Review

Artificial Intelligence in the Histopathological Diagnosis of Neoplasms: An Integrative Review of Diagnostic Accuracy and Clinical Efficacy (2019–2026)

Tiago José Bonomini , Najara Estefani Pereira dos Santos , Sônia Sales Vicente , Sandrina Kassouf , Dayssy Lorena Franco Torres , Stefhanie da Silva Pereira , Tainer Jordão de Farias , Alcides Chaux *

Posted Date: 27 April 2026

doi: 10.20944/preprints202604.1876.v1

Keywords: Artificial Intelligence; deep learning; digital pathology; computational histology; diagnostic accuracy; digital biomarkers; precision oncology

 doi:10.20944/preprints202604.1876.v1

Review

Artificial Intelligence in the Histopathological Diagnosis of Neoplasms: An Integrative Review of Diagnostic Accuracy and Clinical Efficacy (2019–2026)

Tiago Jose Bonomini [1], Najara Estefani Pereira dos Santos [1], Sonia Sales Vicente [1], Sandrina Kassouf [1], Dayssy Lorena Franco Torres [1], Stefhanie da Silva Pereira [1], Tainer Jordao de Farias [1] and Alcides Chaux [1,2,*]

[1] Universidad Central del Paraguay, filial Ciudad del Este, Paraguay

[2] ChauxLab Institute, Asunción, Paraguay

* Correspondence: alcides.chaux@central.edu.py or alcides.chaux@chauxlab.com

Abstract

Introduction: Conventional oncological pathology practice faces critical challenges stemming from interobserver variability and an ever-growing clinical workload. This review evaluates the technological maturity and clinical utility of artificial intelligence (AI) as a diagnostic and predictive support tool in histopathology. **Methods:** An integrative review of the literature (2019–2026) was conducted in PubMed, Scopus, and IEEE Xplore, following the methodology of Whittemore and Knafl. Studies on the diagnostic accuracy of deep learning algorithms in neoplasia histopathology were selected, with methodological quality assessed using QUADAS-2. **Results:** The evidence confirms that convolutional neural networks (CNNs) achieve diagnostic accuracy comparable to or exceeding that of pathologists in binary classification tasks, consistently reporting areas under the curve (AUC) > 0.98 in lung, breast, and prostate cancer. A disruptive finding is the validation of predictive computational histology, capable of inferring genotypic alterations—such as *EGFR* mutations or microsatellite instability—directly from standard hematoxylin and eosin (H&E) images, offering a cost-effective alternative for molecular screening. The evidence strongly supports the "augmented intelligence" model, in which the pathologist–AI synergy surpasses individual performance and mitigates visual fatigue. **Conclusions:** AI has transcended the experimental phase to become a robust technology for triage and digital phenotyping. Its definitive clinical adoption requires prioritizing multicenter external validation and the development of explainable AI (XAI) interfaces to overcome the "black box" barrier.

Keywords: Artificial Intelligence; deep learning; digital pathology; computational histology; diagnostic accuracy; digital biomarkers; precision oncology

1. Introduction

The histopathological examination of tissue samples remains, to date, the indisputable gold standard for the diagnosis, staging, and prognosis of neoplasms. However, conventional pathology practice confronts critical challenges: an exponential increase in workload, the growing complexity of tumor classifications, and, most fundamentally, the interobserver variability inherent to human interpretation [1]. This subjectivity can result in significant diagnostic discrepancies, a phenomenon that has been historically documented in tumor grading, particularly in prostate adenocarcinoma (Gleason system) and breast cancer, where interexpert concordance kappa indices frequently fall short of agreement thresholds [2,3].

In response to these limitations, digital pathology and artificial intelligence have emerged as transformative tools. The transition toward whole slide imaging (WSI) has enabled deep learning (DL) algorithms—and specifically convolutional neural networks (CNNs)—to analyze

 doi:10.20944/preprints202604.1876.v1

morphological patterns with unprecedented precision. Recent studies suggest that these algorithms can not only detect neoplastic lesions with sensitivity exceeding 90% but also predict molecular subtypes and genetic mutations directly from standard hematoxylin and eosin staining, a task visually impossible for the human eye [4,5]. State-of-the-art investigations report AUC values approaching 0.99 in binary classification tasks, occasionally surpassing the performance of generalist pathologists in experimental settings [6].

Despite this technological enthusiasm, the clinical implementation of AI is not without controversy. Although the literature abounds with successful technical validations, critical questions persist regarding the generalization of these models in the face of preanalytical variability—differences in scanner hardware and staining protocols—and the robustness of curated datasets [7]. Additionally, the "black box" nature of deep learning algorithms poses ethical and clinical trust challenges that limit their routine adoption in hospital workflows [8]. A broader concern is that AI's clinical utility depends not only on technical performance but also on transparent integration into existing laboratory information and picture archiving systems [9].

It is therefore imperative to synthesize the current evidence and determine whether AI has transcended technical validation to become a tool of genuine clinical utility. This integrative review aims to evaluate the diagnostic accuracy of artificial intelligence algorithms in neoplasia histopathology between 2019 and 2025, analyzing their comparative efficacy against the human pathologist, the impact of human-in-the-loop approaches, and the barriers to definitive clinical implementation.

2. Methods

2.1. Study Design

An integrative review of the literature was conducted following the methodological framework proposed by Whittemore and Knafl [10]. This approach allows the inclusion of studies with diverse methodologies — both experimental and observational — to achieve a holistic understanding of the phenomenon of interest: the diagnostic accuracy of artificial intelligence in histopathology. Reporting of the review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) statement, adapted to the integrative nature of the study [11].

