Fully Automated Deep Learning Model to Detect Clinically Significant Prostate Cancer at MRI

Jason C. Cai, MD • Hirotsugu Nakai, MD, PhD • Shiba Kuanar, PhD • Adam T. Froemming, MD • Candice W. Bolan, MD • Akira Kawashima, MD, PhD • Hiroaki Takahashi, MD, PhD • Lance A. Mynderse, MD • Chandler D. Dora, MD • Mitchell R. Humphreys, MD • Panagiotis Korfiatis, PhD • Pouria Rouzrokh, MD, MPH, MHPE • Alexander K. Bratt, MD • Gian Marco Conte, MD, PhD • Bradley J. Erickson, MD, PhD • Naoki Takahashi, MD

From the Departments of Radiology (J.C.C., H.N., S.K., A.T.F., H.T., P.K., P.R., A.K.B., G.M.C., B.J.E., N.T.) and Urology (L.A.M.), Mayo Clinic, 200 First St SW, Rochester, MN 55905; Department of Radiology, Massachusetts General Hospital, Boston, Mass (J.C.C.); Departments of Radiology (C.W.B.) and Urology (C.D.D.), Mayo Clinic, Jacksonville, Fla; and Departments of Radiology (A.K.) and Urology (M.R.H.), Mayo Clinic, Scottsdale, Ariz. Received October 4, 2023; revision requested December 13; final revision received April 10, 2024; accepted April 25. **Address correspondence to** N.T. (email: *takahashi.naoki@mayo.edu*).

Supported by a grant from the Mayo Foundation for Education and Research.

Conflicts of interest are listed at the end of this article.

See also the editorial by Johnson and Chandarana in this issue.

Radiology 2024; 312(2):e232635 • https://doi.org/10.1148/radiol.232635 • Content codes: GU MR AI

Background: Multiparametric MRI can help identify clinically significant prostate cancer (csPCa) (Gleason score ≥7) but is limited by reader experience and interobserver variability. In contrast, deep learning (DL) produces deterministic outputs.

Purpose: To develop a DL model to predict the presence of csPCa by using patient-level labels without information about tumor location and to compare its performance with that of radiologists.

Materials and Methods: Data from patients without known csPCa who underwent MRI from January 2017 to December 2019 at one of multiple sites of a single academic institution were retrospectively reviewed. A convolutional neural network was trained to predict csPCa from T2-weighted images, diffusion-weighted images, apparent diffusion coefficient maps, and T1-weighted contrast-enhanced images. The reference standard was pathologic diagnosis. Radiologist performance was evaluated as follows: Radiology reports were used for the internal test set, and four radiologists' PI-RADS ratings were used for the external (ProstateX) test set. The performance was compared using areas under the receiver operating characteristic curves (AUCs) and the DeLong test. Gradient-weighted class activation maps (Grad-CAMs) were used to show tumor localization.

Results: Among 5735 examinations in 5215 patients (mean age, 66 years ± 8 [SD]; all male), 1514 examinations (1454 patients) showed csPCa. In the internal test set (400 examinations), the AUC was 0.89 and 0.89 for the DL classifier and radiologists, respectively ($P = .88$). In the external test set (204 examinations), the AUC was 0.86 and 0.84 for the DL classifier and radiologists, respectively ($P = .68$). DL classifier plus radiologists had an AUC of 0.89 ($P < .001$). Grad-CAMs demonstrated activation over the csPCa lesion in 35 of 38 and 56 of 58 true-positive examinations in internal and external test sets, respectively.

Conclusion: The performance of a DL model was not different from that of radiologists in the detection of csPCa at MRI, and Grad-CAMs localized the tumor.

Supplemental material is available for this article.

Prostate cancer is the second most common cancer in men worldwide (1). Multiparametric MRI of the prostate is a standard imaging technique to diagnose clinically significant prostate cancer (csPCa) and facilitate lesion targeting during biopsy (2). Prostate Imaging Reporting and Data System (PI-RADS) version 2.1 is a standardized interpretation and reporting approach (3). However, lesion classification using PI-RADS is prone to intra- and interobserver variability (4), and a high level of expertise is needed.

The most common machine learning approach to detect csPCa at MRI is training a model on regions of interest drawn on MRI scans with reference to the corresponding biopsy findings. This can be achieved by using either classic machine learning (5,6) or deep learning (DL) (7,8). The lesion-level area under the receiver operating characteristic curve (AUC) for detecting csPCa has ranged from 0.81 to 0.93 (5–8). In general, the prediction can be made to the area of abnormality detected by the radiologist. The probability of csPCa at each individual voxel within the entire prostate can be calculated by applying the model to small image patches centered at each voxel to generate color overlay maps (8). The patient-level AUC for detecting csPCa has ranged from 0.78 to 0.87 using an internal test set (6,8).

Another approach is to train a segmentation model that gives predictions for each voxel. For the ground truth, a region of interest can be drawn on MRI scans with reference to the corresponding biopsy findings (9–13), a combination of the region of interest approach and the use of results from random biopsy (10,14,15) or whole-mount pathologic specimens (16,17). The patient-level AUC for

Abbreviations

AUC = area under the receiver operating characteristic curve, csPCa = clinically significant prostate cancer, DL = deep learning, FPR = false-positive rate, Grad-CAM = gradient-weighted class activation map, PI-RADS = Prostate Imaging Reporting and Data System, PSA = prostate-specific antigen, TPR = true-positive rate

Summary

The performance of a deep learning model was not different from that of radiologists in the detection of clinically significant prostate cancer at MRI, and gradient-weighted class activation maps localized the tumor.

Key Results

- In a retrospective study of 5215 patients (5735 examinations) who underwent multiparametric MRI for prostate cancer evaluation, deep learning (DL) model performance in clinically significant prostate cancer (csPCa) detection was not different from that of experienced radiologists (area under the receiver operating characteristic curve [AUC], 0.86 vs 0.84; $P = .68$).
- The combined DL model plus radiologists performed better than radiologists alone on the external test set (AUC, 0.89 vs 0.84; $P < .001$).
- For positive examinations, gradient-weighted class activation maps consistently highlighted the csPCa lesion.

detecting csPCa has ranged from 0.81 to 0.91 using an internal test set (9–17). Hosseinzadeh et al (10) reported a patient-level AUC of 0.85 using an external test set.

A major drawback of these approaches is that the lesion needs to be annotated by a radiologist or pathologist, which is resource intensive and limits the size of the data set. This applies not only at the time of initial model development but also at the time of model re-evaluation and retraining after clinical implementation.

The aim of this study was to develop a DL model to predict the presence of csPCa using patient-level labels (the presence or absence of csPCa) without information about tumor location and compare its performance with that of radiologists. Because the output from the proposed model does not include the tumor location, leveraging interpretability approaches to identify the lesion was proposed. Specifically, the use of a gradient-weighted class activation map (Grad-CAM) (18) was implemented to provide visualization maps that assisted with tumor localization.

