Volume 3, Number 2, 2026

Pages 377- 407

Bioscience Insights

ISSN: 3080-3489 (Print)
ISSN: 3080-3497 (Online)

BiosciInsights.com

# DEEP LEARNING FOR MRI-BASED DETECTION OF CLINICALLY SIGNIFICANT PROSTATE CANCER: A CRITICAL SYSTEMATIC REVIEW OF DIAGNOSTIC PERFORMANCE, VALIDATION QUALITY, AND CLINICAL TRANSLATION

**Mohommad Riyal Kalidha Farween[*1], Aamina Hanifa[2], Hewa Dikkumburage Tharindu Madhuranga[3], Nirmani Samarakoon[4], Abid Yahya[5]**

*[*1]Department of Basic Science and Social Science for Nursing, Faculty of Nursing, University of Colombo, Colombo, Sri Lanka*

*[2,4]Department of Biomedical Science, Faculty of Health Sciences, KIU University, Colombo, Sri Lanka*

*[3]College of Veterinary Medicine and Research Institute of Veterinary Medicine, Chungnam National University, Daejeon, Republic of Korea*

*[5]Department of Electrical and Communications Systems Engineering, Botswana International University of Science and Technology, Palapye, Botswana*

Corresponding Author: kalisada98@gmail.com

**Article Information**

***Article History***

| | |
|---|---|
| Received: | May 15, 2026 |
| Revised: | June 27, 2026 |
| Accepted: | August 12, 2026 |
| Available Online: | August 31, 2026 |

**Keywords:**

*prostate cancer, multiparametric MRI, biparametric MRI, deep learning, convolutional neural network, U-Net, external validation*

**Abstract**

Prostate MRI is central to the diagnostic pathway for suspected organ-confined prostate cancer; however, interpretation remains reader-dependent and sensitive to variability in scanners, protocols, annotations, and reference standards. Deep learning models have therefore been proposed for lesion detection, segmentation, grading, biopsy support, and workflow optimization. This review critically evaluated MRI-based deep learning studies for clinically significant prostate cancer detection, focusing on diagnostic performance, validation quality, reference standards, risk of bias, explainability, and clinical readiness. A systematic search was conducted across PubMed, ScienceDirect, SpringerLink, the Cochrane Library, and Google Scholar for English-language literature published from 2012 to April 2026. Eligible studies used prostate MRI as input and reported artificial intelligence or deep learning methods for detection, segmentation, classification, risk prediction, or MRI workflow optimization. Reporting followed PRISMA 2020 and PRISMA-S. Risk of bias and applicability were assessed using domains adapted from QUADAS-2, PROBAST+AI, CLAIM 2024, and TRIPOD+AI. Thirty MRI-based model studies were included in the primary evidence synthesis. Dominant architectures included convolutional neural networks, 3D CNNs, U-Net/nnU-Net variants, Faster R-CNN-type detectors, ensemble models, and multimodal MRI-clinical systems. Reported AUROC values frequently ranged from approximately 0.80 to 0.95. However, evidence was limited by retrospective designs, inadequate external validation, inconsistent patient-level separation, heterogeneous

reference standards, incomplete calibration reporting, and limited clinical-utility analyses. MRI-based deep learning is not yet suitable for autonomous diagnosis but shows value in radiologist-supported lesion localization, triage, targeted-biopsy planning, and quality assurance. Prospective multicentre validation, transparent reporting, pathology-linked endpoints, calibration, decision-curve analysis, failure-mode reporting, and domain-shift assessment remain essential.

## INTRODUCTION

Prostate cancer remains a major global malignancy among men. GLOBOCAN 2022 estimated 1,467,854 new prostate cancer cases and 397,430 deaths worldwide, confirming that the disease burden is large but clinically heterogeneous [1]. The central diagnostic challenge is not merely to detect any cancer; it is to detect clinically significant prostate cancer (csPCa), usually defined in imaging-AI studies as ISUP Grade Group 2 or higher, while reducing overdiagnosis of indolent disease.

PSA testing and systematic biopsy have improved early detection, but PSA lacks cancer-specificity and systematic biopsy can miss MRI-visible aggressive lesions while detecting clinically insignificant disease. Contemporary pathways increasingly use MRI before biopsy for men with suspected organ-confined disease, and MRI-targeted biopsy has shown advantages over purely systematic sampling in appropriate settings [2,3]. Nevertheless, MRI interpretation remains dependent on reader expertise, PI-RADS familiarity, prostate zonal anatomy, image quality, lesion conspicuity, and institutional biopsy standards [4].

This review focuses on MRI-based deep learning evidence for csPCa detection and related diagnostic tasks. The emphasis is not that deep learning is promising in general; the emphasis is whether published models are technically valid, externally generalizable, and clinically usable.

The review addresses four questions: (i) which MRI-based deep learning architectures have been used for csPCa detection, segmentation, grading, and biopsy support; (ii) what diagnostic performance has been reported; (iii) how strong the validation and reference standards are; and (iv) what methodological gaps prevent clinical translation.

## RESEARCH METHODS

### Review design and protocol

The review was restructured as a critical systematic review of MRI-based deep learning studies. The manuscript now follows PRISMA 2020 for reporting the identification, screening, eligibility assessment, and synthesis of records, and PRISMA-S for transparent search reporting [5,6]. A prospective registration record was not available for the original work; this is explicitly acknowledged as a limitation rather than implied as completed.

**Eligibility criteria**

**Table 1. Eligibility structure used for the revised synthesis.**

| Domain | Eligibility decision |
|---|---|
| Population | Men investigated for suspected or confirmed prostate cancer. Studies focused exclusively on non-prostate cancers were excluded. |
| Index input | Prostate MRI, including mpMRI, bpMRI, T2WI, DWI/ADC, DCE, or MRI-derived image-quality/reconstruction data. Pathology-only, ultrasound-only, CT-only, PET-only, or generic image-classification studies were excluded from the primary synthesis. |
| AI method | Deep learning or AI models applied to detection, segmentation, classification, grading, risk prediction, biopsy decision support, artifact reduction, or MRI protocol optimization. |
| Reference standard | Histopathology, targeted/systematic biopsy, prostatectomy, radiologist lesion annotation, PI-RADS assessment, or clinical follow-up. The reference standard was not treated as equivalent across studies. |
| Outcomes | AUROC/AUC, sensitivity, specificity, accuracy, PPV, NPV, F1-score, Dice similarity coefficient, calibration metrics, decision-curve analysis, reader comparison, and external-validation results. |
| Exclusion | Conference abstracts without sufficient methods, non-English publications, duplicate reports, studies without MRI input, studies without an AI method, studies without extractable validation or performance information, and articles unrelated to prostate cancer. |

**Information sources and search strategy**

Database-level search strings were specified to make the review auditable. The search period was set from 1 January 2012 to 22 April 2026 because modern medical image deep learning accelerated after 2012. Google Scholar was treated as a supplementary source and screened using a predefined cut-off, because unlimited Google Scholar screening is not reproducible.

