scientific reports

Check for updates

OPEN

Deep learning for fully automatic detection, segmentation, and Gleason grade estimation of prostate cancer in multiparametric magnetic resonance images

Oscar J. Pellicer-Valero[1]✉, José L. Marenco Jiménez[2], Victor Gonzalez-Perez[3], Juan Luis Casanova Ramón-Borja[2], Isabel Martín García[4], María Barrios Benito[4], Paula Pelechano Gómez[4], José Rubio-Briones[2], María José Rupérez[5] & José D. Martín-Guerrero[1]

Although the emergence of multi-parametric magnetic resonance imaging (mpMRI) has had a profound impact on the diagnosis of prostate cancers (PCa), analyzing these images remains still complex even for experts. This paper proposes a fully automatic system based on Deep Learning that performs localization, segmentation and Gleason grade group (GGG) estimation of PCa lesions from prostate mpMRIs. It uses 490 mpMRIs for training/validation and 75 for testing from two different datasets: ProstateX and Valencian Oncology Institute Foundation. In the test set, it achieves an excellent lesion-level AUC/sensitivity/specificity for the GGG≥2 significance criterion of 0.96/1.00/0.79 for the ProstateX dataset, and 0.95/1.00/0.80 for the IVO dataset. At a patient level, the results are 0.87/1.00/0.375 in ProstateX, and 0.91/1.00/0.762 in IVO. Furthermore, on the online ProstateX grand challenge, the model obtained an AUC of 0.85 (0.87 when trained only on the ProstateX data, tying up with the original winner of the challenge). For expert comparison, IVO radiologist's PI-RADS 4 sensitivity/specificity were 0.88/0.56 at a lesion level, and 0.85/0.58 at a patient level. The full code for the ProstateX-trained model is openly available at https://github.com/OscarPellicer/prostate_lesion_detection. We hope that this will represent a landmark for future research to use, compare and improve upon.

Prostate cancer (PCa) is the most frequently diagnosed malignancy in males in Europe and the USA and the second in the number of deaths[1]. Magnetic resonance imaging (MRI) is a medical imaging technique that employs very strong magnetic fields (typically 1.5–3T) to obtain three-dimensional (3D) images of the body; multi-parametric MRI (mpMRI) extends MRI by combining several MRI sequences into a multi-channel 3D image, each sequence providing different information on the imaged tissue. mpMRI has drastically changed the diagnostic approach of PCa: The traditional pathway includes screening based on the determination of prostate serum antigen (PSA) levels and digital rectal examination followed by a systematic random transrectal biopsy[2]. However, in recent years, the introduction of pre-biopsy mpMRI has enabled better selection of patients for prostate biopsy[3], increasing the diagnostic yield of the procedure[4] and allowing for more precise fusion-guided

[1]Intelligent Data Analysis Laboratory, Department of Electronic Engineering, ETSE (Engineering School), Universitat de València (UV), Av. Universitat, sn, 46100 Bujassot, Valencia, Spain. [2]Department of Urology, Fundación Instituto Valenciano de Oncología (FIVO), Beltrán Báguena, 8, 46009 Valencia, Spain. [3]Department of Medical Physics, Fundación Instituto, Valenciano de Oncología (FIVO), Beltrán Báguena, 8, 46009 Valencia, Spain. [4]Department of Radiodiagnosis, Fundación Instituto, Valenciano de Oncología (FIVO), Beltrán Báguena, 8, 46009 Valencia, Spain. [5]Instituto de Ingeniería Mecánica y Biomecánica, Universitat Politècnica de València (UPV), Camino de Vera, sn, 46022 Valencia, Spain. ✉email: Oscar.Pellicer@uv.es

biopsy examinations and focal therapies as compared with cognitive fusion approaches[5]. Additionally, mpMRI-derived parameters, such as tumor volume or PSA density (PSA divided by prostate volume) have proven helpful prognosis and stratification tools[6].

To promote global standardization in the interpretation of prostate mpMRI examinations, the Prostate Imaging Reporting and Data System (PI-RADS) in its latest 2.1 version combines available evidence to assign scores to objective findings in each sequence[7]. However, mpMRI interpretation is time-consuming, expertise dependent[8], and is usually accompanied by a non-negligible inter-observer variability[9]. This is particularly the case outside of expert high-volume centers[10]. Although promising alternative mpMRI scoring criteria are being developed, such as Likert[11], PI-RADS remains still the most widely used criterion for both clinical and academic purposes.

Computer-aided diagnosis (CAD) systems have been broadly defined as "the use of computer algorithms to aid the image interpretation process"[12]. In this sense, CAD is one of the most exciting lines of research in medical imaging and has been successfully applied to interpret images in different medical scenarios[13]. CAD poses several theoretical advantages, namely speeding up the diagnosis, reducing diagnostic errors, and improving quantitative evaluation[14]. On the topic of mpMRI-based PCa CAD, different methods have been proposed since the early 2000s[15]. These pioneered the field but were nonetheless limited in some important aspects (e.g. they lacked proper evaluation, expert comparison, and large enough datasets). In 2014, Litjens et al.[16] proposed the first CAD system able to provide candidate regions for lesions along with their likelihood for malignancy using pharmacokinetic and appearance-derived features from several MRI sequences using classical (non-Deep Learning) voxel-based classification algorithms and evaluated the results on a large cohort of 347 patients.

Since the advent of Deep Learning[17], however, Deep Convolutional Neural Networks (CNNs) have quickly dominated all kinds of image analysis applications (medical and otherwise), phasing out classical classification techniques. In the context of the prostate, the turning point can be traced back to the ProstateX challenge in 2016[16,18,19]. The challenge consisted in the classification of clinically significant PCa (csPCa) given some tentative locations on mpMRI. More importantly, a training set of 204 mpMRIs (330 lesions) was provided openly for training the models, hence enabling many researchers to venture into the problem (further details of this dataset can be found in "Data description" section). At the time, half of the contestants employed classical classification methods[20] and the other half CNNs[21]. In all cases, a patch (or region of interest, ROI) of the mpMRI around the lesion was extracted, and a machine learning algorithm was trained to classify it as either csPCa or not. The second-highest-scoring method[21], with a receiver operating characteristic -ROC- curve (AUC) of 0.84, used a simple VGG-like[22] CNN architecture trained over the mpMRI ROIs to perform classification. The main limitation of all these approaches is that ROIs have to be manually located beforehand (even after the model has been trained), hence limiting their interest and applicability to clinical practice.

In 2019, Cao et al.[23] employed a slice-wise segmentation CNN, FocalNet, not only to predict csPCa but also a to obtain a map of the Gleason grade group (GGG)[24,25] of the prostate. Very briefly, GGG is a standard 1–5 grading system for PCa, where GGG1 cancer cells look normal and are likely to grow slowly (if at all), while GGG5 cells look very abnormal and are likely to grow very quickly. Segmentation-based models are a step up from previous patch classification approaches because they provide a csPCa map of the prostate; however, they cannot directly identify lesions as individual entities and assign a score to each one, as is common procedure in clinical practice. This is natively solved in an instance detection+segmentation framework, which is very common in natural image detection tasks[26]; but has never been applied to csPCa detection. Additionally, two-dimensional (2D) slice-wise CNNs are known to generally underperform as compared with actual 3D CNNs in lesion detection tasks[27]. Indeed, in 2020 several authors turned to 3D CNNs, such as Arif et al.[28] or Aldoj et al.[29].

To the best of our knowledge, the model we propose is the first to leverage a proper instance detection and segmentation network, the 3D Retina U-Net[27], to simultaneously perform detection, segmentation, and Gleason Grade estimation from mpMRIs to a state-of-the-art performance level. It is also one of the few works that combines two very different mpMRI datasets into a single model: the ProstateX dataset and the IVO (Valencian Institute of Technology Foundation) dataset (view "Data description" section), achieving similarly excellent results in both. It uses prior prostate zonal segmentation information, which is provided by an automatic segmentation model, and leverages an automatic non-rigid MRI sequence registration algorithm, among other subsystems, allowing for a fully automatic system that requires no intervention. The code of this project has been made available online at https://github.com/OscarPellicer/prostate_lesion_detection.