Materials and Methods

Internal Data Set

This Health Insurance Portability and Accountability Act–compliant retrospective study was approved by the Mayo Clinic Institutional Review Board (no. 18-011143) and conducted at multiple sites of a single academic institution. The study sample included consecutive patients without a history of csPCa (Gleason score 7 or higher) undergoing prostate MRI for suspicion of cancer between January 2017 and December 2019 (n = 7854). Of these, patients who had pathologic confirmation within 1 year after MRI were included. In addition, patients without a history of low-grade prostate cancer and who had normal MRI findings (PI-RADS 1 or 2) were included so that the study sample had a csPCa prevalence similar to that of a screening population (19,20). Patients who did not provide research consent and those whose examinations had poor image quality were excluded (Fig 1). The institutional review board approved a waiver of the requirement to obtain informed consent.

Prostate-specific antigen (PSA) values were extracted from the radiology reports and the electronic medical records.

MRI Scan Acquisition

All examinations met PI-RADS version 2.0 or 2.1 technical specifications. Details on MRI scan acquisition are described in Table S1.

Pathology

Pathologic diagnosis was obtained from the pathology reports. Transrectal or transperineal US-guided biopsy was performed using MRI/US fusion software (fusion software: UroNav version 2.0–4.0, InVivo; US: bk3000 or Flex Focus 400, BK Medical) for lesions identified at MRI. Systematic biopsies were also performed at the same time. In cases of normal MRI findings, if clinical suspicion of csPCa was high, systematic biopsies were performed. csPCa was defined as a Gleason score of 7 or higher. Patients who had PI-RADS 1 or 2 lesions at screening MRI without pathologic confirmation were considered negative cases.

Data Set Partition

The internal data set was randomly split into 5035 examinations for training, 300 for validation, and 400 for the test set at the patient level. Patients in the test set had only one examination. In addition, ProstateX (n = 204) was used as an external test set (21). The data set was collected and curated for research in computer-aided diagnosis at prostate MRI by Radboud University Medical Centre in the Netherlands and includes T2-weighted, apparent diffusion coefficient, diffusion-weighted, and dynamic contrast-enhanced images. Table 1 summarizes the patient characteristics. The ProstateX data set analyzed during the study is available at *https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=23691656.*

Radiologist Performance Evaluation

For the internal test set, a patient-level PI-RADS score obtained from the radiology report was used. For the external test set, four board-certified abdominal radiologists with at least 10 years of experience (A.T.F., C.W.B., A.K., and N.T.) blinded to the pathologic diagnosis reviewed 51 examinations each and individually assigned a patient-level PI-RADS score.

Preprocessing and Input Data

From each examination, the following axial sequences were retrieved and used as input data: T2-weighted images, apparent diffusion coefficient maps, high-b-value diffusion-weighted images, and T1-weighted dynamic contrast-enhanced images. Preprocessing of images is described in Appendix S1. Additional input data included PSA, whole-gland PSA

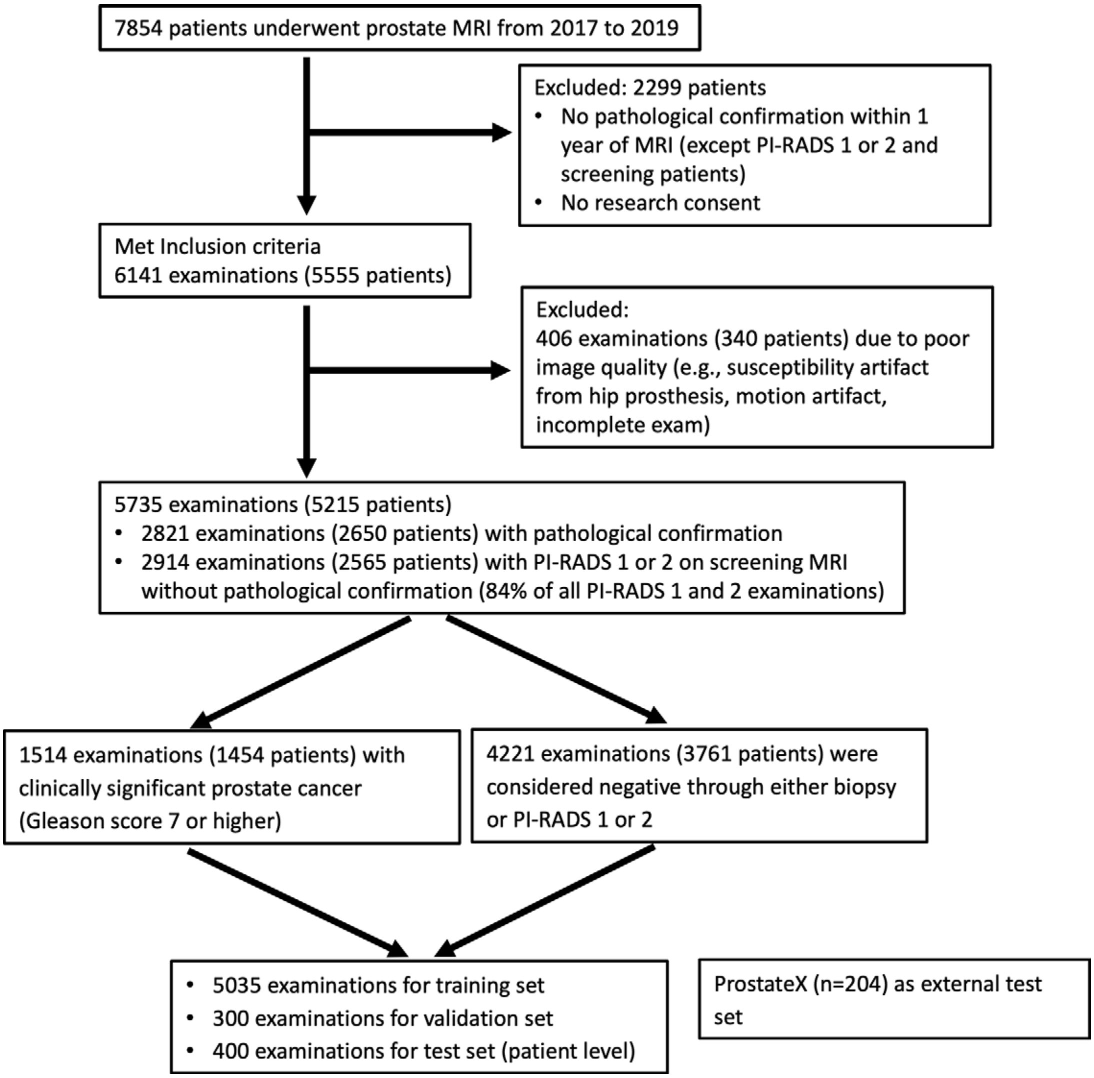
7854 patients underwent prostate MRI from 2017 to 2019
Excluded: 2299 patients
• No pathological confirmation within 1 year of MRI (except PI-RADS 1 or 2 and screening patients)
• No research consent
Met Inclusion criteria
6141 examinations (5555 patients)
Excluded:
406 examinations (340 patients) due to poor image quality (e.g., susceptibility artifact from hip prosthesis, motion artifact, incomplete exam)
5735 examinations (5215 patients)
• 2821 examinations (2650 patients) with pathological confirmation
• 2914 examinations (2565 patients) with PI-RADS 1 or 2 on screening MRI without pathological confirmation (84% of all PI-RADS 1 and 2 examinations)
1514 examinations (1454 patients) with clinically significant prostate cancer (Gleason score 7 or higher)
4221 examinations (3761 patients) were considered negative through either biopsy or PI-RADS 1 or 2
• 5035 examinations for training set
• 300 examinations for validation set
• 400 examinations for test set (patient level)
ProstateX (n=204) as external test set

Figure 1: Flowchart shows inclusion and exclusion criteria and patient characteristics. PI-RADS = Prostate Imaging Reporting and Data System.