**Table 2. Database search strings for reproducible reporting.**

| Source | Search string / screening rule |
|---|---|
| PubMed | ("prostatic neoplasms"[MeSH Terms] OR "prostate cancer"[Title/Abstract] OR "clinically significant prostate cancer"[Title/Abstract]) AND ("magnetic resonance imaging"[MeSH Terms] OR MRI[Title/Abstract] OR mpMRI[Title/Abstract] OR bpMRI[Title/Abstract] OR "diffusion-weighted"[Title/Abstract] OR ADC[Title/Abstract]) AND ("deep learning"[Title/Abstract] OR "convolutional neural network"[Title/Abstract] OR CNN[Title/Abstract] OR U-Net[Title/Abstract] OR nnU-Net[Title/Abstract] OR transformer[Title/Abstract] OR "artificial intelligence"[Title/Abstract] OR "machine learning"[Title/Abstract]) AND (detection OR diagnosis OR classification OR segmentation OR grading OR biopsy OR PI-RADS) |
| ScienceDirect | ("prostate cancer" OR "clinically significant prostate cancer") AND (MRI OR mpMRI OR bpMRI OR "diffusion weighted" OR ADC) AND ("deep learning" OR CNN OR "convolutional neural network" OR U-Net OR transformer OR "artificial intelligence") AND (detection OR diagnosis OR segmentation OR classification OR grading) |
| SpringerLink | "prostate cancer" AND MRI AND ("deep learning" OR CNN OR U-Net OR "artificial intelligence") AND (detection OR segmentation OR classification OR diagnosis) |
| Cochrane Library | "prostate cancer" AND MRI AND ("artificial intelligence" OR "deep learning" OR "machine learning") |
| Google Scholar | First 300 results sorted by relevance using: "prostate cancer" "MRI" "deep learning" detection; "clinically significant prostate cancer" "CNN" "mpMRI"; "prostate MRI" "U-Net" segmentation. |

**Study selection and extraction**

Titles and abstracts were screened against the eligibility criteria, followed by full-text assessment. Duplicate records were removed before screening. Data were extracted independently by two reviewers using a standardized data extraction form developed for this review. For each eligible model study, the extraction fields were author/year, MRI sequence, clinical task, architecture, training/validation design, reference standard, patient-level or lesion-level analysis, reported metrics, external-validation status, explainability, calibration, and main clinical limitation. If the published report did not clearly state patient-level separation or external validation, the item was rated as unclear rather than assumed to be adequate. Any discrepancies between the reviewers were resolved through discussion, and when consensus could not be reached, a third reviewer was consulted. Where necessary, corresponding authors of the included studies were contacted to obtain missing or clarify unclear information. No automation tools were used during the data extraction process.

**Risk of bias, reporting quality, and applicability**

The quality framework separated diagnostic accuracy, prediction modelling, and AI imaging reporting concerns. QUADAS-2 domains were used for diagnostic accuracy studies, PROBAST+AI concepts were used for prediction/risk models, and CLAIM 2024/TRIPOD+AI domains were used to judge whether AI methods were reported sufficiently for reproducibility [7-10].

**Table 3. Risk-of-bias and reporting-quality domains.**

| Domain | What was judged |
|---|---|
| Patient selection | Was the cohort consecutive or clearly sampled? Was there spectrum bias? Were exclusions justified? |
| Index test/model | Were preprocessing, architecture, training, validation, hyperparameters, and thresholds reported? Was data leakage prevented at patient level? |
| Reference standard | Was histopathology, targeted biopsy, prostatectomy, or clinical follow-up used? Were PI-RADS or radiologist labels used as imperfect surrogates? |
| Flow and timing | Were MRI, biopsy, pathology, and model evaluation linked at patient/lesion level? Were missing data and exclusions handled transparently? |

| | |
|---|---|
| External validity | Was testing conducted on an independent institution, scanner, vendor, or time period? Was domain shift assessed? |
| Clinical utility | Were calibration, decision curves, reader studies, biopsy avoidance, workflow burden, and failure cases reported? |
| Transparency | Were code, model weights, preprocessing rules, annotation protocol, and full metrics available? |

**Synthesis approach**

A meta-analysis was not performed because the included studies differed materially in clinical endpoint, MRI protocol, unit of analysis, reference standard, dataset size, threshold selection, internal/external validation, and performance metrics. Instead, a structured qualitative synthesis was performed by clinical task and architecture. The synthesis gives more weight to studies with independent external validation, pathology-linked endpoints, patient-level separation, and clinical utility analysis.

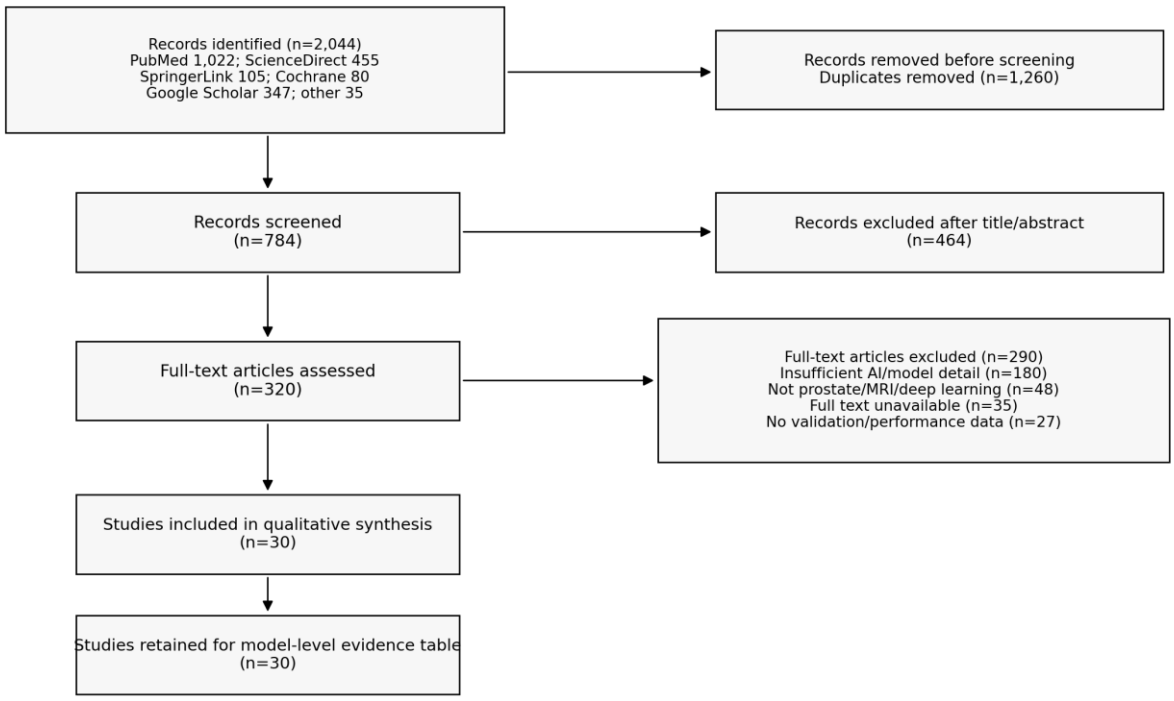



***Figure 1. PRISMA 2020 flow diagram for the review. The final synthesis retains 30 MRI-based deep learning studies with extractable model-level information.***

## Clinical and imaging context

### Disease burden and diagnostic target

The diagnostic target for this review is csPCa rather than any histological prostate cancer. This distinction is essential because model performance can appear inflated when indolent lesions are grouped with aggressive disease or when slice-level labels are treated as independent samples. Models intended for clinical decision support should be evaluated on patient-level or lesion-level endpoints that matter for biopsy and treatment decisions.

