Weakly Supervised Learning-Driven Digital Pathology AI: A Systematic Review from Algorithmic Innovation to Clinical Translation

Yuxi Zhu

School of Biomedical Engineering, Yancheng Normal University, Yancheng, China
2376415233@qq.com

Abstract Digital pathology is undergoing a paradigm shift from conventional microscopic diagnosis to computational pathology, with weakly supervised learning (WSL) serving as a key technology for reducing annotation costs. However, a significant gap remains between algorithmic innovation and clinical deployment. Addressing the core question of how WSL can advance digital pathology AI toward clinical use, this review systematically traces the technological evolution of WSL paradigms, clarifies the relationship between pathology foundation models and WSL, and examines the clinical applications and translational bottlenecks of weakly supervised pathology AI. The review indicates that while WSL substantially lowers the annotation cost barrier, the ultimate clinical value still depends on generalization capability, trustworthiness, and real-world validation.

Keywords weakly supervised learning; digital pathology; foundation models; clinical translation; explainability

1 Introduction

As the "gold standard" of disease diagnosis, pathology plays an irreplaceable role in healthcare systems worldwide. Yet pathology diagnosis currently faces severe challenges: a global shortage of pathologists, substantial inter-observer variability arising from diagnostic subjectivity, low efficiency of conventional microscopic diagnosis, and ever-higher demands driven by the rapid development of molecular pathology and precision medicine [1]. Artificial intelligence (AI) offers transformative solutions to these challenges. Deep learning models have demonstrated the potential to outperform human experts in several pathological diagnosis tasks—in the 2017 CAMELYON16 challenge, deep learning algorithms achieved an AUC of 0.994 for lymph node metastasis detection, significantly surpassing the average performance of pathologists (AUC 0.810) [2].

However, the broad adoption of AI in pathology faces a fundamental bottleneck: extremely high data annotation costs. The gigapixel resolution of whole slide images (WSIs) makes pixel-level annotation practically unscalable, and the fine-grained annotations required for pathological diagnosis demand substantial time from experienced pathologists [3]. Weakly supervised learning (WSL) emerged precisely in this context: by training on data with slide-level labels alone, it greatly reduces the dependence on fine-grained annotation [4].

This review aims to systematically summarize research progress in WSL for digital pathology AI, constructing an integrated knowledge framework from algorithmic innovation to clinical translation. Core application and translational studies covered in this review focus on the period 2019–2026, while a small number of earlier foundational methodological papers (e.g., attention-based MIL from 2018 and the CAMELYON16 evaluation from 2017) are retained to provide theoretical background. Emphasis is placed on work published in top-tier journals such as Nature Medicine, Nature, and The Lancet Oncology.

2 WSL Paradigms and Technological Evolution

In digital pathology, WSL is realized mainly through two families of strategies: slide-level label utilization centered on multiple instance learning (MIL), and weakly supervised localization based on class activation mapping (CAM) and its derivatives. In parallel, self-supervised and contrastive learning serve as auxiliary representation-learning routes that support—but are not equivalent to—weak supervision, and the combination of generative models with WSL is also emerging and is briefly intro-

duced at the end of this section. These strategies are not mutually exclusive but rather show a growing trend of integration.

MIL is the most mature paradigm in weakly supervised pathology AI. In 2018, Ilse et al. proposed attention-based deep MIL (AbMIL), which assigns a learnable weight to each instance through an attention mechanism, replacing fixed pooling strategies and markedly improving both performance and interpretability [5]. In 2021, Lu et al. proposed CLAM (Clustering-constrained Attention Multiple Instance Learning), which further introduces instance-level clustering constraints; it demonstrated excellent performance in renal cell carcinoma subtyping, non-small cell lung cancer subtyping, and lymph node metastasis detection, while localizing morphological features familiar to pathologists without any pixel-level annotation [4]. Subsequently, Shao et al. proposed TransMIL, which introduces Transformer self-attention into the MIL aggregator to capture inter-instance correlations, outperforming conventional AbMIL in multiple cancer classification tasks [6]. CLAM screens the tissue regions of a WSI, extracts patch-level features, and performs attention-based aggregation to identify subregions of high diagnostic value [4].

CAM and its derivatives (e.g., Grad-CAM) initially emerged as post-hoc explanation tools and have since developed into important methods for weakly supervised localization: using only image-level labels, they can generate lesion localization maps, which in pathology have been used to guide instance selection and aggregation calibration in MIL and to provide spatial interpretability for slide-level predictions [7]. Note that such localization results constitute model-driven saliency estimates; their pathological validity must be carefully evaluated together with downstream task performance.

Self-supervised learning (SSL) is inherently distinct from weak supervision. Instead, it acts as an auxiliary representation-learning paradigm for weakly supervised learning (WSL). By learning informative representations from unlabeled data via elaborately designed pretext tasks, SSL establishes a robust foundation for feature extraction in weakly supervised downstream models. Kang et al. conducted systematic benchmarks of multiple SSL approaches across varied pathological datasets and showed that self-supervised pretraining achieves substantially better performance than ImageNet transfer learning [8]. Among pathology-tailored approaches is cross-slide contrastive learning (SCL-WC), which builds positive sample pairs leveraging semantic consistency across different slides acquired from the same patient [9].

