

Automated Gleason Grading and Prostate Cancer Classification from Histopathology Images: A Systematic Review of Deep Learning Models, Generalizability, Failure Modes, and Challenges in Low-Resource Settings

Habert Tumwesigye¹

htumwesigye@must.ac.ug

Mbarara University of Science and Technology

Phionah Nabimanya,

Mbarara University of Science and Technology

Abraham Birungi

Mbarara University of Science and Technology

Dr. Samuel Tumwesigire

Mbarara University of Science and Technology

Dr. Yekosan Mitala

Mbarara University of Science and Technology

Dr. Raymond Atwine

Mbarara University of Science and Technology

Systematic Review

Keywords: prostate cancer, Gleason grading, digital pathology, artificial intelligence, deep learning, computational pathology, whole-slide imaging, external validation, domain shift, low-resource settings

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11258913/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Background

Digital pathology and deep learning have produced rapid advances in automated prostate cancer detection and Gleason grading. However, headline performance from internal benchmark datasets does not necessarily translate to reliable clinical performance across institutions, scanners, staining protocols, specimen types, or patient populations. This review synthesizes evidence on automated Gleason grading and prostate cancer classification from histopathology images, with particular emphasis on external validation, generalizability, failure modes, human–AI interaction, and deployment in low-resource settings.

Methods

The review was structured according to PRISMA 2020 principles. PubMed-indexed studies, publisher records, systematic reviews, and methodological guidance were examined, with emphasis on original studies of digital histopathology AI for prostate cancer and Gleason grading. Fifty key publications were selected for the evidence synthesis, supplemented conceptually by broader literature identified in prior systematic reviews. Because the evidence is heterogeneous in specimen type, endpoint, annotation strategy, architecture, and validation design, a narrative synthesis was used rather than an unverified pooled estimate.

Results

The evidence demonstrates that deep learning can approach or exceed the agreement of general pathologists and can achieve expert-level performance in selected external cohorts. The PANDA challenge demonstrated strong cross-continental performance using 10,616 digitized prostate biopsies, while subsequent real-world evaluations showed wider performance variation among algorithms. External validation on prostatectomy specimens has also demonstrated strong agreement for a biopsy-trained commercial model. Prospective multicenter evidence published in 2026 indicates that AI assistance can alter Gleason grading in routine practice and increase diagnostic confidence. Persistent weaknesses include domain shift, tissue and scanning artifacts, class imbalance, difficulty distinguishing adjacent Gleason patterns, limited prospective validation, dataset leakage, opaque reference standards, and underrepresentation of low-resource settings.

Conclusion

AI-assisted prostate histopathology has moved beyond proof-of-concept research, but clinical readiness should not be inferred from accuracy on a single dataset. Future studies should prioritize independent

multicenter validation, prespecified clinical endpoints, transparent reference standards, calibration and uncertainty reporting, robustness testing, prospective workflow studies, and implementation strategies compatible with low-resource pathology laboratories.

1. Introduction

Prostate cancer grading remains a central determinant of prognosis and clinical management. The modern Gleason system and subsequent International Society for Urological Pathology (ISUP) Grade Group framework standardized the interpretation of architectural patterns, but grading remains subject to interobserver variability, particularly for borderline or morphologically challenging patterns. The distinction between Gleason Pattern 3 and Pattern 4 is clinically important because relatively small changes in grade grouping can affect risk classification and management [10, 11].

Digital pathology has created an opportunity to convert histological interpretation into computationally analyzable whole-slide images (WSIs). Convolutional neural networks, transformer architectures, multiple-instance learning, weak supervision, and hybrid segmentation-classification systems have been applied to cancer detection, Gleason-pattern recognition, Grade Group assignment, tissue segmentation, and prognostic prediction [12–17, 20–22].

Early studies demonstrated proof of concept using relatively small datasets, whereas later investigations used thousands of biopsies or slides. The 2020 population-based study by Ström et al. used thousands of biopsy slides and reported high discrimination for cancer detection and grading performance within the range of expert pathologists [15]. Bulten et al. subsequently demonstrated automated Gleason grading of biopsies in a diagnostic study [12], and their work was followed by evidence that pathologists may improve their agreement with expert reference standards when AI assistance is available [14].

