

[SYSTEMATIC REVIEW]

OPEN ACCESS

AI Models for Predicting Bone Metastasis in Prostate Cancer: A Systematic Review

Raisa Mubarak^{1*}, Zarin Tasnim Huda², Tasnia Ferdous³, Rezvee Muhammed Abdullah⁴, Nusrat Jahan⁵, Istiak Ahmed Dibbo⁶, Nahida Islam⁷, Tahmina Islam⁸, Abdullah Al Masud⁹

1 North-East Medical College and Hospital, South Surma, Sylhet, Bangladesh

2 Shaheed Tajuddin Ahmad Medical College and Hospital, Gazipur, Bangladesh

3 Holy Family Red Crescent Medical College, Eskaton Garden Road, Dhaka, Bangladesh

4 Mymensingh Medical College and Hospital, Mymensingh, Bangladesh

5 Shaheed Syed Nazrul Islam Medical College and Hospital, Kishorgonj, Bangladesh

6 Shaheed Suhrawardy Medical College and Hospital, Sher-E-Bangla Nagar, Dhaka, Bangladesh

7 Holy Family Red Crescent Medical College and Hospital, Dhaka, Bangladesh

8 Tangail Medical College, Bangladesh

9 Mymensingh Medical College and Hospital, Mymensingh, Bangladesh

* Correspondence: Raisa Mubarak, mubarakraisa13@gmail.com · ORCID: 0009-0008-8455-0953

Abstract

Background: Bone is the dominant hematogenous metastatic site in prostate cancer, and identifying patients at risk before or at staging could support more selective use of advanced imaging and earlier treatment planning. Artificial intelligence (AI), machine learning (ML), radiomics, and deep-learning approaches have increasingly been used to estimate bone-metastasis status from clinical variables and prostate imaging, but the evidence has not been synthesized around model validity and clinical readiness.

Methods: A systematic search was conducted to 31 July 2026 using PubMed/MEDLINE, scholarly web indexes, targeted publisher searches, and backward/forward citation checking. Eligible studies developed or validated an AI/ML/deep-learning or radiomics-based model for predicting the presence or risk of bone metastasis in adults with prostate cancer. Studies limited to detecting already visible bone lesions, prognosis after established bone metastasis, or inseparable non-bone metastatic outcomes were excluded. Reporting followed PRISMA 2020 principles. Model characteristics, predictors, validation strategy, and discrimination were narratively synthesized; methodological quality was appraised using PROBAST-informed domains.

Results: Fifteen studies published from 2019 to 2026 were included, representing 211,310 reported participant records, although 207,137 came from a single SEER-based study and several institutional cohorts may overlap. Twelve studies used MRI-derived radiomics, deep-learning features, or multimodal imaging; three were predominantly clinical ML models. Best reported validation/test discrimination ranged from 0.80 to 0.962. Four studies used an independent external cohort, with reported AUCs of 0.903-0.962. PSA, Gleason/ISUP grade, T and N stage, alkaline phosphatase, and radiomic signatures were recurrent predictors. Multimodal models frequently improved internal discrimination, but gains were less consistent on external data. Most studies were retrospective, geographically concentrated, and vulnerable to overfitting, class imbalance, non-independent cohorts, and inconsistent reference standards.

Conclusions: AI models show promising discrimination for prostate-cancer bone-metastasis risk, particularly when clinical variables are integrated with MRI-derived information. However, current evidence supports risk stratification and imaging triage rather than replacement of definitive staging. Prospective multicentre validation, standardized reference standards, calibration assessment, transparent reporting, and clinical-impact studies are required before routine deployment.

Keywords: prostate cancer; bone metastasis; artificial intelligence; machine learning; radiomics; deep learning; magnetic resonance imaging; prediction model

Received: 16 August 2026

Revised: 25 August 2026

Accepted: 29 August 2026

Published: 01 September 2026

Cite this article

Mubarak R, Huda ZT, Ferdous T, Abdullah RM, Jahan N, Dibbo IA, et al. AI Models for Predicting Bone Metastasis in Prostate Cancer: A Systematic Review. J Evid-Based Med Res. 2026;2(3):47–57.

© The Author(s) 2026. Published by Acadentium Research. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Bone is the predominant hematogenous metastatic destination of prostate cancer. In a large autopsy series, skeletal involvement was present in 90% of men with hematogenous metastatic spread, and contemporary clinical data continue to show that metastatic site is strongly associated with survival and disease burden [1,2]. Bone metastases can produce pain, fractures, spinal cord compression, marrow compromise and substantial

treatment complexity. Accordingly, accurate identification of skeletal dissemination is central to staging, treatment selection and prognosis.

Conventional imaging determines whether metastatic deposits are already present, whereas a prediction model addresses a different clinical question: how likely is bone metastasis given information available from the primary tumour and routine clinical assessment? Such models could serve as a pre-imaging triage layer, prioritize patients for definitive staging, or identify unexpectedly

high-risk patients despite apparently limited disease. This distinction is especially relevant where advanced molecular imaging is scarce, costly or selectively indicated.

Artificial intelligence (AI) offers several routes to this task. Conventional machine-learning algorithms can learn nonlinear interactions among PSA, Gleason or International Society of Urological Pathology (ISUP) grade, tumour stage, nodal status and biochemical markers. Radiomics converts magnetic resonance imaging (MRI) into high-dimensional quantitative descriptors of tumour phenotype, while deep learning can derive latent representations directly from images or histopathology. More recent approaches combine clinical, radiomic, pathomic and transformer-derived features in multimodal models. However, high apparent accuracy is not equivalent to clinical validity. Retrospective sampling, small event counts, feature-selection leakage, class rebalancing, scanner dependence and lack of external validation can substantially inflate performance.

The evidence base has expanded rapidly since the first prostate-MRI texture models in 2019, but studies differ in target population, imaging sequence, reference standard, algorithm, validation design and reporting. The present review therefore aimed to systematically identify and critically synthesize AI-based models for predicting bone metastasis in prostate cancer. The prespecified questions were: (1) which data modalities and AI methods have been used; (2) what discrimination has been achieved, particularly on independent external data; (3) which predictors recur across models; and (4) what methodological limitations currently prevent routine clinical implementation?

