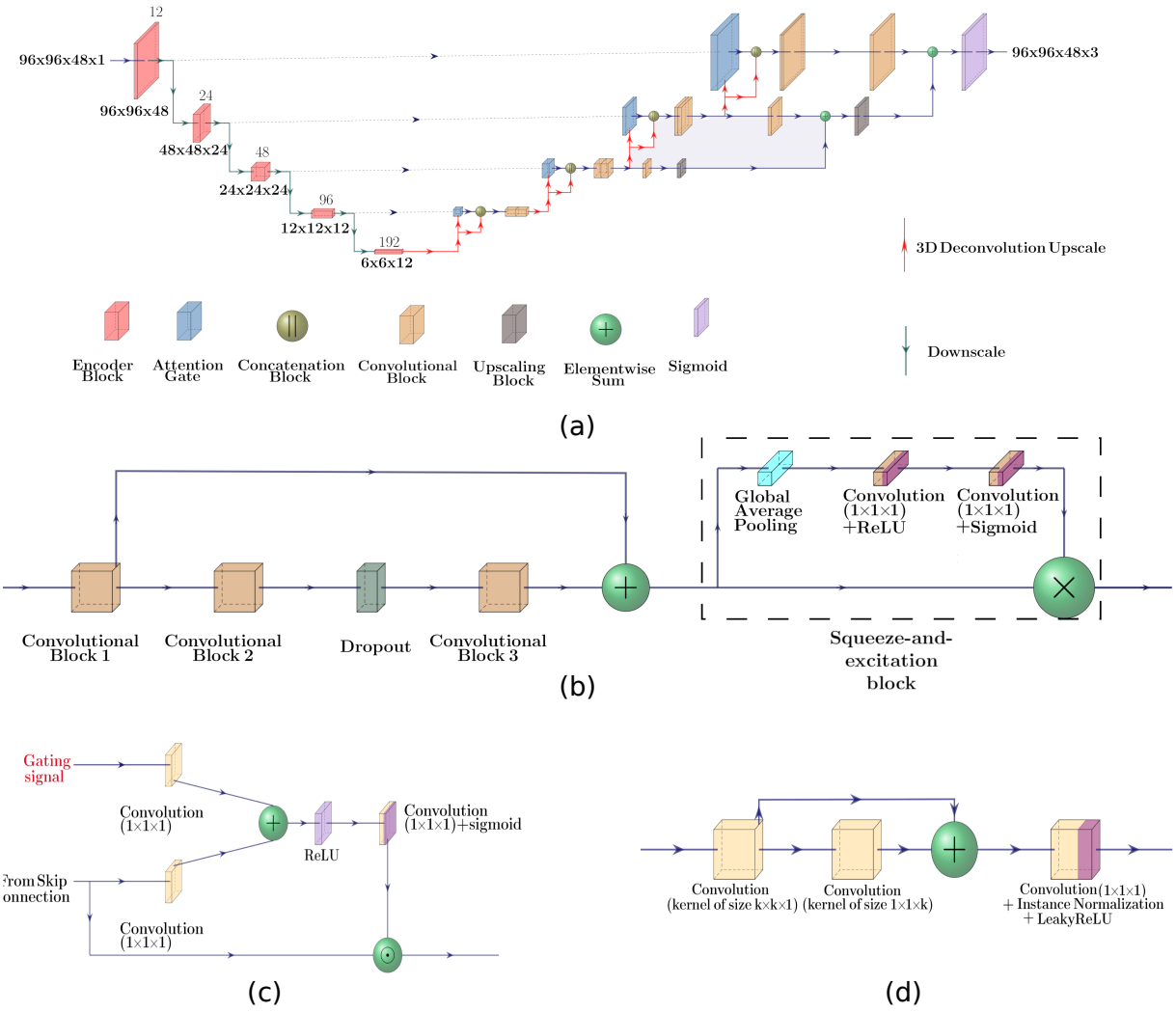


Fig. 4.6.: Architecture of the zonal segmentation network (a) and its components: the encoder block (b), the attention gate (c) and the used convolutional block (d). Values above layers in (a) correspond to the number of output filters in the convolutions performed in this layer. Architecture of global location network is similar but with a lower number of parameters.

4.2.3 Attention mechanisms

Attention in deep learning consists in encouraging the network to focus on some specific parts of the data, deemed with particularly relevant information for its task, and to downplay the importance of the rest of the data. The information can be highlighted based on its spatial location (spatial attention) or on the characteristics of the feature maps that contains it (channel attention). Here, we combined both channel attention and spatial attention through two different methods: respectively Squeeze-and-Excitation modules and attention gates [HSS18; Okt+18].

Squeeze-and-Excitation modules The objective of Squeeze-and-Excitation modules [HSS18] is to put more focus on the feature maps that provides useful information for the segmentation task. Given $\boldsymbol{x}$ the input of the module, of size $W \times H \times D \times C$, we define $\widetilde{\boldsymbol{x}} = [\widetilde{\boldsymbol{x}}_1, \widetilde{\boldsymbol{x}}_2, \ldots, \widetilde{\boldsymbol{x}}_C]$ (size: $W \times H \times D \times C$) the output of the module as

$$\boldsymbol{f_{sq}} = \mathrm{ReLU}(\boldsymbol{W_1}^T \mathrm{GAP}(\boldsymbol{x})) \tag{4.1}$$

$$\boldsymbol{f_{ex}} = \sigma(\boldsymbol{W_2}^T \boldsymbol{f_{sq}}) \tag{4.2}$$

$$\forall c \in [1, C],\ \widetilde{\boldsymbol{x}}_c = (\boldsymbol{f_{ex}})_c \cdot \boldsymbol{x}_c \tag{4.3}$$

GAP() is a 3D Global Average Pooling module, $\sigma()$ is the sigmoid function and $\boldsymbol{W_1}$, $\boldsymbol{W_2}$ are convolutional kernels (See Fig.6.b). Both $\boldsymbol{f_{sq}}$ and $\boldsymbol{f_{ex}}$ have a size of 1×1×1× C. We used Squeeze-and-Excitation modules on the encoder part of our network.

Attention gates Introduced in biomedical segmentation by Oktay et al. [Okt+18] , the principle of attention gates is to highlight in skip connections spatial zones with the more informative content. More precisely, if $\boldsymbol{x}$ and $\boldsymbol{g}$ are respectively the feature maps from skip connection and the gating signal with coarser information, we define

$$\boldsymbol{q_{att}} = \psi^T \left(\mathrm{ReLU}(\boldsymbol{W_g}^T \boldsymbol{g} + \boldsymbol{W_x}^T \boldsymbol{x} + \boldsymbol{b_g})\right) + \boldsymbol{b_\psi} \tag{4.4}$$

$$\boldsymbol{\alpha} = \sigma(\boldsymbol{q_{att}});\ \widehat{\boldsymbol{x}} = \boldsymbol{\alpha} \odot \boldsymbol{x} \tag{4.5}$$

with $\boldsymbol{W_g}$ and $\boldsymbol{W_x}$ convolutional kernels, ψ a 1x1x1 convolutional kernel, $\boldsymbol{b_\psi}$ and $\boldsymbol{b_g}$ biases and $\odot$ the Hadamard product. A visualization of the architecture is available in Fig.6.c. We used attention gates on each layer of our network.

Combination of both attention methods has already been used for PCa detection [Zha+19; SHH21] but to the best of our knowledge it is the first time it is applied to prostate zonal segmentation.

4.2.4 Loss functions for the zonal segmentation network

We used as the main loss function the mean of 1-DSC for WG, for TZ and for PZ.

Several approaches have been proposed in prior works to enforce the partition of the prostate such that WG = TZ∪PZ and TZ∩PZ = ∅. First, one can only segment WG and TZ and build PZ by subtracting TZ to WG [Run+19]. Second, one can learn to segment WG and both zones, and rely on postprocessing to enforce the partition [Mey+21]. Another approach is to segment WG and to classify its voxels as either TZ or PZ [Bar+21]. In this chapter we propose another approach only based on segmentation, with partition loss functions dedicated to obtain a partition of the prostate. If we consider $p^{\mathrm{WG}}, p^{\mathrm{TZ}}, p^{\mathrm{PZ}} \in [0,1]^N$ the probabilistic segmentation by the zonal segmentation network of respectively WG, TZ and PZ, then we define the two losses:

$$\mathcal{L}^1_{aux}(p^{\mathrm{WG}}, p^{\mathrm{TZ}}, p^{\mathrm{PZ}}) = \frac{1}{N}\sum_{i=1}^{N}(p_i^{\mathrm{WG}} - p_i^{\mathrm{TZ}} - p_i^{\mathrm{PZ}})^2 \tag{4.6}$$

and

$$\mathcal{L}^2_{aux}(p^{\mathrm{WG}}, p^{\mathrm{TZ}}, p^{\mathrm{PZ}}) = \frac{\sum_{i=1}^{N}(p_i^{\mathrm{TZ}}.p_i^{\mathrm{PZ}})}{(\sum_{i=1}^{N} p_i^{\mathrm{WG}}) + \epsilon}. \tag{4.7}$$

The objective of $\mathcal{L}^1_{aux}$ is to ensure that the segmentation obtained by the network is coherent i.e. that the segmentations for TZ and for PZ are within the limits of the segmentation of WG, and that they totally cover it. For this reason, we chose to penalize not only segmented voxels outside the whole gland segmentation but also the voxels that are included in both the TZ and PZ segmentations. The objective of $\mathcal{L}^2_{aux}$ is to enforce this lack of intersection.

4.2.5 Sector map construction

Another objective of our work is also to estimate the efficiency of our method to correctly assess the sectorial position of lesions. To this end, we designed an algorithm taking as input the zonal segmentation of a prostate and constructing the associated sector map. We chose to base our sector map on the 27 regions of interest sector map defined in Dickinson et al. [Dic+11]. We defined the limits between sectors as follows:

- According to the longitudinal axis: We split the prostate on three equal-sized parts corresponding to the apex, to the midgland and to the base, taking as extremes

points the lowest and highest positions of WG segmentation. Following axes are computed separately for each third of the prostate.

- According to the antero-posterior axis: We split each part across its median sagittal slice x_{mid}.
- According to the transverse axis: For each part we take the mean of the extreme positions of their slices (their leftest and rightest positions) according to the transverse axis x_{left} and x_{right}, and we define the positions of the inner subdivisions as: $x_{\text{midleft}} = 0.4(x_{\text{mid}} - x_{\text{left}}) + x_{\text{left}}$ and $x_{\text{midright}} = 0.6(x_{\text{right}} - x_{\text{mid}}) + x_{\text{mid}}$.

The main difference between our constructed sector map and the 27 regions of interest sector map [Dic+11] is the absence of the 3 sectors related to the anterior fibromuscular stroma, as we did not segment this particular zone but included it into TZ. For this reason we included lesions located in the anterior fibromuscular stroma among the TZ lesions. An illustration is provided in Fig.4.7

The zonal location of a lesion in the prostate is defined as the zone (PZ or TZ) with the highest proportion of lesions' voxels, and the sectorial position of the lesion as the sector within the considered zone with the highest proportion of lesion's voxels.

4.3 Experimental design

4.3.1 Training of the network

Training on the private dataset was performed using an Intel(R) Xeon(R) Gold 6246R CPU and a NVIDIA Tesla V100 SXM2 32GB GPU, a NVIDIA Tesla T4 16GB GPU having also been used for the ProstateX dataset. We used Keras [Cho+15] and Tensor Flow 2.4.1 [Aba+15] as a deep learning framework.
The training of the global location network has been done with a batch size of 4, using RMSProp as a gradient optimizer with an initial learning rate of 5e-4. The training of the zonal segmentation network has been done with a batch size of 8, using Adam as a gradient optimizer with an initial learning rate of 5e-4, and attributing loss weights of 1, 0.1 and 0.01 respectively to $\mathcal{L}$, $\mathcal{L}^1_{\text{aux}}$ and $\mathcal{L}^2_{\text{aux}}$. Several values of parameters (learning rate, batch size...) were tested before choosing those values.
For both networks we adopted a maximal number of epochs of 500, with a policy of early stopping if the validation loss did not improve for 70 epochs. We also adopted a policy of reduction of the learning rate with a multiplication by 0.2 in case of stagnation of the validation loss for 30 epochs. The dropout rate was set to 0.3.
To improve the performance of the network we artificially increased the number of images used during the training thanks to data augmentation through the Python module batchgenerators [Ise+20]. This module allowed us to apply several transformations such as rotation according to the longitudinal axis, mirror transform along the antero-posterior

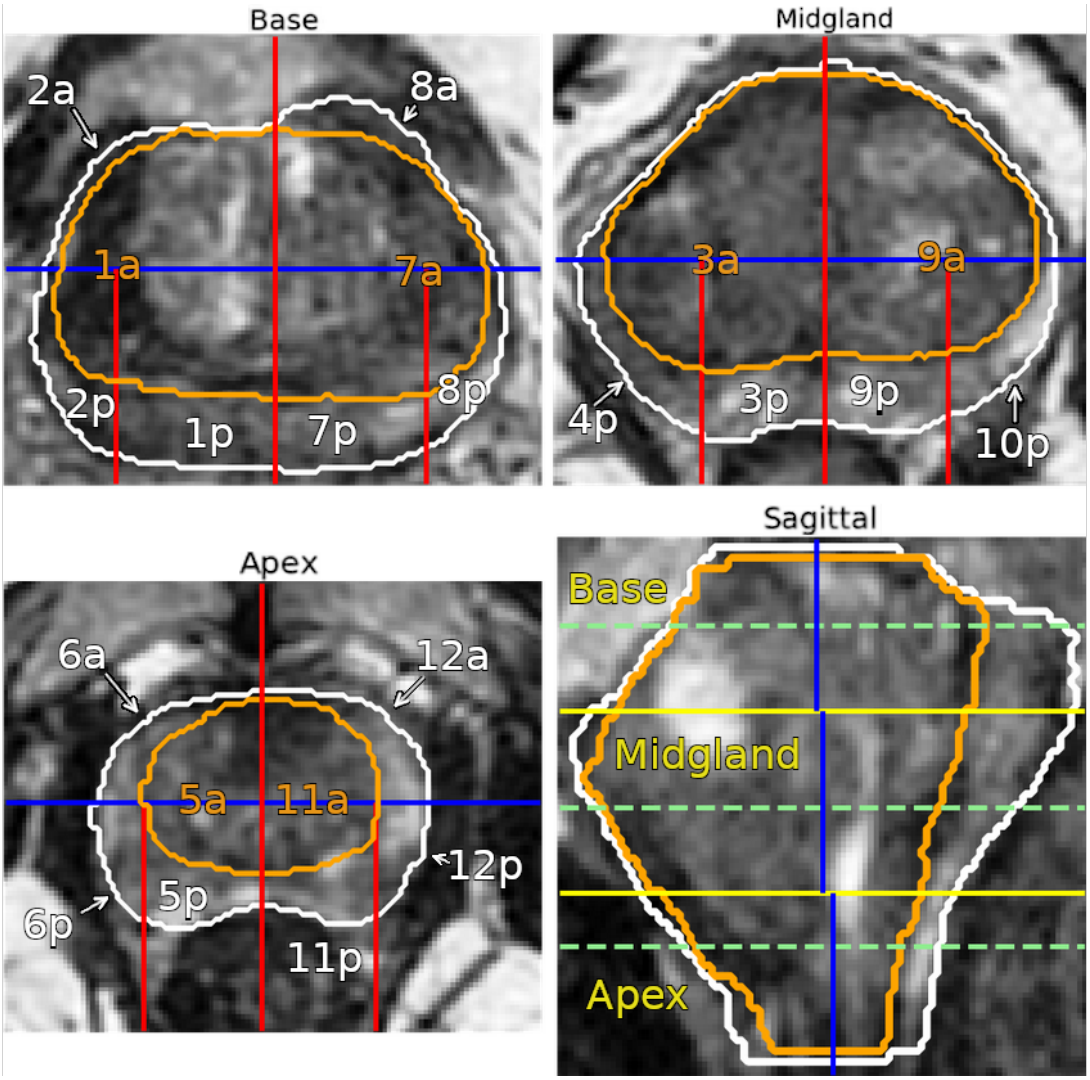


Fig. 4.7.: Sector map of the prostate according to axial views in the base (top left), the midgland (top right) and the apex (bottom left), and to the sagittal view (bottom right). White: whole gland, orange: transition zone. The blue axis separates the anterior from the posterior of the prostate, the red axes are left-right based separations and the yellow axes represent the separations between the base (top left), the midgland (top right) and the apex (bottom left). The green dashed lines on the sagittal view indicate the location of the different axial views.

axis, elastic transform or intensity transforms such as gamma transform. The intensity of each sequence has also been normalized via the subtraction of its voxels' mean value and a division by their standard deviation.

4.3.2 Test and postprocessing

We used test-time augmentation [Wan+19b; Sha+20], which consists in applying different transformations to an image during the test procedure, to use the network on each of these transformed images and then to revert those transforms and to combine the obtained results onto one final prediction by taking their mean, to improve the final segmentation and the robustness of the process. Transformation applied were all combinations of flip along the antero-posterior axis with a rotation of ±10°, for a total of 6 images.
In postprocessing, to apply our method on the whole image we applied a sliding window strategy, where patches of size 96x96x48 were extracted with steps of (24, 24, 12) voxels according to each dimension and where the contribution of each patch to a specific voxel is divided by the number of patches contributing to this voxel. Finally, after reconstruction of the segmentation from the patches, we applied a threshold of 0.5 to obtain WG segmentation that we restrained to its largest connected component according to the longitudinal direction. Within WG segmentation we defined TZ segmentation as the voxels for which the probability to be in TZ was higher than the probability to be in PZ, and conversely for PZ.

4.4 Results

The main metrics we used to estimate the performance of our network were DSC and 95% Hausdorff distance (HD95%). We mainly compared three networks for the zonal segmentation network: a 3D UNet as presented in Isensee et al. [Ise+18], introducing context and localization modules and which served as a basis for other networks, the network UNetV2 with elements presented in section 4.2.2 to take into account the image anisotropy and UFNet which adds deep supervision, attention modules and the partition losses. These 3 networks have respectively 2.15M, 2.28M and 2.32M parameters. A UFNet with 1.03M parameters was used as the global location network.
To improve the segmentations and their stability, inspired by Isensee et al. [Ise+21], we combined the results of 5 neural networks obtained through cross-validation to provide a final, more precise ensembled segmentation by taking the mean of their prediction before postprocessing, including test-time augmentation. These networks are named with the suffix -E.

4.4.1 Results on private dataset

On this dataset we compared the outputs of the networks to the consensus obtained from the 7 radiologists and corrected as described in section 4.2.1 (i.e. slicewise restriction of PZ to its largest components). The same correction was applied to the outputs of the network.

Segmentation results

Results obtained for the metrics on WG, TZ and PZ are given in Table 4.1. They are illustrated in Fig.4.8, along with statistical differences between the performance of the networks. The global location network provided an adequate bounding box of the prostate, i.e. which surrounds the prostate without crossing it, for all images. Some examples of zonal segmentation are provided in Fig.4.9 and Fig.4.10. The mean time to process one patch was 0.8s, and the mean time to process a sequence on this dataset was 4.5s.

Model	Zone	DSC (in %)	p-val/UNet-E	HD95% (in mm)	p-val/UNet-E
UFNet-E (5×2.32M parameters)	WG	**93.90 ± 2.85**	***	**2.59 ± 0.98**	**
	TZ	**91.00 ± 4.34**	***	**3.20 ± 0.90**	***
	PZ	**79.08 ± 7.08**	***	**3.87 ± 1.65**	*
UNetV2-E (5×2.28M parameters)	WG	93.65 ± 2.46	*	2.73 ± 0.80	*
	TZ	90.73 ± 4.09	*	3.42 ± 0.94	*
	PZ	78.11 ± 7.58	*	4.26 ± 2.36	>.05
UNet-E (5×2.15M parameters)	WG	93.48 ± 2.54	-	2.81 ± 0.80	-
	TZ	90.47 ± 4.28	-	3.53 ± 0.97	-
	PZ	77.51 ± 7.95	-	4.09 ± 1.93	-
Model	Zone	DSC (in %)	p-val/UNet	HD95% (in mm)	p-val/UNet
UFNet	WG	93.45 ± 2.89	***	2.81 ± 0.93	**
	TZ	90.45 ± 4.51	***	3.41 ± 0.93	***
	PZ	77.80 ± 7.79	***	4.15 ± 1.92	*
UNetV2	WG	93.16 ± 2.51	>.05	2.91 ± 0.80	>.05
	TZ	90.24 ± 4.16	*	3.52 ± 0.93	>.05
	PZ	76.52 ± 8.03	*	4.37 ± 2.09	>.05
UNet	WG	92.87 ± 2.79	-	3.21 ± 1.32	-
	TZ	89.85 ± 4.56	-	3.83 ± 1.17	-
	PZ	75.95 ± 8.47	-	4.38 ± 1.91	-

Tab. 4.1.: Comparison between our method and UNet on our private dataset after correction. Top: Results for the ensembling of 5 networks of 3 different types, obtained from cross-validation. Best results for each considered metric are in bold. Bottom: Mean of the results from the 5 networks used in the ensemble version. Signed-rank Wilcoxon test with Bonferroni-Holm correction has been used to assess statistically significant differences and to compute p-values for ensembled networks and for mean of networks on each fold. Significant differences are indicated (* : p-value ≤ 0.05; **: p-value ≤ 0.01; ***: p-value ≤ 0.001)

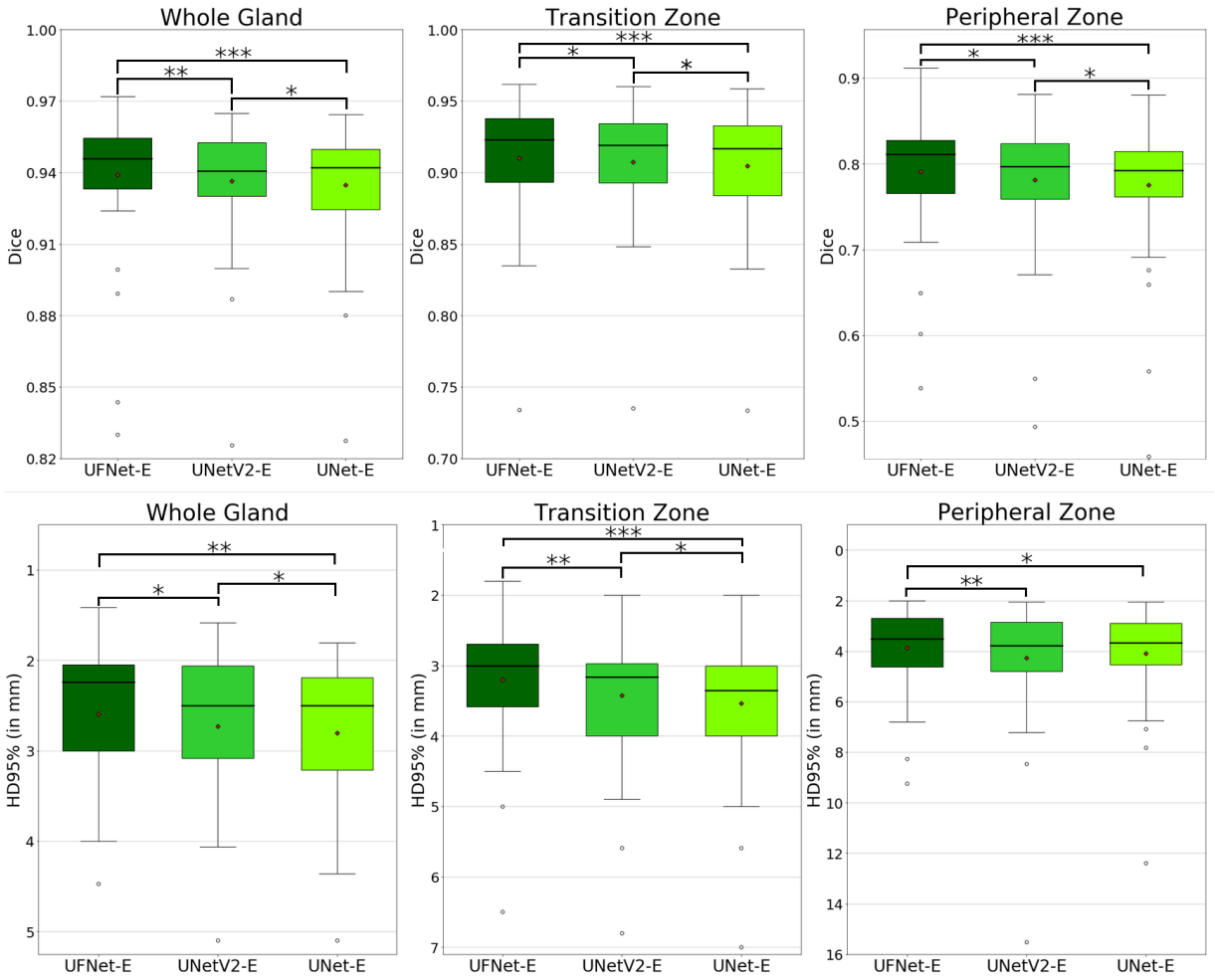


Fig. 4.8.: Metrics between network segmentation and consensus segmentation on private dataset for all three networks. Top: Dice for whole gland, transition zone and peripheral zone. Bottom: Hausdorff distance for whole gland, transition zone and peripheral zone. Black line and red point are respectively median and mean. Signed-rank Wilcoxon test with Bonferroni-Holm correction has been used to assess statistically significant differences and to compute p-values. Significant differences are indicated (* : p-value ≤ 0.05; **: p-value ≤ 0.01; ***: p-value ≤ 0.001)

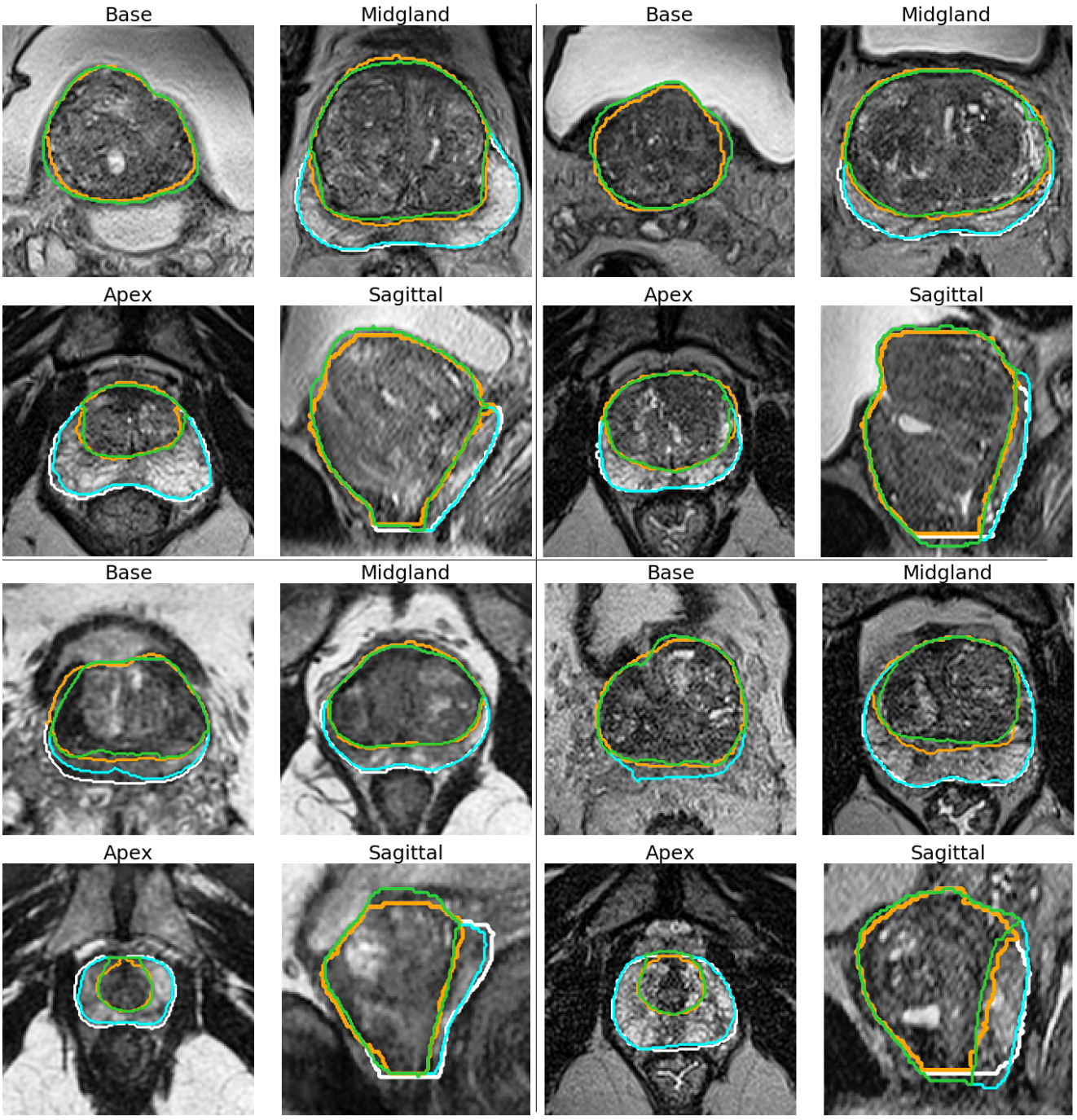


Fig. 4.9.: Good segmentations on the private dataset, with axial views of the base (top left), the midgland (top right) and the apex (bottom left), and sagittal view (bottom right). White: Ground truth whole gland, Orange: Ground truth transition zone, Cyan: Network-segmented whole gland, Green: Network-segmented transition zone.

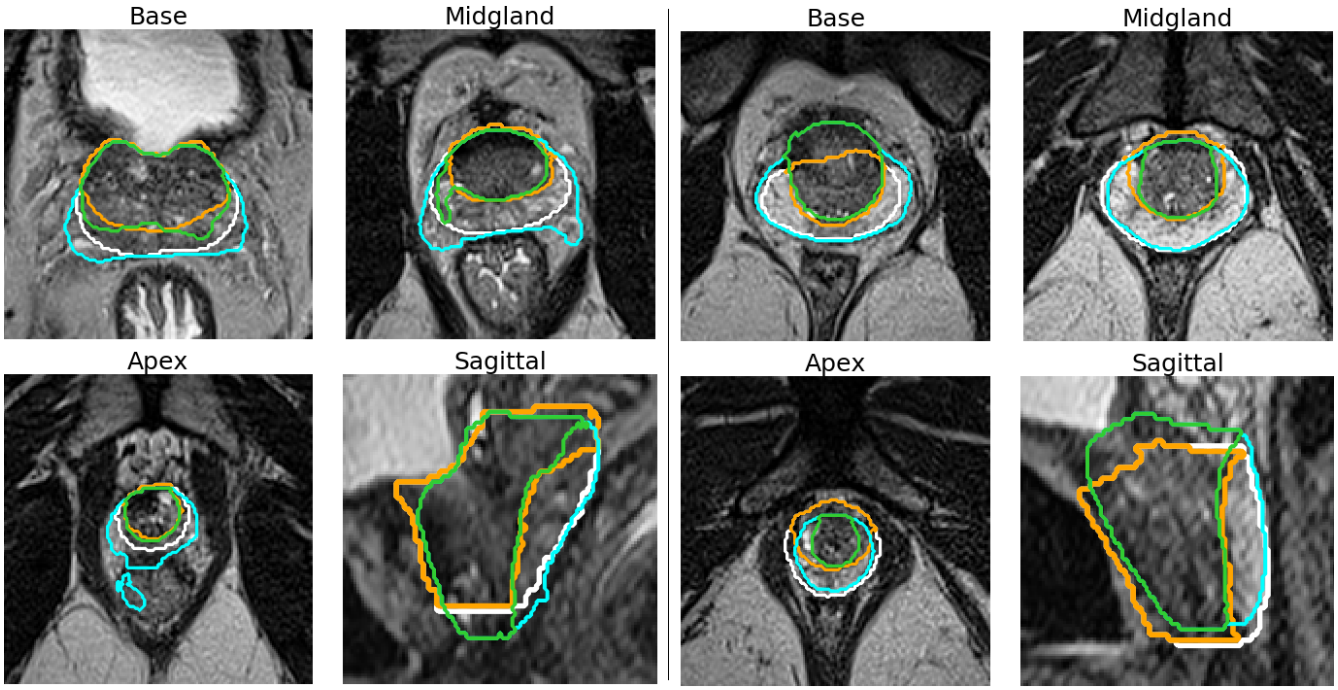


Fig. 4.10.: Poor segmentations on the private dataset, with axial views of the base (top left), the midgland (top right) and the apex (bottom left), and sagittal view (bottom right). White: Ground truth whole gland, Orange: Ground truth transition zone, Cyan: Network-segmented whole gland, Green: Network-segmented transition zone.

Comparison with radiologists

We compared these results to the segmentations provided by the 7 radiologists on the same dataset, as can be seen in Fig.4.11. While not being as good as the best radiologists i.e. those with the closest segmentations to the consensus, UFNet-E obtained results similar to radiologists'. If we rank UFNet-E and all radiologists according to how close they are to the consensus for all metrics according to their means, UFNet-E is ranked between the 3th place and the 7th place, with a global 5th place.

4.4.2 Lesion positions

Test set

On the 17 prostates of the test set with a lesion, we applied our sector map construction algorithm and determined the location of the lesion using the sector map derived from the ground truth as the true sector map. On the zonal location of the lesion, we obtained a 100% accuracy, whereas on the sectorial position we obtained an accuracy of 88% (15 out 17 cases). Observed lesion position errors are due to differences on the delineations of the apex, midgland and base between the ground truth segmentation and the segmentation from the network. Examples of lesion locations are provided in Fig.4.12.

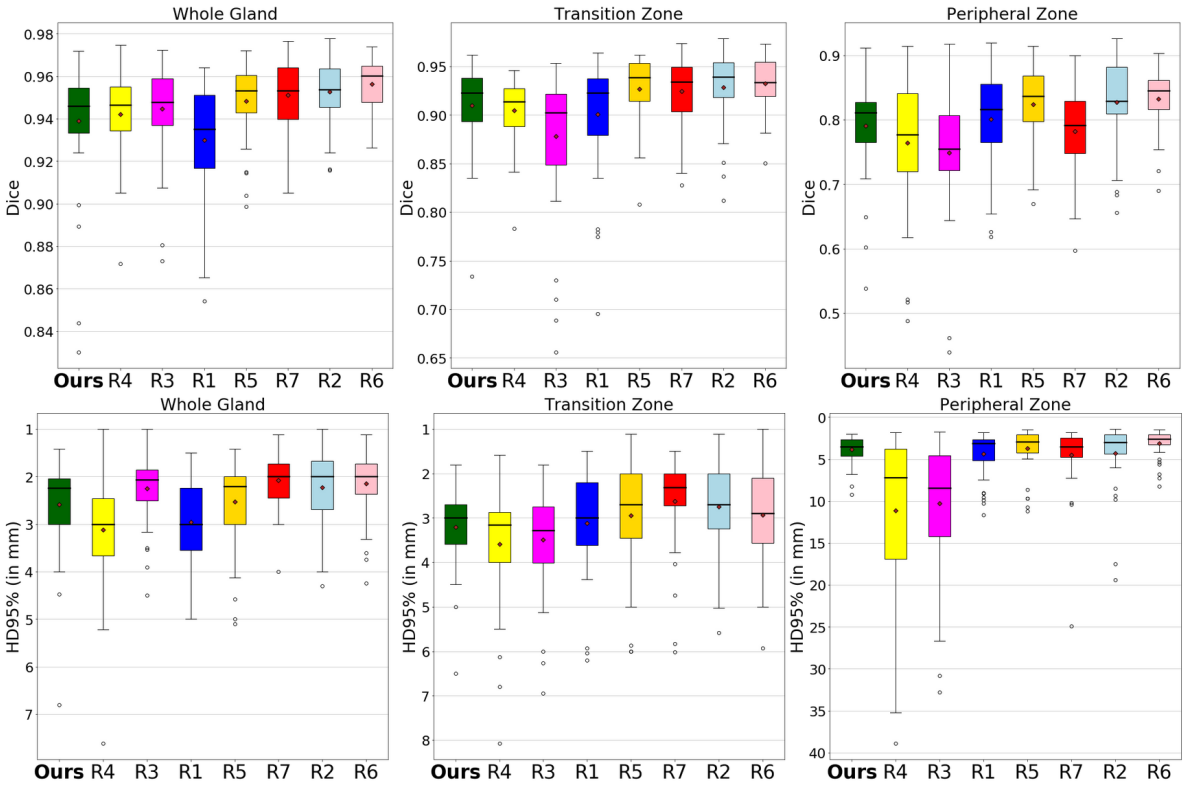


Fig. 4.11.: Metrics between UFNet-E (Ours) segmentation and the consensus segmentation on private dataset, side to side with the metrics for raters' segmentation ranked in increasing order of performance (the rater k being written as Rk). Top: Dice for whole gland, transition zone and peripheral zone. Bottom: 95% Hausdorff distance for whole gland, transition zone and peripheral zone. Black line and red point are respectively median and mean values.

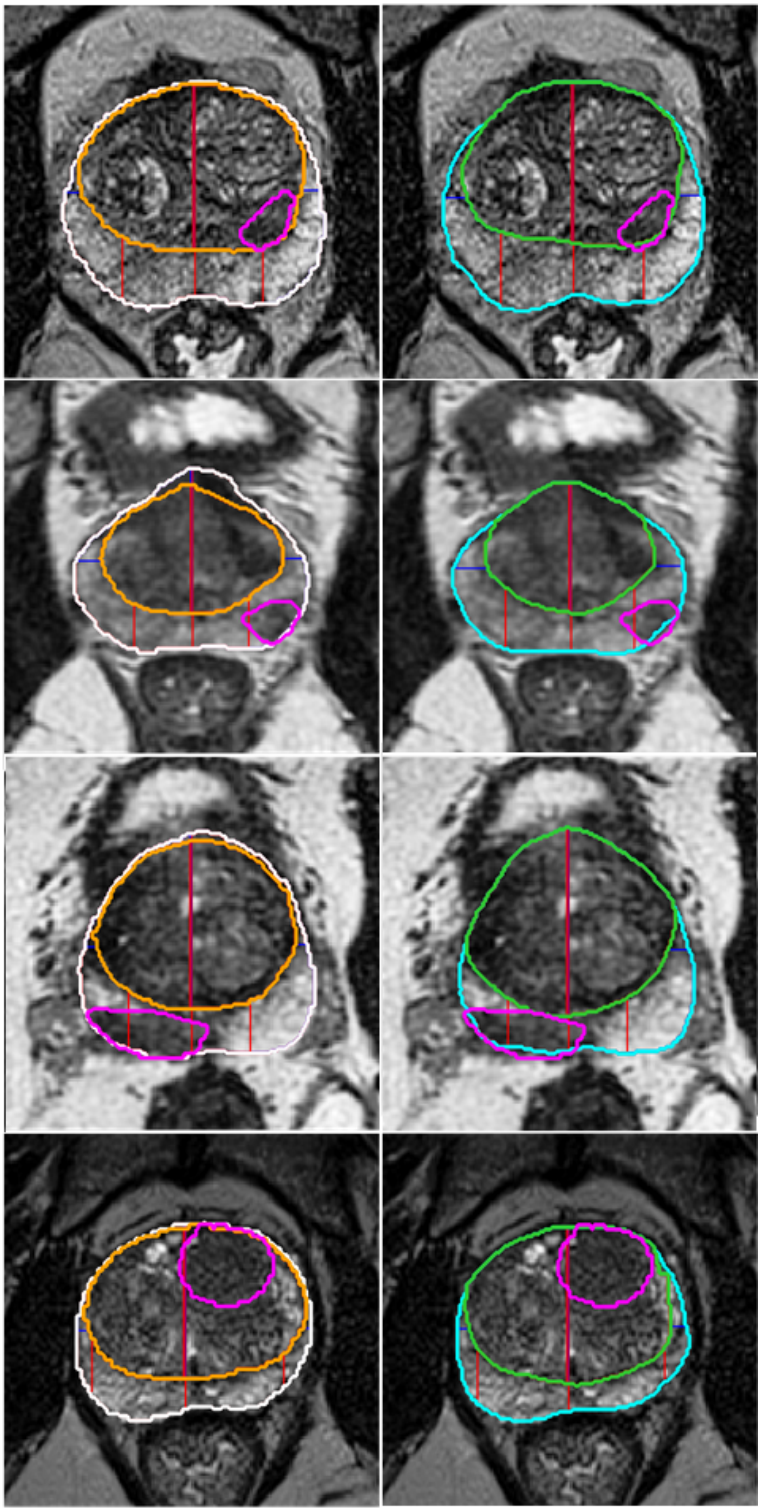


Fig. 4.12.: Examples of correct lesion placement on the test set. Left: placement derived from the true segmentation (white: whole gland, orange: transition zone). Right: placement on the sector map computed from the network segmentation (cyan: whole gland, green: transition zone). Separations between sectors are in red (antero-posterior direction) and in blue (transverse direction). Lesions are in magenta (consensus segmentation from 5 radiologists).

