Artificial intelligence in digital pathology diagnosis and analysis: Technologies, clinical integration, and future prospects (Review)

ABDUL-MOHSEN ALHEJAILY[1] and DOAA ALGHAMDI[2]

[1]Scientific Publication Support Unit and Academic Operations Administration, King Fahad Medical City, Riyadh Second Health Cluster, Riyadh 11525, Saudi Arabia; [2]Department of Anatomical Pathology, King Fahad Medical City, Riyadh Second Health Cluster, Riyadh 11525, Saudi Arabia

Received February 16, 2026; Accepted May 26, 2026

DOI: 10.3892/br.2026.2182

Abstract. The integration of artificial intelligence (AI) into digital pathology is perhaps the most revolutionary leap forward in modern diagnostic medicine. The present review analyzes the existing context of AI pathology systems, particularly diagnostic precision, clinical validation, and technical systems such as convolutional neural networks and transformers and discusses the integration challenge in clinical workflows for these systems. AI systems have achieved pathologist-level performance in controlled settings, including diagnostic accuracy >99% and area under the receiver operating characteristic curve values exceeding 0.97. However, translating research into clinical adoption is riddled with several challenges attributable to computational requirements, data standardization issues, regulatory hurdles and limitations in generalizability. Moreover, Vision Transformers are widely popular as powerful alternatives to conventional convolutional models, delivering high performance in certain domains while also imposing a novel computational burden. Overcoming these challenges is a prerequisite for the successful integration of AI into pathology practice and the realization of its full diagnostic potential.

Contents

Correspondence to: Dr Abdul-Mohsen Alhejaily, Scientific Publication Support Unit and Academic Operations Administration, King Fahad Medical City, Riyadh Second Health Cluster, Aldabbab Street, Riyadh 11525, Saudi Arabia
E-mail: aalhejaily@kfmc.med.sa

Key words: pathology, artificial intelligence, area under the receiver operating characteristic curve, clinical decision support system, diagnostic accuracy, convolutional neural networks, deep learning

1. Introduction

Over the last two decades, digital pathology has undergone an extraordinary transformation from a marginal research area to a clinically valuable practice with the potential to transform diagnostic medicine (1). The digitalization of histopathology slides into whole-slide images, aided by advances in deep learning (DL), presents novel avenues for automated analysis, quantitative diagnosis, and decision support in pathology practice (2). These developments come amid mounting challenges for pathology departments worldwide, including increased work volume, the necessity for more accurate diagnoses, and ongoing issues reconciling interpretations across institutions (3). DL-enabled artificial intelligence (AI) systems have demonstrated impressive performance on fine-grained gigapixel histopathology images, revealing subtle visual patterns that might be difficult to read or even unacceptable for experienced pathologists (4). The detection of cancer metastases in lymph nodes, or even the identification of molecular biomarkers directly from tissue morphology, are among the potential applications in which AI systems can augment human expertise, minimize diagnostic variability, and enable novel forms of computational pathology (5).

The domain evolved from proof-of-concept studies to mass multi-institutional validation studies, and the first AI systems were approved for clinical applications (6,7). However, the transition from algorithmic innovation to clinical application is complex. Despite the significant performance noted in literature, using these results in everyday diagnostic processes requires addressing basic questions about efficient computing, generalization across a wide range of patients and technologies, regulatory requirements, and interoperability with current clinically supported systems (7). Furthermore, the rapid progress of AI architectures from classical convolutional neural networks (CNNs) to transformer-based architectures has raised new concerns regarding computational efficiency, interpretability, and practical implementation (8). The present review evaluates the state of AI in digital pathology across three critical dimensions: Diagnostic accuracy and clinical

validation evidence supporting AI applications, the technical architectures supporting them, and real-world hurdles to their successful clinical implementation (Fig. 1).

2. Clinical validation and diagnostic accuracy

Performance metrics across cancer types. The diagnostic performance of AI in digital pathology has been well studied across various cancer types and the field has been observed to advance rapidly in tackling challenges. In renal cell carcinoma subtyping, numerous studies have shown the highly reproducible performance of AI, with all major architectural methods exhibiting area under the receiver operating characteristic curve (AUROC) values >0.90 in cross-validation and external validation conditions (9-11). This consistency indicates that histological tumor subtyping is a well-solved problem, for which AI systems can reliably distinguish morphological patterns. However, performance varies widely for more complex tasks, such as mutation prediction. In colorectal cancer, AUROC values of 0.937 and 0.919 were obtained using Vision Transformer (ViT)- and Residual Network (ResNet)-based workflows for microsatellite instability (MSI) prediction, respectively, which outperformed multiple-instance learning methods (9). This was confirmed elsewhere by external validation on large-scale cohorts, such as the Cancer Genome Atlas datasets, and using one full transformer-based pipeline with a sensitivity of 0.97, an MSI prediction negative predictive value of 0.99 for surgical resection, and a performance down to a mean AUROC of 0.91 in biopsy samples (12). These findings illustrate that sample quality and tissue type largely influence AI performance.

For prostate cancer, AI systems have previously achieved pathologist-level performance in cross-country cohort studies. A large validation project involving almost 100,000 digitized core needle biopsies from 7,342 patients in 15 laboratories across 11 countries demonstrated that the readings of AI models are highly generalizable and sensitivity-based setups identified all malignant slides without false negatives while reducing immunohistochemistry utilization by 44.4% (11,13). Regarding Gleason pattern classification, self-supervised ViT models achieved a kappa (κ) value of 0.841 overall, and external validation scores across large datasets ranged from 0.774 to 0.888 (13). Another area where AI has demonstrated great potential is in diagnosing lymphoma. A DL model for diffuse large B-cell lymphoma achieved at least 100% diagnostic accuracy at two hospitals and at least 99.71% at a third hospital, with 100% sensitivity (14). Notably, this system performed comparatively higher than human pathologists, with only a 74.39% accuracy. The accuracy of cross-hospital testing was previously reported to fall from 90.50 to 82.09%, attributed to technical differences, and after correcting for these differences, it reached 100% accuracy, highlighting the utmost significance of data standardization. For the detection of metastasis, systems based on AI have exceeded all previous benchmarks for image AI. For example, a CNN-based framework achieved 92.4% sensitivity with eight false positives per image on the Camelyon16 dataset, significantly improved than the previous best automated approach (82.7%) and human pathologists (73.2%) (15). Image-level AUC scores exceeded 97%, indicating that the model performed well on gigapixel images (Table I).