Results

Lesion detection, segmentation, and classification. *Quantitative results.* A comprehensive quantitative evaluation of the trained model on the ProstateX and IVO test sets has been compiled in Table 1 (showing sensitivity and specificity) and in Supplementary Table 1 (showing positive predictive value and negative predictive value). The computation procedure for patient- and lesion-level metrics is explained in "Lesion matching and evaluation" section. For the evaluation of sensitivity and specificity, the model-predicted scores were thresholded at two working points (computed a posteriori on the test data): maximum sensitivity and balanced (similar sensitivity and specificity). Furthermore, radiologist-assigned pre-biopsy PI-RADS scores for all IVO patients with no missing sequences and with PI-RADS information available (N = 106 patients, 111 lesions) has also been included in Table 3 for comparison. Please notice that PI-RADS$\geq$ 3 is omitted since all IVO lesions were assigned at least a PI-RADS 3 score, and hence PI-RADS$\geq$ 3 acts just as a naïve classifier that considers all samples as positive (sensitivity 1 and specificity 0). A graphical representation of the area under the receiver operating characteristic (ROC) curve for the main significance criterion (GGG$\geq$2) can be found in Fig. 1. Also, Supplementary Table 2 uses a single threshold for all tests (but different for IVO and ProstateX datasets), computed a priori from the training data; this table might be a better proxy for the prospective performance of the model.

(Dataset) & Significance criterion	Level	AUC	Max. sensitivity			Balanced			PI-RADS≥4		PI-RADS=5	
			t	Sens.	Spec.	*t*	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
IVO GGG≥1	Lesion (13/33)	0.892	0.027	1.000	0.350	0.105	0.923	0.700	0.741	0.604	0.328	0.962
	Patient (15/30)	0.920	0.253	1.000	0.667	0.301	0.867	0.800	0.710	0.649	0.290	0.973
IVO GGG≥2	Lesion (8/33)	0.945	0.173	1.000	0.800	0.301	0.875	0.920	0.882	0.558	0.441	0.922
	Patient (9/30)	0.910	0.219	1.000	0.762	0.262	0.889	0.810	0.850	0.576	0.400	0.924
IVO GGG≥3	Lesion (3/33)	0.856	0.301	1.000	0.800	0.315	0.667	0.867	0.727	0.440	0.455	0.840
	Patient (3/30)	0.840	0.301	1.000	0.778	0.315	0.667	0.852	0.727	0.432	0.455	0.832
ProstateX GGG≥1	Lesion (17/69)	0.898	0.028	0.941	0.788	0.053	0.824	0.865	–	–	–	–
	Patient (16/45)	0.866	0.108	1.000	0.138	0.104	0.938	0.655	–	–	–	–
ProstateX GGG≥2	Lesion (13/69)	0.959	0.028	1.000	0.786	0.108	0.923	0.911	–	–	–	–
	Patient (13/45)	0.865	0.028	1.000	0.375	0.108	0.923	0.688	–	–	–	–
ProstateX GGG≥3	Lesion (7/69)	0.751	0.195	0.714	0.887	0.195	0.714	0.887	–	–	–	–
	Patient (7/45)	0.767	0.016	1.000	0.395	0.026	0.857	0.500	–	–	–	–

Table 1. Quantitative results for IVO (top) and ProstateX (bottom) test data evaluated with different Gleason Grade Group (GGG) significance criteria (e.g. lesions with GGG≥1, 2, or 3 are considered positive), at lesion- and patient-level ($N_{positives}/N_{total}$), and at two thresholds (t): maximum sensitivity and balanced. For IVO data, results are compared with radiologist-assigned pre-biopsy PI-RADS scores for all IVO patients with no missing sequences and with PI-RADS information available (N=106 patients, 111 lesions). AUC: Area under the ROC curve.

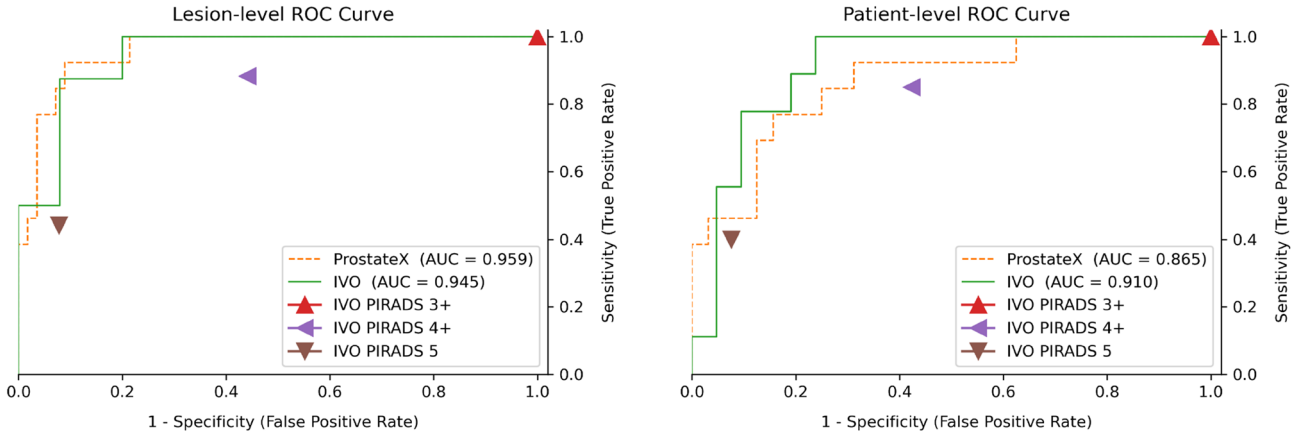

Figure 1. ROC curve of the model for significance criterion Gleason Grade Group ≥2, evaluated at the lesion level (left) and the patient level (right). For comparison, triangular marks represent the radiologist-assigned pre-biopsy PI-RADS. AUC: area under the ROC curve.

Focusing on the results for the GGG≥2 significance criterion, at the highest sensitivity working point, the model achieves a perfect lesion-level sensitivity of 1 (no csPCa is missed) and a specificity of 0.786 and 0.875 for ProstateX and IVO, respectively (AUCs: 0.959 and 0.945). At the patient level, the specificity falls to 0.375 and 0.762 for each dataset (AUCs: 0.865 and 0.910).

For the GGG≥1 significance criterion, the model achieves a lesion-/patient-level maximum sensitivity of 0.941 (spec. 0.788)/1 (spec. 0.138) in the ProstateX dataset, and a maximum sensitivity of 1 (spec. 0.350)/1 (spec. 0.667) in the IVO dataset. In summary, no GGG≥1 patient was missed, although at a cost of low specificity. Using the GGG≥3 significance criterion the model reaches a lesion- and patient-level sensitivity of 0.714 (spec. 0.887)/1 (spec.: 0.395) in the ProstateX dataset, and a maximum sensitivity of 1 (spec. 0.800)/1 (spec. 0.778) in the IVO dataset.

Regarding lesion segmentation performance, the mean DSC across all patients for segmenting any type of lesion irrespective of their GGG (including GGG0 benign lesions), was 0.276/0.255 for the IVO/ProstateX dataset when evaluated at the 0.25 segmentation threshold, and 0.245/0.244 when evaluated at 0.5.

Qualitative results. Figure 2 shows the output of the model evaluated on two IVO test patients and three ProstateX test patients. For the sake of clarity, GGG0 (benign) bounding boxes (BBs) are not shown and, for highly overlapped detections (Intersection over Union, -IoU- > 0.25), only the highest-scoring BB is drawn. Detections with confidence below the GGG≥2 lesion-wise maximum sensitivity threshold (0.173 for IVO, and 0.028