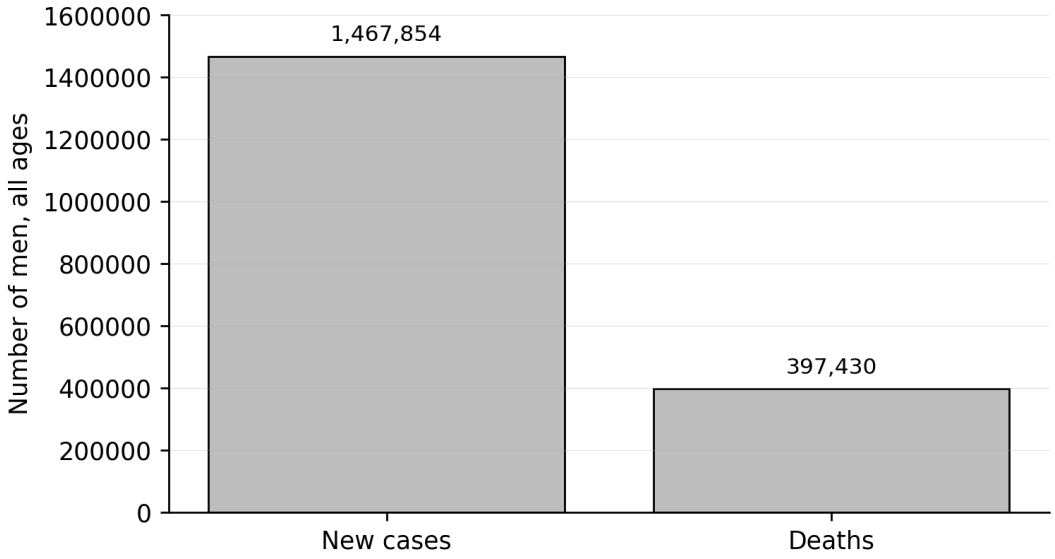



***Figure 2. Global prostate cancer burden based on GLOBOCAN 2022 estimates: 1,467,854 new cases and 397,430 deaths among men worldwide [1].***

### MRI protocols and AI relevance

Only MRI content needed for model interpretation is retained here. Prostate AI models are strongly affected by sequence selection because T2WI, DWI/ADC, DCE, and bpMRI encode different biological and acquisition-dependent information.

**Table 4. MRI sequences and their relevance to deep learning models for prostate cancer.**

| Sequence | Information captured | AI relevance | Main limitation |
|---|---|---|---|
| T2WI | Zonal anatomy, capsule, transition-zone morphology, lesion shape | Supports gland segmentation, zonal localization, capsule-contact assessment, and anatomical context learning | Limited specificity for aggressiveness; benign prostatic hyperplasia can mimic lesions |

| | | | |
|---|---|---|---|
| DWI | Diffusion restriction and cellularity at multiple b-values | Often the most informative sequence for csPCa detection and lesion conspicuity | Sensitive to distortion, motion, b-value protocol, and scanner differences |
| ADC | Quantitative diffusion map derived from DWI | Useful for csPCa classification, lesion grading, and intensity-based features | ADC values are not fully standardized across vendors, field strengths, and protocols |
| DCE | Contrast enhancement kinetics and vascularity | May improve peripheral-zone lesion characterization and equivocal cases | Requires contrast, adds time/cost, and introduces protocol variability |
| bpMRI | T2WI plus DWI/ADC without DCE | Reduces acquisition time and computational complexity; increasingly used for AI triage | May lose complementary functional information in selected cases |
| mpMRI | Integrated anatomical and functional protocol | Most common input for DL detection, segmentation, and PI-RADS comparison | Longer acquisition, sequence registration issues, and larger domain shift across institutions |

**Preprocessing and leakage control**

Preprocessing choices are not neutral technical details. Intensity normalization, prostate cropping, multimodal registration, denoising, resampling, and augmentation can improve stability, but they can also create leakage or distort lesion appearance. Augmentation increases the number of images, not the number of independent patients. Therefore, all splits must be performed at patient level before augmentation. Slice-level splitting is unacceptable for performance claims because adjacent MRI slices from the same lesion are highly correlated.

**Table 5. Minimum technical controls expected in prostate MRI deep learning studies.**

| Control | Reason |
|---|---|
| Patient-level split before augmentation | Prevents the same patient from contributing augmented images to both training and testing. |
| Protocol-specific normalization | Reduces intensity variation but must be reported sufficiently for replication. |
| Registration of sequences | Improves spatial alignment but may introduce interpolation artifacts near small lesions. |
| Lesion/prostate segmentation masks | Useful for localization and radiologist workflow, but annotation variability must be measured. |
| External testing by institution/vendor/time | Best practical test of generalizability and domain shift. |
| Calibration and decision analysis | Required when model outputs are used as risk probabilities or biopsy decision tools. |

**Deep learning model taxonomy**

The revised taxonomy separates architectures by the way they handle prostate anatomy, volumetric context, labels, and clinical data. This prevents the manuscript from treating all AI methods as interchangeable.

**Table 6. Architecture-level comparison for MRI-based prostate cancer AI.**

| Architecture | Typical input | Main strength | Main limitation |
|---|---|---|---|
| Radiomics + classical ML | Handcrafted intensity, texture, shape, and ADC features with SVM/RF/logistic models | Small datasets and interpretable features | Depends on manual feature design; limited representation learning |
| 2D CNN | MRI slices or patches | Lower memory demand and easier training | Loses through-plane context; vulnerable to slice-level leakage |
| 2.5D CNN | Adjacent slices or orthogonal views | Balances spatial context and computational cost | Still approximates 3D anatomy; fusion strategy must be clear |

| | | | |
|---|---|---|---|
| 3D CNN | Volumetric prostate or lesion ROI | Captures anatomical continuity and lesion volume | Requires larger annotated datasets and careful memory management |
| U-Net / nnU-Net | Voxel-level gland or lesion masks | Best suited for segmentation and localization | Performance depends strongly on annotation quality and inter-reader agreement |
| Detector networks | Candidate lesion detection and localization | Clinically intuitive output for radiologist review | False positives increase in prostatitis, artifacts, and transition-zone nodules |
| Transformer-based models | Patch/token representations and long-range context | May learn broader spatial relationships | Data hungry; often requires pretraining and external validation |
| Multimodal MRI-clinical models | MRI plus PSA, prostate volume, age, PI-RADS, biopsy history, pathology | Patient-level risk stratification and biopsy support | Harder to interpret; missing clinical variables can limit deployment |

**Detection and classification models**

Detection/classification models estimate whether csPCa is present at slice, lesion, prostate, or patient level. The unit of analysis must be explicitly reported. A slice-level AUROC cannot be interpreted as patient-level diagnostic accuracy. Studies that report very high accuracy without external testing, threshold specification, confidence intervals, and reference-standard clarity should be treated cautiously.

