Multimodal AI Combining Clinical and Imaging Inputs Improves Prostate Cancer Detection

Roest, Christian; Yakar, Derya; Rener Sitar, Dorjan Ivan; Bosma, Joeran S.; Rouw, Dennis B.; Fransen, Stefan Johannes; Huisman, Henkjan; Kwee, Thomas C.

Published in:
Investigative Radiology

DOI:
10.1097/RLI.0000000000001102

OPEN

Multimodal AI Combining Clinical and Imaging Inputs Improves Prostate Cancer Detection

Christian Roest, MSc, Derya Yakar, MD, PhD, Dorjan Ivan Rener Sitar, BSc, Joeran S. Bosma, MSc, Dennis B. Rouw, MD, Stefan Johannes Fransen, MSc, Henkjan Huisman, PhD, and Thomas C. Kwee, MD, PhD

Objectives: Deep learning (DL) studies for the detection of clinically significant prostate cancer (csPCa) on magnetic resonance imaging (MRI) often overlook potentially relevant clinical parameters such as prostate-specific antigen, prostate volume, and age. This study explored the integration of clinical parameters and MRI-based DL to enhance diagnostic accuracy for csPCa on MRI.

Materials and Methods: We retrospectively analyzed 932 biparametric prostate MRI examinations performed for suspected csPCa (ISUP ≥2) at 2 institutions. Each MRI scan was automatically analyzed by a previously developed DL model to detect and segment csPCa lesions. Three sets of features were extracted: DL lesion suspicion levels, clinical parameters (prostate-specific antigen, prostate volume, age), and MRI-based lesion volumes for all DL-detected lesions. Six multimodal artificial intelligence (AI) classifiers were trained for each combination of feature sets, employing both early (feature-level) and late (decision-level) information fusion methods. The diagnostic performance of each model was tested internally on 20% of center 1 data and externally on center 2 data (n = 529). Receiver operating characteristic comparisons determined the optimal feature combination and information fusion method and assessed the benefit of multimodal versus unimodal analysis. The optimal model performance was compared with a radiologist using PI-RADS.

Results: Internally, the multimodal AI integrating DL suspicion levels with clinical features via early fusion achieved the highest performance. Externally, it surpassed baselines using clinical parameters (0.77 vs 0.67 area under the curve [AUC], $P < 0.001$) and DL suspicion levels alone (AUC: 0.77 vs 0.70, $P = 0.006$). Early fusion outperformed late fusion in external data (0.77 vs 0.73 AUC, $P = 0.005$). No significant performance gaps were observed between multimodal AI and radiologist assessments (internal: 0.87 vs 0.88 AUC; external: 0.77 vs 0.75 AUC, both $P > 0.05$).

Conclusions: Multimodal AI (combining DL suspicion levels and clinical parameters) outperforms clinical and MRI-only AI for csPCa detection. Early information fusion enhanced AI robustness in our multicenter setting. Incorporating lesion volumes did not enhance diagnostic efficacy.

Received for publication April 8, 2024; and accepted for publication, after revision, May 21, 2024.

From the Department of Radiology, Medical Imaging Center, University Medical Center Groningen, Groningen, the Netherlands (C.R., D.Y., D.I.R.S., S.J.F., T.C.K.); Department of Radiology, Netherlands Cancer Center Antoni van Leeuwenhoek, Amsterdam, the Netherlands (D.Y.); Department of Radiology, Radboud University Medical Center, Nijmegen, the Netherlands (J.S.B., H.H.); and Department of Radiology, Martini Ziekenhuis Groningen, Groningen, the Netherlands (D.B.R.).

Conflicts of interest and sources of funding: C.R., T.C.K., D.Y., and H.H. are receiving a grant from Siemens Healthineers. H.H. is receiving a grant from Canon Medical Systems. For the remaining authors, none were declared.

Data availability: All experimental and statistical code used for this study will be publicly available at the following repository following publication: https://github.com/0xC4/multimodal-ai. Data will be provided upon reasonable request.

Correspondence to: Christian Roest, MSc, Department of Radiology, Medical Imaging Center, University Medical Center Groningen, Hanzeplein 1, 9713 GZ, Groningen, the Netherlands. E-mail: c.roest@umcg.nl.

Supplemental digital contents are available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.investigativeradiology.com).

ISSN: 0020-9996/24/5912–0854
DOI: 10.1097/RLI.0000000000001102

Key Words: artificial intelligence, magnetic resonance imaging, prostate cancer

(*Invest Radiol* 2024;59: 854–860)

Prostate cancer (PCa) is a common disease, affecting an annual 1.4 million men worldwide.[1] Magnetic resonance imaging (MRI) and clinical parameters (including patient age, prostate-specific antigen [PSA] levels, PSA density, and prostate volume) have diagnostic value for detecting clinically significant prostate cancer (csPCa; International Society of Urological Pathology [ISUP] ≥2).[2,3]