The PANDA challenge represented a major change in scale and reproducibility. It used 10,616 digitized prostate biopsies and independently evaluated algorithms on US and European cohorts, providing evidence that several independently developed systems could achieve pathologist-level agreement on external datasets [1]. However, later evaluations of algorithms in more heterogeneous real-world material have shown wider ranges of agreement, reinforcing that benchmark success and clinical generalizability are not synonymous [23].

The field has also expanded beyond biopsy grading. External validation of a biopsy-trained commercial algorithm on whole-mount prostatectomy specimens showed strong agreement with expert assessment despite differences in specimen type and patient population [2]. In parallel, prospective multicenter studies have begun to examine how AI affects pathologists in actual diagnostic workflows rather than only measuring standalone algorithm performance [3]. The purpose of this updated review is therefore not simply to summarize model accuracy. It is to examine whether automated prostate histopathology systems are reproducible, externally valid, interpretable, robust to failure modes, and realistically deployable in settings where specialist pathology and digital infrastructure are limited.

2. Review Objectives

To summarize deep-learning approaches used for prostate cancer detection, Gleason-pattern classification and Grade Group assignment.

To compare evidence from internal validation, external validation, real-world evaluation and prospective clinical studies.

To identify major technical and clinical failure modes that may compromise generalizability.

To assess the implications of dataset composition, specimen type, scanner and staining variation, and annotation strategy.

To examine the evidence and practical requirements for deployment in low-resource pathology environments.

To identify methodological priorities for future multicenter and prospective validation.

3. Methods

3.1 Review framework

The review was structured using PRISMA 2020 principles and diagnostic-accuracy reporting concepts. QUADAS-2 and PROBAST/PROBAST + AI were used as conceptual frameworks for considering risk of bias and applicability [5–9]. The objective was to produce a clinically focused evidence synthesis rather than a fabricated quantitative meta-analysis.

3.2 Eligibility focus

Priority was given to peer-reviewed original research using prostate histopathology images, especially H&E-stained biopsy or prostatectomy WSIs, in which AI or deep learning was used for cancer detection, Gleason-pattern classification, Grade Group assignment, segmentation, or clinically relevant pathology prediction. Systematic reviews, consensus papers and reporting-methodology papers were retained when they materially informed interpretation of the primary evidence. MRI-only radiomics studies were not treated as primary evidence for automated histopathology grading.

3.3 Evidence verification

Bibliographic records were checked against PubMed and publisher records where available. The updated evidence base emphasizes studies published through 2026, including recent prospective and external-validation work. The review does not claim a complete census of every paper ever published; instead, it presents 50 key papers selected to represent the major evidence streams and methodological developments.

3.3 a Protocol and registration

No prospective protocol registration number is claimed in this version of the manuscript. The authors should register the protocol prospectively, where feasible, before describing the work as a completed systematic review, and any protocol amendments should be documented transparently. This statement is included to avoid implying a registration that does not exist.

3.4 Data synthesis

The extracted domains were publication year, task, specimen type, dataset scale, model family, annotation strategy, validation design, reference standard, principal performance measure, external validation, human comparison, and reported limitations. Because these domains were heterogeneous, numerical performance values were not pooled across incomparable studies.

4. Results and Evidence Synthesis

4.1 Evolution of the evidence base

The literature evolved from small patch-level classification experiments to large WSI-based systems. Early studies demonstrated that deep residual networks and multiscale CNNs could distinguish benign tissue and individual Gleason patterns, but their limited sample sizes and internal validation designs constrained generalizability [12, 17]. Subsequent studies expanded to large biopsy cohorts, automated localization, whole-slide grading and independent institutions [1, 2, 13–15].

A 2023 systematic review identified 77 deep-learning studies in prostate digital pathology, with most addressing diagnosis or grading and recurrent concerns about lack of external validation. A separate 2023 systematic review and meta-analysis included 24 studies and found high reported diagnostic performance but substantial statistical heterogeneity, illustrating why pooled accuracy should be interpreted cautiously [4, 21]. A 2024 systematic comparison further showed that optimal learning strategies depend on whether the task is local Gleason grading or whole-slide scoring and on the availability of detailed labels [22].

The principal evidence streams identified in this review are summarized in Table 1, highlighting the progression from small proof-of-concept studies to larger and externally evaluated systems.