2.2. Search Strategy

The bibliographic search was conducted across three major electronic databases: PubMed (MEDLINE), Scopus, and IEEE Xplore, covering the period from January 1, 2019, to Apr 25, 2026. This time interval was selected to ensure the inclusion of state-of-the-art deep learning algorithms, whose architectures have evolved significantly since 2018, and to capture emerging evidence, including the consolidation of foundation models as a new architectural paradigm in computational pathology.

The search strategy combined Medical Subject Headings (MeSH) terms and free-text language using Boolean operators (AND, OR). The base search string, adapted to the syntax of each database, was as follows: ("Artificial Intelligence" OR "Deep Learning" OR "Machine Learning" OR "Convolutional Neural Networks") AND ("Histopathology" OR "Whole Slide Imaging" OR "Digital Pathology") AND ("Neoplasms" OR "Cancer" OR "Carcinoma") AND ("Sensitivity and Specificity" OR "Diagnostic Accuracy" OR "Area Under Curve").

Additionally, a manual snowballing search was performed on the reference lists of the most relevant systematic reviews retrieved to identify additional primary studies.

2.3. Eligibility Criteria

Study selection was governed by the PICOS criteria (population, intervention, comparator, outcome, study design) detailed in Table 1.

 doi:10.20944/preprints202604.1876.v1

Table 1. Eligibility criteria (PICOS framework).

Criterion	Inclusion	Exclusion
Population	Digitized human tissue samples (biopsies, resections) with a neoplasia diagnosis (WSI).	Animal model studies, *in vitro* cell lines, or genomic analyses without histopathological correlation.
Intervention	AI algorithms (ML, DL, CNN) applied for tumor detection, classification, or grading.	Algorithms applied exclusively to radiology, non-histological cytology, or clinical dermatoscopy.
Comparator	Conventional histopathological diagnosis (human pathologist), expert consensus, or molecular/IHC tests (gold standard).	Software development technical studies without clinical comparative validation or without diagnostic performance metrics.
Outcomes	Diagnostic performance metrics (sensitivity, specificity, AUC, F1-score, accuracy).	Studies focused solely on nuclear or stromal segmentation without direct diagnostic implication.
Study Design	Diagnostic validation studies, retrospective and prospective cohorts.	Editorials, letters to the editor, expert opinions, book chapters, and conference abstracts without full text.
Language	English, Spanish, Portuguese.	Other languages.

2.4. Study Selection and Data Extraction

The selection process was conducted in two stages. First, titles and abstracts were screened to eliminate duplicates and clearly irrelevant studies. Second, the full text of potentially eligible articles was evaluated. Data extraction was systematized using a predefined template that included: author, year, country, neoplasia type, sample size (number of WSI/patients), algorithm architecture (e.g., ResNet, Inception), reference standard used, and performance metrics (AUC, sensitivity, specificity).

2.5. Methodological Quality Assessment

The QUADAS-2 tool (Quality Assessment of Diagnostic Accuracy Studies-2), the international standard for diagnostic accuracy studies, was used to assess risk of bias and applicability of included studies [12]. This tool evaluates four key domains: patient selection, index test (AI), reference standard, and flow and timing. Each domain was rated as "low," "high," or "unclear" risk of bias. Given the nature of AI studies, particular attention was paid to the independence between training and validation datasets to avoid overfitting.

2.6. Data Synthesis

Due to clinical heterogeneity (diverse cancer types) and methodological heterogeneity (different algorithm architectures and reported metrics), a statistical meta-analysis was not feasible. Consequently, results are presented through a structured narrative synthesis, grouping findings by neoplasia type and clinical application (detection, subtyping, grading).

3. Results

The systematic analysis of the literature selected between 2019 and 2026 reveals substantial progress in the technological maturity and clinical validation of artificial intelligence algorithms in digital pathology. Data from primary studies and systematic reviews were identified and synthesized, with findings categorized across five principal domains.

3.1. Methodological Quality Assessment (QUADAS-2)

The quality of included studies was assessed using QUADAS-2. Overall, risk of bias was considered low for the "reference standard" domain, given that most studies used diagnoses

 doi:10.20944/preprints202604.1876.v1

confirmed by expert pathologist consensus or immunohistochemistry, preserving the gold standard hierarchy. However, areas of concern were identified in the "patient selection" and "applicability" domains.

Regarding selection bias, a significant proportion of studies (approximately 40%) trained and validated their algorithms using curated public datasets (e.g., TCGA, CAMELYON16/17), which often exhibit artificially enriched disease prevalence and do not reflect the preanalytical "noise" — tissue artifacts, bubbles, folds — or the real epidemiological case distribution of clinical routine [13,14]. Regarding external validation, although the most recent studies (2023–2025) have improved in this respect, several earlier works lacked independent external validation cohorts from different institutions. This raises the risk of overfitting and limiting the generalization of results to centers with different scanners and digital infrastructure [15].

3.2. Global Diagnostic Accuracy and Evolution of Architectures

The accumulated evidence documents a radical and irreversible paradigm shift in algorithmic design applied to digital pathology. The field has transitioned from traditional machine learning methods—which relied on manual hand-crafted features, restricted to quantifying rigid parameters such as nuclear morphometry, stromal texture, or staining intensity—toward end-to-end deep learning models [16].

3.2.1. Dominance of Convolutional Neural Networks

In this new ecosystem, CNNs have consolidated their position as the methodological gold standard. Deep architectures such as ResNet (Residual Networks), Inception, and DenseNet have demonstrated superior capacity to capture the phenotypic heterogeneity of neoplasms. Unlike their predecessors, these algorithms learn complex hierarchical representations directly from image pixels: initial layers detect edges and colors, while deeper layers recognize complex histological patterns—tubular formation, nuclear atypia, desmoplasia—that often elude semantic human definition [7,16]. Furthermore, the widespread adoption of transfer learning—where models pretrained on large generic datasets (such as ImageNet) are fine-tuned for histopathology—has enabled high performance even with limited medical datasets [17].