**Segmentation models**

Segmentation models delineate the prostate gland, zones, or suspicious lesions. Dice similarity coefficient is useful but incomplete: a high gland Dice score does not prove accurate csPCa localization, and a lesion Dice score can be unstable for small lesions. Segmentation studies should report lesion-level sensitivity, false-positive burden, inter-reader variability, and performance by zone.

### Grading and aggressiveness prediction

MRI-based grading is more difficult than detection because Gleason/ISUP grade depends on microscopic architecture that may not be fully visible on MRI. Grading models require strong pathology linkage, ideally prostatectomy-based mapping or carefully registered targeted-biopsy labels. Models trained only on radiologist labels or PI-RADS categories should not be presented as histological grading models.

### Multimodal risk prediction

Multimodal models that combine MRI with PSA, prostate volume, PI-RADS, biopsy history, age, and pathology can be clinically useful if they are calibrated and externally validated. However, adding clinical variables does not automatically improve deployability. Missing data, variable definitions, and local biopsy practice can all reduce transportability.

**Algorithm 1. Minimum acceptable validation workflow for prostate MRI AI studies**

| Step | Requirement |
|---|---|
| 1 | Define the diagnostic endpoint before training: csPCa, lesion segmentation, grade group, biopsy avoidance, or workflow optimization. |
| 2 | Lock patient-level train/validation/test separation before augmentation, patch extraction, or slice generation. |
| 3 | Specify MRI sequences, field strength, vendors, b-values, preprocessing, registration, and intensity normalization. |
| 4 | Use an appropriate reference standard and state whether labels are patient-, lesion-, or slice-level. |
| 5 | Report AUROC with confidence intervals plus sensitivity, specificity, PPV, NPV, and threshold selection. |
| 6 | For segmentation, report Dice plus lesion-level detection, false positives per case, and performance by prostate zone. |
| 7 | Evaluate calibration using calibration curves, Brier score, or expected calibration error when probabilities are presented. |
| 8 | Test the final locked model on an independent institution, vendor, or temporal cohort. |
| 9 | Report failure cases, domain-shift effects, and whether performance changes by scanner, sequence quality, and patient subgroup. |

| | |
|---|---|
| 10 | If intended for biopsy decisions, provide decision-curve analysis or a reader study that reflects clinical workflow. |

## RESULTS

### Study selection

The search identified 2,044 records. After removal of duplicates, 784 records were screened and 320 full-text articles were assessed. Applying model-level eligibility criteria retained 30 MRI-based deep learning studies for the primary synthesis.

### Study characteristics

The retained studies covered csPCa detection, lesion classification, prostate/lesion segmentation, Gleason or aggressiveness assessment, false-positive reduction, MRI image reconstruction, artifact reduction, and biopsy decision support. Most studies were retrospective. A smaller number used multicenter or external validation. Public datasets, single-center clinical datasets, and institutional retrospective cohorts were common; however, dataset reporting was often incomplete, particularly for test-set size, lesion-level labels, and confidence intervals.

### Diagnostic performance and evidence strength

Reported performance was frequently high in internal validation, with AUROC values often between approximately 0.80 and 0.95. This range should not be interpreted as settled clinical accuracy. Performance was strongly dependent on the validation design, reference standard, class balance, MRI sequence quality, and whether the unit of analysis was slice, lesion, prostate, or patient. The evidence is strongest for models evaluated on external or multicenter cohorts and weakest for single-center retrospective models using unclear split methods or surrogate labels.

### Segmentation and localization

U-Net and 3D CNN-based models showed practical value for prostate and lesion localization. Their clinical value lies in workflow support: drawing attention to suspicious regions, assisting targeted biopsy planning, and reducing reader variability. However, segmentation accuracy must be interpreted against lesion size and annotation consistency. Transition-zone lesions, prostatitis, benign hyperplasia, motion artifact, and low-contrast lesions remain common causes of false-positive or false-negative output.

### Comparison with PI-RADS and radiologists

Several studies compared deep learning output with PI-RADS or radiologist assessment. The credible interpretation is not that AI replaces radiologists; rather, AI may reduce variability,

improve triage, or increase consistency in selected tasks. Reader studies remain limited, and most published comparisons do not fully account for real-time reporting pressure, medicolegal responsibility, biopsy availability, or local practice patterns.

**External validation and domain shift**

External validation was the central weakness of the evidence base. Models trained on one center can degrade when tested on another due to scanner vendor, magnetic field strength, coil configuration, b-values, reconstruction software, annotation style, patient demographics, and disease prevalence. Same-vendor multicenter testing is useful but still weaker than independent multi-vendor, multi-region validation.