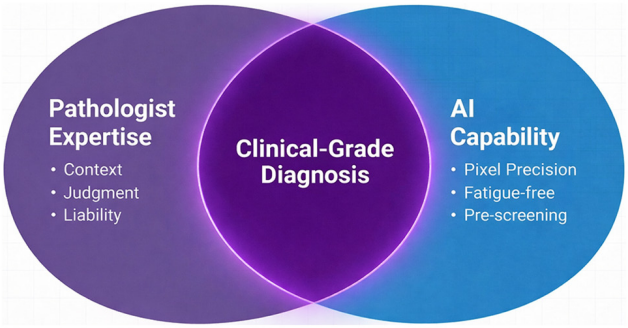

Figure 1. Venn diagram illustrates the synergistic collaboration between human expertise and artificial intelligence in medicine. By merging pathologist nuanced judgment, contextual understanding, and professional liability with pixel-precision, fatigue-free performance, and the ability to run pre-screening, the standard of clinical-grade diagnosis can be raised to a whole new level. AI, artificial intelligence.

Taken together, these findings reveal a clear performance gradient that maps closely to task complexity rather than to cancer type per se. Morphology-driven subtyping of well-characterized entities, such as renal cell carcinoma subtyping or detection of overt tumor in lymph nodes, behaves as a largely solved problem, with AUROC values consistently exceeding 0.90 and approaching 0.98 in external validation, irrespective of architectural choice (9,15,16). Slide-level diagnostic tasks with binary or near-binary outputs, such as benign-vs.-malignant prostate classification or large B-cell lymphoma identification, similarly cluster at the upper end of the performance spectrum, with sensitivities at or near 100% in optimized pipelines (11,13). By contrast, tasks that require inferring molecular or genetic features from morphology alone, exemplified by MSI prediction in colorectal cancer, exhibit a much wider performance band (AUROC, 0.91-0.97) that is highly sensitive to tissue type and sample quality: Resection specimens consistently outperform biopsies, and full transformer pipelines outperform multiple-instance learning by several percentage points within the same cohort (9,10). Fine-grained grading tasks, such as Gleason pattern classification, occupy an intermediate tier (κ=0.77-0.89), reflecting the inherent inter-observer variability of the reference standard itself (12). This gradient has direct clinical implications: AI is now arguably ready for autonomous or near-autonomous deployment in well-defined screening and triage tasks, whereas mutation prediction and grading remain decision-support applications whose performance must be interpreted in light of preanalytical variables and the specific tissue source.

Multi-center validation and generalization. Generalizing AI models across institutions, scanners, and patient populations is a key validation challenge. However, large-scale multicenter studies involving over 9,000 patients across 10 colorectal cancer cohorts demonstrate that appropriately designed transformer-based biomarker prediction models are stable across multiple settings (10). Likewise, a prostate cancer validation study conducted across 15 laboratories

Table I. Key milestones of AI in pathology.

Year	Innovation	Significance
2017	Philips IntelliSite Pathology Solution	First whole slide image system FDA-cleared for primary diagnosis, establishing digital pathology as regulated.
2021	Paige prostate	Initial AI software for prostate cancer detection, integrating AI directly into diagnostics.
2025	ArteraAI prostate	First multimodal AI combining images and clinical data for prognosis, expanding capabilities.
2025	PathAI AIM-MASH	FDA-qualified AI for MASH biomarkers in trials, advancing AI in drug development.
2025	Indica labs HALO AP Dx	FDA clearance for enterprise digital pathology platform, enhancing workflow integration.
2026	PathPresenter Clinical Viewer	510(k) clearance for primary diagnosis with specific scanners, boosting vendor-agnostic tools.

AI, artificial intelligence; FDA, Food and Drug Administration; MASH, metabolic dysfunction-associated steatohepatitis.

in 11 countries showed that appropriate preprocessing can lead to clinically relevant generalization of AI models (11). Nevertheless, these successful outcomes require careful attention to the technical specifications. Another study of domain generalization in lymph node segmentation reported that strategic sampling achieved 0.93 sensitivity and 0.90 Dice similarity, compared with 0.82 sensitivity and 0.83 Dice for random sampling (16). External validation investigations have frequently shown that transferability is challenging. Employing multicentre data, a study of AI methods and their inter-institutional transferability demonstrated that stain normalization improved performance by 13% in CNNs and a 10% increase in ViT performance in applied settings (17).

Comparison with human pathologists. In-depth benchmarking between AI systems and human pathologists remains essential for assessing clinical relevance. In prostate cancer diagnosis, AI models demonstrated performance similar to an independent pathologist (12), where a κ score of 0.752 was obtained in a study that compared an AI system and a pathologist on an external dataset to classify four Gleason pattern classes. The evidence of AI superiority was most impressive in the lymphoma diagnosis study with an AI platform at 100% accuracy, while pathologists reached 74.39% (14). In metastasis detection, the AI system achieved 92.4% sensitivity, outperforming human pathologists by 19.2% (15) (Fig. 2). Higher accuracy in diagnoses has significant clinical relevance, considering the grave consequences of missed metastases, with major implications for staging and treatment. Of particular relevance is a key finding from the clinical validation of Paige Prostate, which received the first U.S. FDA authorization of an AI system for pathology, showing that pathologists achieved improved screening and selection performance when using the AI-augmented system (18). A total of 18 pathologists reviewed 610 whole-slide images of prostate needle core biopsies from 218 institutions. Their analysis revealed an effective framework in which AI enhances human expertise rather than replacing it (Fig. 3).