3.2.2. Performance Metrics and Clinical Robustness

In quantitative terms, the gap between human and computational visual capacity has narrowed significantly. State-of-the-art algorithms consistently report performance metrics approaching technical perfection for binary classification tasks (benign vs. malignant tissue). A paradigmatic finding is a prospective double-blinded study using AI-driven reflectance confocal microscopy for in vivo oral cancer diagnosis, which achieved a sensitivity of 90% and a specificity of 98.3% [6]. This demonstrates that high accuracy can be reproduced in real clinical settings beyond curated datasets. The clinical implication of AUC values approaching 0.99 is profound: they position AI not merely as a support tool, but as a potential autonomous triage mechanism capable of ruling out benign cases with high confidence.

3.2.3. The Generalization Challenge: Stain Normalization

Despite outstanding metrics, domain generalization remains the primary technical obstacle. Variations in H&E staining protocols across laboratories—due to differences in reagent manufacturers, immersion times, or scanner models—can drastically alter the digital appearance of a slide (the "batch effect"). To mitigate this, the most robust studies from 2023–2025 integrate advanced stain normalization techniques (e.g., Macenko or Vahadane methods) or color augmentation during training. These strategies compel the algorithm to ignore irrelevant chromatic variations and focus exclusively on tissue morphology, ensuring that diagnostic accuracy is reproducible in multicenter environments [18,19].

 doi:10.20944/preprints202604.1876.v1

3.2.4. The Weakly Supervised Learning Revolution

Campanella et al. set a landmark by validating multiple instance learning (MIL) at massive scale (44,732 slides) [4]. This finding was subsequently reinforced by Lu et al. with the CLAM (Clustering-constrained Attention Multiple Instance Learning) model, achieving AUC > 0.98 for prostate, skin, and breast cancer without requiring pixel-level manual annotations [20]. By training the model using only the patient's final diagnostic label ("cancer" vs. "no cancer"), these studies demonstrated that it is possible to unlock the potential of vast historical hospital archives without incurring the prohibitive cost of manual curation, thereby facilitating clinical scalability.

3.2.5. The Emergence of Whole-Slide Foundation Models

The most architecturally significant development in the search period is the consolidation of pathology foundation models (FMs) as the next evolutionary step beyond task-specific CNNs. Whereas conventional deep learning requires training a separate model for each diagnostic task—a process demanding large annotated datasets and yielding algorithms that generalize poorly across institutions—FMs are pretrained on massive histological corpora and can be adapted to new tasks with minimal additional data. Ding et al. published in Nature Medicine a multimodal whole-slide foundation model capable of simultaneously processing visual and textual information from WSIs, achieving strong performance across diverse cancer types without task-specific retraining [21]. This architecture represents a qualitative departure from the CNN paradigm and directly addresses the domain generalization problem that has historically limited clinical deployment. Complementary benchmark evidence from Yuan et al. confirms that foundation models such as UNI, when fine-tuned on histopathological datasets, achieve AUC of 0.999 in binary breast cancer classification and 0.998 in eight-class subtype classification — performance that matches or exceeds the best available CNN architectures [22]. Critically, zero-shot performance without fine-tuning remains limited, indicating that minimal task-specific adaptation remains necessary before clinical deployment.

3.3. Performance by Neoplasia Type

The stratified analysis by organ system demonstrates that the utility of AI is not uniform but adapts to the specific diagnostic challenges of each pathology.

3.3.1. Thoracic Neoplasms (Lung Cancer)

In this domain, AI has transcended mere detection to address complex classification issues with direct therapeutic impact. Coudray et al. [5] and subsequently Wei et al. [23] trained deep neural networks (e.g., Inception V3) that distinguished between the two main subtypes of non-small cell lung cancer—adenocarcinoma and squamous cell carcinoma—with an AUC of 0.97. This performance is comparable to that of experienced pathologists and carries critical clinical relevance: by making this distinction based solely on standard H&E images, the algorithm reduces the need for confirmatory immunohistochemical stains (such as p63 or TTF-1). This not only accelerates diagnosis but also preserves valuable tissue in small biopsies, allowing the remaining tissue to be used for mandatory molecular and genetic testing (e.g., EGFR mutations, ALK translocations).

3.3.2. Breast and Lymph Node Pathology

The most mature and validated application in this field is the detection of metastases in sentinel lymph nodes, consolidated after the CAMELYON16 challenge. The search for micrometastases (< 2 mm) across multiple tissue levels is a tedious and error-prone task due to visual fatigue. CNN-based algorithms have consistently demonstrated sensitivity exceeding 90% in this task, acting as an infallible safety net that drastically reduces false negatives [8,24,25]. Moreover, AI is resolving the subjectivity problem in tumor grading. The quantification of the Ki-67 proliferation index and mitotic figure counts—essential components of the Nottingham histological grade—have historically shown

 doi:10.20944/preprints202604.1876.v1

low interobserver reproducibility. Automated image analysis tools provide standardized and reproducible counts, resulting in more precise risk stratification [19].

3.3.3. Uropathology (Prostate and Bladder)

In prostate cancer, the central challenge is variability in Gleason score assignment, especially in the distinction between pattern 3 (well-formed glands) and pattern 4 (cribriform/fused), which determines whether a patient proceeds to surgery or active surveillance. High-impact studies from the PANDA consortium [26] and the Karolinska study [27] indicate that AI systems achieve kappa concordance exceeding 0.80 with urological pathology subspecialists, standardizing grading at a global level. Concurrently, in bladder cancer, Zheng et al. [28] presented a DL model capable of predicting not only the diagnosis but also overall patient survival with high accuracy. This suggests that AI identifies biological aggressiveness features in the histological slide not contemplated in current WHO grading systems.

