



OPEN Effective reduction of unnecessary biopsies through a deep-learning-assisted aggressive prostate cancer detector

Nuno M. Rodrigues^{1,2,26}✉, José Guilherme de Almeida^{1,26}, Ana Sofia Castro Verde¹, Ana Mascarenhas Gaivão³, Carlos Bireiro³, Inês Santiago³, Joana Ip³, Sara Belião³, Celso Matos¹, Leonardo Vanneschi⁴, Manolis Tsiknakis^{5,6}, Kostas Marias^{6,7}, Daniele Regge^{8,9}, Sara Silva^{2,27}, The ProCancer-I Consortium* & Nickolas Papanikolaou^{1,10,27}

Despite being one of the most prevalent cancers, prostate cancer (PCa) shows a significantly high survival rate, provided there is timely detection and treatment. Currently, several screening and diagnostic tests are required to be carried out in order to detect PCa. These tests are often invasive, requiring either a biopsy (Gleason score and ISUP) or blood tests (PSA). Computational methods have been shown to help this process, using multiparametric MRI (mpMRI) data to detect PCa, effectively providing value during the diagnosis and monitoring stages. While delineating lesions requires a high degree of experience and expertise from the radiologists, being subject to a high degree of inter-observer variability, often leading to inconsistent readings, these computational models can leverage the information from mpMRI to locate the lesions with a high degree of certainty. By considering as positive samples only those that have an ISUP ≥ 2 we can train aggressive index lesion detection models. The main advantage of this approach is that, by focusing only on aggressive disease, the output of such a model can also be seen as an indication for biopsy, effectively reducing unnecessary biopsy screenings. In this work, we utilize both the highly heterogeneous ProstateNet dataset, and the PI-CAI dataset, to develop accurate aggressive disease detection models.

Prostate cancer (PCa) is the most prevalent cancer in men and the second most prevalent across genders¹. However, PCa is also characterized by a low mortality rate provided there is early detection, a key factor in ensuring positive treatment outcomes. While biopsies constitute an essential step in diagnosing and stratifying prostate cancer, false positives or incorrect risk assessments can lead to over-treatment. Together with treatment side effects, this may result in a loss of quality of life for the patients, making it imperative to carefully consider treatment choices². The development of computer-aided diagnosis (CAD) models capable of providing “virtual biopsies” assisted by biparametric MRI (bpMRI) has the potential to reduce unnecessary biopsies and improve the risk assessment process. Indeed, the typical process for the recommendation of a biopsy consists of the analysis by an expert radiologist who will recommend a biopsy based on a positive (>2) or negative (<3) Prostate Imaging-Reporting and Data System (PI-RADS) score³, a process with a high rate of false positives⁴.

¹Computational Clinical Imaging Group, Champalimaud Foundation, Lisbon, Portugal. ²LASIGE, Faculty of Sciences, LASIGE, Faculdade de Ciências, Universidade de Lisboa, 1749--016 Lisboa, Portugal, Lisbon, Portugal. ³Radiology Department, Champalimaud Clinical Center, Champalimaud Foundation, Lisbon, Portugal. ⁴NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal. ⁵Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), 700 13, Heraklion, Greece. ⁶Department of Electrical and Computer Engineering, Hellenic Mediterranean University, 710 04, Heraklion, Greece. ⁷Computational BioMedicine Laboratory (CBML), Institute of Computer Science, Foundation for Research and Technology-Hellas (FORTH), Heraklion, Greece. ⁸Department of Radiology, Candiolo Cancer Institute, FPO-IRCCS, Strada Provinciale 142 Km 3.95, 10060 Candiolo, Italy. ⁹Department of Surgical Sciences, University of Turin, 10124 Turin, Italy. ¹⁰Department of Radiology, Royal Marsden Hospital, Sutton, UK. ²⁶Nuno M. Rodrigues and José Guilherme de Almeida contributed equally to this work. ²⁷These authors jointly supervised this work: Sara Silva and Nickolas Papanikolaou. *A list of authors and their affiliations appears at the end of the paper. ✉email: nuno.mvrodrigues1@gmail.com

While the performance of automated systems is seldom as good as that of expert radiologists⁵, the latter commonly suffer from inter- and intra-expert variability^{6,7}, which can be a limiting factor in deciding between performing or not performing a biopsy or even in choosing an appropriate treatment. Computational models have the benefit of producing consistent results provided the input data is identical, with the caveat that performance degradation is common when transferring models between scanner manufacturers⁸ or, in the case of prostate bpMRI, scanner manufacturers and the use of endorectal coil. However, some works have explored the benefits of using large multi-centric heterogeneous datasets to improve the robustness and performance of the models, effectively reducing the effects of domain-shift^{9–11}.

Recent CAD models have shown potential in several clinical applications for PCa, from disease aggressiveness classification^{12–14} to lesion segmentation and detection^{9,15–25}. However, these works seldom focus on unnecessary biopsy reduction, a clinical endpoint which has direct implications for patient care. Additionally, they tend to make use of single-centric datasets and rarely include a prospective validation of the developed models. Here, we make use of the publicly available PI-CAI^{25,26}, as well as ProstateNet (<https://prostatenet.eu>), a large-scale multi-centric dataset of multiparametric prostate MRI to train aggressive lesion segmentation models. We show that using heterogeneous datasets leads to improved segmentation and lesion detection performance, and validate it using a hold-out test set. Through a simulated clinical feasibility analysis, we show how the combination of medical recommendations with our fully automatic models can lead to an effective reduction in the number of unnecessary biopsies with no significant reduction in Recall, effectively reducing the number of false positives. Finally, we validate all aspects of this approach using prospective data.

Methods

Data

In this study, two different datasets were used: PI-CAI²⁶ and ProstateNet (also referred to as PNet). Each dataset is composed of a retrospective cohort, with ProstateNet also having a prospective cohort. The following are the descriptions of the datasets:

- *PI-CAI* is a collection of Biparametric MRI volumes that include T2W, DWI and ADC sequences. These samples were acquired by three Dutch clinical centers (Radboud University Medical Center (RUMC), Ziekenhuis Groep Twente (ZGT), University Medical Center Groningen (UMCG)), and one Norwegian center (Norwegian University of Science and Technology (NTNU)), plus the additional inclusion of 329 cases from the ProstateX dataset²⁷. These clinical centers used only Siemens Healthineers or Philips Medical Systems-based 1.5T or 3T MRI scanners with surface coils to acquire the images, following the Biparametric prostate MRI protocol²⁸. As stated in the official document of the dataset²⁶, ISUP values of 0 represent confirmed negatives or cases without the required 3-year follow-up. In total, 1009 biparametric sequences were used.
- *ProstateNet* (PNet) is a collection of Biparametric MRI volumes that include T2W, DWI and ADC sequences. These samples were acquired by 12 clinical partners of the ProCancer-I project. These partners used Siemens (Aera, Skyra, Sola, Avanto, VIDA, Tim, Prisma, Veri, Symphony, Osirix), Philips (Ingenia, Achieva, Multiva) and GE scanners (Optima, Signa, DISCOVERY). Given that each centre has specific acquisition protocols, no single one was used across all mpMRI studies done. All labels were acquired manually, and for each sample, the label consists of the index lesion (mandatory) and additional lesions that the patient has (optional). ISUP values of 0 represent cases confirmed negative after 1 year of follow-up or non-confirmed cases. In total, 1484 biparametric sequences were used.

To maximize data variability, both datasets were combined into a global one, dubbed PNetCAI. Table 1 shows the composition of the different retrospective datasets regarding scanner manufacturers and ISUP grades, while Table 2 does the same for the prospective cohort. The prospective cases were downloaded from the ProstateNet platform on February 26th 2024. From these numbers, 15% of the samples were used as a hold-out test set, and the remaining were used for training, following a 5-fold cross-validation (CV) strategy.

A connected component analysis was conducted on the training labels of both datasets (Fig. 1), revealing that 16 samples from the PI-CAI datasets that were labelled as aggressive (ISUP ≥ 2) were empty. This was cross-checked with the files present in their repository. A comparison between the size of the lesions on both datasets and their effect on the Dice score is presented in the “Results” section (3).

Biparametric data processing

In order to use all mpMRI sequences as a single volume, both DWI and ADC sequences were resampled to the same space and size of the T2W sequences. Both T2W and DWI images were normalized using Z-scoring normalization, while ADC images were normalized by clipping the intensity values to the 0.5 and 99.5 percentiles, followed by subtracting the mean and dividing by the standard deviation.