Artificial intelligence (AI) for PCa diagnosis on MRI is a promising area of research, with recent studies reporting diagnostic performances approaching that of radiologists.[4–6] A previous systematic review on diagnostic AI for prostate MRI by Syer et al[7] revealed that only 2 out of 27 included studies integrated clinical parameters beyond the radiologist score, whereas the remainder of studies were restricted to image-based analyses. In clinical practice, however, other relevant parameters including PSA, prostate volume, and age are considered alongside MRI.[8] The availability of this information from different sources allows the development of multimodal AI, which incorporates multiple predictors to increase overall diagnostic accuracy.[9] A review by Lipkova et al[10] describes 3 distinct methods of integrating multimodal information (referred to as "information fusion"), with each method referring to integration at a different stage in the pipeline: early fusion, in which all data sources are simultaneously provided to a single model; intermediate fusion, where feature extractors for each modality are updated during training through backpropagation based on the combined prediction; and late fusion, describing the aggregation of predictions of multiple unimodal predictors. Studies involving other organs than prostate have demonstrated that various information fusion methods can enhance diagnostic performance.[11,12] However, the optimal fusion strategy is strongly dependent on the task, and no previous studies have systematically compared these strategies in the context of csPCa detection.

Therefore, this study aimed to determine if the addition of clinical parameters (ie, PSA levels, prostate volume, age) into a multimodal AI system improves the detection of csPCa on MRI. We compared the effect of early versus late information fusion on the diagnostic performance across internal and external datasets. Additionally, we evaluated the effect of incorporating volumetric information derived from lesion segmentations.

MATERIALS AND METHODS

Study Sample

This retrospective study included patients from 2 independent healthcare institutions: University Medical Center Groningen ("center 1"); and Martini Ziekenhuis Groningen ("center 2"). The institutional review boards of the respective institutions waived the requirement for informed consent for this retrospective study (center 1: IRB-2018/597, center 2: IRB-2019-056).

All men who underwent biparametric or multiparametric MRI of the prostate because of clinical suspicion of csPCa based on (1) high PSA levels, (2) abnormal digital rectal examination, (3) lower urinary

TABLE 1. Overview of Magnetic Resonance Imaging Characteristics

	Center 1		Center 2	
	0%		0%	
Endorectal Coil	**T2w**	**DWI**	**T2w**	**DWI**
In-plane resolution (mm)	0.31 ± 0.12	1.46 ± 0.41	0.35 ± 0.01	1.1 ± 0.28
Spacing between slices (mm)	3.3 ± 0.13	3.3 ± 0.20	3.04 ± 0.21	3.03 ± 0.16
Number of averages	2.97 ± 0.29	5.58 ± 2.66	1.02 ± 0.14	6.69 ± 3.02
Echo train length	25.13 ± 4.16	52.42 ± 6.99	20.12 ± 1.03	65.97 ± 9.36
Computed b-value	-	1400 (100%)	-	2000 (96%) 1400 (4%)
Acquired b-values	-	50-500-1000 (95%) 50-400-800 (5%)	-	0-50-400-800 (96%) 0-500-800-1000 (4%)

Continuous values are presented as mean ± standard deviation.
DWI, diffusion-weighted imaging; T2w, T2-weighted imaging.

tract symptoms, and/or (4) follow-up of low-risk PCa were eligible for inclusion. We excluded patients who had (1) undergone previous treatment of the prostate, (2) missing PSA levels, and (3) missing histopathological confirmation following positive (ie, Prostate Imaging Reporting and Data System [PI-RADS] ≥3) MRI.[13]

Reference Standard

The primary outcome of this study was the detection of csPCa on MRI. The ground truth label ("csPCa" or "no csPCa") was established through systematic biopsies (8–12 cores), targeted biopsies (2–4 cores per lesion, with option for 4–10 additional systematic cores), and/or prostatectomy, with ISUP grade ≥2 considered csPCa. Additionally, due to the high sensitivity of PI-RADS ≥3 for csPCa lesions, we considered negative MRI examinations (ie, PI-RADS ≤2) without further histopathological examination as "no csPCa," unless subsequent histopathological examination was recorded within 2 years after MRI, in which case the histopathological reference standard was preferred.[14] PI-RADS scores were retrospectively assigned by a radiologist (center 1: D.Y. with >15 years of experience reading prostate MRI, center 2: D.B.R. with 13 years of experience reading prostate MRI).

Data Extraction

This study explored a biparametric MRI protocol, consisting of axial T2-weighted (T2w) imaging, apparent diffusion coefficient (ADC) maps, and calculated high b-value (≥1400) diffusion-weighted imaging. Center 1 scanners included 3 T Skyra (63%), 3 T Prisma (29%), 1.5 T Aera (5%), and other (3%), Siemens Healthineers, Erlangen, Germany. Center 2 scanners included 3 T Ingenia (95%), 1.5 T Intera (2%), and 1.5 T Achieva dStream (2%), Philips Medical Systems, Best, the Netherlands. Detailed scan parameters are provided in Table 1. Scan quality was prospectively verified by the radiologist, who ordered a repeat scan if the quality was deemed insufficient for diagnosis. Clinical parameters (PSA, prostate volume, and patient age) were extracted from electronic patient files. Prostate volumes were obtained from MRI, aligning with the MRI-first approach to PCa diagnosis.