Private lesion set

On the private lesion set, we obtained an accuracy for the zonal location of lesions of 91% (42 out of 46), and of 74% for their sectorial positions (34 out of 46).
All 4 cases where the lesion has been placed in the wrong zone are cases with a configuration similar to the configuration presented on the right image of Fig.4.13, i.e. with a lesion in the anterior part of PZ near the border with TZ, and with a voxel intensity closer to TZ's than to PZ's. Cases with a sectorial position error correspond to either lesions located according to our algorithm across two adjacent sectors - including the true sectorial position - but with a majority of voxels in the wrong sector (4 cases), or a lesion well located according to transverse and antero-posterior directions but not according to longitudinal direction, for example lesions located in the midgland whereas the true position is in the base (4 cases).

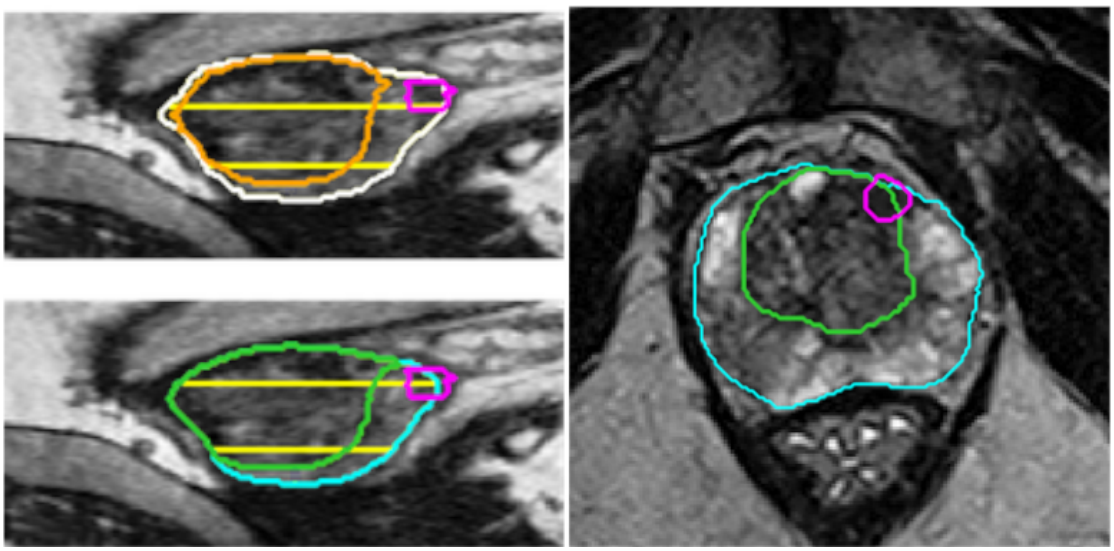


Fig. 4.13.: Left: Example of wrong lesion placement on the test set. Top: placement derived from the true segmentation (white: whole gland, orange: transition zone), in the base. Bottom: placement on the sector map computed from the network segmentation, in the midgland (cyan: whole gland, green: transition zone). Limits between base/midgland/apex are in yellow, the lesion is in magenta. Right: Example from the private lesion set with the lesion (in magenta) located in TZ by the network whereas the radiologists located it in PZ.

4.4.3 Results on ProstateX

Segmentation results

Results obtained on ProstateX are given on Table 4.2, and are illustrated in Fig.4.16. A boxplot graph illustrating the performances according to the different metrics is available in Fig. 4.14. The global location network provided an adequate bounding box of the prostate for all images. The mean inference time of the network for each patch was 0.09s, adding up to 3.5s on the whole sequence.

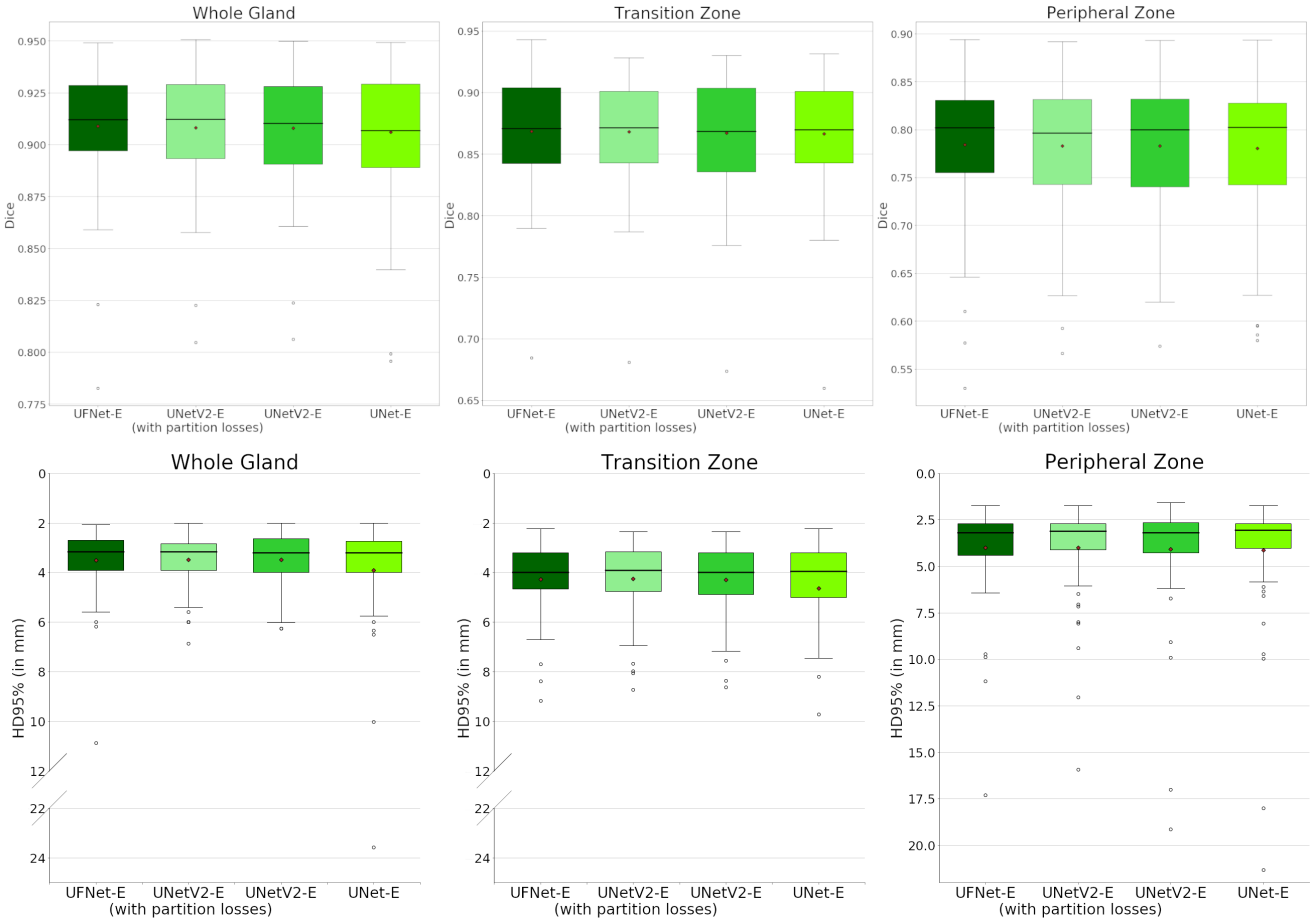


Fig. 4.14.: Metrics between network segmentation and ground truth segmentation on ProstateX dataset for all three networks. Top: Dice for whole gland, transition zone and peripheral zone. Bottom: 95% Hausdorff distance for whole gland, transition zone and peripheral zone. Black line and red point are respectively median and mean values. Signed-rank Wilcoxon test with Bonferroni-Holm correction has been used to assess statistically significant differences, but no differences were found.

Model	Zone	DSC (in %)	HD95% (in mm)
UFNet-E	WG	**90.90 ± 2.94**	3.50 ± 1.36
	TZ	**86.84 ± 4.33**	4.27 ± 1.40
	PZ	**78.40 ± 7.31**	**4.00 ± 2.54**
UNetV2-E (with partition losses)	WG	90.83 ± 2.81	**3.48 ± 1.12**
	TZ	86.82 ± 4.53	**4.25 ± 1.42**
	PZ	78.29 ± 7.14	4.01 ± 2.48
UNetV2-E (without partition losses)	WG	90.81 ± 2.82	3.48 ± 1.11
	TZ	86.73 ± 4.31	4.29 ± 1.43
	PZ	78.31 ± 7.12	4.08 ± 3.04
UNet-E	WG	90.59 ± 3.09	3.91 ± 2.89
	TZ	86.66 ± 4.56	4.64 ± 3.02
	PZ	78.04 ± 7.60	4.14 ± 3.34
UFNet	WG	90.62 ± 2.92	3.58 ± 1.15
	TZ	86.45 ± 4.45	4.38 ± 1.41
	PZ	77.81 ± 7.35	4.06 ± 2.38
UNetV2 (with partition losses)	WG	90.48 ± 2.86	3.59 ± 1.11
	TZ	86.38 ± 4.39	4.37 ± 1.37
	PZ	77.45 ± 7.42	4.26 ± 3.46
UNetV2 (without partition losses)	WG	90.47 ± 2.88	3.66 ± 1.19
	TZ	86.31 ± 4.58	4.44 ± 1.44
	PZ	77.53 ± 7.54	4.32 ± 3.94
UNet	WG	90.25 ± 3.04	3.86 ± 1.93
	TZ	86.19 ± 4.58	4.60 ± 2.10
	PZ	77.27 ± 7.75	4.24 ± 3.26

Tab. 4.2.: Comparison between our method and UNet on the ProstateX dataset.
Top: The -E signals the use of an ensemble of 5 networks. Best results for each considered metric are in bold. No statistical differences were found between the ensembles of networks.
Bottom: Mean of the results from the 5 networks used in the ensemble version.

Evolution of Dice score during training

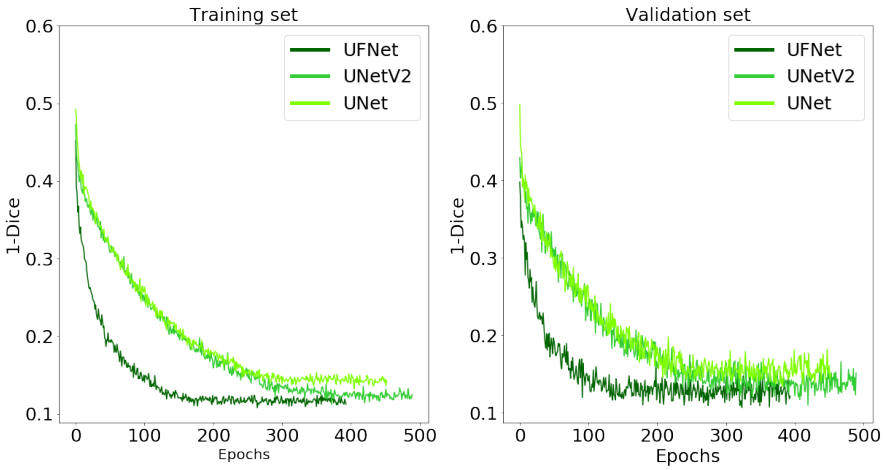


Fig. 4.15.: Evolution of the Dice coefficient on both training and validation sets during training (on one fold of ProstateX).

We also estimated on ProstateX the impact of the different architectures on the evolution of DSC during the training. These results are illustrated in Fig.4.15. It appeared that taking into account the anisotropy does not have an important impact on the course of the training, contrary to the introduction of attention modules which help to speed up the training.

4.5 Discussion

In this work, we used a framework composed of two successive neural networks to generate a zonal segmentation of the prostate on both 2D and 3D T2W sequences. The global location network has been able to provide an adapted bounding box of the prostate, and the zonal segmentation network has been able to generate an accurate zonal segmentation of the prostate. This framework not only allows us to detect the prostate despite its relatively small size (around 5% of the image on the private dataset) but also to take advantage of the better resolution on 3D T2W sequences while only requiring a reasonable amount of memory.

The results for the segmentation of the different zones of the prostate are comparable with the results of the state of the art, as for example Bardis et al. [Bar+21] obtained a mean Dice of 94.0% for the whole prostate, 91.4% for the transition zone and 77.6% for the peripheral zone. On ProstateX we obtained DSC similar or superior to Cuocolo et al. [Cuo+21a] with segmentations from the same dataset, showing the efficiency of our method. Those differences could be explained by the use of 3D networks (compared to 2D used by Cuocolo et al. [Cuo+21a]), of attention modules, and by the difference in training set sizes and number of training epochs.

The use of attention modules and anisotropy-adapted modules appeared to have a positive impact, with fewer epochs for training with attention modules and overall better

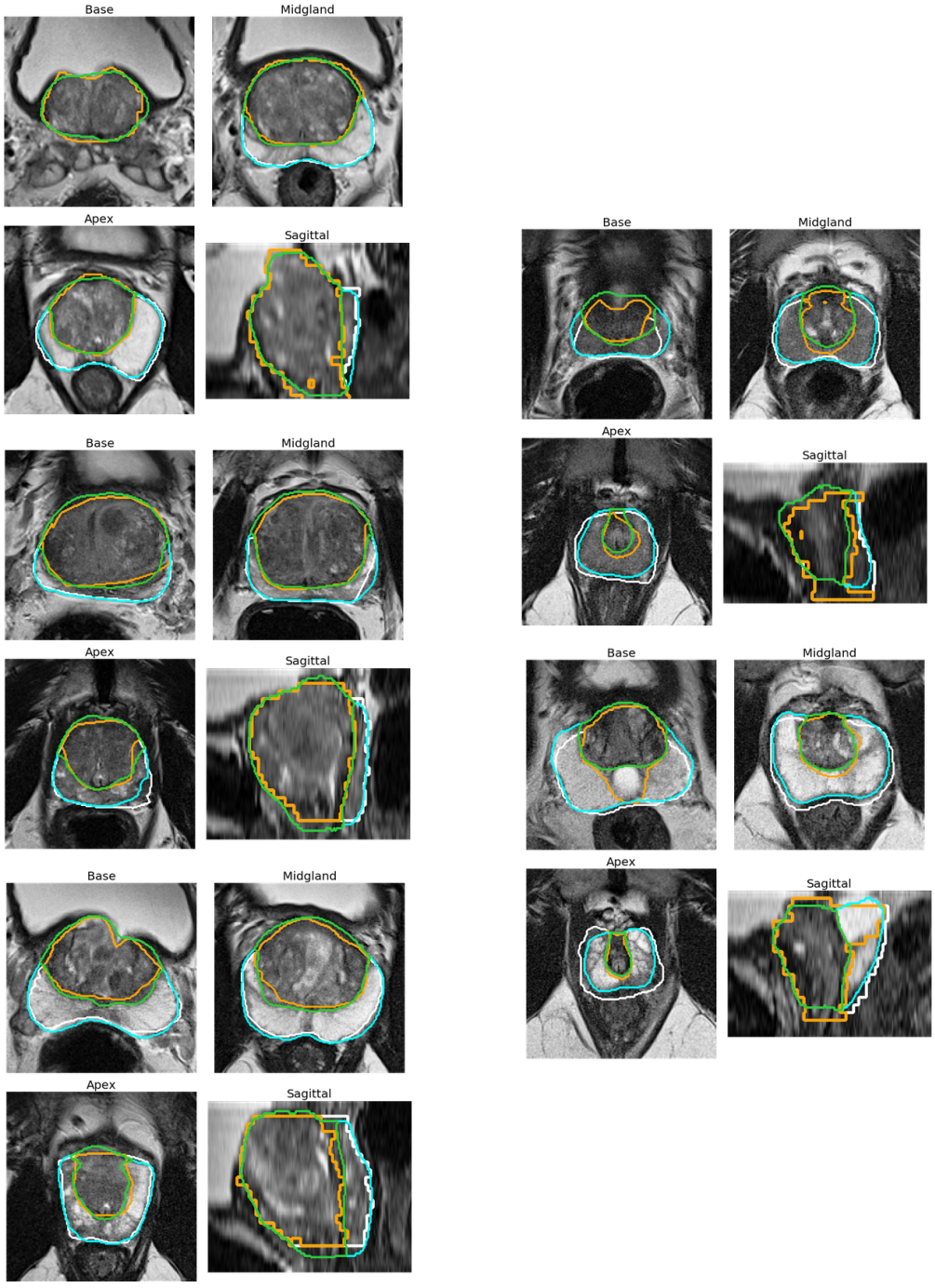


Fig. 4.16.: Segmentations on the 2D dataset ProstateX, with axial views of the base (top left), the midgland (top right) and the apex (bottom left), and sagittal view (bottom right). Left column shows good segmentations, right column shows poor segmentations. White: Ground truth whole gland, Orange: Ground truth transition zone, Cyan: Network-segmented whole gland, Green: Network-segmented transition zone.

results for both modifications, especially on 3D T2W sequences with p-values<.05 between the classic UNet, the version with the anisotropy modules and the version with both additional modules on almost all considered metrics. On the other hand, the use of partition losses seemed to have a positive but moderate impact on the Hausdorff Distance and on TZ segmentation, whereas slightly deteriorated results for PZ.

The apex and the base of the prostate are the most difficult regions to segment, which is in accordance with previous publications, involving human [Bec+19] (cf. Chapter 2) or automatic raters [Ald+20; Sha+17] (See Figs.4.9, 4.10 and 4.16). The reasons behind this complexity include, but are not limited to, the heterogeneity of the tissues on these regions, the possible ambiguities of the borders, or exotic shapes of both WG and TZ at their top and bottom slices.
Results obtained by the network appeared to be on the same level of performance than those of the radiologists, who can be separated in three categories: the experts (raters 3, 6 and 7), the seniors (raters 1 and 2) and the juniors (raters 4 and 5). The network obtained results comparable to those of radiologists in the middle of the pack and significantly better results than the rater 3 (an expert radiologist) for both TZ and PZ (p-val < .05).

We have observed on the private lesion dataset that in more than 90% of the cases, the automatic segmentation preserved the zone in which the lesions were located. The four cases of zonal location errors correspond to a very specific situation that was not present in the training-validation set. Also, we compare favourably with the inter-rater variability on sectorial positions observed by Greer et al. [Gre+18] with 74% of agreement between the radiologists and our method, even though we rely on a slightly simpler sector map. Moreover, the observed errors on sectorial positions would only have a minor clinical impact, as the global position of the lesion (left/right and anterior/posterior) has been preserved in each case. It may also be more appropriate for clinical use to provide the top two sectors on which each lesion lies, since in almost all cases the true sector is among those two. Those results validate the use of neural network segmentation as a second reader for automatic PCa diagnosis to detect in which zone the suspected lesion is located. This is important as it determines which sequence to use to determine the PI-RADS score, as the importance of the different sequences depends on which zone the suspected lesion is located in.

Our study faces several limitations. First the quality of the ProstateX segmentations are uneven, as can be seen in Fig.4.16. Nevertheless we chose to work with this annotation set for the sake of comparing ourselves with other methods. On the private dataset, we had to correct PZ masks computed as WG-TZ with an ad hoc method to avoid the occurrence of thin isolated lines or voxels (see Fig. 4.4). Indeed, without this processing, PZ-associated metrics and especially PZ HD95% were affected. However, the impact was

the same on the three compared networks and it did not modify their relative performances. This processing was not required on the ProstateX segmentation.
In addition, the algorithm to create the prostate sector map is based on debatable hypotheses. In particular boundaries were determined part-wise and not slice-wise, and we arbitrarily defined inner antero-posterior boundaries positions since there exists no formal definition to define them. Those choices have a direct impact on sectorial positions, since peripheral sectors on the apex (especially sectors 6p and 12p) can have a very small area with our method. Moreover, the sector map used in this study, inspired by the 27 region of interest sector map defined in Dickinson et al. [Dic+11], does not correspond to the PI-RADS 2.1 standard sector map [Tur+19] which is based on 39 regions of interest. Nevertheless, in practice, the differences between the two sector maps have little to no practical effects.

In Fig.4.10, we can see that some of the segmentations may have very unusual shapes, with for instance "outgrowths" on the generated segmentation or WG split into two parts. To enforce coherent shapes of WG and TZ, it may be possible to project the prostate on a restricted learned shape space [Kar+18; Tan+19] in order to correct aberrant segmentations.

4.6 Conclusion

We proposed a deep learning-based method for the zonal segmentation of the prostate from T2W sequences, taking into account the anisotropy of these sequences, with attention modules and enforcing the partition of the prostate. This method can be applied on both 2D and 3D T2W sequences. The obtained results not only are similar to the results from the state of the art but are also coherent with the results obtained by radiologists and globally preserving zonal locations and sectorial positions of the lesions, making our method suitable as a first step tool for an automated system dedicated to diagnosis and grading of prostate cancer, as done in Hosseinzadeh et al. [Hos+21] or in Mehta et al. [Meh+21].

5 Weak and Mixed supervision for prostate cancer detection through radiological annotations

Contents

Abstract Automatic method for prostate cancer (PCa) detection are often trained on relatively small datasets, as the accurate segmentation and characterization of cancerous lesions is a time-consuming task. As a way to circumvent this problem, we introduce a weak supervision approach that only requires the sectorial location and the size of each lesion instead of a precise delineation of each lesion. More precisely, our method automatically generates a pseudo-segmentation mask from those information, and use them during the training with segmentation loss functions. We used this method to train a deep learning-based method for PCa on a large weakly-annotated prostate dataset called PAIMRI. To estimate the relevance of this method

to generate pseudo-masks, we compared it to an alternative strategy only based on sectorial location and to the use of precise annotations from the PI-CAI dataset. In addition, we evaluate the impact of mixed supervision combining weak and precise annotations on network performances.
We compared those networks on three datasets - one subset of PI-CAI, one fully-annotated subset of PAIMRI and the external dataset Prostate-158, at both patient and lesion-levels. Our intensity-based pseudo-mask generation was shown to have improved results compared to sector-based pseudo-masks. In addition, using our approach, networks trained on larger weakly-annotated datasets perform generally better than smaller, fully-annotated datasets. Those results validate the interest of weak labels for PCa detection and support the development of large, weakly-annotated datasets.

This chapter will be submitted to a journal [Ham+23c].

5.1 Introduction

5.1.1 Clinical context

Prostate cancer (PCa) is one of the most frequent cancers in the world. Between 2015 and 2019, the incidence rate in the USA was estimated at 109.9 per 100.000 population, the second highest among all types of cancers behind breast cancer[SMJ20]. In addition, it has been estimated that 1 American man over 8 will develop it during his lifetime. This widespread presence makes it an major public health concern with economic impacts[RB11], despite its high 5-year survival rate (97%). Currently, the screening process of PCa can be decomposed into the following stages: first, the clinician proceeds to a digital rectal exam (DRE) and a measurement of the level of PSA (Prostate-Specific Antigen) in the blood . Then, in case of suspicious results, a multiparametric MRI (mpMRI) composed of T2-weighted sequences (T2W), diffusion-weighted images (DWI) and dynamic contrast-enhanced (DCE) imaging is performed. Finally, if the mpMRI confirms clinician's suspicions, a biopsy is performed to confirm the diagnosis. As this last step is invasive and time-consuming, it is important to have a sensitive and specific mpMRI diagnosis in order to avoid useless biopsy exams while preserving the chances of the patient for detecting cancer lesions. However, the interpretation of mpMRI can be complex. Despite the existence of the Prostate Imaging-Reporting and Data System grading standard (PI-RADS, [Tur+19]) to report suspected lesions in a standardized scale from 1 to 5, the variability of its estimation between radiologists is significant [Smi+19; Gre+19; Mus+19], especially for grade 3 lesions (the threshold to undergo a biopsy). In addition, determination of the PI-RADS score can be a time-consuming task.
Therefore, the issue of computer-aided detection of PCa lesions from mpMRI raised the interest of teams all across the world and is the subject of several studies. Yet, the

majority of them are only based on small datasets: Sunoqrot et al.[Sun+22] referenced 3369 publicly available prostate MRIs distributed between 17 public datasets. The largest of those datasets, PI-CAI with 1500 mpMRI, also represents 67% of all publicly available mpMRI cases and the most of lesion-segmented mpMRIs. This is an important drawback to the development of robust and accurate diagnosis methods. This is mentioned by Hosseinzadeh et al.[Hos+21] who state that a large number of cases (>> 2000 MRIs) was necessary to obtain radiologist-level results. However, fully annotating a large dataset requires to gather highly skilled radiologists for several weeks or months. For this reason, in this paper we consider the possibility of using weak segmentations only on a large dataset as a trade-off between the ability to have large datasets and the time needed to annotate them.

5.1.2 Related works

Prostate cancer detection For years, Computer Aided-diagnosis (CAD) tools were conceived to help radiologists detect prostate cancer from mpMRI. For example, in 2012, Niaf et al.[Nia+12] extracted from multiparametric MRI grey-level, texture, gradient and functional features that were then selected (for example using t-test) and used to train classic machine learning methods (such as support vector machines) for the detection of prostate cancer in peripheral zone. However, for the past few years, the majority of those CAD tools are based on deep learning, as in the list of submitted models to the ProstateX[Arm+18] and PI-CAI[Sah+22] challenges for prostate cancer detection. For example Saha et al.[SHH21] conceived an end-to-end 3D framework based on Unet++[Zho+18], taking as inputs the biparametric MRI (bpMRI, corresponding to the mpMRI minus DCE) and a precomputed anatomical prior to predict clinically significant lesions. In addition, the obtained probability heatmap is refined by a classification-based method at the patch level to reduce the number of false positives. Vente et al. [Ven+21] used 2D U-Nets on bpMRI with zonal information to segment tumoral lesions and predict their ISUP score, using Soft-Label Ordinal Regression to leverage the ordinal aspects of those grades. On the PI-CAI challenge, Debs et al [Deb+] used a combination of nnU-Net[Ise+21], Retina U-Net[Jae+18] and normalized PSA density to detect lesions and generate a segmentation as well as the associated level of confidence.

Weakly supervised learning Weak supervision consists in learning a task while having access to only limited or partial. The definition of "limited information" is rather large and gathers several scenarios: learning with noisy labels[Son+22], with bags of labels rather than individual labels - often referred as Multiple Instance Learning[DLLP97], with imprecise labels... A few papers dealt with the detection of prostate cancers using imprecise labels for the training, even though in various settings. Thus, Duran

et al.[DJL20] extended a previously-existing size constraint loss function[Ker+21] to segment and characterize prostate lesions using only randomly sampled scribbles as labels. Patel et al.[PD22] generated lesion segmentations from only classification labels via class activation map (CAM)-based methods[Zho+16; Sel+20], using specific loss functions to enforce inter-modalities information exchange and equivariance constraints. In some cases, this limited information can be combined with a subset for which all the information are available, that case is called mixed supervision [Mly+19]. For example, Mlynarski et al. [Mly+19] designed a U-Net-based network with two decoder branchs: one for tumor segmentation, and one for classification. This network was then trained on both fully-annotated and weakly-annotated image. On the prostate, Bosman et al [Bos+22] used a semi-supervised strategy, based on clinical reports to determine cases with lesions and a network trained on fully-annotated data to segment corresponding candidate lesions, that are then used to augment the training set.

Pseudo-mask generation As a surrogate for real annotations in case of non-availability, several teams decided to generate pseudo-masks from available information. In case of semi-supervised learning, a common strategy is to train on a fully annotated subset and then to use the trained network to generate the pseudo-masks on the non-annotated data [Lee13; Sei+22]. However, when no masks are available at all, it is still possible to construct pseudo-masks from image-level or region-level annotations. For example, Hou et al. [Hou+16] used Gaussian mixture of patch classification and Expectation-Maximization based patch extraction to generate tumor masks on Whole Slide Tissue images. Silva-Rodriguez et al.[SCN21] extracted feature maps after softmax and before Global Aggregation layers for Gleason Grade Prediction of the patches. Several methods used CAM or its derivatives[Zho+16; Sel+20] to produce pseudo-masks[ACK19; Wu+19] then used as a supervision tool during the training. If those methods can be refined to incorporate prior knowledge [Yan+22], by design they rely on the network ability to discriminate the right areas of interest. More simply, Dang et al.[Dan+22] used K-means on 2D patches identified by the user as containing vessels to generate the corresponding pseudo-segmentations for supervision.

5.1.3 Contributions

In this work, we study the impact of large weakly annotated databases on the performance of deep learning-based PCa detection. Our objective is to show the efficiency of our pseudo-mask generation strategy to create segmentation masks by comparing it to a more basic pseudo-mask generation method and to a fully supervised method trained on a smaller dataset. Finally, we want to show how mixed supervision and dataset merging can improve network performances.

Our contributions are:

- a method generating pseudo segmentation mask from weak labels for prostate cancer detection, taking into account clinical information such as lesion' sectorial location and size
- the evaluation of the impact of weak and mixed supervised methods on the detection of prostate lesions on the training of deep-learning based method

5.2 Datasets

In this study, in order to estimate size effects, to compare our method with supervised learning and to test generalization behaviour of our trained networks, we used three datasets: a large dataset with incomplete information, a smaller dataset but with lesion segmentations, and one unrelated, fully-annotated dataset. In the next section we present them and detail their characteristics, but a summary can be found in Tab. 5.1.

	Training	Test	Clinical criteria
PAIMRI	Cases: 5290 Positive cases: 2850 (54%)	Cases: 146 Positive cases: 67 (46%)	PI-RADS≥3
PI-CAI	Cases: 1423 Positive cases:374 (26%)	Cases: 49 Positive cases: 18 (37%)	ISUP≥2
Prostate 158	-	Cases: 138 Positive cases: 82 (59%)	PI-RADS≥4

Tab. 5.1.: Description of datasets used in this work. Red color indicates that no segmentations were available for this dataset.

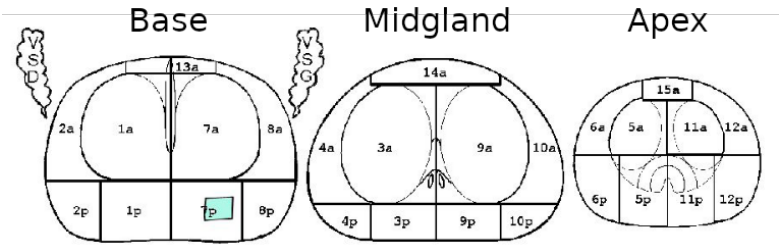


Fig. 5.1.: 27-sector map of a prostate as defined in PI-RADS v1, extracted from a clinical report. The blue square on the schema corresponds to a rough tumor location (not exploited in this work).

5.2.1 PAIMRI-WA dataset

In this work, we analyze a large private set of 10791 patients coming from two different centers (characteristics of each center being available in Tab.5.2a), between March 2009 and March 2022. Among those 10791 cases, 5290 of them were exploitable for automatic detection. The flowchart of their selection is detailed in Appendices (Fig. 5.11). 2850 of them have at least one PI-RADS≥ 3 lesion. For each MRI, clinicians indicated if lesions

were present and for each of them their corresponding PI-RADS score (as defined in PI-RADS v2[Tur+19]) and Likert score (a subjective scale for lesion characterization, not exploited in this study), its maximal diameter and its main sectorial location. This last information was determined using the 27-sector map defined in PI-RADS 1.0[Bar+12] and represented in Fig.5.1. Only one sector was reported, even if the lesion can expand on other sectors. All those information are common information available in clinical reports. Distribution of tumors according to their PI-RADS score and to their zonal location is available in Tab.5.2b. We will refer to this dataset as PAIMRI-WA (WA for weakly annotated).

	PAIMRI-WA		PAIMRI-FA	
Center	Pitié-Salpêtrière	Tenon	Pitié-Salpêtrière	Tenon
Constructor	Siemens	GE	Siemens	GE
Machines	• Skyra (3T) • Aera (1.5T)	• SIGNA Architect (3T) • Optima (1.5T)	• Skyra (3T) • Aera (1.5T)	SIGNA Architect (3T)
#MRI (3T/1.5T)	3752 (2992/760)	1538 (1496/42)	82 (70/12)	64
Avg. T2 resolution (mm)	$0.36 \times 0.36 \times 0.85$	$0.547 \times 0.547 \times 0.5$	Variable	$0.547 \times 0.547 \times 0.5$
#Positive cases (%)	2130 (57%)	720 (47%)	32 (39%)	35 (55%)
#Lesions	3055	1007	44	50

(a) Technical information and population for both PAIMRI datasets.

		<3	3	4	5
PAIMRI-WA	PZ	248	384	2232	737
	TZ	49	225	224	261
PAIMRI-FA	PZ	-	5	43	10
	TZ	-	7	9	13

(b) Distribution of tumors according to their zones and their PI-RADS score for both PAIMRI datasets. Lesions on the anterior fibromuscular stroma were considered as TZ lesions.

Tab. 5.2.: Characteristics of PAIMRI datasets. A more detailed distribution by sectors is available in Appendix.

5.2.2 Other datasets

In addition to this massive but weakly annotated private dataset, we make use of smaller, fully annotated datasets:

- A private dataset of 146 cases, distinct from the one previously described but originating from the same centers, and that we used in a previous study[Ham+22b]. For 60 cases of this dataset, a lesion with a PI-RADS$\geq$ 3 was detected and segmented. We will refer to this dataset as PAIMRI-FA (for fully annotated). Its description is also available in Tab. 5.2b.

- The PI-CAI challenge dataset[Sah+22] is composed of 1500 publicly available biparametric MRI sequences, with their associated prostate segmentation (whole-gland only, produced by a neural network). In addition, for 425 cases out of 1500, lesions with ISUP score $\geq$ 2 were detected. Only 220 of them were segmented by a human expert, while other lesions were segmented by a neural network. 23 of

those cases were discarded for inaccurate lesion segmentations. All information on this dataset can be found online[1].

- The publicly available Prostate-158 dataset, composed of 139 monocentric biparametric MRI sequences, their zonal segmentations and their lesion segmentation if appropriate. We discarded one case for incomplete DWI sequence. Clinically significant lesions were defined as PI-RADS ≥ 4 lesion confirmed by biopsy, and were present in 82 out of 138 cases. More details on this dataset can be found in Adams et al.[Ada+22]

Description of both public datasets can be found in Tab. 5.3.

	PI-CAI	Prostate-158
Centers	• Radboud University Medical Center • Ziekenhuisgroep Twente • University Medical Center Groningen	Charité University Hospital Berlin
Constructors	Siemens (3T/1.5T) Philips (3T/1.5T)	Siemens (3T)
#MRI	1500	138
Avg. T2 resolution (in mm)	Variable	$0.47 \times 0.47 \times 3$
#Positive cases (%)	425 (28%)	83 (60%)
#Lesions	455	89

Tab. 5.3.: Information on PI-CAI and Prostate-158 datasets.

5.3 Methods

5.3.1 Pseudo-mask generation

In this work, we propose to perform a weak supervision of a segmentation task by generating a pseudo-mask, in order to train machine learning models as if exact lesion segmentations were available. We made this choice as it seems to be the most intuitive way to exploit the specific radiological information provided by the clinicians in their reports, i.e. the main sector location of the lesion and its diameter.

Sectorial pseudo-mask

A first method consists in generating pseudo-masks of the main prostate sector in which the lesion has been indicated in. To create them, we generated automatic zonal segmentations P_i of prostates into their Peripheral Zone (PZ) and Transition Zone (TZ) using a previously developed deep-learning based method[Ham+22b]. These zonal segmentations were then used as a basis to create sector maps with a deterministic algorithm we developed in a previous work[Ham+22b]. More precisely, the sectors are defined based on the TZ and PZ as well as anatomically specific planes defined as follows:

[1] https://pi-cai.grand-challenge.org

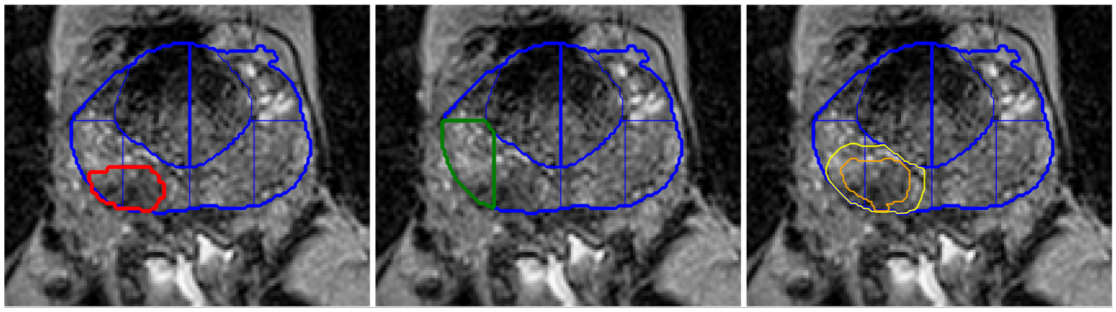


Fig. 5.2.: a) Sector map of the prostate (in blue) and the ground truth lesion segmentation (in red). b) In green, mask of the sector 4p indicated by the radiologist as the main sector where the lesion is localized. c) Probabilistic mask generated by our intensity-based method (yellow: contours of probabilities between 0.1 and 0.5, orange: contours of the binary mask).

- At each third of the prostate for the axial planes. Those planes help us define apex, midgland and base of the prostate.
  Following planes are then computed independently for each of those thirds:
- Median coronal plane - distinction anterior/posterior
- Median sagittal plane - distinction left/right
- Sagittal planes at respectively 40% and 60% of the third for the limits between "lateral" peripheral zone sectors and "medial" peripheral zone sectors. This is used as a surrogate to the central zone segmentation (not available here).

Furthermore, we decided to include lesions noted from anterior fibromuscular stroma (AFMS), corresponding in Fig.5.1 to sectors 13a, 14a and 15a, into their corresponding in-plane TZ sectors for two reasons. First, as AFMS is a non-glandular zone and the majority of AFMS lesions arises in the TZ, tumors from AFMS and TZ are often gathered together[Tav+18]. Second, our zonal segmentation method does not segment AFMS. Example of sector maps generated by our method are given in Figs. 5.2a and 5.3 and . A sector-based pseudo-mask is shown in Fig. 5.2b.

Intensity-based pseudo-mask

Sectorial pseudo-masks delineate roughly the lesion position and they do not take into account information on lesion size, nor the possibility of cross-sectors lesions. For this reason, we propose another pseudo-mask strategy taking those criteria into account. Besides, we constrain the pseudo-mask to never lie at the same time on both TZ and PZ. Indeed, lesions in each zone have different characteristics according to their zone of origin [Vis+12; Gre+91], and consequently PI-RADS [Tur+19] rules for the two zones differ: PI-RADS assessment of lesions on TZ (resp. PZ) are done using primarily T2W (resp. diffusions) sequences. So, zonal locations of the lesions influence their

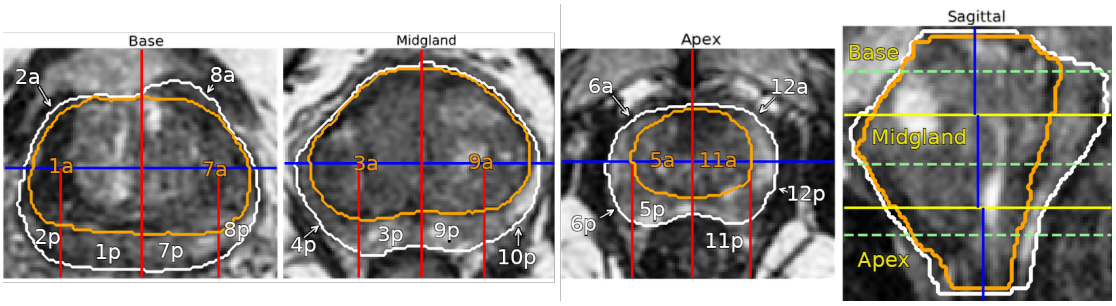


Fig. 5.3.: Sector map of the prostate according to axial views in the base, the midgland and the apex, and to the sagittal view. Segmentation of the whole gland is in white, segmentation of the transition zone is in orange. The blue axis separates the anterior from the posterior of the prostate, the red axes are sagittal planes and the yellow axial planes separate the base, the midgland and the apex. The green dashed lines on the sagittal view indicate the location of the different axial views. Reproduced from [Ham+22b] with authorization of SPIE.

characteristics and their methods of diagnosis, in consequence cross-zonal lesions must not be considered as they hardly correspond to a real case and make the diagnosis process more difficult.

For a patient i with a lesion, the construction of the associated pseudo-mask y_i is done via the following scheme:

- Determination of the target sector indicated by the radiologist.
- Filtering the ADC sequence using an averaging filter (kernel size: $3 \times 7 \times 7$) then determination of C the voxel of the filtered image within the target sector with the lowest intensity.
- Computation of a ball centered on C and with a diameter equal to the lesion diameter reported by the radiologist. This ball is then restrained to the zone to which the indicated sector belongs to. Indeed, a prostate sector belongs to only one zone, and we forbid cross-zonal lesions for previously mentioned reasons.
- In this ball, use of Otsu's thresholding method [Ots79] to determine the potential lesion as voxels of values below the computed threshold. The final pseudo-mask S_i is the restriction of this set to its largest connected component using a 26-neighborhood.