2. Methods

2.1 Review design and reporting

This systematic review was structured in accordance with the PRISMA 2020 reporting framework [3]. Because the included literature comprised heterogeneous diagnostic and risk-prediction models rather than intervention studies, a qualitative synthesis was planned a priori. A formal quantitative meta-analysis was not undertaken when model endpoints, predictors, patient populations, validation methods, and potential cohort overlap made a pooled discrimination estimate clinically misleading.

2.2 Search strategy

A structured electronic search was performed from database inception to 31 July 2026. PubMed/MEDLINE was the principal bibliographic database and was supplemented by scholarly web-index searches, targeted publisher searches, and backward/forward citation checking of eligible papers. The core search concept was: ("prostate cancer" OR "prostatic neoplasm") AND ("bone metastas*" OR "osseous metastas*" OR "skeletal metastas*") AND ("artificial intelligence" OR "machine learning" OR "deep learning" OR radiomic* OR "texture analysis" OR XGBoost OR "random forest" OR "neural network"). Search terms were adapted to source syntax. Titles identified through citation chaining were verified against the final peer-reviewed article and DOI. Only

reports available as complete peer-reviewed articles by the cutoff date were included.

2.3 Eligibility criteria

Studies were eligible when they: enrolled adults with histologically or clinically confirmed prostate cancer; developed, tested, or externally validated a multivariable AI/ML/deep-learning model or a radiomics-based computational prediction pipeline; used information available before or around the relevant staging decision; and reported bone metastasis as a separable target outcome. Both prediction of bone-metastasis status at initial diagnosis and prediction of subsequent bone metastasis after radical treatment were eligible because both estimate patient-level metastatic risk from non-bone predictors.

Studies were excluded if the algorithm merely detected or segmented metastases already visible on bone scintigraphy, CT, PET or MRI; analyzed prognosis only among patients with established bone metastasis; combined bone with other distant metastatic sites without a separable bone outcome; addressed a non-prostate primary; or were reviews, editorials, conference abstracts, preprints, or accepted manuscripts without a final peer-reviewed version. Conventional regression-only clinical nomograms were excluded unless embedded within an explicit radiomics or AI/ML pipeline. Radiomics pipelines using LASSO and multivariable classification were retained because they represent high-dimensional computational feature learning and formed the earliest evidence in this field.

2.4 Study selection and data extraction

Study selection used two sequential passes: a relevance screen of titles/abstracts followed by full-text eligibility verification. For each included report, data were extracted on publication year, setting, sample size, bone-metastasis prevalence where available, predictor modality, feature selection, algorithm, training/testing strategy, independent external validation, and the best reported discrimination estimate for the bone-metastasis task. When several algorithms were compared, both the numerically best result and clinically favored model were recorded where these differed. Performance reported on external cohorts was prioritized over internal validation; otherwise, held-out test or validation performance was used.

2.5 Methodological appraisal and synthesis

Methodological quality was considered using the domains of PROBAST—participants, predictors, outcome and analysis—rather than a numerical score [4]. Particular attention was given to retrospective sampling, event prevalence, independence of validation cohorts, possible institutional cohort overlap, data leakage risk, calibration, decision-curve analysis, and model interpretability. Because several reports originated from the same institutions over overlapping recruitment periods, summed participant counts were treated as study records rather than unique individuals. Results were synthesized by model family: predominantly clinical ML, radiomics/quantitative MRI, and deep-learning or multimodal fusion approaches.

3. Results

3.1 Study yield and characteristics

Fifteen peer-reviewed studies published between 2019 and July 2026 met the eligibility criteria (Figure 1; Table 1) [5–19]. Reported sample sizes ranged from 116 to 207,137 patients (median 214). The summed study-level count was 211,310, but one SEER-based analysis contributed 207,137 records [7]; the remaining 14 studies together reported 4,173 participants. Consequently, the apparent volume of evidence is heavily influenced by a single registry dataset. Most studies were conducted in China; one MRI/clinical study was from Türkiye, and the SEER study used a United States registry with Chinese external validation.

The field evolved in three broad phases (Figure 2). Early work used MRI texture analysis and radiomics nomograms [5,6]. A large registry study in 2021 demonstrated that standard clinicopathological variables could support high-performing ML models [7]. From 2024 onward, studies increasingly compared multiple classifiers, incorporated biparametric MRI, automated segmentation, deep transfer learning, pathomics, habitat analysis, peri-prostatic adipose-tissue radiomics, SHAP interpretation and Vision Transformers [8–19]. Twelve of the 15 studies were imaging-based or multimodal; three were primarily clinical ML models [7,11,18].

3.2 Clinical and laboratory machine-learning models

Liu et al. analyzed 207,137 SEER records and externally validated six algorithms in 644 patients [7]. XGBoost performed best, with an external AUC of 0.962, accuracy of 0.884, sensitivity of 0.906 and specificity of 0.879. Gleason score, PSA, T stage, N stage and age were the dominant predictors. This study provided the largest dataset and one of the strongest external results, although transportability from registry coding to other health systems remains an important consideration.

Jin et al. evaluated six algorithms in 404 newly diagnosed patients [11]. K-nearest neighbours produced the numerically highest test AUC (0.939), followed by random forest (0.929), but the authors favored logistic regression (AUC 0.921) because of interpretability and clinical usability. Gleason score, T stage, N stage, PSA and alkaline phosphatase (ALP) were retained, whereas inflammatory and nutritional indicators contributed little. In 2026, Wang et al. developed seven models in 327 patients from two hospitals [18]. Random forest achieved AUCs of 0.902 in training and 0.906 in internal validation, with SHAP identifying clinical T stage, Gleason score, total PSA, ALP, nodal metastasis and fibrinogen as influential. These findings reinforce that relatively accessible clinical variables can provide strong baseline discrimination.