Deep learning model specification

All 3D deep-learning (DL) detection models that were trained were full resolution nnUNet models (nnUNet)²⁹ that use deep supervision³⁰. The networks are implemented in Pytorch³¹ and were trained for 1000 epochs (250 mini-batches per epoch). To train the nnUNet models, we used the provided 3D full resolution architecture. This framework uses stochastic gradient descent with Nesterov momentum ($\mu = 0.99$), a maximum initial learning rate of 0.001, and polynomial³² learning rate policy which reduces the learning rate by a factor of $(1 - epoch/epoch_{max})^{0.9}$ in each epoch. Initial tests showed that the default learning rate of the nnUNet (0.01) was too high, resulting in underfitting on some of the folds, the reason why we decided to use a lower, more common, value. The loss function was a simple average of Dice and cross-entropy losses and the batch size was

Scanners							
	Total	Siemens	Philips	GE	Toshiba	No data	
ProstateNet	1009	364	403	198	2	42	
PI-CAI	1484	1208	276	0	0	0	
ISUPs							
	0	1	2	3	4	5	
ProstateNet	519	228	141	69	20	31	
PI-CAI	847	228	223	98	39	49	
ISUPs	0	1	2	3	4	5	
Train							
ProstateNet	442 0.52	192 0.22	118 0.14	61 0.07	15 0.02	29 0.03	# samples proportion
PI-CAI	719 0.57	194 0.15	189 0.15	84 0.07	33 0.03	42 0.03	
Test							
ProstateNet	77 0.51	36 0.24	23 0.15	9 0.06	5 0.03	2 0.01	
PI-CAI	128 0.57	34 0.15	34 0.15	14 0.06	5 0.03	7 0.03	

Table 1. Stratification of samples of the retrospective data cohort. On the left, number of samples by scanner manufacturer and by ISUP score for the retrospective cohorts. On the right, number and proportion of samples on the training and test sets.

Scanners						
Total	Siemens	Philips	GE	Toshiba	No data	
73	34	14	25	0	0	
ISUPs						
Total	0	1	2	3	4	5
73	36	16	7	9	5	0

Table 2. Stratification of samples by scanner manufacturer and ISUP score for the prospective cohort of ProstateNet.

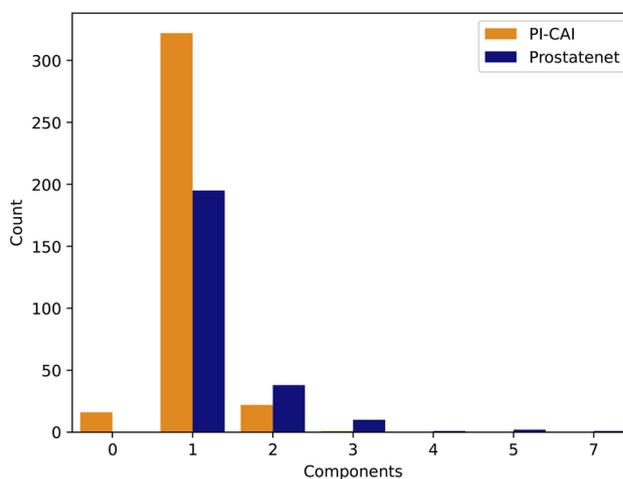


Fig. 1. Connected component analysis. Connected components analysis for both aggressive (ISUP ≥ 2) label masks of the ProstateNet and PI-CAI datasets.

2 sequences per iteration. The nnUNet applies automatic preprocessing based on the dataset fingerprint, and therefore the models for each dataset worked on data with slightly different spatial structures:

- ProstateNet: spacing = $0.5 \times 0.5 \times 3.0\text{mm}$; crop size = $256 \times 256 \times 30$ voxels
- PI-CAI: spacing = $0.4 \times 0.4 \times 3.0\text{mm}$; crop size = $384 \times 384 \times 21$ voxels
- PNetCAI: spacing = $0.5 \times 0.5 \times 3.0\text{mm}$; crop size = $384 \times 384 \times 23$ voxels

Based on recent work^{11,33}, no transformer-based models (ViT) were evaluated, as they were shown to perform significantly worse than nnUNet models. This is further justified by the original ViT paper, which states the need for very large datasets (over 1 million images) to train a ViT model from scratch³⁴.

Network calibration

Previous work³⁵ and prior experiments conducted by us for whole gland segmentation have shown that calibrating segmentation models significantly improves their performance. Given this, we decided to use the findings from Murugresan et al.³⁵ and change the nnUNet loss function to include both label-smoothing³⁶ and margin loss. We applied an α smoothing factor of 0.2 and a margin of 10 to the loss function.

Technical specifications

To train the models for this project, we used a machine with the following specifications: $2 \times$ NVIDIA RTX A6000 GPUs, AMD Ryzen Threadripper 3990X 64-Core Processor, and 64GB DDR4 RAM with 2200MHz clock speed. Each fold of each model took approximately 13h to finish.

Model evaluation

During the 5-fold CV, each model was evaluated based on its Dice Score (DS) and Recall when comparing the predicted output mask to that of the ground truth. When evaluating the performance on both the retrospective hold-out test and the prospective cohort, the same metrics were not computed on the vanilla output of the model, but on the candidate lesions obtained by following the subsequent methodology:

1. Taking the probability maps that the model outputs, a threshold of 10% was defined, clipping all voxels with a probability lower than 10%, generating a soft blob;
2. Taking those soft blobs, we employed the heuristics proposed by Bosma³⁷ and assigned all lesion candidates to their respective ground truth through a linear sum assignment algorithm;
3. All candidates that had a confidence above 10% (the confidence is the maximum probability within the candidate) were kept and turned into hard blobs (binary segmentation masks). All other candidates (i.e. candidates with a confidence below 10%) were excluded and not analyzed any further. This threshold was selected as it reflects what has been used previously in the literature for prostate lesion candidate selection³⁷;
4. Lastly, all hard blobs that had an intersection with the prostate gland of less than 10% (meaning they should be almost entirely outside the prostate, while still accounting for extracapsular extension) were classified as negative. The segmentations for the prostate gland were obtained using the whole gland segmentation model dubbed ProstateAll from Rodrigues et al.¹¹;
5. In order to perform a more rigorous assessment, only hard blobs with at least 10% intersection with the original lesion masks were considered positive, regardless of having located any other lesion present in the same sample. This assessment, despite lowering some of the scores as opposed to simply locating any lesion, provides a more realistic clinical application scenario.

Each model was tested in all available retrospective hold-out sets and on the prospective cohort. The training/testing setup is summarized in Figure 2.

Additionally, we also calculated the Hausdorff Distance (HD), Average Symmetric Surface distance (ASSD), and Relative Absolute Volume Difference (RAVD) during quality assessment of the model, as these metrics provide a quantitative measure of the spatial accuracy by considering the shape and volume of the segmented regions³⁸ (both distance metrics were calculated using MedPy³⁹). The evaluations and details of each metric are available in the Supplementary Methods (A.1).

Results

Model performance is affected by train-test similarity

As previously mentioned in “Model evaluation”, we follow a two-step process in order to select the most appropriate lesion candidates: lesion candidates are selected similarly to what has been described in³⁷, followed by a lesion filtering process that keeps only lesions with a 10% overlap with the whole prostate gland. Table 3 presents the cross-validation results of all developed models. Given that the models were trained as regular index lesion segmentation models, the resulting low Dice scores are a likely consequence of the heterogeneous nature of lesion annotation for the datasets used during training. We also note that bpMRI models outperform T2W models; this is expected, as both DWI and ADC sequences provide information in the form of hyper- and hypo-intense areas, which is much more relevant for lesion localization when compared to T2W sequences. The Recall also shows that bpMRI models, in particular the PI-CAI and PNetCAI models, can detect almost all lesions, achieving a maximum Recall score of 0.9 (90%), while their respective T2W counterparts can only locate approximately 65% of the lesions.