In addition, we must take into account that there can be discrepancies between the sector indicated by the radiologist and the sector determined by our method. To do so, we attribute to voxels outside of the pseudo-mask S_i a value based on their distance to the pseudo-mask: $y_i = \sigma(500S_i - DM(S_i))$, with $\sigma(.)$ the sigmoid function.
In this equation, $DM(S_i) = (1-\lambda)DM_E(S_i) + \lambda DM_I(S_i; P_i)$ is the combination between the classical Euclidean distance map to the mask, allowing us to take into account the proximity to the pseudo-mask S_i, and DM_I a penalization map equal to 0 on the prostate zone (PZ or TZ) in which the lesion is located in, and > 0 on both the other zone and the

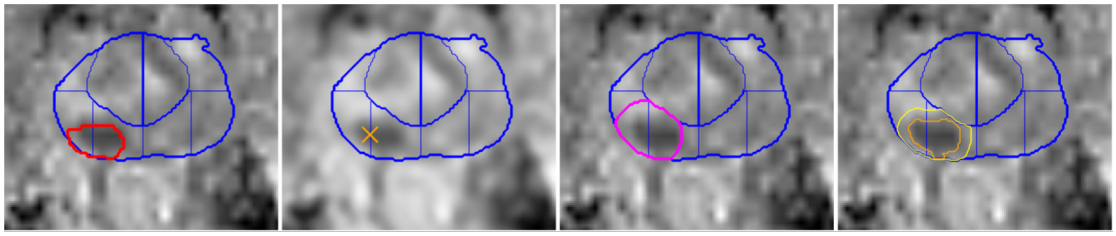


Fig. 5.4.: Construction of the intensity-based pseudo-mask. a): Sector map (in blue) and true lesion segmentation (in red) for reference). b) Determination of the center of the zone of lowest intensity after average filtering (orange cross). c) Construction of the ball centered on this region (magenta). d) Otsu's thresholding for final pseudo-masks (orange) and area of uncertainty based on distance maps (in yellow, contours of probabilities with a value between 0.1 and 0.5.

background. By design, values outside of S_i are lower than 0.5. The combination of the two distances leads to: i) higher probabilities for voxels close to S_i that are in the same zone than the lesion, and ii) lower probabilities for voxels in the other zone or far from the lesion. Here, DM_I was computed as a geodesic distance map based on the zonal segmentation P_i, with specific values for background and prostatic zones.

An example of this intensity-based pseudo-mask y_i is presented in Fig. 5.2c, and the different stages are presented in Fig.5.4. The process of creation of the uncertainty area is detailed in Fig. 5.5

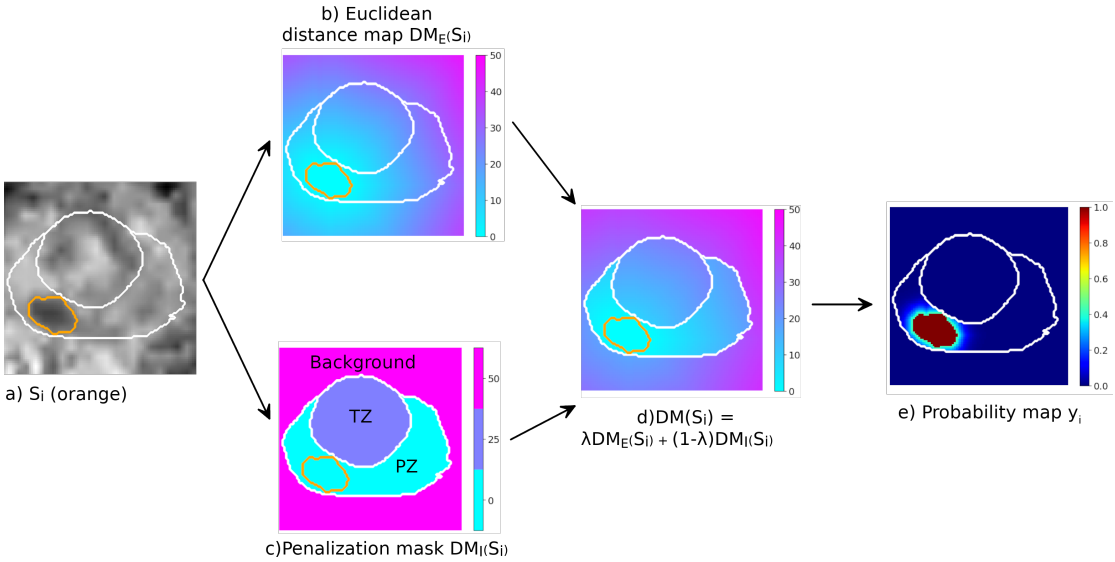


Fig. 5.5.: Construction of the intensity-based pseudo-mask y_i. a) ADC sequence with zonal segmentation (in white) and S_i (in orange). b) Euclidean distance map to the pseudo-mask c) Penalization map in the case of a PZ lesion d) Final distance map e) Corresponding pseudo-mask y_i

We chose to use lowest intensity regions on ADC to determine possible position of tumors for two reasons:

- tumors appear as hyposignal on ADC sequence images and ADC is one of the two sequences used in PI-RADS to detect PZ lesions
- experiments showed that looking for hypersignal in DWI would lead to similar results than ADC for the generation of pseudo-masks for PZ lesions but not for TZ lesions. As for T2W, the sequence recommended on TZ by PI-RADS, the low intensity of bladder walls and fibromuscular stroma is not compatible with simple intensity-based strategies.

This strategy has limitations, as hyposignal in the TZ on the ADC sequence can also be due to benign fibrodysplastic nodules. However, the restriction of the search for low-intensity regions to a single sector and the use of a spatially large averaging filter helps the generation method to focus on the right region to determine the pseudo-mask.

5.3.2 Neural network

To take into account the three modalities, we used a U-Net based network inspired by the work of Saha et al. [SHH21], but with three encoders parts: one for each MR sequence (T2W, ADC, DWI). Each of those encoders takes as input the corresponding sequences as well as the zonal segmentation of the prostate automatically generated by a prostate segmentation framework previously developed [Ham+22b]. Outputs at each layer of the encoders are concatenated, and then given to the decoder part. Attention gates [Okt+18] and Squeeze-and-Excitation modules [Run+19; HSS18] are used to improve respectively spatial and feature discriminative power of the network. A representation of this network is available in Fig. 5.6.
For the loss function we used the Focal Loss [Lin+17] since the structures we have to find are very small compared to the image size and are not present in some cases. We picked as hyperparameters $\alpha = 0.75$ and $\gamma = 2$.

5.4 Results

5.4.1 Accuracy of the generated pseudo-masks

We compared the generated pseudo-masks (sector-based and intensity-based) with ground truth segmentation masks on the PAIMRI-FA dataset. For those comparisons, we used soft Dice and cross-entropy losses. For 12 cases from PAIMRI-FA, in addition to the segmented lesions, the radiologist's main sectorial location and their estimated maximal diameter was available as in PAIMRI-WA. This subset, coined PAIMRI-RA, allowed us to compare both pseudo-masks generation strategies in real settings. However, due to its limited size, we also compared the generated pseudo-masks on the whole PAIMRI-FA dataset but using the main prostate sector and lesion diameter information as provided by the automatic zonal segmentation algorithm and the ground truth lesion masks.
To compute the distance maps for intensity-based pseudo-masks, we used the method

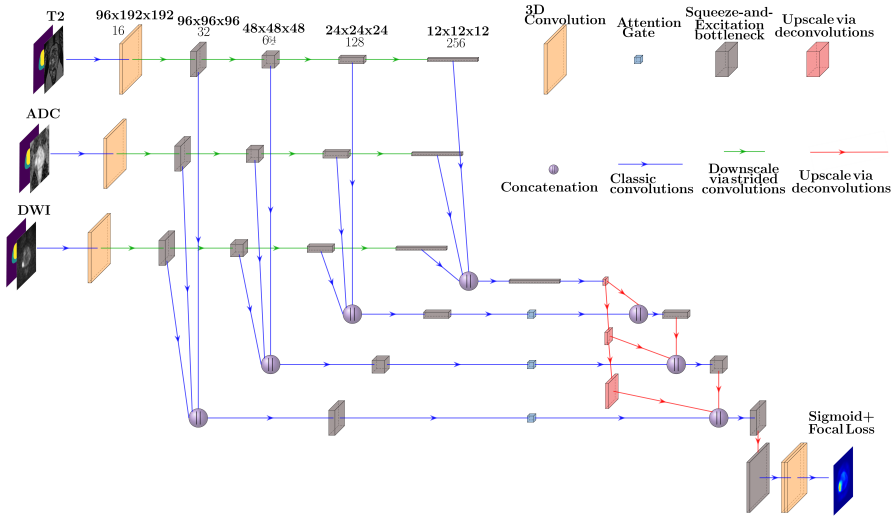


Fig. 5.6.: Lesion detection network architecture, inspired by Saha et al.[SHH21], using Squeeze-and-Excitation bottlenecks [HSS18] and attention gates [Okt+18]. All convolutional layers use LeakyReLU (with $\alpha = 0.1$) as loss functions. Dropout nodes (p = 0.50) are connected at each scale of the decoder to limit overfitting.

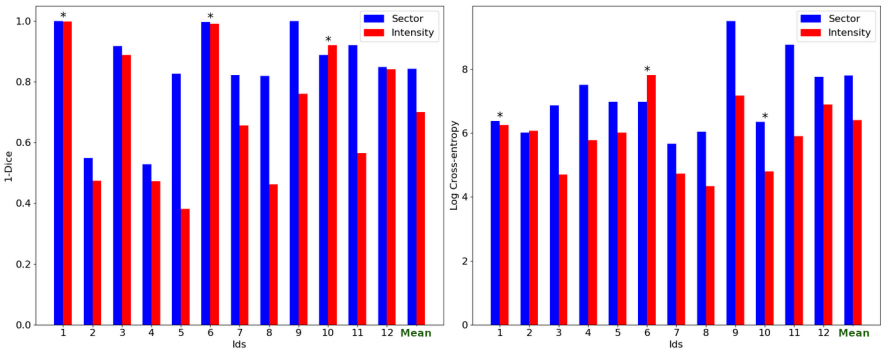
presented by Criminisi et al. [CSB08] and implemented in FastGeodis [ADV22], using as parameters $A = 500$, $\lambda = 0.33$. For the computation of the geodesic distance based on the zonal segmentation, we attributed to each voxel in the background a value of 0, 50 for those in the PZ and 75 for those in the TZ.
Results are shown in Fig. 5.7. For both cases, mean metrics are minimized when using the intensity-based pseudo-mask (even though not for all cases). For comparisons on the whole PAIMRI-FA set we computed the p-value using the Wilcoxon signed-rank test, and for both metrics the p-value was < 0.001.

5.4.2 Experimental design for the impact of weakly annotated databases

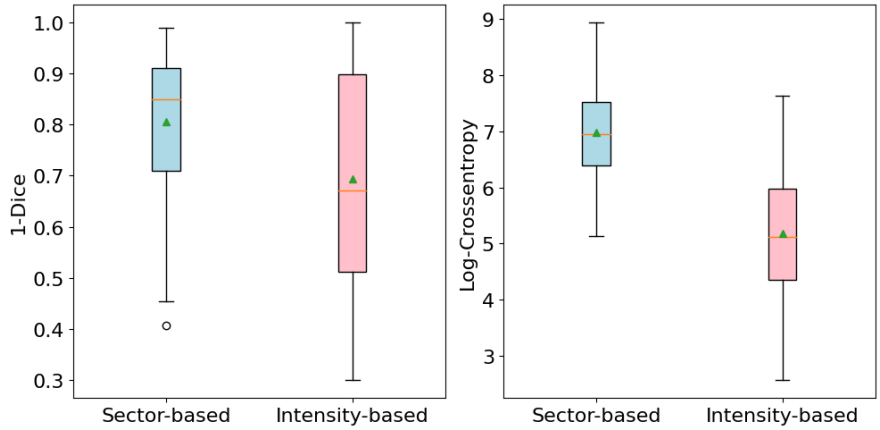
We compare the same network architectures trained on different databases and pseudo-mask generation methods:

- A network trained only on the fully annotated dataset PI-CAI (trained on 1125 cases).
- A network trained only on PAIMRI-WA, (trained on 4025 cases) with sector or intensity based pseudo masks
- A network trained on both PAIMRI-WA (with sectorial pseudo-masks or intensity-based pseudo-masks for PAIMRI-WA) and PI-CAI (with corresponding masks). Trained on 5150 cases.

The trained networks were then tested on PAIMRI-FA (n=146), on a extracted test set of PI-CAI that we'll call PI-CAI-Te (n=49) and on the Prostate-158 (n=138) datasets. In all



(a) Soft Dice loss (left) and log-cross entropy loss (right) for each case of PAIMRI-RA. Stars indicate cases where the radiologist-indicated sector does not correspond to the sector estimated by our segmentation network.



(b) Soft Dice loss (left) and log-cross entropy loss (right) on PAIMRI-FA (N=88)

Fig. 5.7.: Comparison between ground truth lesion and suggested pseudo-masks for Soft Dice loss (=1 - Soft Dice) and log cross-entropy on both PAIMRI-RA (clinical information by a radiologist) and PAIMRI-FA (automatically determined clinical information). On both datasets, the average soft dice loss and log-cross entropy is lower using the intensity-based pseudo-mask rather than the sector mask. On both images, sector-based pseudo-masks' measurements are in blue, intensity-based pseudo-masks' are in red.

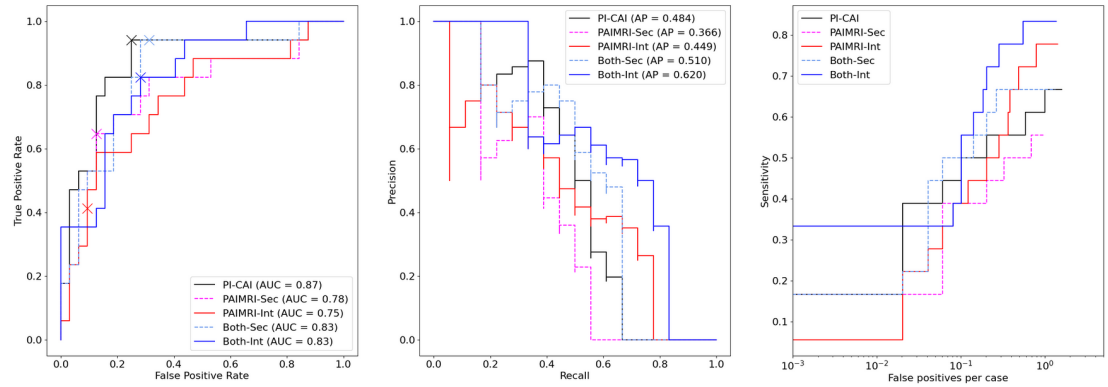
settings we used the Adam optimizer with a initial learning rate of 1e-6, a batch size of 4 and we trained the network for approximately 90 epochs, with the exception of the fully supervised network which was trained for 180 epochs to compensate for its smaller size. All images were resampled to a resolution of $0.5 \times 0.5 \times 1$mm. T2, ADC and DWI were rigidly registered using SimpleElastix[Mar+16]. When possible, b-values for PAIMRI DWI images were set to 2500. B-values on other datasets or when the b-value was not accessible were unchanged.
To improve models performance, data augmentation such as flipping, in-plane rotation, translation and scaling were applied on datasets with the Kornia package [Rib+19].

5.4.3 Performance metrics

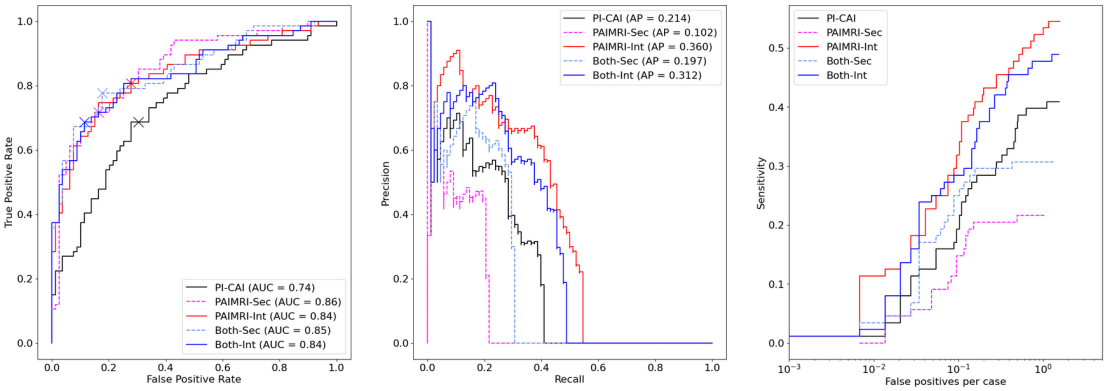
To estimate the efficiency of the different methods, we provide several performance metrics. First, the Area Under Receiver Operating Characteristic (AUROC) curve at the patient-level provides an indication on the performance of the methods to detect patients with clinically significative lesions irrespective of the choice of a specific threshold. Second, the Average Precision (AP) at the lesion-level, giving the indication on how efficient our method is to detect each lesion separately. In its computation, the considered threshold for the detection of a lesion was an Intersection over Union between the mask M and the ground truth G: $IoU(G, M) = \frac{|G \cap M|}{|G \cup M|} \geq 0.1$. Those metrics are similar to those used in the PI-CAI challenge. In addition, we also provide the sensitivity level (or True Positive Rate, TPR) of each method for a threshold corresponding to 0.5 false positive per patient. Finally, since the main radiological indication provided by radiologists is sector location, we also want to assess the accuracy of the trained networks to detect lesions at the right sector location. To do so, we computed sectorial accuracy (SA) by using the previously mentioned automatic sectorization algorithm to determine the sectorial location of each ground truth lesion and of each binarized prediction. For the latter, we defined this sectorial location as the sector with the largest intersection with the predicted lesion mask. To binarize the predictions, we used as a threshold the value maximizing the Youden's index (=sensitivity+specificity-1) on the patient-level ROC curve.

5.4.4 Metrics results

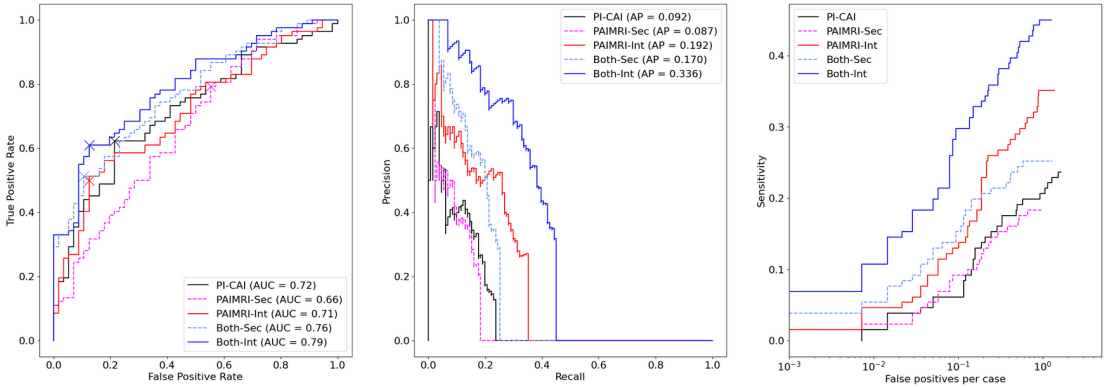
In Table 5.4 we show the performance metrics of the networks on the different test datasets. In addition, corresponding patient-level ROC curves , lesion-level precision-recall curves (related to AP) and FROC curves are available in Fig. 5.8. Visual examples of good predictions are displayed in Fig. 5.9, and examples of false positives and false negatives are shown in Fig. 5.10.



(a) PI-CAI-Te

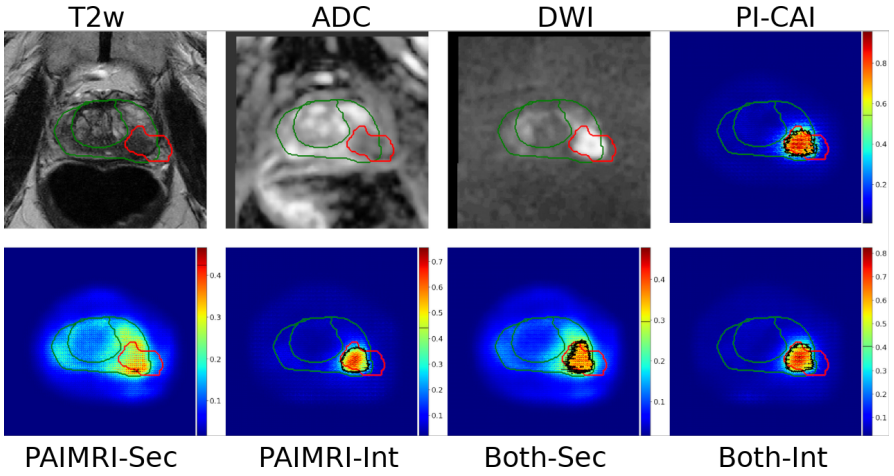


(b) PAIMRI-FA

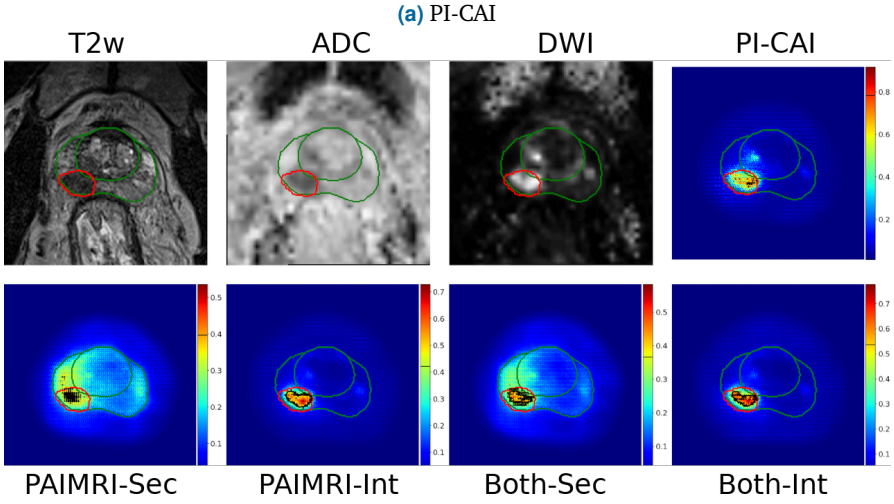


(c) Prostate-158

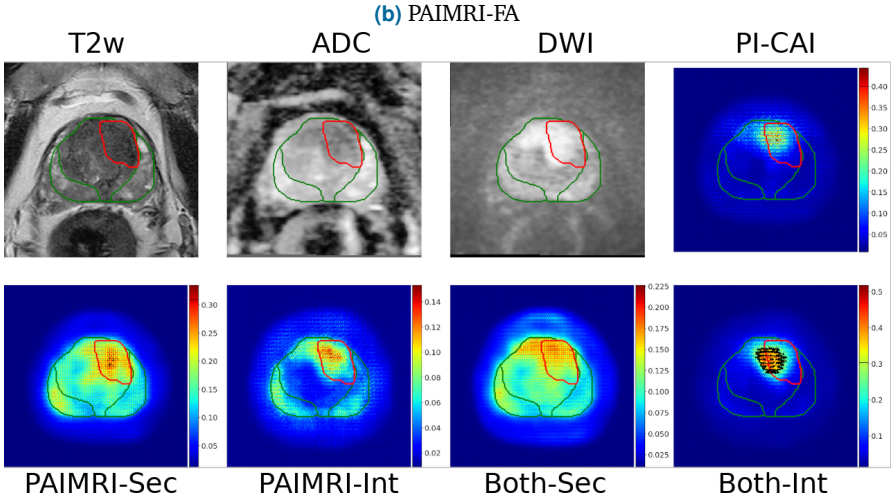
Fig. 5.8.: Result curves for the different datasets: PI-CAI (top), PAIMRI-FA (middle) and Prostate-158 (bottom). On the left: Patient-level ROC curve, with the best performance (i.e. maximizing the Youden's index) indicated by a cross. In the center: Lesion-level Precision-Recall curve. On the right: FROC curve.



(a) PI-CAI



(b) PAIMRI-FA

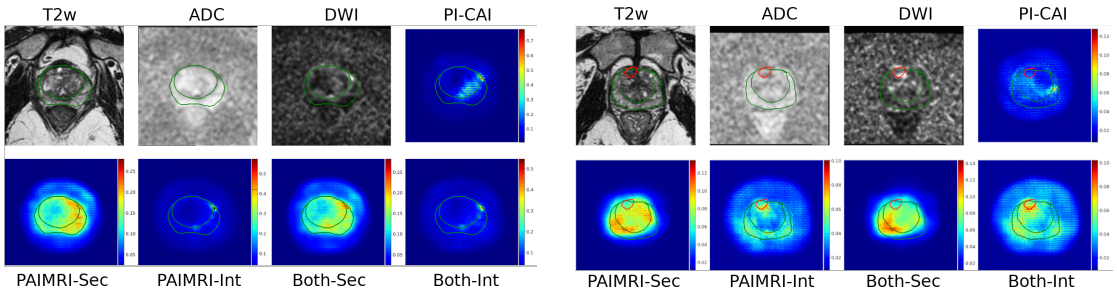


(c) Prostate-158

Fig. 5.9.: Examples of results on PI-CAI (top), PAIMRI-FA (middle) and Prostate158 (bottom) datasets. True segmentations are in red, zonal segmentations are in green, values above the threshold are contoured in black.

Test set / Train set	PI-CAI-Te (n=49)				PAIMRI-FA (n=146)				P158 (n=138)			
	AUROC	AP	Sect. Acc.	TPR@ 0.5FP	AUROC	AP	Sect. Acc.	TPR@ 0.5FP	AUROC	AP	Sect. Acc.	TPR@ 0.5FP
PI-CAI-Tr (n=1135)	**86.76%**	48.37%	**50%**	56%	74.12%	21.35%	32%	36%	71.65%	9.21%	19%	19%
PAIMRI-Sec (n=4025)	78.49%	36.64%	22%	50%	**85.62%**	10.17%	25%	36%	66.33%	8.66%	18%	16%
PAIMRI-Int	75.37%	44.88%	44%	72%	83.79%	**35.97%**	**42%**	**48%**	70.80%	19.19%	15%	29%
Both-Sec (n=5160)	83.27%	51.02%	44%	67%	84.77%	19.74%	40%	31%	76.42%	16.98%	15%	23%
Both-Int	82.54%	**62.03%**	**50%**	**78%**	83.96%	31.16%	41%	45%	**78.88%**	**33.63%**	**27%**	**40%**

Tab. 5.4.: Results of the networks according to the used traininig and test sets. Best metrics on each test set are in **bold**. -Sec indicates the use of sector-based pseudo-masks. -Int indicates the use of intensity-based pseudo-masks.



(a) False positive **(b)** False negative

Fig. 5.10.: Examples of false negative (left) and false positive (right) on the PAIMRI-FA dataset. Ground truth segmentations are in red, zonal segmentations are in green, values above the threshold are contoured in black.

5.5 Discussion

Impact of the labelling method In Table 5.4, we compare the results obtained on the same dataset between the two conceived pseudo-labelling methods: intensity-based and sector-based. We can see that, all other things being equal (test dataset, presence or not of fully annotated data, ...), the intensity-based pseudo-labelling method appears in general as the most relevant, with better Average Precision and sectorial accuracy in all cases, and close results in terms of patient-level AUROC. It means that even if both sectorial-based and intensity-based methods have similar results for patient-level prediction, the intensity-based mask leads to a better assessment of the lesion. However, in terms of sectorial accuracy, sector-based pseudo-masks can hold the same level of precision than intensity-based pseudo-mask, but this result is not persistent across all datasets and settings, and the associated sectorial accuracy can drop as low as half of the one obtained with intensity-based pseudo-masks.

On the examples of Fig. 5.9 we can see that the probability heatmaps are more focused for intensity-based pseudo-mask networks compared to those trained using sector-based pseudo-masks (without considering if those probabilities are above the threshold or not). We can also notice the effect of the addition of the PI-CAI cases here, which helped the sector-based network to produce even more focused heatmaps.

Comparisons between fully-annotated and weakly-annotated methods The difference of performance between the network trained only on PI-CAI and the one trained on PAIMRI using intensity-based pseudo-masks can be noted. Each of these networks outperforms the other ones on the test set related to their training set. Yet, except for the AUROC, the difference of performance between the two networks is more important on the PAIMRI dataset (and at the benefit of the network trained on PAIMRI) than on the PI-CAI dataset. In particular, PI-CAI-trained network has a high false positive rate compared to other networks at equal sensitivity. Similarly, the level of performance of both networks on the Prostate-158 dataset is either relatively close or at the advantage of the PAIMRI-trained dataset. Several reasons could explain those results. One of them is the size difference between the two datasets (training sets of size 1135 vs 4025). Another explanation is the chosen criteria of clinical significance for each dataset. Indeed, the criteria for both Prostate-158 and PAIMRI is the radiological-based PI-RADS score, whereas the criteria for PI-CAI is the histological-based ISUP.

Mixed supervision We can note that the use of mixed supervision by combining the PI-CAI dataset with the PAIMRI dataset using our weak supervision strategy hold among the highest metrics on all datasets, and especially on the Prostate-158 dataset where the difference was significant (+14% in AP, +8% in SA and +11% in TPR@0.5FP compared to other methods). This confirms that the use of multicentric datasets is a good way to obtain network with better generalization abilities. However, this combination seems to be mainly occurring when using the intensity-based pseudo-mask method, rather than using sector-based PAIMRI pseudo-masks with PI-CAI labelled cases. These results are on par with those obtained by Bosma et al. [Bos+22]. They showed that the inclusion of ~4500 cases segmented by an automatic method to a set of ~3000 manually annotated MRIs improved their AUROC on an external set by 2%. Similarly, metrics linked to FROC and Precision-Recall curves were also improved by the inclusion of automatically annotated cases. However, the relevance of their method in cases where the fully-annotated and the automatically-annotated sets have different characteristics can be questioned. For example, here, performance of the network trained on PI-CAI and tested on the PAIMRI-FA dataset was low and so using these segmentation to train a model may lead to bad results. Additional tests on this topic must be conducted.

We want to emphasize that, if our network did not obtain results similar to the most efficient methods developed for the PI-CAI challenge [Sah+22] (best models on the PI-CAI challenge have AUROC around 90% and AP between 60% and 70%), the deep learning method used in this paper is far simpler than those models. Indeed, contrary to many of them, the evaluated method does not use ensembling of multiple networks, nor false-positive reduction postprocessing, nor inclusion of clinical information (such as PSA density)... However we believe that results observed here have no reasons not to hold when using more complex strategies for lesion detection.

Limitations This study suffers from several limitations. First, the conditions of clinical significance between PI-CAI (ISUP ≥ 2) and PAIMRI (PI-RADS ≥ 3) or 4 are not equivalent. Indeed, PI-RADS score is only a radiological score, whereas Gleason score is an histological score which indicates the properties of the cancerous cells. Thus, the PI-RADS is only an imperfect proxy - a high PI-RADS score does not always mean a high Gleason score. It was shown in Drost et al. [Dro+19] that PI-RADS-based detection has a very high sensitivity (0.91) but a low specificity (0.37), meaning not only that PI-RADS based labels can result in many histological false positives, but also that ~10% of lesions are not detected by radiologists on the mpMRI. Since only one score (PI-RADS or Gleason) was provided for each dataset, we choose to binarize them as clinically significant/not clinically significant as a way to mitigate those differences. Nevertheless, if PI-RADS's low specificity questions its use as a labelling tool, the possibility to correctly assess Gleason grade of all lesions solely from mpMRI has yet to be proved.
Another drawback of this study is the lack of data curation of the PAIMRI dataset. Manual curation of large datasets is time-consuming, especially for the segmentation of tumors (quality of pseudo-masks, matching between indicated sector and true sectorial location on PAIMRI; automatic tumor segmentation on PI-CAI) and of prostate zones (done by a neural network in this study). There are several strategies to perform this curation, from the simple exploitation of clinical information (prostate volume, zonal intensities) to more complex strategies such as shape analysis to detect and remove outliers segmentations [SG02].

Future works Several extensions of our work can be considered. First, for the generation of pseudo-masks, refinement for better handling of TZ lesions should be proposed. For instance, taking into account the T2W sequence or differentiating AFMS from TZ in the mask generation.
Another possible extension is to exploit more information from clinical reports. For example, in Fig. 5.1 we see that there is an indication of tumor location and extent (in Fig. 5.1, this information is provided by the blue square). These sketches , often found in clinical reports, can help the generation of more precise pseudo-masks, along the other textual information available in the clinical report.
In addition, in this chapter we did not properly study the impact of the dataset size and/or proportion of weakly-vs-fully data on the network performance. Finally, additional studies should be done on multicentric datasets. As seen in Tab. 5.2a, images from PAIMRI come from two different datasets with different imaging manufacturers. Similarly, images from PI-CAI originate from three centers and two different imaging manufacturers. However, we did not include specific methods to analyze how networks trained on MRI coming from one center can generalize to those from the other center. This question has already been investigated in the context of prostate segmentation[ZR+20; Gib+18], but studies on a larger scope and on tumor detection should be launched.

5.6 Conclusion

In this work we presented a new method to generate pseudo-masks for prostate cancer detection from radiological information. We showed that using this method to train a network on a large weakly-annotated dataset allowed us to a more efficient network than training it on a smaller, fully-annotated dataset. In addition, combining both datasets helped to improve the generalization ability of this network. This confirms the relevance of our pseudo-labelling method and calls for the release of large prostate datasets even with limited information on cancer locations.

5.7 Appendices

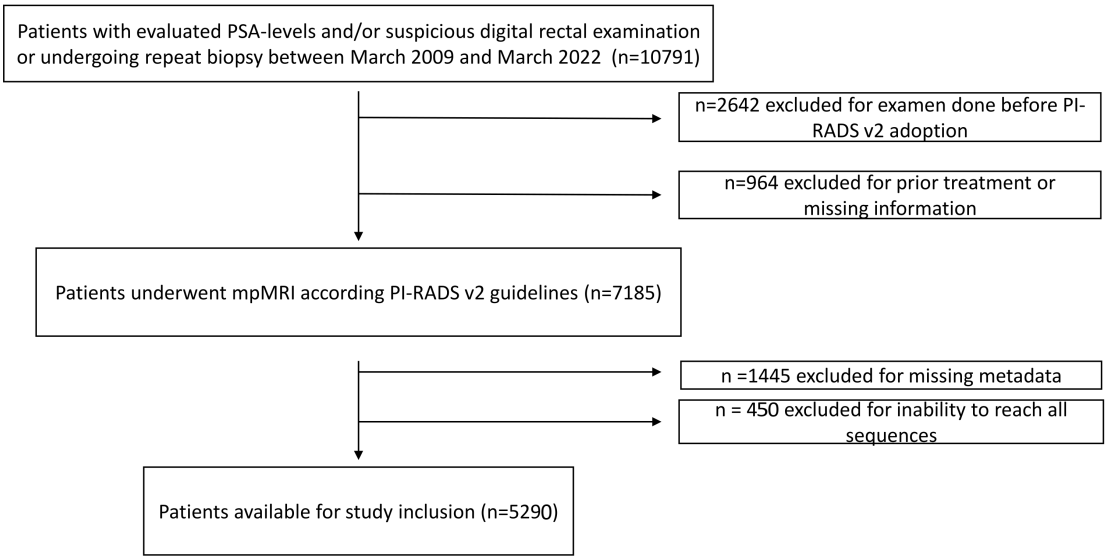


Fig. 5.11.: Diagram shows inclusion of patients into PAIMRI-WA. PSA = prostate-specific antigen, mpMRI = multiparametric MRI

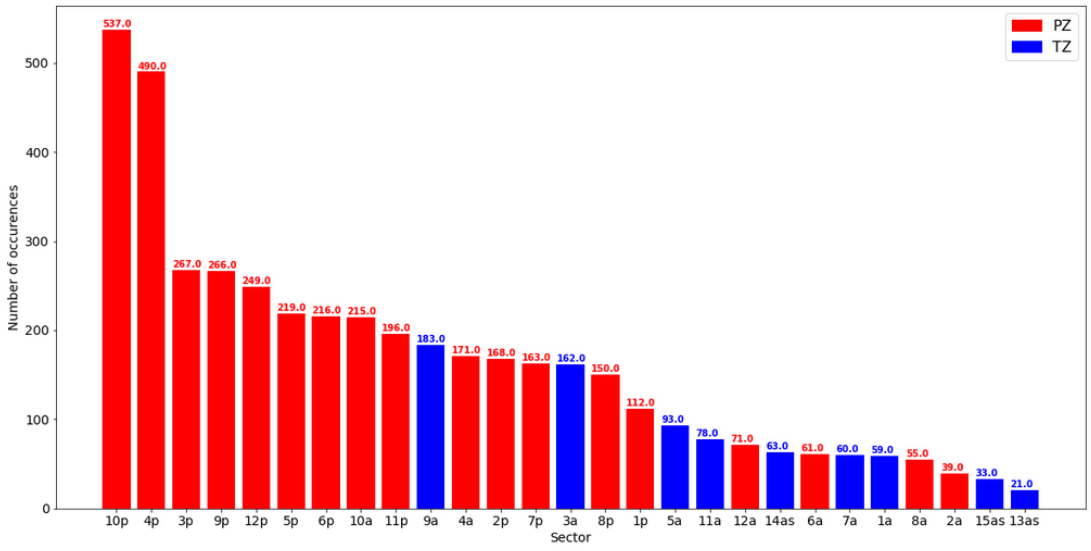


Fig. 5.12.: Number of lesions for each sector. Their corresponding zone is also indicated.

6 Conclusion

In this thesis, we introduced several methods involved at different stages of prostate cancer detection from multiparametric MRI, while focusing on important challenges : variability of the prostate shape, volume and rater performance, correct estimation of a segmentation consensus from binary masks for datasets design, prostate zonal segmentation and possible clinical consequences, construction of large annotated datasets. Below we summarize our contributions and consider additional challenges that should be addressed in future studies.

6.1 Main contributions

Evaluation of inter-rater variability of prostate volumes and segmentations In Chapter 2 we studied the inter-rater variability of prostate zonal segmentation and volume estimation on a dataset of 40 T2W prostate images segmented by 7 raters of various levels (3 expert, 2 intermediate, 2 junior). We showed that segmentation inter-rater variability was globally low for whole gland and slightly higher for transition zone. This variability was mainly concentrated at the extremities of the organ and influenced by prostate properties such as its volume and its PZ-to-TZ intensity ratio, but not by the radiologists' level of expertise. Despite this observed variability on segmentation masks, manual planimetry appears as the volume estimation method with the lowest variability between raters. Yet, ellipsoid-based methods provide a good approximation of prostate volume while being also highly reproducible and simpler to compute.
In addition, we also showed that, for inter-rater segmentation variability studies, using either pairwise metrics or metrics with respect to a consensus was leading to similar trends. Finally, as developed in Appendix B we also studied the impact of the number of raters in prostate segmentation inter-rater variability studies and showed that for both metric methods (either pairwise or with respect to a consensus), the optimal number of raters to consider is 3.

Background-independent region-level consensus segmentation In Chapter 3 we showed that the classic STAPLE method to compute consensus segmentation from binary masks was dependent on the background size and the choice of the prior probability law. In particular, we demonstrated that in case of large background size the obtained consensus tends towards the set of all voxels segmented by a specific number of raters, this number depending on the prior law. As an alternative we designed MACCHIatO, a method

based on the approximation of local Fréchet means of Jaccard or Dice distances through morphological-based heuristics to compute a consensus segmentation. This method is independent from background size, contrary to STAPLE, and computed at a larger scale than the voxel, contrary to mean averaging. Both MACCHIatO and averaging are special cases of a framework we formalized to compute consensus segmentations, but with different choice of distances between binary masks. Consensuses produced by MACCHIatO are generally of intermediary size between those produced by majority Voting and those produced by STAPLE, and differ significantly from those produced by averaging in case of high inter-rater variability.

Automatic zonal segmentation and clinical evaluation In chapter 4 we developed a deep learning based method for the zonal segmentation of prostate from T2W sequences. This method is composed of two networks. The former one, based on U-Net, determines a segmentation of the whole gland from a low resolution T2W MRI. The bounding box corresponding to this segmentation is then extracted and given as input to a second, larger neural network computing the zonal segmentation from this restrained T2W image but at a higher resolution. This second network incorporates spatial and feature-wise attention modules to improve its performance. We evaluated our method on one public (n=131) and one private(n=204) dataset, with consistent results. We also compared the obtained results with those obtained by 7 radiologists, showing that our method had results similar to those obtained by them - being in the middle of the pack. We also checked the impact on the sectorial location of tumors, using the produced zonal segmentation. To this end, we conceived a deterministic algorithm based on PI-RADS rules to construct from a zonal segmentation the corresponding sector map, and compared the estimated location of lesions based on our computed zonal segmentation with the ones provided by a radiologist. We obtained that the correct zones and sectors were conserved in a large majority of cases, validating their possible use in clinical setting as the PI-RADS scale depends on this location.

