HHS Public Access

Author manuscript

Acad Radiol. Author manuscript; available in PMC 2023 October 13.

Published in final edited form as:

Acad Radiol. 2022 August ; 29(8): 1159–1168. doi:10.1016/j.acra.2021.08.019.

A Cascaded Deep Learning–Based Artificial Intelligence Algorithm for Automated Lesion Detection and Classification on Biparametric Prostate Magnetic Resonance Imaging

Sherif Mehralivand, MD,
Molecular Imaging Branch, National Cancer Institute, National Institutes of Health, 10 Center Dr., MSC 1182, Building 10, Room B3B85, Bethesda, Maryland

Dong Yang, PhD,
NVIDIA Corporation, Santa Clara, California

Stephanie A. Harmon, PhD,
Molecular Imaging Branch, National Cancer Institute, National Institutes of Health, 10 Center Dr., MSC 1182, Building 10, Room B3B85, Bethesda, Maryland

Daguang Xu, PD,
NVIDIA Corporation, Santa Clara, California

Ziyue Xu, PhD,
NVIDIA Corporation, Santa Clara, California

Holger Roth, PhD,
NVIDIA Corporation, Santa Clara, California

Samira Masoudi, PhD,
Molecular Imaging Branch, National Cancer Institute, National Institutes of Health, 10 Center Dr., MSC 1182, Building 10, Room B3B85, Bethesda, Maryland

Thomas H. Sanford, MD,
Department of Urology, SUNY Upstate Medical University, Syracuse, New Yor

Deepak Kesani, DO,
Molecular Imaging Branch, National Cancer Institute, National Institutes of Health, 10 Center Dr., MSC 1182, Building 10, Room B3B85, Bethesda, Maryland

Nathan S. Lay, PhD,
Molecular Imaging Branch, National Cancer Institute, National Institutes of Health, 10 Center Dr., MSC 1182, Building 10, Room B3B85, Bethesda, Maryland

Maria J. Merino, MD,

Address correspondence to: B.T., turkbeyi@mail.nih.gov.

This project is based on a Cooperative Research and Development Agreement (CRADA) between the National Institutes of Health and the NVIDIA Corporation.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.acra.2021.08.019.

Laboratory of Pathology, National Cancer Institute, National Institutes of Health, Bethesda, Maryland

Bradford J. Wood, MD,
Center for Interventional Oncology, National Cancer Institute, National Institutes of Health, Bethesda, Maryland; Department of Radiology, Clinical Center, National Institutes of Health, Bethesda, Maryland

Peter A. Pinto, MD,
Urologic Oncology Branch, National Cancer Institute, National Institutes of Health, Bethesda, Maryland

Peter L. Choyke, MD,
Molecular Imaging Branch, National Cancer Institute, National Institutes of Health, 10 Center Dr., MSC 1182, Building 10, Room B3B85, Bethesda, Maryland

Baris Turkbey, MD
Molecular Imaging Branch, National Cancer Institute, National Institutes of Health, 10 Center Dr., MSC 1182, Building 10, Room B3B85, Bethesda, Maryland

Abstract

Rationale and objectives: Prostate MRI improves detection of clinically significant prostate cancer; however, its diagnostic performance has wide variation. Artificial intelligence (AI) has the potential to assist radiologists in the detection and classification of prostatic lesions. Herein, we aimed to develop and test a cascaded deep learning detection and classification system trained on biparametric prostate MRI using PI-RADS for assisting radiologists during prostate MRI read out.

Materials and Methods: T2-weighted, diffusion-weighted (ADC maps, high b value DWI) MRI scans obtained at 3 Tesla from two institutions (n = 1043 in-house and n = 347 Prostate-X, respectively) acquired between 2015 to 2019 were used for model training, validation, testing. All scans were retrospectively reevaluated by one radiologist. Suspicious lesions were contoured and assigned a PI-RADS category. A 3D U-Net-based deep neural network was used to train an algorithm for automated detection and segmentation of prostate MRI lesions. Two 3D residual neural network were used for a 4-class classification task to predict PI-RADS categories 2 to 5 and BPH. Training and validation used 89% (n = 1290 scans) of the data using 5 fold cross-validation, the remaining 11% (n = 150 scans) were used for independent testing. Algorithm performance at lesion level was assessed using sensitivities, positive predictive values (PPV), false discovery rates (FDR), classification accuracy, Dice similarity coefficient (DSC). Additional analysis was conducted to compare AI algorithm's lesion detection performance with targeted biopsy results.

Results: Median age was 66 years (IQR = 60-71), PSA 6.7 ng/ml (IQR = 4.7-9.9) from in-house cohort. In the independent test set, algorithm correctly detected 111 of 198 lesions leading to 56.1% (49.3%-62.6%) sensitivity. PPV was 62.7% (95% CI 54.7%-70.7%) with FDR of 37.3% (95% CI 29.3%-45.3%). Of 79 true positive lesions, 82.3% were tumor positive at targeted biopsy, whereas of 57 false negative lesions, 50.9% were benign at targeted biopsy. Median DSC for lesion segmentation was 0.359. Overall PI-RADS classification accuracy was 30.8% (95% CI 24.6%-37.8%).

Conclusion: Our cascaded U-Net, residual network architecture can detect, classify cancer suspicious lesions at prostate MRI with good detection, reasonable classification performance metrics.

Keywords

Prostate cancer; biparametric; magnetic resonance imaging; detection; artificial intelligence

INTRODUCTION

Prostate cancer is the most common cancer type and second leading cause of cancer-related deaths among men in the United States (1). Thus, its detection is a major public health concern and a matter of debate for several decades. Lack of specificity in prostate specific antigen (PSA) serum testing has led to a significant increase in unnecessary prostate biopsies and resultant major public criticism of prostate cancer screening. Multiparametric magnetic resonance imaging (mpMRI) of the prostate enables the visualization of prostate cancer lesions and targeted biopsies (2,3). This approach improves the detection of clinically significant prostate cancer and reduces the upgrading rate of radical prostatectomy specimens with better risk assessment prior to therapy (4,5). However, unlike serum biomarkers which can be measured objectively with full automation, medical imaging needs to be analyzed by human readers with varying expertise resulting in significant inter-reader variability. This is especially the case in mpMRI of the prostate due to the complexity of its technique. To overcome these challenges, expert-consensus-based Prostate Imaging-Reporting and Data System (PI-RADS) guidelines which aim to standardize patient preparation, scanning technique and interpretation have been released (6,7). Lesions detected on mpMRI are assigned an assessment category on a 5-tier scale based on their likelihood of being clinically significant prostate cancer. Nevertheless, even with more standardized approaches, inter-reader variability in the detection and classification of prostate lesions continued to be significant (8).