3.3 Radiomics and quantitative MRI models

The earliest eligible study by Wang et al. extracted 976 features from T2-weighted and dynamic contrast-enhanced MRI in 176 patients [5]. A combined imaging biomarker plus free PSA, age and Gleason score achieved the highest AUC of 0.916, exceeding either MRI sequence alone. Zhang et al. subsequently developed a multiparametric MRI radiomics nomogram in 116

patients; the integrated radiomics-clinical model reached an AUC of 0.92 in a 35-patient validation cohort [6].

Biparametric MRI became a major focus in 2024. In 348 patients from two institutions, Song et al. compared logistic regression and seven ML classifiers [8]. XGBoost achieved an external AUC of 0.921 and accuracy of 0.876, while random forest yielded an external AUC of 0.911. A related two-centre study from the same research network enrolled 414 patients and integrated automated segmentation, radiomics, ResNet features and clinical variables [10]. The combined ResNet-C model achieved an internal AUC of 0.934 and external AUC of 0.903; importantly, the radiomics-only external AUC was also 0.903, suggesting that added deep-learning complexity did not improve discrimination on the external cohort.

Other 2025 radiomics studies showed variable generalization. Gündoğdu et al. integrated DCE-MRI perfusion features, ISUP grade and PSA density in 122 patients; XGBoost produced validation and test AUCs of 0.915 and 0.926 [12]. Nie et al. used T2 radiomics and random forest in 141 patients, with a training AUC of 0.88 but a lower test AUC of 0.80, illustrating a substantial generalization gap [13]. Yang et al. combined T2 radiomics with clinical information in 1,161 post-prostatectomy patients, only 38 of whom developed bone metastasis [14]. XGBoost showed training AUC 0.926, internal validation AUC 0.706, cross-validated/test estimates around 0.88–0.89, and an external AUC of 0.91. The marked variation across internal analyses highlights the instability that can arise from severe class imbalance and small event counts.

Two studies from Zhongshan Hospital of Traditional Chinese Medicine explored more detailed heterogeneity representations. Automated habitat analysis combining whole-prostate subregions, lesion radiomics and clinical data achieved a reported AUC of 0.900 for bone-metastasis prediction in 214 patients [15]. A separate 168-patient study found that an XGBoost model combining clinical and MRI radiomic features achieved a test AUC of 0.875, compared with 0.732 for radiomics alone [16]. Recruitment periods and institutional sources overlap, so these studies should not be assumed to represent independent populations.

Liu et al. shifted the region of interest from the tumour to peri-prostatic adipose tissue (PPAT) [17]. In 151 newly diagnosed patients, a nomogram integrating a 16-feature PPAT radiomics score, ALP and clinical N stage achieved a C-index of 0.908 and bootstrap-validated C-index of 0.903. This approach is mechanistically interesting because it attempts to capture the tumour microenvironment rather than the tumour alone, but independent external validation is still required.

3.4 Deep learning and multimodal fusion

Zhang et al. combined T2-weighted, diffusion-weighted and ADC radiomics with ResNet-50 deep-transfer features and H&E pathomics in 211 patients [9]. The multimodal model achieved an internal validation AUC of 0.93, compared with 0.89 for deep-transfer learning, 0.86 for radiomics and 0.85 for pathomics. In 2026, Li et al. evaluated clinical, ADC-radiomics and Vision Transformer (ViT) models in 220 patients [19]. The fusion model yielded AUCs of 0.944 in training and 0.894 in validation,

compared with 0.872 validation AUC for ViT alone, 0.842 for radiomics and 0.781 for the clinical model. SHAP indicated that the ViT-derived probability contributed most strongly to the fusion model. These studies support multimodal complementarity, but both relied on internal rather than independent external validation.

3.5 Overall discrimination and validation

Across studies, the best reported validation, test or bootstrapped discrimination ranged from 0.80 to 0.962 (Figure 3). Four studies clearly reported an independent external cohort [7,8,10,14], with external AUCs from 0.903 to 0.962. Most remaining studies used random hold-out splits, cross-validation, or bootstrapping within the source cohort. Strong internal AUCs were common, but external performance did not uniformly increase with model complexity. This was most evident in the semi-automatic bpMRI study, where radiomics and the integrated ResNet-C model both achieved external AUC 0.903 [10].

3.6 Recurrent predictors and methodological quality

PSA, Gleason/ISUP grade, tumour T stage, nodal status and ALP recurred across clinical and imaging studies (Table 2). MRI-derived texture and radiomic signatures added information related to tumour heterogeneity, while newer models used automated habitats, deep features, pathomics, PPAT characteristics and transformer embeddings. However, the methodological appraisal identified recurrent limitations (Table 3): all included studies were retrospective; independent external validation was uncommon; many cohorts were small relative to the dimensionality of imaging features; reference standards varied among bone scintigraphy, PET/CT, CT/MRI confirmation and follow-up; and possible cohort overlap existed between publications from the same institutions. Calibration, decision-curve analysis and explainability were reported increasingly often but were not universal. No prospective impact study demonstrated that model-guided imaging improved patient outcomes or reduced unnecessary testing.

4. Discussion

4.1 Principal findings

This systematic review shows that AI-based prediction of bone metastasis in prostate cancer has moved from handcrafted MRI texture features toward multimodal and interpretable learning systems within seven years. The overall signal is encouraging: most models reported discrimination near or above AUC 0.90, and the best externally validated models reached 0.903–0.962. Nevertheless, the evidence is better interpreted as proof of predictive potential than proof of clinical utility. High AUCs were achieved under heterogeneous and often retrospective conditions, while only four studies clearly tested performance in a genuinely independent external cohort.

A particularly important finding is that model complexity was not consistently rewarded on external data. Large-scale clinical ML based on routinely available variables achieved the highest external AUC in the review [7]. In the two-centre semi-automatic bpMRI study, the integrated deep-learning model and radiomics model had identical external AUCs [10]. Conversely, multimodal fusion

improved internal performance in the radiomics-pathomics and ViT studies [9,19]. These patterns suggest that imaging and deep features may add biologically meaningful information, but their incremental value must be demonstrated against strong clinical baselines on external cohorts rather than inferred from training performance.