3. Technological environments: From CNNs to transformers

CNN design methods. The principles behind using CNNs are based on their ability to capture hierarchical spatial features in images. ResNet architectures, particularly ResNet18 and ResNet50, have been extensively deployed owing to their balance of performance and computational efficiency. In extensive benchmarking studies, ResNet-based workflows performed well, achieving AUROC values above 0.90 for tumor subtyping and 0.919 for MSI prediction (9,19). More broadly, the optimized network depth-width-resolution scaling in EfficientNet architectures shows considerable potential. EfficientNet-b7 achieved a single state-of-the-art AUROC of 0.983 for clear cell renal cell carcinoma detection in external validation, a preferred result for this problem (19).

The ability of the EfficientNet family to maintain performance over time with fewer parameters makes it attractive for deployment in resource-limited settings. Ensemble methods with multiple CNN architectures show improved robustness. After resolving technical discrepancies, a transfer-learning platform was globally optimized (14), comprising 17 pretrained CNNs, including AlexNet, GoogleNet, ResNet variants, VGG networks, Inception, DenseNet, MobileNet, and more, achieving 100% diagnostic performance for lymphoma (14). However, this ensemble strategy leverages the complementary benefits of diverse architectures, albeit at a higher computational cost. As for hybrid architectures, the combination of CNNs and other components has emerged as one of the most promising directions. A ResNet50-based hybrid vision model incorporating a ViT module achieved 99% classification accuracy for prostate cancer histologic assessment (20). This architecture uses the ResNet50 backbone (for local feature extraction). The ViT module can capture long-range dependencies through multi-head self-attention.

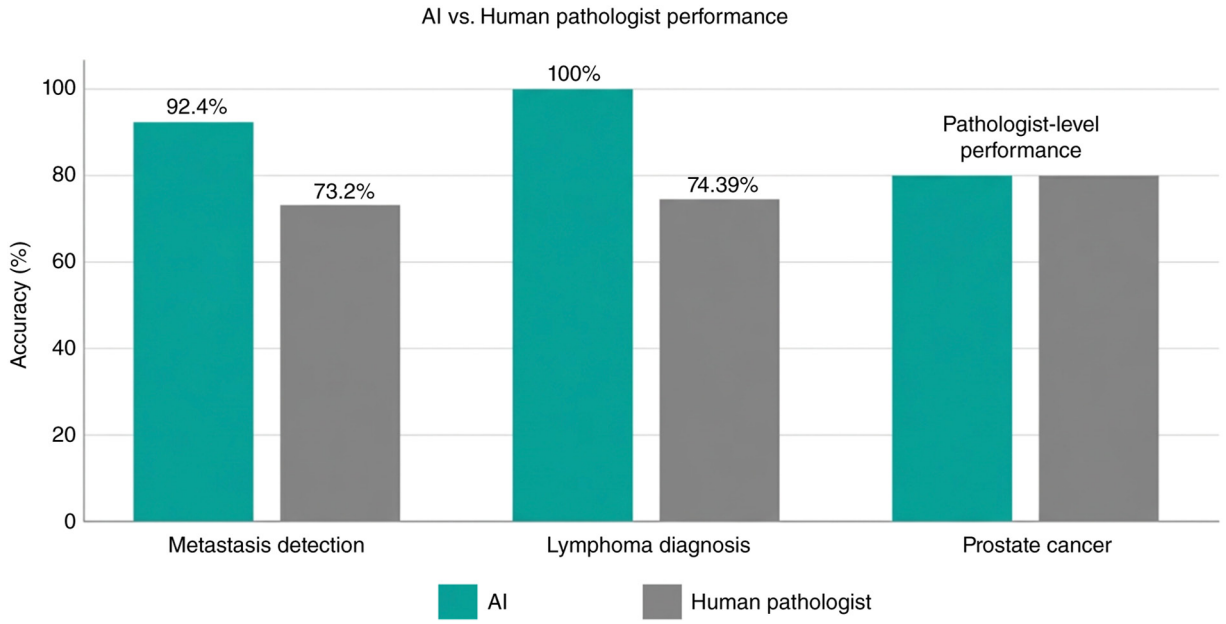

Figure 2. Illustration of the comparative diagnostic performance between AI systems and human pathologists across different oncological tasks, including metastasis detection, lymphoma diagnosis, and prostate cancer. AI systems achieved a sensitivity that significantly outperformed human pathologists in the first two cases and was comparable to that of pathologists in prostate cancer. Diagnostic Performance of AI Systems vs. Human Pathologists: The chart shows that AI can match or exceed human accuracy in important diagnostic tasks, such as detecting metastasis (92.4% vs. 73.2% sensitivity), diagnosing lymphoma (up to 100% vs. 74.39% accuracy), and diagnosing prostate cancer (pathologist-level performance with κ scores up to 0.752). AI, artificial intelligence.

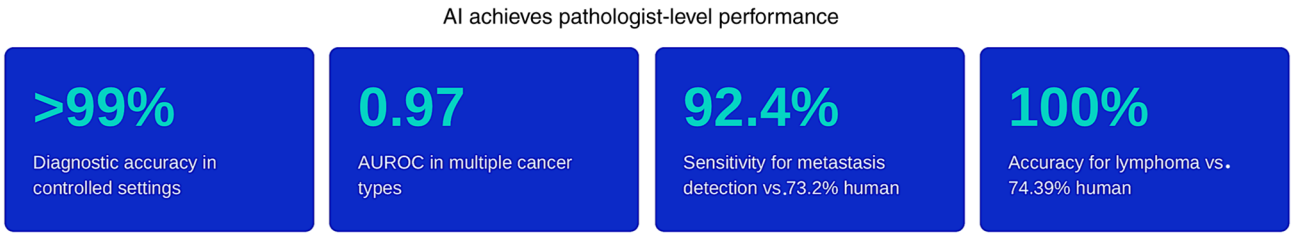

Figure 3. Representative results from several key AI-powered diagnostic studies in cancer management are displayed. Notably, >99% accurate diagnostics were observed *in vitro* in one context, AUROC was measured as 0.97 with six types of cancer involved, an AI model reached a sensitivity level of 92.4% for detecting metastasis, whereas pathologists only identified 73.2% on average. Lastly, the AI model distinguished lymphoma subtypes with 100% accuracy in that context compared with 74.39% accuracy for human counterparts. AI generally provides superior diagnostic outcomes relative to pathologists in a variety of cancer detection challenges. AI, artificial intelligence; AUROC, area under the receiver operating characteristic curve.