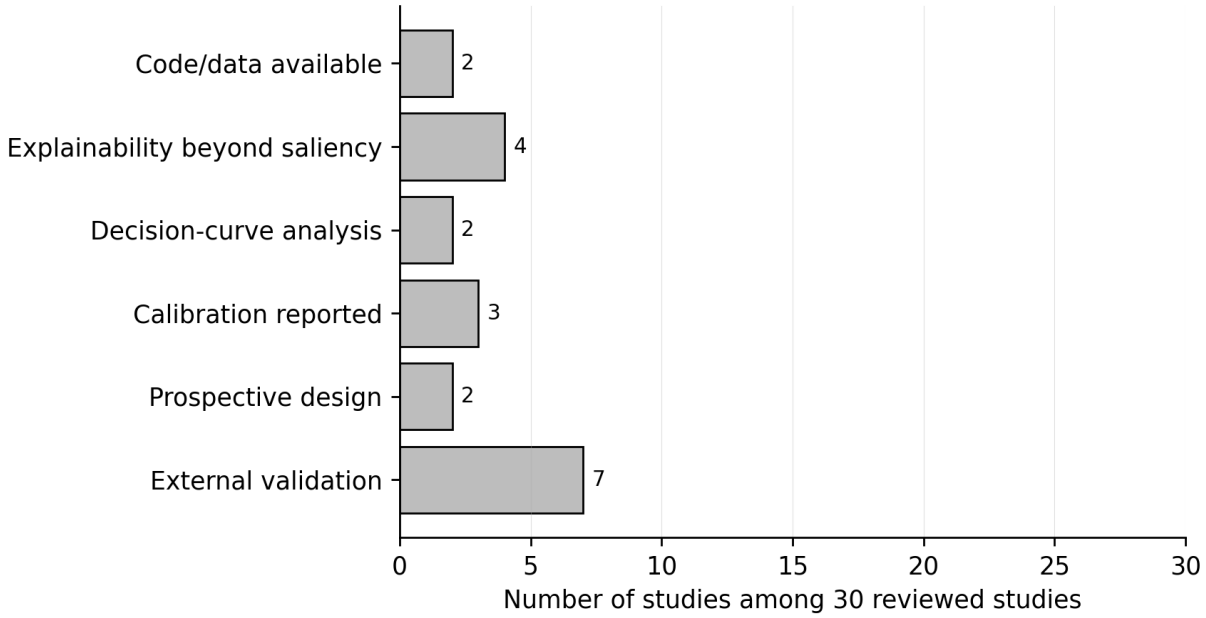



***Figure 3. Validation maturity indicators among the 30 studies retained in the evidence synthesis.***

**Risk of bias and reporting quality**

The main bias concerns were not the choice of neural network alone; they were dataset construction, labeling, splitting, and reporting. Common issues included small or homogeneous datasets, absent external validation, unclear patient-level separation, inconsistent reference standards, no calibration, no decision-curve analysis, and limited code/model availability. Studies that used radiologist labels or PI-RADS as a ground truth were judged less robust than pathology-linked studies for csPCa detection.

**Table 7. Summary of major risk-of-bias concerns in the evidence synthesis.**

| Domain | Judgment from the revised review |
|---|---|
| Patient selection | Moderate to high concern in many studies because cohorts were retrospective, single-center, or enriched with suspicious lesions. |
| Index test/model | Moderate concern because many reports lacked full preprocessing, hyperparameter, threshold, and patient-level split details. |
| Reference standard | Variable concern; pathology-linked labels were stronger than PI-RADS or radiologist-only labels. Biopsy labels may still suffer from sampling mismatch. |
| Flow and timing | Often unclear because lesion-to-pathology linkage and timing between MRI and biopsy/prostatectomy were inconsistently reported. |
| External validation | High concern for most studies; only a minority used independent external or multicenter test sets. |
| Clinical utility | High concern because calibration, decision-curve analysis, reader workflow, and biopsy avoidance were rarely assessed together. |
| Explainability | Limited in most studies; Grad-CAM/saliency alone was treated as insufficient clinical explanation. |

**Clinical translation**

The most defensible clinical role for MRI-based deep learning is decision support, not autonomous diagnosis. A clinically useful system should present lesion-level evidence, uncertainty, and an interpretable relationship to PI-RADS, PSA, prostate volume, and biopsy history. It should also explain when it should not be trusted: poor DWI quality, severe rectal artifact, post-biopsy hemorrhage, atypical anatomy, under-represented scanners, and image protocols outside the training domain.

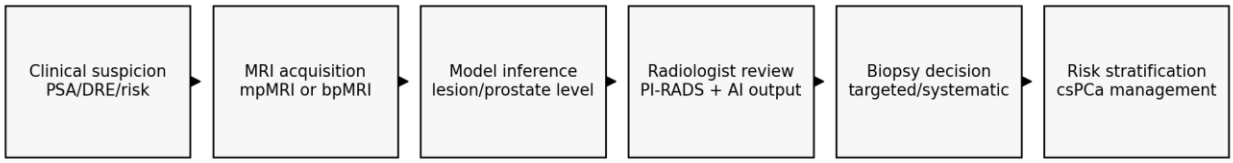



***Figure 4. Proposed clinical positioning of MRI-based deep learning as a support layer within the prostate cancer diagnostic pathway.***

**Biopsy decision support**

Biopsy-support models should be evaluated with decision-curve analysis, missed csPCa rate, avoided biopsy rate, false-positive burden, and reader interaction. AUROC alone is insufficient because biopsy decisions occur at clinically selected thresholds. A model that improves AUROC but increases unnecessary biopsy or misses anterior high-grade lesions is not clinically acceptable.

**Explainability and uncertainty**

Saliency maps, Grad-CAM, and LIME can help identify image regions associated with model output, but they do not prove biological correctness. A reliable clinical explanation should combine localization maps, probability calibration, uncertainty estimation, error analysis, and a failure-case library. For risk models, calibration curves, Brier score, expected calibration error, and decision curves are more clinically relevant than decorative heatmaps alone.

**Regulatory, workflow, and equity concerns**

Deployment requires integration with PACS/RIS, clear responsibility between radiologist and software output, model monitoring after implementation, and documentation of performance across scanners and patient subgroups. Dataset representativeness is especially important because global prostate cancer incidence, MRI access, and diagnostic pathways vary across regions. Privacy-preserving multicenter strategies, including federated learning, may support broader validation without central transfer of raw MRI data, but they do not remove the need for standardized labels and prospective testing.

**DISCUSSION**

The evidence does not support a broad claim that deep learning has already solved MRI-based prostate cancer diagnosis. It supports a narrower claim: deep learning can assist prostate MRI

interpretation under controlled conditions, especially for lesion detection, segmentation, and risk stratification, but clinical reliability depends on external validation, reference-standard quality, and deployment-specific testing.

A recurring weakness is that many studies report internal performance without demonstrating transportability. For medical imaging AI, this is a serious limitation because MRI is not a standardized numeric measurement. It is a vendor-, scanner-, protocol-, and patient-dependent signal. A model trained on one distribution may learn local imaging signatures instead of prostate cancer biology. This risk is amplified by slice-level training, augmentation before splitting, inconsistent prostate/lesion annotation, and outcome labels derived from imperfect biopsy sampling.

Another problem is the inconsistent use of reference standards. Prostatectomy pathology is stronger than targeted biopsy for spatial ground truth, but it reflects a selected surgical population. Biopsy is clinically relevant but subject to sampling error. PI-RADS and radiologist labels are useful for workflow studies but are not equivalent to histopathology. These differences explain why performance values should not be pooled casually across studies.

The review also identifies a reporting problem. Many studies describe architectures but under-report threshold choice, calibration, missing data handling, scanner distribution, lesion matching, and confidence intervals. CLAIM 2024, TRIPOD+AI, and PROBAST+AI should be used not as decorative citations but as working reporting standards. A future journal submission should include a completed checklist as supplementary material.

**Limitations of this review**

This review has limitations. A registered protocol was not available. The synthesis is qualitative rather than meta-analytic because of heterogeneous endpoints, MRI protocols, model types, validation units, and reference standards. English-language restriction may also introduce language bias.

Future research and reporting priorities

- Define the target endpoint clearly and avoid claiming more than the evidence supports.
- Deposit search strings, screening decisions, and extraction sheets as supplementary material.
- Keep the included-study number consistent across the abstract, methods, PRISMA figure, results, and evidence table.
- Provide study-level risk-of-bias judgments using formal domains and make them available in supplementary material.