Data Preprocessing

MRI scans were extracted as raw DICOM files from each institution's picture archive and subsequently converted to NIFTI volumes for further processing. The in-plane image resolutions were normalized by resampling each scan to standardized voxel spacing (0.5×0.5 mm^2). We refrained from resampling along the longitudinal axis due to low vertical resolution (3–5 mm) in axial sequences, which could introduce significant blurring artifacts. Sequences were spatially aligned using real-world coordinates and centrally cropped to $192 \times 192 \times 24$ voxels. The dynamic ranges of T2w and high-b value sequences were normalized using Z-score normalization. ADC map dynamic ranges were adjusted by division with a constant (3000) to prevent gradient issues during training while maintaining the numerical significance of clinically relevant ADC values.

Data Split

MRI examinations from center 1 were randomly allocated to either the training (80%) or internal test set (20%). Examinations from center 2 were held out during model development and served as external test data.

Deep Learning Lesion Detection

Each MRI examination included in the study underwent AI analysis to detect csPCa using a previously published deep learning (DL) model by Bosma et al.[15] Briefly, it comprised a self-configuring nnU-Net trained on 7756 prostate MRI examinations of men scanned at Radboudumc (Nijmegen, the Netherlands; center 3) with clinical suspicion of csPCa.[15,16] For a given MRI scan, the model produces a heat map containing voxel-level likelihood scores for the presence of csPCa.

This model was used to generate a csPCa heat map for each MRI examination. We then obtained AI delineations for up to 3 detected lesions with the highest degree of suspicion of csPCa on the heat map by applying a dynamic threshold.[15] This involved iteratively identifying peak points in the heat map and extending the segmentation to include neighboring voxels with suspicion levels ≥75% of the peak. These peaks served as csPCa suspicion scores (continuous from 0%–100%) for each lesion, and lesion volumes (in cc) were calculated as the voxel count in each segmentation multiplied by the volume of a single voxel. For examinations where fewer than 3 lesions were detected, suspicion and volume were recorded as zeros, indicating missingness. A visual example of lesion detection and segmentation is shown in Figure 1.

Feature Sets

The collected data were used to create 3 sets of features: (1) DL suspicion levels for identified lesions; (2) clinical parameters (PSA, prostate volume, and age); and (3) lesion volume features, comprising lesion volume estimates for each detected lesion based on DL segmentations.

Information Fusion

We compared 2 of the fusion techniques described in Lipkova et al[10]: early fusion, where DL-based risk levels, lesion volumes, and clinical parameters are combined into a combined dataset and used to train a single AI classifier, and late fusion, which integrates separately trained unimodal classifiers through an aggregation of predictions. Figure 2 shows a schematic overview of early and late fusion.

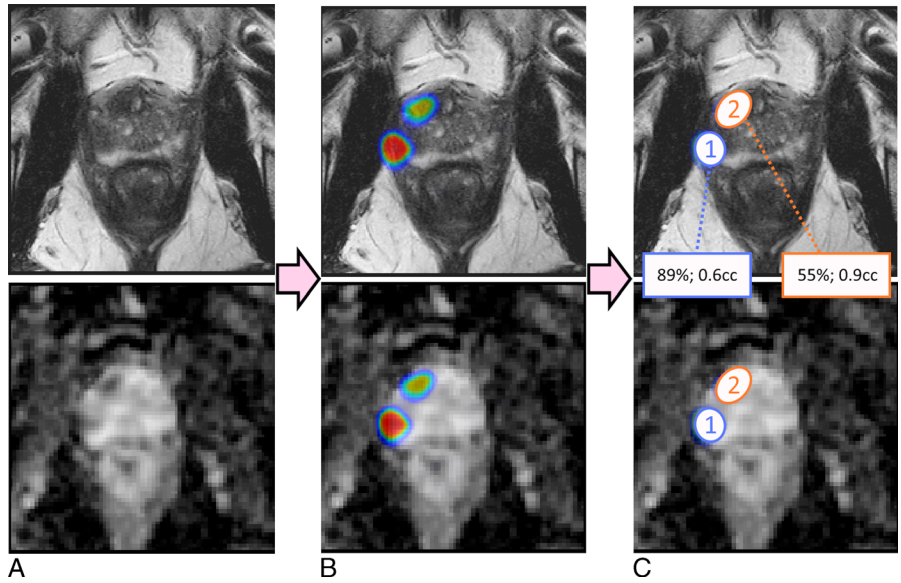

FIGURE 1. An overview of the lesion detection and segmentation approach. A, Each MRI scan is automatically analyzed by the detection AI to produce a csPCa heat map (B, shown as colored overlay). A dynamic threshold[15] is applied to the heat map to obtain segmentations for individual detections (C). For each detection, the suspicion level and volume are extracted.

Model Training

We trained early and late fusion AI models and optimized hyperparameters to maximize diagnostic value while minimizing the influence of hyperparameter selection. For classification layers, we considered 5 widely used algorithms: logistic regression, decision trees, multilayer perceptrons, support vector machines, and gradient boosting. The choice of classification layer and respective hyperparameters was simultaneously optimized within a specified range through 100 trials per experiment, using the 5-fold cross-validated area under the curve (AUC) as the optimization objective. A tree-structured Parzen estimator suggested promising hyperparameter combinations based on objective values from previous iterations, avoiding computationally expensive grid searches.[17] Finally, the hyperparameter configuration with the best cross-validated AUC was retrained using the full training dataset and evaluated on internal and external test datasets. Hyperparameter optimization was implemented using Optuna version 3.3.0.[18] An overview of the model training is shown in Figure 3. The model and hyperparameter search space are given in Supplemental Table 1, http://links.lww.com/RLI/A951.