Pseudo-mask generation method for large weakly-annotated datasets In chapter 5, we presented a new method to generate pseudo-masks for prostate cancer detection, based on simplified radiological information provided by radiologists. To this end we used information present in clinical reports such as the sectorial location of the lesion and its diameter with automatic methods to construct zonal segmentations and sector maps of prostate. In addition, these information were used to lead an intensity-based method to construct a pseudo-mask for the indicated lesion. The generated pseud-masks were shown to be more precise than using only as a pseudo-mask the unique sector provided by radiologists. The objective of this method is to allow to train networks on large datasets while minimizing time to create annotations to supervise them. We experimented that by comparing methods trained on a large weakly-annotated dataset (n=5290) with those trained on a smaller, fully-annotated dataset (n=1500) and showed that methods

trained using our pseudo-mask generation strategy obtained similar or better results than those trained on the smaller, fully-annotated dataset. In addition, a method trained on a mixed-supervised way performed generally better that methods trained on only one dataset, especially on external datasets. Those results demonstrated the relevance of our method for weak annotations and call for the release of large, multicentric datasets even if they only have weak annotations.

Each of those contributions may help the improvement of future prostate cancer CAD methods.

6.2 Perspectives for the future

Several subjects linked to prostate cancer detection were not studied in this thesis, and could lead to future studies extending the themes evoked here.

Exploitation of other clinical information In this thesis, the detection of PCa was done using biparametric MRI, i.e. without considering DCE. Reasons behind this choice are twofold: first, contrary to T2W and diffusion-weighted sequences that lead to a single volumetric image, information coming from DCE are stored in time series of volumetric images, and its processing before use in deep learning is less trivial. Second, its role on PI-RADS determination is very specific since it only serves to choose between a PI-RADS score of 3 and 4 for some PZ lesions[Wei+16], and the debate on its benefit/cost ratio is ongoing in the radiology community. Recent studies [Woo+18; Chr+20] suggest that bpMRI and mpMRI have similar levels of performance for trained radiologists but that DCE sequences may be useful for less experienced radiologists [Col+22] or for some specific lesions [Tag+19]. However, more studies regarding the overall benefits of DCE are required before a conclusion can be drawn. Similarly, some articles explored the relevance of DCE in CAD methods [Meh+21; Bra+21; Che+22], and they were not conclusive with respect to this question.
Moreover, in addition to imaging data, other clinical information such as PSA density (PSAd), prostate volume or patient age could be included into the different models in order to improve their performance. In particular, PSAd seems worthy of interest [Ben+92; Yus+20], and its inclusion into CAD methods is under evaluation [Meh+21; Deb+].

Active surveillance Prostate is a non-vital organ and, contrary to the vast majority of cancers, PCa are mostly non-lethal. Combined with the advanced age of the majority of patients (more than 60% of PCa are detected in people older than 65 [Key]), surgical and radiotherapy-based treatment can be more detrimental to the patient's health than the tumor in its current state. In this situation, corresponding to non-aggressive lesions with a ISUP$\geq$1 or $=$2 with a small volume, the guidelines [Mot+21; Bju+20] recommend active surveillance, i.e. their monitoring via regular screenings: PSA every 3 to 6 months,

DRE every year, mpMRI and biopsies every 2 to 3 years [Mot+21]. The benefits of deep-learning based active surveillance methods are similar to those for PCa detection: assisting radiologists and avoiding as much as possible the performance of prostate biopsy [Alg+18]. However, contrary to simple PCa detection, deep learning-based active surveillance must handle longitudinal data, in order to predict the evolution of the tumor and to correctly assess risks [Lee+22].

Large multicentric datasets One of the identified methods to improve the performances of CAD methods is the construction of very large datasets to train them. Indeed, the training set size has been identified as having an important impact on the method's performance [Sun+22]. If such large datasets are currently not publicly available, it could be the case in the coming years. First, as seen in Chapter 5 those datasets would not have to be fully annotated - reducing the needed time to create them. Second, such datasets with a semi-public access already exist: the PI-CAI challenge [Sah+22] gave access to their best participants in a second phase to a large dataset of 11,000 cases.
However, one aspect of the problem we only barely scratched is the question of multicentric datasets. Indeed, in addition to the prostate variability, characteristics of the MRI machine (constructor, field strength...) and acquisition parameters impact the image properties and hence the possible performance of the deep learning methods. For example, [Zav+20] showed that a neural network trained on GE images had worse performance on Siemens images, and conversely. The first and most obvious option to improve generalization abilities of PCa detection methods is to include directly images from different datasets into the training set, as done in [Zav+20] and by ourselves in chapter 5. However, the normalization of the images should be carefully studied. Instead of classic normalizations like whitening or min-max normalization, which would not help mitigate domain shift impact, more specific methods such as Nyul histogram matching [NUZ00] could be considered. Ways to extend those methods on previously unseen datasets should also be considered - either by only trusting the model and its generalization abilities, by using adapted normalization processes or by using unsupervised domain adaptation [Tol+20].
The robust analysis of cases coming from multiple centers is a difficult task. For PAIMRI we circumvented the problem by gathering data coming from different hospitals but belonging to the same hospital cluster (Assistance Publique - Hôpitaux de Paris, AP-HP). To perform such an analysis, the Health Data Hub (with whom we collaborated during this PhD) is currently hosting the DAICAP[1] project, result of the collaboration between AP-HP, French national research agency for informatics (Inria), a private company (Incepto) and several hospitals from different clusters to create a large multicentric dataset mixing fully annotated retrospective data with clinical and histological information with prospective data. However, in general case, this gathering is a long and tedious task requiring multi-

[1] https://www.health-data-hub.fr/partenariats/daicap

ple authorizations from different entities. Federated learning [Sil+19] could be a way to alleviate the administrative process and to only focus on the technological problems.

Score prediction and explainability of prostate cancer diagnosis In this study, we only tried to determine if a lesion was clinically significant, using radiological criteria (PI-RADS≥ 3) or histological ones (ISUP≥2). However, it is also possible to consider the refinement of determining the exact score from the mpMRI. A few studies have tried to determine the PI-RADS score [San+20a; Yil+22], but the majority of those scoring methods skips the PI-RADS and directly tries to determine the Gleason or ISUP score [Ven+21; Dur+22; Alq+20; Cao+19], generally exploiting the ordinal aspect of those scores in their learning process. This choice to focus on Gleason score is motivated by its histological origin whereas PI-RADS is more of a radiological proxy measurement used when biopsy information is not available. As a proxy, if PI-RADS allow to detect the majority of cancerous lesions (sensitivity of 0.91), it also detects a high number of false positive (specificity of 0.37) [Dro+19]. However, today, the assessment of lesions' Gleason grade solely from mpMRI remains an open question. To the best of our knowledge, the use of PI-RADS for Gleason prediction has not been considered in the literature - except in Alqahtani et al. [Alq+20].
But all those methods for PCa detection and grading will only have a clinical application if they address one of the major drawbacks of deep learning: the black box aspects of this technology. Knowing what caused a specific output of a neural network is usually complex, whereas the GDPR requires that in the medical field (among others), an explanation to a decision should be available to the patient if required [Cou16a]. Some methods exist to provide visual explanation [GL20; Sel+20], and even our methods could be modified to improve their explainability via the exploitation of the attention modules. However, in segmentation cases, visual explanation may not be informative, since they will probably highlight the zones already distinguished by the segmentation, and in case of score prediction visual explanation is not be enough to understand the reason behind a specific scoring. However, in the specific case of PCa diagnosis case, determination of PI-RADS (or Gleason) score follows specific rules [Wei+16; Eps+16] that we have not explicitly exploited in this thesis. They could be used to a model that not only highlights the clinically significant lesions but also their score with an associated explanation [Ham+23a].

Appendix A: Flowchart for prostate cancer detection and treatment

A

In this chapter we present a flowchart of the process for detecting, assessing the risk and treating prostate cancer as recommended by the European Association of Urology[Mot+21]. This flowchart has been reproduced with authorization of Knowuro[San22]. For clarity reasons, possible adjuvant therapies and procedures in case of metastasis were not included.

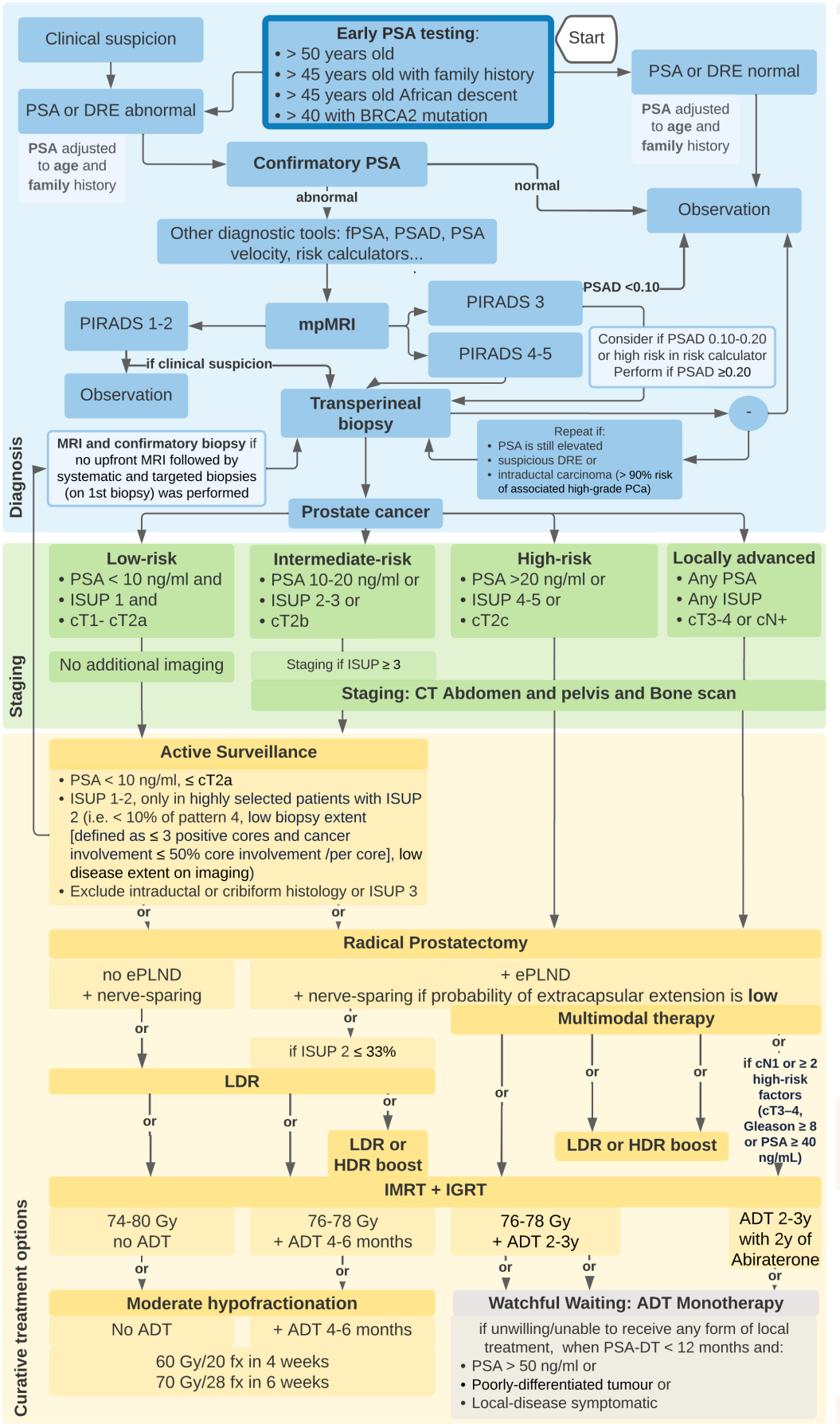


Fig. A.1.: Flowchart of EAU guidelines for prostate cancer diagnosis and treatment

B

Appendix B: Reference standard for evaluation of automatic segmentation algorithms: quantification of inter observer variability of manual delineation of prostate contour on MRI

Abstract Objective: Multiple readers are necessary to establish a reference standard of organ segmentation to develop robust and generalizable automatic segmentation algorithms. The objective of this study is to quantify inter observer variability, regarding manual prostate contour segmentation, in order to propose an ideal number of readers necessary to establish a reference standard.

Methods: Seven radiologists with various experience independently performed a manual segmentation of the prostate contour (whole-gland (WG) and transition zone (TZ)) on 40 MRI. To quantify inter observer variability, a comparative analysis of the delineations was performed using standard metrics (Dice, Hausdorff and volume-based metrics), and impact of the number of raters (from 2 to 7) on segmentation variability using two measurement strategies - pairwise metrics (consistency) and metrics with respect to a reference segmentation (conformity)– was evaluated.

Results: The average segmentation Dice score (DSC) for 2 readers in pairwise comparison was 0.919 (WG) / 0.876 (TZ). The variability decreases rapidly with the number of readers: the interquartile range of the DSC was 0.076 (WG) / 0.021 (TZ) for configurations with 2 readers, 0.005 (WG) / 0.012 (TZ) for 3 readers, and 0.002 (WG) / 0.0037 (TZ) for 6 raters. The evolution of interquartile range according to the number of raters was similar between 2 and 3 raters than between 3 and 6 raters. When using consensus methods, variability often reached its minimum with three readers (with STAPLE, WG: DSC=0.96, min-max=0.945-0.971, TZ: DSC=0.94, min-max=0.912-0.957), and interquartile range has often been found minimal for 3 raters, indicating that consensus between three readers might represent an optimal reference.

Conclusion: The number of readers impacts the inter-reader variability, in both term of inter-reader consistency and conformity to a reference. Variability has often been

found to be minimal for 3 raters, or 3 raters represent a tipping point in the variability evolution, with both pairwise-based metrics or metrics with respect to a reference. It results that three readers may represent an optimal number to determine references for AI applications.

The abstract of this appendix has been accepted to RSNA 2023 [Mol+23]

B.1 Introduction

Segmentation of the prostate on MRI plays a crucial role in numerous clinical applications, including prostate cancer detection and staging, for MRI-US biopsy fusion, and for optimal treatment planning and delivery. Repeatable and accurate prostate and/or lesion segmentation is crucial for comparing images acquired at multiple time points, for longitudinal measurements as for active surveillance or for treatment monitoring of focal therapy. It is also essential for deriving precise biomarkers to improve prostate cancer characterization and prognostic. Finally prostate contour segmentation is performed in clinical routine for multimodal fusion, ultrasound and MRI for prostate-guided biopsy, for treatment planning, more particularly in the context of radiotherapy and focal treatment.

There is currently no consensus as to the optimal technique for delineating prostate contours. Manual segmentation is one of the approaches, in which the physician determines the organ outline on the basis of visual perception of the organ border. While manual image segmentation is considered to be the gold standard, it is a tedious and time-consuming process that is subject to variability . More specifically for prostate contour, inter observer variability may depends on objective and subjective factors such as type of sequence and their quality, prostate morphology and volume, partial volume effect at the base and apex, readers experience and attentiveness. Automated algorithms have been sought in order to remove the variability introduced by raters. A good automated algorithm should require less time to apply and have better precision than segmentation by experts. However automatic structure delineation is subject to algorithm and programming bias which can be induced by a ground truth of insufficient quality. Performance of such tools, as performance of human segmentation, is difficult to quantify because a true segmentation is often not accessible. A true reference standard can only be available from phantom studies, but such phantoms do not reflect the full range of normal and anatomical variability of clinical imaging [Van+13; Gun+22; Gao+07]. A recent consensus from ESR and EORTC [DeS+22] recommend that the reference standard used, when training algorithms, should be based on segmentation by «multiple» trainer's observers. However, the ideal number of readers (and segmentations) to limit variability is still unclear.

The aim of this study is to quantify the inter-observer variability of manual delineation of prostate contour on MRI for a group of 7 independent readers and to determine the optimal number of reader needed to establish the full range of inter-observer variability. To do so, we used different set of measures and a double methodology based either on pairwise measurements or on a comparison to a reference segmentation.

Repeatable and accurate prostate and/or lesion segmentation is crucial for comparing images acquired at multiple time points, for longitudinal measurements as for active surveillance or for treatment monitoring of focal therapy. It is also essential for deriving precise biomarkers, such as PSA density or quantitative MRI parameters to improve prostate cancer characterization and prognostic. Finally prostate contour segmentation is performed in clinical routine for multimodal fusion, ultrasound and MRI for prostate-guided biopsy, for treatment planning, more particularly in the context of radiotherapy and focal treatment.

There is currently no consensus as to the optimal technique for delineating prostate contours. Manual segmentation is one of the approaches, in which the physician determines the organ outline on the basis of visual perception of the organ border. While manual image segmentation is considered to be the gold standard, it is a tedious and time-consuming process that is subject to variability [Par+20]. More specifically for prostate contour, inter-observer variability may depends on objective and subjective factors such as type of sequence and their quality, prostate morphology and volume, partial volume effect at the base and apex, readers experience and attentiveness.

Automated algorithms have been sought in order to remove the variability introduced by raters. A good automated algorithm should require less time to apply and have better precision than segmentation by experts. However automatic structure delineation is subject to algorithm and programming bias which can be induced by a ground truth of insufficient quality. Performance of such tools, as performance of human segmentation, is difficult to quantify because a true segmentation is often not accessible. A true reference standard can only be available from phantom studies, but such phantoms do not reflect the full range of normal and anatomical variability of clinical imaging, and an anatomical reference may not easily obtained [Pup+18; Gun+22; Gao+07].

An alternative and practical approach involves generating a consensus segmentation by combining the masks provided by multiple raters. However, opting for this solution presents two important issues that need to be addressed. First, determining which readers should participate in the segmentation process and how many of them should be involved to obtain a representative sample of observations that accurately reflects reality. Second, producing a consensus segmentation that captures the variability present in multiple observations. While the latter is typically approached from a computer science perspective, the former poses a clinical challenge.

Previous studies have explored the impact of readers' expertise on their segmentations [San+22; AS+23b; Mon+21], but the influence of the number of readers on the resulting consensus segmentation remains unexplored. It should be noted that simply adding more readers may not always enhance the quality of the consensus segmentation due to various reader-associated factors, including subjectivity, experience, hand-eye coordination, preferences, and motivation to dedicate time to the task. Furthermore, the number of readers is constrained by practical limitations such as time and funding.

Recognizing the significant issue of inter-reader variability, a recent consensus reached by ESR and EORTC [DeS+22] recommends that the reference standard used for training algorithms should be based on segmentations generated by "multiple" expert observers. However, an ideal number of readers remains unclear. To address this challenge, an objective assessment of the consistency between readers' segmentations and their agreement with the consensus is needed. The aim of this study is to quantify the inter-observer variability of manual delineation of prostate contour on MRI for a group of 7 independent readers and to determine the optimal number of readers needed in order to establish a reference standard for volumetric measurements and for evaluation of automatic segmentations algorithms.

B.2 Material and Methods

B.2.1 Dataset

This work was supported by the Clinical Data Warehouse of the AP-HP (Assistance Publique-Hôpitaux de Paris) and was approved by our joint institutional review boards. We compiled a cohort of 40 patients from a larger cohort/dataset (in house, n= 1200) of treatment naive patients who underwent prostate MRI before the first round of biopsy, for clinical suspicion of PCa between October 2013 and July 2019. This dataset included patients fulfilling the inclusion criterion for clinical indication of prostate MRI for suspicion of PCa (elevated prostate-specific antigen (PSA), positive DRE, genetic susceptibility) with a standardized PI-RADS V2.1 score. 150 patients were first randomly selected in the bigger dataset for an automatic segmentation project [Ham+22b], the patients included in the present study correspond to a subset of 40 randomly selected patients

B.2.2 MRI protocol

MRI exams were performed using a 3 Tesla clinical system (SIGNA ^{TM} Architect, GE Healthcare, and MAGNETOM ^{TM} Skyra, Siemens Healthcare) using a 32-channel phased-array torso coil, and 1.5 Tesla MR imaging system (MAGNETOM ^{TM} Aera, Siemens Healthcare) using a pelvic phased-array coil with 18 channels. Patients were advised to perform bowel preparation before the exam and to empty their bladder; 1mg Glucagon

	3T SIGNA™ Architect, GE Healthcare, Chicago, IL*	3T MAGNETOM™ Skyra, Siemens Healthcare, Erlangen, Germany**
Parameter	3D T2WI	3D T2WI
Sequence type	Spin echo Cube	SPACE
Field of view (mm)	280	230
Acquisition matrix	512x512	230x320
Repetition time (ms)	1602	1550
Echo time (ms)	102.87	173
Flip angle (degrees)	-	115
Slice thickness (mm)	1	0.85
Image reconstruction matrix (pixels)	0.8x0.8x1	0.4x0.4x0.85
Time for acquisition (min:s)	5min11	5min35

* Receiver frequency coils: 16-channel phased array body small coil and 32-channel spine coil.
** Receiver frequency coils: 18-channel phased array body coil and 32-channel spine coil.
T2WI = T2-weighted imaging, SPACE = Sampling Perfection with Application optimized Contrasts using different flip angle Evolution.

Tab. B.1.: MRI acquisition parameters

was administered intra muscularly to reduce peristaltic motion. All MRI protocol included 3D T2W images. All information are on Table B.1.

B.2.3 Image processing

Seven radiologists independently performed a manual segmentation of the prostate on 40 MRI (n = 280); 3 experts (>1000 prostate MRI interpreted), 2 seniors (500 prostate MRI) and 2 juniors (<100 prostate MRI). A training meeting with the 7 readers was organized before the beginning of the study, to reach an agreement on segmentation criteria. The basic zonal anatomy of the prostate was reviewed (especially base and apex limits, and the distinction between the TZ and PZ at the base). The readers were instructed to segment the whole gland (WG) and then the transition zone (TZ) on the axial plane of the 3D T2w sequence. For segmentation we used the freely available MedInria software in a polygon mode. This software allows the possibility to draw the contours as accurately as possible in multi-incidence (axial, sagittal and coronal). The mask volumes were computed using the python package SimpleITK.

B.2.4 Variability analysis

Metrics used for comparisons between segmentations

In our work, the fundamental measure involves comparing two segmentations, whether they are produced by different readers or by a reader and a reference segmentation. For that, we computed both spatial overlap-based (Dice-Sørensen coefficient or F1-score),

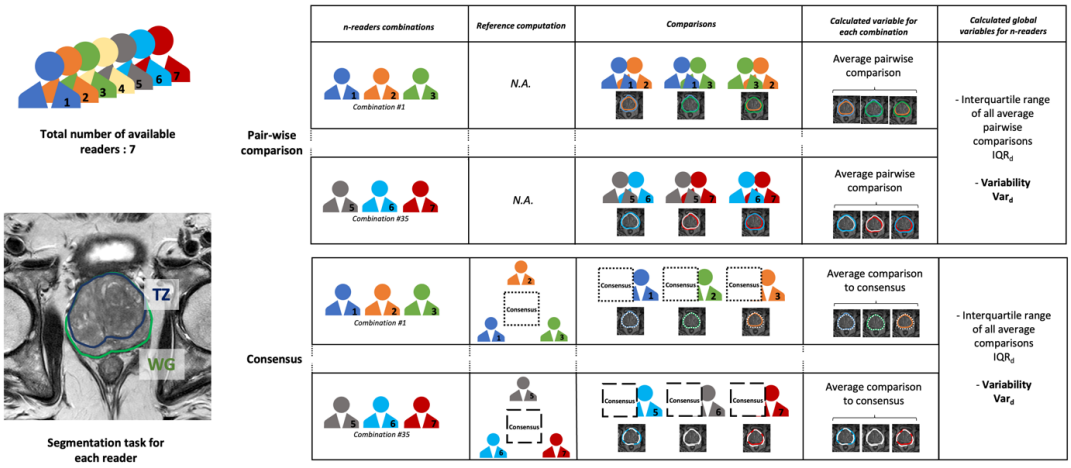


Fig. B.1.: Visual representation of the analyses made in this study. The total number of available readers was 7. Each reader manually segmented whole prostate and transition zone. We then performed separate variability analysis for pair-wise comparison and comparison to consensus. In this figure are listed the comparisons made between readers's segmentation, for n=3 readers, as well as the calculated variables.

distance-based (Hausdorff distance and average symmetric surface distance) and volume-based metrics [TH15]. These metrics are complementary, since two sets may have a large overlap but parts distant from each other. More detailed information about these metrics can be found in Annex 1.

Impact of the number of readers on overall segmentation variability

The evaluation involved two separate analyses based on the chosen metrics. Firstly, the consistency of segmentations between readers was examined by systematically comparing random pairs of readers. This analysis aimed to assess the agreement and variability between individual readers' segmentations. Secondly, the conformity of each reader's segmentation to a consensus segmentation was investigated. This analysis focused on how well each reader's segmentation aligned with the consensus. These analyses were performed considering a variable number of readers, ranging from 2 to 7. A summary of the aggregated analyses is provided in the subsequent section (more detailed information can be found in Annex 2). A visual representation of the analyses is illustrated Fig. B.1.

Consistency of readers' segmentations For each combination of n readers (ranging from 2 to 7), we calculated the average value of a specific metric d (Dice coefficient, Hausdorff distance, or ASSD) for all possible pairs of readers in this combination. This

value reflects the consistency of segmentations from a given combination of readers and represents a single data point. Since there are multiple possible combinations of readers, we further obtained an aggregated measure of this consistency across all possible combinations of n readers, which we refer to as Var_d(n): we computed the interquartile range (IQR) of the data points for each n, denoted as IQRd(n), and normalized it by the average difference in segmentations when all available raters are combined (VF_d). Similarly, we defined another measure to estimate the spread of variability: we calculated the range of the metric for each combination of n raters and divided it by VF_d. This measure, referred to as Dif_d, helps us quantify the extent of variability among the raters' measurements.

Conformity of segmentation to reference segmentations For each 3D segmentation, and each n-uplet of readers between 2 and 7, we computed two reference segmentations, one with Majority Voting and one using the STAPLE algorithm [WZW04], which estimates the consensus segmentation by weighting each input by its level of performance. The first method simply consists in selecting voxels segmented by more than half of the radiologists into the consensus. The second method, among the most popular ones to construct a consensus from binary masks, is an Expectation-Maximization algorithm where each rater is characterized by their sensitivity and specificity. Those values are used to compute the posterior probability of each voxel to belong to the structure and are then updated using the constructed consensus, until convergence. We then reproduced the steps above to obtain the average comparison to the reference segmentation, for a given metric and a given number of readers, representing a single datapoint, as well as the aggregated measures for all possible combinations, Var_d(n) and Dif_d.

Evolution of segmentation volumes according to the number of readers Finally, inspired by Joskowicz et al. [Jos+19], we evaluated the impact of the number of raters on the obtained consensus. For a given n number of readers, we computed the following volumes:

- The union of all segmentations by n readers: corresponding to all the possible volumes
- The intersection of all segmentations by n readers: corresponding to the minimum volume agreed upon by these readers.
- The consensus segmentation obtained by STAPLE [WZW04] and majority voting by n readers

All these values were normalized by the average volume obtained by n readers.

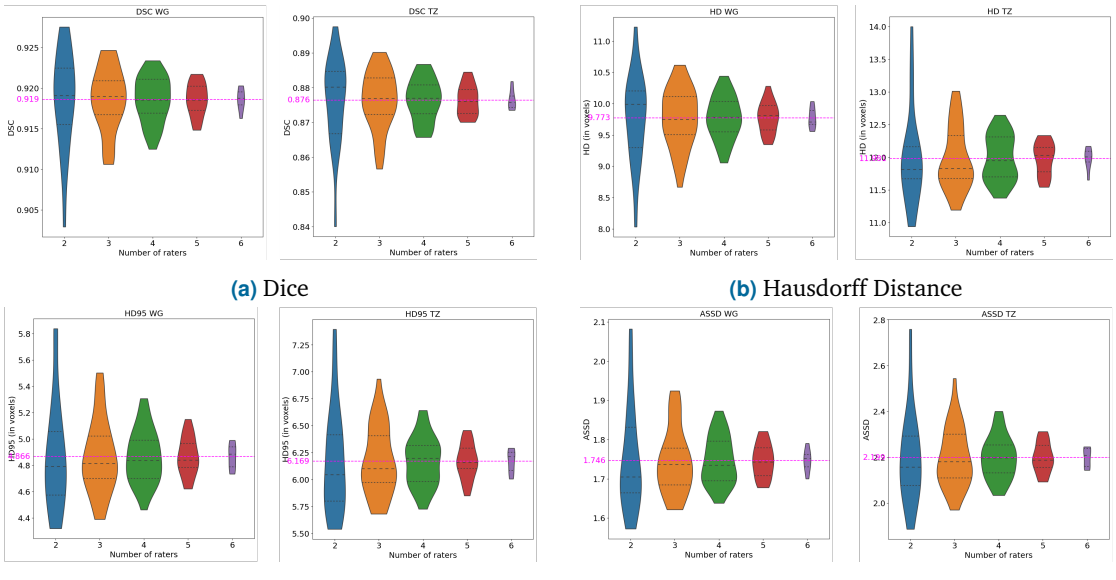
B.3 Results

B.3.1 Impact of the number of readers on overall segmentation variability

Consistency between all prostate segmentations (whole gland and transition zone) according to the number of readers.

The average segmentation Dice score (DS) for 2 readers in pairwise comparison was 0.919 (WG) / 0.876 (TZ).

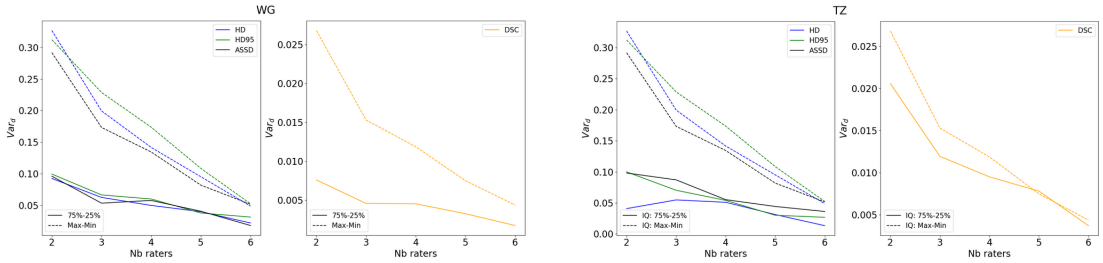
Figure B.2 and B.3 illustrates the relationship between the number of annotators and the inter-reader segmentation variability (consistency), as measured through pairwise comparisons. As the number of annotators increases, the range of values for inter-reader variability estimation became narrower (Figure B.2), and the interquartile value decreased (Figure B.3). The interquartile range of the Dice score was 0.0076 (WG) / 0.021 (TZ) for configurations with 2 readers, 0.005 (WG) / 0.012 (TZ) for 3 readers, and 0.002 (WG) / 0.0037 (TZ) for 6 readers. Overall segmentation consistency remains high in all readers configurations: when considering combinations of only 2 raters, the inter-reader variability (measured using the Dice metric) ranged from 0.903 to 0.928 for WG and from 0.840 to 0.898 for TZ.



(a) Dice **(b)** Hausdorff Distance

(c) 95% Hausdorff Distance **(d)** Average Symmetric Surface Distance

Fig. B.2.: Violinplot of mean pairwise metrics according to the number of raters for all four metrics (DSC, HD, HD95% and ASSD) for both WG and TZ. Final value obtained with the 7 raters is indicated on the graph in purple. The width of each violinplot corresponds with the approximate frequency of data points for each value of the corresponding metric. Inside each violinplot, horizontal lines also show the data distribution, with a central longdashed line indicating the median value and two other dashed lines indicating the range of the central 50% of the data.



(a) WG **(b)** TZ

Fig. B.3.: Evolution of Var_d according to the number of raters on both structures with pairwise metrics. Full lines correspond to Var_d (=75%-25%), dotted lines correspond to Dif_d (=max-min)

Conformity to a reference segmentation (STAPLE and Majority Voting).

Figure B.4 and B.5 illustrate the relationship between the number of annotators and the inter-reader segmentation variability, as assessed through comparisons to a reference segmentation (conformity). Using the STAPLE [WZW04] consensus method (Fig. B.4a, b), the conformity between readers' segmentation and the consensus segmentation was comparable with 2 and 3 annotators, the better for 3 annotators and decreased for 4 or more annotators. Additionally,as illustrated in Fig. B.5 the minimal range of variability was seen for cases with n = 3 (WG: DSC=0.959, min-max=0.945-0.971, TZ: DSC=0.938, min-max=0.911-0.957). Likewise, using the majority voting method (Fig. B.4c, d), the conformity was the higher with 2 and 3 annotators (similar values), even though the results were not as marked as with STAPLE consensus. Similar results were obtained using Hausdorff distances as well as average surface distance, with the highest conformity between readers' segmentation and the consensus segmentation obtained for 2 and 3 raters. As for the range of those variabilities, illustrated in Figure 5, a decrease of variability (better conformity) could be observed with all methods and all metrics between 2 and 3 raters, before either stabilizing (Majority Voting) or increasing (STAPLE).

B.3.2 Evolution of segmentation volumes according to the number of readers

The evolution of consensus volume according to the number of raters is available in Fig. B.6. As anticipated, the union volume of all segmentations (i.e., all voxels included in at least one reader's annotation) grew at a faster rate than the average volume as the number of readers increases, while the intersection of all segmentations (i.e., all voxels included in every reader's annotation) decreased. We showed that the increase in the union volume ratio for the whole gland was substantial when the number of readers increased from 1 to 3, with a nearly 15% increase observed. However, we also observed that the increase was much less significant when the number of readers increased from 3 to 5, resulting in less than a 5% increase in the union volume ratio. We observed a

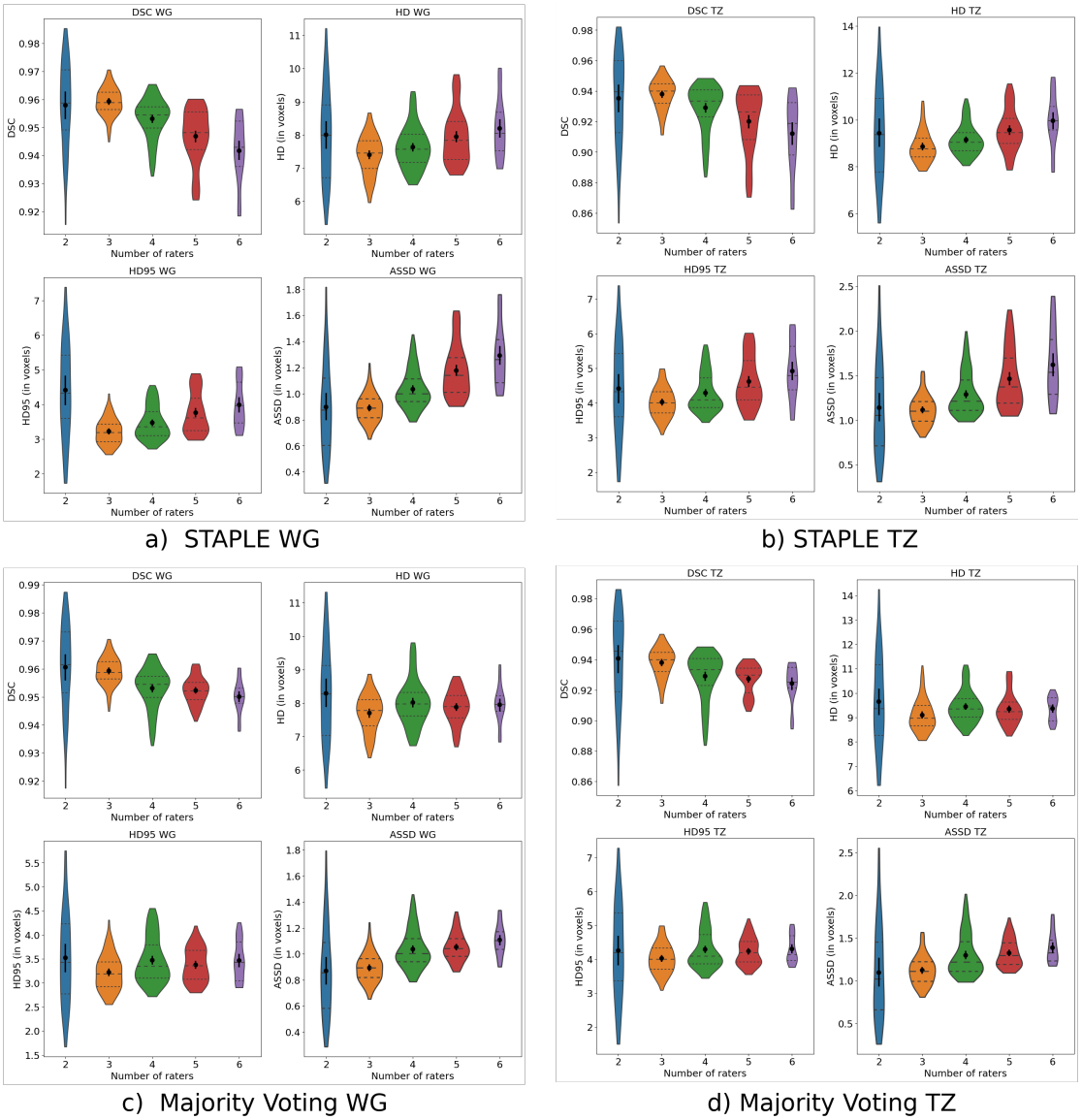


Fig. B.4.: Violinplot of mean pairwise metrics according to the number of raters for all four metrics (DSC, HD, HD95% and ASSD) for both WG and TZ. The purple line shows the Vf_d value for the corresponding metric.

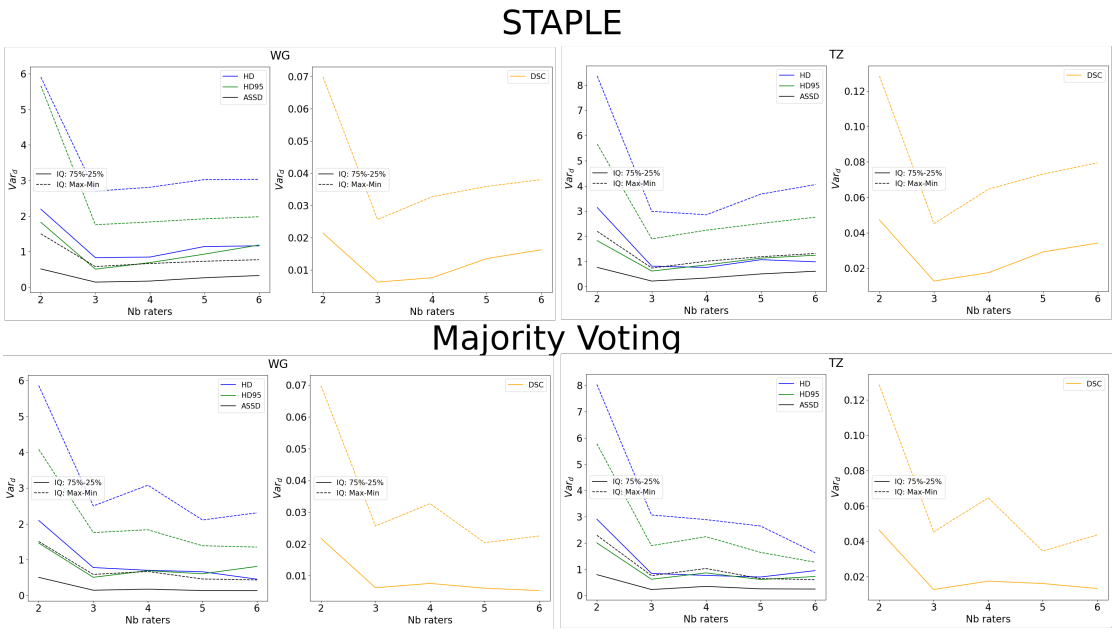


Fig. B.5.: Evolution of Var_d and Dif_d according to the number of raters on both structures with pairwise metrics. Full lines correspond to Var_d (=75%-25%), dotted lines correspond to Dif_d (=max-min)

similar pattern for transition zone segmentation. When comparing the volume resulting from the STAPLE consensus method to that resulting from majority voting, we noted that the STAPLE volume increased at a faster rate, although this increase was relatively limited and only observed with more than four annotators.