4.2 Clinical interpretation

The most plausible near-term role for these models is risk-adapted staging. An AI score could be calculated from routine clinical data and, where available, prostate MRI before definitive metastatic imaging. A low predicted risk might support avoidance or deferral of low-yield testing in carefully defined populations, whereas a high score could prompt expedited or more sensitive imaging. Such a model should not be framed as a replacement for bone scintigraphy, PSMA PET/CT, MRI or other definitive staging methods. Prediction estimates probability; imaging establishes or refutes the presence and distribution of disease.

The recurrence of PSA, grade, T stage, N stage and ALP across independent model families is clinically coherent and strengthens confidence in these variables as core anchors. Radiomics then appears to act as a quantitative heterogeneity layer. Importantly, the best-performing algorithm differed by dataset: XGBoost dominated several studies, but random forest, KNN, Extra Trees and logistic regression also performed well. Algorithm branding is therefore less important than cohort quality, outcome definition, leakage prevention, calibration and transportability.

4.3 Why current performance may be optimistic

Several design features can inflate reported discrimination. First, high-dimensional radiomic feature spaces were often much larger than the number of metastatic events, creating substantial overfitting risk even when LASSO or cross-validation was used. Second, oversampling or other balancing procedures can alter apparent prevalence and calibration if applied before data splitting or evaluated on rebalanced data. Third, manual segmentation and scanner-specific acquisition can limit reproducibility across institutions. Fourth, some studies selected features and tuned models within a single institutional ecosystem and then assessed performance on a small hold-out subset, which is weaker than geographic external validation.

Reference standards also differed. Some studies used whole-body bone scintigraphy, others PET/CT or ECT, and some adjudicated equivocal findings with CT, MRI or follow-up. These tests do not have identical sensitivity, especially for early or small-volume disease. A model trained against a less sensitive reference can learn the probability of a positive test rather than the biological presence of metastasis. Future studies should prespecify a clinically appropriate reference standard and ensure that outcome ascertainment is blinded to predictor information where feasible.

Cohort non-independence is another concern. The two Song et al. studies used overlapping institutions and partially overlapping recruitment periods [8,10]. The two Zhongshan radiomics studies also arose from the same

institution with overlapping time windows [15,16]. Studies by Zhang et al. and Jin et al. both drew on Gansu Provincial Hospital across overlapping years [9,11]. These publications answer different methodological questions and therefore merit inclusion, but they should not be treated as independent patient samples in a pooled meta-analysis. This was a central reason for using narrative synthesis.

4.4 Priorities for clinical translation

Future model development should move from retrospective discrimination studies toward prospective validation and clinical impact. At minimum, studies should use geographically and temporally independent cohorts, retain natural event prevalence in evaluation data, publish calibration intercepts/slopes or calibration plots, report sensitivity and negative predictive value at clinically meaningful thresholds, and perform decision-curve analyses against explicit standard-of-care strategies. Reporting should make the full preprocessing sequence transparent, particularly whether feature selection, normalization, imputation and oversampling occurred within each training fold.

For imaging models, multicentre harmonization is essential. Scanner vendor, field strength, reconstruction, slice thickness, b-values and segmentation practice can alter radiomic features. External testing should therefore include heterogeneous scanners rather than a single closely related protocol. Reproducibility can be improved by publishing feature definitions, code, trained weights and locked inference pipelines. Deep-learning systems should be compared with simpler clinical models to quantify incremental benefit and should include interpretable outputs that allow clinicians to identify implausible predictions.

The most clinically valuable endpoint may also need refinement. Several included studies predicted bone-metastasis status at diagnosis, while one focused on metastasis after prostatectomy. Future research should define whether the intended use is to omit staging in very-low-risk patients, escalate staging in otherwise intermediate-risk patients, predict future skeletal dissemination after local therapy, or select surveillance intensity. Each use case requires a distinct cohort, time horizon, threshold and consequence analysis (Table 4).

4.5 Strengths and limitations of this review

This review focused narrowly on patient-level prediction of prostate-cancer bone metastasis and deliberately excluded lesion-detection algorithms that answer a different clinical question. It also separated external from internal validation and highlighted probable cohort overlap rather than pooling non-independent AUCs. Limitations remain. The search prioritized PubMed/MEDLINE and scholarly web indexes rather than every subscription database; some relevant non-English or unpublished studies may therefore have been missed. Study reporting was heterogeneous, preventing uniform extraction of calibration and threshold metrics. Reported AUCs were not recalculated from individual participant data, and the review did not pool discrimination because doing so would obscure major differences in population, endpoint, model and validation design. Finally, likely

cohort overlap could not always be quantified from published reports.

5. Conclusions

AI models for predicting bone metastasis in prostate cancer demonstrate consistently promising discrimination across clinical ML, MRI radiomics and multimodal deep-learning approaches. External AUCs above 0.90 have been reported, and recurrent predictors—PSA, tumour grade, T/N stage, ALP and quantitative MRI signatures—provide a coherent foundation for risk stratification. However, most evidence remains retrospective and internally validated, with substantial heterogeneity and possible cohort overlap. Current models are best viewed as candidate triage tools that may complement, not replace, definitive metastatic imaging. Prospective multicentre external validation, standardized reference standards, calibration at clinically meaningful thresholds, transparent reproducible pipelines and randomized or pragmatic impact studies are needed before routine clinical adoption.

Declarations

Funding: None.

Conflicts of interest: None.

Data availability: All data synthesized in this review were derived from published studies cited below.

Ethics approval and consent

Not applicable.

Acknowledgements

We are especially grateful to Dr. Abdullah Al Masud, MD, FACP and his organization Dream USMLE, 16523 Wildhorse Creek 63017, Chesterfield, MO, USA, who gave us the technical guidance.

Author contributions

Raisa Mubarak: Conceptualization, Methodology, Project administration, Supervision, Writing-original draft, Writing-review and editing.

Zarin Tasnim Huda: Methodology, Investigation, Data curation, Formal analysis, Writing-original draft, Writing-review and editing.

Tasnia Ferdous: Investigation, Data curation, Validation, Writing-original draft, Writing-review and editing.

Rezvee Muhammed Abdullah: Investigation, Data curation, Validation, Writing-review and editing.