Table 1
Major evidence streams and their principal methodological implications

Evidence stream	Representative evidence	Main implication
Small supervised CNN studies	Patch-level classification and multiscale CNNs [12, 17]	Proof of concept; limited external validity
Large biopsy studies	Ström; Bulten; Nagpal [13–15]	Strong evidence for detection and grading assistance
PANDA challenge	10,616 biopsies; independent US/European validation [1]	Major evidence for cross-cohort reproducibility
AI + pathologist	Reader studies and prospective clinical evaluation [14, 3]	Human–AI synergy may matter more than standalone accuracy
Real-world algorithms	Crowdsourced heterogeneous data [2]	Performance varies across algorithms and real-world cases
Prostatectomy external validation	DeepDx evaluation [2]	Biopsy-trained AI may generalize, but specimen shift must be tested
Weak/semisupervised learning	MIL, CLAM, transformers and newer frameworks [2]	Reduces dependence on pixel-level annotation
Low-resource deployment	Digital pathology infrastructure literature [18, 19]	Infrastructure and implementation are as important as model accuracy

4.2 Deep-learning architectures

CNNs remain important because they efficiently learn local morphological features from image patches. ResNet-like architectures have been used extensively, while transfer learning from general-purpose networks has been explored in smaller datasets [8, 12]. Multiscale approaches attempt to capture both cellular and glandular architecture, which is particularly relevant because Gleason grading is fundamentally an architectural assessment [9].

More recent work has used attention mechanisms, multiple-instance learning (MIL), transformers and hybrid networks. MIL is attractive for whole-slide grading because slide-level labels can be used without exhaustive pixel-level annotation. A systematic comparison across nine datasets found that fully supervised approaches performed strongly for local grading, while CLAM and related MIL approaches were effective for slide-level Gleason scoring [22]. Newer studies have also investigated federated learning, explainability and frequency-domain/multiscale representations [32,45,46]. The representative model families and their methodological roles are summarized in Table 2.

Table 2
Representative model families and their methodological roles

Model family	Typical task	Strength	Key limitation
ResNet/CNN	Patch classification, Gleason pattern	Strong local feature extraction	Requires representative training data
Multiscale CNN	Gleason pattern/score	Combines local and contextual information	Higher computational/data complexity
U-Net/segmentation networks	Tissue/gland/epithelium segmentation	Explicit morphological localization	Segmentation errors propagate to grading
Attention/MIL/CLAM	WSI Grade Group scoring	Uses slide-level labels	May miss subtle focal morphology
Transformer/hybrid models	WSI or patch classification	Long-range contextual modeling	Data and computational demands
Federated learning	Multi-site training	Can reduce direct data sharing	Complex harmonization and validation
Explainable AI	Decision support	May improve interpretability	Saliency does not prove causal reasoning

4.3 Automated cancer detection

Cancer detection is generally easier than precise Gleason grading. Large studies have reported very high discrimination for benign-versus-malignant classification, including AUC values near 0.99 in independent cohorts [13, 15, 21]. This reflects the fact that malignant tissue can contain relatively distinctive architectural patterns. Nevertheless, false positives from benign mimics and false negatives from small or unusual lesions remain clinically important. Systematic reviews similarly report strong diagnostic discrimination while emphasizing heterogeneity and limited external validation [20, 21].

The clinical value of automated detection may therefore be greatest as a triage or second-review mechanism, particularly in laboratories where pathology workload is high. However, a high AUC does not establish safe deployment because prevalence, case mix, threshold selection and calibration influence clinical consequences.

4.4 Automated Gleason grading

Gleason grading is more difficult because the target is ordinal, morphology-dependent and affected by expert disagreement. Nagpal et al. showed that a deep-learning system trained on whole-slide prostatectomy images could outperform the mean accuracy of general pathologists relative to a

specialist reference standard [12]. Bulten et al. demonstrated diagnostic-grade automated grading on prostate biopsies [12].

The PANDA challenge strengthened the evidence base by providing a large common dataset and blinded external validation. Across US and European external cohorts, leading submitted algorithms achieved quadratic-weighted kappa values of approximately 0.86–0.87 against expert uropathologists [1]. These results demonstrate that the task is technically tractable, but they should not be interpreted as proof that every model performs similarly.

The difficulty of distinguishing adjacent patterns remains important. In a pilot study, the largest errors occurred between Gleason 3 and 4 and between 4 and 5 [8]. Independent gland-level work similarly found substantially lower discrimination for Gleason 3 versus 4 than for benign-versus-cancer discrimination [33]. This is clinically relevant because the boundary between Grade Group 1 and Grade Group 2 can affect treatment and surveillance decisions.