The similarity between training and testing data (i.e., training and testing models on training and hold-out datasets constructed from the same dataset) can also be an important factor affecting performance. While T2W models trained on PNet data perform well only on data from PNet (Dice = 0.34 and Dice = 0.13 for

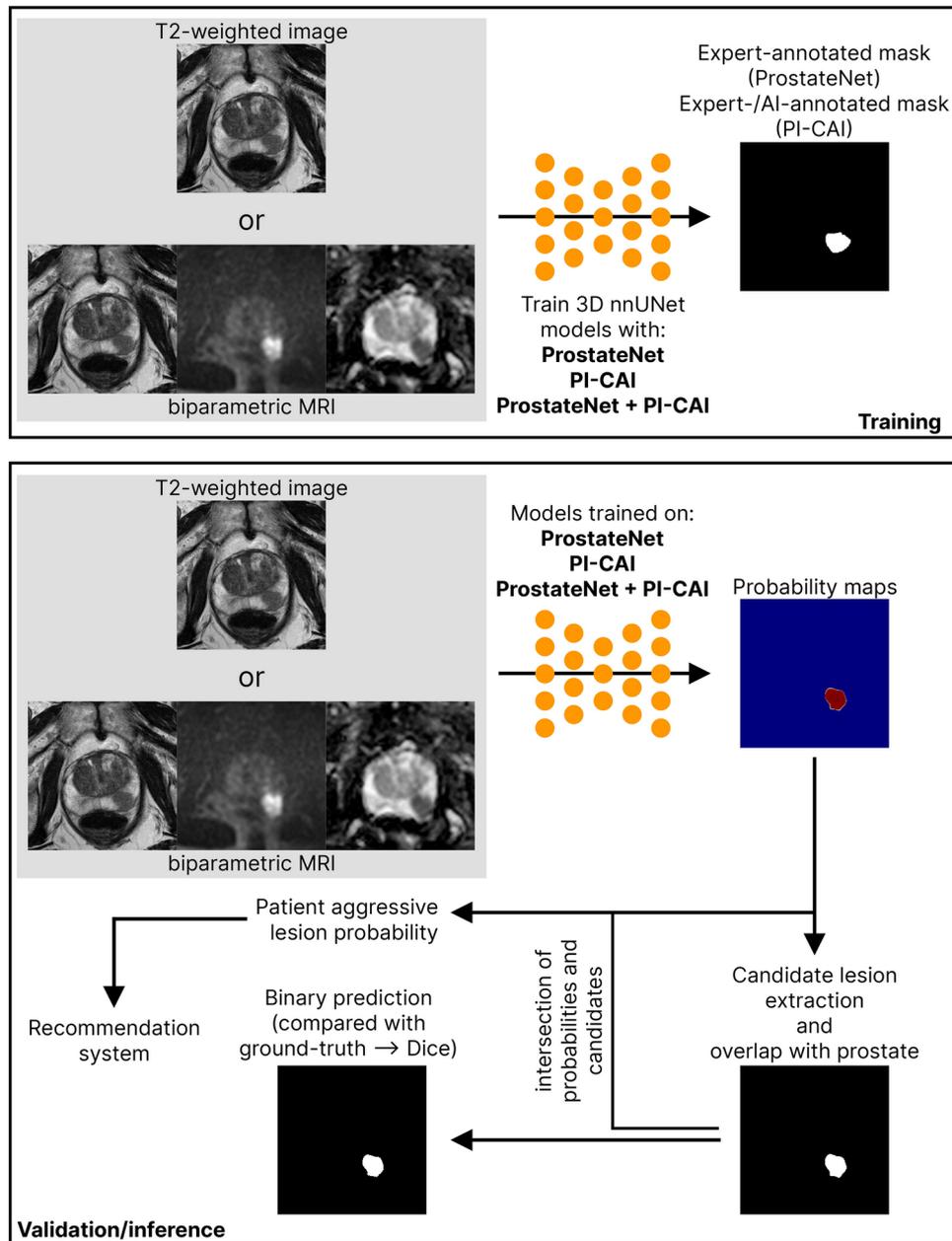


Fig. 2. Visualization of the training and validation/inference protocol for the models described in this work. Training was performed using either T2-weighted or biparametric MRI studies belonging to either ProstateNet (PNet), PI-CAI or ProstateNet + PI-CAI (PNetCAI) to detect lesions annotated by radiologists. The validation/inference protocol consists in detecting lesions, extracting the most relevant lesion candidates³⁷ and considering only lesions with an overlap of at least 10% with the whole prostate gland as inferred by a deep-learning model for prostate segmentation¹¹. The patient aggressive lesion probability is then used in a recommendation system, while the binary/probabilistic prediction is used for visualization.

T2W PNet models tested on PNet and PI-CAI, respectively), PI-CAI are more consistent (Dice = 0.34 and Dice = 0.30 for T2W PI-CAI models tested on PNet and PI-CAI, respectively; Tables 4, 5), an effect which is also consistent for Recall. However, using bpMRI leads to considerably worse performance in terms of both Dice and Recall for PI-CAI models tested on PNet data (Tables 4, 5); indeed, for bpMRI models, which outperform T2W models, performance is only consistently good for PNetCAI models. In other words, models perform consistently better only when there is some similarity between training and testing data.

This can be further observed in Table 6, where the bpMRI PNetCAI excels the bpMRI PNet model on its hold-out test set, while differing only in 2 lesions from the bpMRI PI-CAI model on its test set. Furthermore, after a manual analysis of these missed cases, we discovered that both were from out-of-distribution samples with very large fields of view.

	Dice	Hausdorff	RAVD	ASSD	Recall	<i>p</i> -value	
PNet	0.32 ± 0.02	48.44 ± 4.38	-0.23 ± 0.04	12.28 ± 2.28	0.57 ± 0.03	6.97e-5	T2W
PCAI	0.31 ± 0.01	176.8 ± 14.13	0.57 ± 0.18	54.07 ± 4.96	0.65 ± 0.02	4.27e-6	
PNetCAI	0.35 ± 0.01	167.23 ± 11.95	0.58 ± 0.18	48.64 ± 4.12	0.67 ± 0.03	1.05e-5	
PNet	0.37 ± 0.02	54.75 ± 4.59	-0.12 ± 0.1	12.8 ± 2.11	0.69 ± 0.03	0.0004	bpMRI
PCAI	0.61 ± 0.01	163.32 ± 16.75	1.04 ± 0.76	29.32 ± 3.83	0.9 ± 0.03	0.0004	
PNetCAI	0.5 ± 0.01	153.5 ± 13.96	0.46 ± 0.11	31.51 ± 3.39	0.83 ± 0.0	-	

Table 3. CV results. For each dataset, the average Dice, Hausdorff, RAVD, ASSD and Recall performance, along with their respective standard deviations, are presented. The highest recall value per sequence combination is highlighted in bold for easier comparison. *p*-values for the T-test significance comparing the Dice score between bpMRI PNetCAI results and each other model are also shown, with significant differences (*p*-value < 0.01) marked as green or red if the bpMRI PNetCAI results are better or worse, respectively.

Trained on	PNet	PI-CAI	PNetCAI	
	0.34 ± 0.05	0.13 ± 0.03	0.21 ± 0.03	Dice
	0.62	0.34	0.45	Recall
PNet	0.47	0.33	0.4	Precision
	0.34 ± 0.05	0.3 ± 0.03	0.32 ± 0.03	
	0.59	0.66	0.63	
PI-CAI	0.24	0.24	0.24	
	0.43 ± 0.05	0.29 ± 0.04	0.35 ± 0.03	
	0.74	0.64	0.68	
T2W	PNetCAI	0.25	0.24	0.24
	0.38 ± 0.05	0.12 ± 0.02	0.22 ± 0.03	
	0.72	0.34	0.49	
PNet	0.5	0.23	0.33	
	0.1 ± 0.03	0.6 ± 0.04	0.41 ± 0.04	
	0.33	0.85	0.64	
PI-CAI	0.28	0.28	0.28	
	0.41 ± 0.04	0.53 ± 0.04	0.49 ± 0.03	
	0.79	0.83	0.82	
bpMRI	PNetCAI	0.28	0.28	0.28

Table 4. Hold-out test set results. For each pairwise evaluation, the average Dice, Recall and Precision performances are presented. The best Recall result for each dataset per sequence combination is highlighted in bold for easier comparison.

	(a) Tested on PNet			(b) Tested on PI-CAI			(c) Tested on PNetCAI			bpMRI
	PNet	PI-CAI	PNetCAI	PNet	PI-CAI	PNetCAI	PNet	PI-CAI	PNetCAI	
PNet		2.5e-6	0.584		1.4e-18	1.1e-14		5.7e-5	3.4e-10	
PI-CAI	0.919		4.7e-8	0.0002		0.237	0.009		0.083	
PNetCAI	0.184	0.223		0.0004	0.890		0.0009	0.478		

T2W

Table 5. T-test *p*-values for the pairwise comparison of the Dice scores presented in Table 4. Significant differences (*p*-value < 0.01) marked as green.

Trade-off between avoiding biopsies and dangerous underestimates

To understand whether the best performing model—trained on bpMRI PNetCAI data—could be used as a CAD system for the effective reduction of biopsies (i.e. correctly predicting when an individual has no aggressive lesions), we first determined how many lesions were present in each case and calculated the number of detected lesions for all models. We then performed a simple experiment assigning lesions to one of six categories:

- Correct + avoided biopsy: if no lesions were present and the model correctly estimated this (i.e. recommended avoiding an unnecessary biopsy);
- Correct: if one or more lesions were present and the model correctly estimated the number of lesions
- Overestimate: if one or more lesions were present and the model overestimated the number of lesions

Trained on	Tested on									
	PNet					PI-CAI				
T2W	PNet	14	5	3	2	12	3	1	4	
	PI-CAI	11	6	5	2	21	7	5	6	
	PNetCAI	15	7	5	2	22	3	3	6	
bpMRI	PNet	13	8	5	2	11	4	2	3	
	PI-CAI	7	5	1	0	27	13	5	6	
	PNetCAI	15	9	5	2	27	11	5	6	
# lesions		23	9	5	2	33	14	6	7	
ISUP		2	3	4	5					

Table 6. Hold-out test set results stratified by the ISUP grade of the lesions. For each pairwise evaluation, the number of predicted lesions is compared to the total number of lesions. The best-performing model (most successful detections) for each dataset per sequence combination is highlighted in bold.