Nusrat Jahan: Formal analysis, Visualization, Writing-review and editing.

Istiaq Ahmed Dibbo: Investigation, Data curation, Writing-review and editing.

Nahida Islam: Investigation, Data curation, Writing-review and editing.

Tahmina Islam: Writing-review and editing.

Abdullah Al Masud: Writing-review and editing.

All authors reviewed and approved the final version of the manuscript.

Generative AI use

For paraphrasing of citations and basic structure of outline, actually very limited use of Quillbot has been made and all have been edited properly after review.

References

- Bubendorf L, Schopfer A, Wagner U, Sauter G, Moch H, Willi N, et al. Metastatic patterns of prostate cancer: an autopsy study of 1,589 patients. *Hum Pathol*. 2000;31(5):578-583. doi:10.1053/hp.2000.6698.
- Gandaglia G, Karakiewicz PI, Briganti A, Passoni NM, Schiffmann J, Trudeau V, et al. Impact of the site of metastases on survival in patients with metastatic prostate cancer. *Eur Urol*. 2015;68(2):325-334. doi:10.1016/j.eururo.2014.07.020.
- 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.
- Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58. doi:10.7326/M18-1376.
- Wang Y, Yu B, Zhong F, Guo Q, Li K, Hou Y, et al. MRI-based texture analysis of the primary tumor for pre-treatment prediction of bone metastases in prostate cancer. *Magn Reson Imaging*. 2019;60:76-84. doi:10.1016/j.mri.2019.03.007.
- Zhang W, Mao N, Wang Y, Xie H, Duan S, Zhang X, et al. A radiomics nomogram for predicting bone metastasis in newly diagnosed prostate cancer patients. *Eur J Radiol*. 2020;128:109020. doi:10.1016/j.ejrad.2020.109020.
- Liu WC, Li MX, Qian WX, Luo ZW, Liao WJ, Liu ZL, et al. Application of machine learning techniques to predict bone metastasis in patients with prostate cancer. *Cancer Manag Res*. 2021;13:8723-8736. doi:10.2147/CMAR.S330591.
- Song X, Zhang S, Shen T, Hu X, Wang Y, Du M, et al. A machine learning radiomics model based on bpMRI to predict bone metastasis in newly diagnosed prostate cancer patients. *Magn Reson Imaging*. 2024;107:15-23. doi:10.1016/j.mri.2023.12.009.
- Zhang YF, Zhou C, Guo S, Wang C, Yang J, Yang ZJ, et al. Deep learning algorithm-based multimodal MRI radiomics and pathomics data improve prediction of bone metastases in primary prostate cancer. *J Cancer Res Clin Oncol*. 2024;150(2):78. doi:10.1007/s00432-023-05574-5.
- Song X, Shen T, Hu X, Zhang S, Wang Y, Du M, et al. A semi-automatic deep learning model based on biparametric MRI scanning strategy to predict bone metastases in newly diagnosed prostate cancer patients. *Front Oncol*. 2024;14:1298516. doi:10.3389/fonc.2024.1298516.
- Jin T, An J, Wu W, Zhou F. Development and validation of a machine learning model for bone metastasis in prostate cancer: based on inflammatory and nutritional indicators. *Urology*. 2024;190:63-70. doi:10.1016/j.urology.2024.05.027.
- Gundogdu H, Panc K, Sekmen S, Er H, Gurun E. Enhancing bone metastasis prediction in prostate cancer using quantitative mpMRI features, ISUP grade and PSA density: a machine learning approach. *Abdom Radiol (NY)*. 2025;50(5):2221-2231. doi:10.1007/s00261-024-04667-0.
- Nie S, Fan B, Gui S, Zou H, Lan M. Predictive impact of T2-MRI radiomics model on initial diagnosis of bone metastasis in prostate cancer patients. *BMC Med Imaging*. 2025;25(1):106. doi:10.1186/s12880-025-01642-z.
- Yang H, Wei C, Zhou S, Mao F. Machine learning-based identification of high-risk bone metastasis factors after radical prostatectomy in prostate cancer. *Front Oncol*. 2025;15:1549851. doi:10.3389/fonc.2025.1549851.
- Yang Y, Zheng B, Zou B, Liu R, Yang R, Chen Q, et al. MRI radiomics and automated habitat analysis enhance machine learning prediction of bone metastasis and high-grade Gleason scores in prostate cancer. *Acad Radiol*. 2025;32(9):5303-5316. doi:10.1016/j.acra.2025.05.059.
- Yang Y, Zou B, Zheng B, Guo Y, Yu S, Chen B. Predicting bone metastasis and high-grade Gleason scores in prostate cancer: a retrospective study integrating clinical features and magnetic resonance imaging radiomics. *Transl Androl Urol*. 2025;14(10):2844-2858. doi:10.21037/tau-2025-412.
- Liu B, Cai Q, Zhao X, Su H, Lin Z, Wu J, et al. A nomogram based on radiomic features from peri-prostatic adipose tissue for predicting bone metastasis in first-time diagnosed prostate cancer patients. *Adipocyte*. 2025;14(1):2517583. doi:10.1080/21623945.2025.2517583.
- Wang L, He W, Zhao C, Lv Q, Qiu T, Zhai J, et al. Development and validation of a machine learning-based early warning model for bone metastasis in newly diagnosed prostate cancer. *Front Oncol*. 2026;16:1815465. doi:10.3389/fonc.2026.1815465.
- Li G, Liu L, Xu Z, Huang Z, Zheng T, Liu Z, et al. A multimodal fusion model integrating Vision Transformer, radiomics, and clinical features for predicting bone metastasis in prostate cancer. *Front Oncol*. 2026;16:1841761. doi:10.3389/fonc.2026.1841761.