B.4 Discussion

In this study, our objective was to assess the inter-observer variability in the manual delineation of the prostate contour by involving a group of 7 independent observers. We employed various metrics and algorithms to quantify this variability and determine

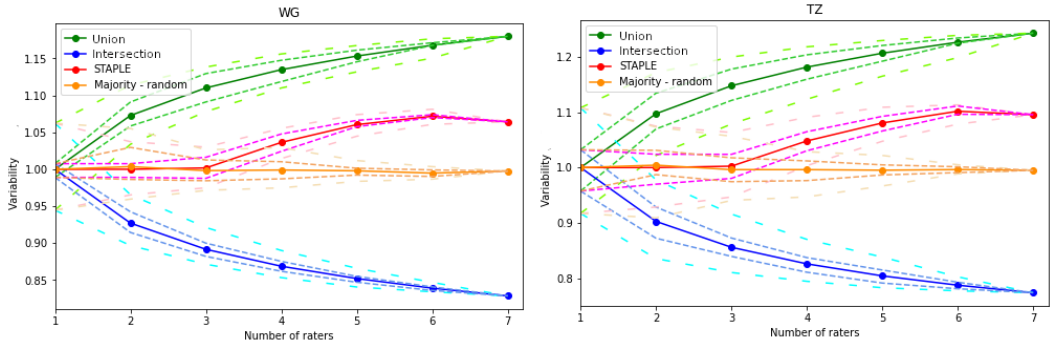


Fig. B.6.: Evolution of the consensus size according to the number of raters involved in the consensus computation. Left: WG, Right: TZ. Y-axis represent the ratio between the consensus volume and the segmentations' average volume. Dashed lines represent 1 and 3rd quartiles, longdashed lines represent min-max.

the optimal number of readers required to achieve a consensus. Through exhaustive combinations and the use of different metrics, we discovered that a combination of three readers yielded the most reliable reference standard with minimal delineation variability for volumetric measurements and the evaluation of automatic algorithms.

Several studies have evaluated inter-rater variability of manual segmentation of different structures. A non-exhaustive list of studies dedicated to prostate segmentation can be found in Table B.2.

Recently, a systematic review of the literature was conducted to comprehensively analyze and compare the applicability and efficiency of published methods for automatic segmentation of the prostate gland [Wu+22]. Among the 33 studies analyzed, the majority (31 teams) reported employing multiple readers using various approaches such as splitting, stratification, and blinding. The number of readers ranged from 2 readers in 6 studies to 12 readers in 1 study. Surprisingly, only 4 studies assessed inter-rater variability, revealing that it was generally low for the segmentation of the entire gland. However, greater variability was observed at the extreme base and apex of the gland. It is noteworthy that none of these studies specifically investigated the optimal combination of readers to establish a consensus.

Sharp et al. [Sha+14b] referenced more than 14 studies estimating inter-rater segmentation variability on 25 organs, a list far from being exhaustive. Comparisons between all those studies are complex; number of raters are variable, organ and/or lesion to be segmented are different, as well as the chosen imaging modalities. For example, on CT, a state-of-the-art method for automatic segmentation [Ise+21] has been able to perform liver segmentation from with a mean Dice score of 0.96 but could not perform better than a Dice score of 0.66 on hepatic vessels segmentation. Those reasons make it difficult to establish one final and definitive conclusion on inter-rater variability, as it is wide and differs significantly between structures evaluated and modality used.

To the best of our knowledge, the study conducted by Joskowicz et al. [Jos+19], stands as the sole investigation that evaluates inter-rater variability of segmentation of different structures (liver, lung tumors, kidney contours, and brain hematoma on CT) by the same readers. The study involved a large number of independent raters, specifically 11 individuals with varying degrees of expertise. The primary objective was to establish a reference standard for the evaluation of automatic segmentation tools. The authors conducted a thorough assessment of manual tracing variability for each case. Additionally, they performed pair-wise comparisons between random pairs of observers, taking into account the stratification by case and structure type. The expertise of the observers and groups of observers were also considered during the analysis. Interestingly, the authors introduced two noteworthy group-wise metrics in their study. They computed a volume called “consensus” including voxels that are included in all delineations, a volume called “possible” including voxels that are included in at least one delineation and the difference between these two volumes, referred to as the “variability”. These metrics

Article	Zonal	N. raters	N. cases	Pairwise/ Consensus	Impact of experience	Other studied parameters
Adair Smith et al. [AS+23a]	No	2	6	Pairwise	No difference between radiographers and clinicians	-
Salvi et al. [Sal+22]	No	2	15	Pairwise	-	-
Meyer et al. [Mey+19]	Yes	2-3	20-10	Pairwise	-	-
Pathmanathan et al. [Pat+19b]	No	3	10	Pairwise	-	Imaging
Shahedi et al. [Sha+17]	No	3	10	Both	-	Thirds
Shahedi et al. [Sha+14a]	No	3	10	Both	-	Thirds
Gardner et al. [Gar+15]	No	5	10	Consensus	-	-
Khalvati et al. [Kha+16]	No	5	15	Pairwise	-	-
Pathmanathan et al. [Pat+19a]	No	5	15	Pairwise	-	-
Sanders et al. [San+22]	No	7	25	Pairwise	Difference between radiation oncologists and clinical observers	Imaging, volume
Adair Smith et al. [AS+23b]	No	10	10	Consensus	No difference between radiographers and clinicians	-
Sabater et al. [Sab+21]	No	9	10	Consensus	-	-
Nyholm et al. [Nyh+13]	No	10	25	Pairwise	-	Volume
Padgett et al. [Pad+19]	Yes	2	30	Pairwise	-	Thirds, imaging
Shahedi et al. [Sha+16]	No	3	10	Consensus	-	-
Liu et al. [Liu+12]	No	5	23	Consensus	-	Thirds
Becker et al. [Bec+19]	Yes	6	80	Pairwise	Difference between radiologists and non-radiologists	Thirds
Montagne et al. [Mon+21]	Yes	7	40	Both	No influence of experience within radiologists	Thirds, volume, intensity ratio

Tab. B.2.: Selected studies on inter-rater prostate segmentation variability

were calculated for different number of readers, from 2 to 10. Authors demonstrated that the volume overlap variability for a large group of delineations is wide and differs significantly between structures (minimal for kidney contours). It increases as a function of the number of observers in groups. Indeed, the volume of “possible” segmentations (pixels segmented by at least one reader) monotonically increased with the number of observers: 37% for two observers, 53%, 72% and 85% for 3, 5 and 8 delineations (after nine delineations the contribution of each additional delineations was less than 5%).

In our study on prostate segmentation, we also observed an increase in the maximum “possible” volume with the number of observers. However, in contrast to Joskowicz’s findings, the rate of increase was significantly lower for more than 3 readers. Specifically, the increase in volume between 3 and 5 readers was less than 5%.

Our study was the first to investigate the influence of the number of readers on consensus formation. It is important to note that the term "consensus" volume used by Joskowicz et al. does not refer to a practical method of consensus formation, but rather represents the minimal set of pixels on which all readers agree. In contrast, our study employed two reference segmentation methods: the STAPLE method and the Majority Voting approach. The STAPLE algorithm provided a probabilistic estimation of the true contour by incorporating the manually drawn contours contributed by all raters. This allowed us to evaluate the consistency of each individual segmentation with respect to the state-of-the-art in consensus formation. Our results indicated that the conformity to the consensus segmentation was highest for k=2 and k=3 readers. However, the variability range and interquartile were consistently lower for k=3, with only marginal improvements observed for higher k (as illustrated in Figure 5). It is worth noting that the observed high variability for k=2 arises from a specific case in consensus computation, where the consensus is determined by the intersection of two segmentations. This particular case is more dependent on the individual choices made by radiologists compared to other cases. These findings were consistent across both overlap-based metrics (such as Dice coefficient) and Hausdorff distances, suggesting that the optimal number of readers may be the same despite their differences.

Another fundamental distinction between our work and that of Joskowicz et al. lies in the types of organs or lesions that were segmented. In their study, the authors considered kidney contouring as the simplest task due to its lower volume variability, while other tasks involved contouring pathological processes with more ambiguous contours, resulting in higher variability. Among these tasks, the segmentation of brain hematoma exhibited the highest level of variability. The segmentation of the prostate on MRI presents challenges similar to those encountered in segmenting other abdominal or pelvic organs. Specifically, these challenges are related to the inconsistent visibility of the prostatic capsule , the difficulty in distinguishing between different prostatic zones and the important variability of prostate gland morphology.

Several limitations can be found in this study. First, our study was only done on a small size set – 40 cases. However, we believe that the high number of radiologists who provided segmentations allow us to derive conclusions despite this small number. Second, all the segmented cases included in this study as well as the annotators are from the same institution. In consequence, there could be hidden biases due to similar imaging acquisition or in-house methodologies in the segmentations. Moreover, contrary to Joskowicz et al [Jos+19], in our study we only focused on one organ: the prostate. The results we obtained may be expanded to organs with similar characteristics but can not be expanded to all medical segmentations (such as tumor segmentations, for example). A methodological limit to this study is the absence of statistical significance between the different measures according to the number of raters. This is due to the complexity of such a study, from a statistical point of view: each data point can share a certain number of raters, and thus data for different numbers of raters are neither independent nor can be paired (for example, each data point for the case n=2 has 2 raters in common with 5 data points of the case n=3, and 1 rater in common with 20 more of them.)

B.5 Conclusion

The number of readers impacts the inter-reader variability, in both term of inter-reader consistency and conformity to a reference. Variability has been found to be minimal for 3 raters, or 3 raters represent a tipping point in the variability evolution, with both pairwise-based metrics or metrics with respect to a reference. It results that three readers may represent an optimal number in order to establish a reference standard for volumetric measurements of the prostate gland and for evaluation of automatic segmentations algorithms

B.6 Appendices

B.6.1 Metrics used for comparisons between 2 segmentations

The Dice-Sørensen coefficient (DSC, also known as F1-score) is one of the most used metrics for comparing segmentations. For two binary sets A and B, it is defined as $DSC(A,B) = \frac{2|A \cap B|}{|A \cup B|}$. A Dice coefficient of 0 indicates no overlap between the two sets, an DSC of 1 indicates a perfect overlap (i.e. A=B)

The Hausdroff distance (HD) corresponds to the maximal distance between the two sets, equal to 0 if the two sets are equal. For two binary sets A and B it is defined as $HD(A,B) = \max(\max_{a \in A} \min_{b \in B} d(a,b); \max_{b \in B} \min_{a \in A} d(b,a))$, with d the classic Euclidean distance. However, it is sensitive to outliers, as only one point can heavily impact it. In consequence, the Hausdorff distance 95% (HD95), defined as as the 95th percentile of the ensemble $\{d(a,B) \forall a \in A\} \cup \{d(A,b) \forall b \in B\}$, is more robust to outliers than the classic one. Its combination with Dice coefficient allows to consider at the same time the extent of the overlap between two segmentations and the "maximal distance" between them.

The Average Symmetric Surface distance (ASSD) is defined as the average of all the distances from points on the boundary of the first segmentation to the boundary of the other segmentation, and vice versa:$ASSD(A,B) = \frac{1}{|A|+|B|} \sum_{a \in A} \min_{b \in B} d(a,b) + \sum_{b \in B} \min_{a \in A} d(b,a)$. It provides information on the correspondance of both sets' global shape, as it is defined on all points from both sets.

Finally, segmentation volume of the binary set A corresponds to the number of segmented voxels, converted into cc.

B.6.2 Assessment of the consistency between readers' segmentations

For each metric d and for a number n of readers between 2 and 7, we went through the following steps:
We first determined all the possible combinations of n readers (n-uplets):

Raters	2	3	4	5	6
C(7, n)	21	35	35	21	7

For each combination of n readers we then defined the average pairwise comparison, as the mean of the pairwise comparisons, between all couples of readers present in the n-uplet, using the metric d (Dice, Hausdrauff distance or ASSD):

$$d(A_1, \ldots, A_n) = \frac{n(n_1)}{2} \sum_{i,j=1,i<j}^{n} d(A_i, A_j)$$

Thus, each datapoint of this analysis represents a specific combination of readers and corresponds to the average difference between two segmentations of this readers group.

For any given number n of readers, considering all possible combinations of n readers, we computed a normalized robust estimation of the variability: $Var_d(n) = \frac{IQR_d(n)}{Vf_d}$ with $IQR_d(n)$ the interquartile range of the measures with the metric d for n raters and Vf_d the average difference of segmentations in the group containing all the seven raters altogether: $Vf_d = d(A_1, \ldots, A_7)$.

Similarly, we defined the normalized range of variability: $Dif_d(n) = \frac{range(n)}{Vf_d}$ with range(n) the interquartile range of the measures with the metric d for n raters.

It should be noticed that, by design, the mean value is always equal to Vf_d no matter the number of raters, and that the obtained values converge towards Vf_d. However, in this case the subject of interest is the speed of convergence of both Dif_d and Var_d.

Appendix C: Automatic segmentation of prostate zonal anatomy on MRI: a systematic review of the literature

C

Objectives: Accurate zonal segmentation of prostate boundaries on MRI is a critical prerequisite for automated prostate cancer detection based on PI-RADS. Many articles have been published describing deep learning methods offering great promise for fast and accurate segmentation of prostate zonal anatomy. The objective of this review was to provide a detailed analysis and comparison of applicability and efficiency of the published methods for automatic segmentation of prostate zonal anatomy by systematically reviewing the current literature. Methods: A Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) was conducted until June 30, 2021, using PubMed, ScienceDirect, Web of Science and EMBase databases. Risk of bias and applicability based on Quality Assessment of Diagnostic Accuracy Studies 2 (QUADAS-2) criteria adjusted with Checklist for Artificial Intelligence in Medical Imaging (CLAIM) were assessed. Results: A total of 458 articles were identified, and 33 were included and reviewed. Only 2 articles had a low risk of bias for all four QUADAS-2 domains. In the remaining, insufficient details about database constitution and segmentation protocol provided sources of bias (inclusion criteria, MRI acquisition, ground truth). Eighteen different types of terminology for prostate zone segmentation were found, while 4 anatomic zones are described on MRI. Only 2 authors used a blinded reading, and 4 assessed inter-observer variability. Conclusions: Our review identified numerous methodological flaws and underlined biases precluding us from performing quantitative analysis for this review. This implies low robustness and low applicability in clinical practice of the evaluated methods. Actually, there is not yet consensus on quality criteria for database constitution and zonal segmentation methodology.

This appendix has been pusblished in Insights into Imaging [Wu+22].

C.1 Introduction

Magnetic resonance imaging (MRI) is the first imaging choice for detecting and localizing prostate cancer [Mot+21; Roz+20], based on the Prostate Imaging Reporting and Data

System (PI-RADS) scoring system [Tur+19] and depending on zonal anatomy. Zonal segmentation of the prostate plays a crucial role for prostate cancer detection as the PI-RADS score differs depending on the areas studied, based on diffusion-weighted imaging (DWI) for peripheral zone lesions and T2-weighted (T2W) imaging for transitional zone lesions, but also for multiple clinical application such as reproducible prostate volume and Prostate Specific Antigen (PSA) density evaluation [Ben+92], MRI-ultrasound fusion biopsy, radiotherapy, or focal planning.

Zonal segmentation of the prostate is usually performed manually on T2W images by contouring the prostate in a slice-by-slice manner. It is extremely time-consuming, tedious, and prone to inter and intraobserver variability due to the subjective human interpretation of organ boundaries and large variability in prostate anatomy and gland intensity heterogeneity across patients [Kor+15]. There is a real need to develop automatic methods to accelerate the whole process and offer robust and accurate prostate segmentation.

Automatic zonal segmentation of the prostate is a challenging task for multiple reasons. Prostate gland is subject to large morphological variation, intra-prostatic heterogeneity, and poor contrast with adjacent tissues, making delineation of prostatic zonal contours laborious. Multi-institutional applicability can be difficult to evaluate as there is a wide technically induced variability in the image acquisition, as MRI signal intensity is not standardized and image characteristics are strongly influenced by acquisition protocol, field strength, scanner type, coil type, etc. [Zav+20].

Finally, the performances of an automated segmentation method depend in part on the database (heterogeneity of the data used, knowledge of possible selection biases), quality of ground truth (manual delineation of the prostate performed by human experts), training time and hardware requirements. First commonly used methods were based on machine learning methods, such as atlas-based registration models in which several reference images with corresponding labels are registered and deformed onto the target image [Lit+12; Pad+19] or C-means clustering models [Chi+16; MBC14]. Most common methods described after 2017 are based on deep learning with convolutional neural networks (CNN) allowing automatic extraction of features and semantic image segmentation. Common architectures such as U-net [RFB15], V-net [MNA16] or ResNet [He+15] have been extensively used. Modification and fine tuning of existing models, by either combining multiple U-nets [Zhu+19; Zab+19; Cla+17], adding attention modules such as squeeze and excitation [Run+19], feature pyramid attention [Liu+19], adding blocks [Kha+19], transition layers or up-sampling strategies [Nai+20], allowed either improving accuracy of classical CNN or obtaining same accuracy with reduced memory and storage requirements.

The primary objective of this review was to provide a detailed analysis and comparison of applicability and efficiency of the published methods for automatic segmentation of prostate zonal anatomy by systematically outlining, analyzing, and categorizing the

relevant publications in the field to date. We also aimed to identify methodological flaws and biases to demonstrate the need for a consensus on quality criteria for database constitution and prostate zonal segmentation methodology.

C.2 Materials and methods

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement (PRISMA) [Pag+21]. The methods for performing this systematic review were registered on PROSPERO [Boo+11] database (registration number CD42021265371), and were agreed by all authors before the start of the review process to avoid bias. This study was exempt from ethical approval at our institution because the analysis involved only deidentified data.

C.2.1 Data sources and search

Medical literature published in the English language published until 30 June 2021 was searched in multiple databases (Medline, Science direct, Embase and Web of Science) using the following terms:

`(prostatic OR prostate) AND (automated OR automatic) AND (segmentation OR segmented) AND (zone OR zonal) AND ("magnetic resonance" OR mri OR \"magnetic resonance\"OR mri OR mr) AND (\"artificial intelligence\" OR \"deep learning\" OR \"machine learning \")` and all possible combinations.

No beginning date was applied.

C.2.2 Study selection

Full-text selection was independently performed by two radiologists, one experimented radiologist specialized in uroradiology and prostate imaging (S.M., 5 years in prostate imaging, with more than 1000 cases of prostate MRI per year) and one radiology fellow specialized in uroradiology and prostate imaging (C.W., 1 year in prostate imaging, with more than 1000 cases of prostate MRI per year). A third experimented professor of radiology specialized in prostate imaging (R.R.P., 15 years in prostate imaging, with more than 1000 cases of prostate MRI per year) intervened in case of disagreement. We summarized search strategy details for each database in Fig. C.1.

We imported all articles retrieved into the reference manager Zotero and removed all duplicates. The same two radiologists (C.W., S.M.) then independently and manually screened titles and abstracts of the resultant database to ensure relevance. Articles that were obviously out of the scope of the research topic were excluded at this stage. Subsequently, all the remaining articles full texts were retrieved and read, applying inclusion and exclusion criteria (explained below) with conflicts resolved by consensus

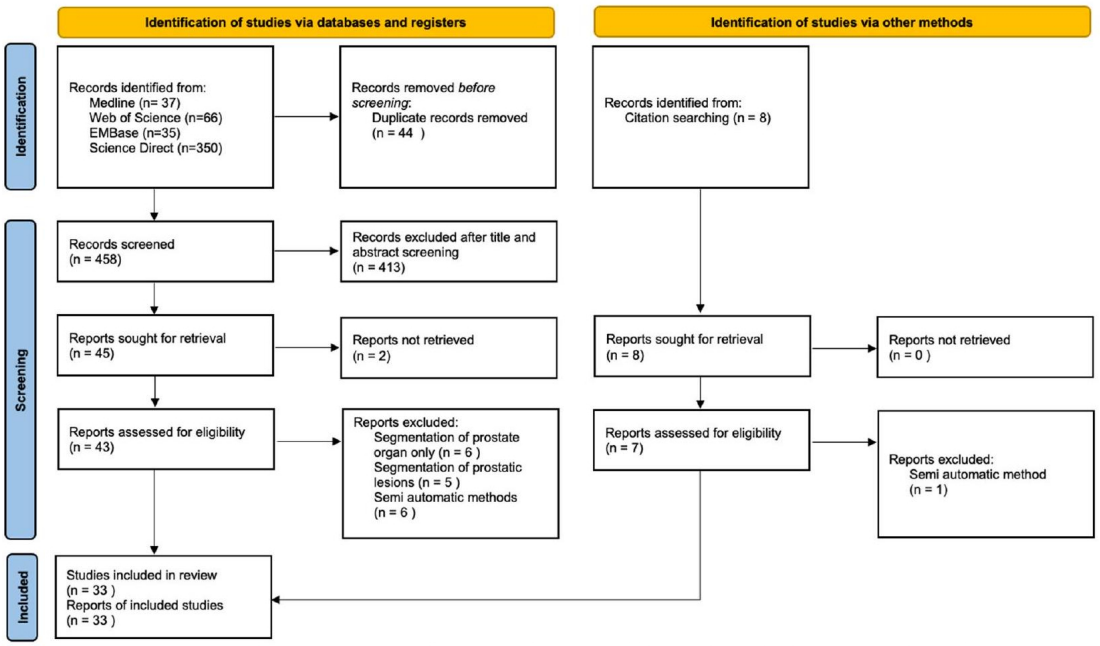


Fig. C.1.: Flow diagram based on Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) recommendations for systematic reviews

with the third reviewer. Reference lists of these relevant articles were also reviewed for possible papers missed in the primary search, and those papers were screened using

C.2.3 Selection criteria

Inclusion criteria Articles were included if they were original articles, used machine learning or deep learning algorithms and aimed to segment prostate human MRI images by zonal anatomy, using a fully automated method with manual segmentation as ground truth.

Exclusion criteria Articles were excluded if they were commentaries, editorials, letters, case reports or abstracts. Were also excluded articles with semi-automated segmentation methods, no description of segmentation method, segmentation of the whole gland (WG), or prostate cancer without zonal anatomy, absence of similarity metrics or of evaluation against ground truth segmentations.

Data collection and extraction process The qualifying papers were then reviewed, and various data of the studies were extracted and tabulated prior to analysis (Table C.1).

Assessment of methodological quality The two same radiologists (C.W., S.M.) independently assessed and extracted data from each of the included articles, using the Quality Assessment of Diagnostic Accuracy Studies tool-2 (QUADAS-2) framework [FW+11]

Sources	Patients	Data	Flow and timing	Reference standard	Test
Scientific database	Public or in-house database	Vendor	Cross-validation	Type of annotation	Validation or test on external data
Title	Eligibility criteria: inclusion and exclusion criteria	Field	Splitting in training, validation and test set	Annotation tool if used	Performance metrics
Authors	Sample size	Array		Number of annotators	Results based on DSC
Year of publication	Ethic consent	Field of view		Ground truth segmentation and rationale	
Journal name	Presence of benign prostate hypertrophia	Pre-processing		Measurements of inter- and intra-rater variability if any	
	Presence of prostate cancer	Post-processing		Type of annotators	
	Percentage of prostate cancer	Number of vendors		Experience of annotators	
	Uni or multicentric	Slice thickness			
	Prospective or retrospective	Type of slice and sequence			
		Cross-validation			

DSC = Dice Similarity Coefficient

Tab. C.1.: Data extraction.

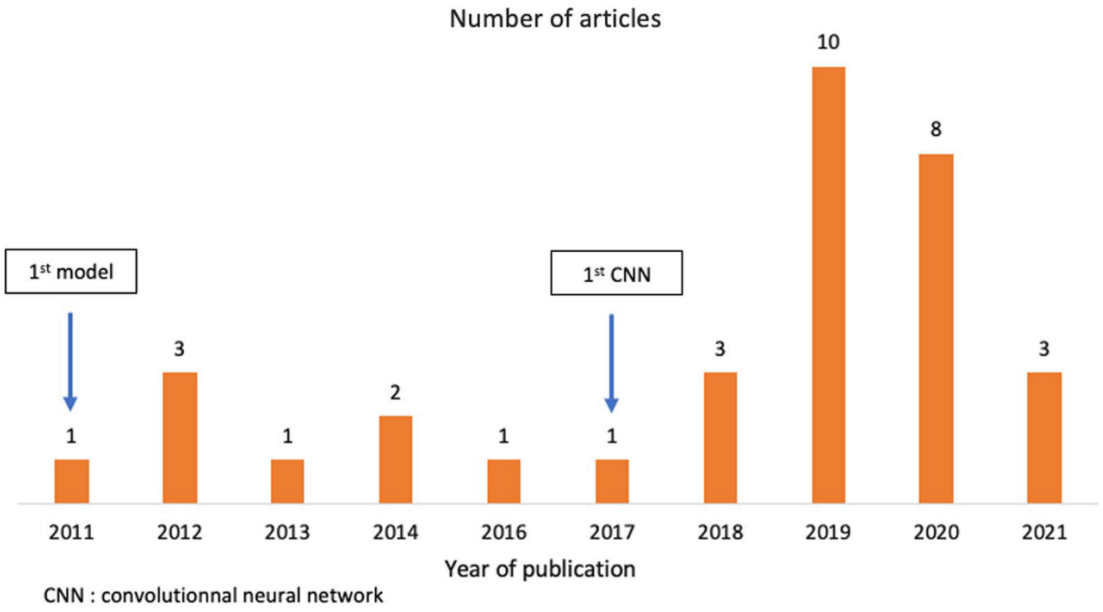


Fig. C.2.: Chronological distribution of the 33 reviewed articles. 1st model for prostate zonal anatomy segmentation was published in 2011. 1st convolutional neural network (CNN) was published in 2017

adjusted with topics from the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [MMK20] to evaluate the risk of bias and applicability for each selected study, with conflicts resolved by consensus with the third reviewer.

Extracted data were tabulated, synthesized, and evaluated for methodological flaws and applicability of the proposed techniques.

C.3 Results

After removing duplicates, 458 articles were remaining. Final consensus was reached yielding a total of 33 articles [Zav+20; Lit+12; Pad+19; Chi+16; MBC14], [Zhu+19; Zab+19; Cla+17; Run+19; Liu+19; Kha+19; Nai+20], [Mak+11; Yin+12; MG12; Tot+13; Chi+14; Can+18; MBH18; Che+17; Jen+19; Ham+19; Run+20; Mey+19;

Mot+20; Qin+20; Liu+20; Lee+20; Ald+20; San+20b; Lai+21; Bar+21; Cuo+21a] (Figs. C.1,C.2).

C.3.1 Datasets

Training, validation, and test sets All articles used retrospective datasets.

Wide heterogeneity in training, validation and test datasets was found (Table C.2).

Performance testing of the algorithms can be done on same source than for the development or use different source of data, and based on either public data, private data or a combination of both. Public data were used in 15/33 articles for testing. Only 7 studies [Zav+20; Chi+16; Cla+17; Che+17; Run+20; Qin+20; Liu+20] used both private and public data for testing, allowing better generalization of their algorithms. None of them used prospective data for validation and testing.

Most used public datasets were PROSTATEx [Arm+18], NCI-ISBI 2013 [NB15] and PROMISE12 [Lit+14b] (Additional file 1: Table S1).

Eight authors applied cross-validation, using a subset of available dataset as training set, while the remaining data constituted the test set to evaluate the segmentation performance and accuracy. Nine reported using cross-validation for testing, averaging the results from the different rounds, hence adding bias.

Technique We identified major technical differences in datasets regarding the number of vendors, field strength, type of coils, sequences, slice thickness, field of view (FOV) and input data used for automatic segmentation (Table C.3). Less than half (14/33) studies used more than one type of vendors and 7/33 used both 1.5 T and 3 T MRI machines. More than 2/3 (24/33) used mono-modal input, mainly T2-weighted planes, in combination with apparent diffusion coefficient (ADC) map in one study [Zab+19] or with multiparametric and multi-incidence MR images in another [Chi+16]. The slice thickness of T2-weighted axial planes was consistent with the PI-RADS v2.1 recommendations in 13/33 studies ($\leq$ 3 mm), which was not the case for the public data base PROSTATEx (3.6 mm). Only 7 studies provided sequence details (type of sequence, slice thickness, FOV) used for ground truth manual segmentation.

C.3.2 Zonal anatomy

We found 18 different types of very heterogeneous and unclear terminologies of zonal anatomy (Fig. C.3, Additional file 1: Fig. S1). Out the 33 articles reviewed, less than 1/4(8/33) [Mak+11; MG12; Ham+19; Mey+19; Qin+20; Liu+20; San+20b; Cuo+21a], provided precise terminology and segmentation protocol. Frequently the inappropriate term "central gland" (CG) was used, with ambiguous definition of central zone (CZ) and anterior fibro-muscular stroma (AFMS) alternatively included in peripheral

zone (PZ) or transition zone (TZ), or mainly not described at all. Two studies mis-used the term "central zone" to refer to the "central gland" [Chi+14; Ald+20].

C.3.3 Ground truth

Manual delineation of the prostate gland performed by human experts was used to generate ground truth (Table C.4).

Annotation tool Twenty studies (61%) reported using manual contouring, while a third (11/33) reported using annotation tools. One team [Jen+19] specified that the radiologist did not delineate zones on all slices but relied on interpolation performed by their annotation tools. Two studies [Ham+19; Run+20] did not provide any information.

C.4 Qualifications of annotators

Most studies (27/33, 81%) reported a radiologist or a radiation oncologist as human expert. In 3 papers, no detail was provided on annotators qualification, although one [Run+19] specified using an "expert" reader. Definition of an "expert" reader was mostly unclear with no specification of number of MRI they interpreted, for example [MBC14; Run+19; Tot+13; Jen+19; Mey+19; Ald+20].

C.5 Number of readers

Number of readers and their experience are described in Table C.4. Number of readers was not available in two studies. While 2/3 of teams (22/33) reported using more than one reader, with splitted, stratified or blinded reading approaches, 7 did not provide information on reading approach.

C.6 Intra and inter-rater variability

Inter-rater variability for annotations was rated in only 4 studies [Lit+12; MBC14; Mak+11; Ald+20]. Some studies used alternative techniques to approach better homogeneity of ground truth. In [Zab+19], the four radiologists met for a training session and together segmented two example patients to achieve a similar methodology for the rest of the dataset, using only experienced radiologists. In [Zav+20], the contours segmented by three radiologists were cross-checked and reviewed by two radiation oncologists, resulting in better homogeneity of ground truth. In [Nai+20], the initial prostate masks were drawn by two students who were trained in segmenting prostate zones.

C.6.1 Risk of bias and quality assessment

The detailed results are presented in Fig. C.4 and Additional file 1: Table S2.

Regarding patient selection, we considered a low risk of bias if there were clear data inclusion and exclusion criteria, inclusion of patients with and without PCa. Models were considered less applicable if datasets were composed of only one type of scanners or if no information was specified. For reference standard, number of readers and type of reading for ground truth segmentation were reviewed. Clear partitioning of the database (into training, validation, and test sets) was needed to waive risk of bias for flow and timing. Some articles used cross-validation methods without keeping a clear independent test dataset [Zav+20; Lit+12; Pad+19; Run+19; MG12; Tot+13; Che+17; Run+20] [Qin+20] Overall, all 33 included studies were judged to have a low risk of bias in the domain "index test" and 22 of 33 (67%) of the studies were judged to have a low risk of bias considering "flow and timing". However, only 1/4 of the studies (8/33) were judged to have a low risk of bias in the domain "patient selection", 1/3(10/33) in the domain "reference standard". Only 2 articles were judged to have a low risk of bias in all four domains.

C.6.2 AI methodology

Before 2017, authors mostly used machine learning-based methods for automatic segmentation of prostatic zones. After 2017, almost all publications were based on deep learning with convolutional neural networks (CNN) (72%, 24/33). Common architectures such as U-net [RFB15] have been extensively used, with modification and fine tuning of existing models, allowing either improved accuracy of classical networks or reduced memory and storage requirements.

Dice coefficient (DSC) and Hausdorff distance [TH15] were commonly used metrics. Almost all authors found inferior results for PZ than WG, CG or TZ segmentation, attributing this to the more complex shape and structure of PZ, especially within the anterior bundles. Eleven authors subsequently stratified their DSC results based on prostate height, with various methods:in three equal parts [Zab+19], in 25% apex, 50% mid gland and 25% base [Ald+20] in 30%, 40% and 30%, respectively [Jen+19]. Five authors did not provide any details on how they divided the volume.

These results as well as the remaining metrics are summarized in Table C.5.

C.7 Discussion

Our systematic review highlights the high prevalence of deficiencies in methodology in the literature on automatic segmentation of prostate gland on MRI.

Table 2 Overview of types of databases used with training, validation and test sets distribution

First author, year of publication	Inclusion criteria	Presence of PCa	Number of patients (total)	Training			Validation		Test		
				Total	Public data	In-house data	Cross-validation	Validation data	Total	Public data	In-house data
Cuocolo et al. [43]	✓	✓	204	79	79[(A)]	0	Fivefold	20	105	105[(A)]	0
Bardis et al. [42]	✓	✓	242	146	0	145	0	48	48	0	48
Lai et al. [41]	✓	✓	115	80	80[(A)]	0	Fivefold	20	15	15[(A)]	0
Nai et al. [18]	✓	✓	160	120	120[(A)]	0	0	20	20	20[(A)]	0
Sanford et al. [40]	✓	✓	1054	518 + 162[§]	0	680	0	130 + 42[§]	202	0	202
Aldoj et al. [39]	✓	✓	188	106	106[(A)]	0	Fourfold	35	47	20[(A)]	0
Zavala-Romero et al. [6]	✓	✓	550	198 or 297	297[(A)]	198	0	0	variable	33[(A)]	22
Lee et al. [38]	✓	✓	330	260 *(for WG)* or 162 *(for TZ)*	0	260 *(for WG)* or 162 *(for TZ)*	0	0	70 *(for WG)* or 50 *(for TZ)*	0	50
Liu et al. [37]	✓	✓	351	218	218[(A)]	0	0	45	92	45[(A)]	47
Qin et al. [36]	×	✓	240	162 + 45	45 [(B)]	162	0	0	33	15 [(B)]	18
Motamed et al. [35]	×	Unknown	681	291 *(source)* + variable *(target)*	0	406	0	97	145 *(source)* + 33 *(target)*	0	178
Zabihollahy et al. [13]	✓	✓	225	80	0	80	0	20	125	0	125
Padgett et al. [8]	✓	✓	61	Variable	0	Variable	0	0	Variable	0	1
Rundo et al. [15][1]	×	✓	80	Variable	Variable[(B)]	Variable	0	0	Variable	Variable[(B)]	Variable
Meyer et al. [34]	✓	✓	98	58	58[(A)]	0	Fourfold	20	20	20[(A)]	0
Liu et al. [16]	✓	✓	359	200	200[(A)]	0	Fivefold	50	110	63[(A)]	46
Rundo et al. [33][2]	×	✓	40 [†]	Variable [†]	[†(C)]	Variable	0	0	Variable	0	Variable
Hambarde et al. [32]	×	Unknown	52	42	0	42	0	0	10	0	10
Jensen et al. [31]	✓	✓	40	32	0	32	Fivefold	2	8	0	8
Khan et al. [17]	×	✓	80	35	35[(B)]	0	0	15	30	30[(B)]	0
Cheng et al. [30]	×*	✓	225	116 +/− [†]	8[(A)]	108	0	0	Variable	Variable[(A+C)]	27
Zhu et al. [12]	✓	✓	163	76	0	76	0	36	51	0	51
Mooij et al. [29]	0	Unknown	53	36	0	36	Fivefold	9	8	0	8
Can et al. [28]	0	Unknown	29	12	12[(B)]	0	0	7	10	10[(B)]	0
Clark et al. [14]	×*	✓	154	115	78[(C)]	37	0	0	38	12[(C)]	26
Chilali et al. [9]	✓	✓	55	30	30 *(Prostatlas)*	0	0	0	25	13[(C)]	12

Table 2 (continued)

First author, year of publication	Inclusion criteria	Presence of PCa	Number of patients (total)	Training			Validation		Test		
				Total	Public data	In-house data	Cross-validation	Validation data	Total	Public data	In-house data
Makni et al. [10]	✓	✓	31	? *(simulated images)*	0	0	0	0	31	0	31
Chi et al. [27]	✓	Unknown	8	4	0	4	0	0	4	0	4
Toth et al. [26]	✓	✓	40	Variable	0	Variable	0	0	Variable	0	Variable
Litjens et al. [7]	×	Unknown	48	48	0	47	0	0	1	0	1
Moschidis and Graham [25]	✓	×	22	Variable	0	Variable	0	0	Variable	0	Variable
Yin et al. [24]	×	✓	522 *(images)*	261 *(images)*	0	261 *(images)*	Fivefold	52 *(images)*	261 *(images)*	0	261 *(images)*
Makni et al. [23]	✓	✓	31	?	0	0	0	0	31	0	31

*Not specified for in-house data

[†] +/− 50 patients from PROMISE12 dataset used for pre-training of WG segmentation

[§] Pre-training data + data for transfer learning

[(A)] Public data used is PROSTATE-X

[(B)] Public data used is NCI-ISBI

[(C)] Public data used is PROMISE12

[1] Rundo et al., USE-Net: incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets [30]

[2] Rundo et al., CNN-based Prostate Zonal Segmentation on T2-weighted MR Images: A Cross-dataset Study[27]

PCa prostate cancer, *WG* whole gland, *TZ* transition zone

Table 3 Input MRI parameters (number of vendors, type of field, type of coil, input sequences)

First author, year of publication	Vendors				Type of coil		Field strenght		Data input		
	Number of vendors	Philips	GE	Siemens	ERC	SC	1.5T	3T	Mono-parametric	Sequence	Slice thickness*
Cuocolo et al. [43]	1	–	–	✓	–	✓	–	✓	✓	Axial T2W	3.6
Bardis et al. [42]	2	✓	–	✓	–	✓	–	✓	✓	Axial T2W	3.0
Lai et al. [41]	1	–	–	✓	–	✓	–	✓	×	Axial T2W + DWI + ADC	3.6
Nai et al. [18]	1	–	–	✓	–	✓	–	✓	×	Axial T2W + DWI + ADC	3.6
Sanford et al. [40]	3	✓	✓	✓	✓	✓	✓	✓	✓	Axial T2W	3.0
Aldoj et al. [39]	1	–	–	✓	–	✓	–	✓	✓	Axial T2W	3.6
Zavala-Romero et al. [6]	2	–	✓	✓	–	✓	–	✓	✓	3 planes T2W	3.6
Lee et al. [38]	1	–	–	✓	–	✓	–	✓	✓	Axial + sagittal T2W	3.0
Liu et al. [37]	1	–	–	✓	–	✓	–	✓	✓	Axial T2W	3.0–3.6
Qin et al. [36]	at least 2	✓	?	✓	✓	✓	✓	✓	×	Axial T2W + ADC	3.6
Motamed et al. [35]	2	✓	–	✓	?	?	?	?	✓	DWI	3.0
Zabihollahy et al. [13]	1	–	✓	–	–	✓	–	✓	×	Axial T2W + ADC	3.0–4.0
Padgett et al. [8]	2	–	✓	✓	?	?	–	✓	✓	Axial T2W	2.5
Rundo et al. [15][1]	2	✓	–	✓	–	✓	–	✓	✓	Axial T2W	1.25–4.0
Meyer et al. [34]	1	–	–	✓	–	✓	–	✓	✓	Axial T2W	3.0
Liu et al. [16]	1	–	–	✓	–	✓	–	✓	✓	Axial T2W	3.6
Rundo et al. [33][2]	2	✓	–	✓	–	✓	–	✓	✓	Axial T2W	1.25–3.0
Hambarde et al. [32]	1	✓	–	–	?	?	✓	–	✓	Axial T2W	5.0
Jensen et al. [31]	2	–	✓	✓	✓	✓	✓	✓	✓	Axial T2W	1.5–3.0
Khan et al. [17]	2	✓	–	✓	✓	✓	✓	✓	✓	Axial T2W	3.0–4.0
Cheng et al. [30]	multiple	?	?	✓	✓	✓	?	✓	✓	Axial T2W	3.0
Zhu et al. [12]	1	✓	–	–	–	✓	–	✓	×	Axial T2W + DWI	4.0
Mooij et al. [29]	?	?	?	?	?	?	?	?	✓	3D T2W	3.6
Can et al. [28]	2	✓	–	✓	✓	✓	✓	✓	✓	Axial T2W	3.0–4.0
Clark et al. [14]	multiple	✓	?	?	✓	✓	✓	✓	✓	DWI	?
Chilali et al. [9]	3	✓	✓	✓	✓	✓	✓	✓	✓	Axial T2W	3.0–4.0
Makni et al. [10]	1	✓	–	–	?	?	✓	–	×	Axial T2W + DWI + CE	1.25
Chi et al. [27]	1	–	–	✓	–	✓	–	✓	×	Axial T2W + ADC	3.3–3.75
Toth et al. [26]	?	?	?	?	✓	–	–	✓	✓	Axial T2W	3.0
Litjens et al. [7]	?	–	–	–	?	?	?	?	×	Axial T2W + ADC	4.0

Table 3 (continued)

First author, year of publication	Vendors				Type of coil		Field strenght		Data input		
	Number of vendors	Philips	GE	Siemens	ERC	SC	1.5T	3T	Mono-parametric	Sequence	Slice thickness*
Moschidis and Graham [25]	1	✓	–	–	–	✓	✓	–	✓	3D T2W	?
Yin et al. [24]	1	✓	–	–	✓	✓	–	✓	✓	Axial T2W	3.0
Makni et al. [23]	1	✓	–	–	–	✓	✓	–	×	Axial T2W + DWI + CE	2.5