Table 1: Patient Demographic Characteristics in the Data Sets

Characteristic	Training Set	Validation Set	Internal Test Set	External Test Set
No. of examinations	5035	300	400	204
Age (y)*	66 ± 8	66 ± 8	67 ± 8	64 ± 7
PSA availability	3872 (76.9)	228 (76.0)	303 (76)	NA
Pathology-proven cases	2455 (48.8)	138 (46.0)	228 (57)	NA
Benign or Gleason score of 6[†]	3729 (74.1)	224 (74.7)	268 (67)	134 (65.7)
Gleason score ≥7	1306 (25.9)	76 (25.3)	132 (33)	70 (34.3)
PI-RADS score				
1 or 2	3072 (61.0)	185 (61.7)	212 (53.0)	91 (44.6)[‡]
3	461 (9.2)	25 (8.3)	41 (10.3)	19 (9.3)[‡]
4	913 (18.1)	47 (15.7)	81 (20.3)	43 (21.1)[‡]
5	589 (11.7)	43 (14.3)	66 (16.5)	51 (25.0)[‡]

Note.—Unless otherwise specified, data are numbers of examinations, with percentages in parentheses. NA = not applicable, PI-RADS = Prostate Imaging Reporting and Data System, PSA = prostate-specific antigen.

* Data are means ± SDs.

[†] Benign includes PI-RADS 1 and 2 cases without pathologic proof in the screening population.

[‡] PI-RADS scores were assigned by radiologists from our institution.

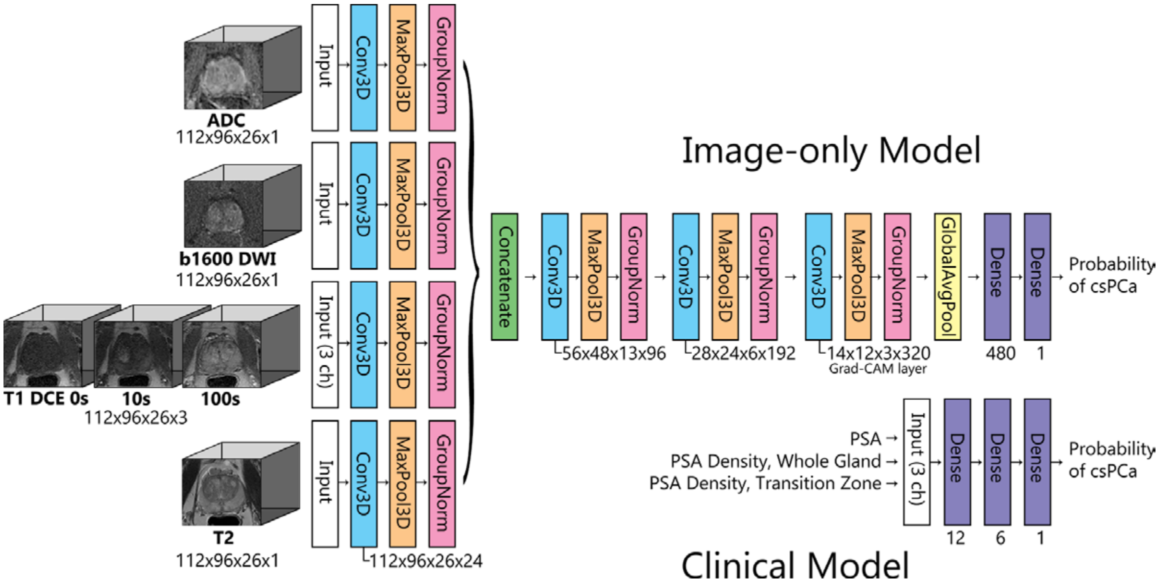

Figure 2: Diagram shows the architecture of the image-only model and clinical model. The image-only model consisted of one set of three-dimensional (3D) convolutions (3D convolutional kernel [Conv3D], maximum pooling [MaxPool3D], and group normalization [GroupNorm]) for each input volume (T2, diffusion-weighted imaging [DWI], apparent diffusion coefficient [ADC], and dynamic contrast-enhanced [DCE]), followed by concatenation, three additional sets of 3D convolutions, global average pooling (GlobalAvgPool), and two fully connected layers. The clinical model consists of a neural network of two fully connected layers with prostate-specific antigen (PSA) level, whole-gland PSA density, and transition zone PSA density as input. The output from each model was the probability of clinically significant prostate cancer (csPCa). ch = channel, Grad-CAM = gradient-weighted class activation map.

density (PSA divided by whole-gland volume), and transition zone PSA density (PSA divided by transition zone volume). Whole-gland and transition zone volumes were calculated from segmentations of the T2-weighted image generated by U-Net (22). These automated preprocessing steps were applied to every input volume, and the segmentations were not manually corrected.

Model Training

A convolutional neural network–based DL model to predict csPCa from MRI (image-only model) and a neural network–based model to predict csPCa from clinical data (clinical model) were developed. The details of the models are described in Appendix S1 and illustrated in Figure 2.

Subsequently, the predictions of the clinical and image-only models (probabilities) were combined using logistic regression (23) to predict presence of csPCa (hereafter, image+clinical model). In addition, the predictions (probabilities) from the image+clinical model and PI-RADS scores (integers 2–5) were combined using logistic regression (hereafter, image+clinical+radiologist model). The external data set (ProstateX) did not include PSA values; thus, the predictions of the image-only models and PI-RADS scores were combined (hereafter, image+radiologist model). The logistic regression models were trained on the validation set.

Statistical Analysis

Both internal and external test sets were used to evaluate the models (internal test set: image-only, image+clinical, image+clinical+radiologist, and clinical models; external test set: image-only and image+radiologist models). AUCs were used to quantify the performance of the models, and performance was compared with that of radiologists with use of the DeLong test.

The Youden *J* statistic was used to determine optimal threshold values on the validation set, and true-positive rate (TPR) and false-positive rate (FPR) were calculated on the test set. Additionally, FPR was calculated for the corresponding TPR equivalent to that of radiologists at the PI-RADS 3 threshold. TPR and FPR at optimal threshold and PI-RADS 3 equivalent threshold were compared with that of radiologists with use of the McNemar test.

Because the sensitivity of PI-RADS has been reported to be 87% (24), results were also reported before and after excluding patients who had a normal MRI interpretation (PI-RADS 1 or 2) without pathologic confirmation for the internal test set.

Calibration plots of the image-only model on internal and external test sets were constructed.

All statistical analyses were performed by authors (J.C.C., H.N., and N.T.) using Python 3.9–3.11. An alpha level of .05 was used for comparison.