Table 1. Characteristics of all included studies

Study	Setting / n	Prediction data	Model / algorithm	Validation	Best reported discrimination	Key predictors / finding	Notes
Wang 2019 [5]	China; n=176	T2WI + DCE-T1 radiomics; clinical	Texture features + regression/LASSO	Internal	AUC 0.916	MRI signature + fPSA, age, Gleason	Earliest eligible MRI texture model
Zhang 2020 [6]	China; n=116 (81/35)	T2-FS, DWI, DCE-T1 radiomics + clinical	LASSO radiomics nomogram	Hold-out validation	AUC 0.920	12-feature signature + clinical risk factors	Small validation cohort
Liu 2021 [7]	SEER n=207,137; external China n=644	Clinical / registry	DT, RF, MLP, LR, NB, XGB	Independent external	AUC 0.962	Gleason, PSA, T, N, age	Largest study; web predictor
Song 2024 bpMRI [8]	China, 2 centres; n=348	T2WI + DWI radiomics	LR, NB, XGB, RF, GBDT, SVM, AdaBoost, KNN	Independent external n=64	AUC 0.921 (XGB)	bpMRI radiomics	Likely overlaps research network with [10]
Zhang 2024 [9]	China; n=211	T2WI/DWI/ADC radiomics + H&E pathomics + DL	ResNet50 DTL + ML fusion	Internal validation n=42	AUC 0.930	Multimodal fusion	Possible institutional-period overlap with [11]
Song 2024 ResNet-C [10]	China, 2 centres; n=414	bpMRI + clinical + radiomics + DL	UNETR/ResNet; ResNet-C fusion	Independent external n=96	AUC 0.903	Integrated radiomics + DL + clinical	Radiomics alone also external AUC 0.903
Jin 2024 [11]	China; n=404	Clinical/laboratory	LR, RF, XGB, NB, KNN, DT	Held-out test	AUC 0.939 KNN; LR 0.921	Gleason, T, N, PSA, ALP	LR favored for interpretability
Gundogdu 2025 [12]	Türkiye; n=122	DCE-MRI quantitative features + ISUP + PSAD	Multiple ML; XGB best	Internal validation/test	AUC 0.926 test	IAUGC, MaxSlope, ISUP, PSAD	Oversampling used; no external cohort
Nie 2025 [13]	China; n=141	T2WI radiomics	LASSO + RF	Held-out test	AUC 0.800	T2 texture signature	Generalization drop from train AUC 0.88
Yang 2025 post-RP [14]	China; n=1,161; BM n=38	T2 radiomics + clinical/postoperative	XGB, RF, SVM, KNN	Independent external n=361	AUC 0.910 external	Radscore, Gleason, tumour volume, ALP, PSA, ECE	Severe class imbalance; internal estimates varied
Yang 2025 habitat [15]	China; n=214	T2-FS habitat + lesion radiomics + clinical	nnUNet segmentation + Extra Trees	Internal evaluation	AUC 0.900	PSA + habitat/lesion radiomics	Probable cohort overlap with [16]
Yang 2025 TAU [16]	China; n=168	MRI radiomics + clinical	Multiple ML; XGB best	Held-out test	AUC 0.875	fPSA + radiomics	Combined model > radiomics-only AUC 0.732
Liu 2025 PPAT [17]	China; n=151	PPAT T2 radiomics + clinical	LASSO Radscore + logistic nomogram	Bootstrap internal	C-index 0.903	Radscore, ALP, cN	Novel microenvironment ROI; no external validation
Wang 2026 [18]	China, 2 hospitals; n=327	Clinical/laboratory	Boruta + 7 ML models; RF best	Internal validation n=98	AUC 0.906	cT, Gleason, tPSA, ALP, nodes, fibrinogen	SHAP + online tool; no independent external cohort
Li 2026 [19]	China; n=220	ADC radiomics + clinical + ViT	Clinical, radiomics, ViT, fusion	Internal validation n=66	AUC 0.894	ViT probability dominant in SHAP	Fusion > clinical/radiomics/ViT alone

Abbreviations: ADC, apparent diffusion coefficient; ALP, alkaline phosphatase; AUC, area under the receiver operating characteristic curve; BM, bone metastasis; cN/cT, clinical nodal/tumour stage; DCE, dynamic contrast enhanced; DTL, deep transfer learning; ECE, extracapsular extension; ECE, extracapsular extension; fPSA, free prostate-specific antigen; LR, logistic regression; PPAT, peri-prostatic adipose tissue; PSAD, PSA density; RF, random forest; ViT, Vision Transformer; XGB, extreme gradient boosting.

Table 2. Recurrent predictor domains and their interpretation

Predictor domain	Examples repeatedly selected	Model families	Interpretation for BM risk	Evidence pattern
Tumour burden / biology	PSA, fPSA, PSA density	Clinical ML; radiomics; fusion	Higher biochemical tumour burden and aggressive phenotype	Most consistent predictor family across studies
Histologic grade	Gleason score / ISUP grade	Clinical ML; MRI + clinical	Higher grade reflects aggressive metastatic potential	Repeatedly retained in clinical and multimodal models
Anatomic stage	T stage, N stage, extracapsular extension	Clinical ML; fusion	Local extension and nodal disease proxy systemic dissemination	Important in registry, postoperative and nomogram studies
Bone-turnover / laboratory	ALP; fibrinogen in one model	Clinical ML; nomogram	ALP may reflect occult skeletal involvement or disease burden	ALP recurrent; inflammatory/nutritional markers otherwise weak
Primary-tumour radiomics	Texture, wavelet, GLSZM/GLDM, ADC/T2/DWI signatures	Radiomics / ML	Quantifies intratumoural heterogeneity not captured by routine assessment	Frequently improves over single clinical or single-sequence models
Microenvironment / habitat	PPAT radiomics; automated prostate habitats	Radiomics / clustering	Captures peritumoural or spatial heterogeneity	Promising but currently single-centre and internally validated
Deep image representations	ResNet DTL; ViT probability	Deep learning / fusion	Learns latent image phenotype directly from MRI or pathology	Improved internal fusion models; incremental external value uncertain

Table 3. Validation and methodological quality indicators (PROBAST-informed)