ERC endorectal coil, *SC* Surface coil, *T2W* T2-weighted, *DWI* diffusion-weighted imaging, *ADC* apparent diffusion coefficient, CE contrast-enhanced

*Slice thickness in mm, when axial T2W slices used

?Not reported

[1] Rundo et al., USE-Net: incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets [30]

[2] Rundo et al., CNN-based Prostate Zonal Segmentation on T2-weighted MR Images: A Cross-dataset Study[27]

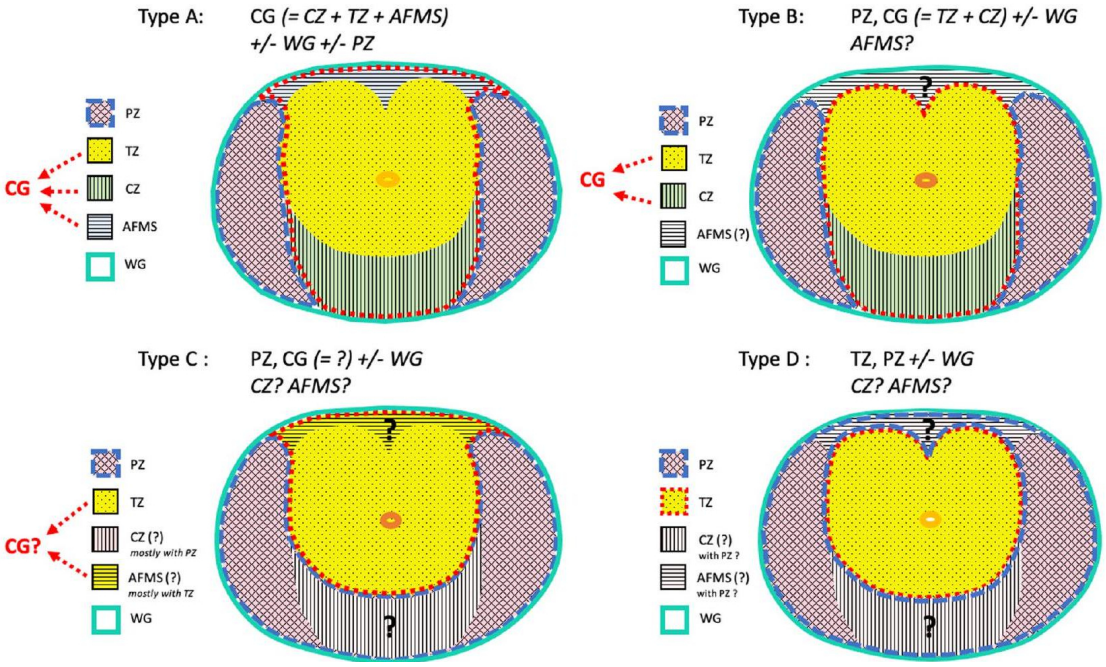


Fig. C.3.: Schematic of the four major types of protocol of zonal segmentation. Type A: articles for which "central gland" included CZ, TZ and AFMS. Type B: articles for which "central gland" included TZ and CZ. No details for AFMS. Type C: articles which did not provide details for AFMS, CZ or CG. CZ seemed to be mostly segmented PZ, while AFMS seemed to be mostly segmented with TZ, usually called "CG". Type D: articles which did not provide details for AFMS or CZ. CZ and AFMS seemed to be mostly segmented with PZ. CZ central zone, TZ transition zone, AFMS anterior fibro-muscular stroma, PZ peripheral zone, CG central gland

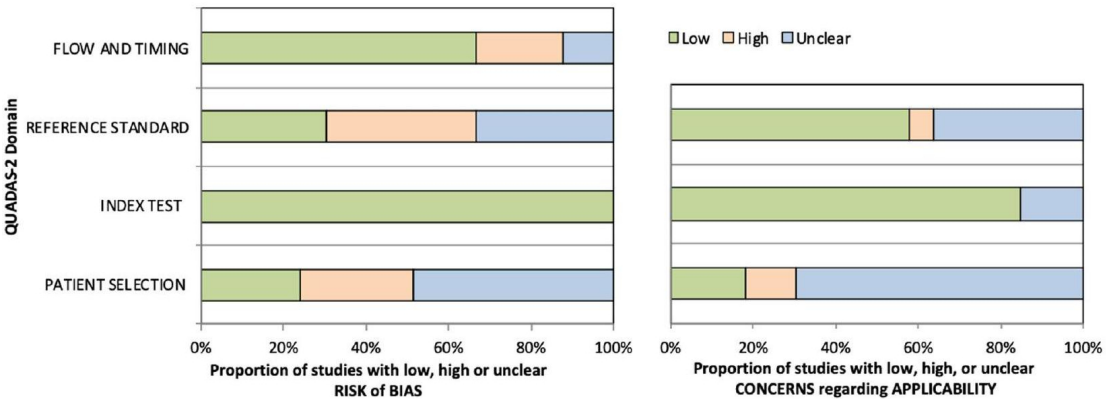


Fig. C.4.: Stacked bar charts showing results of quality assessment for risk of bias and applicability of included studies. QUADAS-2 scores for methodologic study quality are expressed as the percentage of studies that met each criterion. For each quality domain, the proportion of included studies that were determined to have low, high, or unclear risk of bias and/or concerns regarding applicability is displayed in green, orange, and blue, respectively. QUADAS-2: Quality Assessment of Diagnostic Accuracy Studies 2

First author, year of publication	Annotation		Annotators			
	Type	Tool	Qualification	Number	Type of reading	Experience*
Cuocolo et al. [Cuo+21a]	Software	itk-SNAP	(A)	4	Splitted[¶]	2 to 5
Bardis et al. [Bar+21]	Software	In house	(A)	12	Stratified	10
Lai et al. [Lai+21]	Manual	–	(A)	1	–	10
Nai et al. [Nai+20]	Software	MITK	Medical physicist	4	Stratified	2 to 10
Sanford et al. [San+20b]	Software	pseg	(A)	1	–	10
Aldoj et al. [Ald+20]	Manual	–	(A)	1	–[★]	"Expert"
Zavala-Romero et al. [Zav+20]	Manual	–	$(A)+(B)$	3	Stratified	10
Lee et al. [Lee+20]	Manual	-	(A)	2	?	4
Liu et al. [Liu+20]	Software	Osirix	$(C)+(A)$	More than 2	Stratified	10 and 19
Qin et al. [Qin+20]	Manual	–	$(A)^{\dagger}$	?	–	?
Motamed et al. [Mot+20]	Manual	–	(A)	2	?	4 and 6
Zabihollahy et al. [Zab+19]	Software	itk-SNAP	(A)	4	Splitted	5 and 14
Padgett et al. [Pad+19]	Manual	–	(B)	2	Blinded [§]	10 and 26
Rundo et al. [Run+19][1]	Manual	-	?	Multiple	?	"Expert"
Meyer et al. [Mey+19]	Software	3DSLICER	Medical student + urologist + (A)	4	Stratified	"Expert"
Liu et al. [Liu+19]	Software	Osirix	(C) +(A)	7	Stratified	10 – 15
Rundo et al.[Run+20][2]	?	?	(A)	Multiple	?	?
Hambarde et al. [Ham+19]	?	?	(A)	Multiple	?	?
Jensen et al. [Jen+19]	Software	?	(A)	1	–	"Expert"
Khan et al. [Kha+19]	Manual	-	(A)	3	?	?
Cheng et al. [Che+17]	Software	pseg	(A)	1	-	10
Zhu et al. [Zhu+19]	Manual	–	?	2	?	More than 5
Mooij et al. [MBH18]	Manual	–	?	?	?	?
Can et al. [Can+18]	Manual	-	(A)	3	?	?
Clark et al. [Cla+17]	Manual	–	(A)	1	?	?
Chilali et al. [Chi+16]	Manual	–	(A)	1	–	15
Makni et al. [MBC14]	Manual	–	(A)	3	Blinded	"Expert"
Chi et al. [Chi+14]	Manual	-	(A)	1	–	5
Toth et al. [Tot+13]	Software	3DSLICER	(A)	1	–	"Expert"
Litjens et al. [Lit+12]	Manual	–	(A)	3	?	?
Moschidis and Graham [MG12]	Manual	–	(A)	2	?	?
Yin et al. [Yin+12]	Manual	–	"Radiologist-trained operators"	2	Splitted	?
Makni et al. [Mak+11]	Manual	_	(A)	3	Blinded	4.6 and 9

(A) Radiologist
(B) Radiation oncologist
(C) Research fellow
? Data not reported
* Experience of reader(s), in years
† Unclear for PROMM, in-house data
¶ Splitted but consensus per binome resident-senior
★ Only one reading for ground truth segmentation but evaluation of intra and inter observator variability on some masks
§ Measure of inter- observator variability for 10 masks
Splitted: database is divided such as each set of images is read only once, resulting in an equivalent of single reader
(Stratified: first reading (mostly by a less experienced reader) subsequently corrected by a more experience reader
Blinded: blinded reading by at least 2 readers
[1] Rundo et al., USE-Net: incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets [Run+19]
[2] Rundo et al., CNN-based Prostate Zonal Segmentation on T2-weighted MR Images: A Cross-dataset Study [Run+20]

Tab. C.4.: Type of ground truth segmentation

Since 2011, 33 studies proposed new or fine-tuned existing approaches for automatic prostatic zonal segmentation. Many studies are hampered by issues with limitation of the dataset used in the model, methodological mistakes, poor reproducibility, and biases in study design. Most studies focused on achieving the best accuracy for their algorithms, sometimes putting aside validity and applicability in clinical practice. Indeed, only two articles presented with an overall low risk of bias.

The common limitations concerned datasets used for the model development, definition of the ground truth for evaluation of the model and strategies used for model evaluation.

Regarding the datasets used, some are private, and some are public open source. For private databases, advanced technical characteristics of images (e.g., imaging sequence, field of view, noise) used and patient's inclusion and exclusion criteria were poorly or not described. Most databases lacked representability of patients' variability as prostate volume, prostate tissue heterogeneity, prostatic pathology as PCa or benign hypertropia. Open-source prostate MRI databases also have several limitations such as selection bias, limited annotations, low-resolution images, unclear terminology, lack of demographic statistics and of precise histologic data.

This can have a direct impact on the generalizability of the model developed. Indeed, it has been shown for example that prostate morphological differences contribute to segmentation variability: Montagne et al. [Mon+21], showed that the smaller the prostate volume was, the higher the variability was; several authors [Nai+20; Ald+20; Cuo+21a] found poorer performance of their model applied on special cases such as history of trans-urethral-resection of prostate (TURP), while most databases lacked representativity of patients variability.

Even though it is tedious and time-consuming, reference segmentation should require at least two trained readers because inter- and intra-rater variability can be significant. Quality of images (slice thickness, partial volume artifacts), apex or base location [Mon+21; Bec+19] or prostate morphological differences [Mon+21] have been shown to decrease accuracy of segmentation. Meyer et al. [Mey+19] showed that training on segmentation obtained by a single reader introduced bias into the training data. Indeed, performance was higher when obtained from the expert who created the training data in comparison with evaluation against other expert segmentation. Aldoj et al. [Ald+20] emphasized the need for finely annotated sets as they improved overall performances of their algorithms, showing the greater importance of well annotated databases compared to large and coarsely annotated databases.

Quality of the resulting auto segmentation is evaluated against the corresponding reference segmentation, so called the ground truth. The main approach is manual delineation of the prostate zones performed by human experts. We found a great heterogeneity on the segmentation protocols and terminology used. Eighteen different types of prostate delineation were found; each anatomical zone was segmented directly or obtained by

First author, year of publication	Type	DSC results †				Stratification by gland height	Pre-processing details	Post-processing details
		WG	TZ	PZ	CG			
Cuocolo et al. [Cuo+21a]	CNN	0.9063*	-	0.7142*	0.8692*	x	✓	x
Bardis et al. [Bar+21]	CNN	0.94	0.91	0.774	–	x	✓	x
Lai et al. [Lai+21]	CNN	–	0.93	0.7004	-	x	✓	x
Nai et al. [Nai+20]	CNN	0.89*	–	0.712*	0.856*	✓	✓	x
Sanford et al. [San+20b]	CNN	0.915	0.89	–	-	x	✓	x
Aldoj et al. [Ald+20]	CNN	0.921*	–	0.781*	0.895*	✓	√	x
Zavala-Romero et al. [Zav+20]	CNN	0.825[a] 0.892[b]	-	0.788[a] 0.811[b]	-	x	✓	✓
Lee et al. [Lee+20]	CNN	0.8712	0.7648	–	–	x	✓	x
Liu et al. [Liu+20]	CNN	–	0.89*[c] 0.87*[d]	0.80*[c] 0.79*[d]	-	✓	✓	x
Qin et al. [Qin+20]	CNN	-	–	0.806	0.901	x	✓	✓
Motamed et al. [Mot+20]	CNN	0.89[e] 0.85[f]	0.86[e] 0.84[f]	-	–	x	x	✓
Zabihollahy et al. [Zab+19]	CNN	0.9533[g] 0.9209[h]	–	0.8678[g] 0.861[h]	0.9375[g] 0.8989[h]	✓	✓	✓
Padgett et al. [Pad+19]	Atlas	0.83*	0.75*	0.59*	–	✓	x	x
Rundo et al.[Run+19][1]	CNN	–	-	0.919[i] 0.831[j] 0.801[k]	0.871[i] 0.886[j] 0.937[k]	x	✓	✓
Meyer et al. [Mey+19]	CNN	–	0.876	0.798	–	x	✓	✓
Liu et al. [Liu+19]	CNN	-	0.86[c] 0.79[d]	0.74[c] 0.74[d]	-	✓	✓	x
Rundo et al.[Run+20][2]	CNN	-	–	0.91* (with pre-training)	0.85* (with pre-training)	x	✓	✓
Hambarde et al. [Ham+19]	CNN	–	–	0.8733	–	x	✓	x
Jensen et al. [Jen+19]	CNN	-	–	0.692	0.794	✓	✓	✓
Khan et al. [Kha+19]	CNN	–	–	0.703*	0.88*	×	x	x
Cheng et al. [Che+17]	CNN	0.9235*	–	–	0.9006*	✓	✓	✓
Zhu et al. [Zhu+19]	CNN	0.927	–	0.793	–	✓	✓	×
Mooij et al. [MBH18]	CNN	–	0.85*	0.6*	–	x	✓	x
Can et al. [Can+18]	CNN	–	–	0.722*	0.89*	×	x	x
Clark et al. [Cla+17]	CNN	0.886[c] 0.862[d]	0.847[c]	–	–	×	✓	x
Chilali et al. [Chi+16]	C means + Atlas	0.9478	0.7023	0.62	–	x	✓	x
Makni et al. [MBC14]	C means	–	0.88	0.78	–	×	✓	x
Chi et al. [Chi+14]	Gaussian model	0.8	–	0.53	0.83	×	x	x
Toth et al. [Tot+13]	Active appearance model	0.81	-	0.68[l] 0.60[m]	0.79[l] 0.72[m]	✓	✓	x
Litjens et al. [Lit+12]	Atlas	–	–	0.75	0.8	x	x	x
Moschidis and Graham [MG12]	Random Forrest + Graph Cuts	-	-	–	–	x	✓	x
Yin et al. [Yin+12]	Graph Cuts	-	–	–	0.81	×	x	x
Makni et al. [Mak+11]	C means	-	-	0.76[l]	0.87[l]	¶	x	x

CNN convolutional neural network

† Dice similarity coefficient (DSC) for whole gland (WG), transition zone (TZ), peripheral zone (PZ) or central gland (CG) (means)

* Best results if several models were tested

¶ no Dice Similary Coefficien (DSC) provided

[a,b] Trained on combined datasets and, respectively, tested on G^a or Siemens [b]

[c,d] Respectively for testing on internal [c] or external [d] data

[e,f] Respectively for source [e] or target [f] with 115 patients for training (best results)

[g,h] Respectively for T_2-weighted [g] and apparent diffusion coefficient (ADC) map [h]

[i,j] Trained on combined datasets and, respectively, tested on dataset #1[i], #2[j] or #3[k]

[l, m] Using pre-segmented whole gland (WG)', or with whole process [m]

[1] Rundo et al., USE-Net: incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets [Run+19]

[2] Rundo et al., CNN-based Prostate Zonal Segmentation on T2-weighted MR Images: A Cross-dataset Study [Run+20]

Tab. C.5.: Overview of segmentation methods with performance based on DSC. Number of articles reporting stratification by gland height, and reporting pre- or post-processing steps

subtraction from one region to another (resulting in CZ, AFMS and PZ, which can be obtained either by delineation or by subtraction of WG and TZ). Terminology used was extremely variable from one study to another and did not always respect the one used and referenced in the PIRADS [Tur+19; Wei+16] (for example, use of "central gland" instead of CZ or TZ).

Number of readers, level of expertise, inter- and intravariability evaluation were mostly absent, limiting the generalizability of the developed models due to interobserver variability. Only 2/33 studies [MBC14; Mak+11] used blinded reading for ground truth. Nonetheless, prostate segmentation is a very challenging task. The prostate gland usually has fuzzy boundaries. Pixel intensities are heterogeneous both inside and outside the prostate, and contrasts and pixel intensities are very similar for prostate and non-prostate regions. The manual delineation of the prostate zones is therefore limited by the subjective interpretation of the organ boundaries. Becker et al. [Bec+19] found in a

multi-reader study a higher variability at the extreme part of the gland (apex and base) and for the TZ delineation. Similar results were found by Padgett et al. [Pad+19] who found a difference of DSC from 0.88 to 0.81 for WG and TZ. Meyer et al. [Mey+19] showed that training on segmentation obtained by a single reader introduced bias into the training data.

Strategies used for model evaluation were limited by the lack of external validation Only 7 studies [Zav+20; Chi+16; Cla+17; Che+17; Run+20; Qin+20; Liu+20] used both private and public data to evaluate their model. The absence of an external testing dataset is a critical limitation to the clinical applicability of the developed models. Data augmentation and transfer learning were also used to help addressing this issue [Zav+20; Cla+17; Run+19; Liu+19; MBH18; Jen+19; Run+20; Mot+20; Qin+20; Liu+20; Lee+20; Ald+20; San+20b; Lai+21; Cuo+21a; Mey+21]. It is important to note that some bias cannot be balanced-out by increasing the sample size by data augmentation or repetition of training. For example, data augmentation of a dataset constituted without prostate cancer patients cannot decrease risk of bias induced by the more homogeneous contours it provides.

Even without data augmentation, MRI images contains wide heterogeneity and most of the times pre-processing steps involving intensity normalization or noise reduction to remove confounding features and improve image quality are necessary [GSJ20]. Some authors [Zav+20; Zab+19]15, 31, 33, 35, 51] also reported post-processing. Not reporting some of the pre- or post-processing steps can affect reproducibility and sufficient detail enables readers to determine the quality and generalizability of the work. While several checklists can be used such as those from Enhancing the Quality and Transparency Of health Research (EQUATOR) Network guidelines [Equ], the use of the recently published Checklist for Artificial Intelligence in Medical Imaging [MMK20] would be helpful to lower risk of bias of ongoing work.

In the future, there is a need for well-sampled databases including large number of representative cases for the anatomical variability of the prostate gland and technical specificities (2D T2 versus 3D T2, slice thickness, FOV, vendors) to account for the anatomical, disease related, acquisition related variabilities, with a multi-readers segmentations and a well-defined delineation guideline of the prostate (as it is already done for example in organs at risk for radiotherapy planning [Vrt+20]).

Constitution of quality database should be based on latest PI-RADS recommendations, by associating quality criteria such as the consensual quality requirements ESUR/ESUI [de+20] or Prostate Imaging Quality (PI-QUAL) [Gig+21] score to guarantee essential image quality for zonal segmentation and tumor detection.

The main limitation of this review is the absence of details of technical information used; each study making its own contribution for networks with countless hyperparameters,

sometimes without enough details to be gathered. This precluded us from comparing models' accuracy without bias.

Some other relevant papers also could be missing because of incongruences between search terms, article keywords, or indexing in the databases, such as for conference proceedings papers. In particular, databases such as ArXiv were not searched as it also provides access to preprints, without peer review.

C.8 Conclusion

This review systematically synthesizes published automatic prostate zonal segmentation methods using MRI. We found that no papers in the literature currently have both sufficiently documented datasets selection and segmentation criteria and enough external validation.

This underlines the critical need for higher quality datasets, a documented reproducible method and terminology for zonal segmentation and sufficient external dataset to develop the best quality methods free from biases: an essential step for future development of automatic detection of prostate cancer.

Acronyms

ADC: Apparent Diffusion Coefficient

AFMS: Anterior Fibromuscular Stroma

AHD: Average Hausdorff Distance

AUC: Area Under the Curve

BPEF: Biproximate Ellipsoid Formula

bpMRI: Biparametric MRI

CAD: Computer-Aided Diagnosis/Detection

CADe: Computer-Aided Detection

CADx: Computer-Aided Diagnosis

CI: Confidence Interval

CNN: Convolutional Neural Networks

CZ: Central Zone

DCEI: Dynamic Contrast-Enhanced Imaging

DSC: Dice Score

DRE: Digital Rectal Exam

DWI: Diffusion-weighted Imaging

GGG: Gleason Grade Group

GDPR: General Data Protection Regulation

HD: Hausdorff Distance

ICC: Intraclass Correlation Coefficient

IQR: Interquartile Range

LMSD: Local Mean Squared Distance

MPM: Manual Planimetry Method

MRI: Magnetic Resonance Imaging

MV: Majority Voting

mpMRI: multiparametric MRI

PCa: Prostate Cancer

PSA: Prostate-Specific Antigen

PSAd: PSA density

PV: Prostate Volume

PZ: Peripheral Zone

(r)STD: (relative) Standard Deviation

T2WI: T2-weighted Imaging

TEF: Traditional Ellipsoid Formula

TRUS: Transrectal Ultrasonography

TZ: Transition Zone

WG: Whole Gland

Bibliography

[Aba+15] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Arxiv eprints. Software available from https://www.tensorflow.org/. 2015 (cit. on p. 85).

[ACK19] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. "Weakly Supervised Learning of Instance Segmentation With Inter-Pixel Relations". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 2209–2218 (cit. on p. 104).

[Ada+22] Lisa C. Adams, Marcus R. Makowski, Günther Engel, Maximilian Rattunde, Felix Busch, Patrick Asbach, Stefan M. Niehues, Shankeeth Vinayahalingam, Bram van Ginneken, Geert Litjens, and Keno K. Bressem. "Prostate158 - An Expert-Annotated 3T MRI Dataset and Algorithm for Prostate Cancer Detection". In: *Computers in Biology and Medicine* 148 (2022-09), p. 105817 (cit. on p. 107).

[ADV22] Muhammad Asad, Reuben Dorent, and Tom Vercauteren. "FastGeodis: Fast Generalised Geodesic Distance Transform". In: *Journal of Open Source Software* 7.79 (2022), p. 4532 (cit. on p. 112).

[Ahm+17] Hashim U. Ahmed, Ahmed El-Shater Bosaily, Louise C. Brown, Rhian Gabe, Richard Kaplan, Mahesh K. Parmar, Yolanda Collaco-Moraes, Katie Ward, Richard G. Hindley, Alex Freeman, Alex P. Kirkham, Robert Oldroyd, Chris Parker, Mark Emberton, and PROMIS study group. "Diagnostic Accuracy of Multi-Parametric MRI and TRUS Biopsy in Prostate Cancer (PROMIS): A Paired Validating Confirmatory Study". In: *Lancet* 389.10071 (2017-02), pp. 815–822 (cit. on p. 8).

[AL12] Andrew Asman and Bennett Landman. "Formulating Spatially Varying Performance in the Statistical Fusion Framework". In: *Medical Imaging, IEEE Transactions on* 31 (2012-06), pp. 1326–1336 (cit. on pp. 48, 49).

[AL13] Andrew Asman and Bennett Landman. "Non-local statistical label fusion for multi-atlas segmentation". In: *Medical Image Analysis* 17.2 (2013), pp. 194–208 (cit. on pp. 48, 49).

[Ald+20] N. Aldoj, F. Biavati, F. Michallek, S. Stober, and M. Dewey. "Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like U-net". In: *Sci. Rep.* 10 (2020) (cit. on pp. 21, 76, 99, 154–156, 162–165).

[Alg+18] Ahmad Algohary, Satish Viswanath, Rakesh Shiradkar, Soumya Ghose, Shivani Pahwa, Daniel Moses, Ivan Jambor, Ronald Shnier, Maret Böhm, Anne-Maree Haynes, Phillip Brenner, Warick Delprado, James Thompson, Marley Pulbrock, Andrei S Purysko, Sadhna Verma, Lee Ponsky, Phillip Stricker, and Anant Madabhushi. "Radiomic Features on MRI Enable Risk Categorization of Prostate Cancer Patients on Active Surveillance: Preliminary Findings". In: *Journal of Magnetic Resonance Imaging* 48.3 (2018), pp. 818–828 (cit. on p. 126).

[Alj+09] P. Aljabar, R.A. Heckemann, A. Hammers, J.V. Hajnal, and D. Rueckert. "Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy". In: *NeuroImage* 46.3 (2009), pp. 726–738 (cit. on p. 48).

[Alq+20] Saeed Alqahtani, Cheng Wei, Yilong Zhang, Magdalena Szewczyk-Bieda, Jennifer Wilson, Zhihong Huang, and Ghulam Nabi. "Prediction of Prostate Cancer Gleason Score Upgrading from Biopsy to Radical Prostatectomy Using Pre-Biopsy Multiparametric MRI PIRADS Scoring System". In: *Scientific Reports* 10.1 (2020-05), p. 7722 (cit. on p. 127).

[Arm+18] Samuel G. Armato, Henkjan Huisman, Karen Drukker, Lubomir Hadjiiski, Justin S. Kirby, Nicholas Petrick, George Redmond, Maryellen L. Giger, Kenny Cha, Artem Mamonov, Jayashree Kalpathy-Cramer, and Keyvan Farahani. "PROSTATEx Challenges for Computerized Classification of Prostate Lesions from Multiparametric Magnetic Resonance Images". In: *J. Med. Imag.* 5.04 (2018-11), p. 1 (cit. on pp. 15, 76, 103, 154).

[AS+23a] Gillian Adair Smith, Alex Dunlop, Sophie E. Alexander, Helen Barnes, Francis Casey, Joan Chick, Ranga Gunapala, Trina Herbert, Rebekah Lawes, Sarah A. Mason, Adam Mitchell, Jonathan Mohajer, Julia Murray, Simeon Nill, Priyanka Patel, Angela Pathmanathan, Kobika Sritharan, Nora Sundahl, Alison C. Tree, Rosalyne Westley, Bethany Williams, and Helen A. McNair. "Evaluation of Therapeutic Radiographer Contouring for Magnetic Resonance Image Guided Online Adaptive Prostate Radiotherapy". In: *Radiotherapy and Oncology* 180 (2023-03), p. 109457 (cit. on p. 143).

[AS+23b] Gillian Adair Smith, Alex Dunlop, Sophie E. Alexander, Helen Barnes, Francis Casey, Joan Chick, Ranga Gunapala, Trina Herbert, Rebekah Lawes, Sarah A. Mason, Adam Mitchell, Jonathan Mohajer, Julia Murray, Simeon Nill, Priyanka Patel, Angela Pathmanathan, Kobika Sritharan, Nora Sundahl, Rosalyne Westley, Alison C. Tree, and Helen A. McNair. "Interobserver Variation of Clinical Oncologists Compared to Therapeutic Radiographers (RTT) Prostate Contours on T2 Weighted MRI". In: *Technical Innovations & Patient Support in Radiation Oncology* 25 (2023-03), p. 100200 (cit. on pp. 134, 143).

[Aud+20] Benoît Audelan, Dimitri Hamzaoui, Sarah Montagne, Raphaële Renard-Penna, and Hervé Delingette. "Robust Fusion of Probability Maps". In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020*. Ed. by Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz. Cham: Springer International Publishing, 2020, pp. 259–268 (cit. on pp. 17, 48, 68).

[Aud+22] Benoît Audelan, Dimitri Hamzaoui, Sarah Montagne, Raphaële Renard-Penna, and Hervé Delingette. "Robust Bayesian Fusion of Continuous Segmentation Maps". In: *Medical Image Analysis* 78 (2022), p. 102398 (cit. on p. 17).

[Bar+12] Jelle O. Barentsz, Jonathan Richenberg, Richard Clements, Peter Choyke, Sadhna Verma, Geert Villeirs, Olivier Rouviere, Vibeke Logager, and Jurgen J. Fütterer. "ESUR Prostate MR Guidelines 2012". In: *European Radiology* 22.4 (2012), pp. 746–757 (cit. on pp. 8, 106).

[Bar+21] M. Bardis, R. Houshyar, C. Chantaduly, K. Tran-Harding, A. Ushinsky, C. Chahine, M. Rupasinghe, D. Chow, and P. Chang. "Segmentation of the Prostate Transition Zone and Peripheral Zone on MR Images with Deep Learning". In: *Radiol. Imaging Cance* 3.3 (2021) (cit. on pp. 76, 84, 97, 154, 162, 164).

[Bat+20] Tharakeswara K. Bathala, Aradhana M. Venkatesan, Jingfei Ma, Priyadarshini Bhosale, Wei Wei, Rajat J. Kudchadker, Jihong Wang, Mitchell S. Anscher, Chad Tang, Teresa L. Bruno, Steven J. Frank, and Janio Szklaruk. "Quality comparison between three-dimensional T2-weighted SPACE and two-dimensional T2-weighted turbo spin echo magnetic resonance images for the brachytherapy planning evaluation of prostate and periprostatic anatomy". In: *Brachytherapy* 19.4 (2020), pp. 484–490 (cit. on p. 76).

[Bec+19] Anton S. Becker, Krishna Chaitanya, Khoschy Schawkat, Urs J. Muehlematter, Andreas M. Hötker, Ender Konukoglu, and Olivio F. Donati. "Variability of Manual Segmentation of the Prostate in Axial T2-weighted MRI: A Multi-Reader Study". In: *Eur J Radiol* 121 (2019-12), p. 108716 (cit. on pp. 12, 21, 38, 39, 42, 61, 74, 99, 143, 163, 164).

[Ben+92] Mitchell C. Benson, Ihn Seong Whang, Allan Pantuck, Kenneth Ring, Steven A. Kaplan, Carl A. Olsson, and William H. Cooner. "Prostate Specific Antigen Density: A Means of Distinguishing Benign Prostatic Hypertrophy and Prostate Cancer". In: *The Journal of Urology* 147.3, Part 2 (1992-03), pp. 815–816 (cit. on pp. 20, 125, 150).

[Bez+18] Adam Bezinque, Andrew Moriarity, Crystal Farrell, Henry Peabody, Sabrina L. Noyes, and Brian R. Lane. "Determination of Prostate Volume: A Comparison of Contemporary Methods". In: *Academic Radiology* 25.12 (2018-12), pp. 1582–1587 (cit. on pp. 21, 23, 41).

[Bha+22] Indrani Bhattacharya, Yash S. Khandwala, Sulaiman Vesal, Wei Shao, Qianye Yang, Simon J.C. Soerensen, Richard E. Fan, Pejman Ghanouni, Christian A. Kunder, James D. Brooks, Yipeng Hu, Mirabela Rusu, and Geoffrey A. Sonn. "A Review of Artificial Intelligence in Prostate Cancer Detection on Imaging". In: *Therapeutic Advances in Urology* 14 (2022-01), p. 17562872221128791 (cit. on p. 12).

[Bju+20] Marc A. Bjurlin, Peter R. Carroll, Scott Eggener, Pat F. Fulgham, Daniel J. Margolis, Peter A. Pinto, Andrew B. Rosenkrantz, Jonathan N. Rubenstein, Daniel B. Rukstalis, Samir S. Taneja, and Baris Turkbey. "Update of the Standard Operating Procedure on the Use of Multiparametric Magnetic Resonance Imaging for the Diagnosis, Staging and Management of Prostate Cancer". In: *J. Urol.* 203.4 (2020), pp. 706–712 (cit. on pp. 74, 125).

[Boo+11] Alison Booth, Mike Clarke, Davina Ghersi, David Moher, Mark Petticrew, and Lesley Stewart. "An International Registry of Systematic-Review Protocols". In: *The Lancet* 377.9760 (2011-01), pp. 108–109 (cit. on p. 151).

[Bos+22] Joeran S. Bosma, Anindo Saha, Matin Hosseinzadeh, Ilse Slootweg, Maarten de Rooij, and Henkjan Huisman. *Annotation-Efficient Cancer Detection with Report-Guided Lesion Annotation for Deep Learning-Based Prostate Cancer Detection in bpMRI*. Arxiv eprints. 2022-02 (cit. on pp. 104, 118).

[Bra+21] Valentina Brancato, Marco Aiello, Luca Basso, Serena Monti, Luigi Palumbo, Giuseppe Di Costanzo, Marco Salvatore, Alfonso Ragozzino, and Carlo Cavaliere. "Evaluation of a Multiparametric MRI Radiomic-Based Approach for Stratification of Equivocal PI-RADS 3 and Upgraded PI-RADS 4 Prostatic Lesions". In: *Scientific Reports* 11.1 (2021-01), p. 643 (cit. on p. 125).

[Bri+23] Nina Brillat-Savarin, Carine Wu, Laurène Aupin, Camille Thoumin, Dimitri Hamzaoui, and Raphaële Renard-Penna. "3.0 T Prostate MRI: Visual Assessment of 2D and 3D T2-weighted Imaging Sequences Using PI-QUAL Score". In: *European Journal of Radiology* 166 (2023-09), p. 110974 (cit. on p. 17).

[Bul+12] Julie C. Bulman, Robert Toth, Amish D. Patel, B. Nicolas Bloch, Colm J. McMahon, Long Ngo, Anant Madabhushi, and Neil M. Rofsky. "Automated Computer-Derived Prostate Volumes from MR Imaging Data: Comparison with Radiologist-Derived MR Imaging and Pathologic Specimen Volumes". In: *Radiology* 262.1 (2012-01), pp. 144–151 (cit. on pp. 20, 23, 40, 41).

[CAAW12] Olivier Commowick, Alireza Akhondi-Asl, and Simon K. Warfield. "Estimating A Reference Standard Segmentation with Spatially Varying Performance Parameters: Local MAP STAPLE". In: *IEEE Transactions on Medical Imaging* 31.8 (2012-08), pp. 1593–1606 (cit. on pp. 48, 49, 54).

[Cac+23] Giovanni E. Cacciamani, Daniel I. Sanford, Timothy N. Chu, Masatomo Kaneko, Andre L. De Castro Abreu, Vinay Duddalwar, and Inderbir S. Gill. "Is Artificial Intelligence Replacing Our Radiology Stars? Not Yet!" In: *European Urology Open Science* 48 (2023-02), pp. 14–16 (cit. on p. 9).

[Can+18] Yigit B. Can, Krishna Chaitanya, Basil Mustafa, Lisa M. Koch, Ender Konukoglu, and Christian F. Baumgartner. "Learning to Segment Medical Images with Scribble-Supervision Alone". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Ed. by Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R.S. Tavares, Andrew Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 236–244 (cit. on pp. 153, 162, 164).

[Cao+19] Ruiming Cao, Amirhossein Mohammadian Bajgiran, Sohrab Afshari Mirak, Sepideh Shakeri, Xinran Zhong, Dieter Enzmann, Steven Raman, and Kyunghyun Sung. "Joint Prostate Cancer Detection and Gleason Score Prediction in Mp-MRI via FocalNet". In: *IEEE Transactions on Medical Imaging* 38.11 (2019-11), pp. 2496–2506 (cit. on p. 127).

[CCH06] William Crum, Oscar Camara, and Derek Hill. "Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis". In: *IEEE transactions on medical imaging* 25 (2006-12), pp. 1451–61 (cit. on p. 67).

[Che+17] Ruida Cheng, Holger Roth, Nathan Lay, Le Lu, Baris Turkbey, William Gandler, Evan McCreedy, Tom Pohida, Peter Pinto, Peter Choyke, Matthew Mcauliffe, and Ronald Summers. “Automatic magnetic resonance prostate segmentation by deep learning with holistically nested networks”. In: *J. Med. Imaging* 4.4 (2017-08), p. 041302 (cit. on pp. 75, 153, 154, 156, 162, 164, 165).

[Che+22] Tong Chen, Zhiyuan Zhang, Shuangxiu Tan, Yueyue Zhang, Chaogang Wei, Shan Wang, Wenlu Zhao, Xusheng Qian, Zhiyong Zhou, Junkang Shen, Yakang Dai, and Jisu Hu. “MRI Based Radiomics Compared With the PI-RADS V2.1 in the Prediction of Clinically Significant Prostate Cancer: Biparametric vs Multiparametric MRI”. In: *Frontiers in Oncology* 11 (2022) (cit. on p. 125).

[Chi+14] Y. Chi, H. Ho, Y. M. Law, Q. Tian, H. J. Chen, K. J. Tay, and J. Liu. “A Compact Method for Prostate Zonal Segmentation on Multiparametric MRIs”. In: *Medical Imaging 2014: Image-Guided Procedures, Robotic Interventions, and Modeling*. Vol. 9036. SPIE, 2014-03, pp. 163–171 (cit. on pp. 153, 155, 162, 164).

[Chi+16] O. Chilali, P. Puech, S. Lakroum, M. Diaf, S. Mordon, and N. Betrouni. “Gland and Zonal Segmentation of Prostate on T2W MR Images”. In: *Journal of Digital Imaging* 29.6 (2016-12), pp. 730–736 (cit. on pp. 150, 153, 154, 162, 164, 165).

[Cho+15] F. Chollet et al. *Keras*. https://github.com/fchollet/keras. 2015 (cit. on p. 85).

[Chr+20] Charlotte Christophe, Sarah Montagne, Stéphanie Bourrelier, Morgan Roupret, Eric Barret, François Rozet, Eva Comperat, Jean François Coté, Olivier Lucidarme, Olivier Cussenot, Benjamin Granger, and Raphaële Renard-Penna. “Prostate Cancer Local Staging Using Biparametric MRI: Assessment and Comparison with Multiparametric MRI”. In: *European Journal of Radiology* 132 (2020-11), p. 109350 (cit. on p. 125).

[Cla+13] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, and F. Prior. “The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository”. In: *J. Digit. Imaging* 26.6 (2013), pp. 1045–1057 (cit. on pp. 76, 78).

[Cla+17] Tyler Clark, Junjie Zhang, Sameer Baig, Alexander Wong, Masoom A. Haider, and Farzad Khalvati. “Fully Automated Segmentation of Prostate Whole Gland and Transition Zone in Diffusion-Weighted MRI Using Convolutional Neural Networks”. In: *Journal of Medical Imaging* 4.4 (2017-10), p. 041307 (cit. on pp. 150, 153, 154, 162, 164, 165).

[Col+22] Alexander P. Cole, Bjoern J. Langbein, Francesco Giganti, Fiona M. Fennessy, Clare M. Tempany, and Mark Emberton. “Is Perfect the Enemy of Good? Weighing the Evidence for Biparametric MRI in Prostate Cancer”. In: *BJR* 95.1131 (2022-03), p. 20210840 (cit. on p. 125).

[Com+18] Olivier Commowick et al. “Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure”. In: *Scientific Reports* 8.1 (2018-12), p. 13650 (cit. on p. 61).

[Cou16a] Council of European Union. *Recital 71*. 2016 (cit. on p. 127).

[Cou16b] Council of European Union. *REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. 2016 (cit. on p. 13).

[Cox+12] Brett W. Cox, Daniel E. Spratt, Michael Lovelock, Mark H. Bilsky, Eric Lis, Samuel Ryu, Jason Sheehan, Peter C. Gerszten, Eric Chang, Iris Gibbs, Scott Soltys, Arjun Sahgal, Joe Deasy, John Flickinger, Mubina Quader, Stefan Mindea, and Yoshiya Yamada. "International Spine Radiosurgery Consortium Consensus Guidelines for Target Volume Definition in Spinal Stereotactic Radiosurgery". In: *Int. J. Radiat. Oncol. Biol. Phys.* 83.5 (2012), pp. 597–605 (cit. on p. 79).

[CSB08] Antonio Criminisi, Toby Sharp, and Andrew Blake. "GeoS: Geodesic Image Segmentation". In: *Computer Vision – ECCV 2008*. Ed. by David Forsyth, Philip Torr, and Andrew Zisserman. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 99–112 (cit. on p. 112).

[Cuo+19] Renato Cuocolo, Maria Brunella Cipullo, Arnaldo Stanzione, Lorenzo Ugga, Valeria Romeo, Leonardo Radice, Arturo Brunetti, and Massimo Imbriaco. "Machine Learning Applications in Prostate Cancer Magnetic Resonance Imaging". In: *European Radiology Experimental* 3.1 (2019-12), p. 35 (cit. on p. 9).