Grad-CAM Evaluation

The output of the model did not include the location of the suspected tumor. Therefore, the voxels that contributed to a positive prediction were visualized using Grad-CAM from activations in the final layer (18). Of the 400 cases in the internal test set, 250 were used to calibrate how the Grad-CAM was displayed. The calibrated Grad-CAM was evaluated using the remaining 150 examinations from the internal test set and the ProstateX test set. A detailed description of the calibration process of Grad-CAM is found in Appendix S1. The Grad-CAMs were subsequently reviewed by an abdominal

Table 2: Model Performance: AUC, True-Positive Rate, and False-Positive Rate of Models

Data Set and Model	AUC*	*P* Value	TPR and FPR at Optimal Threshold (%)†	*P* Value	TPR and FPR at TPR Equivalent to PI-RADS Score ≥3 (%)†	*P* Value
Internal test set, all cases						
Image-only	0.89 [0.85, 0.93]	.88	79 (104/132) [71, 85] 16 (43/268) [12, 21]	.08	91 (120/132) [85, 95] 45 (120/268) [39, 51]	<.001
Image+clinical	0.91 [0.88, 0.94]	.42	77 (102/132) [68, 84] 11 (30/268) [8, 16]	.63	90 (119/132) [84, 94] 33 (89/268) [28, 39]	.051
Image+clinical+radiologist	0.94 [0.91, 0.96]	<.001	82 (108/132) [74, 87] 7 (19/268) [5, 11]	.04	91 (120/132) [85, 95] 25 (68/268) [21, 31]	>.99
Radiologist	0.89 [0.86, 0.93]		85 (112/132) [78, 90] 13 (35/268) [10, 18]		91 (120/132) [85, 95] 25 (68/268) [21, 31]	
Clinical	0.81 [0.76, 0.86]					
Internal test set, pathology-proven cases						
Image-only	0.85 [0.80, 0.90]	.01	79 (104/132) [71, 85] 26 (25/96) [18, 36]	.89	90 (119/132) [84, 94] 58 (56/96) [48, 68]	.21
Image+clinical	0.88 [0.83, 0.92]	.001	77 (102/132) [69, 84] 19 (18/96) [12, 28]	.42	90 (119/132) [84, 94] 44 (42/96) [34, 54]	.003
Image+clinical+radiologist	0.87 [0.82, 0.92]	<.001	82 (108/132) [74, 87] 20 (19/96) [13, 29]	.04	91 (120/132) [85, 95] 60 (58/96) [50, 70]	.06
Radiologist	0.78 [0.72, 0.83]		85 (112/132) [78, 90] 36 (35/96) [28, 46]		91 (120/132) [85, 95] 71 (68/96) [61, 79]	
Clinical	0.77 [0.70, 0.84]					
External test set						
Image-only	0.86 [0.80, 0.91]	.68	83 (58/70) [72, 90] 27 (36/134) [20, 35]	.45	93 (65/70) [84, 97] 40 (54/134) [32, 49]	.43
Image+radiologist	0.89 [0.84, 0.93]	<.001	91 (64/70) [83, 96] 23 (31/134) [17, 31]	.33	93 (65/70) [84, 97] 34 (45/134) [26, 42]	.55
Radiologist	0.84 [0.79, 0.90]		87 (61/70) [77, 93] 25 (33/134) [18, 33]		93 (65/70) [84, 97] 36 (48/134) [28, 44]	

Note.—*P* values are in comparison with radiologists. AUC = area under the receiver operating characteristic curve, FPR = false-positive rate, TPR = true-positive rate.

* Data are AUCs, with 95% CIs in brackets.

† Data are percentages, with numbers of cases in parentheses and 95% CIs in brackets.

radiologist with more than 20 years of experience (N.T.). True-positive examinations determined by the image-only model at optimal threshold were reviewed. If there was an overlap with the location of the csPCa lesion with the pathologic report as a reference, it was considered a successful localization. Subanalyses in examinations with a PI-RADS score of 1 or 2 and with multiple csPCa lesions were also performed.

Codes used in developing the DL model and statistical analysis are available on GitHub *(https://github.com/jasonccai/prostate)*.

Results

Patient Characteristics

A total of 6141 examinations from 5555 individual patients ultimately met the inclusion criteria. A total of 524 patients underwent more than one examination during screening or before biopsy, accounting for 1110 examinations. A total of 406 examinations were excluded due to missing sequences or poor image quality (such as the presence of severe susceptibility and motion artifacts). The final sample contained 5735 examinations (from 5215 individual patients [mean age, 66 years ± 8 {SD}; all male]), of which 1514 examinations (1454 patients) showed csPCa. In this sample, pathologic confirmation was performed following 2821 examinations (2650 patients), while 2914 examinations (2565 patients) were classified as PI-RADS 1 or 2 during screening, for which pathologic confirmation was not subsequently performed (Table 1). The patient inclusion and exclusion criteria are summarized in Figure 1.

Model Evaluation

Data of model evaluation are summarized in Table 2.

When the entire internal test set was evaluated ($n = 400$), the image-only model had an AUC of 0.89 (95% CI: 0.85, 0.93; $P = .88$), while radiologists had an AUC of 0.89 (95% CI: 0.86, 0.93). *P* values are in comparison with the radiologists' AUC. The image+clinical model had an AUC of 0.91 (95% CI: 0.88, 0.94; $P = .42$), and the image+clinical+radiologist model had an AUC of 0.94 (95% CI: 0.91, 0.96; $P < .001$).