3.3.4. Neuropathology (Glioblastoma)

Research in brain tumors has revealed AI capabilities that go beyond human vision. Zheng et al. [29] demonstrated that the analysis of spatial cellular architecture—the distance relationships and neighborhood organization between cells—via graph algorithms predicts prognosis in glioblastoma patients with greater precision than traditional clinical variables. While the pathologist observes cell types, AI observes the ecosystem and complex spatial interactions that correlate with aggressive molecular phenotypes.

3.3.5. Dermatopathology and Gastrointestinal Pathology

The distinction between adenoma (low-grade dysplasia) and early adenocarcinoma in gastrointestinal biopsies is a common source of discrepancy. Iizuka et al. [30] reported models with AUC > 0.96 for classifying gastric and colonic lesions, providing decisive decision support in borderline cases. In digital pathology more broadly, the "black box" concern has been a major barrier. A systematic review and survey of over 150 papers by Abdelsamea et al. underscores that explainable AI (XAI) techniques—including attention maps and feature visualization—are technically feasible and essential to broadening the use of AI algorithms in the clinical setting [31]. Crucially, these approaches highlight the specific morphological regions driving algorithmic conclusions, substantially increasing pathologist confidence in predictive output.

3.4. Comparative Efficacy: From "Replacement" to Augmented Intelligence and the Human-AI Synergy

The exhaustive review of the literature enables the dismantling of the pathologist replacement myth. The evidence does not support a competitive scenario but rather a collaborative one of augmented intelligence, where the inherent weaknesses of human cognition are compensated by computational power, and vice versa.

One of the most consistent findings is AI's capacity to standardize diagnostic practice. Human morphological evaluation is intrinsically subjective; classic studies cite kappa concordance indices as low as 0.40–0.60 in critical tasks such as the differentiation of ductal atypia in breast or the assignment of intermediate Gleason scores (3+4 vs. 4+3) [3]. Bera et al. emphasize that the introduction of decision-support algorithms drastically reduces this variability [1]. By providing objective and reproducible quantification of histological features (e.g., percentage of fused glands), AI acts as a real-time calibrated second opinion, elevating the concordance of general pathologists to levels comparable with subspecialist experts. Crucially, AI does not tire, has no cognitive bias from prior cases, and maintains constant criteria throughout the working day, mitigating human error due to fatigue.

3.5. The Human-in-the-Loop Model: Superior Synergy

 doi:10.20944/preprints202604.1876.v1

Comparative validation studies consistently demonstrate that maximum diagnostic performance is achieved neither with AI alone nor with the pathologist alone, but through integration of both. Steiner et al. demonstrated that deep learning assistance increased pathologists' sensitivity for detecting nodal metastases and significantly reduced slide review time for negative cases [32]. Similarly, in simulated workflows, Al-Nafjan et al. report that the hybrid model achieves a significant reduction in error rates [7]. A prospective study by Lin et al. [33] provides some of the most concrete quantitative evidence for this synergy to date: a transformer-based AI model for classifying histological growth patterns in colorectal liver metastases elevated junior pathologist accuracy from 85.9% to 94.7% while simultaneously reducing diagnostic time by 36%, with the AI-assisted group outperforming senior pathologists working without assistance. AI excels at high-sensitivity tasks—exhaustive screening, micrometastasis detection, mitosis counting—acting as a safety net that alerts to regions of interest that the human eye might miss. The human pathologist, in turn, provides contextual specificity, ruling out false positives generated by technical artifacts (tissue folds, ink, debris) that often confuse algorithms. This symbiosis results in a more robust, efficient, and patient-safe diagnostic system.

Beyond pure accuracy, AI is transforming laboratory logistics. Validated algorithms are being implemented to perform intelligent triage of the daily workload. AI can preanalyze slides overnight, classifying cases as “probably benign” or “high malignancy suspicion,” thereby allowing urgent cases to be prioritized at the start of the working day when pathologist mental acuity is at its peak, reducing turnaround times for critical cancer diagnoses (Ström et al., 2020). Baxi et al. (2022) similarly document that a complete digital pathology workflow incorporating AI increases efficiency and operational utility without compromising diagnostic quality.

3.6. Beyond Morphology: Molecular and Genomic Prediction

One of the most disruptive and promising findings of this review is the demonstrated capacity of AI to infer complex genotypic alterations directly from standard histological images (H&E). This field, termed “predictive computational histology,” suggests that the tumor genotype imprints a subtle phenotypic signature on tissue architecture—invisible to the human eye but decipherable by deep neural networks.

Coudray et al. established a historical precedent by demonstrating that a deep neural network could predict driver gene mutations in non-small cell lung cancer [5]. Their model achieved AUC of 0.85 for *STK11*, 0.75 for *EGFR*, and 0.81 for *TP53* in independent validation cohorts. The capacity to predict *EGFR* status is particularly relevant, as it determines eligibility for therapies with tyrosine kinase inhibitors (TKIs), proposing that AI could serve as a rapid pre-test while awaiting molecular confirmation.

Expanding on these findings, Kather et al. published a fundamental multicenter study that validated AI's capacity to detect microsatellite instability (MSI) directly from H&E slides in colorectal, gastric, and endometrial cancer [34]. MSI is a critical biomarker predicting response to immunotherapy (checkpoint inhibitors). The algorithm achieved an AUC of 0.84 in the external validation cohort, significantly exceeding chance. Subsequent results from Echle et al. [35] demonstrated that these models can generalize across diverse clinical cohorts, suggesting that implementation of this tool would allow universal, rapid, and economical identification of immunotherapy candidates. This would resolve the problem of MSI underdiagnosis in centers with limited resources for reflex molecular testing.