4.5 External validation and generalizability

External validation is the central unresolved methodological issue. Models may learn scanner-specific, staining-specific or institution-specific signals in addition to tumor morphology. Differences in tissue processing, section thickness, staining intensity, scanner optics, compression, magnification and population characteristics can therefore cause domain shift [23, 24].

The 2024 'in the wild' evaluation is particularly informative because it tested five leading public PANDA algorithms and two commercial systems on a heterogeneous crowdsourced dataset. Pathologist pairwise quadratic kappa ranged from 0.777 to 0.916, while algorithm agreement ranged from 0.617 to 0.900 [23]. The broad range demonstrates that benchmark ranking does not necessarily predict performance on heterogeneous clinical material.

Conversely, the external validation of DeepDx Prostate on 150 whole-mount prostatectomy specimens showed quadratic-weighted kappa of 0.89 for Grade Group classification, with sensitivity of 0.997 and specificity of 0.88 for cancer-versus-noncancer tile classification [2]. This provides evidence that some systems can transfer across specimen types, but it also illustrates the importance of testing the exact deployment domain rather than assuming transferability.

4.6 Human–AI collaboration

Standalone AI performance is not the only clinically meaningful endpoint. In a controlled reader study, Bulten et al. found that AI assistance increased agreement between pathologists and an expert reference standard, and AI-assisted pathologists outperformed both unassisted pathologists and the standalone AI system at the group level [14].

More recently, a prospective multicenter study published in 2026 evaluated AI assistance in 130 consecutive diagnostic biopsy cases at two Dutch hospitals. AI assistance changed Gleason Grade Group assessments in 16.5% of biopsy entries and increased pathologists' reported confidence; tumor-

detection differences were largely confined to Grade Group 1 cases [3]. This represents a meaningful transition from retrospective algorithm validation toward evaluation of AI as a clinical workflow intervention.

4.7 Failure modes

The principal failure modes, their mechanisms, potential consequences, and mitigation strategies are summarized in Table 3.

Table 3
Common failure modes in automated prostate histopathology AI

Failure mode	Mechanism	Potential consequence	Mitigation
Gleason 3 vs 4 ambiguity	Morphologic overlap	Grade Group misclassification	Expert adjudication; hard-negative testing
Artifacts/foreign material	Algorithm responds to non-tumor features	False positives/negatives	Artifact-rich external test sets
Staining/scanner shift	Different color/optical distributions	Performance degradation	Multisite training; normalization; validation
Tissue detection errors	Benign tissue or folds treated as tumor	Incorrect downstream grade	Dedicated tissue QC
Class imbalance	Rare grades underrepresented	Poor minority-class sensitivity	Balanced sampling; class-aware metrics
Data leakage	Related slides/patients cross partitions	Inflated performance	Patient-level split; independent institutions
Reference-standard noise	Expert disagreement or uncertain labels	Unstable targets	Consensus/adjudication; report uncertainty
Specimen shift	Biopsy vs prostatectomy differences	Unexpected transfer failure	Separate validation by specimen type
Shortcut learning	Model uses stain/scanner/site cues	Poor portability	Counterfactual/site-held-out testing
Overreliance by users	Automation bias	Failure to challenge incorrect output	Human factors training; confidence/uncertainty display

4.8 Data leakage and dataset design

Dataset leakage is a major concern in computational pathology. If tiles from the same patient, slide or tissue block appear in both training and test sets, the model may encounter highly correlated morphology during evaluation. This can substantially overestimate generalization. Patient-level separation should therefore be mandatory, and external testing should ideally occur at a different institution.

The same principle applies to preprocessing and stain normalization. If normalization parameters are estimated using the complete dataset before partitioning, information from the test set may indirectly enter the training process. Reporting should therefore state exactly how preprocessing, augmentation, feature extraction and threshold selection were performed.

4.9 Low-resource settings

Low-resource pathology settings may face a different problem from the one assumed in high-income benchmark studies. The limiting factor may not be model architecture but the absence of scanners, reliable networks, storage, computational hardware, validated workflows, maintenance support and trained digital-pathology personnel. A 2023 review specifically discussing AI-enabled digital pathology in low-resource organizations emphasized infrastructure, workflow and governance requirements [38]. Broader African digital-pathology literature similarly highlights uneven access to digitization and computational resources [39].

A realistic low-resource prostate AI pathway should therefore be modular: image acquisition and quality control; local or regional slide storage; efficient inference; uncertainty flagging; pathologist review; and confirmatory review for discordant or high-risk cases. Where local computing is limited, centralized inference with secure telepathology may be feasible, but data governance, connectivity and turnaround time must be considered.