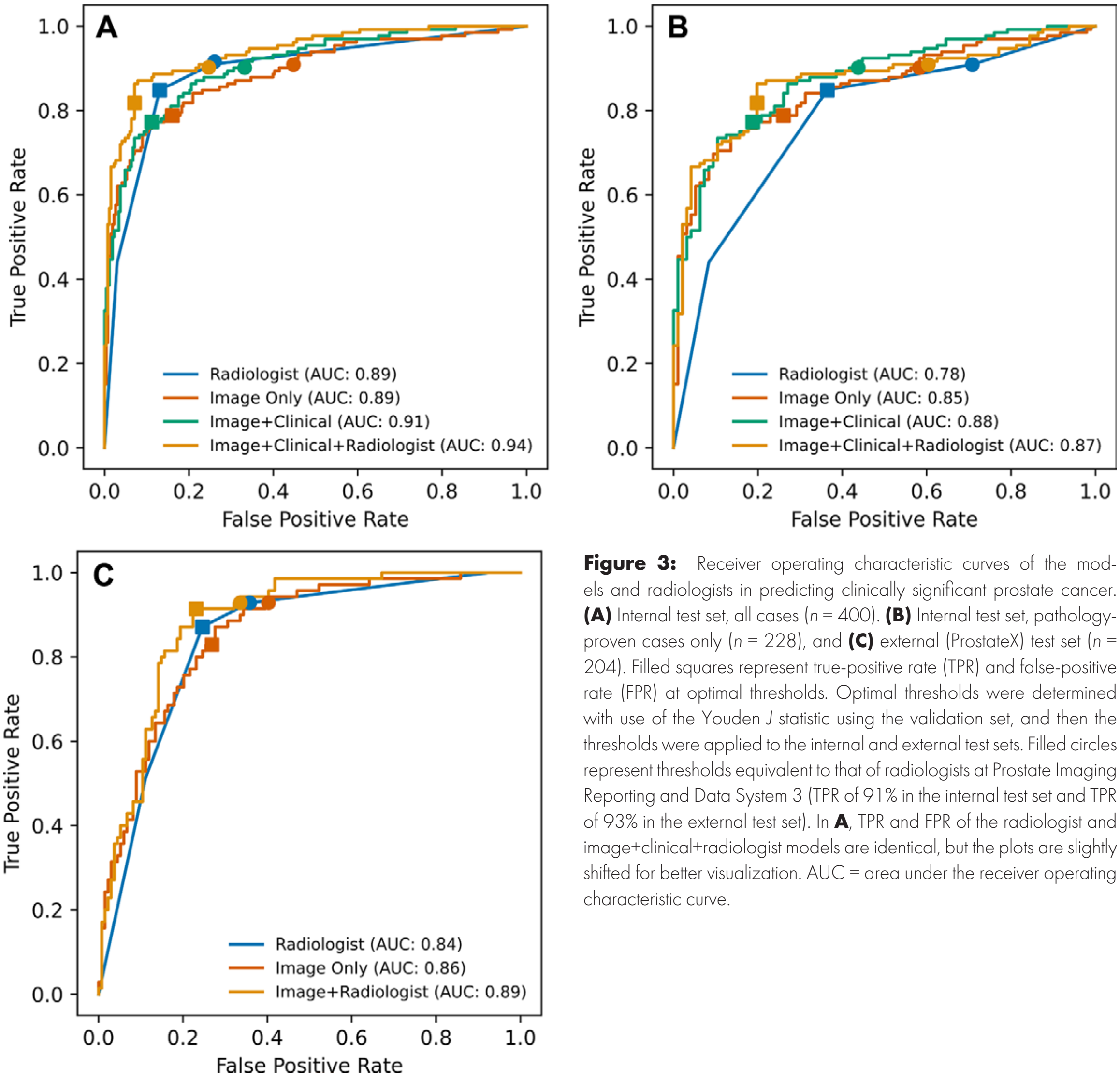

Figure 3: Receiver operating characteristic curves of the models and radiologists in predicting clinically significant prostate cancer. **(A)** Internal test set, all cases (*n* = 400). **(B)** Internal test set, pathology-proven cases only (*n* = 228), and **(C)** external (ProstateX) test set (*n* = 204). Filled squares represent true-positive rate (TPR) and false-positive rate (FPR) at optimal thresholds. Optimal thresholds were determined with use of the Youden *J* statistic using the validation set, and then the thresholds were applied to the internal and external test sets. Filled circles represent thresholds equivalent to that of radiologists at Prostate Imaging Reporting and Data System 3 (TPR of 91% in the internal test set and TPR of 93% in the external test set). In **A**, TPR and FPR of the radiologist and image+clinical+radiologist models are identical, but the plots are slightly shifted for better visualization. AUC = area under the receiver operating characteristic curve.

The receiver operating characteristic curves are shown in Figure 3. The clinical model had an AUC of 0.81 (95% CI: 0.76, 0.86). At optimal thresholds, TPR and FPR were 79% (104 of 132 examinations; 95% CI: 71, 85) and 16% (43 of 268 examinations; 95% CI: 12, 21) (*P* = .08) for the image-only model; 77% (102 of 132; 95% CI: 68, 84) and 11% (30 of 268; 95% CI: 8, 16) (*P* = .63) for the image+clinical model; and 82% (108 of 132; 95% CI: 74, 87) and 7% (19 of 268; 95% CI: 5, 11) (*P* = .04) for the image+clinical+radiologist model. *P* values are in comparison with the radiologists' TPR and FPR of 85% (112 of 132; 95% CI: 78, 90) and 13% (35 of 268; 95% CI: 10, 18), respectively, at optimal threshold (PI-RADS 4 or greater).

When patients who had MRI findings with no suspicious focal lesion (PI-RADS 1 or 2) without pathologic confirmation were excluded from the AUC analysis (*n* = 228), the image-only model had an AUC of 0.85 (95% CI: 0.80, 0.90; *P* = .01) for identifying csPCa in the internal test set, while radiologists had an AUC of 0.78 (95% CI: 0.72, 0.84). *P* values are in comparison with the radiologists' AUC. The image+clinical model had an AUC of 0.88 (95% CI: 0.83, 0.92; *P* = .001), and the image+clinical+radiologist model had an AUC of 0.87 (95% CI: 0.82, 0.92; *P* < .001). The clinical model had an AUC of 0.77 (95% CI: 0.70, 0.84).

At optimal thresholds, TPR and FPR were 79% (104 of 132 examinations; 95% CI: 71, 85) and 26% (25 of 96 examinations; 95% CI: 18, 36) (*P* = .89) for the image-only model; 77% (102 of 132; 95% CI: 69, 84) and 19% (18 of 96; 95% CI: 12, 28) (*P* = .42) for the image+clinical model; and 82% (108 of 132; 95% CI: 74, 87) and 20% (19 of 96; 95% CI: 13, 29) (*P* = .04) for the image+clinical+radiologist model. *P* values are in comparison with the radiologists' TPR and FPR of 85% (112 of 132; 95% CI: 78, 90) and 36% (35 of 96; 95% CI: 28, 46), respectively, at optimal threshold (PI-RADS 4 or greater).

For the external data set, the image-only model had an AUC of 0.86 (95% CI: 0.80, 0.91; *P* = .68), while radiologists had an AUC of 0.84 (95% CI: 0.79, 0.90). *P* values are in comparison with the radiologists' AUC. The image+radiologist model had an AUC of 0.89 (95% CI: 0.84, 0.93; *P* < .001).

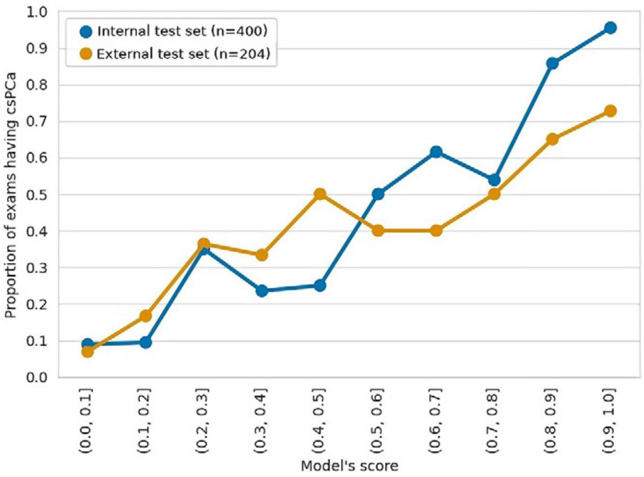

Figure 4: Calibration plot of the image-only model. The output of the image-only model (probability score) was categorized into 10 bins with increments of 0.1, and the proportion of clinically significant prostate cancer (csPCa) was calculated for each bin.

At optimal thresholds, TPR and FPR were 83% (58 of 70 examinations; 95% CI: 72, 90) and 27% (36 of 134 examinations; 95% CI: 20, 35) (*P* = .45) for the image-only model and 91% (64 of 70; 95% CI: 83, 96) and 23% (31 of 134; 95% CI: 17, 31) (*P* = .33) for the image+radiologist model. *P* values are in comparison with the radiologists' TPR and FPR of 87% (61 of 70; 95% CI: 77, 93) and 25% (33 of 134; 95% CI: 18, 33), respectively, at optimal threshold (PI-RADS 4 or greater).

Calibration plots of the image-only model on internal and external test sets are shown in Figure 4.