- Overestimate + unnecessary biopsy: if no lesions were present and the model overestimated the number of lesions (i.e. recommended an unnecessary biopsy)
- Underestimate: if two or more lesions were present and the model estimated a number of lesions between one and excluding the correct number of lesions
- Dangerous underestimate: if two or more lesions were present and the model detected no lesions (i.e. recommended avoiding a necessary biopsy)

This categorization system leads to a consistent trade-off between overestimating the number of lesions while recommending an unnecessary biopsy and avoiding unnecessary biopsies (Fig. 3); in other words, these systems could have the potential of reducing the number of biopsies but this set up has to be carefully considered as it could also result in avoiding biopsies for patients who would require them. A concerning aspect of this analysis is that only in one instance—PNetCAI models tested on PNet data—does it fulfill the task of reducing the number of biopsies without missing any relevant predictions (Table 7).

Additionally, there is a consistently large number of recommended unnecessary biopsies—indeed, for bpMRI PNetCAI models tested on PNet data, 54.05% of cases ($n = 120$) would have an unnecessary biopsy recommended, while only 17.76% of cases ($n = 27$) would avoid an unnecessary biopsy. This can have a negative impact on the well-being of individuals who have to undergo these unnecessary biopsies.

Prospective validation of a simulated clinical decision system

As noted above, an automated system based solely on our models would either lead to dangerous underestimates (i.e. no lesion detected when a lesion was present) or an excess of unnecessary biopsies. To curtail these negative aspects, we devised a clinical decision protocol requiring the interaction of two different decisions, one made by a radiologist (i.e. determine that an individual should have a follow-up biopsy) and the other made by our CAD system: (i) if a radiologist does not recommend a follow-up biopsy, none is performed; (ii) if a radiologist recommends a follow-up biopsy and our model recommends no follow-up biopsy, this is not performed; and (iii) if a radiologist and our model recommend a follow-up biopsy, a biopsy is performed. In effect, this is the ideal case scenario for a model which is highly sensitive but whose specificity is relatively low (i.e. the model produces an excess of false positives).

To avoid the self-fulfilling prophecy of developing models and testing them on the same data, we used a ProstateNet prospective cohort of 73 cases (21 aggressive PCa) to determine whether such a strategy could be beneficial. In terms of prospective segmentation and detection performance, these models perform similarly to those trained and tested with retrospective data (Table 8). Lastly, and most importantly, our results show that using a combined CAD system as described above would indeed lead to a reduction of unnecessary biopsies (21.9% of cases [$n=16$]; Fig. 4) without increasing the dangerous underestimates.

Finally, we assess whether these models are capable of performing reasonably well across different confidence thresholds and whether they can be reliably used at the lesion level. As highlighted in Fig. A.3, these models perform better when confidence thresholds are lower (AUROC is consistently higher when such is the case). Additionally, there is limited applicability for these models as lesion segmentation tools due to their relatively high number of false positives.

Determinants of performance

To better understand performance (Dice scores), we analysed distinct factors—lesion size and whether annotations were derived by an AI or by a radiologist. ProstateNet and PI-CAI have different distributions of lesion size (Fig. 5B), with ProstateNet presenting lesions larger than those in PI-CAI. Indeed, at a significance threshold of 0.05, there is a significant Dice difference between below and above median lesions for both datasets (Fig. 5A). While more evident in the ProstateNet dataset, both sets of data exhibit a size bias where larger lesions are easier to segment. Given that some lesions in PI-CAI are generated by an AI model²⁶, we compared the Dice scores between lesions annotated by AI and by radiologists, showing that the former lead to higher Dice scores than the latter ($p = 7.6e - 5$; Fig. 5C). In Fig. 5D, we highlight a more comprehensive vision of these results.

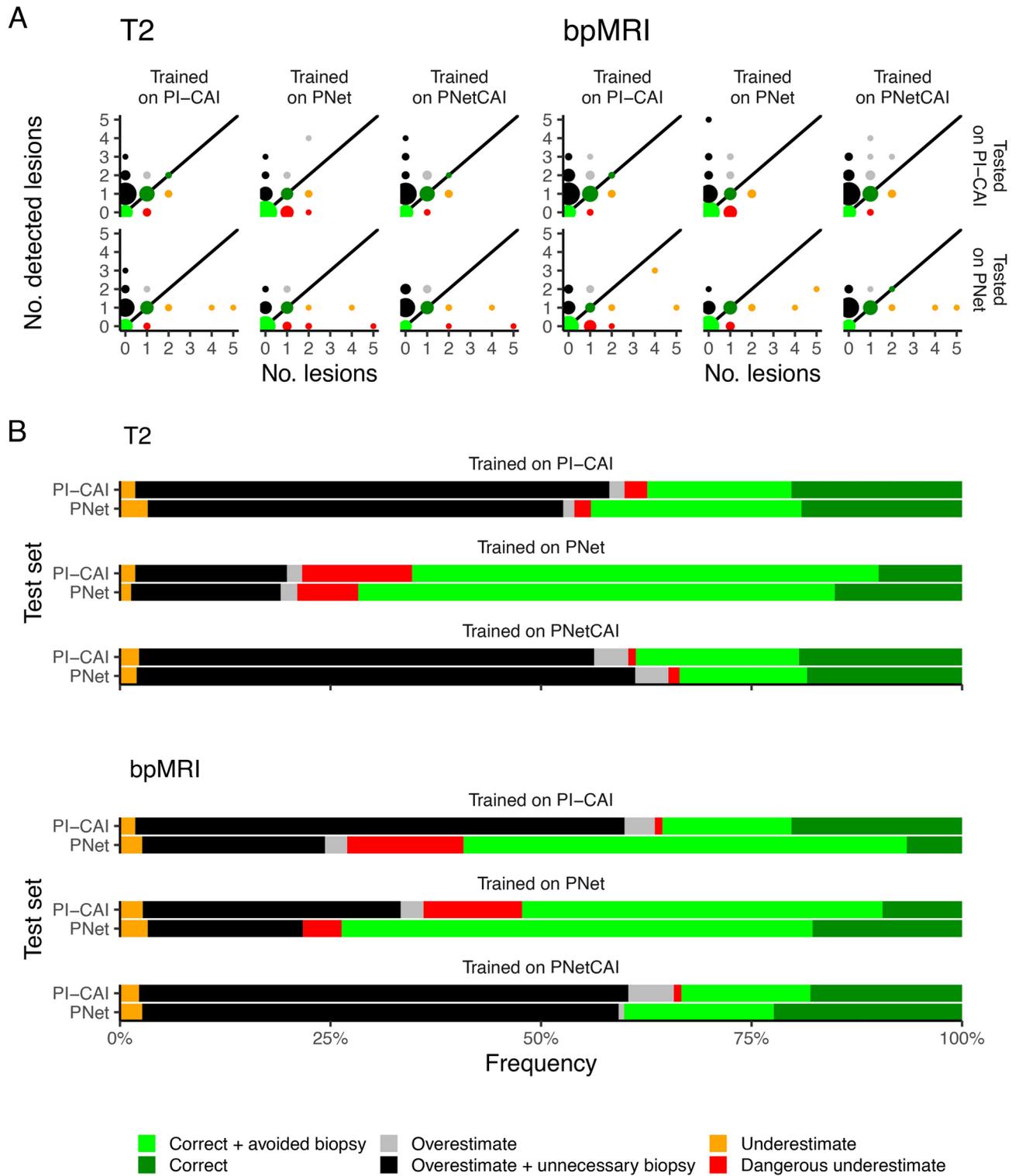


Fig. 3. Distribution of CAD recommendations, stratified by training and testing dataset. **(A)** Distribution of annotated (no. of lesions in x-axes) and detected (no. of detected lesions in y-axes) lesions. **(B)** Relative frequencies of different predictions from the CAD system. For both **(A,B)** the colors correspond to a classification relating to whether or not this recommendation would lead to a change in the diagnostic algorithm proposed to the patient.