Experiments

Three experiments were performed to study the benefit of multimodal analysis. First, we obtained the optimal multimodal model by comparing the diagnostic performances between AI models trained with different feature sets and fusion methods. Second, we compared the diagnostic efficacy of the optimal multimodal AI with both clinical and DL baseline models, the latter utilizing the conventional approach of taking the maximum DL suspicion score at the patient level. Third, we compared the AI diagnostic performance to that of a radiologist using PI-RADS.

Statistical Analysis

Baseline patient characteristics were summarized and tested for differences between cohorts using Wilcoxon tests. Receiver operating characteristic analysis was performed to compare the diagnostic performances between models. Diagnostic accuracy was quantified using the AUC. Differences in AUC were assessed for statistical significance using 2-sided DeLong's tests, with P values <0.05 considered statistically significant

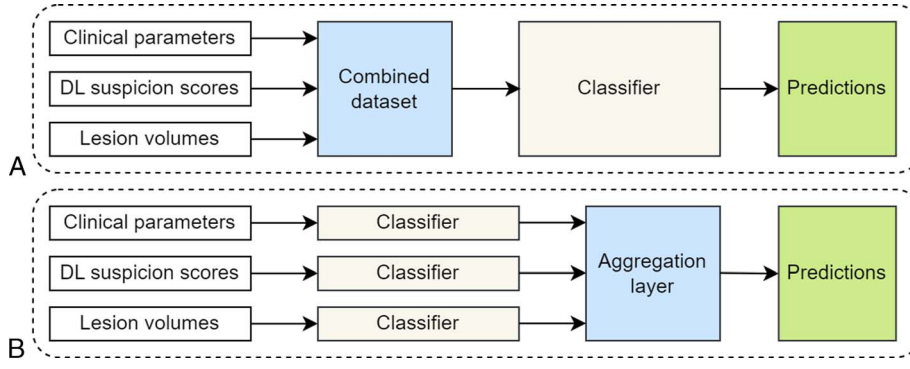

FIGURE 2. A schematic overview showing (A) early fusion and (B) late fusion. In early fusion, separate information sources are combined before classifiers are trained. In late fusion, an aggregation layer combines predictions from unimodal classifiers. DL, deep learning.

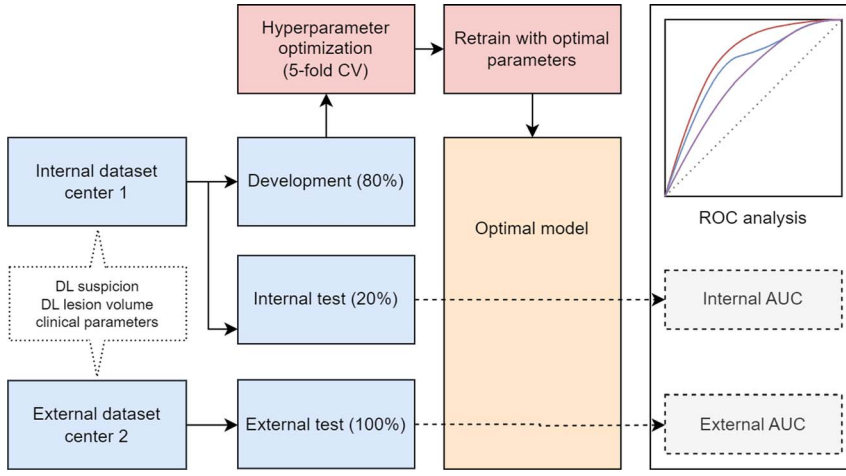

FIGURE 3. An overview of the methodology used to train each multimodal AI model. AUC, area under the receiver operating characteristic curve; CV, cross-validation; DL, deep learning; ROC, receiver operating characteristic.

results.[19] The 95% confidence intervals (CIs) for diagnostic accuracies were calculated using DeLong's method. *P* values were adjusted for multiple comparisons using Holm's method.[20] All statistical analyses were performed in R version 4.2.2 using the pROC package.

RESULTS

Baseline Characteristics

Inclusion criteria identified 1052 potentially eligible MRI examinations of 953 patients. After exclusion, 932 examinations of 810 patients were finally included in our study (Fig. 4). Baseline characteristics for each dataset partition are presented in Table 2. Statistical comparisons revealed a significant difference in the distribution of reference standards between centers ($P < 0.001$). No further statistically significant differences were found between baseline characteristics from center 1 and center 2.

Optimal Multimodal Configuration

The highest diagnostic performance on the internal validation set was reached by the multimodal AI combining DL suspicion levels with clinical parameters (PSA, prostate volume, age) using early fusion (0.87 AUC, 95% CI: 0.78–0.96) and a logistic classifier, which was selected as the optimal configuration. Externally, this model achieved an AUC of 0.77 (95% CI: 0.73–0.82).