[Cuo+21a] Renato Cuocolo, Albert Comelli, Alessandro Stefano, Viviana Benfante, Navdeep Dahiya, Arnaldo Stanzione, Anna Castaldo, Davide Raffaele De Lucia, Anthony Yezzi, and Massimo Imbriaco. "Deep Learning Whole-Gland and Zonal Prostate Segmentation on a Public MRI Dataset". In: *Journal of Magnetic Resonance Imaging* 54.2 (2021), pp. 452–459 (cit. on pp. 76, 79, 97, 154, 162–165).

[Cuo+21b] Renato Cuocolo, Arnaldo Stanzione, Anna Castaldo, Davide Raffaele De Lucia, and Massimo Imbriaco. "Quality Control and Whole-Gland, Zonal and Lesion Annotations for the PROSTATEx Challenge Public Dataset". In: *European Journal of Radiology* 138 (2021-05), p. 109647 (cit. on pp. 15, 79).

[Dan+22] Vien Ngoc Dang, Francesco Galati, Rosa Cortese, Giuseppe Di Giacomo, Viola Marconetto, Prateek Mathur, Karim Lekadir, Marco Lorenzi, Ferran Prados, and Maria A. Zuluaga. "Vessel-CAPTCHA: An efficient learning framework for vessel annotation and segmentation". In: *Medical Image Analysis* 75 (2022), p. 102263 (cit. on p. 104).

[DD16] Michel Marie Deza and Elena Deza. "Distances and Similarities in Data Analysis". In: *Encyclopedia of Distances*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016, pp. 327–345 (cit. on pp. 55, 56).

[DDW13] Bin Dai, Shilin Ding, and Grace Wahba. "Multivariate Bernoulli distribution". In: *Bernoulli* 19.4 (2013), pp. 1465–1483 (cit. on p. 55).

[de +20] Maarten de Rooij, Bas Israël, Marcia Tummers, Hashim U. Ahmed, Tristan Barrett, Francesco Giganti, Bernd Hamm, Vibeke Løgager, Anwar Padhani, Valeria Panebianco, Philippe Puech, Jonathan Richenberg, Olivier Rouvière, Georg Salomon, Ivo Schoots, Jeroen Veltman, Geert Villeirs, Jochen Walz, and Jelle O. Barentsz. "ESUR/ESUI Consensus Statements on Multi-Parametric MRI for the Detection of Clinically Significant Prostate Cancer: Quality Requirements for Image Acquisition, Interpretation and Radiologists' Training". In: *European Radiology* 30.10 (2020-10), pp. 5404–5416 (cit. on p. 165).

[Deb+] Noelie Debs, Alexandre Routier, Clement Abi-Nader, Arnaud Marcoux, Francois Nicolas, Alexandre Bone, and Marc-Michel Rohe. *Deep Learning for Detection and Diagnosis of Prostate Cancer from bpMRI and PSA: Guerbet's Contribution to the PI-CAI 2022 Grand Challenge* (cit. on pp. 103, 125).

[DeS+22] Nandita M DeSouza, Aad van Der Lugt, Christophe M Deroose, Angel Alberich-Bayarri, Luc Bidaut, Laure Fournier, Lena Costaridou, Daniela E Oprea-Lager, Elmar Kotter, Marion Smits, et al. "Standardised lesion segmentation for imaging biomarker quantitation: a consensus recommendation from ESR and EORTC". In: *Insights into Imaging* 13.1 (2022), p. 159 (cit. on pp. 132, 134).

[Dic+11] Louise Dickinson, Hashim U. Ahmed, Clare Allen, Jelle O. Barentsz, Brendan Carey, Jurgen J. Futterer, Stijn W. Heijmink, Peter J. Hoskin, Alex Kirkham, Anwar R. Padhani, Raj Persad, Philippe Puech, Shonit Punwani, Aslam S. Sohaib, Bertrand Tombal, Arnauld Villers, Jan van der Meulen, and Mark Emberton. "Magnetic Resonance Imaging for the Detection, Localisation, and Characterisation of Prostate Cancer: Recommendations from a European Consensus Meeting". In: *Eur. Urol.* 59.4 (2011), pp. 477–494 (cit. on pp. 80, 84, 85, 100).

[Dis+17] Florian A. Distler, Jan P. Radtke, David Bonekamp, Claudia Kesch, Heinz-Peter Schlemmer, Kathrin Wieczorek, Marietta Kirchner, Sascha Pahernik, Markus Hohenfellner, and Boris A. Hadaschik. "The Value of PSA Density in Combination with PI-RADS™ for the Accuracy of Prostate Cancer Prediction". In: *The Journal of Urology* 198.3 (2017-09), pp. 575–582 (cit. on p. 20).

[DJL20] A. Duran, P-M Jodoin, and C. Lartizien. "Prostate Cancer Semantic Segmentation by Gleason Score Group in bi-parametric MRI with Self Attention Model on the Peripheral Zone". In: *MIDL*. 2020 (cit. on p. 104).

[DLLP97] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. "Solving the multiple instance problem with axis-parallel rectangles". In: *Artificial Intelligence* 89.1 (1997), pp. 31–71 (cit. on p. 103).

[Dro+19] Frank-Jan H. Drost, Daniël F. Osses, Daan Nieboer, Ewout W. Steyerberg, Chris H. Bangma, Monique J. Roobol, and Ivo G. Schoots. "Prostate MRI, with or without MRI-targeted Biopsy, and Systematic Biopsy for Detecting Prostate Cancer". In: *The Cochrane Database of Systematic Reviews* 4.4 (2019-04), p. CD012663 (cit. on pp. 119, 127).

[Dur+22] Audrey Duran, Gaspard Dussert, Olivier Rouvière, Tristan Jaouen, Pierre-Marc Jodoin, and Carole Lartizien. "ProstAttention-Net: A Deep Attention Model for Prostate Cancer Segmentation by Aggressiveness in MRI Scans". In: *Medical Image Analysis* 77 (2022-04), p. 102347 (cit. on p. 127).

[DV+15] Anne-Sophie Dewalle-Vignion, Nacim Betrouni, Clio Baillet, and Maximilien Vermandel. "Is STAPLE algorithm confident to assess segmentation methods in PET imaging?" In: *Physics in Medicine and Biology* (2015-11) (cit. on p. 49).

[Elk+19] F. Elkhoury, E. Felker, L. Kwan, A. Sisk, M. Delfin, S. Natarajan, and L. Marks. "Comparison of Targeted vs Systematic Prostate Biopsy in Men Who Are Biopsy Naive: The Prospective Assessment of Image Registration in the Diagnosis of Prostate Cancer (PAIREDCAP) Study". In: *JAMA Surg.* 154.9 (2019), pp. 811–818 (cit. on p. 74).

[Eps+16] Jonathan I. Epstein, Lars Egevad, Mahul B. Amin, Brett Delahunt, John R. Srigley, Peter A. Humphrey, and Grading Committee. "The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason Grading of Prostatic Carcinoma: Definition of Grading Patterns and Proposal for a New Grading System". In: *Am J Surg Pathol* 40.2 (2016-02), pp. 244–252 (cit. on pp. 9, 11, 127).

[Equ] *EQUATOR Network | Enhancing the QUAlity and Transparency Of Health Research*. https://www.equator-network.org/ (cit. on p. 165).

[Eri+02] L. M. Eri, H. Thomassen, B. Brennhovd, and L. L. Håheim. "Accuracy and Repeatability of Prostate Volume Measurements by Transrectal Ultrasound". In: *Prostate Cancer and Prostatic Diseases* 5.4 (2002), pp. 273–278 (cit. on p. 26).

[Eve+15] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. "The Pascal Visual Object Classes Challenge: A Retrospective". In: *Int J Comput Vis* 111.1 (2015-01), pp. 98–136 (cit. on p. 13).

[FW+11] Penny F. Whiting, Anne W.S. Rutjes, Marie E. Westwood, Susan Mallett, Jonathan J. Deeks, Johannes B. Reitsma, Mariska M.G. Leeflang, Jonathan A.C. Sterne, Patrick M.M. Bossuyt, and the QUADAS-2 Group*. "QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies". In: *Annals of Internal Medicine* (2011-10) (cit. on p. 152).

[Gao+07] Zhanrong Gao, David Wilkins, Libni Eapen, Christopher Morash, Youssef Wassef, and Lee Gerig. "A study of prostate delineation referenced against a gold standard created from the visible human data". In: *Radiotherapy and Oncology* 85.2 (2007), pp. 239–246 (cit. on pp. 132, 133).

[Gar+14] Brian Garvey, Barış Türkbey, Hong Truong, Marcelino Bernardo, Senthil Periaswamy, and Peter L. Choyke. "Clinical Value of Prostate Segmentation and Volume Determination on MRI in Benign Prostatic Hyperplasia". In: *Diagnostic and Interventional Radiology* 20.3 (2014), pp. 229–233 (cit. on pp. 20, 23).

[Gar+15] Stephen J. Gardner, Ning Wen, Jinkoo Kim, Chang Liu, Deepak Pradhan, Ibrahim Aref, Richard Cattaneo, Sean Vance, Benjamin Movsas, Indrin J. Chetty, and Mohamed A. Elshaikh. "Contouring Variability of Human- and Deformable-Generated Contours in Radiotherapy for Prostate Cancer". In: *Physics in Medicine and Biology* 60.11 (2015-06), pp. 4429–4447 (cit. on p. 143).

[Gat+19] Marco Gatti, Riccardo Faletti, Giorgio Calleris, Jacopo Giglio, Claudio Berzovini, Francesco Gentile, Giancarlo Marra, Francesca Misischi, Luca Molinaro, Laura Bergamasco, Paolo Gontero, Mauro Papotti, and Paolo Fonio. "Prostate cancer detection with biparametric magnetic resonance imaging (bpMRI) by readers with different experience: performance and comparison with multiparametric (mpMRI)". In: *Abdom. Radiol.* 44.5 (2019), 1883—1893 (cit. on p. 74).

[Gha+21] Soleen Ghafoor, Anton S. Becker, Sungmin Woo, Pamela I Causa Andrieu, Daniel Stocker, Natalie Gangai, Hedvig Hricak, and Hebert Alberto Vargas. "Comparison of PI-RADS Versions 2.0 and 2.1 for MRI-based Calculation of the Prostate Volume". In: *Academic Radiology* 28.11 (2021-11), pp. 1548–1556 (cit. on pp. 21, 40, 41).

[Gib+18] Eli Gibson, Yipeng Hu, Nooshin Ghavami, Hashim U. Ahmed, Caroline Moore, Mark Emberton, Henkjan J. Huisman, and Dean C. Barratt. "Inter-Site Variability in Prostate Segmentation Accuracy Using Deep Learning". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Ed. by Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 506–514 (cit. on p. 119).

[Gig+21] Francesco Giganti, Alex Kirkham, Veeru Kasivisvanathan, Marianthi-Vasiliki Papoutsaki, Shonit Punwani, Mark Emberton, Caroline M. Moore, and Clare Allen. "Understanding PI-QUAL for Prostate MRI Quality: A Practical Primer for Radiologists". In: *Insights into Imaging* 12.1 (2021-05), p. 59 (cit. on p. 165).

[Gig+23] Francesco Giganti, Valeria Panebianco, Clare M. Tempany, and Andrei S. Purysko. "Is Artificial Intelligence Replacing Our Radiology Stars in Prostate Magnetic Resonance Imaging? The Stars Do Not Look Big, But They Can Look Brighter". In: *European Urology Open Science* 48 (2023-02), pp. 12–13 (cit. on p. 9).

[GL20] Damien Garreau and Ulrike Luxburg. "Explaining the Explainer: A First Theoretical Analysis of LIME". In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. PMLR, 2020-06, pp. 1287–1296 (cit. on p. 127).

[Gre+18] Matthew D. Greer, Joanna H. Shih, Tristan Barrett, Sandra Bednarova, Ismail Kabakus, Yan Mee Law, Haytham Shebel, Maria J. Merino, Bradford J. Wood, Peter A. Pinto, Peter L. Choyke, and Baris Turkbey. "All Over the Map: An Interobserver Agreement Study of Tumor Location Based on the PI-RADSv2 Sector Map". In: *J Magn Reson Imaging* 48.2 (2018-08), pp. 482–490 (cit. on pp. 12, 74, 99).

[Gre+19] Matthew D. Greer, Joanna H. Shih, Nathan Lay, Tristan Barrett, Leonardo Bittencourt, Samuel Borofsky, Ismail Kabakus, Yan Mee Law, Jamie Marko, Haytham Shebel, Maria J. Merino, Bradford J. Wood, Peter A. Pinto, Ronald M. Summers, Peter L. Choyke, and Baris Turkbey. "Interreader Variability of Prostate Imaging Reporting and Data System Version 2 in Detecting and Assessing Prostate Cancer Lesions at Prostate MRI". In: *AJR Am J Roentgenol* (2019-03), pp. 1–8 (cit. on pp. 12, 102).

[Gre+91] Damian R. Greene, Thomas M. Wheeler, Shin Egawa, J. Kay Dunn, and Peter T. Scardino. "A Comparison of the Morphological Features of Cancer Arising in the Transition Zone and in the Peripheral Zone of the Prostate". In: *The Journal of Urology* 146.4 (1991-10), pp. 1069–1076 (cit. on p. 108).

[GS18] Alonso Gragera and Vorapong Suppakitpaisarn. "Relaxed triangle inequality ratio of the Sørensen–Dice and Tversky indexes". In: *Theoretical Computer Science* 718 (2018), pp. 37–45 (cit. on p. 56).

[GSJ20] Jatin Gupta, Sumindar Kaur Saini, and Mamta Juneja. "Survey of Denoising and Segmentation Techniques for MRI Images of Prostate for Improving Diagnostic Tools in Medical Applications". In: *Materials Today: Proceedings*. International Conference on Aspects of Materials Science and Engineering 28 (2020-01), pp. 1667–1672 (cit. on p. 165).

[Gun+22] Deepa Darshini Gunashekar, Lars Bielak, Leonard Hägele, Benedict Oerther, Matthias Benndorf, Anca-L Grosu, Thomas Brox, Constantinos Zamboglou, and Michael Bock. "Explainable AI for CNN-based prostate tumor segmentation in multi-parametric MRI correlated to whole mount histopathology". In: *Radiation Oncology* 17.1 (2022), pp. 1–10 (cit. on pp. 132, 133).

[Haa+17] Matthias Haas, Karsten Günzel, Kurt Miller, Bernd Hamm, Hannes Cash, and Patrick Asbach. "Is the Ellipsoid Formula the New Standard for 3-Tesla MRI Prostate Volume Calculation without Endorectal Coil?" In: *Urologia Internationalis* 98.1 (2017), pp. 49–53 (cit. on p. 42).

[Ham+19] Praful Hambarde, Sanjay N. Talbar, Nilesh Sable, Abhishek Mahajan, Satishkumar S. Chavan, and Meenakshi Thakur. “Radiomics for Peripheral Zone and Intra-Prostatic Urethra Segmentation in MR Imaging”. In: *Biomedical Signal Processing and Control* 51 (2019-05), pp. 19–29 (cit. on pp. 153–155, 162, 164).

[Ham+22a] Dimitri Hamzaoui, Sarah Montagne, Benjamin Granger, Alexandre Allera, Malek Ezziane, Anna Luzurier, Raphaëlle Quint, Mehdi Kalai, Nicholas Ayache, Hervé Delingette, and Raphaële Renard-Penna. “Prostate Volume Prediction on MRI: Tools, Accuracy and Variability”. In: *Eur Radiol* 32.7 (2022-07), pp. 4931–4941 (cit. on pp. 13, 16, 20).

[Ham+22b] Dimitri Hamzaoui, Sarah Montagne, Raphaële Renard-Penna, Nicholas Ayache, and Hervé Delingette. “Automatic Zonal Segmentation of the Prostate from 2D and 3D T2-weighted MRI and Evaluation for Clinical Use”. In: *Journal of Medical Imaging* 9.2 (2022-03), p. 024001 (cit. on pp. 14, 16, 74, 106, 107, 109, 111, 134).

[Ham+22c] Dimitri Hamzaoui, Sarah Montagne, Raphaele Renard-Penna, Nicholas Ayache, and Hervé Delingette. “MOrphologically-aware Jaccard-based ITerative Optimization (MOJITO) for Consensus Segmentation”. In: *MICCAI Workshop UNSURE 2022: Uncertainty for Safe Utilization of Machine Learning in Medical Imaging*. 2022-09 (cit. on pp. 14, 16, 48).

[Ham+23a] Charlie A. Hamm, Georg L. Baumgärtner, Felix Biessmann, Nick L. Beetz, Alexander Hartenstein, Lynn J. Savic, Konrad Froböse, Franziska Dräger, Simon Schallenberg, Madhuri Rudolph, Alexander D. J. Baur, Bernd Hamm, Matthias Haas, Sebastian Hofbauer, Hannes Cash, and Tobias Penzkofer. “Interactive Explainable Deep Learning Model Informs Prostate Cancer Diagnosis at MRI”. In: *Radiology* (2023-04), p. 222276 (cit. on p. 127).

[Ham+23b] Dimitri Hamzaoui, Sarah Montagne, Raphaële Renard-Penna, Sébastien Moliere, Nicholas Ayache, and Hervé Delingette. *Morphologically-Aware Consensus Computation via Heuristics-based IterATive Optimization (MACCHIatO)*. Submitted to MELBA - Journal of Medical Imaging. 2023 (cit. on pp. 14, 16, 48).

[Ham+23c] Dimitri Hamzaoui, Raphaële Renard-Penna, Sarah Montagne, Sébastien Moliere, Nicholas Ayache, and Hervé Delingette. *Weakly and Mixed supervision for prostate cancer detection through radiological annotations*. In preparation. 2023 (cit. on pp. 14, 16, 102).

[He+15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. *Deep Residual Learning for Image Recognition*. Arxiv eprints. 2015-12 (cit. on p. 150).

[Hos+21] Matin Hosseinzadeh, Anindo Saha, Patrick Brand, Ilse Slootweg, Maarten de Rooij, and Henkjan Huisman. “Deep Learning–Assisted Prostate Cancer Detection on Bi-parametric MRI: Minimum Training Data Size Requirements and Effect of Prior Knowledge”. In: *Eur. Radiol.* 32 (2021-11) (cit. on pp. 100, 103).

[Hou+16] Le Hou, Dimitris Samaras, Tahsin M. Kurc, Yi Gao, James E. Davis, and Joel H. Saltz. “Patch-Based Convolutional Neural Network for Whole Slide Tissue Image Classification”. In: *Proceedings. IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2016 (2016), pp. 2424–2433 (cit. on p. 104).

[HSS18] J. Hu, L. Shen, and G. Sun. “Squeeze-and-Excitation Networks”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7132–7141 (cit. on pp. 83, 111, 112).

[Hum04] Peter A. Humphrey. “Gleason Grading and Prognostic Factors in Carcinoma of the Prostate”. In: *Mod Pathol* 17.3 (2004-03), pp. 292–306 (cit. on pp. 9, 11).

[Ins22] Institut National du Cancer. *Panorama Des Cancers En France*. 2022 (cit. on p. 7).

[Ise+18] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. “Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge”. In: *Brain lesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Ed. by Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Bjoern Menze, and Mauricio Reyes. Vol. 10670. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 287–297 (cit. on p. 87).

[Ise+20] F. Isensee, P. Jäger, J. Wasserthal, D. Zimmerer, J. Petersen, S. Kohl, J. Schock, A. Klein, T. Roß, S. Wirkert, P. Neher, S. Dinkelacker, G. Köhler, and K. Maier-Hein. *Batchgenerators - a Python framework for data augmentation*. https://github.com/MIC-DKFZ/batchgenerators. 2020 (cit. on p. 85).

[Ise+21] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation”. In: *Nature Methods* 18 (2 2021), pp. 203–211 (cit. on pp. 21, 48, 75, 87, 103, 142).

[Jae+18] Paul F. Jaeger, Simon A. A. Kohl, Sebastian Bickelhaupt, Fabian Isensee, Tristan Anselm Kuder, Heinz-Peter Schlemmer, and Klaus H. Maier-Hein. *Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection*. Arxiv eprints. 2018-11 (cit. on p. 103).

[Jen+19] Carina Jensen, Kristine Storm Sørensen, Cecilia Klitgaard Jørgensen, Camilla Winther Nielsen, Pia Christine Høy, Niels Christian Langkilde, and Lasse Riis Østergaard. “Prostate Zonal Segmentation in 1.5T and 3T T2W MRI Using a Convolutional Neural Network”. In: *Journal of Medical Imaging* 6.1 (2019-02), p. 014501 (cit. on pp. 153, 155, 156, 162, 164, 165).

[Jeo+08] Chang Wook Jeong, Hyoung Keun Park, Sung Kyu Hong, Seok-Soo Byun, Hak Jong Lee, and Sang Eun Lee. “Comparison of Prostate Volume Measured by Transrectal Ultrasonography and MRI with the Actual Prostate Volume Measured after Radical Prostatectomy”. In: *Urologia Internationalis* 81.2 (2008), pp. 179–185 (cit. on pp. 20, 23).

[Jia+19] Haozhe Jia, Yang Song, Heng Huang, Weidong Cai, and Yong Xia. “HD-Net: Hybrid Discriminative Network for Prostate Segmentation in MR Images”. In: *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019*. Ed. by Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan. Vol. 11765. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 110–118 (cit. on p. 75).

[Jon+06] Sara Jonmarker, Alexander Valdman, Anna Lindberg, Magnus Hellström, and Lars Egevad. “Tissue Shrinkage after Fixation with Formalin Injection of Prostatectomy Specimens”. In: *Virchows Archiv: An International Journal of Pathology* 449.3 (2006-09), pp. 297–301 (cit. on p. 42).

[Jos+19] Leo Joskowicz, D Cohen, N Caplan, and Jacob Sosna. “Inter-observer variability of manual contour delineation of structures in CT”. In: *European Radiology* 29 (2019), pp. 1391–1399 (cit. on pp. 137, 142, 145).

[Kar+18] Davood Karimi, Golnoosh Samei, Claudia Kesch, Guy Nir, and Septimiu Salcudean. "Prostate segmentation in MRI using a convolutional neural network architecture and training strategy based on statistical shape models". In: *Int. J. Comput. Assist. Radiol. Surg.* 13.8 (2018), pp. 1211–1219 (cit. on p. 100).

[Kas+18] V. Kasivisvanathan et al. "MRI-Targeted or Standard Biopsy for Prostate-Cancer Diagnosis". In: *New Engl. J. Med.* 378.19 (2018), pp. 1767–1777 (cit. on pp. 8, 74).

[Kas+19] V. Kasivisvanathan, A. Stabile, J.B. Neves, F. Giganti, M. Valerio, Y. Shanmugabavan, K.D. Clement, D. Sarkar, Y. Philippou, D. Thurtle, J. Deeks, M. Emberton, Y. Takwoingi, and C.M. Moore. "Magnetic Resonance Imaging-targeted Biopsy Versus Systematic Biopsy in the Detection of Prostate Cancer: A Systematic Review and Meta-analysis". In: *Eur. Urol.* 76.3 (2019), pp. 284–303 (cit. on p. 74).

[Ker+21] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. "Boundary loss for highly unbalanced segmentation". In: *Med. Image Anal.* 67 (2021), p. 101851 (cit. on p. 104).

[Key] *Key Statistics for Prostate Cancer | Prostate Cancer Facts*. https://www.cancer.org/cancer/prostate-cancer/about/key-statistics.html (cit. on p. 125).

[Kha+16] Farzad Khalvati, Aryan Salmanpour, Shahryar Rahnamayan, Masoom A. Haider, and H. R. Tizhoosh. "Sequential Registration-Based Segmentation of the Prostate Gland in MR Image Volumes". In: *Journal of Digital Imaging* 29.2 (2016-04), pp. 254–263 (cit. on p. 143).

[Kha+19] Zia Khan, Norashikin Yahya, Khaled Alsaih, and Fabrice Meriaudeau. "Zonal Segmentation of Prostate T2W-MRI Using Atrous Convolutional Neural Network". In: *2019 IEEE Student Conference on Research and Development (SCOReD)*. 2019-10, pp. 95–99 (cit. on pp. 150, 153, 162, 164).

[Kor+15] Anne Sofie Korsager, Valerio Fortunati, Fedde van der Lijn, Jesper Carl, Wiro Niessen, Lasse Riis Østergaard, and Theo van Walsum. "The Use of Atlas Registration and Graph Cuts for Prostate Segmentation in Magnetic Resonance Images". In: *Medical Physics* 42.4 (2015-04), pp. 1614–1624 (cit. on pp. 20, 150).

[Kos19] Sven Kosub. "A note on the triangle inequality for the Jaccard distance". In: *Pattern Recognition Letters* 120 (2019), pp. 36–38 (cit. on p. 56).

[Kra88] Dieter Kraft. *A software package for sequential quadratic programming*. Tech. rep. DFVLR-FB 88-28. Koln, Germany: DLR German Aerospace Center – Institute for Flight Mechanics, 1988 (cit. on p. 60).

[KS19] Davood Karimi and Septimiu E Salcudean. "Reducing the Hausdorff Distance in Medical Image Segmentation with Convolutional Neural Networks". In: *IEEE Transactions on medical imaging* 39.2 (2019), pp. 499–513 (cit. on pp. 67, 68).

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Vol. 25. Curran Associates, Inc., 2012 (cit. on p. 75).

[Lai+21] Chih-Ching Lai, Hsin-Kai Wang, Fu-Nien Wang, Yu-Ching Peng, Tzu-Ping Lin, Hsu-Hsia Peng, and Shu-Huei Shen. “Autosegmentation of Prostate Zones and Cancer Regions from Biparametric Magnetic Resonance Images by Using Deep-Learning-Based Neural Networks”. In: *Sensors* 21.8 (2021-01), p. 2709 (cit. on pp. 154, 162, 164, 165).

[LC07] Jae Seok Lee and Byung Ha Chung. “Transrectal Ultrasound versus Magnetic Resonance Imaging in the Estimation of Prostate Volume as Compared with Radical Prostatectomy Specimens”. In: *Urologia Internationalis* 78.4 (2007), pp. 323–327 (cit. on p. 20).

[Lee13] Dong-Hyun Lee. “Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks”. In: *Workshop on Challenges in Representation Learning, ICML*. Vol. 3. Atlanta, 2013, p. 896 (cit. on p. 104).

[Lee+20] Dong Kyu Lee, Deuk Jae Sung, Chang-Su Kim, Yuk Heo, Jeong Yoon Lee, Beom Jin Park, and Min Ju Kim. “Three-Dimensional Convolutional Neural Network for Prostate MRI Segmentation and Comparison of Prostate Volume Measurements by Use of Artificial Neural Network and Ellipsoid Formula”. In: *American Journal of Roentgenology* 214.6 (2020-06), pp. 1229–1238 (cit. on pp. 154, 162, 164, 165).

[Lee+22] Changhee Lee, Alexander Light, Evgeny S. Saveliev, Mihaela van der Schaar, and Vincent J. Gnanapragasam. “Developing Machine Learning Algorithms for Dynamic Estimation of Progression during Active Surveillance for Prostate Cancer”. In: *npj Digital Medicine* 5.1 (2022-08), pp. 1–7 (cit. on p. 126).

[Lem+15] Guillaume Lemaître, Robert Martí, Jordi Freixenet, Joan C. Vilanova, Paul M. Walker, and Fabrice Meriaudeau. “Computer-Aided Detection and Diagnosis for Prostate Cancer Based on Mono and Multi-Parametric MRI: A Review”. In: *Computers in Biology and Medicine* 60 (2015-05), pp. 8–31 (cit. on p. 9).

[LG07] Andrew R. Leach and Valerie J. Gillet. “Similarity Methods”. In: *An Introduction To Chemoinformatics*. Dordrecht: Springer Netherlands, 2007, pp. 99–117 (cit. on p. 56).

[Lin+15] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. *Microsoft COCO: Common Objects in Context*. Arxiv eprints. 2015-02 (cit. on p. 13).

[Lin+17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. *Focal Loss for Dense Object Detection*. 2017 (cit. on p. 111).

[Lit+12] Geert Litjens, Oscar Debats, Wendy van de Ven, Nico Karssemeijer, and Henkjan Huisman. “A Pattern Recognition Approach to Zonal Segmentation of the Prostate on MRI”. In: *Medical image computing and computer-assisted intervention: MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention* 15.Pt 2 (2012), pp. 413–420 (cit. on pp. 150, 153, 155, 156, 162, 164).

[Lit+14a] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman. “Computer-aided detection of prostate cancer in MRI”. In: *IEEE Trans. Med. Imaging* 33 (2014), pp. 1083–1092 (cit. on pp. 76, 78).

[Lit+14b] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, R. Strand, F. Malmberg, Y. Ou, C. Davatzikos, M. Kirschner, F. Jung, J. Yuan, W. Qiu, Q. Gao, P. Edwards, B. Maan, F. van der Heijden, S. Ghose, J. Mitra, J. Dowling, D. Barratt, H. Huisman, and A. Madabhushi. "Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge". In: *Med. Image Anal.* 18.2 (2014), pp. 359–373 (cit. on pp. 21, 39, 75, 154).

[Lit+17] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman. *ProstateX Challenge data*. The Cancer Imaging Archive (TCIA). Data available at https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=23691656. 2017 (cit. on pp. 76, 78).

[Liu+] Siqi Liu, Daguang Xu, S. Kevin Zhou, Olivier Pauly, Sasa Grbic, Thomas Mertelmeier, Julia Wicklein, Anna Jerebko, Weidong Cai, and Dorin Comaniciu. ""3D Anisotropic Hybrid Network: Transferring Convolutional Features from 2D Images to 3D Anisotropic Volumes". In: ed. by Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger (cit. on p. 80).

[Liu+12] Derek Liu, Nawaid Usmani, Sunita Ghosh, Wafa Kamal, John Pedersen, Nadeem Pervez, Don Yee, Brita Danielson, Albert Murtha, John Amanie, and Ron S. Sloboda. "Comparison of Prostate Volume, Shape, and Contouring Variability Determined from Preimplant Magnetic Resonance and Transrectal Ultrasound Images". In: *Brachytherapy* 11.4 (2012), pp. 284–291 (cit. on p. 143).

[Liu+19] Yongkai Liu, Guang Yang, Sohrab Afshari Mirak, Melina Hosseiny, Afshin Azadikhah, Xinran Zhong, Robert E. Reiter, Yeejin Lee, Steven S. Raman, and Kyunghyun Sung. "Automatic Prostate Zonal Segmentation Using Fully Convolutional Network With Feature Pyramid Attention". In: *IEEE Access* 7 (2019), pp. 163626–163632 (cit. on pp. 150, 153, 162, 164, 165).

[Liu+20] Yongkai Liu, Guang Yang, Melina Hosseiny, Afshin Azadikhah, Sohrab Afshari Mirak, Qi Miao, Steven S. Raman, and Kyunghyun Sung. "Exploring Uncertainty Measures in Bayesian Deep Attentive Neural Networks for Prostate Zonal Segmentation". In: *IEEE Access* 8 (2020), pp. 151817–151828 (cit. on pp. 154, 162, 164, 165).

[Low+13] Bradley Lowekamp, David Chen, Luis Ibanez, and Daniel Blezek. "The Design of SimpleITK". In: *Frontiers in Neuroinformatics* 7 (2013) (cit. on pp. 23, 25, 61, 80).

[Ma+21] Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L. Martel. "Loss odyssey in medical image segmentation". In: *Medical Image Analysis* 71 (2021), p. 102035 (cit. on p. 56).

[Mak+11] N. Makni, A. Iancu, O. Colot, P. Puech, S. Mordon, and N. Betrouni. "Zonal Segmentation of Prostate Using Multispectral Magnetic Resonance Images: Zonal Segmentation of Prostate Using Multispectral MR Images". In: *Medical Physics* 38.11 (2011-10), pp. 6093–6105 (cit. on pp. 153–155, 162, 164).

[Mar+16] Kasper Marstal, Floris Berendsen, Marius Staring, and Stefan Klein. "SimpleElastix: A User-Friendly, Multi-lingual Library for Medical Image Registration". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2016, pp. 574–582 (cit. on p. 114).

[Maz+15] Yousef Mazaheri, Debra A. Goldman, Pier Luigi Di Paolo, Oguz Akin, and Hedvig Hricak. "Comparison of Prostate Volume Measured by Endorectal Coil MRI to Prostate Specimen Volume and Mass after Radical Prostatectomy". In: *Academic Radiology* 22.5 (2015-05), pp. 556–562 (cit. on pp. 21, 40, 41).

[MBC14] Nasr Makni, Nacim Betrouni, and Olivier Colot. “Introducing Spatial Neighbourhood in Evidential C-Means for Segmentation of Multi-Source Images: Application to Prostate Multi-Parametric MRI”. In: *Information Fusion*. Special Issue on Information Fusion in Medical Image Computing and Systems 19 (2014-09), pp. 61–72 (cit. on pp. 150, 153, 155, 162, 164).

[MBH18] Germonda Mooij, Ines Bagulho, and Henkjan Huisman. *Automatic Segmentation of Prostate Zones*. Arxiv eprints. 2018-06 (cit. on pp. 153, 162, 164, 165).

[McN68] J. E. McNeal. “Regional Morphology and Pathology of the Prostate”. In: *American Journal of Clinical Pathology* 49.3 (1968-03), pp. 347–357 (cit. on p. 21).

[McN83] J. E. McNeal. “The Prostate Gland, Morphology and Pathobiology”. In: *Monographs in urology* 4 (1983), pp. 3–37 (cit. on p. 39).

[McN+88] J. E. McNeal, E. A. Redwine, F. S. Freiha, and T. A. Stamey. “Zonal Distribution of Prostatic Adenocarcinoma. Correlation with Histologic Pattern and Direction of Spread”. In: *The American Journal of Surgical Pathology* 12.12 (1988-12), pp. 897–906 (cit. on p. 9).

[Meh+21] Pritesh Mehta, Michela Antonelli, Hashim U. Ahmed, Mark Emberton, Shonit Punwani, and Sébastien Ourselin. “Computer-aided diagnosis of prostate cancer using multiparametric MRI and clinical features: A patient-level classification framework”. In: *Med. Image Anal.* 73 (2021), p. 102153 (cit. on pp. 100, 125).

[Mey+18] Anneke Meyer, Alireza Mehrtash, Marko Rak, Daniel Schindele, Martin Schostak, Clare Tempany, Tina Kapur, Purang Abolmaesumi, Andriy Fedorov, and Christian Hansen. “Automatic high resolution segmentation of the prostate from multi-planar MRI”. In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 177–181 (cit. on p. 76).

[Mey+19] A. Meyer, M. Rakr, D. Schindele, S. Blaschke, M. Schostak, A. Fedorov, and C. Hansen. “Towards Patient-Individual PI-Rads v2 Sector Map: CNN for Automatic Segmentation of Prostatic Zones From T2-Weighted MRI”. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019, pp. 696–700 (cit. on pp. 21, 76, 143, 153–155, 162–165).

[Mey+21] A. Meyer, G. Chlebus, M. Rak, D. Schindele, M. Schostak, B. van Ginneken, A. Schenk, H. Meine, H.K. Hahn, A. Schreiber, and C. Hansen. “Anisotropic 3D Multi-Stream CNN for Accurate Prostate Segmentation from Multi-Planar MRI”. In: *Comput. Methods Programs Biomed.* 200 (2021), p. 105821 (cit. on pp. 80, 84, 165).

[MG12] Emmanouil Moschidis and Jim Graham. “Automatic Differential Segmentation of the Prostate in 3-D MRI Using Random Forest Classification and Graph-Cuts Optimization”. In: *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*. 2012-05, pp. 1727–1730 (cit. on pp. 153, 154, 156, 162, 164).

[Mly+19] Pawel Mlynarski, Hervé Delingette, Antonio Criminisi, and Nicholas Ayache. “Deep Learning with Mixed Supervision for Brain Tumor Segmentation”. In: *Journal of Medical Imaging* 6.3 (2019-08), p. 034002 (cit. on p. 104).

[MMK20] John Mongan, Linda Moy, and Charles E. Kahn. “Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers”. In: *Radiology: Artificial Intelligence* 2.2 (2020-03), e200029 (cit. on pp. 153, 165).

[MNA16] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation”. In: *2016 Fourth International Conference on 3D Vision (3DV)*. 2016, pp. 565–571 (cit. on pp. 21, 75, 150).

[Mol+23] Sébastien Moliere, Dimitri Hamzaoui, Sarah Montagne, Alexandre Allera, Malek Ezziane, Anna Luzurier, Raphaelle Quint, Mehdi Kalai, Nicholas Ayache, Hervé Delingette, and Raphaële Renard-Penna. *Reference standard for evaluation of automatic segmentation algorithms: quantification of inter observer variability of manual delineation of prostate contour on MRI*. Abstract accepted to RSNA 2023. 2023 (cit. on pp. 13, 16, 132).

[Mon+21] Sarah Montagne, Dimitri Hamzaoui, Alexandre Allera, Malek Ezziane, Anna Luzurier, Raphaelle Quint, Mehdi Kalai, Nicholas Ayache, Hervé Delingette, and Raphaële Renard-Penna. “Challenge of Prostate MRI Segmentation on T2-weighted Images: Inter-Observer Variability and Impact of Prostate Morphology”. In: *Insights Imaging* 12.1 (2021-06), p. 71 (cit. on pp. 12, 13, 16, 20, 61, 134, 143, 163).

[Mot+20] Saman Motamed, Isha Gujrathi, Dominik Deniffel, Anton Oentoro, Masoom A. Haider, and Farzad Khalvati. *A Transfer Learning Approach for Automated Segmentation of Prostate Whole Gland and Transition Zone in Diffusion Weighted MRI*. Arxiv eprints. 2020-10 (cit. on pp. 154, 162, 164, 165).

[Mot+21] Nicolas Mottet, Roderick C. N. van den Bergh, Erik Briers, Thomas Van den Broeck, Marcus G. Cumberbatch, Maria De Santis, Stefano Fanti, Nicola Fossati, Giorgio Gandaglia, Silke Gillessen, Nikos Grivas, Jeremy Grummet, Ann M. Henry, Theodorus H. van der Kwast, Thomas B. Lam, Michael Lardas, Matthew Liew, Malcolm D. Mason, Lisa Moris, Daniela E. Oprea-Lager, Henk G. van der Poel, Olivier Rouvière, Ivo G. Schoots, Derya Tilki, Thomas Wiegel, Peter-Paul M. Willemse, and Philip Cornford. “EAU-EANM-ESTRO-ESUR-SIOG Guidelines on Prostate Cancer-2020 Update. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent”. In: *Eur Urol* 79.2 (2021-02), pp. 243–262 (cit. on pp. 8, 26, 74, 125, 126, 129, 149).

[Mus+19] Thais Caldara Mussi, Fernando Ide Yamauchi, Cássia Franco Tridente, Adriano Tachibana, Victor Martins Tonso, Débora Rachello Recchimuzzi, Layra Ribeiro de Souza Leão, Daniel Calich Luz, Tatiana Martins, and Ronaldo Hueb Baroni. “Interobserver Agreement and Positivity of PI-RADS Version 2 Among Radiologists with Different Levels of Experience”. In: *Academic Radiology* 26.8 (2019-08), pp. 1017–1022 (cit. on pp. 12, 102).

[MW96] Kenneth O. McGraw and S. P. Wong. “Forming Inferences about Some Intraclass Correlation Coefficients”. In: *Psychological Methods* 1 (1996), pp. 30–46 (cit. on p. 25).

[Nai+20] Ying-Hwey Nai, Bernice W. Teo, Nadya L. Tan, Koby Yi Wei Chua, Chun Kit Wong, Sophie O’Doherty, Mary C. Stephenson, Josh Schaefferkoetter, Yee Liang Thian, Edmund Chiong, and Anthonin Reilhac. “Evaluation of Multimodal Algorithms for the Segmentation of Multiparametric MRI Prostate Images”. In: *Computational and Mathematical Methods in Medicine* 2020 (2020-10), e8861035 (cit. on pp. 150, 153, 155, 162–164).

[NB15] Anant Madabhushi Nicholas Bloch. *NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures*. 2015 (cit. on p. 154).

[Nia+12] E. Niaf, O. Rouvière, F. Mège-Lechevallier, F. Bratan, and C. Lartizien. "Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI". In: *Physics in Medicine and Biology* 57.12 (2012), pp. 3833–3851 (cit. on p. 103).

[NUZ00] L.G. Nyul, J.K. Udupa, and Xuan Zhang. "New Variants of a Method of MRI Scale Standardization". In: *IEEE Transactions on Medical Imaging* 19.2 (2000-02), pp. 143–150 (cit. on p. 126).

[Nyh+13] Tufve Nyholm, Joakim Jonsson, Karin Söderström, Per Bergström, Andreas Carlberg, Gunilla Frykholm, Claus F. Behrens, Poul Flemming Geertsen, Redas Trepiakas, Scott Hanvey, Azmat Sadozye, Jawaher Ansari, Hazel McCallum, John Frew, Rhona McMenemin, and Björn Zackrisson. "Variability in Prostate and Seminal Vesicle Delineations Defined on Magnetic Resonance Images, a Multi-Observer, -Center and -Sequence Study". In: *Radiation Oncology* 8.1 (2013-05), p. 126 (cit. on p. 143).