Low-cost computational pathology initiatives demonstrate that open-source and locally manufacturable infrastructure can reduce hardware barriers [40]. However, reducing hardware cost does not remove the need for clinical validation, quality management, cybersecurity, maintenance and continuing performance monitoring.

4.10 Explainability and clinical trust

Explainability is increasingly incorporated into prostate AI systems, including saliency maps, attention maps and interpretable feature representations. These tools can help a pathologist understand where the algorithm concentrated its attention, but visual explanations should not automatically be interpreted as proof of causal reasoning. Explanation quality must be evaluated against pathology-relevant structures and failure cases.

Recent work has also explored AI-derived risk groupings that move beyond conventional Gleason Grade Groups. Such systems may ultimately provide additional prognostic information, but they should be evaluated separately from the narrower task of reproducing the established Gleason system [20, 21].

5. Evidence Matrix: 50 Key Papers

The following 50 papers were selected to represent the major evidence streams relevant to this review. They are not presented as a claim that only 50 papers exist. Rather, they provide a focused evidence set spanning grading, detection, external validation, human–AI collaboration, methodological comparison, risk prediction, digital pathology implementation and reporting standards. The complete focused evidence matrix is presented in Table 4.

Table 4
Evidence matrix of 50 key papers included in the focused evidence set

No.	First author / year	Evidence contribution
1	Allsbrook 2001	Interobserver reproducibility of Gleason grading
2	Egevad 2012	Implications of modified Gleason system
3	Epstein 2016	2014 ISUP consensus Gleason grading
4	van Leenders 2020	2019 ISUP consensus grading
5	Nagpal 2019	Deep learning for prostatectomy Gleason scoring
6	Kott 2021	Deep residual network for biopsy grading
7	Karimi 2020	Multiscale CNN and augmentation
8	Yang 2020	Segmentation and grading with multiphoton microscopy
9	Ström 2020	Population-based AI detection and grading
10	Bulten 2020	Automated deep learning on prostate biopsies
11	Bulten 2021	AI assistance improves pathologist grading
12	Nagpal 2020	Deep-learning biopsy grading validation
13	PANDA/Bulten 2022	Large multinational AI challenge
14	Mun 2021	Weakly supervised YAAGGS
15	Qiu 2022	Pyramid semantic parsing for grading
16	Kartasalo 2021	Status and next steps for AI grading
17	Rabilloud 2023	Systematic review of prostate digital pathology DL
18	Morozov 2023	Systematic review/meta-analysis of diagnostic accuracy
19	Dominguez-Morales 2024	Systematic comparison of DL methods
20	Classifying malignancy 2023	Independent gland-level validation
21	Schmidt 2025	External validation on prostatectomy specimens
22	Wild evaluation 2024	Public and commercial algorithms in heterogeneous data
23	Shao 2024	Digital histopathology risk stratification
24	Tward 2024	Multimodal deep-learning risk stratification
25	Federated learning 2024	Multi-site federated prostate diagnosis/grading
26	RAPHIA 2024	MRI–histopathology registration

No.	First author / year	Evidence contribution
27	Eminaga 2024	Interpretable AI-derived grading
28	Müller 2024	Comparison of 11 deep neural architectures
29	Zaheer 2025	Attention-enhanced segmentation/explainability
30	Xie 2025	Multiscale Fourier MIL
31	Kabir 2025	Classification/segmentation framework
32	van Hees 2026	Prospective multicenter AI assistance
33	Egevad 2025	Review of AI in prostate pathology
34	Ahmed 2026	Semi-supervised grading under imperfect data
35	Sica 2023	Systematic review of DL prostate pathology
36	Critical evaluation 2024	Critical assessment of AI as digital twin
37	Tissue detection 2026	Impact of tissue detection on AI
38	Zehra 2023	AI-enabled digital pathology in low-resource organizations
39	Michelow 2026	Digital diagnostics in Africa
40	Choudhury 2024	Low-cost open-source histopathology workstation
41	Page 2021	PRISMA 2020 reporting standard
42	McInnes 2018	PRISMA-DTA
43	Whiting 2011	QUADAS-2
44	Moons 2025	PROBAST + AI
45	Bossuyt 2015	STARD 2015
46	Schaafsma 2020	Clinical implementation considerations for AI pathology
47	Tizhoosh & Pantanowitz 2018	AI in digital pathology overview
48	Litjens 2017	Deep learning in medical image analysis
49	Campanella 2019	Clinical-grade weakly supervised pathology
50	Ehteshami Bejnordi 2017	Deep learning and breast histopathology benchmark methodology

6. Risk of Bias and Applicability

Across the evidence base, the most recurrent concerns are retrospective study design, single-center datasets, unclear patient-level partitioning, limited external validation, inconsistent reference standards and selective reporting of performance metrics. The systematic review literature independently identifies lack of external validation as a recurring weakness [4, 20, 21].