Applicability extends to breast pathology. Shamai et al. [36] and Couture et al. [37] developed models to predict estrogen receptor (ER), progesterone receptor (PR) expression, and intrinsic molecular classification (luminal A/B vs. basal) based solely on morphology. While accuracy (AUC approximately 0.70–0.78) does not yet reach the level to replace immunohistochemistry, these algorithms offer an automated quality assurance mechanism to detect technical errors in IHC staining or sample confusion.

 doi:10.20944/preprints202604.1876.v1

More recent investigations—such as Fu et al. [38]—have demonstrated that this predictive capacity is a pan-cancer phenomenon applicable to more than ten tumor types with variable precision. Furthermore, the integration of AI with spatial transcriptomics data is enabling intratumoral gene expression mapping at near-cellular resolution, as suggested by the work of Zheng et al. in glioblastoma [29], linking spatial morphology with molecular aggressiveness.

The validation of these models has profound implications for the democratization of precision oncology. Tissue conservation is enabled by obtaining presumptive molecular information without exhausting the paraffin block—crucial in small biopsies. Concerning cost-effectiveness, an H&E slide (cost < 5 USD) is transformed into a molecular screening tool, reserving next-generation sequencing (NGS, cost > 500–1000 USD) for confirmatory cases, thereby optimizing healthcare system resources.

4. Discussion

The critical synthesis of evidence accumulated between 2019 and 2026 indicates that histopathology stands on the threshold of an irreversible paradigmatic transformation, comparable in magnitude and disruption to the introduction of immunohistochemistry in the late twentieth century. The results of this review confirm, without ambiguity, that deep learning algorithms have transcended the experimental proof-of-concept stage to reach a robust preclinical maturity. Studies such as the prospective double-blinded evaluation of AI-driven confocal microscopy by Zanoni et al. [6] demonstrate that CNNs can achieve near-perfect diagnostic accuracy (specificity 98.3%, sensitivity 90%) in real clinical conditions, surpassing human visual discrimination capacity for subtle and repetitive patterns. Moreover, 2025 has witnessed a structural shift in the field: the emergence of multimodal whole-slide foundation models [21] signals that the dominant design paradigm is transitioning from single-task CNNs toward general-purpose architectures capable of performing diverse diagnostic functions from a single pretrained representation. This is analogous to the transition from specialized expert systems to large language models in natural language processing.

Nevertheless, it is imperative to maintain healthy scientific skepticism: this computational supervision is not without critical vulnerabilities. While algorithmic sensitivity is outstanding, external generalization remains the Achilles' heel of the technology. Dependence on curated training datasets and extreme sensitivity to preanalytical variations—subtle differences in staining protocols, reagent degradation, or optical characteristics of different scanner manufacturers, as reported by Tellez et al. [39] and Koohbanani et al. [18]— underscores that robustness in a real laboratory environment is a considerably more complex challenge than in silico validation. An algorithm that excels on a clean dataset like TCGA may fail catastrophically on a slide with folds, air bubbles, or excessive margin ink—artifacts that the human eye unconsciously ignores but which can induce computational false positives.

Beyond the mere replication of traditional morphological diagnosis, the most disruptive finding of this period is the consolidation of predictive computational histology. The validated capacity to infer complex genotypic alterations—such as driver mutations in *EGFR* or microsatellite instability—directly from standard H&E slides [5,34] redefines the intrinsic value of the conventional biopsy. This suggests a fascinating biological hypothesis: the tumor genotype imprints a phenotypic signature on tissue architecture—possibly through nuclear chromatin patterns or stromal arrangement—that is invisible to the human eye but mathematically decipherable by deep neural networks.

This advance has profound ethical and economic implications for global health equity. It transforms a low-cost glass slide into a mass molecular screening tool, offering a viable solution for patient prioritization in health systems where universal next-generation sequencing is economically unsustainable [35] or logistically slow. The impact of obtaining predictive biomarker status within minutes of digitization—rather than waiting weeks for molecular results—represents a concrete pathway toward more equitable cancer care. Nevertheless, the implementation of these digital biomarkers must be handled with extreme clinical caution, functioning strictly as triage tests with high negative predictive value to rule out cases (conserving resources), rather than replacing the established confirmatory molecular tests that dictate definitive therapy.

 doi:10.20944/preprints202604.1876.v1

The antagonistic narrative of "human vs. machine" has been definitively displaced by the collaborative model of augmented intelligence. The reviewed evidence consistently demonstrates that the symbiosis between the pathologist and the algorithm is superior to either component operating independently [7,32]. AI contributes an indefatigable capacity for high-volume screening and the mathematical standardization of subjective tasks, mitigating interobserver variability historically documented by Elmore et al. [3] and Bera et al. [1]. This enables a cognitive offloading, where the pathologist delegates the tedious search for rare events—such as micrometastases—to the algorithm. The human pathologist, in turn, contributes the indispensable clinical integration, contextual judgment, and ethical accountability necessary to manage diagnostic ambiguity, borderline cases, and technical artifacts that confuse the machine.