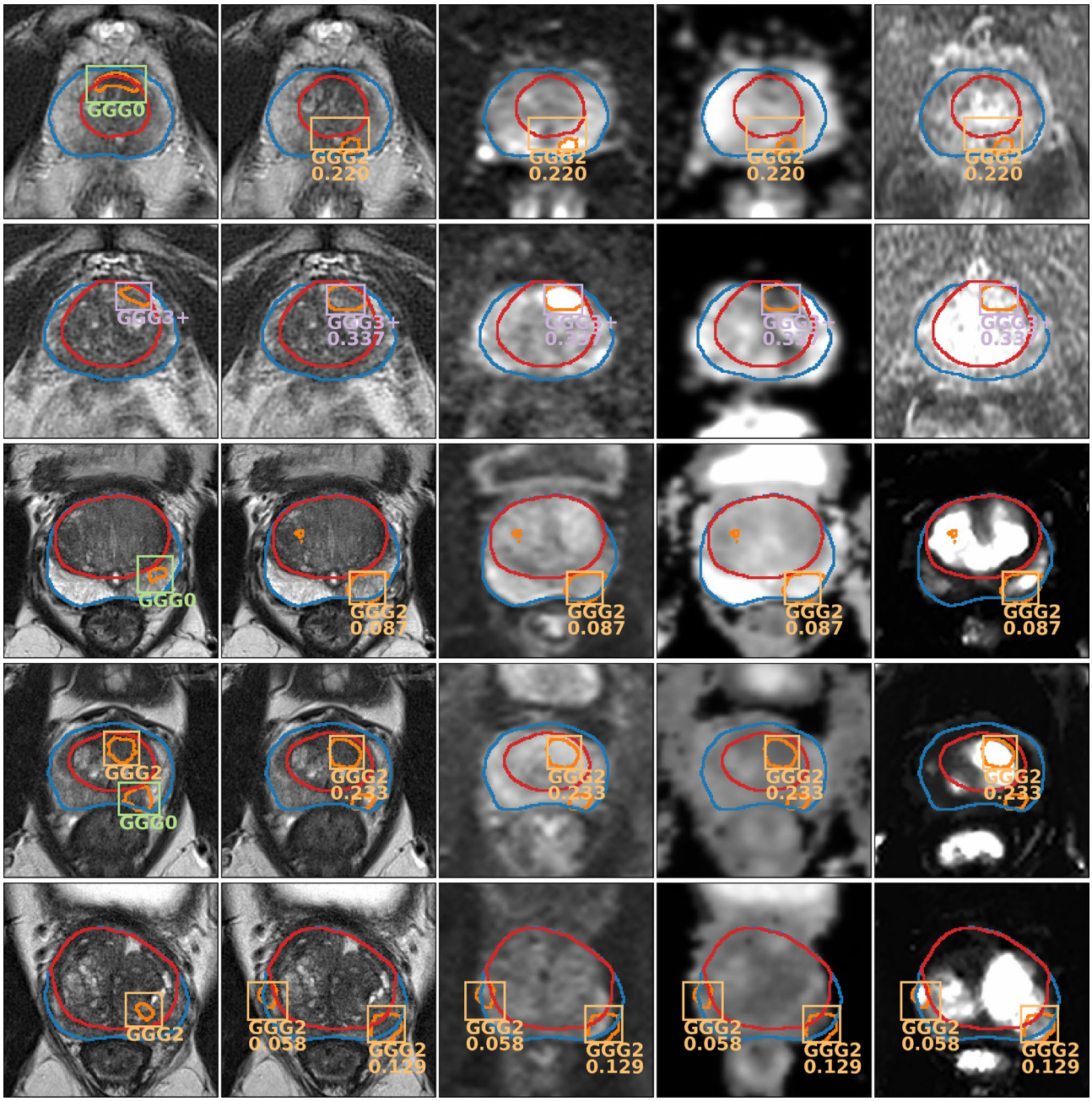

Figure 2. Output of the model (every row corresponds to a different patient) evaluated on two IVO test patients (first two rows) and three ProstateX test patients (last three rows). For each patient, first image from the left shows the ground truth on the T2 sequence; the rest show the output predictions of the model on different sequences (from left to right: T2, b800, ADC, K^{trans} -IVO- / DCE $t = 30$ -ProstateX-). Gleason Grade Group (GGG) 0 -benign- bounding boxes (BBs) are not shown and only the highest-scoring BB is shown for sets of highly overlapped detections (intersection over union > 0.25). Detections with confidence below the GGG≥2 lesion-wise maximum sensitivity threshold (0.173 for IVO, and 0.028 for ProstateX) are not shown either.

for ProstateX) are not shown either. The first IVO patient (Fig. 2, row 1) is of special interest, as it is one of the relatively few IVO cases where the targeted biopsy did not find csPCa (as evidenced by the GGG0 BB in the GT image to the left), but the massive biopsy (20–30 cylinders) detected GGG2 csPCa. As can be seen, the model was able to detect this GGG2 lesion while ignoring the benign GGG0 one, hence outperforming the radiologists for this particular instance. For the second IVO patient (Fig. 2, row 2) a GGG3+ GT lesion (GGG4 specifically) was properly detected by the model with very high confidence.

The first ProstateX patient (Fig. 2, row 3) is a case of failure, where the model detects a non-existent GGG2 lesion, albeit with relatively low confidence; in fact, it would have been ignored at the balanced sensitivity setting ($t = 0.108$). For the next patient (Fig. 2, row 4), the model has been able to segment both GT lesions; however, only the csPCa lesion is detected, while the other is ignored (actually, the model correctly detected the other

MRI sequence dropped	ProstateX		IVO	
	Lesion	Patient	Lesion	Patient
None (Baseline)	0.959	0.865	0.945	0.910
b400/500	0.944	0.861	0.940	0.868
b800/1000/1400	0.946	0.873	0.895	0.783
All b-numbers	0.951	0.844	0.845	0.720
ADC	0.905	0.870	0.940	0.836
K^{trans}	0.894	0.865	–	–
All DCE	–	–	0.895	0.820
All but T2	0.804	0.808	0.782	0.545

Table 2. Area under the ROC curve after dropping one (or several) particular sequences (i.e.: setting the value to 0) in test time for the Gleason Grade Group ≥2 significance criterion.

Dataset	N	Mean DSC		
		Prost.	CG	PZ
Private train	80	0.941	0.935	0.866
Private test	12	0.915	0.915	0.833
NCI-ISBI train	60	0.894	0.860	0.690

Table 3. Results for the prostate zonal segmentation model. DSC:Sørensen-Dice similarity coefficient.

lesion as a GGG0, but BBs for those lesions are not shown). For the third patient (Fig. 2, row 5), the model could correctly identify the GGG2 GT lesion but also identified an additional GGG2 lesion. This might be a mistake or might show a real lesion that was missed by the radiologists (we cannot know, as no massive biopsy information is available for the ProstateX dataset). Due to this uncertainty, lesion-level evaluation should not penalize detections for which GT information was not available (such as this one), as discussed in "Lesion matching and evaluation" section.

Sequence ablation tests. In "Model training and validation" section, Random Channel Drop is presented as a training-time data augmentation technique that should help alleviate the problem of missing sequences. For a model trained in such a fashion, we can assess the individual importance of the different sequences by dropping them (i.e.: setting them to 0) at test time and analyzing the performance penalty that the model incurs. The AUCs after dropping different sequences (or combinations of them) are shown in Table 2.

As can be seen, removing the low b-valued (b400 for ProstateX/b500 for IVO) DW sequence seems to have minimal impact on both datasets, as is to be expected. Conversely, while removing the high b-valued (b800 for ProstateX/b1000 or b1400 for IVO) DW sequences has little impact on the ProstateX data, it severely affects the performance on the IVO data, likely due to the higher b values employed in this dataset (which may prove more informative). Furthermore, removing all DW sequences severely affects the IVO dataset, but has almost no impact on ProstateX. The removal of the ADC map has a similar negative impact on both datasets, although the results vary depending on how they are analyzed (lesion- or patient-wise). Likewise, dropping the K^{trans} sequence on the ProstateX data or the DCE sequences on the IVO data clearly harms the performance. For the final test, all sequences are dropped except for the T2; despite it, the model still has a commendable performance, especially in the ProstateX set, which might indicate that the proposed Random Channel Drop augmentation has served its purpose of making the model more robust to missing sequences.

Prostate zonal segmentation. Regarding the prostate zonal segmentation model, which was developed with the sole purpose of automating the PCa detection system (view "Pre-processing" section), the results for all datasets can be found in Table 3, with mean Sørensen-Dice similarity coefficient (DSC) ranging from 0.894 to 0.941. DSC is a metric between 0 and 1, employed to assess the relative overlap between predicted and ground truth (GT) segmentations. Some qualitative results for this segmentation model can be seen in Figs. 2, 4, and 5.

Discussion

Despite mpMRI interpretation being time-consuming and observer-dependent, it is a major clinical decision driver and poses great clinical relevance. In this paper we presented a CAD system developed with two main MRI datasets integrating T2, DW, b-value, and ADC maps in both of them as well as K^{trans} for ProstateX and DCE for the IVO dataset. These were compared against fusion and transperineal template biopsies, which is considered the pre-operative gold standard to evaluate prostate cancer extent[30].

Different outcomes can be measured for this system. Regarding lesion detection as exposed in "Lesion detection, segmentation, and classification" section, the results for lesions GGG≥2 significance criterion can

be considered optimal: all csPCa lesions were detected while maintaining a very high specificity, except for the patient-level ProstateX evaluation, and a great AUC ranging from 0.865 to 0.959. Furthermore, the IVO results outperform the PI-RADS scores, especially at the high sensitivity setting (PI-RADS$\geq$ 4) which is of most interest in clinical practice. This can be seen in Fig. 1, where the ROC is at all instances above and to the left of the PI-RADS scores. For further comparison, several studies have reported radiologist sensitivities/specificities for the detection of csPCa from mpMRI at a patient level of 0.93/0.41[4], or 0.58–0.96/0.23–87 as shown in a systematic review[31]. The results vary wildly due to their single-center nature, their differing criteria for the definition of csPCa, and the often-inaccurate reference standards employed.

Considering GGG$\geq$3 significance criterion, caution is required when interpreting these results due to the very low number of positive cases (e.g.: only three in the IVO test set). Furthermore, the 0.714 patient-level sensitivity does not mean that the model missed GGG3 lesions, but rather that they were assigned to a lower GGG (such as GGG2) and were therefore ignored for the GGG$\geq$3 classification problem.