ViT architectures. Despite these recent developments, ViTs quickly emerged as strong competitors to convolutional architectures and offered theoretical benefits, such as global receptive fields and the ability to model long-range dependencies. Benchmarking studies demonstrated that ViT-based workflows yield predictions that are comparably accurate as those of CNN-based methods, resulting in an AUROC of 0.937 for MSI predictions in colorectal cancer (9). An external validation test showed a 1.5% higher accuracy for ViTs than for CNNs (17). Self-supervised ViT architectures are particularly promising. For prostate cancer diagnosis, a self-supervised ViT structure provided κ scores (0.967) for internal validation of a benign vs. a malignant model in tissue classification, and κ values for external validation ranged from 0.876 to 0.995 across four new datasets (12). These models, based on a self-supervised pretraining strategy, can be trained on non-annotated histopathology images and learn rich representations, thereby alleviating the limitations of the annotated training dataset to a greater extent. Swin Transformers, which leverage hierarchical feature maps and shifted windowing schemes for high-performance applications, have been successfully implemented for pathology problems.

The Swin Transformer-based CTransPath feature extractor combines hierarchical CNN architectures with global self-attention, achieving top performance in biomarker prediction (10). In recent years, key trends in transformer architectures for pathology have been innovative. The Cascaded Cross-Attention Network solves the quadratic scaling challenge of traditional transformers by introducing a cascaded cross-attention mechanism that scales linearly with the number of input patches (10). Notably, this model yielded mean AUC values of 0.970 for non-small cell lung cancer and 0.985 for renal cell carcinoma classification, respectively, and demonstrated robust performance in low-data regimes. For cell-level analysis, CellViT employs a U-Net-shaped encoder-decoder architecture with ViTs as the encoder networks (21). This approach achieved a mean panoptic quality of 0.51 and an F1-detection score of 0.83 on the

PanNuke dataset and generalized across datasets. In addition, CellNuc-DETR uses hierarchical undercutting with a Swin Transformer and multi-scale deformable transformers rather than segmentation, thereby reducing computational costs and enabling open-source detection (22).

Novel architectures and MI learning. Beyond pure CNN and transformer algorithms, new architectures have emerged. Among the proposed novel architectures, the Sequencer architecture, featuring the interplay between horizontal and vertical bidirectional long short-term memory (LSTM) networks in a BiLSTM2D module, achieved higher performance than the CNN and the standard Transformer in biomarker prediction (23). Cross-cohort validation of this architecture yielded higher AUROC values than those of ResNet, EfficientNet, ViT, and Swin Transformer, while using fewer computational resources.

In recent years, novel but interesting approaches to natural language processing have emerged in the market. NLP models were compared to improve biomarker predictions in colorectal cancer using benchmark dataset, demonstrating improvements in prediction by up to 10% and precision by up to 26% (24). The Sequencer2D model was particularly promising, suggesting that bidirectional LSTM architectures are better suited for capturing sequential dependencies that are overlooked by purely spatial convolutions or attention mechanisms. With thousands of unlabeled patches, the entire slide-image problem has been addressed, making Multiple Instance Learning (MIL) a dominant approach for whole-slide image labeling at the slide level. Nevertheless, extensive benchmarking indicated that classical MIL and even attention-based MIL variants (such as CLAM) are frequently outperformed by end-to-end weakly supervised methods using ResNet- or ViT-based architectures (9,19). The enhanced performance of end-to-end approaches appears to depend on accurately providing high prediction scores to highly informative image regions with reasonable histopathological features (Table II).

4. Clinical workflow integration challenges

Computational needs and scalability. Computational needs and scalability are addressed through novel architectural solutions. The cascaded cross-attention network scales linearly rather than quadratically with the number of patches, resulting in lower computational requirements without compromising performance (10). CellNuc-DETR improves computational efficiency by prioritizing detection over segmentation; distributed processing and lower-resolution inference further enhance its scalability for clinical applications (22).

Beyond architectural optimization, the practical implications of clinical-grade deployment are substantial and frequently underestimated. A single whole-slide image at a magnification of x40 typically occupies 1-4 GB and contains tens of thousands of tiles, so inference at clinically acceptable turnaround times (minutes per case) effectively requires Graphics Processing Unit (GPU) acceleration. Central Processing Unit (CPU)-only inference for a transformer-based model can take 10-30 times longer than GPU inference, rendering it unsuitable for real-time reporting in busy laboratories. For hospitals, this translates into concrete infrastructure decisions. An enterprise-grade GPU server suitable for routine diagnostic deployment generally costs in the range of US $30,000-$80,000 in capital expenditure, with additional ongoing costs for power, cooling, secure storage of multi-petabyte slide archives, and dedicated information technology staff. Cloud-based inference offers an alternative, but introduces recurring per-case costs and raises non-trivial concerns about patient data residency, network bandwidth (whole-slide images may require gigabit connectivity to upload at clinically useful speeds), and compliance with regulations such as the Health Insurance Portability and Accountability Act and the General Data Protection Regulation. These logistical and economic constraints help explain why even validated AI tools have diffused slowly into routine practice, and why community and resource-limited hospitals face a structurally different adoption curve than well-funded academic centers.