Recent advancements in deep learning (DL) have led to impressive results in many different research domains. This was enabled by major developments in DL algorithm theory, as well as increased availability of large datasets and better hardware for training efficiency. DL can achieve or even surpass human performance level for many different tasks (9). However, DL applications in medicine struggle to follow this trend due to inherent limitations in well-annotated and diverse data availability. Nevertheless, several studies have been lately published demonstrating the efficiency of DL in medical imaging even in small single-center patient cohorts (9,10). In 2019, a novel DL-based model was proposed for the classification of prostate lesions on mpMRI (11). However, that work included a relatively small sample size from one institution which can be a significant limitation for DL-based AI approaches. In this study we aimed to develop a fully automated AI algorithm for assisting radiologists during prostate MRI read out during prostate lesion detection, segmentation and PI-RADS classification steps based on cascaded DL Networks, which were trained using a large multi-institutional dataset of mpMRI scans with expert-derived annotations for prostate segmentation, prostate lesion segmentation and PI-RADS category labelling; and to test its performance metrics in an independent test set.

MATERIALS AND METHODS

Patient Population

The dataset of this retrospective study comprises patient cohorts from two different institutions. Patient selection for the final dataset is demonstrated in Figure 1.

PROSTATEx Dataset—This dataset is publicly available and was used for an international contest between November 2016 and January 2017 and consists of 347 exams of 344 patients (three patients with repeated scans) (12,13). All scans were performed at the Radboud University Medical Center in Nijmegen, The Netherlands where patients underwent T2-weighted, proton density-weighted, dynamic contrast enhanced (DCE) and diffusion weighted imaging using the 3 Tesla MAGNETOM Trio and Skyra scanner systems (Siemens Healthineers, Erlangen, Germany) without the use of an endorectal coil (Supplement Table 1). T2-weighted images were obtained using a turbo spin echo sequence and had a resolution of 0.5 mm in plane and a slice thickness of 3.6 mm. The DWI series were obtained with a single-shot echo planar imaging sequence with a resolution of 2 mm in-plane and 3.6 mm slice thickness and with diffusion-encoding gradients in three directions. Three b-values were acquired (50, 400, and 800 s/mm^2), and ADC maps and b = 1400 DWI were calculated by the scanner software. For use in this study, all scans from training and testing cohorts were independently evaluated for lesion segmentation and PI-RADS classification. Biopsy outcomes were not provided to the radiologist at the time of interpretation.

NCI MRI Dataset—This dataset comprises of 1043 scans from patients who underwent mpMRI at our institution for clinical assessment of suspicion for or previously confirmed localized prostate cancer. Patients were scanned between January 2015 and September 2019. This includes patients accrued as part of one or more IRB approved protocols, including radiologic profiling of prostate cancer (Clinical Trials.gov Identifier: NCT03354416), patients undergoing MRI-fusion-guided prostate biopsy (NCT00102544), and/or patients either diagnosed at our institution or referred to us for surgical treatment of intermediate or high risk prostate cancer (NCT02594202) and signed an informed consent form. For patients undergoing radical prostatectomy, these patients' outcomes and correlation to preoperative mpMRI and MRI-targeted biopsies were recently published in a major publication (4). For patients with positive findings on MRI (i.e. detection of PIRADSv2 lesions score 2-5), these data were recently reported in a study evaluating automated classification of PIRADS categorization (14). Importantly, this dataset additionally included patients who did not demonstrate any lesions to prevent selection bias towards prostate cancer patients and more aggressive disease states. mpMRI scans at NCI, NIH were performed using a 3 Tesla magnet (Achieva 3.0-T-TX, Philips Healthcare, Best, the Netherlands) with a 16-channel surface coil (SENSE, Philips Healthcare) and an endorectal coil (BPX-30, Medrad, Pittsburgh, Pennsylvania) inflated with perfluorocarbon (3mol/l Fluorinert) to a volume of 45 ml in 472 patients. T2-weighted turbo-spin-echo acquisition sequences were obtained in the axial, sagittal, and coronal planes. Diffusion-weighted imaging was obtained for production of apparent diffusion coefficient (ADC) maps using a monoexponential decay model and a separate high b-value sequence (2000 s/mm^2 for scans with endorectal coil, 1500 s/mm^2

for scans without endorectal coil). In 571 patients where an endorectal coil was not used, a 32-channel cardiac coil (SENSE, InVivo, Gainesville, Florida) was used instead. MRI acquisition parameters are summarized in Supplemental Table 1.

MRI Evaluation and Lesion Annotation

All MRI scans were evaluated by a single genitourinary radiologist (BT) with over 14 years of experience and reading over 1500 mpMRI exams per year. Suspicious areas were manually outlined using an in-house segmentation software on T2W series, with visual correlation to diffusion-weighted series for lesion volume determination (15). The resulting contours were saved in MIPAV VOI (volume of interest) format and assigned a corresponding PI-RADSv2 category. For all subsequent image processing, analysis, and training, all DICOM images were converted to NIfTI format, maintaining original spatial resolution. Diffusion-weighted series (ADC and high b-value) were then resampled to spatial resolution of T2W imaging. Contours were converted to NIfTI format as binary masks in equivalent spatial resolution as T2W imaging.

AI Algorithm Development

In all of our experiments, only T2-weighted images, high b-value DWI and ADC maps were used for training, validation and testing of the deep neural networks. Briefly, the final algorithm consists of three separate models (Supplemental Figure 1): lesion detection/segmentation, Benign Prostatic Hyperplasia (BPH) nodules filter/removal, and PIRADSv2 classification. The lesion segmentation algorithm was trained for binary segmentation of any lesion PIRADSv2 scores 2-5 from radiologist interpretation. The architecture is a 3D neural network (Supplemental Figure 2) and follows a similar architecture to 3D U-Net (convolutional encoder-decoder), which guarantees the input and output to share the same 3D shape (16). The subsequent classification algorithms (BPH filter and PIRADSv2 classification) are both based on a 3D deep residual network. Full details of algorithm design and image processing are presented in Supplementary Material.