Study	Retrospective	Independent external validation	Calibration / DCA	Interpretability	Primary methodological concern
Wang 2019	Yes	No	Limited / not central	Regression coefficients	Small, high-dimensional single-centre cohort
Zhang 2020	Yes	No	Yes / Yes	Nomogram	Only 35 validation patients
Liu 2021	Yes	Yes (n=644)	Not fully emphasized	Feature importance + web tool	Registry coding / transportability; low BM prevalence
Song 2024 bpMRI	Yes	Yes (n=64)	NR	Algorithm comparison	Small external cohort; probable overlap with related publication
Zhang 2024	Yes	No	Yes / Yes	Multimodal feature weights	Many features relative to validation n=42
Song 2024 ResNet-C	Yes	Yes (n=96)	NR	Component comparison	Auto-segmentation DSC modest; overlap risk
Jin 2024	Yes	No	Model comparison	LR chosen for interpretability	Single-centre; numerically best KNN not preferred clinical model
Gundogdu 2025	Yes	No	NR	Feature-level clinical interpretation	Small n; oversampling can affect calibration
Nie 2025	Yes	No	NR	Radiomic feature selection	Test AUC fell to 0.80; single-centre
Yang 2025 post-RP	Yes	Yes (n=361)	Yes / Yes	SHAP	Only 38 BM events; unstable internal estimates

Study	Retrospective	Independent validation	external	Calibration / DCA	Interpretability	Primary concern	methodological
Yang 2025 habitat	Yes	No		NR	Feature importance	Single-centre; likely cohort overlap with TAU study	
Yang 2025 TAU	Yes	No		NR	Feature weights	Single-centre; likely cohort overlap	
Liu 2025 PPAT	Yes	No (bootstrap)		Yes / Yes	Nomogram	Bootstrap only; novel ROI needs external reproducibility	
Wang 2026	Yes	No independent		Yes / Yes	SHAP + online tool	Two-centre source but random internal split	
Li 2026	Yes	No		Yes / Yes	SHAP	Single-centre internal validation only	

DCA, decision curve analysis; NR, not reported or not sufficiently detailed in the extracted report. “PROBAST-informed” indicates qualitative appraisal using the participants, predictors, outcome and analysis concepts of PROBAST rather than a numerical score.

Table 4. Translational gaps and recommended standards for future studies

Domain	Observed gap	Why it matters	Recommended minimum standard
Target use	Diagnostic status and future risk often mixed	Different decisions require different time horizons and thresholds	State intended use: imaging triage, post-treatment risk, or surveillance; prespecify horizon
Reference standard	Bone scan, PET/CT, CT/MRI adjudication and follow-up varied	Label noise changes what the model actually learns	Use a prespecified high-quality reference and adjudication pathway; blind outcome assessment where feasible
Validation	External validation uncommon	Internal AUC cannot establish transportability	Geographic and temporal external validation before clinical claims
Class imbalance	Some cohorts had very few BM events	Can inflate accuracy and destabilize AUC/calibration	Preserve natural prevalence in test sets; report event counts and threshold metrics
Data leakage	Feature selection/normalization/oversampling workflows variably reported	Leakage can markedly overstate performance	Perform all preprocessing within training folds and document the pipeline
Calibration	Often absent despite good AUC	AUC does not guarantee reliable absolute risk	Report calibration-in-the-large, slope/plot, Brier score when appropriate
Clinical utility	Few comparisons with guideline or imaging strategies	Net benefit determines whether predictions improve decisions	Decision-curve analysis and head-to-head comparison with standard staging rules
Reproducibility	Scanner, segmentation and code details inconsistent	Radiomic features can be protocol dependent	Multivendor testing; robust feature definitions; publish code/weights or locked executable
Impact	No prospective model-guided outcome study	Predictive performance does not prove patient benefit	Prospective pragmatic or randomized impact evaluation

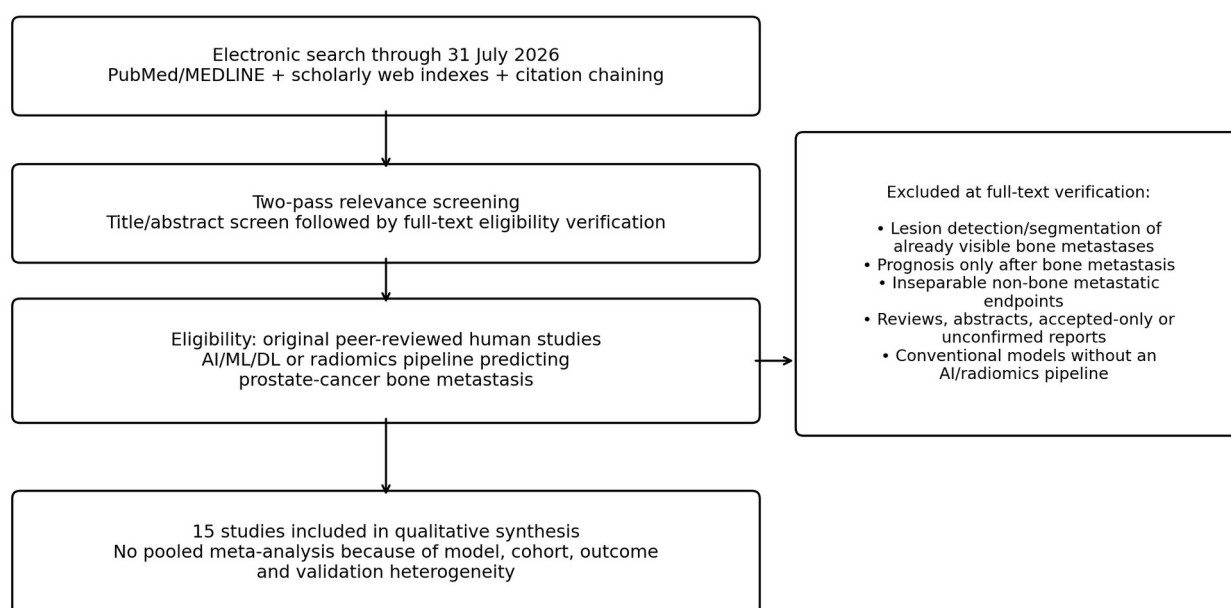


Figure 1. Eligibility and synthesis workflow. The figure deliberately avoids unverified database record counts and reports only the final included-study count.