Closely related to deployment is the persistent challenge of interpretability. The opacity of deep neural networks remains a principal barrier to clinical trust, and a small but maturing toolkit of explainability methods has emerged to address it. Attention maps, derived natively from transformer-based architectures, indicate which image regions the model weighted most heavily when generating a prediction, and have been used in slide-level pipelines such as CTransPath-based biomarker prediction pipelines to localize tumor and stromal regions of interest (10). Class activation mapping and gradient-weighted class activation mapping (Grad-CAM), originally developed for convolutional networks, generate coarse heatmaps that highlight the discriminative regions for a given output class and are now routinely overlaid on whole-slide images to support pathologist review. More fine-grained methods, such as saliency maps and integrated gradients, attribute pixel-level contributions to a prediction, while occlusion-based and Shapley additive explanation approaches probe model behavior by systematically perturbing inputs. In MIL frameworks, attention-pooling weights themselves serve as a form of built-in interpretability by ranking the contribution of individual tiles to the slide-level decision (25,26). Despite this progress, important caveats remain. Heatmap-based explanations can be unstable across model retraining, may align poorly with clinically meaningful features, and risk providing spurious post hoc rationalizations rather than genuine causal insight. Robust validation of explanations against pathologist reasoning, ideally through prospective reader studies, remains an open and pressing research need.

Standardization and generalization of data. In translating AI models from research to clinical practice, data standardization remains a widespread challenge. Transitional variables from slide preparation, staining protocols, and image acquisition significantly limit the performance of AI models in cross-institutional environments. This was especially illustrated in a study on lymphoma diagnosis, in which the accuracy of cross-hospital tests dropped from near-perfect performance to 82-91% due to technical differences (13). In particular, scanner-induced color variability is problematic. Images from different scanner vendors, scanners, and models have different color properties. Physical color calibration has been reported to mitigate this heterogeneity more effectively than computational color normalization based on color-related

Table II. Key metrics and findings for AI models in cancer diagnosis.

Cancer type	Task/Application	AI model architecture	Performance metric and value	Human comparison	Key findings	(Refs.)
Renal Cell	Subtyping and detection	ResNet and EfficientNet-b7	EfficientNet AUROC, 0.983	Not in the source	Consistent subtyping; SOTA results.	(10)
Colorectal	MSI prediction	ViT and Res Net-based	ViT AUROC, 0.937	Not in the source	ViT outperformed ResNet and MIL methods.	(12)
Prostate	Gleason grading and malignancy	Self-supervised ViT	κ, 0.841; sensitivity, 100%	Pathologist κ, 0.752	Pathologist-level; reduced IHC use by 44.4%.	(14)
Lymphoma	Diagnostic accuracy	Ensemble of 17 CNNs	Accuracy, 100%; sensitivity, 100%	Pathologist accuracy, 74.39%	Significantly outperformed pathologists.	(16)
Lung (NSCLC)	Classification	Cascaded cross-attention	Mean AUROC, 0.970	Not in the source	Linear scaling with input patches.	(22)
Breast/ Colorectal	Metastasis detection	CNN-based framework	AUROC, >0.97; sensitivity, 92.4%	Pathologist sensitivity, 73.2%	Outperformed automated and human methods.	(19)
Prostate	Histologic assessment	Hybrid (ResNet50 + ViT)	Accuracy, 99%	Not in the source	Combines local and global dependencies.	(21)

AI, artificial intelligence; ViT, vision transformer; CNN, convolutional neural network; AUROC, area under the receiver operating characteristic curve; IHC, immunohistochemistry; MSI, microsatellite instability; MIL, multiple instance learning; SOTA, State-Of-The-Art; NSCLC, non-small cell lung cancer; ResNet, residual network.

data (11). But physical calibration can be expensive and may require additional equipment and procedures, with limitations to their everyday use. Staining technique variations across laboratories pose major challenges for accurate diagnoses. Stain normalization techniques, such as the Macenko method, are also established as standard preprocessing steps, achieving 13% greater inter-institutional transferability for CNNs and 10% for ViTs (17).

Regulatory approval. Regulatory approvals remain a major milestone for the transition to clinical use. Paige Prostate, an AI application in pathology, was the first to receive U.S. FDA authorization (18). However, the validation process involves rigorous assessment of the effectiveness and safety of the system, including careful evaluation of the various failure modes and their clinical implications. The design of AI system-based frameworks does not automatically align with current regulatory frameworks, leading to difficulties with evaluation and approval. Unlike traditional medical devices, with their usage restricted to specific medical contexts, AI systems can be updated over the long term, enabling improvements while raising questions about whether and how those updates should be regulated. For biomarker prediction algorithms, regulatory approval is an additional hurdle.

Currently, commercial products for MSI detection are approved only for surgical resection tissue, not for biopsy material (10). The slow and sporadic implementation of precision oncology biomarkers is partly due to the complexity, expense, and technical requirements associated with regulatory-compliant validation and sophisticated instrumentation (10). A dearth of studies investigating generalizability to other laboratories, scanners, patient samples, patient populations, and complex cases has hampered the practical application of AI (11,27).

AI-powered precision in pathology. The successful integration of AI into pathology will likely depend on its adoption by pathologists and on the continued development of robust paradigm work using humans and AI. Acceptance of AI by pathologists is driven by several closely related factors, including trust in the technology, knowledge of its applicability and limitations, effects on workflow performance, and professional autonomy and liability. Current AI algorithms generally require clinical supervision and pre-screening to focus on the most relevant aspects rather than operate as fully autonomous diagnostic processes (16). This collaborative model, in which AI supports decision-making rather than replacing the expertise human clinicians, appears to be the most clinically

viable approach. The Paige Prostate validation confirmed that pathologists showed improved diagnostic accuracy when aided by AI (18). However, implementing AI assistance into existing workflows involves careful interface design with training.