In comparison, the early fusion model performed similarly to late fusion in the internal test set (0.874 vs 0.871 AUC, $P = 0.92$), but demonstrated significantly higher performance than late fusion in the external test data (0.77 vs 0.73 AUC, $P = 0.005$). Incorporating DL-based lesion volume features resulted in significantly decreased diagnostic performance in external data (external: 0.77 vs 0.69 AUC, $P < 0.001$), but not in internal data (0.87 vs 0.77 AUC, $P = 0.14$).

Comparison to Unimodal Baselines

The performance of the optimal multimodal AI was significantly higher compared with the unimodal baseline AI using only clinical parameters (internal: 0.87 vs 0.72 AUC, $P = 0.05$; external: 0.77 vs 0.67 AUC, $P < 0.001$). The multimodal AI also reached higher diagnostic accuracy compared with the baseline DL model in internal (0.87 vs 0.78 AUC, $P = 1$) and external test data (0.77 vs 0.70, $P = 0.006$), although the former was not significant.

Comparison to Radiologist

The radiologists achieved a diagnostic performance of 0.88 AUC (95% CI: 0.79–0.97) in the internal validation set and 0.75 AUC (95% CI: 0.71–0.79) in the external test cohort. The radiologists' sensitivity and specificity at PI-RADS ≥3 and ≥4 are shown in Figure 5. No significant differences in AUC were found between the optimal multimodal AI and radiologists (internal: 0.87 vs 0.88 AUC, $P = 1$; external: 0.77 vs 0.75 AUC, $P = 1$). However, in external data,

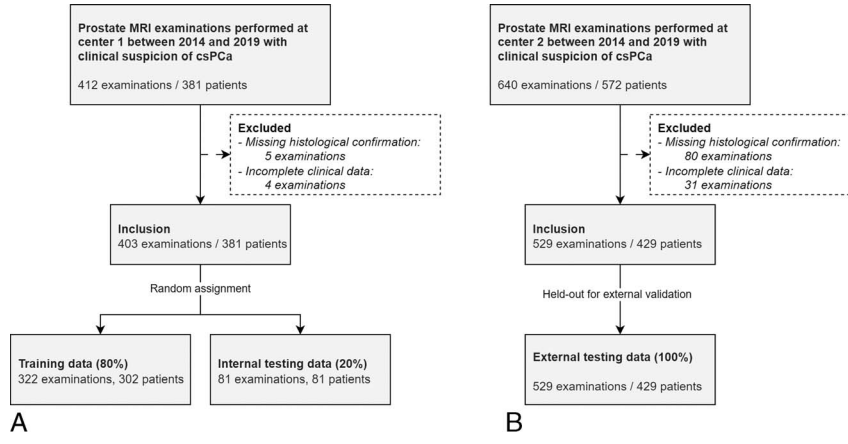

FIGURE 4. Study flow chart for the internal cohort (a) and external cohort (b). Note: csPCa=clinically significant prostate cancer.

TABLE 2. Overview of Baseline Demographic and Clinical Characteristics

	Center 1		Center 2	
Dataset	**Development**	**Test**	**Test**	**P**
n	322	81	529	
Age	70 (64–74)	70 (64–75)	70 (65–74)	0.95*
PSA (µg/L)	8.2 (5.8–11.7)	9.0 (5.4–15)	8.5 (6–12)	0.78*
Prostate volume (mL)	47 (33–71)	52 (32–69)	50 (35–68)	0.36*
PI-RADS (%)				0.06†
1–2	151 (47)	39 (48)	241 (46)	
3	37 (11)	11 (14)	35 (7)	
4	87 (27)	21 (26)	153 (29)	
5	41 (13)	9 (11)	100 (18)	
csPCa (%)				1†
No	227 (70)	61 (75)	383 (72)	
Yes	95 (30)	20 (25)	146 (28)	
Ground truth				<0.001†
SB	61 (19)	12 (15)	264 (50)	
TB	92 (29)	29 (35)	87 (16)	
TB + SB	70 (22)	14 (17)	77 (15)	
Surgery	1 (<1)	1 (1)	101 (19)	
Negative MRI (PI-RADS ≤ 2)	98 (30)	25 (31)	-	

Numerical values are reported as median (quartile 1–quartile 3). Categorical variables are summarized as frequency (%).

*Wilcoxon rank sum test.

†χ^2 test.

csPCa, clinically significant prostate cancer; PI-RADS, Prostate Imaging-Reporting and Data System; PSA, prostate-specific antigen; SB, systematic biopsy; TB, targeted biopsy.

the sensitivity of the AI at PI-RADS ≥4 equivalent specificity of 0.64 was lower than that of the radiologist (0.77 [95% CI: 0.68–0.84] vs 0.80 [95% CI: 0.74–0.86]).

Importance of Predictors

A post hoc jackknife test was conducted to assess the importance of predictors in the final model.[21] Omitting the DL suspicion score of the primary lesion and PSA levels resulted in the most notable decrease in AUC compared with the full model, suggesting their importance as predictors (Fig. 6). These findings were consistent across both internal and external test sets.