Grad-CAM Evaluation

Of the 150 cases from the internal test set reserved for Grad-CAM evaluation, 50 (33%) contained csPCa; of the 204 cases from the ProstateX data set, 70 (34%) contained csPCa. In the internal test set, Grad-CAMs overlapped with the csPCa lesion in 35 of the 38 true-positive examinations. In the ProstateX data set, Grad-CAMs overlapped with the csPCa lesion in 56 of the 58 true-positive examinations (Table 3). Only one of 37 patients had Grad-CAM activation over two csPCa lesions. Among true-positive examinations, four examinations were categorized as PI-RADS 2 by radiologists' evaluation, and three had Grad-CAM activation in the areas of csPCa described in the pathology report. Figures 5 and 6 show examples of the Grad-CAM outputs from cases in the internal and external test sets, respectively.

Table 3: Grad-CAM Performance

Data Set and Category	No. of Examinations
Internal test set	
TPR by image-only model*	38/50 (76)
Grad-CAM overlaps with csPCa	35/38 (92)
Grad-CAM does not overlap with csPCa	3/35 (9)
External test set	
TPR by image-only model*	58/70 (83)
Grad-CAM overlaps with csPCa	56/58 (97)
Grad-CAM does not overlap with csPCa	2/58 (3)

Note.—Data in parentheses are percentages. csPCa = clinically significant prostate cancer, Grad-CAM = gradient-weighted class activation map, TPR = true-positive rate.

* TPR was calculated using the optimal threshold determined using the validation set.

Discussion

MRI can help identify clinically significant prostate cancer (csPCa) but is limited by reader experience and interobserver variability. We developed a deep learning model to predict the presence of csPCa with use of MRI and compare its performance with radiologist performance. The areas under the receiver operating characteristic curves of the image-only model, radiologists, and image+radiologist model on the external test set were 0.86, 0.84, and 0.89, respectively. There was no evidence of a difference in image-only model performance from that of experienced radiologists (*P* = .68), and the image+radiologist model was better than the radiologist alone (*P* < .001). Our approach is unique in that the ground truth labels contained only the presence or absence of csPCa at the patient level without information about tumor location.

Previously reported machine learning or DL models were trained using ground truths that included manually annotated regions of interest obtained with reference to biopsy findings or whole-mount pathologic specimens. However, creating regions of interest is a resource-intensive process, limiting the number of cases that can be used for model development. In addition, only patients undergoing biopsy or prostatectomy can be included, and an imbalance of cases cannot be avoided. Schelb et al (14) supplemented the manually annotated regions of interest with results from the pathology reports of sextant biopsy. Recently, Bosma et al (25) developed a semisupervised method using teacher and student models augmented by radiology reports. However, manual annotation was still needed to train these models. A drawback of using patient-level labels is the need for a larger training set (26). However, use of patient-level labels also allowed us to include PI-RADS 1 and 2 examinations without biopsy as negative cases, reducing the inherent bias of unbalanced data sets.

As the output of our DL model did not include the location of the tumor, we used Grad-CAMs (18) for tumor localization. Grad-CAM analysis uses the gradient of the classification score with respect to the convolutional features determined by the network to understand which parts of the image are most important for classification. Although Grad-CAM analysis reliably localized the csPCa lesion, when multiple lesions were present, the remaining lesions were not highlighted. This may be because the classification model needed to identify only one lesion to perform the classification task correctly. Additionally, the approach was limited by poor spatial resolution, particularly in the z-axis, and could narrow the area of suspicion only to the superior, middle, or inferior third of the prostate. Due to these limitations, the Grad-CAM output cannot be directly used as a region of interest for subsequent MRI-guided (fusion) biopsy, and radiologists still need to draw regions of interest in areas of suspicion. Additionally, previous articles have raised concerns that Grad-CAM analysis and

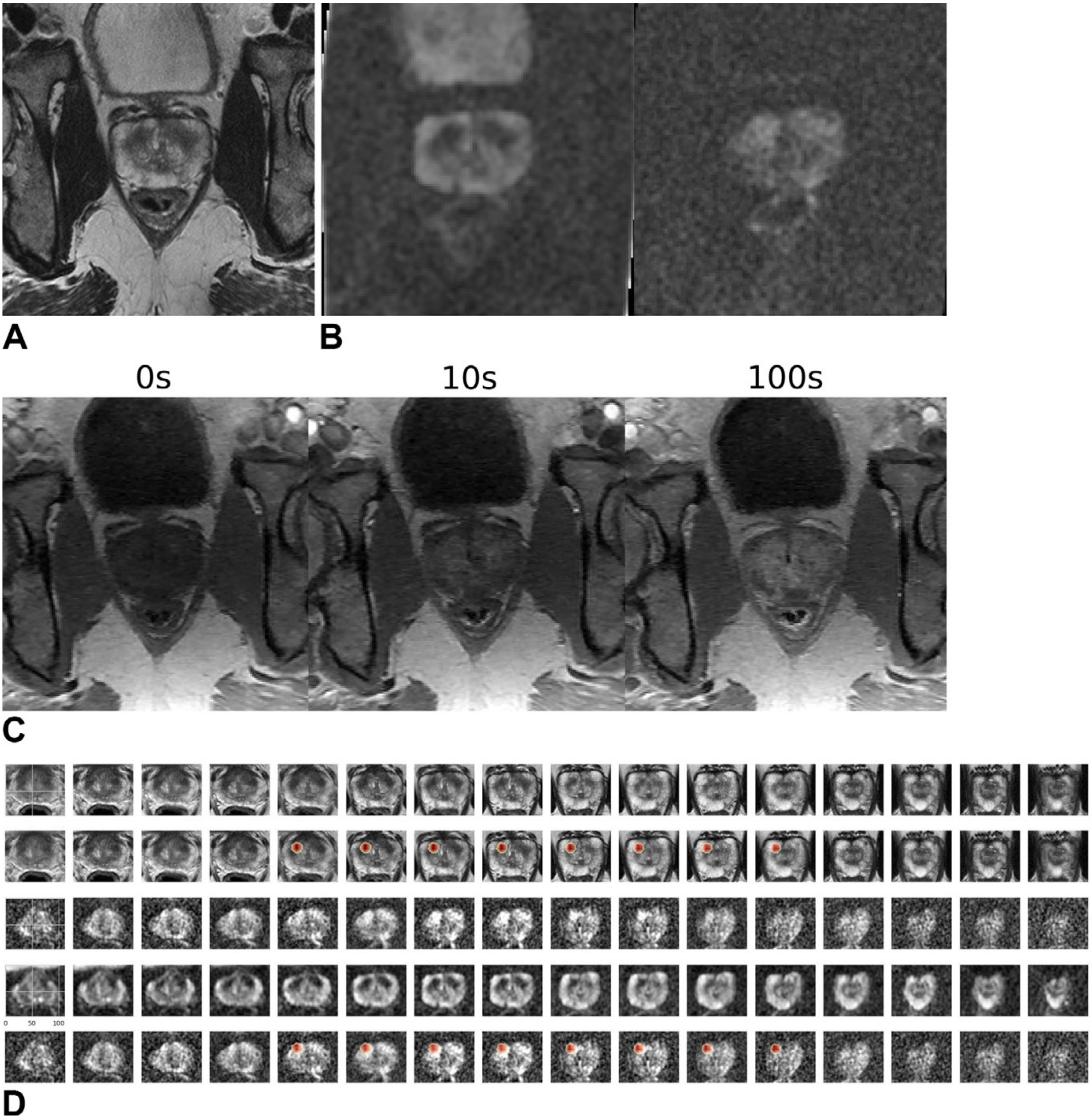

Figure 5: Images in a 59-year-old male patient who underwent MRI for clinical suspicion of prostate cancer (internal test set). The patient subsequently underwent prostatectomy and had a 1.5-cm prostate adenocarcinoma (Gleason score 3+4) in the right mid anterior to bilateral posterior inferior prostate gland. The model's output (patient level probability) was 0.83. Only the lesion in the right lobe was highlighted by the gradient-weighted class activation map (Grad-CAM). The radiologist graded this examination as Prostate Imaging Reporting and Data System (PI-RADS) 4 for the right lobe lesion and PI-RADS 3 for the left lobe lesion. **(A)** T2-weighted image (representative section). **(B)** Apparent diffusion coefficient map (representative section, left) and high-*b*-value diffusion-weighted image (representative section, right). **(C)** T1 dynamic contrast-enhanced images (representative sections). **(D)** Volumetric composite of T2-weighted images (rows 1 and 2), diffusion-weighted images (rows 3 and 5), and apparent diffusion coefficient maps (row 4), with superimposed Grad-CAMs (rows 2 and 5). All images are in the transverse plane.