Pathologists are required to be aware of the areas of trustworthy AI recommendations, which require transparency about the confidence of the system and the rationale for its predictions. Interpretability is a major concern for the clinical acceptance of AI systems. Conventional algorithms are often limited in interpretability, making it difficult to characterize novel human-interpretable biomarkers (21). Integration into the workflow also needs to be based on practical realities, including how AI predictions are issued, presented in pathology reports, and integrated with laboratory information systems. The CellViT framework also proposed an approach to interoperability problems by exporting predictions in JSON format for compatibility with software tools such as QuPath (10,21).

5. Discussion and future trends

AI for digital pathology has now reached sufficient technical maturity based on the study findings to achieve or even exceed the performance of specialist human pathologists at clinical work points. Notably, however, there remains a significant gap between experimental evidence demonstrating this potential and routine clinical implementation. Several limitations, including computational feasibility, generalization, adherence to governance, and integration into existing workflow processes, need to be addressed. No particular architecture has demonstrated definitive superiority.

CNN-based architectures, long the workhorse of vision tasks, remain effective, with efficient designs such as EfficientNet achieving strong accuracy at substantially reduced computational cost through compound scaling of network depth, width, and resolution. More recently, Vision Transformer (ViT) architectures have demonstrated comparable or superior performance across a wide range of problems by modeling long-range spatial dependencies through self-attention rather than the local receptive fields characteristic of convolution. In practice, the choice between these families is rarely absolute: CNNs retain advantages in data efficiency and inductive bias on smaller datasets, whereas ViTs tend to scale more favorably when large volumes of training data are available. This complementarity has inspired interest in hybrid approaches that combine convolutional feature extraction with attention-based representation, aiming to capture the strength of both paradigms.

Nevertheless, it is difficult to say which architectures work best because their performance is context-dependent and depends on the underlying problem, such as dataset size, nature of pathological structures and the clinical goal of analysis. Sequence-based methods such as Perceiver, transformer-based feature approaches (such as TransPath features), more task-specific models such as those reported by Xu *et al* (27), and models developed at Google all reflect a broader shift toward designing tools tailored to specific tasks rather than a single general-purpose vision model.

To be relevant for clinical application, however, models require excellent generalization capabilities. Even with state-of-the-art computer vision techniques, differences in staining, scanner technology or slide preparation lead to large differences in model performance when transferred to new datasets. The highest performance on cross-institutional datasets has been achieved with normalization methods for staining and calibration of color, which improve model transfer capabilities. A major issue in moving AI into practice, then, is how to ensure that models maintain optimal performance in clinical settings that cannot be perfectly replicated during testing, but can only be approximated through statistical variation. Research into this must move into rigorous clinical data with real variations in scanners, preparations, staining procedures, and not just test models with controlled variation or within the limitations of artificial datasets.

The development of foundation models built by training on large, diverse datasets of histopathology slide images from millions of slides collected from different hospitals, clinics and centers, represents a new and exciting area of development. These foundation models can then be adapted to multiple downstream tasks using fine-tuning with relatively small datasets of slides manually labeled for the given purpose. A promising area is that of multimodal AI models, where histopathology images are analyzed together with gene expression data, radiology scans and other clinical patient information. Integration of data from various modalities will likely lead to more precise and robust predictions for patient diagnosis, prognosis, treatment response, and disease progression.

AI does not replace pathologists but instead redeploys them by performing automated, routine tasks such as cell counting and screening. This frees them up for the more challenging aspects of medicine, such as identifying novel markers, providing differential diagnoses, integrating pathology information with other clinical data from radiography and genetics for prognostication and contributing more deeply to patient-focused clinical decisions. However, there are instances where pathologists will always be necessary to serve as a check, especially where models might err in rare cases.

Beyond just performance metrics, several issues will need to be solved. This includes questions of responsibility for diagnoses made using AI tools and training needs for the pathologist of the future. There is also concern that the focus may shift too far toward computational skills, eroding the morphology knowledge that has been passed down for many generations. The path to widely accepted use of AI in diagnostic pathology can be thought of as having three core priorities. First, clear frameworks for standardized validation and ongoing surveillance are needed, as required for any medical device, to establish and ensure safety and continuous performance. Second, pathology education and training programs need to be adapted to ensure that all pathologists and aspiring pathologists are proficient not only in conventional morphology but also in understanding and working with data science tools, enabling them to assess their capabilities correctly and work synergistically. Finally, healthcare systems need to be reconfigured by adapting their workflows and infrastructure, and reimbursement models need to be updated so that AI can seamlessly integrate with clinical practice and be fully utilized for maximum benefit.

Ultimately, the hope is that AI will fundamentally transform pathology into a quantitative, integrated discipline that makes data-driven clinical decisions.

6. Limitations

The presenet study is subject to certain limitations, which deserve to be explicitly stated. These constraints fall into two categories: Limitations of the present review itself, and limitations of the underlying body of evidence that it surveys.

Limitations of the present review. Several constraints inherent to the present review must be acknowledged. First, the present article is a narrative rather than a systematic review. Studies were identified through structured but non-exhaustive searches of PubMed, Google Scholar, and arXiv, combining the terms 'digital pathology', 'artificial intelligence', 'deep learning', 'convolutional neural network', 'vision transformer', and the names of the cancer types covered, and were selected on the basis of methodological rigor, relevance to clinical translation, and representativeness of architectural trends. The authors did not pre-register a protocol, did not apply formal PRISMA criteria, and did not perform a quantitative meta-analysis or risk-of-bias assessment. As a consequence, study selection is inevitably subject to author judgment and to the visibility of high-profile or recent work, and important contributions may have been omitted. Second, the scope of the review is deliberately restricted to a small number of high-impact oncological applications, principally renal, prostate, colorectal, and lymphoid malignancies, together with metastasis detection. Findings should therefore not be extrapolated uncritically to other diagnostic domains, including non-neoplastic disease, infectious pathology, dermatopathology, hematopathology beyond lymphoma, autopsy pathology, or molecular subtyping outside the cancers discussed. Third, the field is evolving at an exceptional pace; foundation models, multimodal architectures, and self-supervised pretraining strategies are advancing on a timescale of months rather than years, and several of the performance metrics and architectures cited here are likely to be superseded by the time of publication. Fourth, the review focuses on diagnostic accuracy and architectural design and does not systematically address health-economic evaluation, prospective real-world implementation studies, or equity considerations relating to deployment in resource-limited settings, all of which are essential complements to technical performance and warrant dedicated treatment elsewhere. Readers are therefore encouraged to consult systematic reviews and live-evidence resources alongside this synthesis when making implementation or policy decisions.