We pseudo-randomly collected 150 patient cases from the data pool as the hold-out test set. In order to prepare the framework for the clinical study, the patient distribution of the test set matches the real population among a large group of patients in terms of PI-RADS category. The remaining 1390 patient cases were used for 5 fold cross validation.

Comparison to Pathology Outcome

For the test set, an additional analysis was conducted to compare AI algorithm's lesion detection and PI-RADS classification vs. pathology. The ground truth for this analysis was based on TRUS/MRI fusion guided biopsies which was performed by a urologist (PAP) and interventional radiologist (BJW) (both with >15 years of experience in performing TRUS/MRI fusion guided biopsies using a commercial platform [UroNav, Philips, Gainesville, FL, USA]).

Statistical Analysis

The detection performance was measured at the patient level and lesion level. We computed the instance numbers of true positive (TP), false positive (FP), and false negative (FN)

at the lesion-level compared to radiologist ground truth to evaluate sensitivity, positive predictive value (PPV), and false discovery rate (FDR). Additionally, lesion-based detection performance was characterized using the Free-Response Operating Characteristic (FROC) curve. Here we considered a prediction as true positive when it has any overlapping area with ground truth labels. Given the use of radiologist-based ground truth, true negatives cannot be evaluated at the lesion-level. The segmentation accuracy was measured with the Dice similarity coefficient score. At the patient-level, a true negative is defined as a PIRADS 1 scan correctly predicted as containing no lesions. Therefore, specificity was additionally calculated at the patient-level.

For the classification algorithm, accuracy in PIRADSv2 scoring was calculated compared to radiologist ground truth. For all regions of interest (ROI) detected, the classification model was used to predict the PI-RADS category of each ROI and compared to radiologist ground truth.

Results were reported separately for validation and testing datasets. Performance metrics in the validation sets were reported as median values across 5 fold cross-validation sets. Performance metrics in the hold-out test set were calculated from the entire set, where 95% confidence intervals (CI) were estimated from 2000 bootstrap samples by random sampling on the patient-level to account for intra and inter-lesion correlations.

RESULTS

Study Population and Data Split

Overall, 1390 patient scans were included in this study. The dataset was randomly subdivided into a training and validation set (n = 1240, 89%) and an independent test set (n = 150, 11%). Patient demographics after data split for all datasets created at our institution are summarized in Table 1.

Validation Set Prostate Lesion Detection Performance

Detection performance metrics for the baseline algorithm on the lesion level are summarized in Table 2 for all five folds. The mean lesion level sensitivity was 65.5% (range 61.5%-70%), mean positive predictive value was 46.7% (range 42.4%-50.9%) and mean false discovery rate was 53.3% (range 49.1%-57.6%). Patient level statistics are presented in Supplemental Table 2, where mean sensitivity was 97.8% (range 95.3%-99.3%) and specificity was 2.5% (range 0%-12.5%).

Detection performance metrics for the BPH filter enhanced algorithm on the lesion level are summarized in Table 2 for all five folds. FROC curves for all five folds are shown in Figure 2. The mean lesion level sensitivity was 61% (range 57.7%-65.9%), mean positive predictive value was 54.7% (range 48.9%-58.6%) and mean false discovery rate was 59.4% (range 51.1%-65.9%). The mean patient level sensitivity was 96.1% (range 97.1%-98.1%) and mean specificity was 10% (range 0%-25%) (Supplemental Table 2).

Test Set Lesion Detection Performance

Detection performance metrics for the baseline and BPH filter enhanced algorithm in the test set are summarized in Table 3. FROC curve for the BPH filter enhanced test set detection is shown in Figure 3.

The baseline algorithm demonstrated a sensitivity of 61.1% (95% CI 54.6%-67.5%), a positive predictive value of 44.8% (95% CI 38.4%-51.4%) and a false discovery rate of 55.2% (95% CI 48.6%-61.6%). This corresponds to a mean of 0.99 false positive lesions per patient.

The BPH filter enhanced algorithm demonstrated a sensitivity of 56.1% (95% CI 49.3%-62.6%), a positive predictive value of 62.7% (95% CI 54.7%-70.7%) and a false discovery rate of 37.3% (95% CI 29.3%-45.3%). This corresponds to a mean of 0.44 false positive lesions per patient.

Sample visual results for algorithm prediction and visualization are shown in Figures 4-6.

Lesion-Based Segmentation Congruence

Mean Dice similarity coefficient scores of lesion detection algorithm for the validation set after 5 fold cross-validation had a mean Dice similarity coefficient score of 0.385 (range 0.347-0.141) in the validation set and 0.359 in the test set (Supplemental Table 3).

PI-RADS Category Classification Performance

Validation Set—In the validation set, overall classification accuracy for correct PI-RADS category classification was 60.3% (135/224). The confusion matrix for ground truth lesion PI-RADS category vs. algorithm PI-RADS category prediction is shown in Supplemental Table 4. Correct classification among PI-RADS categories was distributed as follows: PI-RADS category 2 0% (0/21), PI-RADS category 3 66.7% (46/69), PI-RADS category 4 76.1% (67/88), PI-RADS category 5 47.8% (22/46).

Test Set—The confusion matrix for PI-RADS category predictions in the test set is shown in Supplemental Table 5. Overall classification accuracy for correct PI-RADS category classification was 30.8% (61/198, 95% CI 24.6%-37.8%). Correct classification among PI-RADS categories was distributed as follows: PI-RADS category 2 0% (0/23), PI-RADS category 3 66.7% (26/39), PI-RADS category 4 14.6% (14/96), PI-RADS category 5 52.5% (21/40).