However, full adoption of this hybrid workflow faces the formidable barrier of the "black box." The inherent opacity of the hidden layers of deep learning algorithms creates a clinical and legal trust gap: who is responsible if AI misses a diagnosis? This gap can only be bridged through aggressive development of explainable artificial intelligence (XAI). As Abdelsamea et al. [31] comprehensively document, XAI techniques — including attention maps, feature attribution, and local interpretable model-agnostic explanations (LIME) — are not merely academic exercises but necessary translational steps to make AI-assisted diagnosis not only accurate but transparent, auditable, and ultimately legally defensible before patients and the judicial system. A methodologically compelling advance in this direction is the work of Mittmann et al. [40], who developed an inherently explainable AI—rather than one requiring post-hoc explanation—for Gleason grading in prostate cancer. Trained with annotations from 54 international pathologists using pathologist-defined terminology and soft labels to capture interobserver uncertainty, the model achieves segmentation performance comparable to or superior to black-box alternatives (Dice score 0.713 vs. 0.691) while producing outputs directly interpretable in terms pathologists already use. This approach demonstrates that explainability and accuracy need not be traded against each other and that incorporating domain expert knowledge into model design from the outset is a viable path to clinical trust. The integration of such approaches into clinical digital pathology workflows, as reviewed by Baxi et al. [9] in the context of biomarker discovery and patient stratification, represents the next critical frontier for the field.

In conclusion, AI in histopathology has proven to be a disruptive technology with validated diagnostic accuracy and an unprecedented capacity to extract prognostic and molecular information concealed in basic tissue morphology. Although current algorithms match or surpass human performance on specific, repetitive detection and grading tasks, their real value does not reside in replacing the pathologist. On the contrary, the future of the specialty lies in the formation of an augmented intelligence ecosystem that optimizes operational efficiency, reduces the subjectivity inherent to human perception, and democratizes access to precision medicine through accessible digital biomarkers. The transition from task-specific CNNs toward multimodal whole-slide foundation models, now underway, promises to further reduce the barriers to clinical deployment by enabling general-purpose architectures that require only minimal task-specific adaptation. To achieve effective, safe, and ethical clinical translation, future research must abandon isolated validation in silos of perfect data to focus on prospective multicenter clinical trials that address domain generalization, algorithmic explainability, and seamless integration into existing digital pathology workflows.

Author Contributions: Conceptualization, T.J.B.; methodology, A.C.; formal analysis, A.C.; investigation, T.J.B.; writing—original draft preparation, T.J.B.; writing—review and editing, N.E.P.S., S.S.V, S.K., S.L.F.T., S.S.P., T.J.F., and A.C.; supervision, A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

 doi:10.20944/preprints202604.1876.v1

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bera, K.; Schalper, K.A.; Rimm, D.L.; Velcheti, V.; Madabhushi, A. Artificial Intelligence in Digital Pathology - New Tools for Diagnosis and Precision Oncology. *Nat. Rev. Clin. Oncol.* **2019**, *16*, 703–715, doi:10.1038/s41571-019-0252-y.
2. Allsbrook, W.C., Jr; Mangold, K.A.; Johnson, M.H.; Lane, R.B.; Lane, C.G.; Amin, M.B.; Bostwick, D.G.; Humphrey, P.A.; Jones, E.C.; Reuter, V.E.; et al. Interobserver Reproducibility of Gleason Grading of Prostatic Carcinoma: Urologic Pathologists. *Hum. Pathol.* **2001**, *32*, 74–80, doi:10.1053/hupa.2001.21134.
3. Elmore, J.G.; Nelson, H.D.; Pepe, M.S.; Longton, G.M.; Tosteson, A.N.A.; Geller, B.; Onega, T.; Carney, P.A.; Jackson, S.L.; Allison, K.H.; et al. Variability in Pathologists' Interpretations of Individual Breast Biopsy Slides: A Population Perspective. *Ann. Intern. Med.* **2016**, *164*, 649–655, doi:10.7326/M15-0964.
4. Campanella, G.; Hanna, M.G.; Geneslaw, L.; Miraflor, A.; Werneck Krauss Silva, V.; Busam, K.J.; Brogi, E.; Reuter, V.E.; Klimstra, D.S.; Fuchs, T.J. Clinical-Grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images. *Nat. Med.* **2019**, *25*, 1301–1309, doi:10.1038/s41591-019-0508-1.
5. Coudray, N.; Ocampo, P.S.; Sakellaropoulos, T.; Narula, N.; Snuderl, M.; Fenyö, D.; Moreira, A.L.; Razavian, N.; Tsirigos, A. Classification and Mutation Prediction from Non-Small Cell Lung Cancer Histopathology Images Using Deep Learning. *Nat. Med.* **2018**, *24*, 1559–1567, doi:10.1038/s41591-018-0177-5.
6. Zanoni, D.K.; Demétrio De Souza França, P.; Valero, C.; Peterson, G.; Ardigo, M.; Ghossein, R.; Dusza, S.W.; Matsuura, D.; Scholfield, D.W.; Adilbay, D.; et al. A Prospective Double-Blinded Comparison of Reflectance Confocal Microscopy with Conventional Histopathology for in Vivo Assessment in Oral Cancer. *Clin. Cancer Res.* **2024**, *30*, 2486–2496, doi:10.1158/1078-0432.CCR-23-1361.
7. Al-Nafjan, A.; Aljuhani, A.; Alshebel, A.; Alharbi, A.; Alshehri, A. Artificial Intelligence in Predictive Healthcare: A Systematic Review. *J. Clin. Med.* **2025**, *14*, 6752, doi:10.3390/jcm14196752.
8. Cruz-Roa, A.; Basavanhally, A.; González, F.; Gilmore, H.; Feldman, M.; Ganesan, S.; Shih, N.; Tomaszewski, J.; Madabhushi, A. Automatic Detection of Invasive Ductal Carcinoma in Whole Slide Images with Convolutional Neural Networks. In Proceedings of the Medical Imaging 2014: Digital Pathology; Gurcan, M.N., Madabhushi, A., Eds.; SPIE, March 20 2014; Vol. 9041, p. 904103.
9. Baxi, V.; Edwards, R.; Montalto, M.; Saha, S. Digital Pathology and Artificial Intelligence in Translational Medicine and Clinical Practice. *Mod. Pathol.* **2022**, *35*, 23–32, doi:10.1038/s41379-021-00919-2.
10. Whittemore, R.; Knafl, K. The Integrative Review: Updated Methodology. *J. Adv. Nurs.* **2005**, *52*, 546–553, doi:10.1111/j.1365-2648.2005.03621.x.
11. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* **2021**, *372*, n71, doi:10.1136/bmj.n71.
12. Whiting, P.F.; Rutjes, A.W.S.; Westwood, M.E.; Mallett, S.; Deeks, J.J.; Reitsma, J.B.; Leeflang, M.M.G.; Sterne, J.A.C.; Bossuyt, P.M.M.; QUADAS-2 Group QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. *Ann. Intern. Med.* **2011**, *155*, 529–536, doi:10.7326/0003-4819-155-8-201110180-00009.
13. Janowczyk, A.; Madabhushi, A. Deep Learning for Digital Pathology Image Analysis: A Comprehensive Tutorial with Selected Use Cases. *J. Pathol. Inform.* **2016**, *7*, 29, doi:10.4103/2153-3539.186902.
14. Wolff, R.F.; Moons, K.G.M.; Riley, R.D.; Whiting, P.F.; Westwood, M.; Collins, G.S.; Reitsma, J.B.; Kleijnen, J.; Mallett, S.; PROBAST Group† PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann. Intern. Med.* **2019**, *170*, 51–58, doi:10.7326/M18-1376.
15. Stanley, E.A.M.; Souza, R.; Wilms, M.; Forkert, N.D. Where, Why, and How Is Bias Learned in Medical Image Analysis Models? A Study of Bias Encoding within Convolutional Networks Using Synthetic Data. *EBioMedicine* **2025**, *111*, 105501, doi:10.1016/j.ebiom.2024.105501.