Limitations of the underlying evidence base. AI in digital pathology is evolving rapidly; hence, studies published following the literature search cut-off date may not have been covered. Although the present review aimed to provide direct comparison among AI systems, this was challenging given variability in study design, cohorts and datasets, and evaluation methods employed (9,19), as well as varying levels of quality and reproducibility among studies.

Despite the fact that AI interpretability and explainability are addressed in some studies, they are not explored in detail. This remains a considerable bottleneck to trust from pathologists and regulatory authorities (21). Moreover, the present review mainly focuses on AI in cancer pathology, while less emphasis was placed on its applications in other pathology fields such as non-neoplastic conditions, infectious diseases or autopsy pathology, which represent an important area of research in the future (1,3). Instead of focusing mainly on the technical accuracy of the algorithms and their validation, cost effectiveness studies and other economic evaluations, as well as real-world implementation studies, are generally absent, even though this evidence is important for health care decision-making (7).

AI in pathology remains constrained by data-related limitations that directly impact model reliability and generalizability. A large proportion of these algorithms are only trained using curated data of high quality that may not at all resemble actual clinical samples such as variations in stain protocols, type of scanners, tissue preparation methods, and population diversity. This will inevitably cause a situation where the algorithm degrades sharply outside of the trained environment, a phenomenon known as 'domain shift'. Quality of the annotations is another problem in and of itself. Labels that define what is 'correct' are often decided by an expert pathologist and show substantial amounts of variance between different expert pathologists that introduce noise into the data. There is an absence of samples that show rare diseases or rare histological variants. Limited training datasets also introduce a bias towards certain institutions, causing reliability to be questioned with such tests being performed in hospitals outside of the context from where training data originates, making these tests unreliable or unfairly distributed.

Beyond data issues, significant limitations involving interpretability, regulation and workflow issues are also oberved. Most deep neural networks work as 'black boxes' to doctors and pathologists, which creates major difficulties in explaining why a certain prediction was made. This accountability is absolutely vital in diagnosis and prediction for legal and ethical reasons. To be considered for regulatory approval, these tools have to undergo testing. Yet, there have been few large-scale multi-center studies conducted prospectively to assess their reliability and accuracy in real-world hospital scenarios. The integration of these tools into daily pathology routines requires integration with digital pathology software, standardization of reports and clinician trust. Notably, AI tools lack contextual reasoning. They do not have any capacity for reasoning or awareness to consider patient histories, broader clinical contexts, the perspectives of other doctors or uncommon or edge cases, as only an actual expert pathologist can. As a result, AI is best positioned as an assistive tool rather than a replacement, with its limitations necessitating continuous human oversight.

7. Conclusion

AI has shown transformative potential in digital pathology, delivering diagnostic performance similar to that of human pathologists and in some cases superior in real-world settings across numerous cancer types and diagnoses. Despite the disadvantages of increased computational requirements for capturing long-range dependencies and global context, ViTs have become a powerful alternative to CNNs. Hybrid architectures and novel methods are constantly pushing the boundaries of what is possible through innovations in

efficiency, interpretability, and generalization. However, the translation from research success to clinical impact remains incomplete.

Despite the substantial progress in these directions, conducting research has been difficult. The requirements for computation and data standardization, regulatory and institutional barriers to dataset standardization, and workflow barriers to data integration stand as obstacles to broad application. Achieving this can only be done from a multidimensional perspective: Technical approaches to make things more efficient and consistent, validation research that critically tests generalization across a wide variety of clinical domains, regulatory frameworks that harmonize safety with innovation, and combined approaches that effectively incorporate human and AI collaboration. The field is at a critical intersection of these two trends. The technical foundations are in place, the regulatory pathways are already established, and some initial clinical implementations have demonstrated feasibility.

The next phase of development should focus on closing the gap between research prototypes and clinical tools that consistently enhance patient care across various healthcare delivery contexts. This will mean not just improved algorithms but also more successful inclusion of these algorithms into the tangled web of intricate sociotechnical systems that have become modern pathology practice. As digital pathology becomes routine in our lives and AI capabilities continue to develop, the dream of a clinical-style AI-augmented pathology practice will become a reality. The challenge now is to realize this transformation in ways that do more than just improve diagnostic accuracy, efficiency, and patient outcomes, while preserving human expertise and judgment in medical diagnosis. As technical and translational progress continue to accelerate, AI is expected to fundamentally reshape pathology practice for the better.

Acknowledgements

Not applicable.

Funding

No funding was received.

Availability of data and materials

Not applicable.