Cascaded Algorithm Performance

Combined detection and classification performance for the cascaded algorithm (detection, BPH filtering, classification) are summarized in Table 4. The majority of false positive lesions were classified as PI-RADS categories 3 and 4. The majority of false negative lesions were PI-RADS categories 2,3 and 4. Among the correctly detected lesions, correct classifications were 0% (0/7) for PI-RADS category 2, 58.3% (7/12) for PI-RADS category 3, 66.1% (39/59) for PI-RADS category 4, 58.3% (21/36) for PI-RADS category 5. Overall classification accuracy was 57.7%.

Correlation with Pathology Results

Of the 198 lesions in the test set, 136 had targeted biopsy outcomes (n = 43 benign, n = 19 Gleason Grade [GG] 1, n = 42 GG2, n = 18 GG3, n = 7 GG4, n = 1 GG5). Of 79 true positive lesions, 82.3 were tumor positive at targeted biopsy, whereas of 57 false negative lesions, 50.9% were benign at targeted biopsy. Cancer detection rate could not be assessed since this was a retrospective study, false positive lesions' pathology validation could not be done with targeted biopsy.

DISCUSSION

In this work, we sought to develop a DL-based AI system for biparametric MRI which comprises of cascaded algorithms detection and classification algorithms, simulating the workflow of a human radiologist during the reading process. The algorithm is based on a 3D U-Net and trained to detect and segment cancer-suspicious lesions within the prostate based on expert reader annotations. Following detection, a four-class 3D Residual Network was trained to assign a PI-RADS category to detected regions. Initial experiments resulted in a high number of false positive detections corresponding to BPH nodules. Therefore, we trained a fourth algorithm for BPH classification to reduce false-positive lesion detections. This algorithm was inserted between the lesion detector and classifier. With this approach, the FDR decreased 17.9% for lesion detection task in the independent test set with only 5% penalty in detection sensitivity. Majority of false negative lesions were categories 2, 3 and 4 while only 10% (4/40) of PI-RADS 5 lesions were missed.

There was moderate classification accuracy at 30.8% in the test set with highest agreement in categories 3 and 5. Similar to the detection results, this is mostly attributed to the fact that PI-RADS 5 lesions are by definition large and well demarcated and therefore easier to detect and classify for computer algorithms. Overall, 4 class classifications tasks are inherently more challenging than binary tasks. Furthermore, the categorical PI-RADS system is based on mainly subjective image features which limit reproducibility. This is reflected in the modest inter-reader agreement which is well established in current literature (17,18). We therefore believe PI-RADS category prediction with all its challenges in the human world may not a suitable task for DL based algorithms. Retrospective analysis of biopsy outcomes in the test set, when available, demonstrated majority of cancer positive lesions in this dataset were detected by the algorithm. Future work will focus on modifying this algorithm for automatic detection of prostate cancer lesions based on gold standard histopathology. If accomplished, this system will be able to automatedly detect suspicious lesions and assign a risk category based on prediction probability. This will ensure similar benefits as the PI-RADS system with the advantage of being independent of subjective criteria and reader dependence.

Direct comparison of performance metrics between other published algorithms are difficult and often invalid due to differences in methodology, performance metrics and most importantly patient populations. Several prior works explored the use of DL based AI systems for prostate lesion detection and classification on mpMRI. Schelb *et al.* used a similar approach to develop a model for prostate lesion detection on mpMRI (11). Similar to this work a U-Net architecture was used but with a two-dimensional instead of three-

dimensional structure. Annotations were derived from 312 men for training and testing from a single institution. Detection accuracy was similar between the U-Net and radiologist lesion detections for PI-RADS thresholds ≥ 3 and ≥ 4. Congruency of lesions was good with a mean Dice similarity coefficient score 0.359 for detected prostate lesions. However, ground truth true positivity was defined by any overlap of prediction masks with a prostate sextant containing clinically significant prostate cancer confirmed on biopsy which makes direct comparison to our study difficult. Our model was trained to detect and classify PI-RADS prostate lesions rather than biopsy-confirmed prostate cancer. Arif *et al.* trained a model based on U-Net architecture with 292 patients with low-risk prostate cancer eligible for active surveillance (19). The model was trained to detect prostate cancer lesions of ISUP grade ≥ 2 confirmed by targeted biopsy on multi-parametric prostate MRI including T2-weighted imaging, DWI and ADC maps. True positive lesions were defined if prediction masks and ground truth masks were overlapping > 0.01 cc. Sensitivity ranged from 82% to 92%, Specificity from 43% to 76% and AUC 0.65 to 0.89 depending on lesion sizes. The model showed the best performance on lesions > 0.5 cc. Prostate lesion congruency data was not reported in this publication. In another study, Sanford *et al.* developed a classification model for PI-RADS lesions based on a two-dimensional ResNet-50 architecture (14). They achieved 58% classification accuracy with 80% agreement on PI-RADS category 5 lesions. However, lesion had to be detected first by a radiologist before classification could be performed by the algorithm instead of an end-to-end solution for both intraprostatic lesion detection and classification. The main strengths of our model are its diverse multi-institutional patient population, use of biparametric MRI, a three-dimensional neural network architecture and its large data size.

This study has three limitations. First, we did not train our model based on gold standard labels, which would be histopathologically proven prostate cancers. The focus of this project was to develop an assistance tool for radiologists during prostate MRI read-out via replicating an expert radiologist's diagnostic prostate MRI reads but not an AI model to predict histopathology outcomes. Second, annotations were made by a single highly experienced genitourinary radiologist. Ideally, annotations should be made by multiple readers based on consensus. However, our institutional infrastructure in this field does not permit this and although this may seem like a limitation, a prostate MRI read out assisting AI model trained on a highly focused can be advantageous for consistent radiology evaluations. Third, DCE imaging was not incorporated into the model due to inconsistent image acquisition in our study population. While DCE is not a primary sequence used in PIRADSv2 assessment, it is used for further discrimination between PIRADSv2 category 3 and category 4 lesions in the peripheral zone. Therefore, it is possible underperformance in the classification task of the model was due to lack of enhancement-related information. Finally, true value and applicability of an algorithm needs to be tested in a prospective and comparative study against human performance to determine its practical value when implemented in a clinical scenario.