Trained on	Tested on	Total	Avoided biopsy	Correct	Unnecessary biopsy	Underestimate	Dangerous underestimate	Overestimate
PNet			55.92% (85)	17.76% (27)	18.42% (28)	3.29% (5)	4.61% (7)	0% (0)
PI-CAI	PNet	152	52.63% (80)	6.58% (10)	21.71% (33)	2.63% (4)	13.82% (21)	2.63% (4)
PNetCAI			17.76% (27)	22.37% (34)	56.58% (86)	2.63% (4)	0% (0)	0.66% (1)
PNet			42.79% (95)	9.46% (21)	30.63% (68)	2.70% (6)	11.71% (26)	2.70% (6)
PI-CAI	PI-CAI	222	15.32% (34)	20.27% (45)	58.11% (129)	1.80% (4)	0.90% (2)	3.60% (8)
PNetCAI			15.32% (34)	18.02% (40)	58.11% (129)	2.25% (5)	0.90% (2)	5.41% (12)

Table 7. Absolute and relative frequency of bpMRI AI system recommendations, stratified by training and testing dataset. Counts are displayed between brackets after percentages.

Metric	Modality	PNet	PC	PNetCAI
Recall	T2W	0.71	0.52	0.90
Precision	T2W	0.48	0.25	0.30
# Detected lesions	T2W	15	11	19
Recall	bpMRI	0.86	0.48	1.00
Precision	bpMRI	0.54	0.23	0.34
# Detected lesions	bpMRI	18	10	21
# Lesions				21

Table 8. Prospective cohort results. For each model, per sequence, the average Dice, Recall and Precision performances are presented. The best Recall scores are highlighted in bold for easier comparison.

Finally, to acquire a qualitative understanding of prediction quality, we analyzed a subset of true positive and false negative detections at the lesion level for our best-performing model—trained on bpMRI PNetCAI data. Figure 6 offers a concise overview of our analysis, while Figs. A.1 and A.2 present a comprehensive depiction. As highlighted in Fig. A.1, true positives typically encompass all or nearly all of the lesions as annotated by expert radiologists. This is what is expected of such CAD systems, providing information regarding the general area where it thinks the lesion is located to guide the radiologist. When considering negative examples (Fig. A.2), there is a trend—while the lesion annotated by expert radiologists may be missed, the models identify another likely lesion somewhere else in the prostate. In summary, the conclusions derived from our qualitative analysis are as follows:

- In each case, our model detected additional existing lesions and/or cysts. Although these were marked as missed cases due to insufficient overlap with the ground truth mask, they nonetheless correctly identified other lesions as aggressive, demonstrating significant clinical value for a CAD system.
- In some instances (Fig. A.2), with the fourth example being the only visible one in this set of slices, our model correctly identified the area of interest despite low confidence and probability scores. This demonstrates the utility of our model in guiding radiologists to significant areas regardless of the displayed probability.

Discussion

In this work, we posit a hybrid computer-aided diagnosis (CAD) system combining radiologists and an automatic lesion detection model, which can reduce the number of unnecessary biopsies in the diagnosis of aggressive prostate cancer (ISUP>1) in the general population of patients undergoing biparametric MRI for prostate cancer diagnosis. Through a simulated clinical feasibility scenario, a reduction of approximately 20% of unnecessary biopsies was achieved, with a prospective validation showing that this does not lead to a reduction in the number of detected prostate cancer cases. Ultimately, we highlight how deep-learning methods can assist in the reduction of unnecessary biopsies without leading to decreased sensitivity. This has the potential to reduce patient discomfort and complications following biopsies.

Largely, most CAD systems of the sort seek to solve a similar, albeit separate problem — that of detecting undiagnosed prostate cancer cases with the objective of increasing sensitivity by reducing the amount of false negatives; our approach considers a different problem — that of reducing the number of unnecessary biopsies (i.e. reducing false positives). Indeed, this is also a considerable problem, as a 2019 meta-review showed that the pooled sensitivity for PI-RADS 2.1 was approximately 91% (95% CI=83%-95%)⁴. Works seeking to automate or partially automate prostate cancer diagnosis contemplate strategies focusing either on the detection of lesions with a sufficiently high PI-RADS score (i.e. 3 or 4)⁴⁰ or in the detection of lesions with a confirmed aggressive

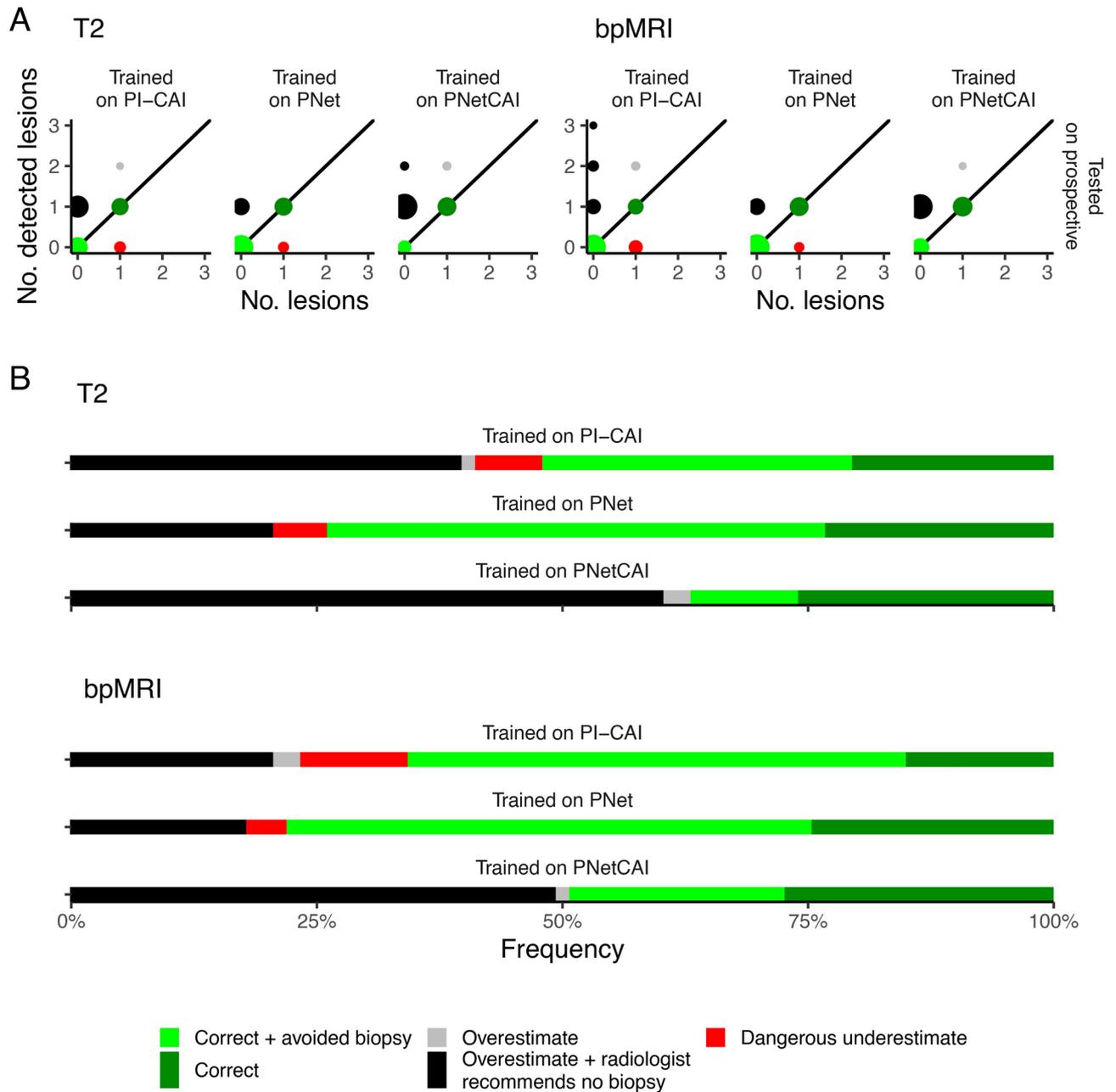


Fig. 4. Distribution of CAD recommendations, stratified by training for the prospective dataset. **(A)** Distribution of annotated (no. of lesions in x-axes) and detected (no. of detected lesions in y-axes) lesions. **(B)** Relative frequencies of different predictions from the CAD system. For both **(A,B)** the colors correspond to a classification relating to whether or not this recommendation would lead to a change in the diagnostic algorithm proposed to the patient.

histological grade (ISUP>1)^{25,41,42}. The former has the obvious advantage of requiring no biopsy for training, but hinders the clinical applicability evidenced by the latter. Some of these strategies also incorporate a human-in-the-loop setup, which is more similar to the study design we introduce here⁴³. The relevant performance metric which we can compare between our work and previous works is the Recall—we observed a Recall of 82% for models trained/tested on PNetCAI, slightly lower to what has been previously reported (87.2%⁴³, 89.4%²⁵, 93%⁴²). However, we note that these studies are trained/tested on a relatively small number of clinical centers (4 or fewer)^{25,42,43} (which greatly reduces the variability of the data), do not provide confirmation of prospective validation, and do not study the impact of using diverse training datasets on performance. Given the previously reported drop in performance when transferring models between different datasets^{10,11,44} and the fact that models (clinical and otherwise) tend to suffer from temporal degradation^{45–47}, such assessments are of paramount importance. Finally, and to the best of our knowledge, our work offers a unique analysis of performance differences when considering lesion size and annotation types, thus better contextualizing results.