Authors' contributions

AMA conceptualized the present study, supervised the review and wrote the present manuscript. DA, a senior anatomical pathology consultant, critically reviewed and approved the information related to pathology in the present review article. Data authentication is not applicable. Both authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Patient consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Bera K, Schalper KA, Rimm DL, Velcheti V and Madabhushi A: Artificial intelligence in digital pathology-new tools for diagnosis and precision oncology. Nat Rev Clin Oncol 16: 703-715, 2019.
2. Niazi MKK, Parwani AV and Gurcan MN: Digital pathology and artificial intelligence. Lancet Oncol 20: e253-e261, 2019.
3. Baxi V, Edwards R, Montalto M and Saha S: Digital pathology and artificial intelligence in translational medicine and clinical practice. Mod Pathol 35: 23-32, 2022.
4. Steiner DF, Chen PC and Mermel CH: Closing the translation gap: AI applications in digital pathology. Biochim Biophys Acta Rev Cancer 1875: 188452, 2021.
5. Madabhushi A: Digital pathology image analysis: Opportunities and challenges. Imaging Med 1: 7-10, 2009.
6. Colling R, Pitman H, Oien K, Rajpoot N, Macklin P; CM-Path AI in Histopathology Working Group; Snead D, Sackville T and Verrill C: Artificial intelligence in digital pathology: A roadmap to routine use in clinical practice. J Pathol 249: 143-150, 2019.
7. Tizhoosh HR and Pantanowitz L: Artificial intelligence and digital pathology: Challenges and opportunities. J Pathol Inform 9: 38, 2018.
8. McGenity C, Clarke EL, Jennings C, Matthews G, Cartlidge C, Freduah-Agyemang H, Stocken DD and Treanor D: Artificial intelligence in digital pathology: A systematic review and meta-analysis of diagnostic test accuracy. NPJ Digit Med 7: 114, 2024.
9. Laleh NG, Muti C, Loeffler CML, Echle A, Saldanha OL, Mahmood F, Lu MY, Trautwein C, Langer R, Dislich B, *et al*: Benchmarking artificial intelligence methods for end-to-end computational pathology. bioRxiv: Aug 10, 2021 (Epub ahead of print).
10. Ghaffari Laleh N, Muti HS, Loeffler CML, Echle A, Saldanha OL, Mahmood F, Lu MY, Trautwein C, Langer R, Dislich B, *et al*: Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. Med Image Anal 79: 102474, 2022.
11. Bulten W, Pinckaers H, van Boven H, Vink R, de Bel T, van Ginneken B, van der Laak J, Hulsbergen-van de Kaa C and Litjens G: Automated deep-learning system for Gleason grading of prostate cancer using biopsies: A diagnostic study. Lancet Oncol 21: 233-241, 2020.
12. Wagner SJ, Reisenbüchler D, West NP, Niehues JM, Zhu J, Foersch S, Veldhuizen GP, Quirke P, Grabsch HI, van den Brandt PA, *et al*: Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. Cancer Cell 41: 1655-1668.e4, 2023.
13. Blilie A, Mulliqi N, Ji X, Szolnoky K, Boman SE, Titus M, Gonzalez GM, Asenjo J, Gambacorta M, Libretti P, *et al*: Artificial intelligence-assisted prostate cancer diagnosis for reduced use of immunohistochemistry. Commun Med (Lond) 5: 425, 2025.
14. Mulliqi N: Towards clinically reliable artificial intelligence for prostate cancer pathology. Karolinska Institutet, 2025.
15. Chaurasia AK, Harris HC, Toohey PW and Hewitt AW: A generalised vision transformer-based self-supervised model for diagnosing and grading prostate cancer using histological images. Prostate Cancer Prostatic Dis 28: 918-926, 2025.
16. Li D, Bledsoe JR, Zeng Y, Liu W, Hu Y, Bi K, Liang A and Li S: A deep learning diagnostic platform for diffuse large B-cell lymphoma with high accuracy across multiple hospitals. Nat Commun 11: 6004, 2020.
17. Mayer RS, Kinzler MN, Stoll AK, Gretser S, Ziegler PK, Saborowski A, Reis H, Vogel A, Wild PJ and Flinner N: The model transferability of AI in digital pathology: Potential and reality. Pathologie (Heidelb) 45: 124-132, 2024 (In German).
18. Dambandkhameneh M: Domain generalization across multi-centre digital pathology images for lymph node segmentation and classification. Toronto Metropolitan University, 2024.

19. Liu Y, Gadepalli K, Norouzi M, Dahl GE, Kohlberger T, Boyko A, Venugopalan S, Timofeev A, Nelson PQ, Corrado GS, *et al*: Detecting cancer metastases on gigapixel pathology images. arXiv: 1703.02442, 2017.
20. Raciti P, Sue J, Retamero JA, Ceballos R, Godrich R, Kunz JD, Casson A, Thiagarajan D, Ebrahimzadeh Z, Viret J, *et al*: Clinical validation of artificial intelligence-augmented pathology diagnosis demonstrates significant gains in diagnostic accuracy in prostate cancer detection. Arch Pathol Lab Med 147: 1178-1185, 2023.
21. Degadwala S, Midhunchakkaravarthy D and Khan S: Automating histologic assessment of prostate cancer with a ResNet50-based hybrid vision model. J Trends Comput Sci Smart Technol 7: 295-311, 2025.
22. Khader F, Kather JN, Han T, Nebelung S, Kuhl C, Stegmaier J and Truhn D: Cascaded cross-attention networks for data-efficient whole-slide image classification using transformers. In: Machine Learning in Medical Imaging. MLMI 2023. Lecture Notes in Computer Science. Cao X, Xu X, Rekik I, Cui Z and Ouyang X (eds). Vol 14349. Springer, Cham, pp417-426, 2023.
23. Cen M, Li X, Guo B, Jonnagaddala J, Zhang H and Xu XS: A novel and efficient digital pathology classifier for predicting cancer biomarkers using sequencer architecture. Am J Pathol 193: 2122-2132, 2023.
24. Pina O, Dorca E and Vilaplana V: Cell nuclei detection and classification in whole slide images with transformers. arXiv: 2502.06307, 2025.
25. Zhou B, Khosla A, Lapedriza A, Oliva A and Torralba A: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp2921-2929, 2016.
26. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D and Batra D: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp618-626, 2017.
27. Xu H, Usuyama N, Bagga J, Zhang S, Rao R, Naumann T, Wong C, Gero Z, González J, Gu Y, *et al*: A whole-slide foundation model for digital pathology from real-world data. Nature 630: 181-188, 2024.