DISCUSSION

This study explored the integration of clinical parameters and MRI-based DL to enhance diagnostic accuracy for csPCa on MRI. Our findings indicate that combining DL lesion suspicion scores with clinical parameters (PSA, prostate volume, age) improves csPCa detection accuracy, surpassing both MRI-only and clinical baseline models. Furthermore, external testing demonstrated superior diagnostic performance by using early fusion compared with late fusion of multimodal information.

Notably, our experiments revealed significant differences between information fusion methods, with early fusion surpassing late fusion in external data. Previous works lacked a comparison of different techniques for integrating DL predictions with clinical variables in the setting of csPCa detection and used direct methods such as logistic regression to combine clinical and DL data. Our results suggest that combining a DL and clinical information using early fusion can be used to achieve superior diagnostic performance compared with late fusion models in external data. This finding appears to contradict Lipkova et al,[10] who suggested that late fusion may improve generalizability by reducing the reliance on a dominant modality, especially when there are differences in information density. The discrepancy suggests that merging at the decision level may be suboptimal for csPCa detection, and that the joint representations of both MRI and clinical data may be crucial for improving diagnostic accuracy. The late-fusion approach could have resulted in weaker unimodal classifiers that struggle to handle challenging cases where joint features could have helped to distinguish between them. Nevertheless, Lipkova et al[10] previously concluded that the optimal fusion technique is highly dataset and task dependent, and may thus be dependent on factors such as scan quality and population characteristics. Thus, we recommend that future investigations into csPCa detection using multimodal AI prioritize early fusion techniques when integrating clinical and DL features, to improve robustness in multicenter datasets.

Potentially, integrating clinical values directly into the DL model could lead to additional enhancements in diagnostic performance. However, this approach poses technical challenges because DL models based on convolutional neural networks typically rely on 3-dimensional operations, which are incompatible with the 1-dimensional nature of clinical features. Overcoming this mismatch would be necessary to effectively incorporate clinical data into the model architecture, which requires further exploration in future studies. On the contrary, the current approach offers the advantage of being easily applied as a cascaded component following any image-based detection model, allowing for seamless integration into existing workflows.

Given previous studies demonstrating diagnostic performances of image-based AI models approaching the performance of radiologists, we hypothesized that optimally incorporating additional parameters could help AI models to move closer to this performance.[7,22] Our best AI model achieved an externally validated diagnostic performance comparable to that of a radiologist (AI: 0.77 AUC; radiologist: 0.75 AUC) and performed similarly at the clinically relevant operating point of PI-RADS ≥3. However, at the PI-RADS ≥4 operating point, the radiologist still achieved higher sensitivity than the AI. The performance of the AI-model might be increased by using more data,[23] incorporating AI-certainty/uncertainty metrics,[24] or adding other sources of information (eg, genomics[25]). Although it may seem unfair to compare an AI model that integrates imaging and clinical information to PI-RADS scores, the scores were prospectively assigned in a real-world setting where radiologists also had access to clinical information, as required by the PI-RADS version 2 guidelines.[13]

Interestingly, the addition of DL lesion volume estimates as a predictor did not improve the diagnostic performance of AI models in our study. This appears to contradict previous evidence by Martorana et al,[26] who observed up to 4 times higher csPCa detection rates for lesions with volume >1 cc compared with lesions <0.5 cc by radiologists using PI-RADS. This discrepancy can be explained by the fact that their study considered lesion volume in isolation, whereas our study only considered it in conjunction with other relevant predictors, including DL suspicion levels. Since the DL lesion detection model used in this study was trained in a large dataset comprising over 7000 MRI examinations, it is reasonable that a correlation between lesion volume and csPCa would be learned by the DL model itself, and therefore does not require incorporation as a separate parameter. The potential collinearity introduced by adding lesion volumes as a predictor (eg, with DL suspicion levels) may also explain the observed drop in performance, as it could lead to an overdependency on correlated features, which is ultimately detrimental to model performance. Finally, although we utilized an effective segmentation algorithm previously reported to achieve a Dice score of 0.67 for report-guided segmentation of csPCa