Applicability concerns are especially relevant when models are trained on carefully curated research images but intended for routine pathology. Real-world slides contain artifacts, variable staining, fragmented tissue, benign mimics and unusual tumor morphologies. The 2024 real-world evaluation provides direct evidence that algorithm performance can vary substantially outside benchmark settings [23].

A clinically credible study should therefore report at least: patient-level data partitioning; institution and scanner information; specimen type; prevalence and grade distribution; reference-standard construction; external validation; calibration or uncertainty; predefined thresholds; subgroup performance; failure-case analysis; and a transparent account of exclusions.

7. Discussion

The updated literature supports a nuanced conclusion. Deep learning can perform highly competitive automated prostate cancer detection and Gleason grading, particularly when trained on large, carefully curated datasets. Large-scale studies and the PANDA challenge have moved the field beyond small proof-of-concept experiments [1, 2, 13–15]. However, the evidence also demonstrates that performance depends strongly on the data distribution, validation setting and intended clinical use.

The most important development since earlier reviews is the shift toward external and real-world evaluation. The 'in the wild' study showed a broad performance range across algorithms on heterogeneous material [2], while external validation on prostatectomy specimens showed that some biopsy-trained commercial systems can generalize across specimen types [2]. These findings are not contradictory: they demonstrate that generalization is model- and context-specific and must be measured rather than assumed.

The emergence of prospective clinical studies is another important advance. The 2026 multicenter study evaluated the effect of AI assistance on actual pathologist decisions and confidence rather than simply comparing an algorithm with a static reference standard [3]. This is closer to the question that matters clinically: whether AI improves the diagnostic process without introducing unacceptable new errors.

The distinction between algorithm performance and clinical utility should therefore remain central. A model can have high accuracy but low clinical utility if it is poorly calibrated, computationally expensive, difficult to integrate into workflow, or prone to silent failure on unfamiliar cases. Conversely, a model with slightly lower standalone accuracy may be valuable if it reliably identifies suspicious cases, communicates uncertainty, and improves pathologist performance.

For low-resource settings, the deployment question is even broader. The literature suggests that AI could help extend specialist pathology expertise, but only if the digital pathway is sustainable. Scanner acquisition, data storage, electricity, connectivity, computational resources, maintenance, staff training and regulatory oversight must be considered alongside model performance [18, 19]. Recent low-resource and African digital-pathology discussions likewise emphasize infrastructure, governance, sustainability and local validation as prerequisites for safe implementation [18, 19].

8. Proposed Framework for Low-Resource Deployment

Phase 1 — Digital readiness: establish validated scanning, image-quality control, storage and secure data-transfer procedures.

Phase 2 — Local validation: test candidate models on representative local H&E slides before clinical use.

Phase 3 — External validation: evaluate performance on an independent laboratory and, where feasible, a second geographic population.

Phase 4 — Human-in-the-loop deployment: use AI as decision support, with pathologist acceptance/rejection of outputs.

Phase 5 — Safety monitoring: record discordant cases, uncertainty, failure modes, scanner changes and performance drift.

Phase 6 — Continuous improvement: retrain or recalibrate only under controlled governance with versioned datasets and independent revalidation.

9. Recommendations for Future Research

Use multicenter, geographically diverse datasets with prespecified external test sets.

Perform patient-level separation and explicitly audit for data leakage.

Report performance by Grade Group and difficult pattern pairs, particularly Pattern 3 versus Pattern 4.

Include artifact-rich and technically poor slides in robustness testing.

Report calibration, uncertainty and clinically relevant error rates in addition to accuracy and AUC.

Compare biopsy and prostatectomy performance when both are relevant to intended use.

Evaluate pathologist–AI collaboration prospectively rather than relying only on standalone model metrics.

Report scanner, staining, tissue-processing and preprocessing details sufficiently to reproduce domain-shift testing.