In addition to the previous tests, the ongoing ProstateX challenge was used for external lesion-level validation, achieving an AUC of 0.85, which would have been the second-best AUC in the original ProstateX challenge[19]. Additionally, an identical model trained only on the ProstateX data (which has been made publicly available alongside this paper), achieved an AUC of 0.87, which would have tied with the best contender in the challenge. There are now higher AUCs in the online leaderboard but, unfortunately, we were unable to find any publications regarding them, and hence no further analysis can be performed. In any case, these results must be also interpreted with caution: on one hand, the proposed system solves a much more complex problem (namely detection, segmentation & classification) than the comparatively simpler ROI classification systems which are typically employed for this task, and it is therefore in a disadvantage compared to them. On the other hand, as indicated in "Data description" section, the ProstateX challenge mpMRIs were used for training the segmentation and detection components of the model, but not the classification head (as GGG information is kept secret by the challenge, and hence unavailable for training). The inclusion of this data was useful for increasing the number of training samples, but it might have introduced some unknown bias for the evaluation of this dataset.

Outside the ProstateX challenge, one of the very first works on the topic by Litjens et al.[18] reported a sensitivity of 0.42, 0.75, and 0.89 at 0.1, 1, and 10 false positives per normal case using a classical radiomics-based model. More recently, Xu et al.[32] used a csPCa segmentation CNN whose output was later matched to GT lesions based on distance (similar to ours). He reported a sensitivity of 0.826 at some unknown specificity; also, despite using the ProstateX data, unfortunately, no ProstateX challenge results were provided. Cao et al.[23] proposed a segmentation CNN that also included GGG classification as part of its output, reporting a maximum sensitivity of 0.893 at 4.64 false positives per patient and an AUC of 0.79 for GGG$\geq$2 prediction. Interestingly, the authors employed histopathology examinations of whole-mount specimens as GT for the model. Aldoj et al.[29] utilized the ProstateX data to perform csPCa classification on mpMRI ROIs around the provided lesion positions, reporting an AUC of 0.91 on their internal 25-patient test set; once again, despite using the ProstateX data exactly as conceived for the challenge, they do not provide any challenge results for comparison.

In an interesting prospective validation study, Schelb et al.[33] obtained a sensitivity/specificity of 0.99/0.24 using a segmentation CNN, a performance that they found comparable to radiologist-derived PI-RADS scores. Woźnicki et al.[34] proposed a classical radiomics-based model (no CNNs involved) achieving an AUC of 0.807. As for patient-level csPCa classification results, Yoo et al.[35] achieved an AUC of 0.84 using slice-wise CNN classifier whose predictions were later combined into a patient-wise total score and Winkel et al.[36] achieved a sensitivity/specificity of 0.87/0.50 on a prospective validation study using a segmentation-based detection system which is most similar to the one proposed here.

Considering lesion segmentation concordance, as exposed in "Lesion detection, segmentation, and classification" section, our results are unfortunately not directly comparable to other papers in the literature (as those focus on segmenting exclusively csPCa and benign lesions are ignored) and were mostly added for completeness. For instance, Schelb et al.[33] reported a DSC of 0.34 for csPCa segmentation, similar to Vente et al.[37]'s 0.37 DSC. Secondly, the reference segmentations for the ProstateX dataset were generated in an automatic manner; hence, the performance for this dataset is not compared against a proper ground truth. Thirdly, mpMRI lesions tend to be small with ill-defined margins and a very high inter-observer variability[38]. For all these reasons, these relatively low DSC metrics must be interpreted with caution. Instead, the previously discussed metrics provide a more objective outlook on the actual performance of the model.

With respect to the ablation tests, there is an ongoing debate regarding the need for DCE sequences. Biparametric MRI (bpMRI) (without DCE sequences) seems to be a more cost- and time-effective alternative to mpMRI, with little detriment to accuracy[39,40]. Likewise, the role of DCE sequences is currently minor in the final score of the PI-RADS system, being used only in peripheral zone regions with value 3 in the DW sequence (which rises to 4 if an early focal uptake is detected in DCE sequences). Conversely, the results of the present study hint towards a greater importance of DCE sequences, which turned out to be the second most important sequences for the model, only behind b-numbers (T2 does not count as it was always included).

Lastly, regarding prostate zonal segmentation, we observed a great concordance between the model's and expert radiologist's prostate segmentation with a DSC that ranged from 0.894 to 0.941 depending on the MRI dataset. As can be seen, the results in the Private test set are extremely good, better in fact than any other model in the literature when evaluated in its internal test set and when evaluated blindly in the NCI-ISBI dataset. In Qin et al.[41], for instance, the authors train one CNN on an internal dataset and another identical CNN on the NCI-ISBI train dataset independently, and evaluate them by cross-validation, achieving a DSC of 0.908 and 0.785 at the CG and PZ in their internal dataset, and a DSC of 0.901 and 0.806 in the NCI-ISBI dataset. For a fairer comparison with our model, in Rundo et al.[42], the authors train their model on two internal datasets (achieving a DSC of 0.829/0.859 in CG segmentation, and 0.906/0.829 in PZ segmentation), which then test blindly in the NCI-ISBI dataset, achieving 0.811 and 0.551 in CG and PZ segmentation, respectively. Finally, Aldoj et al.[43],

training on a larger cohort of 141 patients and evaluating in their internal test set of 47, achieved a DSC of 0.921, 0.895, and 0.781 for whole gland, CG, and PZ segmentation.

The interpretation of mpMRIs based on Artificial Intelligence (AI) represents a very promising line of research that has already been successfully applied to prostate gland segmentation and PCa lesion detection using both transperineal prostate biopsy and radical prostatectomy specimens as GT with varying results[35,36]. We went a step further and developed the first algorithm, to the best of our knowledge, that automatically contours the prostate into its zones, performs well at lesion detection and Gleason Grade prediction (identifying lesions of a given grade or higher), and segments such lesions albeit with a moderate overlapping. The model outperformed expert radiologists with extensive MRI experience and achieved top results in the ProstateX challenge.

The code has been made publicly available, including an automatic prostate mpMRI non-rigid registration algorithm and an automatic mpMRI lesion segmentation model. Most importantly, the fact that the code is online might allow future researchers to use this model as a reference upon which to build or to compare their models.

Our work presents some limitations. Firstly, further validation and prospective blinded trial would be required to compare histological results of targeted biopsies to the lesions identified by the model. Secondly, although the model was successfully trained on two datasets, it still behaves differently on each of them (e.g.: the optimal thresholds vary significantly between them), which is not desirable, but probably unavoidable. Obviously, more data from sources as varied as possible would be ideal to overcome such difficulties and further improve the performance and generality of the model. Thirdly, AI systems have proven cumbersome to integrate into clinical practice for a variety of reasons (costs, rejection, etc.); we hope that by making the code freely available some of these obstacles can be more easily overcome.

In any case, this is yet another step in the foreseeable direction of developing a strong collaborative AI net that progressively incorporates as many mpMRIs with the corresponding GT as possible. The clinical applications of this model are countless, amongst which we could consider assisting radiologists by speeding up prostate segmentation, training purposes as well as a safety net to avoid missing PCa lesions. Further, the ability to detect csPCa can easily highlight which MRIs would require prompt reporting and prioritizing biopsy. Moreover, given the recent trend towards conservative PCa approaches such as focal therapy or active surveillance (usually implying a more dedicated prostate biopsy), predicting the Gleason Grade, as well as the number of lesions pre-biopsy, could identify eligible men that could be offered transperineal targeted biopsy in the first place.

Materials and methods

Data description. For the development and validation of the model, two main prostate mpMRI datasets were employed: ProstateX[16], which is part of an ongoing online challenge at https://prostatex.grand-challenge.org and is freely available for download[18]; and IVO, from the homonymous Valencian Institute of Oncology. The study was approved by the Ethical Committee of the Valencian Institute of Oncology (CEIm-FIVO) with protocol code PROSTATEDL (2019-12) and date 17th of July, 2019. All experiments were performed in accordance with relevant guidelines and regulations. Informed consent was obtained from all participants and/or their legal guardians.

For ProstateX, the data consisted of a total of 204 mpMRIs (one per patient) including the following sequences: T2-weighted (T2), diffusion-weighted (DW) with b-values b50, b400, and b800 s/mm^2, apparent diffusion coefficient (ADC) map (calculated from the b-values), and K^{trans} (computed from dynamic contrast-enhanced -DCE-T1-weighted series). For each of these patients, one to four (1.62 per patient on average) lesion locations (i.e.: a point marking their position) and their GGG are provided (GGG is provided as part of the ProstateX2 challenge, which shares the same data with ProstateX). The lesion locations were reported by or under the supervision of an expert radiologist with more than 20 years of experience in prostate MR and confirmed by MR-guided biopsy. Furthermore, 140 additional mpMRIs are provided as part of the challenge set, including all previous information except for the GGG of the lesions. All mpMRIs were acquired by two different Siemens 3-Tesla scanners.