other pixel attribution methods may be sensitive to random and adversarial data perturbations, which may result in misleading attributions (27,28). For these reasons, further investigation of the effectiveness of pixel attribution methods is required, and their interpretation necessitates expert judgment.

A large proportion of negative cases came from patients with PI-RADS 1 or 2 lesions who did not undergo biopsy. The sensitivity of PI-RADS for detecting prostate cancer is 87% (24). Therefore, a small proportion of our data set contained false-negative labels; such cases may be predicted as positive by the model and erroneously considered false-positive findings. We felt that this was a necessary trade-off to curate a large data set with disease prevalence similar to that of the screening population (19,20), and the benefits of a representative data

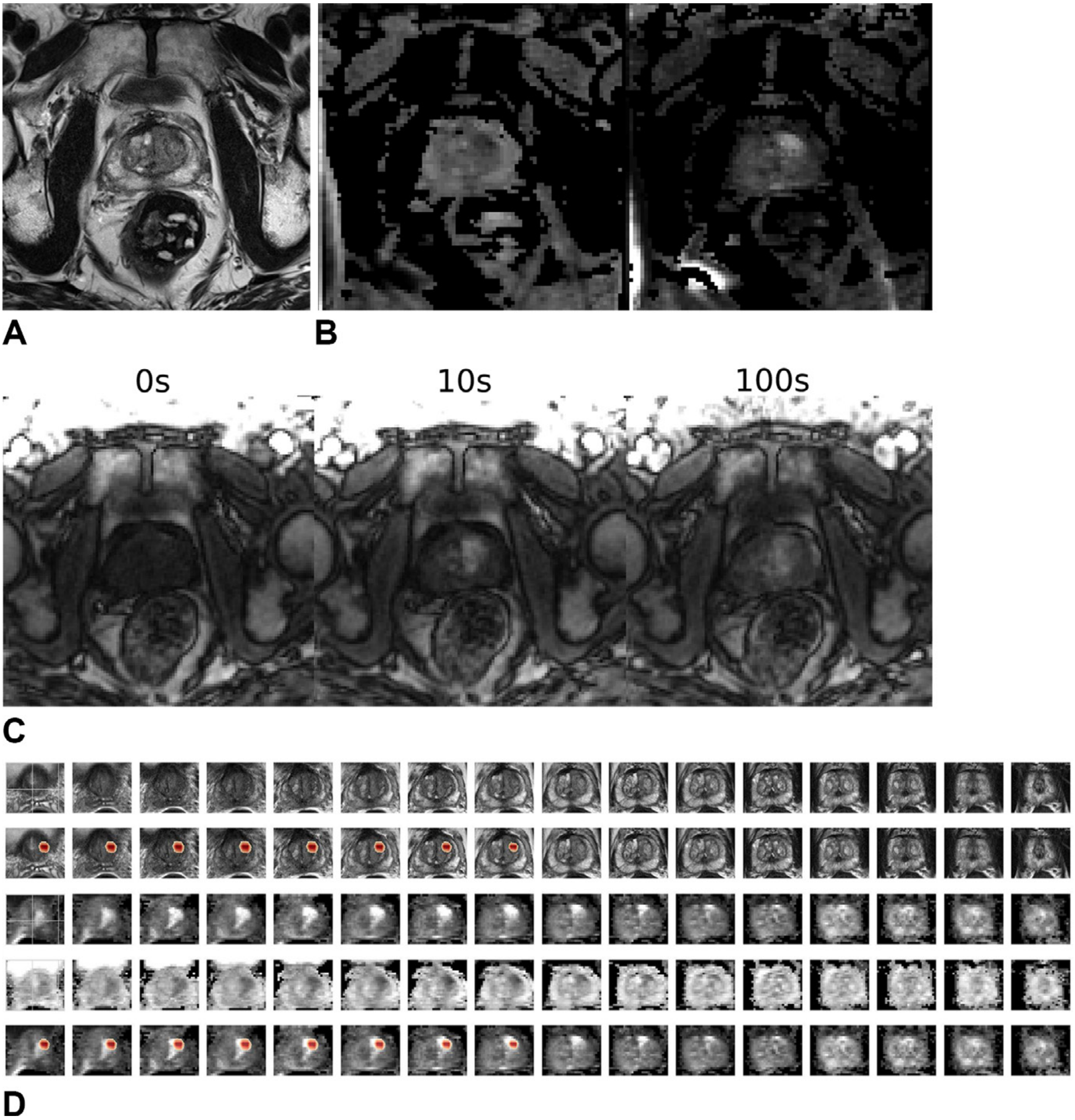

Figure 6: Images in a 64-year-old male patient who underwent MRI for clinical suspicion of prostate cancer (external test set). The patient was subsequently diagnosed with clinically significant prostate cancer in the left anterior transition zone at the base of the prostate gland. The model's output (patient-level probability) was 0.97. The radiologist graded this examination as Prostate Imaging Reporting and Data System 3 for the left lobe lesion. **(A)** T2-weighted image (representative section). **(B)** Apparent diffusion coefficient map (representative section, left) and high-*b*-value diffusion-weighted image (representative section, right). **(C)** T1 dynamic contrast-enhanced images (representative sections). **(D)** Volumetric composite of T2-weighted images (rows 1 and 2), diffusion-weighted images (rows 3 and 5), and apparent diffusion coefficient maps (row 4), with superimposed gradient-weighted class activation maps (rows 2 and 5). All images are in the transverse plane.

set outweighed the drawback of having a small proportion of incorrectly labeled data.

There were other limitations to our study. First, this was a retrospective, single-institution study, although the data set came from multiple sites. Second, we assessed combined model and radiologist performance using a logistic regression of their respective predictions; a formal clinical study to assess how radiologists use the model's predictions will be performed next. Third, only radiologists who specialized in prostate MRI participated in our study, and it is anticipated that the model will perform better than and further improve the diagnostic accuracy of trainees and general radiologists. Last, our model required dynamic contrast-enhanced sequences as input; hence, retraining is needed for biparametric data sets.

In conclusion, we developed a deep learning model to predict the presence of clinically significant prostate cancer (csPCa) with use of MRI. The model's performance was not statistically different from that of experienced radiologists on internal and external test sets. Gradient-weighted class activation mapping localized the csPCa lesion in the majority of positive predictions. We believe that our model has the potential to assist radiologists in identifying csPCa and facilitate lesion biopsy, hence improving the diagnosis of prostate cancer.