Develop low-cost inference pathways and measure real-world turnaround time, computational requirements and maintenance burden.

Use STARD, QUADAS-2/PROBAST + AI and PRISMA-based reporting where appropriate.

Create representative African and other underrepresented datasets rather than assuming transferability from high-income-country cohorts.

Establish governance for model versioning, monitoring, cybersecurity, patient privacy and clinical accountability.

10. Strengths and Limitations of This Review

Strengths include emphasis on external validation, real-world performance, prospective clinical evidence, failure modes and low-resource implementation rather than accuracy alone. The review also separates histopathology AI from MRI-only radiomics and explicitly avoids claiming that all AI systems are equivalent.

The principal limitation is that this updated manuscript presents a focused 50-paper evidence set rather than a newly completed exhaustive systematic search with independently duplicated screening and a registered protocol. Consequently, no fabricated PRISMA flow counts or pooled estimates are presented. For formal systematic-review publication, the next step should be a fully reproducible database search, duplicate screening, study-level extraction and formal risk-of-bias assessment.

11. Conclusion

Deep learning has established a strong technical basis for automated prostate cancer detection and Gleason grading from histopathology images. The strongest evidence comes from large datasets, independent validation and studies that directly compare AI with expert or routine pathologist performance. Nevertheless, generalizability remains a central challenge. Domain shift, artifacts, difficult grade boundaries, specimen differences, reference-standard variability and dataset leakage can materially affect performance. The newest evidence, including prospective multicenter evaluation, supports viewing AI primarily as a clinical decision-support technology whose value must be demonstrated in the workflow in which it will be used.

For low-resource settings, the most defensible pathway is not direct deployment of an externally developed model based solely on published accuracy. Instead, implementation should proceed through local validation, human-in-the-loop use, quality monitoring and infrastructure-aware design. Future research should prioritize transparent, multicenter, prospective and geographically diverse studies capable of demonstrating safety, robustness and clinical utility.

Declarations

Ethics approval and ethical considerations

This review did not require institutional research ethics approval because it did not involve recruitment of human participants, collection of new human specimens, access to identifiable patient-level clinical records, or experimental intervention.

Consent for publication

Not applicable. This review did not involve the collection or publication of identifiable information from individual participants.

Funding

No specific external financial or non-financial funding was received for the preparation of this systematic review.

Competing interests

The authors declare that they have no financial or non-financial competing interests that could have influenced the conduct, interpretation, or reporting of this review.

Data availability

No primary dataset was generated or analyzed for this review. The evidence base was derived from published scientific literature and publicly accessible bibliographic and publisher records.

Author contributions

Dr. Habert Tumwesigye^{1*}: Conceptualization, methodology, literature synthesis, evidence interpretation, manuscript drafting, critical revision, supervision, and final approval of the manuscript.

Phionah Nabimanya¹: Literature identification and screening support, evidence extraction, manuscript review, and final approval.

Birungi Abraham¹: Literature identification and evidence extraction support, interpretation of findings, manuscript review, and final approval.

Dr. Samuel Tumwesigire¹: Clinical/pathological interpretation, critical review of the manuscript, and final approval.

Dr. Yekosani Mitala¹: Methodological and scientific review, interpretation of findings, manuscript revision, and final approval.

Dr. Raymond Atwine¹: Pathology and clinical interpretation, critical manuscript review, and final approval.

All authors approved the final manuscript and agree to be accountable for the work.