For IVO, there were a total of 221 mpMRIs, including the following sequences: T2, DW with b-values b100, b500, and b1000 s/mm^2 (in 1.36% of the cases, b1400 was available, instead of b1000), ADC (4.52% missing) and a temporal series of 30 DCE T1-weighted images (42.53% missing). For each mpMRI, one to two (1.04 per patient) lesions were segmented by one of several radiologists with two to seven years of experience in PCa imaging, and their PI-RADS were provided. The Gleason Score (GS)[24] was assessed by transperineal fusion-guided with two to three cylinders directed to each of the ROIs. Additionally all patients underwent systematic template biopsy comprising 20–30 cylinders to sample the rest of the prostate.

Four PCa classes were considered: GGG0 or benign (57.32% of all lesions), GGG1 (GS 3+3, 17.28%), GGG2 (GS 3+4, 12.70%), and GGG3+ (GS $\geq$ 4+3, 12.70%); therefore, lesions of GGG$\geq$3 were grouped into a single category to try to balance the classes, and also because the protocol for a suspect GGG 3+ lesion would be similar irrespective of its specific grade (i.e.: the lesion would be biopsied for confirmation).

Pre-processing. After collecting them, mpMRIs had to be pre-processed to accomplish three main objectives, namely: (1) homogenize differences within datasets, (2) homogenize differences between datasets, and (3) enrich the images with extra information that might be useful for the model. Additionally, the preprocessing pipeline was designed to require as little human intervention as possible, in pursuit of developing a system easily implementable in clinical practice.

For the first objective, all images were cropped to an ROI around the prostate of size $160 \times 160 \times 24$ voxels with a spacing of (0.5, 0.5, 3)mm, which corresponds with the median (and mode) spacing of the T2 sequences for both datasets. The rest of the sequences were applied the same processing for the sake of homogeneity. B-Spline interpolation of third order was employed for all image interpolation tasks, while Gaussian label interpolation was used for the segmentation masks. For the IVO dataset, the time series of 30 DCE images per

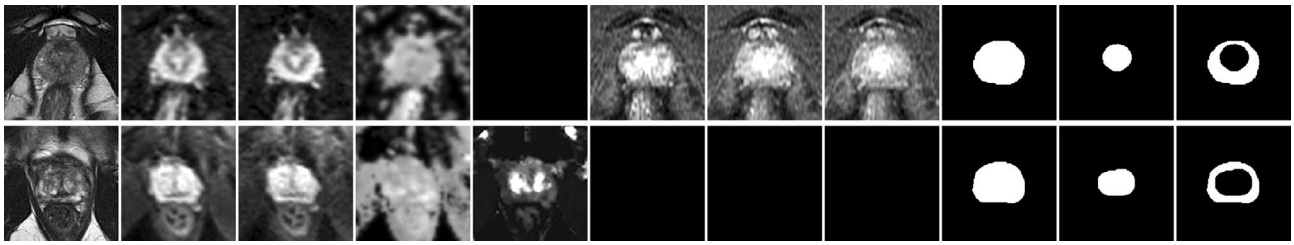

Figure 3. Final pre-processed image from a single patient (top: IVO, bottom: ProstateX). Channels (from left to right): T2, b400/b500, b800/b1000/b1400, ADC, K^{trans}, DCE $t = 10$, DCE $t = 20$, DCE $t = 30$, prostate mask, CG mask and PZ mask.

patient was sampled at times 10, 20, and 30, approximately coinciding with the peak, progression, and decay of the contrast agent. Then, all sequences were combined into a single multi-channel image, in which any missing sequences were left blanks (value of 0), such as the three DCE channels in every ProstateX image, or the K^{trans} channel in every IVO image. The intensity was normalized by applying Equation 1 to every channel of an image I independently, as introduced in Pellicer-Valero et al.[44].

$$I_{new} = \frac{I - percentile(I, 1)}{percentile(I, 99) - percentile(I, 1)} \tag{1}$$

Regarding objective (2), the procedure for homogenizing lesion representations between datasets is described in "Pre-processing" section, and a special data augmentation employed to alleviate the problem of missing sequences is presented in "Model training and validation" section. Additionally, sequences b500 (from IVO) and b400 (from ProstateX) were considered similar enough to conform to the same channel in the final image; likewise, sequences b1000/b1400 (from IVO) and b800 (from ProstateX) were assigned to a single common channel too.

Concerning objective (3), "Pre-processing" section argues that prostate zonal segmentation is an important input for PCa assessment and describes the conception of a model for producing such segmentations automatically. Additionally, DW and ADC sequences were found to be misaligned to the rest of the sequences in several patients; hence an automated registration step was added, which is presented in "Pre-processing" section.

Figure 3 shows the channels of one image from each dataset after all the mentioned pre-processing steps.

Automated lesion growing. To enable training a single model on both datasets, it was mandatory to homogenize how lesion information was to be provided to the model: while the IVO dataset provided the full segmentation mask for each lesion, in ProstateX only the center position of the lesion was available. Although detection systems can be adapted to detect positions, they are typically designed to work with much more semantically rich BBs[26], or segmentations, or both[45].

To solve this inconsistency between the datasets, a similar approach to Liu et al.[46] was employed: for the ProstateX dataset, lesions were automatically segmented by growing them from the provided image position (used as seed), using a threshold level set method from Python library SimpleITK[47]. Concretely, the algorithm was applied independently to sequences T2, b800, and K^{trans}, and all segmented areas present in at least two of these three sequences were kept. Figure 4 shows the process of applying this segmentation algorithm to one image. This figure (and several others in this paper) were generated using Python library plot_lib[48].

Automated prostate zonal segmentation. Following McNeal's criterion[49], the prostate is typically partitioned into two distinct zones: the Central Gland (CG, including both the transition zone and the central zone, which are difficult to distinguish) and the Peripheral Zone (PZ). PCa lesions vary in frequency and malignancy depending on zone[50] and, as such, PI-RADS v2 considers them when assessing mpMRIs[51]. Therefore, just like a radiologist, a model for automated PCa detection and classification will likely benefit from having both CG and PZ mask priors provided as inputs, in addition to the mpMRI.

Accordingly, a cascading system of two segmentation CNNs, similar to the one introduced by Zhu et al.[52], was developed for automatic CG and PZ segmentation. As it can be seen in Supplementary Figure 1, the first CNN -a published model[44] based on the U-Net[53] CNN architecture with dense[54] and residual[55] blocks-, takes a prostate T2 image as input and produces a prostate segmentation mask as output. Then, the second CNN takes both the T2 image and the prostate segmentation mask obtained in the previous step and generates a CG segmentation mask as output. Finally, the PZ segmentation mask can be computed by subtracting the CG from the prostate segmentation mask.

The second CNN employed an architecture identical to the first one but was retrained on 92 prostate T2 images from a private dataset, in which the CG was manually segmented by a radiologist with two years of experience in PCa imaging. To be more precise, 80 of the 92 images were used for training the CG segmentation model, while the remaining 12 were employed for testing. Additionally, this model was also blindly tested (i.e.: with no retraining or adaptation of any kind) against the NCI-ISBI[56] train dataset, which is freely available at http://doi.org/10.7937/K9/TCIA.2015.zF0vlOPv. The results of this prostate zonal segmentation model are very briefly analyzed and compared to others in "Prostate zonal segmentation" section. Once trained and validated, this model was employed to obtain the CG and PZ masks of all the prostates in the current study.