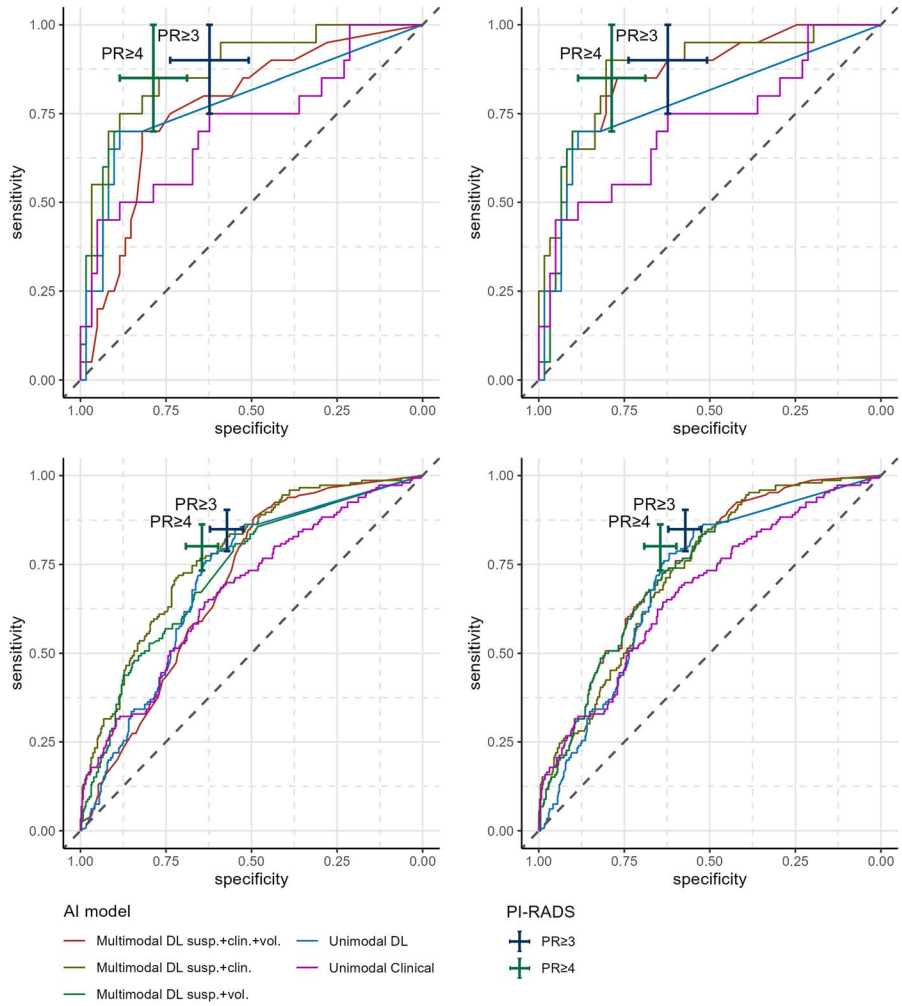

FIGURE 5. Receiver operating characteristic curves showing the diagnostic performance for AI models trained using early (left) and late (right) fusion methods in the internal test set (top row) and external test set (bottom row). Cross markers show the sensitivity and specificity with 95% CIs for the radiologist at PI-RADS ≥3 and ≥4 operating points. clin., clinical parameters; DL, deep learning; PI-RADS/PR, Prostate Imaging-Reporting and Data System; susp., DL suspicion levels; vol., lesion volumes.

lesions, the impact of the choice of segmentation algorithm on the benefit of lesion volumes was not evaluated in this study.[15]

Our results showed a marked difference in the AUCs between internal and external test datasets, with similar performance drops observed for the AI models and the radiologists. Several factors may have contributed to this difference. Importantly, aligning with current guidelines, center 1 avoided biopsies in some MRI-negative patients, resulting in a lack of histopathological confirmation that could refute the radiologist's interpretation. In contrast, all patients at center 2, including those with negative MRI results, had histopathological confirmation available. This resulted in a larger potential for discrepancies in MRI-negative patients at center 2. Additionally, center 2 had a larger proportion of patients undergoing surgery (1% at center 1 vs 19% at center 2), which may also have increased the likelihood of revealing csPCa missed by the radiologist. Further contributors to intercenter AUC differences include population differences and variations in scan

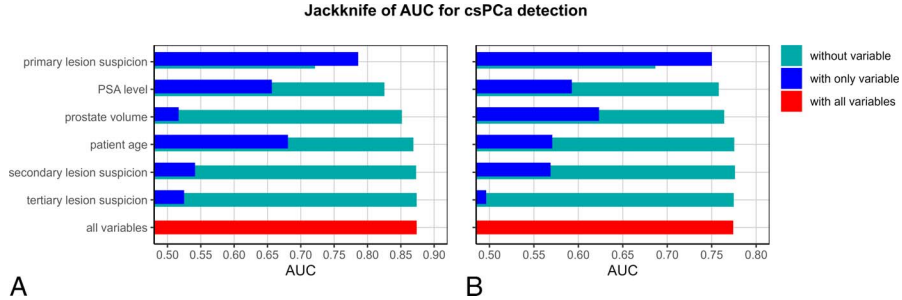

FIGURE 6. Jackknife analysis illustrating the variable importance of predictors in the final model for the internal test set (a) and external test set (b). Variables are ordered based on their impact on the AUC when omitted from the model in the internal dataset. Note: PSA=prostate-specific antigen.

quality. Nevertheless, we believe that these sources of heterogeneity have not impacted the validity of our results, considering that diagnostic accuracies were compared under the same conditions within each dataset.

Our study had some limitations. First, because a large portion of the internal data was reserved for model development, the internal test dataset was relatively small (n = 81). This limited sample size may have reduced the statistical power of certain comparisons. However, external validation in a larger dataset comprising 529 MRI examinations confirmed our findings from the internal evaluation. Second, the DL lesion detection model used in this study was previously trained on MRI scans from an external institution. DL models have previously been shown to be sensitive to variations in vendors and scanners, which could potentially lead to reduced diagnostic accuracy.[27] However, this limitation affected all methods equally, minimizing its impact on the observed performance differences. Lastly, the inclusion of patients with multiple scans may have led to overrepresentation of certain subjects and increased correlation between observations, potentially biasing the model toward specific patient characteristics and limiting its generalizability to new cases.