In conclusion, we developed an end-to-end DL based AI approach which includes a three-dimensional prediction model with promising detection and moderate PI-RADS classification performance for prostate lesions on biparametric MRI. This system can be used as an adjunct tool for radiologists during the reading process. Future work will include

prostate cancer tissue prediction, testing on external data and prospective testing with human interaction.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

FUNDING

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin 2020; 70(1):7–30. [PubMed: 31912902]
2. Ahmed HU, El-Shater Bosaily A, Brown LC, Gabe R, Kaplan R, Parmar MK, et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. Lancet (London, England) 2017; 389(10071):815–822. [PubMed: 28110982]
3. Kasivisvanathan V, Rannikko AS, Borghi M, et al. MRI-targeted or standard biopsy for prostate-cancer diagnosis. The New England journal of medicine 2018; 378(19):1767–1777. [PubMed: 29552975]
4. Ahdoot M, Wilbur AR, Reese SE, et al. MRI-Targeted, systematic, and combined biopsy for prostate cancer diagnosis. The New England journal of medicine 2020; 382(10):917–928. [PubMed: 32130814]
5. Siddiqui MM, Rais-Bahrami S, Turkbey B, et al. Comparison of MR/ultrasound fusion-guided biopsy with ultrasound-guided biopsy for the diagnosis of prostate cancer. JAMA 2015; 313:390–397. [PubMed: 25626035]
6. Weinreb JC, Barentsz JO, Choyke PL, et al. PI-RADS Prostate Imaging - Reporting and Data System: 2015, Version 2. Eur Urol 2016; 69:16–40. [PubMed: 26427566]
7. Turkbey B, Rosenkrantz AB, Haider MA, et al. Prostate imaging reporting and data system version 2.1: 2019 update of prostate imaging reporting and data system version 2. European Urology 2019; 0232:1–12.
8. Park KJ, Choi SH, Lee JS, et al. Interreader Agreement with Prostate Imaging Reporting and Data System Version 2 for Prostate Cancer Detection: A Systematic Review and Meta-Analysis. The Journal of urology 2020; 204(4):661–670. [PubMed: 32552474]
9. McKinney SM, Sieniek M, Godbole V, et al. International evaluation of an AI system for breast cancer screening. Nature 2020; 577(7788):89–94. [PubMed: 31894144]
10. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016; 316(22):2402–2410. [PubMed: 27898976]
11. Schelb P, Kohl S, Radtke JP, et al. Classification of Cancer at Prostate MRI: Deep Learning versus Clinical PI-RADS Assessment. Radiology 2019; 293(3):607–617. [PubMed: 31592731]
12. Litjens G, Debats O, Barentsz J, et al. Computer-Aided Detection of Prostate Cancer in MRI. IEEE Transactions on Medical Imaging 2014; 33 (5):1083–1092. [PubMed: 24770913]
13. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. Journal of Digital Imaging 2013; 26(6):1045–1057. [PubMed: 23884657]
14. Sanford T, Harmon SA, Turkbey EB, et al. Deep-Learning-Based Artificial Intelligence for PI-RADS Classification to Assist Multiparametric Prostate MRI Interpretation: A Development

Study. Journal of magnetic resonance imaging: JMRI 2020; 52(5):1499–1507. [PubMed: 32478955]

15. Turkbey B, Fotin SV, Huang RJ, et al. Fully automated prostate segmentation on MRI: comparison with manual segmentation methods and specimen volumes. AJR Am J Roentgenol 2013; 201:W720–W729. [PubMed: 24147502]
16. Çiçek Ö, Abdulkadir A, Lienkamp SS, et al. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation2016; Cham: Springer International Publishing.
17. Greer MD, Shih JH, Barrett T, et al. All over the map: An interobserver agreement study of tumor location based on the PI-RADSv2 sector map. Journal of Magnetic Resonance Imaging 2018; 48:482–490. [PubMed: 29341356]
18. Rosenkrantz AB, Ginocchio LA, Cornfeld D, et al. Interobserver Reproducibility of the PI-RADS Version 2 Lexicon: A Multicenter Study of Six Experienced Prostate Radiologists. Radiology 2016; 280:793–804. [PubMed: 27035179]
19. Arif M, Schoots IG, Castillo Tovar J, et al. Clinically significant prostate cancer detection and segmentation in low-risk patients using a convolutional neural network on multi-parametric MRI. European radiology 2020; 30(12):6582–6592. [PubMed: 32594208]

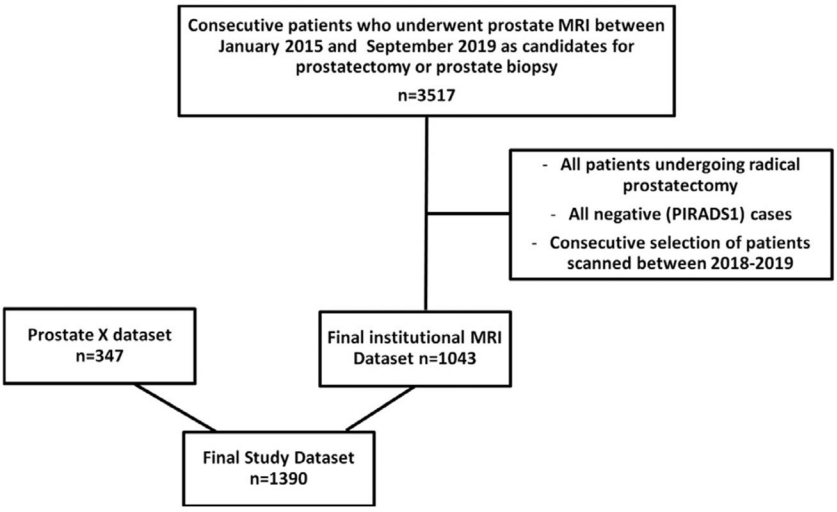

Fig 1.

Study flowchart for dataset creation for training, validation and testing of our algorithms. MRI, Magnetic Resonance Imaging.