Deputy Editor: Vicky Goh
Scientific Editor: Shannyn Wolfe (AJE)

Author contributions: Guarantors of integrity of entire study, **J.C.C., S.K., M.R.H., N.T.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **J.C.C., S.K., C.W.B., A.K., L.A.M., A.K.B., N.T.**; clinical studies, **A.T.F., C.W.B., A.K., L.A.M., M.R.H., A.K.B., N.T.**; experimental studies, **J.C.C., H.N., A.T.F.**; statistical analysis, **J.C.C., S.K., P.R., N.T.**; and manuscript editing, **J.C.C., H.N., S.K., A.T.F., C.W.B., A.K., H.T., L.A.M., C.D.D., M.R.H., P.K., P.R., G.M.C., B.J.E., N.T.**

Data sharing: For internal data, data generated or analyzed during the study are available from the corresponding author by request. For external data, data generated by the authors are available from the corresponding author by request. Data analyzed during the study are available at reference 21.

Disclosures of conflicts of interest: J.C.C. No relevant relationships. **H.N.** No relevant relationships. **S.K.** No relevant relationships. **A.T.F.** No relevant relationships. **C.W.B.** No relevant relationships. **A.K.** No relevant relationships. **H.T.** No relevant relationships. **L.A.M.** No relevant relationships. **C.D.D.** No relevant relationships. **M.R.H.** Grant to institution from Boston Scientific; consulting fees from Applaud Medical, BD Peripheral, and Virtuoso; president of GSD Healthcare. **P.K.** No relevant relationships. **P.R.** No relevant relationships. **A.K.B.** No relevant relationships. **G.M.C.** No relevant relationships. **B.J.E.** Grants from the National Institutes of Health and Department of Defense; research chair for Society for Imaging Informatics in Medicine and advisory board member for the Temetry Centre for AI Research and Education in Medicine (T-CAIREM); stock in VoiceIt, board member for Yunu, and founder of FlowSIGMA. **N.T.** No relevant relationships.

References

1. Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer 2019;144(8):1941–1953.
2. Stabile A, Giganti F, Rosenkrantz AB, et al. Multiparametric MRI for prostate cancer diagnosis: current status and future directions. Nat Rev Urol 2020;17(1):41–61.
3. 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. Eur Urol 2019;76(3):340–351.
4. Smith CP, Harmon SA, Barrett T, et al. Intra- and interreader reproducibility of PI-RADSv2: a multireader study. J Magn Reson Imaging 2019;49(6):1694–1703.
5. Bonekamp D, Kohl S, Wiesenfarth M, et al. Radiomic machine learning for characterization of prostate lesions with MRI: comparison to ADC values. Radiology 2018;289(1):128–137.
6. Lay N, Tsehay Y, Greer MD, et al. Detection of prostate cancer in multiparametric MRI using random forest with instance weighting. J Med Imaging (Bellingham) 2017;4(2):024506.
7. Hiremath A, Shiradkar R, Fu P, et al. An integrated nomogram combining deep learning, Prostate Imaging-Reporting and Data System (PI-RADS) scoring, and clinical variables for identification of clinically significant prostate cancer on biparametric MRI: a retrospective multicentre study. Lancet Digit Health 2021;3(7):e445–e454.
8. Hamm CA, Baumgärtner GL, Biessmann F, et al. Interactive explainable deep learning model informs prostate cancer diagnosis at MRI. Radiology 2023;307(4):e222276.
9. 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 2021;73:102155.
10. Hosseinzadeh M, Saha A, Brand P, Slootweg I, de Rooij M, Huisman H. Deep learning-assisted prostate cancer detection on bi-parametric MRI: minimum training data size requirements and effect of prior knowledge. Eur Radiol 2022;32(4):2224–2234.
11. 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. Eur Radiol 2020;30(12):6582–6592.
12. Pellicer-Valero OJ, Marenco Jiménez JL, Gonzalez-Perez V, et al. Deep learning for fully automatic detection, segmentation, and Gleason grade estimation of prostate cancer in multiparametric magnetic resonance images. Sci Rep 2022;12(1):2975.
13. Ishioka J, Matsuoka Y, Uehara S, et al. Computer-aided diagnosis of prostate cancer on magnetic resonance imaging using a convolutional neural network algorithm. BJU Int 2018;122(3):411–417.
14. 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.
15. Rajagopal A, Redekop E, Kemisetti A, et al. Federated learning with research prototypes: application to multi-center MRI-based detection of prostate cancer with diverse histopathology. Acad Radiol 2023;30(4):644–657.
16. Cao R, Mohammadian Bajgiran A, Afshari Mirak S, et al. Joint prostate cancer detection and Gleason score prediction in mp-MRI via FocalNet. IEEE Trans Med Imaging 2019;38(11):2496–2506.
17. Bhattacharya I, Seetharaman A, Kunder C, et al. Selective identification and localization of indolent and aggressive prostate cancers via CorrSigNIA: an MRI-pathology correlation and deep learning framework. Med Image Anal 2022;75:102288.
18. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis 2016;128(2):336–359.
19. Nordström T, Discacciati A, Bergman M, et al; STHLM3 study group. Prostate cancer screening using a combination of risk-prediction, MRI, and targeted prostate biopsies (STHLM3-MRI): a prospective, population-based, randomised, open-label, non-inferiority trial. Lancet Oncol 2021;22(9):1240–1249.
20. Eldred-Evans D, Burak P, Connor MJ, et al. Population-based prostate cancer screening with magnetic resonance imaging or ultrasonography: the IP1-PROSTAGRAM Study. JAMA Oncol 2021;7(3):395–402.
21. Litjens G, Debats O, Barentsz J, Karssemeijer N, Huisman H. SPIE-AAPM PROSTATEx Challenge Data (Version 2). The Cancer Imaging Archive. https://www.cancerimagingarchive.net/collection/prostatex/. Updated July 5, 2022. Accessed July 5, 2024.
22. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention 2015. Lecture Notes in Computer Science, volume 9351; Springer, Cham: 234–241.
23. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digit Med 2020;3(1):136.
24. Woo S, Suh CH, Kim SY, Cho JY, Kim SH. Diagnostic performance of Prostate Imaging Reporting and Data System version 2 for detection of prostate cancer: a systematic review and diagnostic meta-analysis. Eur Urol 2017;72(2):177–188.
25. Bosma JS, Saha A, Hosseinzadeh M, Slootweg I, de Rooij M, Huisman H. Semisupervised learning with report-guided pseudo labels for deep learning-based prostate cancer detection using biparametric MRI. Radiol Artif Intell 2023;5(5):e230031.
26. Althnian A, AlSaeed D, Al-Baity H, Samha A, Dris A. Impact of dataset size on classification performance: an empirical evaluation in the medical domain. Appl Sci (Basel) 2021;11(2):796.
27. Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile. Proc AAAI Conf Artif Intell 2019, 33(01):3681–3688.
28. Kindermans PJ, Hooker S, Adebayo J, et al. The (Un)reliability of saliency methods. In: Samek W, Montavon G, Vedaldi A, Hansen L, Müller KR, eds. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Lecture Notes in Computer Science, volume 11700. Springer, Cham: 267–280.