CONCLUSIONS

Multimodal AI (combining DL suspicion levels and clinical parameters) outperforms clinical and MRI-only AI for csPCa detection. Early information fusion enhanced AI robustness in our multicenter setting. Incorporating lesion volumes did not enhance diagnostic efficacy.

REFERENCES

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024;74:229–263.
2. Mehralivand S, Bednarova S, Shih JH, et al. Prospective evaluation of PI-RADS™ version 2 using the International Society of Urological Pathology Prostate Cancer Grade Group System. *J Urol*. 2017;198:583–590.
3. van Leenders GJLH, van der Kwast TH, Grignon DJ, et al. The 2019 International Society of Urological Pathology (ISUP) consensus conference on grading of prostatic carcinoma. *Am J Surg Pathol*. 2020;44:e87–e99.
4. Cao R, Zhong X, Afshari S, 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*. 2021;54:474–483.
5. Netzer N, Weißer C, Schelb P, et al. Fully automatic deep learning in bi-institutional prostate magnetic resonance imaging: effects of cohort size and heterogeneity. *Invest Radiol*. 2021;56:799–808.
6. Deniffel D, Abraham N, Namdar K, et al. Using decision curve analysis to benchmark performance of a magnetic resonance imaging–based deep learning model for prostate cancer risk assessment. *Eur Radiol*. 2020;30:6867–6876.
7. Syer T, Mehta P, Antonelli M, et al. Artificial intelligence compared to radiologists for the initial diagnosis of prostate cancer on magnetic resonance imaging: a systematic review and recommendations for future studies. *Cancers (Basel)*. 2021; 13:3318 Published 2021 Jul 1.
8. Mottet N, van den Bergh RCN, Briers E, et al. EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer-2020 update. Part 1: screening, diagnosis, and local treatment with curative intent. *Eur Urol*. 2021;79:243–262.
9. Acosta JN, Falcone GJ, Rajpurkar P, et al. Multimodal biomedical AI. *Nat Med*. 2022;28:1773–1784.
10. Lipkova J, Chen RJ, Chen B, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*. 2022;40:1095–1110.
11. Yap J, Yolland W, Tschandl P. Multimodal skin lesion classification using deep learning. *Exp Dermatol*. 2018;27:1261–1267.
12. Mo S, Cai M, Lin L, et al. Multimodal priors guided segmentation of liver lesions in MRI using mutual information based graph co-attention networks. *Medical Image Computing and Computer Assisted Intervention–MICCAI*. 2020;429–438.
13. 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.
14. Park KJ, Choi SH, Kim MH, et al. Performance of prostate imaging reporting and data system version 2.1 for diagnosis of prostate cancer: a systematic review and meta-analysis. *J Magn Reson Imaging*. 2021;54:103–112.
15. Bosma JS, Saha A, Hosseinzadeh M, et al. Semisupervised learning with report-guided pseudo labels for deep learning–based prostate cancer detection using biparametric MRI. *Radiol Artif Intell*. 2023;5:e230031 Published 2023 Jul 26.
16. Isensee F, Jaeger PF, Kohl SAA, et al. nnU-net: a self-configuring method for deep learning–based biomedical image segmentation. *Nat Methods*. 2021;18:203–211.
17. Bergstra J, Bardenet R, Bengio Y, et al. Algorithms for hyper-parameter optimization. *Adv Neural Inform Process Syst*. 2011;24.
18. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2019.
19. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837–845.
20. Holm S. A simple sequentially rejective multiple test procedure. *Scand J Stat*. 1979;6:65–70.
21. Quenouille MH. Notes on bias in estimation. *Biometrika*. 1956;43(3–4):353–360.
22. Roest C, Fransen SJ, Kwee TC, et al. Comparative performance of deep learning and radiologists for the diagnosis and localization of clinically significant prostate cancer at MRI: a systematic review. *Life (Basel)*. 2022;12:1490.
23. Hosseinzadeh M, Saha A, Brand P, et al. Deep learning–assisted prostate cancer detection on bi-parametric MRI: minimum training data size requirements and effect of prior knowledge. *Eur Radiol*. 2022;32:2224–2234.
24. Alves N, Bosma JS, Venkadesh KV, et al. Prediction variability to identify reduced AI performance in cancer diagnosis at MRI and CT. *Radiology*. 2023;308:e230275.
25. Shao W, Han Z, Cheng J, et al. Integrative analysis of pathological images and multi-dimensional genomic data for early-stage cancer prognosis. *IEEE Trans Med Imaging*. 2020;39:99–110.
26. Martorana E, Pirola GM, Scialpi M, et al. Lesion volume predicts prostate cancer risk and aggressiveness: validation of its value alone and matched with prostate imaging reporting and data system score. *BJU Int*. 2017;120:92–103.
27. Alis D, Yergin M, Alis C, et al. Inter-vendor performance of deep learning in segmenting acute ischemic lesions on diffusion-weighted imaging: a multicenter study. *Sci Rep*. 2021;11:12434.