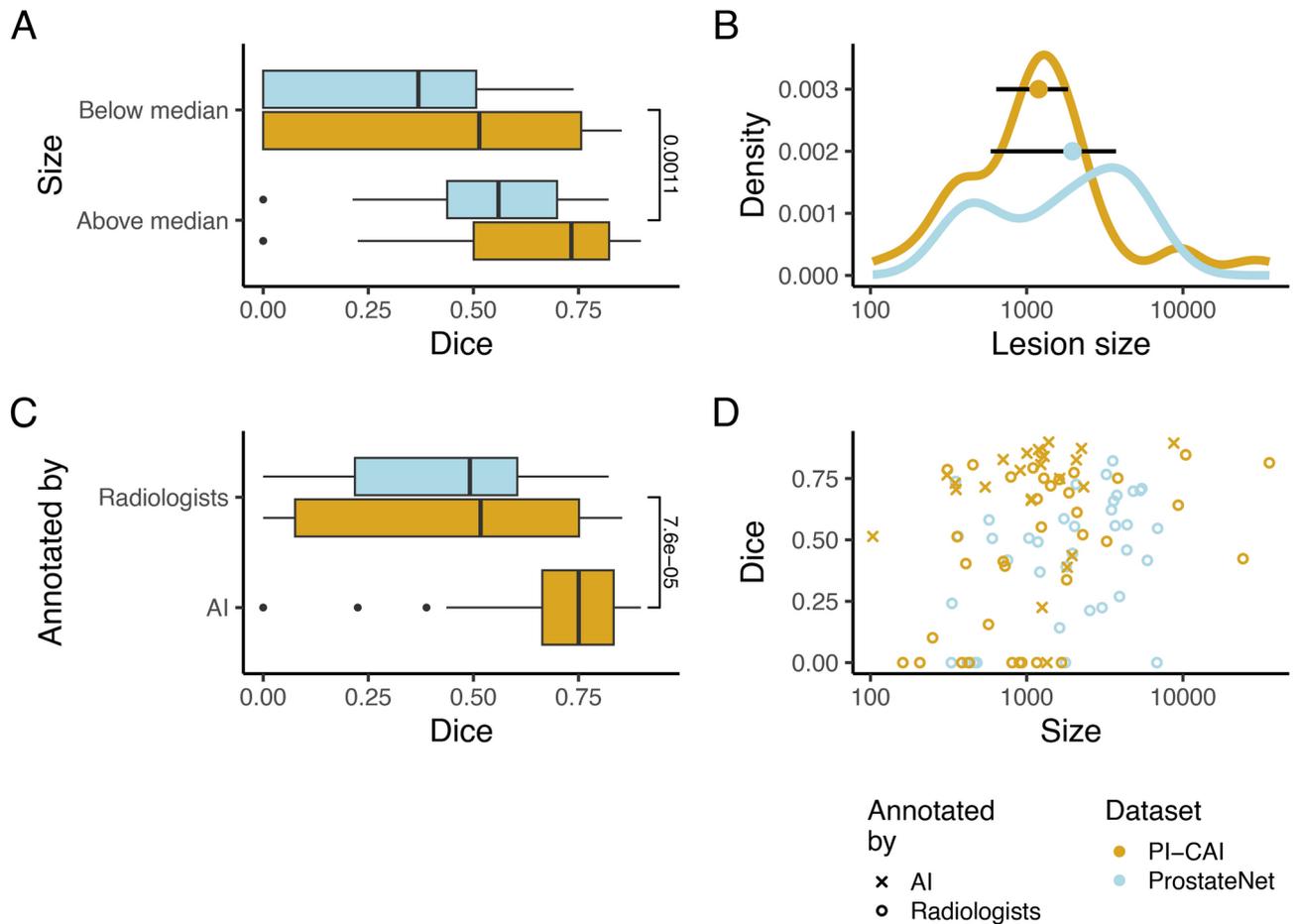


Fig. 5. Effect of lesion size and annotation type on performance for the best performing model (bpMRI). (A) Performance distribution stratified by dataset and lesion size (below or above median). (B) Distribution density for lesion sizes across both datasets. Circles represent the median value while black horizontal lines represent the range between the 1st and 3rd quartiles. (C) Performance distribution stratified by dataset and annotation type (whether the lesion was annotated by a radiologist or by an AI model). (D) Comparison of lesion size with Dice. Each point corresponds to a case, different shapes correspond to different annotation types. Across all plots, golden and blue correspond to PI-CAI and ProstateNet, respectively. p -values in (A,C) correspond to a two-sided Wilcoxon test.

This work has some caveats—the simulated clinical scenario does not allow us to estimate the effect of real-world agents (i.e. medical doctors) interacting with such a CAD system. This may lead to optimistic results as automation bias (when users excessively trust the output of automatic CAD systems⁴⁸) can lead to unforeseen outcomes as radiologists may trust excessively in wrong predictions made by CAD systems⁴⁹. It should also be highlighted that, while the best performing model detects all important cases in ProstateNet both retrospectively and prospectively (Figs. 3, 4), not all index lesions are detected, which can cause confusion when results are interpreted in a clinical setting; this is in part largely associated with how these datasets are annotated — indeed, radiologists are tasked with segmenting at least the index lesion, leading to a fair degree of heterogeneity in the annotations. Additionally, performance is relatively poor when we consider the specificity of these models; while this can be improved through the assistance of a radiologist, it should be noted that additional sources of false positive reduction should be taken, such as an auxiliary classification of lesion candidates⁵⁰ or zone-specific PSA density⁵¹. Furthermore, our approach does not focus on lesion location — particularly, we perform predictions at the patient, rather than at the lesion level — so further studies on this are necessary. Finally, it should be noted that there is no guarantee that nnUNet is the best performing model (“No Free Lunch” theorem) — earlier works have suggested that other models may be better performing than nnUNet for prostate lesion segmentation⁵⁰, so a more comprehensive assessment with other models could be important.