 doi:10.20944/preprints202604.1876.v1

16. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A Survey on Deep Learning in Medical Image Analysis. *Med. Image Anal.* **2017**, *42*, 60–88, doi:10.1016/j.media.2017.07.005.
17. Mormont, R.; Geurts, P.; Maree, R. Comparison of Deep Transfer Learning Strategies for Digital Pathology. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); IEEE, June 2018; pp. 2343–234309.
18. Koohbanani, N.A.; Unnikrishnan, B.; Khurram, S.A.; Krishnaswamy, P.; Rajpoot, N. Self-Path: Self-Supervision for Classification of Pathology Images with Limited Annotations. *IEEE Trans. Med. Imaging* **2021**, *40*, 2845–2856, doi:10.1109/TMI.2021.3056023.
19. Veta, M.; Heng, Y.J.; Stathonikos, N.; Bejnordi, B.E.; Beca, F.; Wollmann, T.; Rohr, K.; Shah, M.A.; Wang, D.; Rousson, M.; et al. Predicting Breast Tumor Proliferation from Whole-Slide Images: The TUPAC16 Challenge. *Med. Image Anal.* **2019**, *54*, 111–121, doi:10.1016/j.media.2019.02.012.
20. Lu, M.Y.; Williamson, D.F.K.; Chen, T.Y.; Chen, R.J.; Barbieri, M.; Mahmood, F. Data-Efficient and Weakly Supervised Computational Pathology on Whole-Slide Images. *Nat. Biomed. Eng.* **2021**, *5*, 555–570, doi:10.1038/s41551-020-00682-w.
21. Ding, T.; Wagner, S.J.; Song, A.H.; Chen, R.J.; Lu, M.Y.; Zhang, A.; Vaidya, A.J.; Jaume, G.; Shaban, M.; Kim, A.; et al. A Multimodal Whole-Slide Foundation Model for Pathology. *Nat. Med.* **2025**, *31*, 3749–3761, doi:10.1038/s41591-025-03982-3.
22. Yuan, B.; Hu, Y.; Liang, Y.; Zhu, Y.; Zhang, L.; Cai, S.; Peng, R.; Wang, X.; Yang, Z.; Hu, J. Comparative Analysis of Convolutional Neural Networks and Transformer Architectures for Breast Cancer Histopathological Image Classification. *Front. Med. (Lausanne)* **2025**, *12*, 1606336, doi:10.3389/fmed.2025.1606336.
23. Wei, J.W.; Tafe, L.J.; Linnik, Y.A.; Vaickus, L.J.; Tomita, N.; Hassanpour, S. Pathologist-Level Classification of Histologic Patterns on Resected Lung Adenocarcinoma Slides with Deep Neural Networks. *Sci. Rep.* **2019**, *9*, 3358, doi:10.1038/s41598-019-40041-7.
24. Ehteshami Bejnordi, B.; Veta, M.; Johannes van Diest, P.; van Ginneken, B.; Karssemeijer, N.; Litjens, G.; van der Laak, J.A.W.M.; the CAMELYON16 Consortium; Hermsen, M.; Manson, Q.F.; et al. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women with Breast Cancer. *JAMA* **2017**, *318*, 2199–2210, doi:10.1001/jama.2017.14585.
25. Golden, J.A. Deep Learning Algorithms for Detection of Lymph Node Metastases from Breast Cancer: Helping Artificial Intelligence Be Seen: Helping Artificial Intelligence Be Seen. *JAMA* **2017**, *318*, 2184–2186, doi:10.1001/jama.2017.14580.
26. Bulten, W.; Pinckaers, H.; van Boven, H.; Vink, R.; de Bel, T.; van Ginneken, B.; van der Laak, J.; Hulsbergen-van de Kaa, C.; Litjens, G. Automated Deep-Learning System for Gleason Grading of Prostate Cancer Using Biopsies: A Diagnostic Study. *Lancet Oncol.* **2020**, *21*, 233–241, doi:10.1016/S1470-2045(19)30739-9.
27. Ström, P.; Kartasalo, K.; Olsson, H.; Solorzano, L.; Delahunt, B.; Berney, D.M.; Bostwick, D.G.; Evans, A.J.; Grignon, D.J.; Humphrey, P.A.; et al. Artificial Intelligence for Diagnosis and Grading of Prostate Cancer in Biopsies: A Population-Based, Diagnostic Study. *Lancet Oncol.* **2020**, *21*, 222–232, doi:10.1016/S1470-2045(19)30738-7.
28. Zheng, Q.; Yang, R.; Ni, X.; Yang, S.; Xiong, L.; Yan, D.; Xia, L.; Yuan, J.; Wang, J.; Jiao, P.; et al. Accurate Diagnosis and Survival Prediction of Bladder Cancer Using Deep Learning on Histological Slides. *Cancers (Basel)* **2022**, *14*, 5807, doi:10.3390/cancers14235807.
29. Zheng, Y.; Carrillo-Perez, F.; Pizurica, M.; Heiland, D.H.; Gevaert, O. Spatial Cellular Architecture Predicts Prognosis in Glioblastoma. *Nat. Commun.* **2023**, *14*, 4122, doi:10.1038/s41467-023-39933-0.
30. Iizuka, O.; Kanavati, F.; Kato, K.; Rambeau, M.; Arihiro, K.; Tsuneki, M. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. *Sci. Rep.* **2020**, *10*, 1504, doi:10.1038/s41598-020-58467-9.
31. Abdelsamea, M.M.; Zidan, U.; Senousy, Z.; Gaber, M.M.; Rakha, E.; Ilyas, M. A Survey on Artificial Intelligence in Histopathology Image Analysis. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2022**, *12*, e1474, doi:10.1002/widm.1474.