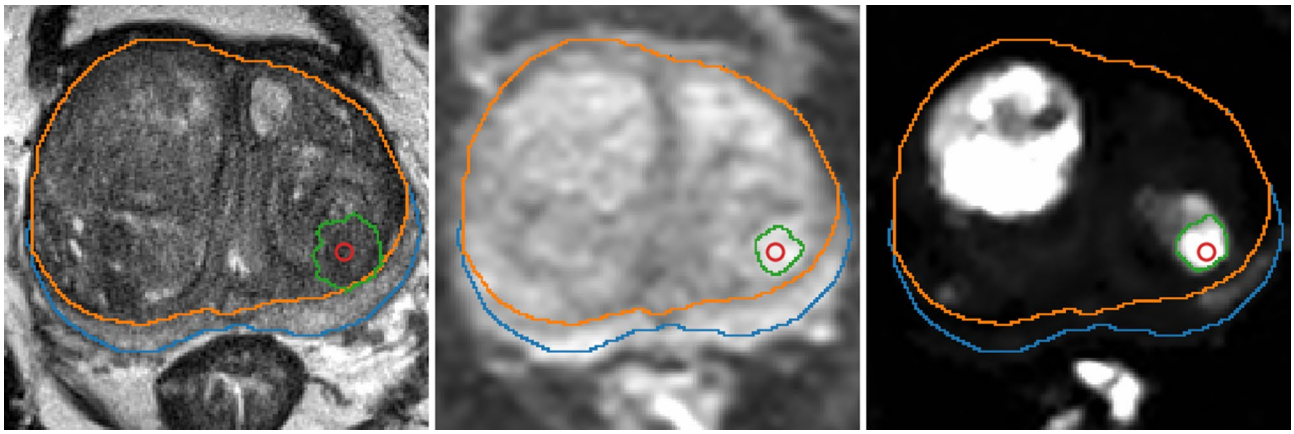

Figure 4. Automatic lesion segmentation for a ProstateX patient in sequences (from left to right: T2, b800 and K^{trans}) before combining them. Prostate zonal segmentation and the original lesion position (in red) are shown for reference.

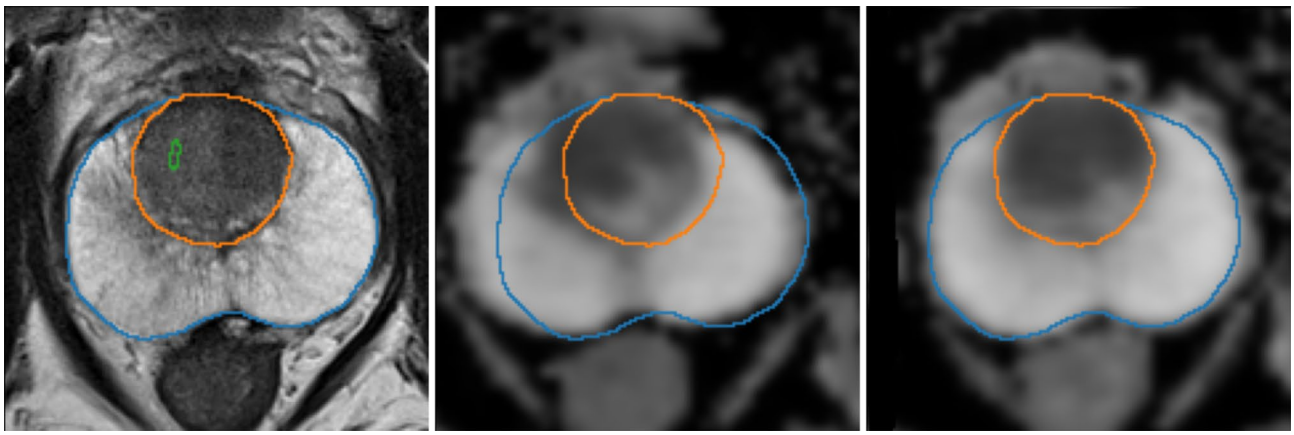

Figure 5. Automatic registration between T2 sequence (left) and ADC map (center: before, right: after) for a sample mpMRI.

Automated sequence registration. In several patients, DW sequences and the ADC map were misaligned to T2 and the other sequences. As a solution, non-rigid registration (based on a BSpline transformation) was applied between the spatial gradient of the T2 and the ADC map using Python library SimpleITK[47], with Mattes Mutual Information[57] as loss function and gradient descent[58] as the optimizer for the BSpline parameters. For every mpMRI, the registration algorithm was run 50 times with different parameter initializations, and the correlation coefficient between the spatial gradient of the T2 sequence and the spatial gradient of the registered ADC map was evaluated at the CG and the PZ areas. These custom metrics allowed to place a bigger emphasis to the areas of interest, as compared to image-wide metrics. Finally, the transformation associated with the run yielding the highest value for the average of all metrics and the loss was chosen as final and applied to both DW and ADC sequences. Figure 5 shows the result of applying this procedure to one mpMRI.

Model training and validation. After pre-processing the data, it was used to train a Retina U-Net[27] CNN architecture, which allows for the simultaneous detection, segmentation, and classification of PCa lesions. "Model training and validation" section provides an overview of this architecture, while "Hyperparameters–Epoch and CV ensembling during testing" sections deal with all engineering decisions related to the model training, validation, and testing.

Architecture: Retina U-Net. The Retina U-Net[27] architecture combines the Retina Net[59] detector with the U-Net segmentation CNN and is specifically designed for application to medical images. On one hand, Retina Net is a one-shot detector, meaning that classification and BB refinement (regression) are directly performed using the intermediate activation maps from the output of each decoder block in the Feature Pyramid Network (FPN) that conforms its backbone[60], making it not only more efficient but also better suited for lesion detection in medical

images, which have distinct characteristics compared to natural images (e.g.: there is no overlap between detections).

Furthermore, in the Retina U-Net, the FPN has been extended with two more high-resolution pyramid levels leading to a final segmentation layer, hence making the extended FPN architecture extremely akin to that of the U-Net. Therefore, the lesions are segmented independently of the detections (unlike other similar detection+segmentation architectures, such as Mask R-CNN[26]). This simplifies the architecture significantly, while still being a sensible choice for segmenting lesions since they all represent a single entity irrespective of their particular classes. Supplementary Figure 2 shows an overview of the Retina U-net architecture applied to the problem of simultaneous PCa detection, classification, and segmentation.

Hyperparameters. An ensemble of five CNNs (see "Model training and validation" section) was trained with the ResNet101-like backbone[55] with batch normalization[61] and a batch size of 6, at 120 batches per epoch, for a total of 115 epochs. Please, refer to "Model training and validation" section for more information on how data was split for training and validating the model. A triangular cyclical learning rate (LR) with exponential decay was employed[62], with LRs oscillating between a minimum of 8×10^{-5} and a maximum of 3.5×10^{-4}. For the BBs, a single aspect ratio of 1 (before BB refinement) was considered sufficient, with scales ranging from $4 \times 4 \times 1$ voxels (i.e.: $2 \times 2 \times 3$ mm), all the way to $28 \times 28 \times 9$ voxels (i.e.: $14 \times 14 \times 27$ mm), depending on the pyramid level on which the detection was performed. The rest of the parameters were left at their default values[27].

In particular, the encoder was a ResNet101-like CNN with the highest-resolution pyramid levels (P_0 and P_1) consisting of a single convolution, and the rest ($P_2, \ldots, P_5$) consisting of [3, 7, 21, 3] residual blocks, respectively. The stride of the last convolution of each pyramid level $P_0, \ldots, P_5$ was set to [1, 2, 2, 2, 2, 2], respectively for the x and y dimensions of the feature maps, and to [1, 1, 1, 2, 2, 2] for the z dimension, to account for the non-uniform voxel spacing. The decoder consisted in a single convolution per pyramid level followed by a simple upsampling; feature maps from the skip connections were merged with the upsampled feature maps by addition. Both the BB regressor head and the classifier head consisted of a stack of five convolutions. Convolution kernels were all of size 3×3 and *relu* non-linearity was used as activation function.

Online data augmentation. To help with regularization and to expand the limited training data, extensive online 3D data augmentation was employed during training using the Python library Batchgenerators[63]. Both rigid and non-rigid transformations, such as scaling, rotations, and elastic deformations were used.

Additionally, a custom augmentation was included to help deal with the issue of missing sequences, either because they never existed (such as K^{trans} images in the IVO dataset), or because they were not available. This augmentation, named Random Channel Drop, consisted in setting any given channel to zero (blanking it) with a certain probability, hence accustoming the model to dealing with missing data. During training, every channel of every image had a 7.5% probability of being dropped, except for the T2 channel and the segmentation masks, which had a probability of 0% (since they are assumed to be always available). The three DCE channels were considered as a whole for the purposes of dropping them (i.e.: they could not be dropped independently of each other).

Data partitioning. The mpMRIs were split into two sets: the train/validation set and the test set. The test set only contained "complete" mpMRIs (with no missing sequences), amounting to 30 IVO patients (23.62% of all complete IVO patients) and 45 ProstateX patients (22.17% of all ProstateX patients). This set was kept secret during the development of the model and was only employed eventually to validate it. Instead, for internal validation, five-fold cross-validation (CV) was employed: the train/validation set was split into five disjoint subsets, and five different instances of the same Retina U-Net model were successively trained on four out of the five subsets and validated on the fifth, hence creating a virtual validation dataset that encompassed the totality of the training data (but not the test data, which were kept apart).

As mentioned in "Data description" section, there was an additional ProstateX challenge set containing 140 mpMRIs with all the same information as the training set, except for the lesion GGG, which was not available. Hence, this dataset could also be employed for training both the segmentation and the BB regressor components of the Retina U-Net (but not the classifier). As such, this dataset was included as part of the training set (but not in the validation sets, as it contained no GT class information), and the classifier had to be modified to ignore any detection belonging to this dataset (i.e.: the loss was not propagated from such detections).