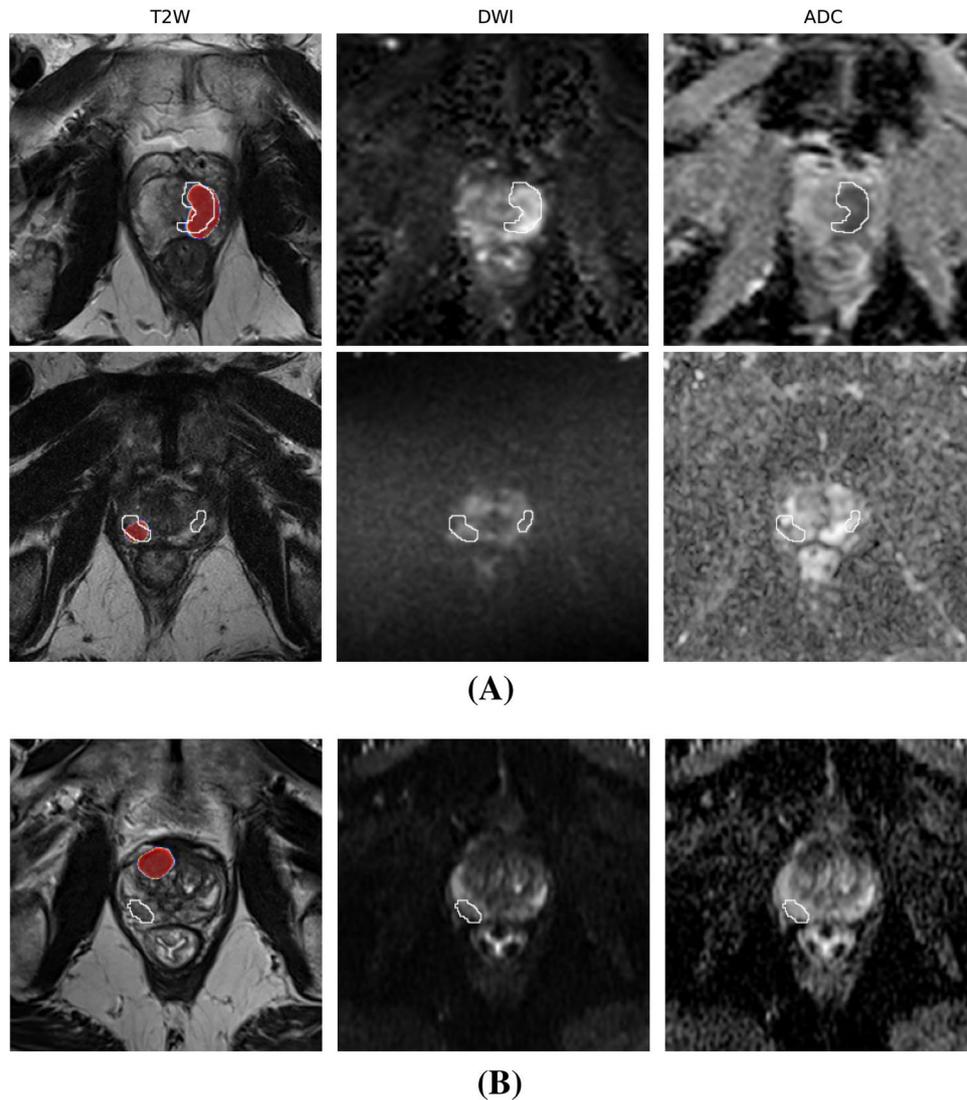


Fig. 6. Examples of correctly detected and missed cases. **(A)** Correctly classified and detected lesions. Each row represents a different case selected at random from the correctly detected samples, and the slices shown are those where the index lesion ground truth is most visible in the sequences, **(B)** missed detected example. The slice choice is the same as the one described previously. For both sets of examples, the ground truth is represented by the white outline, allowing for the view of the target region, and the probability maps are only displayed in the T2W images as to not cover the hyper- and hypo- intense areas of both DWI and ADC sequences.

Data availability

The datasets generated and/or analysed during the current study are available in the PI-CAI repository, <https://zenodo.org/records/6624726>. The datasets generated and/or analysed during the current study are not publicly available due to data privacy laws but are available from the corresponding author on reasonable request.

Received: 18 July 2024; Accepted: 23 April 2025

Published online: 30 April 2025

References

1. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2022. *CA Cancer J. Clin.* **72**, 7–33 (2022).
2. Resnick, M. J. et al. Long-term functional outcomes after treatment for localized prostate cancer. *N. Engl. J. Med.* **368**, 436–445. <https://doi.org/10.1056/NEJMoa1209978> (2013).
3. Scott, R., Misser, S. K., Cioni, D. & Neri, E. PI-RADS v2.1: What has changed and how to report. *SA J. Radiol.* **25**, 2062 (2021).
4. 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**, CD012663 (2019).
5. Cao, R. et al. Performance of deep learning and genitourinary radiologists in detection of prostate cancer using 3-T multiparametric magnetic resonance imaging. *J. Magn. Reson. Imaging* **54**, 474–483 (2021).

6. Steenbergen, P. et al. Prostate tumor delineation using multiparametric magnetic resonance imaging: Inter-observer variability and pathology validation. *Radiother. Oncol.* **115**, 186–190 (2015).
7. Chen, M. Y., Woodruff, M. A., Dasgupta, P. & Rukin, N. J. Variability in accuracy of prostate cancer segmentation among radiologists, urologists, and scientists. *Cancer Med.* **9**, 7172–7182. <https://doi.org/10.1002/cam4.3386> (2020).
8. Kushol, R., Parnianpour, P., Wilman, A. H., Kalra, S. & Yang, Y.-H. Effects of MRI scanner manufacturers in classification tasks with deep learning models. *Sci. Rep.* **13**, 16791 (2023).
9. Netzer, N. et al. Fully automatic deep learning in bi-institutional prostate magnetic resonance imaging: Effects of cohort size and heterogeneity. *Invest. Radiol.* **56**, 799–808 (2021).
10. Meglič, J., Sunoqrot, M. R. S., Bathen, T. F. & Elschof, M. Label-set impact on deep learning-based prostate segmentation on MRI. *Insights Imaging* **14**. <https://doi.org/10.1186/s13244-023-01502-w> (2023).
11. Rodrigues, N. M. et al. Analysis of domain shift in whole prostate gland, zonal and lesions segmentation and detection, using multicentric retrospective data. *Comput. Biol. Med.* **171**, 108216. <https://doi.org/10.1016/j.compbiomed.2024.108216> (2024).
12. Rodrigues, A. et al. Value of handcrafted and deep radiomic features towards training robust machine learning classifiers for prediction of prostate cancer disease aggressiveness. *Sci. Rep.* **13**. <https://doi.org/10.1038/s41598-023-33339-0> (2023).
13. Pachetti, E. & Colantonio, S. 3d-vision-transformer stacking ensemble for assessing prostate cancer aggressiveness from t2w images. *Bioengineering* **10**. <https://doi.org/10.3390/bioengineering10091015> (2023).
14. Bernatz, S. et al. Comparison of machine learning algorithms to predict clinically significant prostate cancer of the peripheral zone with multiparametric mri using clinical assessment categories and radiomic features. *Eur. Radiol.* **30**, 6757–6769 (2020).
15. Pellicer-Valero, O. J. et al. Deep learning for fully automatic detection, segmentation, and Gleason grade estimation of prostate cancer in multiparametric magnetic resonance images. [arXiv:2103.12650](https://arxiv.org/abs/2103.12650) (2022).
16. Dai, Z. et al. Segmentation of the prostatic gland and the intraprostatic lesions on multiparametric magnetic resonance imaging using mask region-based convolutional neural networks. *Adv. Radiat. Oncol.* **5**, 473–481 (2020).
17. Cao, R. et al. Prostate cancer detection and segmentation in multi-parametric MRI via cnn and conditional random field. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 1900–1904. <https://doi.org/10.1109/ISBL2019.8759584> (2019).
18. Hambarde, P. et al. Prostate lesion segmentation in MR images using radiomics based deeply supervised u-net. *Biocybern. Biomed. Eng.* **40**, 1421–1435. <https://doi.org/10.1016/j.bbe.2020.07.011> (2020).
19. Cao, R. et al. Joint prostate cancer detection and Gleason score prediction in MP-MRI via focalnet. *IEEE Trans. Med. Imaging* **38**, 2496–2506. <https://doi.org/10.1109/TMI.2019.2901928> (2019).
20. Hosseinzadeh, M. et al. Deep learning-assisted prostate cancer detection on bi-parametric MRI: Minimum training data size requirements and effect of prior knowledge. *Eur. Radiol.* **32**, 2224–2234 (2022).
21. Seetharaman, A. et al. Automated detection of aggressive and indolent prostate cancer on magnetic resonance imaging. *Med. Phys.* **48**, 2960–2972 (2021).
22. Khan, Z., Yahya, N., Alsaih, K., Ali, S. S. A. & Meriaudeau, F. Evaluation of deep neural networks for semantic segmentation of prostate in T2W MRI. *Sensors* **20**, 3183. <https://doi.org/10.3390/s20113183> (2020).
23. Khan, Z., Yahya, N., Alsaih, K., Al-Hiyali, M. I. & Meriaudeau, F. Recent automatic segmentation algorithms of MRI prostate regions: A review. *IEEE Access* **9**, 97878–97905. <https://doi.org/10.1109/access.2021.3090825> (2021).
24. Bashkanov, O. et al. Automatic detection of prostate cancer grades and chronic prostatitis in biparametric MRI. *Comput. Methods Programs Biomed.* **239**, 107624 (2023).
25. Saha, A. et al. Artificial intelligence and radiologists in prostate cancer detection on MRI (PI-CAI): An international, paired, non-inferiority, confirmatory study. *Lancet Oncol.* (2024).
26. Saha, A. et al. *The PI-CAI Challenge: Public Training and Development Dataset*. <https://doi.org/10.5281/zenodo.6517398> (2022).
27. Armato, S. G. et al. PROSTATEX challenges for computerized classification of prostate lesions from multiparametric magnetic resonance images. *J. Med. Imaging (Bellingham)* **5**, 044501 (2018).
28. Engels, R. R., Israël, B., Padhani, A. R. & Barentsz, J. O. Multiparametric magnetic resonance imaging for the detection of clinically significant prostate cancer: What urologists need to know. Part 1: Acquisition. *Eur. Urol.* **77**, 457–468. <https://doi.org/10.1016/j.eururo.2019.09.021> (2020).
29. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**, 203–211. <https://doi.org/10.1038/s41592-020-01008-z> (2020).
30. Zhu, Q., Du, B., Turkbey, B. I., Choyke, P. L. & Yan, P. Deeply-supervised cnn for prostate segmentation. In *2017 International Joint Conference on Neural Networks (IJCNN)*. 178–184 (2017).
31. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. Vol. 32. 8024–8035 (Curran Associates, Inc., 2019).
32. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. [arXiv:1606.00915](https://arxiv.org/abs/1606.00915) (2017).
33. Rodrigues, N. M., Silva, S., Vanneschi, L. & Papanikolaou, N. A comparative study of automated deep learning segmentation models for prostate MRI. *Cancers* **15**. <https://doi.org/10.3390/cancers15051467> (2023).
34. Dosovitskiy, A. et al. An image is worth 16 x 16 words: Transformers for image recognition at scale. <https://doi.org/10.48550/ARXIV.2010.11929> (2020).
35. Murugesan, B., Liu, B., Galdran, A., Ayed, I. B. & Dolz, J. Calibrating segmentation networks with margin-based label smoothing. *Med. Image Anal.* **87**, 102826. <https://doi.org/10.1016/j.media.2023.102826> (2023).
36. Müller, R., Kornblith, S. & Hinton, G. *When Does Label Smoothing Help?* (Curran Associates Inc., 2019).
37. Bosma, J. S. et al. Semisupervised learning with report-guided pseudo labels for deep learning-based prostate cancer detection using biparametric MRI. *Radiol. Artif. Intell.* **5**, e230031 (2023).
38. Yeghiazaryan, V. & Voiculescu, I. Family of boundary overlap metrics for the evaluation of medical image segmentation. *J. Med. Imaging (Bellingham)* **5**, 015006 (2018).
39. Maier, O. et al. loli/medpy: Medpy 0.4.0. <https://doi.org/10.5281/zenodo.2565940> (2019).
40. Hosseinzadeh, M. et al. Deep learning-assisted prostate cancer detection on bi-parametric MRI: Minimum training data size requirements and effect of prior knowledge. *Eur. Radiol.* **32**, 2224–2234 (2022).
41. Zhao, L. et al. Predicting clinically significant prostate cancer with a deep learning approach: A multicentre retrospective study. *Eur. J. Nucl. Med. Mol. Imaging* **50**, 727–741 (2023).
42. Hamm, C. A. et al. Interactive explainable deep learning model informs prostate cancer diagnosis at MRI. *Radiology* **307**, e222276 (2023).
43. Yu, R. et al. PI-RADS AI: Introducing a new human-in-the-loop AI model for prostate cancer diagnosis based on MRI. *Br. J. Cancer* **128**, 1019–1029 (2023).
44. Yu, A. C., Mohajer, B. & Eng, J. External validation of deep learning algorithms for radiologic diagnosis: A systematic review. *Radiol. Artif. Intell.* **4**, e210064 (2022).
45. Bedoya, A. D. et al. Machine learning for early detection of sepsis: An internal and temporal validation study. *JAMIA Open* **3**, 252–260 (2020).
46. Foote, H. P. et al. Development and temporal validation of a machine learning model to predict clinical deterioration. *Hosp. Pediatr.* **14**, 11–20 (2024).
47. Vela, D. et al. Temporal quality degradation in AI models. *Sci. Rep.* **12**, 11654 (2022).