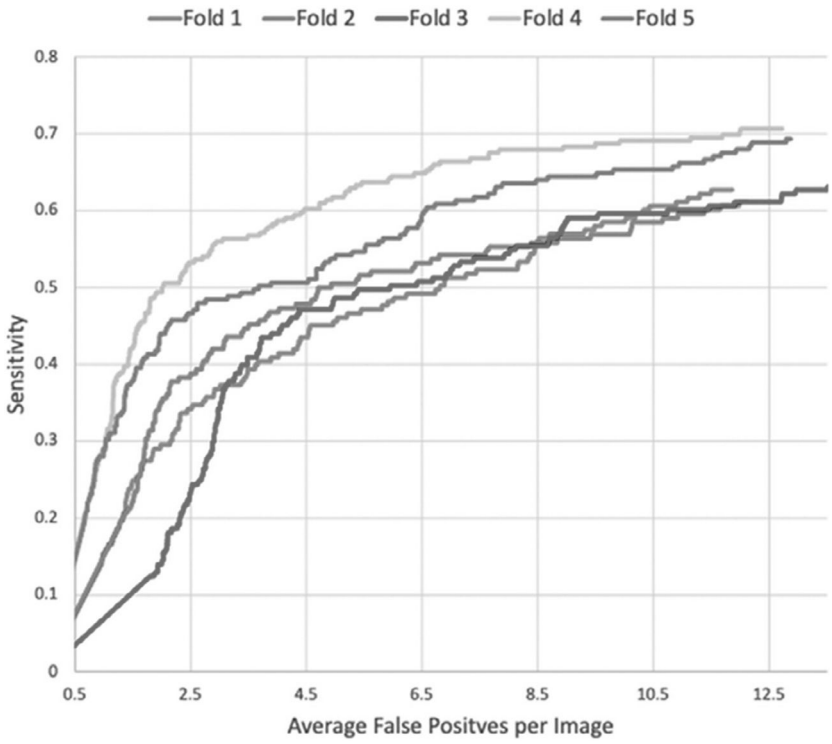

Fig 2.

FROC curves for detection performance with BPH filter in all five folds in the validation set.

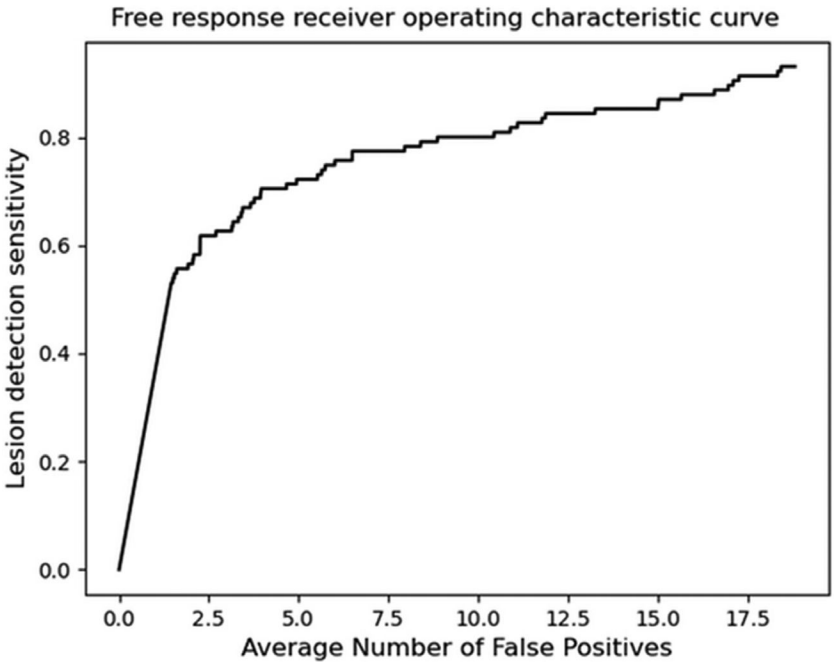

Fig 3.
FROC curves for detection performance with BPH filter in the test set.

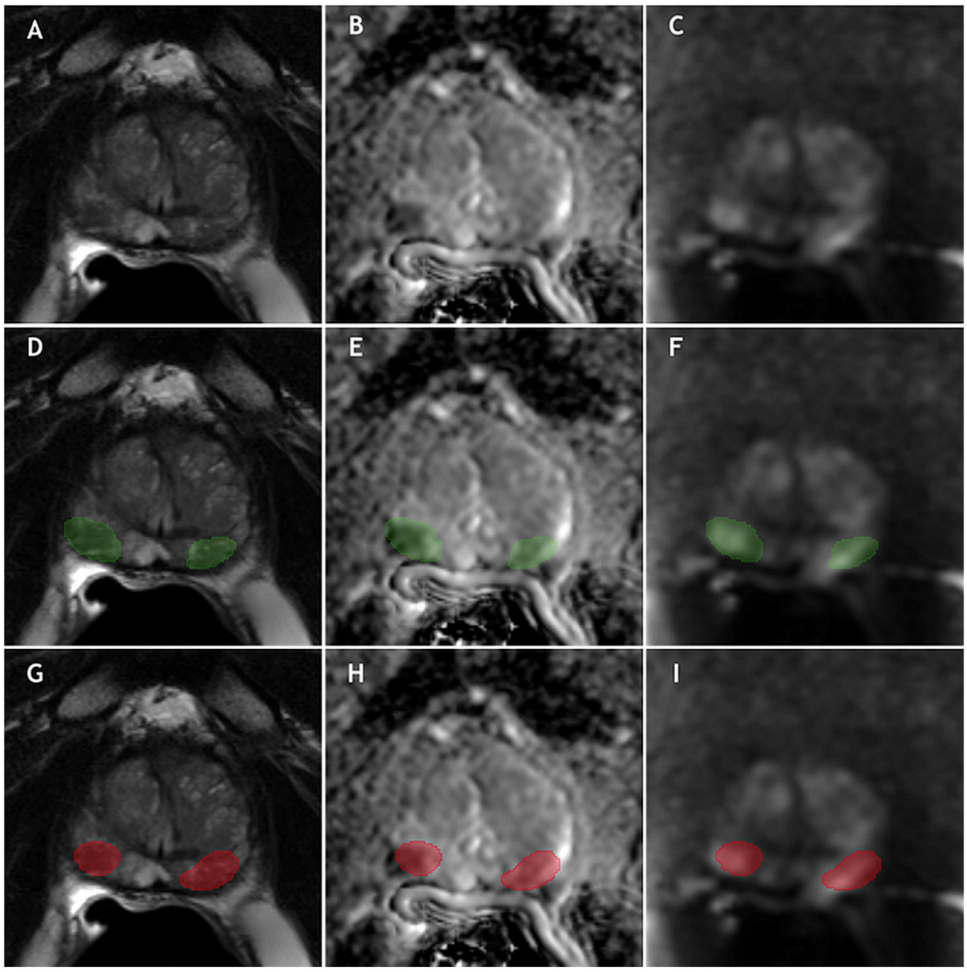

Fig 4.