- Report model performance by validation type: internal, temporal, single-center external, multicenter external, and prospective.
- Do not compare AUROC values across studies unless endpoint, unit of analysis, reference standard, and validation design are comparable.
- Remove unrelated references, especially non-MRI, non-prostate, general image-classification, and NLP references unless explicitly used for methodology background.
- Use only original figures with captions below the figure. Do not place titles inside figures.
- Use precise language: “radiologist support”, “triage”, “lesion localization”, and “biopsy decision support” are more defensible than “clinical replacement” or “revolutionary performance”.

## CONCLUSIONS

Deep learning models for prostate MRI show useful technical capability, particularly for lesion localization, segmentation, csPCa classification, and biopsy-support workflows. However, the evidence base remains limited by retrospective design, small or homogeneous cohorts, heterogeneous MRI protocols, inconsistent reference standards, incomplete reporting, weak calibration analysis, and insufficient external validation. Current models should be positioned as radiologist-support tools rather than autonomous diagnostic systems. The next stage of the field must shift from internal accuracy reporting to prospective multicenter evaluation, transparent methodology, calibrated risk prediction, clinical utility analysis, and rigorous monitoring under real deployment conditions.

## DECLARATIONS

### ETHICS APPROVAL AND CONSENT TO PARTICIPATE

Not applicable. This review used previously published literature and did not involve direct human participant recruitment or individual-level patient data.

### CONSENT FOR PUBLICATION

Not applicable.

### COMPETING INTERESTS

The authors declare no competing interests.

### FUNDING

No specific funding was declared for this review.

## AUTHOR CONTRIBUTIONS

KF contributed to literature screening, extraction, data interpretation, and manuscript drafting. AH, TM, and NS contributed to data interpretation and manuscript preparation. AY supervised the study, provided technical direction, and critically revised the manuscript for methodological, AI, and clinical-translation content.

## ACKNOWLEDGEMENTS

None declared.

## DATA AVAILABILITY

The datasets analysed during the current study were derived from publicly available published studies, all of which are cited within this article. No additional datasets were generated during the current study.

**Abbreviations**

| Abbreviation | Meaning |
|---|---|
| ADC | Apparent diffusion coefficient |
| AI | Artificial intelligence |
| AUC/AUROC | Area under the receiver operating characteristic curve |
| bpMRI | Biparametric magnetic resonance imaging |
| CLAIM | Checklist for Artificial Intelligence in Medical Imaging |
| CNN | Convolutional neural network |
| csPCa | Clinically significant prostate cancer |
| DCE | Dynamic contrast-enhanced imaging |
| DL | Deep learning |
| DRE | Digital rectal examination |
| DWI | Diffusion-weighted imaging |
| ISUP | International Society of Urological Pathology |
| MRI | Magnetic resonance imaging |
| mpMRI | Multiparametric magnetic resonance imaging |
| PI-RADS | Prostate Imaging Reporting and Data System |
| PPV/NPV | Positive/negative predictive value |
| PRISMA | Preferred Reporting Items for Systematic Reviews and Meta-Analyses |

| | |
|---|---|
| PROBAST+AI | Prediction model Risk Of Bias ASsessment Tool extension for artificial intelligence |
| PSA | Prostate-specific antigen |
| QUADAS-2 | Quality Assessment of Diagnostic Accuracy Studies 2 |
| TRIPOD+AI | Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis + AI |
| XAI | Explainable artificial intelligence |

**REFERENCES**

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A, Bray F. 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. doi:10.3322/caac.21834.
2. Kasivisvanathan V, Rannikko AS, Borghi M, Panebianco V, Mynderse LA, Vaarala MH, et al. MRI-targeted or standard biopsy for prostate-cancer diagnosis. N Engl J Med. 2018;378:1767-1777. doi:10.1056/NEJMoa1801993.
3. Rouviere O, Puech P, Renard-Penna R, Claudon M, Roy C, Mege-Lechevallier F, et al. Use of prostate systematic and targeted biopsy on the basis of multiparametric MRI in biopsy-naive patients: the MRI-FIRST study. Lancet Oncol. 2019;20:100-109. doi:10.1016/S1470-2045(18)30569-2.
4. Turkbey B, Rosenkrantz AB, Haider MA, Padhani AR, Villeirs G, Macura KJ, et al. Prostate Imaging Reporting and Data System version 2.1: 2019 update. Eur Urol. 2019;76:340-351. doi:10.1016/j.eururo.2019.02.033.
5. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.
6. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, Koffel JB. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev. 2021;10:39. doi:10.1186/s13643-020-01542-z.
7. Whiting PF, Rutjes AWS, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155:529-536. doi:10.7326/0003-4819-155-8-201110180-00009.

8. Moons KGM, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. doi:10.1136/bmj-2024-082505.
9. Tejani AS, et al. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiol Artif Intell. 2024;6:e240300. doi:10.1148/ryai.240300.
10. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. doi:10.1136/bmj-2023-078378.
11. Pellicer-Valero OJ, Marenco Jimenez JL, Gonzalez-Perez V, Casanova Ramon-Borja JL, Martin Garcia I, Barrios Benito M, et al. Deep learning for fully automatic detection, segmentation, and Gleason grade estimation of prostate cancer in multiparametric MRI. Sci Rep. 2022;12:2975. doi:10.1038/s41598-022-06730-6.
12. Liu S, Zheng H, Feng Y, Li W. Prostate cancer diagnosis using deep learning with 3D multiparametric MRI. Proc SPIE Med Imaging. 2017;10134:1013428. doi:10.1117/12.2277121.
13. Kaya M, Durdag O, Solak M, Coskuncay A, Burakgazi G. Lesion detection using artificial intelligence models in MR images of prostate cancer and prostatitis patients and comparison of model performance. Front Urol. 2026;5:1726795. doi:10.3389/fruro.2025.1726795.
14. Fransen SJ, Roest C, van Lohuizen QY, Bosma JS, Simonis FFJ, Kwee TC, et al. Using deep learning to optimize the prostate MRI protocol by assessing the diagnostic efficacy of MRI sequences. Eur J Radiol. 2024;175:111470. doi:10.1016/j.ejrad.2024.111470.
15. Zhao Y, Xie XL, Zhu X, Huang WN, Zhou CW, Ren KX, et al. FOCUS-DWI improves prostate cancer detection through deep learning reconstruction with IQMR technology. Abdom Radiol. 2025;51:1276-1288. doi:10.1007/s00261-025-05100-w.
16. Roest C, Kwee TC, de Jong IJ, Schoots IG, van Leeuwen P, Heijmink SWTPJ, et al. Development and validation of a deep learning model based on MRI and clinical characteristics to predict risk of prostate cancer progression. Radiol Cancer. 2025;7:e240078. doi:10.1148/rycan.240078.
17. Lee YJ, Moon HW, Choi MH, Jung SE, Park YH, Lee JY, et al. MRI-based deep learning algorithm for assisting clinically significant prostate cancer detection: a bicenter prospective study. Radiology. 2025;314:e232788. doi:10.1148/radiol.232788.