48. Kostick-Quenet, K. M. & Gerke, S. AI in the hands of imperfect users. *NPJ Digit. Med.* **5**, 197 (2022).
49. Dratsch, T. et al. Automation bias in mammography: The impact of artificial intelligence BI-RADS suggestions on reader performance. *Radiology* **307**, e222176 (2023).
50. Saha, A., Hosseinzadeh, M. & Huisman, H. End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effects of attention mechanisms, clinical priori and decoupled false positive reduction. *Med. Image Anal.* **73**, 102155 (2021).
51. Hamm, C. A. et al. Reduction of false positives using zone-specific prostate-specific antigen density for prostate MRI-based biopsy decision strategies. *Eur. Radiol.* (2024).
52. Hamm, C. A. et al. Reduction of false positives using zone-specific prostate-specific antigen density for prostate MRI-based biopsy decision strategies. *Eur. Radiol.* <https://doi.org/10.1007/s00330-024-10700-z> (2024).

Acknowledgements

This work was supported by FCT through the LASIGE Research Unit, ref. UID/000408/2025, and Nuno Rodrigues PhD Grant10.54499/2021.05322.BD (<https://doi.org/10.54499/2021.05322.BD>). Ana Sofia and José Guilherme de Almeida were supported by the European Union H2020: ProCancer-I project (EU grant 952159). This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project UIDB/04152/2020 (<https://doi.org/10.54499/UIDB/04152/2020>) -Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS.

Author contributions

Conceptualization, N.M.R., J.G.A.; methodology, N.M.R., J.G.A.; software, N.M.R., J.G.A.; validation, N.M.R.; formal analysis, N.M.R., J.G.A.; investigation, N.M.R., J.G.A.; resources, N.M.R., J.G.A.; data curation, N.M.R., J.G.A., A.C.V., A.G., C.B., I.S., J.I., S.B., P.C.; writing-original draft preparation, N.M.R., J.G.A.; writing-review and editing, S.S., L.V., N.P., P.C.; visualization, N.M.R., J.G.A.; supervision, S.S., L.V., N.P.; project administration, N.P., C.M., M.T., K.M., D.R.; funding acquisition, N.P., M.T., K.M., D.R.; all authors have read and agreed to the published version of the manuscript.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-99795-y>.

Correspondence and requests for materials should be addressed to N.M.R.

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.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025

The ProCancer-I Consortium

Manolis Tsiknakis⁷, Kostas Marias⁷, Stelios Sfakianakis⁷, Varvara Kalokyri⁷, Eleftherios Trivizakis⁷, Grigorios Kalliatakis⁷, Avtantil Dimitriadis⁷, Dimitris Fotiadis⁷, Nikolaos Tachos⁷, Eugenia Mylona⁷, Dimitris Zaridis⁷, Charalampos Kalantzopoulos⁷, Nikolaos Papanikolaou¹, José Guilherme Almeida¹, Ana Castro Verde¹, Ana Carolina Rodrigues¹, Nuno Rodrigues¹, Miguel Chambel¹, Henkjan Huisman¹¹, Maarten de Rooij¹¹, Anindo Saha¹¹, Jasper J. Twilt¹¹, Jurgen Futterer¹¹, Luis Martí-Bonmati¹², Leonor Cerdá-Alberich¹², Gloria Ribas¹², Silvia Navarro¹², Manuel Marfil¹², Emanuele Neri¹³, Giacomo Aringhieri¹³, Lorenzo Tumminello¹³, Vincenzo Mendola¹³, Deniz Akata¹⁴, Mustafa Özmen¹⁴, Ali Devrim Karaosmanoglu¹⁴, Firat Atak¹⁴, Musturay Karcaaltincaba¹⁴, Joan C. Vilanova¹⁵, Jurgita Usinskiene¹⁶, Ruta Briediene¹⁶, Audrius Untanas¹⁶, Kristina Slidevska¹⁶, Katsaros Vasilis¹⁷, Georgiou Georgios¹⁷, Dow-Mu Koh¹⁸, Robby Emsley¹⁸, Sharon Vit¹⁸, Ana Ribeiro¹⁸, Simon Doran¹⁸, Tiaan Jacobs¹⁸, Gracián García-Martí¹⁹, Daniele Regge⁸, Valentina Giannini⁸, Simone Mazzetti⁸, Giovanni Cappello⁸, Giovanni Maimone⁸, Valentina Napolitano⁸, Sara Colantonio²⁰, Maria Antonietta Pascali²⁰, Eva Pachetti²⁰, Giulio del Corso²⁰, Danila Germanese²⁰, Andrea Berti²⁰, Gianluca Carloni²⁰, Jayashree Kalpathy-Cramer²¹, Christopher Bridge²¹, Joao Correia²², Walter

Hernandez²², Zoi Giavri²³, Christos Pollalis²³, Dimitrios Agraniotis²³, Ana Jiménez Pastor²⁴, Jose Munuera Mora²⁴, Clara Saillant²⁵, Theresa Henne²⁵ & Rodessa Marquez²⁵

¹¹Radboud, Nijmegen, Netherlands. ¹²HULAFE-Biomedical Imaging Research Group, Medical Imaging Department, Instituto de Investigación Sanitaria La Fe, Hospital Universitari i Politècnic La Fe, Valencia, Spain. ¹³Academic Radiology, Department of Translational Research, University of Pisa, Pisa, Italy. ¹⁴Department of Radiology, Hacettepe, Ankara, Turkey. ¹⁵Department of Radiology (IDI), Institute of Biomedical Research of Girona Dr. Josep Trueta (IDIBGI), Girona, Spain. ¹⁶National Cancer Institute, Vilnius, Lithuania. ¹⁷General Anti-Cancer and Oncological Hospital of Athens, Athens, Greece. ¹⁸Radiology & AI Research Hub, Division of Radiotherapy and Imaging, The Institute of Cancer Research, The Royal Marsden NHS Foundation Trust, London, UK. ¹⁹Quirónsalud Hospital/ CIBERSAM, Valencia, Spain. ²⁰Institute of Information Science and Technologies of the National Research Council of Italy, Pisa, Italy. ²¹Mass General Hospital, Boston, USA. ²²B3D, Birmingham, UK. ²³Advantis, Athens, Greece. ²⁴Quibim, S.L., Valencia, Spain. ²⁵Univie, Vienna, Austria.