In summary, the model was trained and five-fold cross-validated with 191 IVO patients (of which only 45.55% were complete) + 159 ProstateX patients (all complete) + 140 ProstateX test patients (those coming from the ProstateX challenge set, for which GGG class information was not available). For testing, a secret subset consisting of 30 IVO patients and 45 ProstateX patients (all complete) was employed. The model was also tested on the ongoing ProstateX challenge.

Epoch and CV ensembling during testing. During the final test set prediction, both epoch and CV ensembling were used to boost the capabilities of the model. In general, ensembling consists in training N models for the same task, using them to predict on a given test set, and then combining all N predictions to achieve a better joint performance than that of each model individually. Hence, the five CV models were used for ensembling and, additionally, for every one of these CV models, the weights from the best (i.e.: highest validation mean -over all classes- Average Precision) five epochs were used as further independent models, totaling an equivalent of 25 virtual models.

Then, the predictions from the ensemble on the test set were combined in the following way: for segmentation masks, the average mask (over all 25 proposals) was computed and, for the BBs, the weighted box clustering (WBC) algorithm with an Intersection over Union threshold of 1×10^{-5} was applied to each class independently. The WBC algorithm is described in the original Retina U-Net paper[27].

Lesion matching and evaluation. The results were evaluated at three lesion significance thresholds (GGG≥1, GGG≥2, and GGG≥3) and two levels: lesion-level and patient-level. Only predicted BBs with a predicted GGG equal or above the chosen significance threshold (e.g.: GGG≥2) were considered, and the rest were completely ignored.

For lesion-level evaluation, each of the GT lesions was first matched with one (or none) of the detected lesions. First, all predicted BBs whose centroid was less than 15 mm away from that of the GT BB were selected as candidates for matching, and assigned a matching score computed as $\widehat{p} + k \cdot (1 - d/15\,\text{mm})$, where $\widehat{p}$ represents the actual score given by the model to that detection, d is the distance between the GT BB centroid and the candidate BB centroid, and $k = 2$. That way, both the model confidence ($\widehat{p}$) and distance to the GT (d) were considered for matching. The parameters for this matching procedure (e.g.: $k = 2$, 15 mm) were adjusted directly on the training set. If no detections existed within a 15 mm radius of a GT BB, a score of 0 was assigned to it. This evaluation method measures the performance of the model only on GT lesions for which biopsy confirmation and GGG are available, without assuming anything about the rest of the prostate, which may or may not contain other lesions. Furthermore, it allows the model to compete in the online ProstateX challenge (despite it not being an ROI classification model) since it can assign a score to every GT lesion.

For patient-level evaluation, the patient score was computed as the highest score from any BB predicted for the patient, and the GT GGG of a patient was computed as the highest GGG among all his GT lesions and among all the 20–30 cylinders obtained in the systematic biopsy (which were only available for patients from the IVO dataset). Hence, for the IVO dataset, a patient without any significant GT lesions might still have csPCa; for ProstateX, however, we do not know, and we must assume that this does not happen.

Data availability

Data from the ProstateX challenge are available at https://doi.org/10.7937/K9TCIA.2017.MURS5CL[18]; data from the Valencian Institute of Oncology is not publicly available, since the ethical committee (CEIm-FIVO) only approved its use for the current study. They might be made available for research purposes on reasonable request from the corresponding author. The code of the project is available at https://github.com/OscarPellicer/prostate_lesion_detection.

Received: 13 July 2021; Accepted: 3 February 2022
Published online: 22 February 2022

References

1. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**(6), 394–424. https://doi.org/10.3322/caac.21492 (2018).
2. Mottet, N. *et al.* EAU - ESTRO - ESUR - SIOG Guidelines on Prostate Cancer. *European Association of Urology*, 12–18 (2017).
3. Mehralivand, S. *et al.* A magnetic resonance imaging-based prediction model for prostate biopsy risk stratification. *JAMA Oncol.* **4**(5), 678–685. https://doi.org/10.1001/jamaoncol.2017.5667 (2018).
4. Ahmed, H. U. *et al.* Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): A paired validating confirmatory study. *Lancet* **389**(10071), 815–822. https://doi.org/10.1016/S0140-6736(16)32401-1 (2017).
5. Marra, G. *et al.*. Controversies in MR targeted biopsy: Alone or combined, cognitive versus software-based fusion, transrectal versus transperineal approach? (2019).
6. Cellini, N. *et al.* Analysis of intraprostatic failures in patients treated with hormonal therapy and radiotherapy: Implications for conformal therapy planning. *Int. J. Radiat. Oncol. Biol. Phys.* **53**(3), 595–599. https://doi.org/10.1016/S0360-3016(02)02795-5 (2002).
7. Turkbey, B. *et al.* Prostate Imaging Reporting and Data System Version 2.1: 2019 Update of Prostate Imaging Reporting and Data System Version 2 (2019).
8. Gaziev, G. *et al.* Defining the learning curve for multiparametric magnetic resonance imaging (MRI) of the prostate using MRI-transrectal ultrasonography (TRUS) fusion-guided transperineal prostate biopsies as a validation tool. *BJU Int.* **117**(1), 80–86. https://doi.org/10.1111/bju.12892 (2016).
9. Sonn, G. A. *et al.* Prostate magnetic resonance imaging interpretation varies substantially across radiologists. *Eur. Urol. Focus* **5**(4), 592–599. https://doi.org/10.1016/j.euf.2017.11.010 (2019).
10. Kohestani, K. *et al.* Performance and inter-observer variability of prostate MRI (PI-RADS version 2) outside high-volume centres. *Scand. J. Urol.* **53**(5), 304–311. https://doi.org/10.1080/21681805.2019.1675757 (2019).
11. Khoo, C. C. *et al.* Likert vs PI-RADS v2: A comparison of two radiological scoring systems for detection of clinically significant prostate cancer. *BJU Int.* **125**(1), 49–55. https://doi.org/10.1111/bju.14916 (2020).
12. Giger, M. L. & Suzuki, K. Computer-aided diagnosis. In *Biomedical Information Technology*, 359–374. Elsevier (2008). ISBN 9780123735836. https://doi.org/10.1016/B978-012373583-6.50020-7
13. Morton, M. J., Whaley, D. H., Brandt, K. R. & Amrami, K. K. Screening mammograms: Interpretation with computer-aided detection-prospective evaluation (2006).
14. Van Ginneken, B., Schaefer-Prokop, C. M. & Prokop, M. Computer-aided diagnosis: How to move from the laboratory to the clinic (2011).
15. Chan, I. *et al.* Detection of prostate cancer by integration of line-scan diffusion, T2-mapping and T2-weighted magnetic resonance imaging; a multichannel statistical classifier. *Med. Phys.* **30**(9), 2390–2398. https://doi.org/10.1118/1.1593633 (2003).
16. Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N. & Huisman, H. Computer-aided detection of prostate cancer in MRI. *IEEE Trans. Med. Imaging* **33**(5), 1083–1092. https://doi.org/10.1109/TMI.2014.2303821 (2014).
17. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**(6), 84–90. https://doi.org/10.1145/3065386 (2012).
18. Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., & Huisman, H. ProstateX Challenge data (2017).