[Okt+18] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N.Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert. *Attention U-Net: Learning Where to Look for the Pancreas*. Arxiv eprints. 2018 (cit. on pp. 83, 111, 112).

[Orc+14] C. Orczyk, S. S. Taneja, H. Rusinek, and A. B. Rosenkrantz. "Assessment of Change in Prostate Volume and Shape Following Surgical Resection through Co-Registration of in-Vivo MRI and Fresh Specimen Ex-Vivo MRI". In: *Clinical Radiology* 69.10 (2014-10), e398–403 (cit. on p. 42).

[Ots79] Nobuyuki Otsu. "A Threshold Selection Method from Gray-Level Histograms". In: *IEEE Transactions on Systems, Man, and Cybernetics* 9.1 (1979), pp. 62–66 (cit. on p. 109).

[Pad+19] Kyle R. Padgett, Amy Swallen, Sara Pirozzi, Jon Piper, Felix M. Chinea, Matthew C. Abramowitz, Aaron Nelson, Alan Pollack, and Radka Stoyanova. "Towards a Universal MRI Atlas of the Prostate and Prostate Zones : Comparison of MRI Vendor and Image Acquisition Parameters". In: *Strahlentherapie Und Onkologie: Organ Der Deutschen Rontgengesellschaft ... [et Al]* 195.2 (2019-02), pp. 121–130 (cit. on pp. 38, 143, 150, 153, 156, 162, 164, 165).

[Pag+21] Matthew J. Page, Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, Roger Chou, Julie Glanville, Jeremy M. Grimshaw, Asbjørn Hróbjartsson, Manoj M. Lalu, Tianjing Li, Elizabeth W. Loder, Evan Mayo-Wilson, Steve McDonald, Luke A. McGuinness, Lesley A. Stewart, James Thomas, Andrea C. Tricco, Vivian A. Welch, Penny Whiting, and David Moher. "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews". In: *BMJ* 372 (2021-03), n71 (cit. on p. 151).

[Par+20] S. Park, L. C. Chu, E. K. Fishman, A. L. Yuille, B. Vogelstein, K. W. Kinzler, K. M. Horton, R. H. Hruban, E. S. Zinreich, D. Fadaei Fouladi, S. Shayesteh, J. Graves, and S. Kawamoto. "Annotated Normal CT Data of the Abdomen for Deep Learning: Challenges and Strategies for Implementation". In: *Diagnostic and Interventional Imaging* 101.1 (2020-01), pp. 35–44 (cit. on p. 133).

[Pas+16] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. *ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation*. Arxiv eprints. 2016 (cit. on p. 76).

[Pat+16] Nicholas R. Paterson, Luke T. Lavallée, Laura N. Nguyen, Kelsey Witiuk, James Ross, Ranjeeta Mallick, Wael Shabana, Blair MacDonald, Nicola Scheida, Dean Fergusson, Franco Momoli, Sonya Cnossen, Christopher Morash, Ilias Cagiannos, and Rodney H. Breau. "Prostate Volume Estimations Using Magnetic Resonance Imaging and Transrectal Ultrasound Compared to Radical Prostatectomy Specimens". In: *Canadian Urological Association Journal = Journal De l'Association Des Urologues Du Canada* 10.7-8 (2016-08), pp. 264–268 (cit. on p. 20).

[Pat+19a] Angela U. Pathmanathan, Helen A. McNair, Maria A. Schmidt, Douglas H. Brand, Louise Delacroix, Cynthia L. Eccles, Alexandra Gordon, Trina Herbert, Nicholas J. van As, Robert A. Huddart, and Alison C. Tree. "Comparison of Prostate Delineation on Multimodality Imaging for MR-guided Radiotherapy". In: *The British Journal of Radiology* 92.1095 (2019-03), p. 20180948 (cit. on p. 143).

[Pat+19b] Angela U. Pathmanathan, Maria A. Schmidt, Douglas H. Brand, Evanthia Kousi, Nicholas J. van As, and Alison C. Tree. "Improving Fiducial and Prostate Capsule Visualization for Radiotherapy Planning Using MRI". In: *Journal of Applied Clinical Medical Physics* 20.3 (2019-03), pp. 27–36 (cit. on p. 143).

[PD22] Gaurav Patel and Jose Dolz. "Weakly Supervised Segmentation with Cross-Modality Equivariant Constraints". In: *Medical Image Analysis* 77 (2022-04), p. 102374 (cit. on p. 104).

[Pol+17] Stephan Polanec, Mathias Lazar, Georg Wengert, Hubert Bickel, Claudio Spick, Martin Susani, Shahrokh Shariat, Paola Clauser, and Pascal Baltzer. "3D T2-weighted imaging to shorten multiparametric prostate MRI protocols". In: *Eur. Radiol.* 28.4 (2017), pp. 1634–1641 (cit. on p. 76).

[Pop+06] Teo Popa, Luis Ibanez, Elliot Levy, Amy White, Jill Bruno, and Kevin Cleary. "Tumor volume measurement and volume measurement comparison plug-ins for VolView using ITK". In: *Medical Imaging 2006: Visualization, Image-Guided Procedures, and Display*. Ed. by Kevin R. Cleary and Jr. Galloway Robert L. Vol. 6141. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. 2006, pp. 395–402 (cit. on p. 79).

[Pra+17] Ferran Prados, John Ashburner, Claudia Blaiotta, Tom Brosch, Julio Carballido-Gamio, Manuel Jorge Cardoso, Benjamin N. Conrad, Esha Datta, Gergely Dávid, Benjamin De Leener, Sara M. Dupont, Patrick Freund, Claudia A.M. Gandini Wheeler-Kingshott, Francesco Grussu, Roland Henry, Bennett A. Landman, Emil Ljungberg, Bailey Lyttle, Sebastien Ourselin, Nico Papinutto, Salvatore Saporito, Regina Schlaeger, Seth A. Smith, Paul Summers, Roger Tam, Marios C. Yiannakas, Alyssa Zhu, and Julien Cohen-Adad. "Spinal cord grey matter segmentation challenge". In: *NeuroImage* 152 (2017), pp. 312–329 (cit. on p. 61).

[Pup+18] L. F. Pupulim, M. Ronot, V. Paradis, S. Chemouny, and V. Vilgrain. "Volumetric Measurement of Hepatic Tumors: Accuracy of Manual Contouring Using CT with Volumetric Pathology as the Reference Method". In: *Diagnostic and Interventional Imaging* 99.2 (2018-02), pp. 83–89 (cit. on p. 133).

[Pur+21] Andrei S. Purysko, Ronaldo H. Baroni, Francesco Giganti, Daniel Costa, Raphaële Renard-Penna, Chan Kyo Kim, and Steven S. Raman. "PI-RADS Version 2.1: A Critical Review, From the AJR Special Series on Radiology Reporting and Data Systems". In: *American Journal of Roentgenology* 216.1 (2021-01), pp. 20–32 (cit. on p. 38).

[Qin+20] Xiangxiang Qin, Yu Zhu, Wei Wang, Shaojun Gui, Bingbing Zheng, and Peijun Wang. “3D Multi-Scale Discriminative Network with Multi-Directional Edge Loss for Prostate Zonal Segmentation in Bi-Parametric MR Images”. In: *Neurocomputing* 418 (2020-12), pp. 148–161 (cit. on pp. 154, 156, 162, 164, 165).

[Rah+92] A. Rahmouni, A. Yang, C. M. Tempany, T. Frenkel, J. Epstein, P. Walsh, P. K. Leichner, C. Ricci, and E. Zerhouni. “Accuracy of In-Vivo Assessment of Prostatic Volume by MRI and Transrectal Ultrasonography”. In: *Journal of Computer Assisted Tomography* 16.6 (1992), pp. 935–940 (cit. on pp. 20, 42).

[RB11] Claus G. Roehrborn and Libby K. Black. “The Economic Burden of Prostate Cancer”. In: *BJU International* 108.6 (2011), pp. 806–813 (cit. on pp. 7, 102).

[Ree+16] Fairleigh Reeves, Wouter Everaerts, Declan G. Murphy, and Anthony Costello. “Chapter 29 - The Surgical Anatomy of the Prostate”. In: *Prostate Cancer (Second Edition)*. Ed. by Jack H. Mydlo and Ciril J. Godec. 2nd Edition. San Diego: Academic Press, 2016, pp. 253–263 (cit. on p. 8).

[Ren+20] Félix Renard, Soulaimane Guedria, Noel De Palma, and Nicolas Vuillerme. “Variability and Reproducibility in Deep Learning for Medical Image Segmentation”. In: *Scientific Reports* 10.1 (2020-08), p. 13724 (cit. on p. 12).

[RFB15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015*. Ed. by Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi. Vol. 9351. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 234–241 (cit. on pp. 75, 150, 156).

[Rib+19] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. *Kornia: An Open Source Differentiable Computer Vision Library for PyTorch*. Arxiv eprints. 2019-10 (cit. on p. 114).

[RM07] Torsten Rohlfing and Calvin R. Maurer. “Shape-Based Averaging”. In: *IEEE Transactions on Image Processing* 16.1 (2007), pp. 153–161 (cit. on p. 48).

[Rom+18] Eduardo Romera, José M. Álvarez, Luis M. Bergasa, and Roberto Arroyo. “ERFNet: Efficient Residual Factorized ConvNet for Real-Time Semantic Segmentation”. In: *IEEE Trans. Intell. Transp. Syst.* 19.1 (2018), pp. 263–272 (cit. on p. 76).

[Ros+10] A. Rosenkrantz, Jeffry M Neil, X. Kong, J. Melamed, J. Babb, S. Taneja, and B. Taouli. “Prostate cancer: Comparison of 3D T2-weighted with conventional 2D T2-weighted imaging for image quality and tumor detection”. In: *AJR. Am. J. Roentgenol.* 194.2 (2010), pp. 446–52 (cit. on p. 76).

[Ros+13] Andrew B. Rosenkrantz, Sooah Kim, Ruth P. Lim, Nicole Hindman, Fang-Ming Deng, James S. Babb, and Samir S. Taneja. “Prostate Cancer Localization Using Multiparametric MR Imaging: Comparison of Prostate Imaging Reporting and Data System (PI-RADS) and Likert Scales”. In: *Radiology* 269.2 (2013), pp. 482–492 (cit. on p. 80).

[Ros+17] Andrew B. Rosenkrantz, Abimbola Ayoola, David Hoffman, Anunita Khasgiwala, Vinay Prabhu, Paul Smereka, Molly Somberg, and Samir S. Taneja. “The Learning Curve in Prostate MRI Interpretation: Self-Directed Learning Versus Continual Reader Feedback”. In: *AJR Am J Roentgenol* 208.3 (2017-03), W92–W100 (cit. on p. 13).

[Rou+19] Olivier Rouvière, Philippe Puech, Raphaële Renard-Penna, Michel Claudon, Catherine Roy, Florence Mège-Lechevallier, Myriam Decaussin-Petrucci, Marine Dubreuil-Chambardel, Laurent Magaud, Laurent Remontet, Alain Ruffion, Marc Colombel, Sébastien Crouzet, Anne-Marie Schott, Laurent Lemaitre, Muriel Rabilloud, Nicolas Grenier, and MRI-FIRST Investigators. "Use of Prostate Systematic and Targeted Biopsy on the Basis of Multiparametric MRI in Biopsy-Naive Patients (MRI-FIRST): A Prospective, Multicentre, Paired Diagnostic Study". In: *Lancet Oncol* 20.1 (2019-01), pp. 100–109 (cit. on pp. 8, 74).

[Rou+22] Olivier Rouvière, Tristan Jaouen, Pierre Baseilhac, Mohammed Lamine Benomar, Raphael Escande, Sébastien Crouzet, and Rémi Souchon. "Artificial Intelligence Algorithms Aimed at Characterizing or Detecting Prostate Cancer on MRI: How Accurate Are They When Tested on Independent Cohorts? – A Systematic Review". In: *Diagnostic and Interventional Imaging* (2022-12) (cit. on p. 9).

[Roz+20] F. Rozet, P. Mongiat-Artus, C. Hennequin, J. B. Beauval, P. Beuzeboc, L. Cormier, G. Fromont-Hankard, R. Mathieu, G. Ploussard, R. Renard-Penna, I. Brenot-Rossi, F. Bruyere, A. Cochet, G. Crehange, O. Cussenot, T. Lebret, X. Rebillard, M. Soulié, L. Brureau, and A. Méjean. "[French ccAFU guidelines - update 2020-2022: prostate cancer]". In: *Progres En Urologie: Journal De l'Association Francaise D'urologie Et De La Societe Francaise D'urologie* 30.12S (2020-11), S136–S251 (cit. on pp. 26, 149).

[RT14] Andrew B. Rosenkrantz and Samir S. Taneja. "Radiologist, Be Aware: Ten Pitfalls That Confound the Interpretation of Multiparametric Prostate MRI". In: *American Journal of Roentgenology* 202.1 (2014-01), pp. 109–120 (cit. on p. 38).

[Run+19] L. Rundo, C. Han, Y. Nagano, J. Zhang, R. Hataya, C. Militello, A. Tangherloni, M.S. Nobile, C. Ferretti, D. Besozzi, M.C. Gilardi, S. Vitabile, G. Mauri, H. Nakayama, and P. Cazzaniga. "USE-Net: Incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets". In: *Neurocomputing* 365 (2019), pp. 31–43 (cit. on pp. 76, 84, 111, 150, 153, 155, 156, 162, 164, 165).

[Run+20] Leonardo Rundo, Changhee Han, Jin Zhang, Ryuichiro Hataya, Yudai Nagano, Carmelo Militello, Claudio Ferretti, Marco S. Nobile, Andrea Tangherloni, Maria Carla Gilardi, Salvatore Vitabile, Hideki Nakayama, and Giancarlo Mauri. "CNN-Based Prostate Zonal Segmentation on T2-Weighted MR Images: A Cross-Dataset Study". In: *Neural Approaches to Dynamics of Signal Exchanges*. Ed. by Anna Esposito, Marcos Faundez-Zanuy, Francesco Carlo Morabito, and Eros Pasero. Smart Innovation, Systems and Technologies. Singapore: Springer, 2020, pp. 269–280 (cit. on pp. 153–156, 162, 164, 165).

[Sab+10] Mert R. Sabuncu, B. T. Thomas Yeo, Koen Van Leemput, Bruce Fischl, and Polina Golland. "A Generative Model for Image Segmentation Based on Label Fusion". In: *IEEE Transactions on Medical Imaging* 29.10 (2010), pp. 1714–1729 (cit. on p. 48).

[Sab+21] S. Sabater, M. R. Pastor-Juan, I. Andres, L. López-Martinez, V. Lopez-Honrubia, M. I. Tercero-Azorin, M. Sevillano, E. Lozano-Setien, E. Jimenez-Jimenez, R. Berenguer, A. Rovirosa, S. Castro-Larefors, M. Magdalena Marti-Laosa, O. Roche, F. Martinez-Terol, and M. Arenas. "MRI Prostate Contouring Is Not Impaired by the Use of a Radiotherapy Image Acquisition Set-up. An Intra- and Inter-Observer Paired Comparative Analysis with Diagnostic Set-up Images". In: *Cancer Radiotherapie: Journal De La Societe Francaise De Radiotherapie Oncologique* 25.2 (2021-04), pp. 107–113 (cit. on p. 143).

[Sah+22] Anido Saha, Jasper J. Twilt, Joeran S. Bosma, Bram van Ginneken, Derya Yakar, Mattijs Elschot, Jeroen Veltman, Jurgen Fütterer, Maarten de Rooij, and Henkjan Huisman. *Artificial Intelligence and Radiologists at Prostate Cancer Detection in MRI: The PI-CAI Challenge*. 2022 (cit. on pp. 15, 103, 106, 118, 126).

[Sal+22] Massimo Salvi, Bruno De Santi, Bianca Pop, Martino Bosco, Valentina Giannini, Daniele Regge, Filippo Molinari, and Kristen M. Meiburger. "Integration of Deep Learning and Active Shape Models for More Accurate Prostate Segmentation in 3D MR Images". In: *Journal of Imaging* 8.5 (2022-05), p. 133 (cit. on p. 143).

[San+20a] Thomas Sanford, Stephanie A. Harmon, Evrim B. Turkbey, Deepak Kesani, Sena Tuncer, Manuel Madariaga, Chris Yang, Jonathan Sackett, Sherif Mehralivand, Pingkun Yan, Sheng Xu, Bradford J. Wood, Maria J. Merino, Peter A. Pinto, Peter L. Choyke, and Baris Turkbey. "Deep-Learning-Based Artificial Intelligence for PI-RADS Classification to Assist Multiparametric Prostate MRI Interpretation: A Development Study". In: *Journal of Magnetic Resonance Imaging* 52.5 (2020-11), pp. 1499–1507 (cit. on p. 127).

[San+20b] Thomas H. Sanford, Ling Zhang, Stephanie A. Harmon, Jonathan Sackett, Dong Yang, Holger Roth, Ziyue Xu, Deepak Kesani, Sherif Mehralivand, Ronaldo H. Baroni, Tristan Barrett, Rossano Girometti, Aytekin Oto, Andrei S. Purysko, Sheng Xu, Peter A. Pinto, Daguang Xu, Bradford J. Wood, Peter L. Choyke, and Baris Turkbey. "Data Augmentation and Transfer Learning to Improve Generalizability of an Automated Prostate Segmentation Model". In: *American Journal of Roentgenology* 215.6 (2020-12), pp. 1403–1410 (cit. on pp. 154, 162, 164, 165).

[San+22] Jeremiah W. Sanders, Henry Mok, Alexander N. Hanania, Aradhana M. Venkatesan, Chad Tang, Teresa L. Bruno, Howard D. Thames, Rajat J. Kudchadker, and Steven J. Frank. "Computer-Aided Segmentation on MRI for Prostate Radiotherapy, Part I: Quantifying Human Interobserver Variability of the Prostate and Organs at Risk and Its Impact on Radiation Dosimetry". In: *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology* 169 (2022-04), pp. 124–131 (cit. on pp. 134, 143).

[San22] Inês C. Santiago. *Urology: the last review: Rapid-revision for FEBU and European board exams*. Knowuro, 2022 (cit. on p. 129).

[Sar+11] S. Sara Mahdavi, Nick Chng, Ingrid Spadinger, William J. Morris, and Septimiu E. Salcudean. "Semi-automatic segmentation for prostate interventions". In: *Med. Image Anal.* 15.2 (2011), pp. 226–237 (cit. on p. 74).

[SCN21] Julio Silva-Rodríguez, Adrián Colomer, and Valery Naranjo. "WeGleNet: A Weakly-Supervised Convolutional Neural Network for the Semantic Segmentation of Gleason Grades in Prostate Histology Images". In: *Computerized Medical Imaging and Graphics* 88 (2021-03), p. 101846 (cit. on p. 104).

[Scr+16] Luca Scrucca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. "Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models". In: *The R Journal* 8.1 (2016-08), pp. 289–317 (cit. on p. 25).

[Sea+93] E. Seaman, M. Whang, C. A. Olsson, A. Katz, W. H. Cooner, and M. C. Benson. "PSA Density (PSAD). Role in Patient Evaluation and Management". In: *The Urologic clinics of North America* 20.4 (1993-11), pp. 653–663 (cit. on p. 20).

[Sei+22] Constantin Marc Seibold, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. "Reference-Guided Pseudo-Label Generation for Medical Semantic Segmentation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.2 (2022-06), pp. 2171–2179 (cit. on p. 104).

[Sel+20] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *International Journal of Computer Vision* 128.2 (2020-02), pp. 336–359 (cit. on pp. 104, 127).

[SG02] Mikkel B Stegmann and David Delgado Gomez. *A Brief Introduction to Statistical Shape Analysis*. 2002 (cit. on p. 119).

[Sha+14a] Maysam Shahedi, Derek W. Cool, Cesare Romagnoli, Glenn S. Bauman, Matthew Bastian-Jordan, Eli Gibson, George Rodrigues, Belal Ahmad, Michael Lock, Aaron Fenster, and Aaron D. Ward. "Spatially Varying Accuracy and Reproducibility of Prostate Segmentation in Magnetic Resonance Images Using Manual and Semiautomated Methods". In: *Medical Physics* 41.11 (2014-11), p. 113503 (cit. on p. 143).

[Sha+14b] Gregory Sharp, Karl D. Fritscher, Vladimir Pekar, Marta Peroni, Nadya Shusharina, Harini Veeraraghavan, and Jinzhong Yang. "Vision 20/20: Perspectives on Automated Image Segmentation for Radiotherapy". In: *Medical Physics* 41.5 (2014-05), p. 050902 (cit. on p. 142).

[Sha+16] Maysam Shahedi, Derek W. Cool, Cesare Romagnoli, Glenn S. Bauman, Matthew Bastian-Jordan, George Rodrigues, Belal Ahmad, Michael Lock, Aaron Fenster, and Aaron D. Ward. "Postediting Prostate Magnetic Resonance Imaging Segmentation Consistency and Operator Time Using Manual and Computer-Assisted Segmentation: Multiobserver Study". In: *Journal of Medical Imaging (Bellingham, Wash.)* 3.4 (2016-10), p. 046002 (cit. on p. 143).

[Sha+17] Maysam Shahedi, Derek Cool, Glenn Bauman, Matthew Bastian-Jordan, Aaron Fenster, and Aaron Ward. "Accuracy Validation of an Automated Method for Prostate Segmentation in Magnetic Resonance Imaging". In: *J. Digit. Imaging* 30.6 (2017), pp. 782–795 (cit. on pp. 25, 76, 99, 143).

[Sha+20] Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. *When and Why Test-Time Augmentation Works*. Arxiv eprints. 2020 (cit. on p. 87).

[SHH21] Anindo Saha, Matin Hosseinzadeh, and Henkjan Huisman. "End-to-End Prostate Cancer Detection in bpMRI via 3D CNNs: Effects of Attention Mechanisms, Clinical Priori and Decoupled False Positive Reduction". In: *Medical Image Analysis* 73 (2021-10), p. 102155 (cit. on pp. 84, 103, 111, 112).

[Sie+23] Rebecca L. Siegel, Kimberly D. Miller, Nikita Sandeep Wagle, and Ahmedin Jemal. "Cancer Statistics, 2023". In: *CA: A Cancer Journal for Clinicians* 73.1 (2023), pp. 17–48 (cit. on pp. 7, 74).

[Sil+19] Santiago Silva, Boris A. Gutman, Eduardo Romero, Paul M. Thompson, Andre Altmann, and Marco Lorenzi. "Federated Learning in Distributed Medical Databases: Meta-Analysis of Large-Scale Subcortical Brain Data". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019-04, pp. 270–274 (cit. on p. 127).

[Smi+19] Clayton P. Smith, Stephanie A. Harmon, Tristan Barrett, Leonardo K. Bittencourt, Yan Mee Law, Haytham Shebel, Julie Y. An, Marcin Czarniecki, Sherif Mehralivand, Mehmet Coskun, Bradford J. Wood, Peter A. Pinto, Joanna H. Shih, Peter L. Choyke, and Baris Turkbey. "Intra and Inter-Reader Reproducibility of PI-radsv2: A Multi-Reader Study". In: *J Magn Reson Imaging* 49.6 (2019-06), pp. 1694–1703 (cit. on pp. 12, 102).

[SMJ20] R.L. Siegel, K. D. Miller, and A. Jemal. "Cancer statistics, 2020". In: *CA Cancer J. Clin.* 70.1 (2020), pp. 7–30 (cit. on p. 102).

[Son+22] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. "Learning From Noisy Labels With Deep Neural Networks: A Survey". In: *IEEE Transactions on Neural Networks and Learning Systems* (2022), pp. 1–19 (cit. on p. 103).

[Sos+03] Jacob Sosna, Neil M. Rofsky, Sandra M. Gaston, William C. DeWolf, and Robert E. Lenkinski. "Determinations of Prostate Volume at 3-Tesla Using an External Phased Array Coil: Comparison to Pathologic Specimens". In: *Academic Radiology* 10.8 (2003-08), pp. 846–853 (cit. on pp. 21, 40, 41).

[Spä81] H Späth. "The Minisum Location Problem for the Jaccard Metric". In: *Operations-Research-Spektrum* 3 (1981), pp. 91–94 (cit. on p. 56).

[Sri+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *J. Mach. Learn. Res.* 15.1 (2014), 1929–1958 (cit. on p. 80).

[Sta+19] A. Stabile, F. Giganti, A. Rosenkrantz, S. Taneja, G. Villeirs, I. Gill, C. Allen, M. Emberton, C. Moore, and V. Kasivisvanathan. "Multiparametric MRI for prostate cancer diagnosis: current status and future directions". In: *Nat. Rev. Urol.* 17 (2019), pp. 41–61 (cit. on p. 74).

[Sud+17] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. "Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations". In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Ed. by M. Jorge Cardoso, Tal Arbel, Gustavo Carneiro, Tanveer Syeda-Mahmood, João Manuel R.S. Tavares, Mehdi Moradi, Andrew Bradley, Hayit Greenspan, João Paulo Papa, Anant Madabhushi, Jacinto C. Nascimento, Jaime S. Cardoso, Vasileios Belagiannis, and Zhi Lu. Cham: Springer International Publishing, 2017, pp. 240–248 (cit. on pp. 67, 80).

[Sui+14] Avan Suinesiaputra, Brett R. Cowan, Ahmed O. Al-Agamy, Mustafa A. Elattar, Nicholas Ayache, Ahmed S. Fahmy, Ayman M. Khalifa, Pau Medrano-Gracia, Marie-Pierre Jolly, Alan H. Kadish, Daniel C. Lee, Ján Margeta, Simon K. Warfield, and Alistair A. Young. "A collaborative resource to build consensus for automated left ventricular segmentation of cardiac MR images". In: *Med. Image Anal.* 18.1 (2014), pp. 50–62 (cit. on p. 79).

[Sun+22] Mohammed R. S. Sunoqrot, Anindo Saha, Matin Hosseinzadeh, Mattijs Elschot, and Henkjan Huisman. "Artificial Intelligence for Prostate MRI: Open Datasets, Available Applications, and Grand Challenges". In: *European Radiology Experimental* 6.1 (2022-08), p. 35 (cit. on pp. 103, 126).

[Tag+19] Mehdi Taghipour, Alireza Ziaei, Francesco Alessandrino, Elmira Hassanzadeh, Mukesh Harisinghani, Mark Vangel, Clare M. Tempany, and Fiona M. Fennessy. "Investigating the Role of DCE-MRI, over T2 and DWI, in Accurate PI-RADS v2 Assessment of Clinically Significant Peripheral Zone Prostate Lesions as Defined at Radical Prostatectomy". In: *Abdominal Radiology (New York)* 44.4 (2019-04), pp. 1520–1527 (cit. on p. 125).

[Tan+19] Zhixian Tang, Kun Chen, Mingyuan Pan, Manning Wang, and Zhijian Song. "An augmentation strategy for medical image processing based on statistical shape model and 3D thin plate spline for deep learning". In: *IEEE Access* 7 (2019), pp. 133111–133121 (cit. on p. 100).

[Tav+18] S. Tavolaro, P. Mozer, M. Roupret, E. Comperat, F. Rozet, E. Barret, S. Drouin, C. Vaessen, O. Lucidarme, O. Cussenot, F. Boudghène, and R. Renard-Penna. "Transition Zone and Anterior Stromal Prostate Cancers: Evaluation of Discriminant Location Criteria Using Multiparametric Fusion-Guided Biopsy". In: *Diagnostic and Interventional Imaging* 99.6 (2018-06), pp. 403–411 (cit. on p. 108).

[TH15] Abdel Aziz Taha and Allan Hanbury. "Metrics for Evaluating 3D Medical Image Segmentation: Analysis, Selection, and Tool". In: *BMC Medical Imaging* 15 (2015-08) (cit. on pp. 25, 136, 156).

[TH22] Baris Turkbey and Masoom A. Haider. "Artificial Intelligence for Automated Cancer Detection on Prostate MRI: Opportunities and Ongoing Challenges, From the *AJR* Special Series on AI Applications". In: *American Journal of Roentgenology* 219.2 (2022-08), pp. 188–194 (cit. on p. 9).

[Tia+18] Zhiqiang Tian, Lizhi Liu, Zhenfeng Zhang, and Baowei Fei. "PSNet: prostate segmentation on MRI based on a convolutional neural network". In: *J. Med. Imaging* 5.2 (2018-01), p. 021208 (cit. on p. 75).

[Tol+20] Marco Toldo, Andrea Maracani, Umberto Michieli, and Pietro Zanuttigh. "Unsupervised Domain Adaptation in Semantic Segmentation: A Review". In: *Technologies* 8.2 (2020-06), p. 35 (cit. on p. 126).

[Tot+13] Robert Toth, Justin Ribault, John Gentile, Dan Sperling, and Anant Madabhushi. "Simultaneous Segmentation of Prostatic Zones Using Active Appearance Models with Multiple Coupled Levelsets". In: *Computer Vision and Image Understanding* 117.9 (2013-09), pp. 1051–1060 (cit. on pp. 153, 155, 156, 162, 164).

[Tur+13] Baris Turkbey, Sergei V. Fotin, Robert J. Huang, Yin Yin, Dagane Daar, Omer Aras, Marcelino Bernardo, Brian E. Garvey, Juanita Weaver, Hrishikesh Haldankar, Naira Muradyan, Maria J. Merino, Peter A. Pinto, Senthil Periaswamy, and Peter L. Choyke. "Fully Automated Prostate Segmentation on MRI: Comparison with Manual Segmentation Methods and Specimen Volumes". In: *AJR. American journal of roentgenology* 201.5 (2013-11), W720–729 (cit. on pp. 21, 41).

[Tur+19] B. Turkbey, A.B. Rosenkrantz, M.A. Haider, A.R. Padhani, G. Villeirs, K.J. Macura, C.M. Tempany, P.L. Choyke, F. Cornud, D.J. Margolis, H.C. Thoeny, S. Verma, J. Barentsz, and J.C. Weinreb. "Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2". In: *Eur. Urol.* 76.3 (2019), pp. 340–351 (cit. on pp. 8, 10, 20, 23, 38, 40, 74, 100, 102, 106, 108, 150, 164).

[UVL17] D. Ulyanov, A. Vedaldi, and V. Lempitsky. *Instance Normalization: The Missing Ingredient for Fast Stylization*. Arxiv eprints. 2017 (cit. on p. 80).

[Van+13] Joris Van de Velde, Emmanuel Audenaert, Bruno Speleers, Tom Vercauteren, Thomas Mulliez, Pieter Vandemaele, Eric Achten, Ingrid Kerckaert, Katharina D'Herde, Wilfried De Neve, and Tom Van Hoof. "An Anatomically Validated Brachial Plexus Contouring Method for Intensity Modulated Radiation Therapy Planning". In: *International Journal of Radiation Oncology*Biology*Physics* 87.4 (2013), pp. 802–808 (cit. on p. 132).

[Ven+21] Coen de Vente, Pieter Vos, Matin Hosseinzadeh, Josien Pluim, and Mitko Veta. "Deep Learning Regression for Prostate Cancer Detection and Grading in Bi-Parametric MRI". In: *IEEE Trans. Biomed. Eng.* 68.2 (2021-02), pp. 374–383 (cit. on pp. 103, 127).

[VGB12] G. Vincent, G. Guillard, and M. Bowes. "Fully Automatic Segmentation of the Prostate using Active Appearance Models". In: *MICCAI Grand Challenge: Prostate MR Image Segmentation 2012*. 2012 (cit. on p. 75).

[Vir+20] Pauli Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272 (cit. on p. 60).

[Vis+12] Satish E. Viswanath, Nicholas B. Bloch, Jonathan C. Chappelow, Robert Toth, Neil M. Rofsky, Elizabeth M. Genega, Robert E. Lenkinski, and Anant Madabhushi. "Central Gland and Peripheral Zone Prostate Tumors Have Significantly Different Quantitative Imaging Signatures on 3 Tesla Endorectal, in Vivo T2-weighted MR Imagery". In: *Journal of Magnetic Resonance Imaging* 36.1 (2012), pp. 213–224 (cit. on p. 108).

[Vrt+20] Tomaž Vrtovec, Domen Močnik, Primož Strojan, Franjo Pernuš, and Bulat Ibragimov. "Auto-Segmentation of Organs at Risk for Head and Neck Radiotherapy Planning: From Atlas-Based to Deep Learning Methods". In: *Medical Physics* 47.9 (2020), e929–e950 (cit. on p. 165).

[Wan+19a] Bo Wang, Yang Lei, Sibo Tian, Tonghe Wang, Yingzi Liu, Pretesh Patel, Ashesh B. Jani, Hui Mao, Walter J. Curran, Tian Liu, and Xiaofeng Yang. "Deeply Supervised 3D Fully Convolutional Networks with Group Dilated Convolution for Automatic MRI Prostate Segmentation". In: *Medical Physics* 46.4 (2019), pp. 1707–1718 (cit. on p. 21).

[Wan+19b] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks". In: *Neurocomputing* 338 (2019-04), 34–45 (cit. on p. 87).

[WBD98] Peter Willett, John M. Barnard, and Geoffrey M. Downs. "Chemical Similarity Searching". In: *Journal of Chemical Information and Computer Sciences* 38.6 (1998), pp. 983–996 (cit. on p. 56).

[Wei+16] Jeffrey C. Weinreb, Jelle O. Barentsz, Peter L. Choyke, Francois Cornud, Masoom A. Haider, Katarzyna J. Macura, Daniel Margolis, Mitchell D. Schnall, Faina Shtern, Clare M. Tempany, Harriet C. Thoeny, and Sadna Verma. "PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2". In: *European Urology* 69.1 (2016-01), pp. 16–40 (cit. on pp. 8, 9, 11, 40, 125, 127, 164).

[Wei99] Joachim Weickert. "Coherence-enhancing diffusion filtering". In: *Int. J. Comput. Vis.* 31.2 (1999), pp. 111–127 (cit. on p. 75).

[Wes+20] Antonio C. Westphalen et al. “Variability of the Positive Predictive Value of PI-RADS for Prostate MRI across 26 Centers: Experience of the Society of Abdominal Radiology Prostate Cancer Disease-focused Panel”. In: *Radiology* 296.1 (2020), pp. 76–84 (cit. on p. 74).

[WNS20] Neil F. Wasserman, Eric Niendorf, and Benjamin Spilseth. “Measurement of Prostate Volume with MRI (A Guide for the Perplexed): Biproximate Method with Analysis of Precision and Accuracy”. In: *Scientific Reports* 10.1 (2020-01), p. 575 (cit. on pp. 21, 23, 41, 42).

[Woo+18] Sungmin Woo, Chong Hyun Suh, Sang Youn Kim, Jeong Yeon Cho, Seung Hyup Kim, and Min Hoan Moon. “Head-to-Head Comparison Between Biparametric and Multiparametric MRI for the Diagnosis of Prostate Cancer: A Systematic Review and Meta-Analysis”. In: *American Journal of Roentgenology* 211.5 (2018-11), W226–W241 (cit. on p. 125).

[Wu+19] Kai Wu, Bowen Du, Man Luo, Hongkai Wen, Yiran Shen, and Jianfeng Feng. “Weakly Supervised Brain Lesion Segmentation via Attentional Representation Learning”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 211–219 (cit. on p. 104).

[Wu+22] Carine Wu, Sarah Montagne, Dimitri Hamzaoui, Nicholas Ayache, Hervé Delingette, and Raphaële Renard-Penna. “Automatic Segmentation of Prostate Zonal Anatomy on MRI: A Systematic Review of the Literature”. In: *Insights into Imaging* 13.1 (2022-12), p. 202 (cit. on pp. 14, 16, 142, 149).

[WZW04] Simon K. Warfield, Kelly H. Zou, and William M. Wells. “Simultaneous Truth and Performance Level Estimation (STAPLE): An Algorithm for the Validation of Image Segmentation”. In: *IEEE Trans Med Imaging* 23.7 (2004-07), pp. 903–921 (cit. on pp. 12, 25, 48, 49, 51, 69, 78, 137, 139).

[XT15] Saining Xie and Zhuowen Tu. “Holistically-Nested Edge Detection”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. 2015, pp. 1395–1403 (cit. on p. 75).

[Yan+18] Ziv Yaniv, Bradley C. Lowekamp, Hans J. Johnson, and Richard Beare. “SimpleITK Image-Analysis Notebooks: A Collaborative Environment for Education and Reproducible Research”. In: *Journal of Digital Imaging* 31.3 (2018-06), pp. 290–303 (cit. on pp. 23, 25).

[Yan+22] Hongxu Yang, Caifeng Shan, Alexander F. Kolen, and Peter H. N. de With. “Weakly-Supervised Learning for Catheter Segmentation in 3D Frustum Ultrasound”. In: *Computerized Medical Imaging and Graphics* 96 (2022-03), p. 102037 (cit. on p. 104).

[Yil+22] Kadir Yildirim, Muhammed Yildirim, Hasan Eryesil, Muhammed Talo, Ozal Yildirim, Murat Karabatak, Mehmet Sezai Ogras, Hakan Artas, and U Rajendra Acharya. “Deep Learning-Based PI-RADS Score Estimation to Detect Prostate Cancer Using Multiparametric Magnetic Resonance Imaging”. In: *Computers and Electrical Engineering* 102 (2022-09), p. 108275 (cit. on p. 127).

[Yin+12] Yin Yin, Sergei V. Fotin, Senthil Periaswamy, Justin Kunz, Hrishikesh Haldankar, Naira Muradyan, Baris Turkbey, and Peter Choyke. “Fully Automated 3D Prostate Central Gland Segmentation in MR Images: A LOGISMOS Based Approach”. In: *Medical Imaging 2012: Image Processing*. Vol. 8314. SPIE, 2012-02, pp. 952–960 (cit. on pp. 153, 162, 164).

[Yus+20] Igor Yusim, Muhammad Krenawi, Elad Mazor, Victor Novack, and Nicola J. Mabjeesh. “The Use of Prostate Specific Antigen Density to Predict Clinically Significant Prostate Cancer”. In: *Scientific Reports* 10.1 (2020-11), p. 20015 (cit. on p. 125).

[Zab+19] Fatemeh Zabihollahy, Nicola Schieda, Satheesh Krishna Jeyaraj, and Eranga Ukwatta. “Automated Segmentation of Prostate Zonal Anatomy on T2-weighted (T2W) and Apparent Diffusion Coefficient (ADC) Map MR Images Using U-Nets”. In: *Medical Physics* 46.7 (2019), pp. 3078–3090 (cit. on pp. 150, 153–156, 162, 164, 165).

[Zav+20] Olmo Zavala-Romero, Adrian L. Breto, Isaac R. Xu, Yu-Cherng C. Chang, Nicole Gautney, Alan Dal Pra, Matthew C. Abramowitz, Alan Pollack, and Radka Stoyanova. “Segmentation of Prostate and Prostate Zones Using Deep Learning: A Multi-MRI Vendor Analysis”. In: *Strahlentherapie und Onkologie* 196.10 (2020-10), pp. 932–942 (cit. on pp. 126, 150, 153–156, 162, 164, 165).

[Zha+19] Guokai Zhang, Weigang Wang, Dinghao Yang, Jihao Luo, Pengcheng He, Yongtong Wang, Ye Luo, Binghui Zhao, and Jianwei Lu. “A Bi-Attention Adversarial Network for Prostate Cancer Segmentation”. In: *IEEE Access* 7 (2019), pp. 131448–131458 (cit. on p. 84).

[Zha+20] Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Cicarrelli, Frederik Barkhof, and Daniel Alexander. “Disentangling Human Error from Ground Truth in Segmentation of Medical Images”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 15750–15762 (cit. on p. 48).

[Zho+16] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. “Learning Deep Features for Discriminative Localization”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, 2016-06, pp. 2921–2929 (cit. on p. 104).

[Zho+18] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. “UNet++: A Nested U-net Architecture for Medical Image Segmentation”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Ed. by Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R.S. Tavares, Andrew Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi. Cham: Springer International Publishing, 2018, pp. 3–11 (cit. on p. 103).

[Zhu+19] Yi Zhu, Rong Wei, Ge Gao, Lian Ding, Xiaodong Zhang, Xiaoying Wang, and Jue Zhang. “Fully Automatic Segmentation on Prostate MR Images Based on Cascaded Fully Convolution Network”. In: *Journal of Magnetic Resonance Imaging* 49.4 (2019), pp. 1149–1156 (cit. on pp. 150, 153, 162, 164).

[ZR+20] Olmo Zavala-Romero, Adrian L. Breto, Isaac R. Xu, Yu-Cherng C. Chang, Nicole Gautney, Alan Dal Pra, Matthew C. Abramowitz, Alan Pollack, and Radka Stoyanova. “Segmentation of prostate and prostate zones using deep learning: A multi-MRI vendor analysis”. In: *Strahlenther Onkol.* 196.10 (2020-10), pp. 932–942 (cit. on pp. 76, 119).

List of Figures

List of Tables