Ground truth masks and segmentation predictions in a test set patient. Green contours represent ground truth lesions and red contours are model predictions. Patient is a 66 year-old man with a PSA of 7.88 ng/ml and prostate volume of 42 ml and first-time multiparametric prostate MRI. The patient was scanned at 3T using an endorectal coil. Two distinct lesions were detected by the radiologist representing the ground truth. Lesion 1 was 0.9 cm in the right apical-mid peripheral zone and assigned PI-RADS category 4. Lesion 2 was 0.9 cm and in the left apical-mid peripheral zone and assigned PI-RADS category 4. Both lesions were correctly detected by the cascaded algorithm representing true positive lesions. (a) Original T2 weighted imaging, (b) Original Apparent Diffusion Map, (c) Original high b-value, (d) T2 weighted imaging with ground truth contours overlaid, (e) Apparent Diffusion Map with ground truth contours overlaid, (f) High b-value with ground truth contours overlaid, (g) T2 weighted imaging with prediction contours overlaid, (h) Apparent Diffusion Map with prediction contours overlaid, (i) High b-value with prediction contours overlaid.

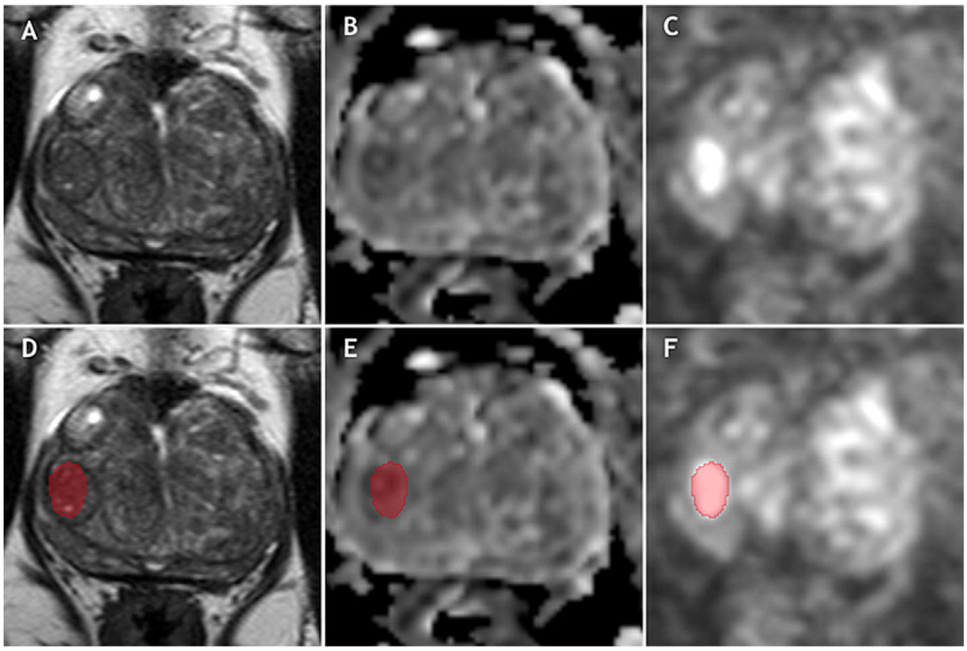

Fig 5.

Segmentation prediction in a test set patient without associated ground truth mask. Red contours are model predictions. 71 year-old man with PSA of 6.56 ng/ml and prostate volume 51 ml. The patient was scanned at 3T using without an endorectal coil. No distinct lesions were detected representing a negative MRI (PI-RADS category 1). One false positive lesion was called by the cascaded algorithm in the right mid transition zone. The false positive lesion was reevaluated and defined as a BPH nodule. (a) Original T2 weighted imaging, (b) Original Apparent Diffusion Map, (c) Original high b-value, (d) T2 weighted imaging with prediction contours overlaid, (e) Apparent Diffusion Map with prediction contours overlaid, (f) High b-value with prediction contours overlaid.