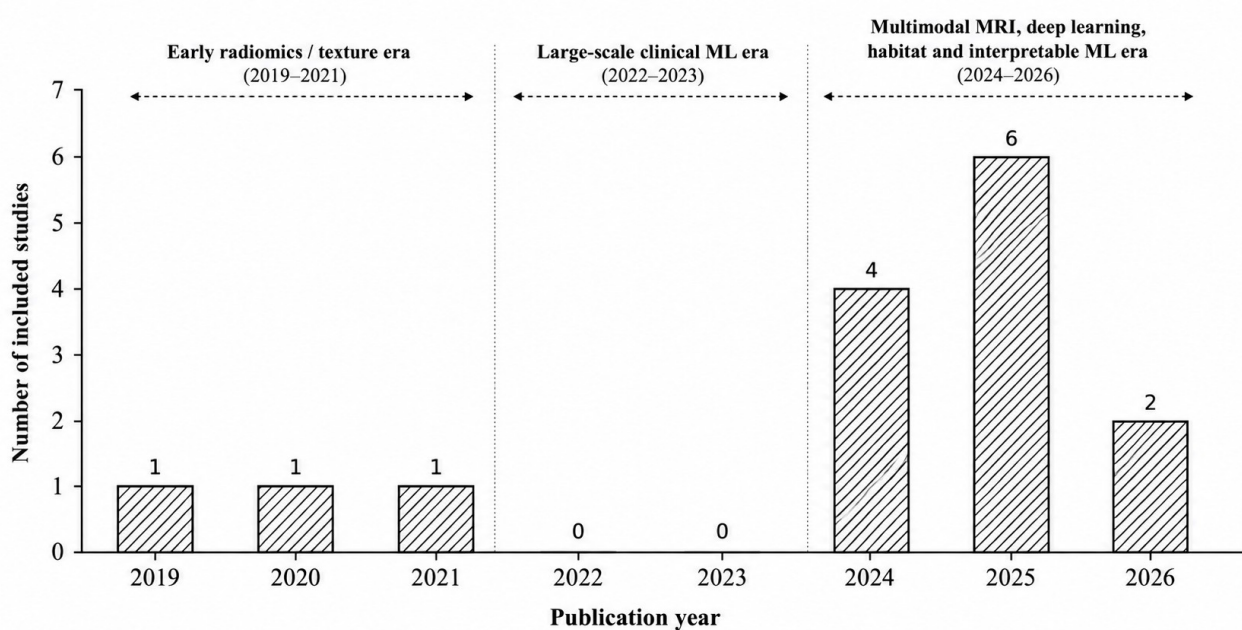


Figure 2. Expansion of AI bone-metastasis prediction studies over time. The increase after 2024 reflects the move from radiomics toward multimodal MRI, deep learning, habitat analysis and interpretable ML.

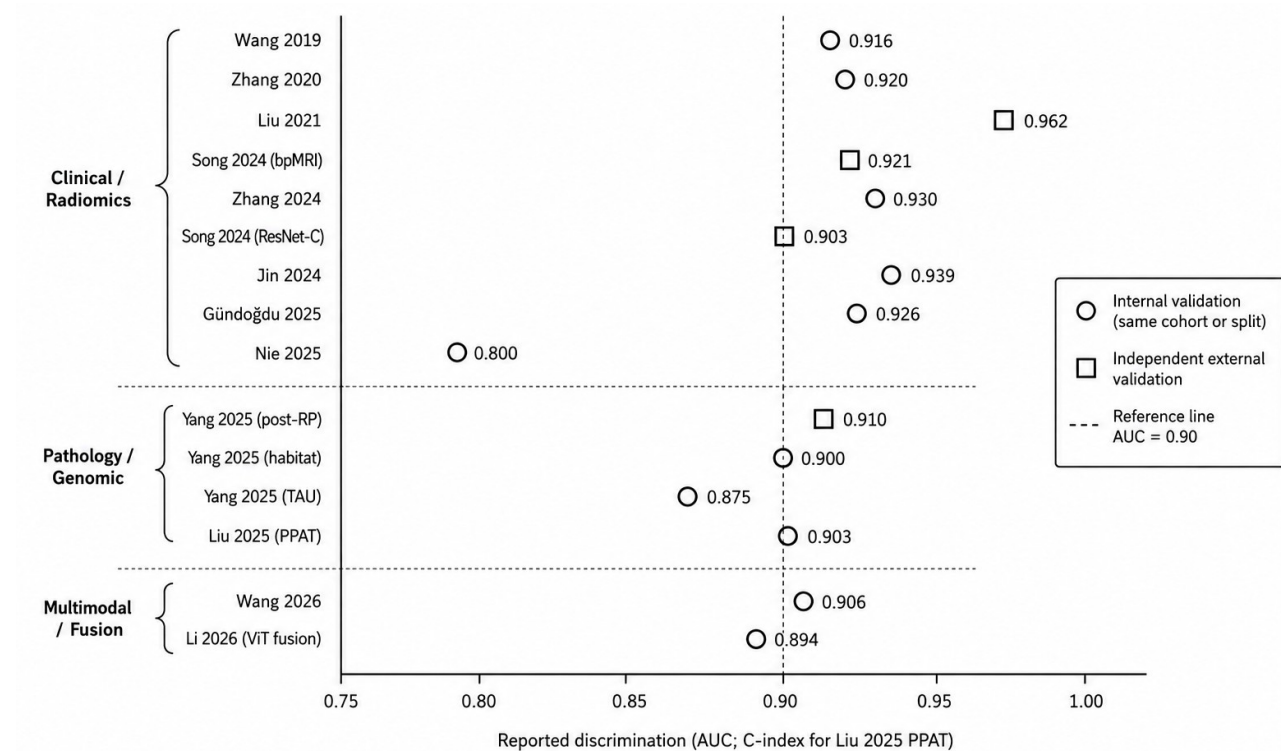


Figure 3. Best reported validation/test discrimination by study. Independent external validation is shown with square markers. For Liu et al. 2025 PPAT, the reported bootstrap C-index (0.903) is plotted on the same discrimination scale. Values should not be pooled because cohorts, endpoints and validation designs differ.