References

1. Bulten W, Kartasalo K, Chen PHC et al (2022) Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat Med* 28(1):154–163. 10.1038/s41591-021-01620-2
2. Schmidt B, Soerensen SJC, Bhambhani HP et al (2025) External validation of an artificial intelligence model for Gleason grading of prostate cancer on prostatectomy specimens. *BJU Int* 135(1):133–139. 10.1111/bju.16464
3. van Hees JE, van Diest PJ, Nguyen TQ et al (2026) Influence of Artificial Intelligence Assistance on Gleason Grading and Prostate Cancer Detection by Urologists in Daily Practice: A Prospective Multicenter Study. *JCO Clin Cancer Inf* 10:e2500352. 10.1200/CCI-25-00352
4. Rabilloud N, Allaume P, Acosta O et al (2023) Deep Learning Methodologies Applied to Digital Pathology in Prostate Cancer: A Systematic Review. *Diagnostics (Basel)* 13(16):2676. 10.3390/diagnostics13162676
5. Page MJ, McKenzie JE, Bossuyt PM et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372:n71. 10.1136/bmj.n71
6. McInnes MDF, Moher D, Thoms BD et al (2018) PRISMA-DTA: Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies. *JAMA* 319(4):388–396. 10.1001/jama.2017.19163
7. Whiting PF, Rutjes AWS, Westwood ME et al (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med* 155(8):529–536. 10.7326/0003-4819-155-8-201110180-00009
8. Moons KGM, Damen JAA, Kaul T et al (2025) PROBAST + AI: an updated guidance for assessment of risk of bias and applicability of prediction models. *BMJ* 388:e082505. 10.1136/bmj-2024-082505
9. Bossuyt PM, Reitsma JB, Bruns DE et al (2015) STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 351:h5527. 10.1136/bmj.h5527
10. Epstein JI, Egevad L, Amin MB et al (2016) The 2014 ISUP Consensus Conference on Gleason Grading of Prostatic Carcinoma. *Am J Surg Pathol.* ;40(2):244–252
11. van Leenders GJLH, van der Kwast TH, Grignon DJ et al (2020) The 2019 ISUP Consensus Conference on Grading of Prostatic Carcinoma. *Am J Surg Pathol.* ;44(11):e87-e99
12. Nagpal K, Foote D, Liu Y et al (2019) Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med* 2:48. 10.1038/s41746-019-0112-2
13. Bulten W, Pinckaers H, van Boven H et al (2020) Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *Lancet Oncol* 21(2):233–241. 10.1016/S1470-2045(19)30739-9
14. Bulten W, Balkenhol M, Awoumou Belinga JJ et al (2021) Artificial intelligence assistance significantly improves Gleason grading of prostate biopsies by pathologists. *Mod Pathol* 34(3):660–671. 10.1038/s41379-020-0640-y

15. Ström P, Kartasalo K, Olsson H et al (2020) Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol* 21(2):222–232. 10.1016/S1470-2045(19)30738-7
16. Mun Y, Paik I, Shin SJ et al (2021) Yet Another Automated Gleason Grading System (YAAGGS) by weakly supervised deep learning. *NPJ Digit Med* 4:99. 10.1038/s41746-021-00469-6
17. Karimi D, Nir G, Fazli L et al (2020) Deep Learning-Based Gleason Grading of Prostate Cancer From Histopathology Images—Role of Multiscale Decision Aggregation and Data Augmentation. *IEEE J Biomed Health Inf* 24(5):1413–1426. 10.1109/JBHI.2019.2944643
18. Zehra T, Parwani A, Abdul-Ghafar J, Ahmad Z (2023) A suggested way forward for adoption of AI-enabled digital pathology in low resource organizations in the developing world. *Diagn Pathol* 18:68. 10.1186/s13000-023-01352-6
19. Choudhury D, Dolezal JM, Dyer E et al (2024) Developing a low-cost, open-source, locally manufactured workstation and computational pipeline for automated histopathology evaluation using deep learning. *EBioMedicine* 107:105276. 10.1016/j.ebiom.2024.105276
20. Frewing A, Gibson AB, Robertson R, Urie PM, Della Corte D (2024) Don't Fear the Artificial Intelligence: A Systematic Review of Machine Learning for Prostate Cancer Detection in Pathology. *Arch Pathol Lab Med* 148(5):603–612. 10.5858/arpa.2022-0460-RA
21. Morozov A, Taratkin M, Bazarkin A et al (2023) A systematic review and meta-analysis of artificial intelligence diagnostic accuracy in prostate cancer histology identification and grading. *Prostate Cancer Prostatic Dis* 26(4):681–692. 10.1038/s41391-023-00673-3
22. Dominguez-Morales JP, Duran-Lopez L, Marini N et al (2024) A systematic comparison of deep learning methods for Gleason grading and scoring. *Med Image Anal* 95:103191. 10.1016/j.media.2024.103191
23. Faryna K, Tessier L, Retamero J et al (2024) Evaluation of Artificial Intelligence-Based Gleason Grading Algorithms 'in the Wild'. *Mod Pathol* 37(11):100563. 10.1016/j.modpat.2024.100563
24. Fuhlert P, Westhaeusser F, Dietrich E et al (2026) A systematic analysis of the impact of data variation on AI-based histopathological grading of prostate cancer. *Med Image Anal* 108:103884. 10.1016/j.media.2025.103884