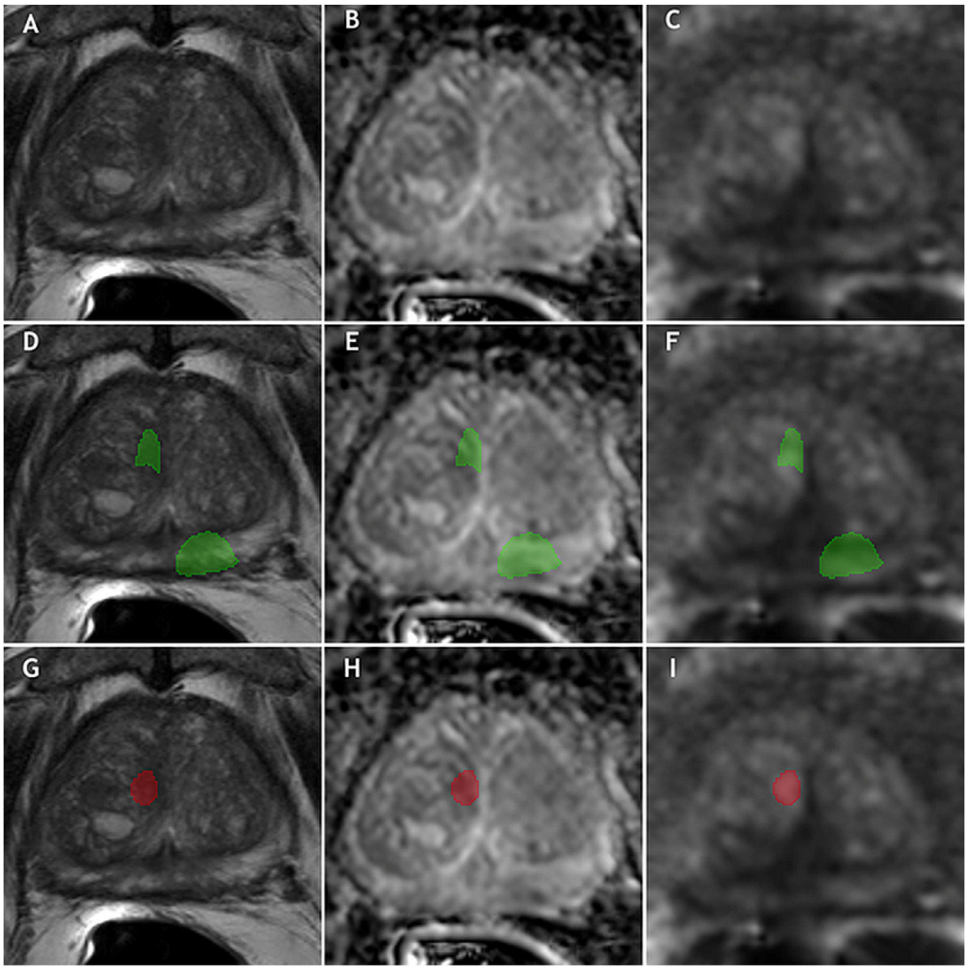

Fig 6.

Ground truth masks and segmentation predictions in a test set patient. Green contours represent ground truth lesions and red contours are model predictions. 70 year-old man with PSA of 8.71 ng/ml and prostate volume of 63.6 ml and first-time multiparametric prostate MRI. The patient was scanned at 3T using an endorectal coil. Two distinct lesions were detected by the radiologist representing the ground truth. Lesion 1 was 1.8 cm in the midline to right apical anterior transition zone and assigned PI-RADS category 5. The second lesion was 1.6 cm in the left mid peripheral zone and assigned PI-RADS category 4. Lesion 1 was correctly detected while lesion 2 was missed by the cascaded algorithm representing a false negative. (a) Original T2 weighted imaging, (b) Original Apparent Diffusion Map, (c) Original high b-value, (d) T2 weighted imaging with ground truth contours overlaid, (e) Apparent Diffusion Map with ground truth contours overlaid, (f) High b-value with ground truth contours overlaid, (g) T2 weighted imaging with prediction contours overlaid, (h) Apparent Diffusion Map with prediction contours overlaid, (i) High b-value with prediction contours overlaid.

TABLE 1.

Patient and lesion characteristics of the entire cohort after data split (numbers for age, serum PSA, prostate volume represent median value and numbers in parenthesis represent interquartile range).

Patient Characteristics*			
Variables	**Overall, *n* = 1043**	**Training, *n* = 943**	**Test, *n* = 100**
Age (years)	66 (60-71)	66 (60-71)	66 (59-71)
Serum PSA (ng/ml)	6.7 (4.7-9.9)	6.7 (4.6-9.8)	7.3 (5.3-11.9)
Prostate volume (ml)	59 (42-87)	60 (44-88)	46 (32-72)
Biopsy	409 (39%)	343 (36%)	66 (66%)
Gleason Grade			
0	41 (10%)	34 (9.9%)	7 (11%)
1	42 (10%)	37 (11%)	5 (7.6%)
2	201 (49%)	164 (48%)	37 (56%)
3	83 (20%)	73 (21%)	10 (15%)
4	40 (9.8%)	33 (9.6%)	7 (11%)
5	2 (0.5%)	2 (0.6%)	0 (0%)
Lesion Characteristics			
Variable **PI-RADS Category**	**Overall, *n* = 1960**	**Training, *n* = 1734**	**Test, *n* = 226**
1	532 (27%)	504 (29%)	28 (12%)
2	163 (8.3%)	140 (8.1%)	23 (10%)
3	402 (21%)	363 (21%)	39 (17%)
4	532 (27%)	436 (25%)	96 (42%)
5	331 (17%)	291 (17%)	40 (18%)

*Patient-level characteristics included for in-house dataset only and does not represent *n* = 347 from publicly available Prostate-X dataset.

TABLE 2.

Lesion level detection performance for all five folds in the validation set without BPH filtering and with BPH filtering (TP, true positive, FP, false positive, FN, false negative, PPV, positive predictive value, FDR, false discovery rate).

	Without BPH Filtering (Baseline)						With BPH Filtering				
	TP	**FP**	**FN**	**Sensitivity (%)**	**PPV (%)**	**FDR (%)**	**TP**	**FP**	**Sensitivity (%)**	**PPV (%)**	**FDR (%)**
Fold 0	139	189	87	61.5	42.4	57.6	133	139	58.8	48.9	51.1
Fold 1	145	144	89	62	50.2	49.8	139	98	59.4	58.6	59.4
Fold 2	177	226	79	69.1	43.9	56.1	161	149	62.9	51.9	62.9
Fold 3	189	182	81	70	50.9	49.1	178	132	65.9	57.4	65.9
Fold 4	152	179	82	65	45.9	54.1	135	104	57.7	56.5	57.7

TABLE 3.

Lesion level detection performance in the test set without BPH filtering and with BPH filtering (TP, true positive, FP, false positive, FN, false negative, PPV, positive predictive value, FDR, false discovery rate, F1 score, harmonic mean of positive predictive value and sensitivity).

	TP	FN	FP	Sensitivity (%)	PPV (%)	FDR (%)	F1 score (%)
Without BPH Filtering (Baseline)	121	77	149	61.1 (54.6-67.5)	44.8 (38.4-51.4)	55.2 (48.6-61.6)	51.7 (46-57.2)
With BPH Filtering	111	87	66	49.3-62.6)	62.7 (54.7-70.7)	37.3 (29.3-45.3)	59.2 (53-65)

TABLE 4.

Confusion matrix for PI-RADS classification of the cascaded algorithm.

		Prediction			
		False Negative	**PI-RADS 3**	**PI-RADS 4**	**PI-RADS 5**
Ground Truth	False Positive	–	31	32	3
	PI-RADS 2	16	3	3	1
	PI-RADS 3	27	7	4	1
	PI-RADS 4	40	15	39	5
	PI-RADS 5	4	6	9	21