19. Armato, S. G. *et al.* PROSTATEx Challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *J. Med. Imaging* **5**(04), 1. https://doi.org/10.1117/1.jmi.5.4.044501 (2018).
20. Kitchen, A. & Seah, J. Support vector machines for prostate lesion classification. In *Medical Imaging 2017: Computer-Aided Diagnosis* (eds. Armato, S. G. & Petrick, N. A.), vol. 10134, p. 1013427. SPIE, 2017. https://doi.org/10.1117/12.2277120
21. Liu, S., Zheng, H., Feng, Y., & Li, W. Prostate cancer diagnosis using deep learning with 3d multiparametric MRI (2017).
22. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR (2015).
23. Cao, R. *et al.* Joint prostate cancer detection and Gleason score prediction in mp-MRI via FocalNet. *IEEE Trans. Med. Imaging* **38**(11), 2496–2506. https://doi.org/10.1109/TMI.2019.2901928 (2019).
24. Epstein, J. I., Allsbrook, W. C., Amin, M. B. & Egevad, L. L. The 2005 International Society of Urological Pathology (ISUP) consensus conference on gleason grading of prostatic carcinoma. *Am. J. Surg. Pathol.* **29**(9), 1228–1242. https://doi.org/10.1097/01.pas.0000173646.99337.b1 (2005).
25. Epstein, J. I. *et al.* 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. *Am. J. Surg. Pathol.* **40**(2), 244–252. https://doi.org/10.1097/PAS.0000000000000530 (2016).
26. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN, 2020. https://ieeexplore.ieee.org/document/8372616/
27. Jaeger, P. F. *et al.* Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection. *Proceedings of Machine Learning Research* **116**, 171–183 (2020).
28. Arif, M. *et al.* Clinically significant prostate cancer detection and segmentation in low-risk patients using a convolutional neural network on multi-parametric MRI. *Eur. Radiol.* **30**(12), 6582–6592. https://doi.org/10.1007/s00330-020-07008-z (2020).
29. Aldoj, N., Lukas, S., Dewey, M. & Penzkofer, T. Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network. *Eur. Radiol.* **30**(2), 1243–1253. https://doi.org/10.1007/s00330-019-06417-z (2020).
30. Drost, F.-J.H. *et al.* Prostate MRI, with or without MRI-targeted biopsy, and systematic biopsy for detecting prostate cancer. *Cochrane Database Syst. Rev.* **4**(4), CD012663. https://doi.org/10.1002/14651858.CD012663.pub2 (2019).
31. Fütterer, J. J. *et al.* Can Clinically Significant Prostate Cancer Be Detected with Multiparametric Magnetic Resonance Imaging? A Systematic Review of the Literature (2015). http://www.europeanurology.com/article/S0302283815000366/fulltexthttp://www.europeanurology.com/article/S0302283815000366/abstracthttps://www.europeanurology.com/article/S0302-2838(15)00036-6/abstract
32. Xu, H., Baxter, J. S. H., Akin, O. & Cantor-Rivera, D. Prostate cancer detection using residual networks. *Int. J. Comput. Assist. Radiol. Surg.* **14**(10), 1647–1650. https://doi.org/10.1007/s11548-019-01967-5 (2019).
33. Schelb, P. *et al.* Simulated clinical deployment of fully automatic deep learning for clinical prostate MRI assessment. *Eur. Radiol.* **1–12**, 2020. https://doi.org/10.1007/s00330-020-07086-z (2020).
34. Woźnicki, P. *et al.* Multiparametric MRI for prostate cancer characterization: Combined use of radiomics model with PI-RADS and clinical parameters. *Cancers* **12**(7), 1767. https://doi.org/10.3390/cancers12071767 (2020).
35. Yoo, S., Gujrathi, I., Haider, M. A. & Khalvati, F. Prostate cancer detection using deep convolutional neural networks. *Sci. Rep.* **9**(1), 2019. https://doi.org/10.1038/s41598-019-55972-4 (2019).
36. Winkel, D. J. *et al.* Autonomous detection and classification of PI-RADS lesions in an MRI screening population incorporating multicenter-labeled deep learning and biparametric imaging: Proof of concept. *Diagnostics* **10**(11), 951. https://doi.org/10.3390/diagnostics10110951 (2020).
37. Vente, C. D., Vos, P., Hosseinzadeh, M., Pluim, J. & Veta, M. Deep learning regression for prostate cancer detection and grading in bi-parametric MRI. *IEEE Trans. Biomed. Eng.* **68**(2), 374–383. https://doi.org/10.1109/TBME.2020.2993528 (2021).
38. Steenbergen, P. *et al.* Prostate tumor delineation using multiparametric magnetic resonance imaging: Inter-observer variability and pathology validation. *Radiother. Oncol.* **115**(2), 186–190. https://doi.org/10.1016/j.radonc.2015.04.012 (2015).
39. Junker, D. *et al.* Comparison of multiparametric and biparametric MRI of the prostate: Are gadolinium-based contrast agents needed for routine examinations?. *World J. Urol.* **37**(4), 691–699. https://doi.org/10.1007/S00345-018-2428-Y (2018).
40. Zawaideh, J. P. *et al.* 30(7):4039–4049 (2020). https://doi.org/10.1007/S00330-020-06782-0
41. Qin, X. *et al.* 3D multi-scale discriminative network with multi-directional edge loss for prostate zonal segmentation in bi-parametric MR images. *Neurocomputing* **418**, 148–161. https://doi.org/10.1016/j.neucom.2020.07.116 (2020).
42. Rundo, L. *et al.* USE-Net: Incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets. *Neurocomputing* **365**, 31–43. https://doi.org/10.1016/j.neucom.2019.07.006 (2019).
43. Aldoj, N., Biavati, F., Michallek, F., Stober, S. & Dewey, M. Automatic prostate and prostate zones segmentation of magnetic resonance images using DenseNet-like U-net. *Sci. Rep.* **10**(1), 1–17. https://doi.org/10.1038/s41598-020-71080-0 (2020).
44. Pellicer-Valero, O. J. *et al.* Robust resolution-enhanced prostate segmentation in magnetic resonance and ultrasound images through convolutional neural networks. *Appl. Sci. (Switzerland)* **11**(2), 1–17. https://doi.org/10.3390/app11020844 (2021).
45. Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031 (2017).
46. Liu, Z. *et al.* A two-stage approach for automated prostate lesion detection and classification with mask R-CNN and weakly supervised deep neural network. *LNCS* **11850**, 43–51. https://doi.org/10.1007/978-3-030-32486-5_6 (2019).
47. Yaniv, Z., Lowekamp, B. C., Johnson, H. J. & Beare, R. SimpleITK image-analysis notebooks: A collaborative environment for education and reproducible research. *J. Digit. Imaging* **31**(3), 290–303. https://doi.org/10.1007/s10278-017-0037-8 (2018).
48. Pellicer-Valero, O. J. OscarPellicer/plot_lib. https://doi.org/10.5281/zenodo.4395271 (2020).
49. Selman, S. H. The McNeal prostate: A review (2011).
50. Haffner, J. *et al.* Peripheral zone prostate cancers: Location and intraprostatic patterns of spread at histopathology. *Prostate* **69**(3), 276–282. https://doi.org/10.1002/pros.20881 (2009).
51. Weinreb, J. C. *et al.* PI-RADS prostate imaging—Reporting and data system: 2015, Version 2. *Eur. Urol.* **69**(1), 16–40. https://doi.org/10.1016/j.eururo.2015.08.052 (2016).
52. Zhu, Y. *et al.* Fully automatic segmentation on prostate MR images based on cascaded fully convolution network. *J. Magn. Reson. Imaging* **49**(4), 1149–1156. https://doi.org/10.1002/jmri.26337 (2019).
53. Ronneberger, O., Fischer, P., & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **9351**, 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28. arXiv:1505.04597
54. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017:2261–2269, 2017. https://doi.org/10.1109/CVPR.2017.243
55. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016:770–778, 2015. https://doi.org/10.1109/CVPR.2016.90. arXiv:1512.03385
56. Bloch, N. *et al.* Challenge: Automated Segmentation of Prostate Structures. *Cancer Imaging Arch.***2015**, https://doi.org/10.7937/K9/TCIA.2015.zF0vlOPv (2013).

57. Mattes, D., Haynor, D. R., Vesselle, H., Lewellyn, T. K. & Eubank, W. Nonrigid multimodality image registration. In *Medical Imaging 2001: Image Processing* (eds. Sonka, M. & Hanson, K. M.), vol. 4322, 1609–1620. SPIE. https://doi.org/10.1117/12.431046 (2001).
58. Ruder, S. An overview of gradient descent optimization algorithms. (2016). arXiv:1609.04747
59. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.***42**(2), 318–327 (2017). arXiv:1708.02002
60. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B. & Belongie, S. Feature pyramid networks for object detection. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017, 936–944 (2017). ISBN 9781538604571. https://doi.org/10.1109/CVPR.2017.106
61. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *32nd International Conference on Machine Learning, ICML 2015*, vol. 1, 448–456. International Machine Learning Society (IMLS) (2015). ISBN 9781510810587.
62. Smith, L. N. Cyclical Learning Rates for Training Neural Networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 464–472. IEEE, 2017. ISBN 978-1-5090-4822-9. https://doi.org/10.1109/WACV.2017.58. arXiv:1506.01186
63. Fabian, I., *et al*. batchgenerators—a python framework for data augmentation (2020). https://zenodo.org/record/3632567

Author contributions

O.J.P.: data curation, methodology, software, visualization, and writing—original draft; J.L.M.J.: methodology, resources, and writing—original draft; V.G.: methodology, resources, and writing—review & editing; J.L.C.R.: supervision; I.M.G.: resources; M.B.B.: resources, validation; P.P.G.: validation; J.R.: supervision; M.J.R.: writing—review & editing; J.D.M.: conceptualization, project administration, supervision, and writing—review & editing.

Funding

This work has been partially supported by a doctoral grant of the Spanish Ministry of Innovation and Science, with reference FPU17/01993.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-06730-6.

Correspondence and requests for materials should be addressed to O.J.P.-V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.