18. Wang Y, Xin Y, Zhang B, Pan F, Li X, Zhang M, et al. Assessment of prostate cancer aggressiveness through the combined analysis of prostate MRI and 2.5D deep learning models. Front Oncol. 2025;15:1539537. doi:10.3389/fonc.2025.1539537.
19. Litjens G, Debats O, Barentsz J, Karssemeijer N, Huisman H. Computer-aided detection of prostate cancer in MRI. IEEE Trans Med Imaging. 2014;33:1083-1092. doi:10.1109/TMI.2014.2303821.
20. Singh SK, Sinha A, Singh H, Mahanti A, Patel A, Mahajan S, et al. A novel deep learning-based technique for detecting prostate cancer in MRI images. Multimed Tools Appl. 2023. doi:10.1007/s11042-023-15793-0.
21. Wang Z, Liu C, Cheng D, Wang L, Yang X, Cheng KT. Automated detection of clinically significant prostate cancer in mp-MRI images based on an end-to-end deep neural network. IEEE Trans Med Imaging. 2018;37:1127-1139. doi:10.1109/TMI.2017.2789181.
22. Schelb P, Kohl S, Radtke JP, Wiesenfarth M, Kickingereder P, Bickelhaupt S, et al. Classification of cancer at prostate MRI: deep learning versus clinical PI-RADS assessment. Radiology. 2019;293:607-617. doi:10.1148/radiol.2019190938.
23. Bhattacharya I, Seetharaman A, Kunder C, Shao W, Chen LC, Soerensen SJC, 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. doi:10.1016/j.media.2021.102288.
24. Matsuoka Y, Ueno Y, Uehara S, Tanaka H, Kobayashi M, Tanaka H, et al. Deep-learning prostate cancer detection and segmentation on biparametric versus multiparametric MRI: added value of dynamic contrast-enhanced imaging. Int J Urol. 2023;30:1103-1111. doi:10.1111/iju.15280.
25. Hu L, Guo X, Zhou D, Wang Z, Dai L, Li L, et al. Development and validation of a deep learning model to reduce the interference of rectal artifacts in MRI-based prostate cancer diagnosis. Radiol Artif Intell. 2024;6:e230362. doi:10.1148/ryai.230362.
26. Cai JC, Nakai H, Kuanar S, Froemming AT, Bolan CW, Kawashima A, et al. Fully automated deep learning model to detect clinically significant prostate cancer at MRI. Radiology. 2024;312:e232635. doi:10.1148/radiol.232635.
27. Wu H, Liu F, Yang Q, Chen H, Wang Y, Yang X, et al. Automated MRI system for clinically significant prostate cancer detection: development, validation, and real-world implementation. Nat Commun. 2025;16:11583. doi:10.1038/s41467-025-66593-z.

28. Alejandro J, Mora-Rubio A, Peran-Teruel M, Navarro Beltran A, Torres CR, et al. Clinically significant prostate cancer detection with deep learning in a multicenter magnetic resonance imaging study. Sci Rep. 2026;16:10976. doi:10.1038/s41598-026-42214-7.

29. Deniffel D, Abraham N, Namdar K, Dong X, Salinas E, Milot L, et al. Using decision curve analysis to benchmark performance of an MRI-based deep learning model for prostate cancer risk assessment. Eur Radiol. 2020;30:6867-6876. doi:10.1007/s00330-020-07030-1.

30. Yang C, Li B, Luan Y, Wang S, Bian Y, Zhang J, et al. Deep learning model for the detection of prostate cancer and classification of clinically significant disease using multiparametric MRI in comparison to PI-RADS score. Urol Oncol. 2024;42:158.e17-158.e27. doi:10.1016/j.urolonc.2024.01.021.

31. Netzer N, Eith C, Bethge O, Hielscher T, Schwab C, Stenzinger A, et al. Application of a validated prostate MRI deep learning system to independent same-vendor multi-institutional data: demonstration of transferability. Eur Radiol. 2023;33:7463-7476. doi:10.1007/s00330-023-09882-9.

32. Hosseinzadeh M, Saha A, Brand P, Slootweg I, de Rooij M, Huisman H. Deep learning-assisted prostate cancer detection on biparametric MRI: minimum training data size requirements and effect of prior knowledge. Eur Radiol. 2021/2022;32:2224-2234. doi:10.1007/s00330-021-08320-y.

33. Sherafatmandjoo H, Safaei AA, Ghaderi F, Allameh F. Prostate cancer diagnosis based on multiparametric MRI, clinical and pathological factors using deep learning. Sci Rep. 2024;14:14951. doi:10.1038/s41598-024-65354-0.

34. Zheng H, Hung ALY, Miao Q, Song W, Scalzo F, Raman SS, et al. AtPCa-Net: anatomical-aware prostate cancer detection network on multiparametric MRI. Sci Rep. 2024;14:5740. doi:10.1038/s41598-024-56405-7.

35. Song Y, Domingo MR, Conlin CC, Do DD, Baxter MT, Dornisch A, et al. Deep learning artificial intelligence and restriction spectrum imaging for patient-level detection of clinically significant prostate cancer on biparametric MRI. Eur Urol Open Sci. 2026;85:49-59. doi:10.1016/j.euros.2026.01.014.

36. Yu X, Lou B, Shi B, Winkel D, Arrahmane N, Diallo M, et al. False positive reduction using multiscale contextual features for prostate cancer detection in multiparametric MRI scans. Proc IEEE ISBI. 2020:1355-1359. doi:10.1109/ISBI45749.2020.9098338.

37. Aldoj N, Lukas S, Dewey M, Penzkofer T. Semi-automatic classification of prostate cancer on multiparametric MR imaging using a multi-channel 3D convolutional neural network. Eur Radiol. 2020;30:1243-1253. doi:10.1007/s00330-019-06417-z.

38. Ishioka J, Matsuoka Y, Uehara S, Yasuda Y, Kijima T, Yoshida S, et al. Computer-aided diagnosis of prostate cancer on MRI using a convolutional neural network algorithm. BJU Int. 2018;122:411-417. doi:10.1111/bju.14397.

39. Yang X, Liu C, Wang Z, Yang J, Min HL, Wang L, et al. Co-trained convolutional neural networks for automated detection of prostate cancer in multiparametric MRI. Med Image Anal. 2017;42:212-227. doi:10.1016/j.media.2017.08.006.

40. Schrader A, Netzer N, Hielscher T, Gortz M, Zhang KS, Schutz V, et al. Prostate cancer risk assessment and avoidance of prostate biopsies using fully automatic deep learning in prostate MRI: comparison to PI-RADS and integration with clinical data in nomograms. Eur Radiol. 2024;34:7909-7920. doi:10.1007/s00330-024-10818-0.

41. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17:195. doi:10.1186/s12916-019-1426-2.

42. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag. 2020;37:50-60. doi:10.1109/MSP.2020.2975749.