 doi:10.20944/preprints202604.1876.v1

32. Steiner, D.F.; MacDonald, R.; Liu, Y.; Truszkowski, P.; Hipp, J.D.; Gammage, C.; Thng, F.; Peng, L.; Stumpe, M.C. Impact of Deep Learning Assistance on the Histopathologic Review of Lymph Nodes for Metastatic Breast Cancer. *Am. J. Surg. Pathol.* **2018**, *42*, 1636–1646, doi:10.1097/PAS.0000000000001151.
33. Lin, R.; Chen, Y.; Li, Y.; Tan, Y.; Wang, C.; Wang, Z.; Sun, M.; Wang, L.; Wu, Y.; Ou, Q.; et al. Artificial Intelligence-Powered Copilots for Precision Diagnosis and Surgical Assessment of Histological Growth Patterns in Resectable Colorectal Liver Metastases: A Prospective Study. *Int. J. Surg.* **2025**, *111*, 7939–7955, doi:10.1097/JS9.0000000000002922.
34. Kather, J.N.; Pearson, A.T.; Halama, N.; Jäger, D.; Krause, J.; Loosen, S.H.; Marx, A.; Boor, P.; Tacke, F.; Neumann, U.P.; et al. Deep Learning Can Predict Microsatellite Instability Directly from Histology in Gastrointestinal Cancer. *Nat. Med.* **2019**, *25*, 1054–1056, doi:10.1038/s41591-019-0462-y.
35. Echle, A.; Rindtorff, N.T.; Brinker, T.J.; Luedde, T.; Pearson, A.T.; Kather, J.N. Deep Learning in Cancer Pathology: A New Generation of Clinical Biomarkers. *Br. J. Cancer* **2021**, *124*, 686–696, doi:10.1038/s41416-020-01122-x.
36. Shamai, G.; Livne, A.; Polónia, A.; Sabo, E.; Cretu, A.; Bar-Sela, G.; Kimmel, R. Deep Learning-Based Image Analysis Predicts PD-L1 Status from H&E-Stained Histopathology Images in Breast Cancer. *Nat. Commun.* **2022**, *13*, 6753, doi:10.1038/s41467-022-34275-9.
37. Couture, H.D.; Williams, L.A.; Geradts, J.; Nyante, S.J.; Butler, E.N.; Marron, J.S.; Perou, C.M.; Troester, M.A.; Niethammer, M. Image Analysis with Deep Learning to Predict Breast Cancer Grade, ER Status, Histologic Subtype, and Intrinsic Subtype. *NPJ Breast Cancer* **2018**, *4*, 30, doi:10.1038/s41523-018-0079-1.
38. Fu, Y.; Jung, A.W.; Torne, R.V.; Gonzalez, S.; Vöhringer, H.; Shmatko, A.; Yates, L.R.; Jimenez-Linan, M.; Moore, L.; Gerstung, M. Pan-Cancer Computational Histopathology Reveals Mutations, Tumor Composition and Prognosis. *Nat. Cancer* **2020**, *1*, 800–810, doi:10.1038/s43018-020-0085-8.
39. Tellez, D.; Litjens, G.; Bándi, P.; Bulten, W.; Bokhorst, J.-M.; Ciompi, F.; van der Laak, J. Quantifying the Effects of Data Augmentation and Stain Color Normalization in Convolutional Neural Networks for Computational Pathology. *Med. Image Anal.* **2019**, *58*, 101544, doi:10.1016/j.media.2019.101544.
40. Mittmann, G.; Laiouar-Pedari, S.; Mehrtens, H.A.; Haggenmüller, S.; Bucher, T.-C.; Chanda, T.; Gaisa, N.T.; Wagner, M.; Klamminger, G.G.; Rau, T.T.; et al. Pathologist-like Explainable AI for Interpretable Gleason Grading in Prostate Cancer. *Nat. Commun.* **2025**, *16*, 8959, doi:10.1038/s41467-025-64712-4.