43. World Health Organization. Ethics and governance of artificial intelligence for health: WHO guidance. Geneva: World Health Organization; 2021.

44. Masoudi S, Harmon SA, Mehralivand S, Walker SM, Raviprakash H, Bagci U, et al. Quick guide on radiology image preprocessing for deep learning applications in prostate cancer research. J Med Imaging. 2021;8:010901. doi:10.1117/1.JMI.8.1.010901.

**Appendix A. Core study extraction matrix**

This matrix includes MRI-based AI/deep learning studies with extractable model-level information. It is organized by study, MRI input, task, model family, validation design, and critical interpretation.

**Table A1. Core MRI-based AI studies retained for the evidence synthesis.**

| Study | MRI input | Task | Model | Validation | Main critical interpretation |
|---|---|---|---|---|---|
| Pellicer-Valero et al., 2022 [11] | mpMRI: T2WI/DWI/ADC | Detection, segmentation, Gleason estimation | 3D U-Net/CNN | Retrospective clinical validation | High internal AUC; no clear external validation; pathology linkage important. |
| Liu et al., 2017 [12] | PROSTATEx mpMRI | Lesion classification | 3D CNN/XmasNet | Public dataset validation | Small public dataset; risk of overfitting and limited external generalizability. |
| Kaya et al., 2025/2026 [13] | Clinical mpMRI | Lesion detection/localization | Faster R-CNN + CNN | Retrospective validation | Very high reported performance requires independent prospective confirmation. |

| | | | | | |
|---|---|---|---|---|---|
| Fransen et al., 2024 [14] | T2WI protocol data | MRI sequence optimization | Deep reconstruction CNN | Experimental/clinical validation | Workflow benefit; not a direct csPCa diagnostic endpoint. |
| Zhao et al., 2025 [15] | FOCUS-DWI | Image enhancement/differentiation | IQMR deep reconstruction | Retrospective validation | Needs scanner- and vendor-diverse external testing. |
| Roest et al., 2025 [16] | MRI + clinical data | Progression prediction | Multimodal deep learning | Development/validation study | Useful patient-level direction; prospective decision impact remains unclear. |
| Lee et al., 2025 [17] | Bicenter mpMRI | csPCa detection assistance | MRI-based DL algorithm | Bicenter prospective study | Stronger design than most studies; wider institutional diversity still needed. |
| Wang et al., 2025 [18] | mpMRI | Aggressiveness classification | 2.5D deep learning | Retrospective validation | MRI-grade relationship requires strong pathology alignment. |
| Litjens et al., 2014 [19] | Prostate MRI | Computer-aided detection | CAD/radiomics ML | Algorithmic validation | Important baseline; predates modern DL |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | and external validation expectations. |
| Singh et al., 2023 [20] | mpMRI | Cancer classification | Hybrid deep learning | Algorithmic validation | Limited external validation and potential dataset-specific performance. |
| Wang et al., 2018 [21] | T2WI/DWI/ADC | csPCa detection | End-to-end CNN | Retrospective validation | Small cohort and limited external testing reduce clinical certainty. |
| Schelb et al., 2019 [22] | T2WI + diffusion MRI | Cancer classification vs PI-RADS | CNN | Retrospective comparison | Useful radiologist comparison; workflow effects not fully tested. |
| Bhattacharya et al., 2022 [23] | MRI-pathology dataset | Tumor localization/aggressiveness | CorrSigNIA CNN + registration | Experimental validation | Registration and pathology alignment are strengths but also potential error sources. |
| Matsuoka et al., 2023 [24] | bpMRI/mpMRI/DCE | Detection and segmentation | DNN | Retrospective validation | Protocol dependence and DCE contribution require generalization testing. |

| | | | | | |
|---|---|---|---|---|---|
| Hu et al., 2024 [25] | Artifact-degraded MRI | Artifact-robust diagnosis | CNN with adversarial training | Retrospective validation | Clinically relevant; external artifact diversity remains limited. |
| Cai et al., 2024 [26] | mpMRI | Fully automated csPCa detection | Deep learning model | Clinical retrospective validation | Reported AUC approximately 0.86-0.89; broader external validation needed. |
| Wu et al., 2025 [27] | Large bpMRI cohort | csPCa detection and implementation | Automated CNN system | Large-scale validation | Large dataset strengthens evidence; explainability and subgroup monitoring remain concerns. |
| Alejandro et al., 2026 [28] | Multicenter mpMRI | Segmentation/classification | nnU-Net + EfficientNet | Multicenter validation | External evidence useful; scanner and protocol subgroup performance should be detailed. |
| Deniffel et al., 2020 [29] | MRI risk dataset | Biopsy risk assessment | 3D DL risk model | Retrospective validation + DCA | Decision-curve analysis is a strength; |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | prospective impact not established. |
| Yang et al., 2024 [30] | mpMRI | Detection/csPCa classification | Ensemble DL | Retrospective validation | Comparison with PI-RADS useful; multicenter testing limited. |
| Netzer et al., 2023 [31] | Independent same-vendor mpMRI | Transferability detection/segmentation | Validated automatic DL system | External same-vendor validation | Shows transportability but also domain-shift sensitivity. |
| Hosseinzadeh et al., 2021 [32] | bpMRI | Detection and training-size analysis | DL with prior knowledge | Algorithmic validation | Helpful for data-efficiency question; limited by training data availability. |
| Sherafatmandjoo et al., 2024 [33] | MRI + clinical/pathology data | Classification | Multi-input ResNet50 | Retrospective validation | High reported AUC; needs prospective and external confirmation. |
| Zheng et al., 2024 [34] | Anatomical mpMRI | Detection/classification | AtPCa-Net | Algorithmic validation | Anatomical awareness is useful; clinical deployment testing limited. |

| | | | | | |
|---|---|---|---|---|---|
| Song et al., 2026 [35] | bpMRI + restriction spectrum imaging | Patient-level csPCa detection | Deep learning AI model | Clinical diagnostic validation | Patient-level endpoint is valuable; external validation still required. |
| Yu et al., 2020 [36] | mpMRI | False-positive reduction | Multiscale CNN | Algorithmic/conference validation | Addresses a real failure mode; limited clinical validation. |
| Aldoj et al., 2019/2020 [37] | mpMRI | Lesion classification | Multi-channel 3D CNN | Retrospective validation | Volumetric approach useful; external validation limited. |
| Ishioka et al., 2018 [38] | MRI | Computer-aided diagnosis | CNN CAD algorithm | Retrospective clinical validation | Early evidence; small scale and limited reporting by current standards. |
| Yang et al., 2017 [39] | mpMRI | Automated detection | Co-trained CNNs | Algorithmic validation | Foundational method; small dataset and high leakage risk if not patient-level. |
| Schrader et al., 2024 [40] | Prostate MRI + clinical data | Biopsy avoidance/risk nomogram | Automatic DL + clinical nomogram | Retrospective comparison | Clinically relevant outcome; prospective |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | decision impact needed. |