An increasingly common engineering pipeline combines these components: obtaining high-quality features through self-supervised pretraining or frozen foundation-model feature extraction, performing slide-level aggregation via MIL, and providing spatial interpretability through CAM/attention maps. The prevalence of this combination in recent studies suggests a gradual convergence of technical routes for weakly supervised pathology AI. Unlike conventional attention-based MIL, which aggregates instances independently, TransMIL further employs a Transformer to model the correlations between patches, thereby enhancing the exploitation of spatial and contextual information in WSIs [6].

As an emerging direction, the combination of generative models with WSL has begun to attract attention: on one hand, generative adversarial networks and diffusion models can synthesize images of rare morphological patterns, providing data augmentation for weakly supervised training under extreme class imbalance [10]; on the other hand, text-conditioned diffusion models such as PathLDM leverage pathology report text to guide image generation, demonstrating the potential of "text–image" weak signals for generative modeling [11]. However, this direction remains at an early stage overall: whether generated data can stably improve slide-level weakly supervised tasks, and whether synthetic images introduce distribution shifts, require systematic validation.

3 Pathology Foundation Models: An Extension of Weakly Supervised Representation Capability

A conceptual boundary must first be clarified: pathology foundation models are typically obtained through self-supervised or vision-language pretraining and are not equivalent to weakly supervised learning per se. The connection between the two lies in the fact that foundation models provide general, transferable image representations—solving the problem of "where representations come from"—while the utilization of slide-level clinical labels still relies on weakly supervised aggregation such as MIL, solving the problem of "how labels are used." Foundation models serve as the representational basis of weakly supervised pathology AI, rather than being weakly supervised methods themselves.

The development of pathology foundation models marks a paradigm shift in computational pathology from "task-specific models" toward "general-purpose pretraining + task-specific fine-tuning." In 2024, Chen et al. proposed the UNI model based on a ViT architecture, pretrained on over one million WSIs, surpassing previous state-of-the-art models across 34 downstream clinical tasks [12]. In the same year, the Virchow model proposed by Vorontsov et al. was trained on over 1.5 million WSIs from multiple institutions, showing improved scalability and robustness in rare cancer and pan-cancer tasks [13]. The GigaPath model proposed by Xu et al. is based on a ViT architecture trained on over one million real-world WSIs, supporting long-sequence modeling to capture global context [14]. Through large-scale self-supervised pretraining on pathology images, UNI obtains general-purpose and highly

ISSN 3106-7018 (Print) ISSN 3106-5791 (Online)

transferable histopathology image representations that can be applied to diverse downstream pathology tasks [12].

Vision-language foundation models further expand the representation space available to weak supervision. CONCH, proposed by Lu et al., was trained via contrastive learning on approximately 1.17 million pathology image-text pairs, demonstrating excellent performance on downstream tasks including classification, retrieval, and weakly supervised prediction [15]. MUSK, proposed by Xiang et al., fuses ViT-L/14 with an LLM, trained on 50 million image patches and one million image-text pairs, focusing on precision oncology tasks and molecular alteration prediction [16].

The value of foundation models ultimately manifests in their transfer learning capability. A widely used paradigm uses a pretrained foundation model as a frozen feature extractor, whose embeddings are fed into a lightweight MIL aggregator for slide-level weakly supervised classification—this directly embodies the aforementioned connection in which "foundation models provide representations while MIL utilizes slide-level labels." The large-scale benchmark by Campanella et al. systematically compared multiple public pathology foundation models as feature extractors for weakly supervised computational pathology, providing an important reference for model selection [17]. The study by Neidlinger et al. further showed that the representation quality of foundation models is closely related to the performance ceiling of downstream weakly supervised tasks [18].

4 Clinical Applications and Translational Bottlenecks of Weakly Supervised Pathology AI

Clinical applications of weakly supervised pathology AI have formed a relatively complete task spectrum. Because the supervision granularity of representative studies varies, we distinguish core WSL evidence from broader clinical background evidence. For diagnostic tasks, cancer detection and grading are the most mature application areas. A weakly supervised system trained with slide-level labels only by Campanella et al. on more than 44,000 WSIs from over 15,000 patients achieved an AUC > 0.98 in prostate cancer, skin cancer, and breast cancer lymph node metastasis detection, and could safely exclude 65%–75% of negative slides at 100% sensitivity [19]. In the PANDA challenge organized by Bulten et al., Gleason grading models built by 1,290 teams with heterogeneous training strategies—ranging from fully supervised to weakly supervised—achieved a Kappa coefficient (approximately 0.86) close to that of expert urological pathologists [20]; as a crowdsourced challenge with mixed supervision, this result is best regarded as clinical background evidence rather than a pure WSL result. Similarly, the clinical validation study by Pantanowitz et al. showed that the Paige Prostate system—a commercial product trained on large-scale data with slide-level diagnostic labels—could improve pathologists' diagnostic sensitivity from 74% to 90% [21], illustrating clinical deployment feasibility rather than a specific WSL methodological advance.

For molecular prediction tasks, Kather et al. first demonstrated that deep learning trained on tiles with slide-level MSI labels (i.e., weak supervision) can directly predict microsatellite instability (MSI) status of gastrointestinal cancers from H&E slides [22]. Building on a weakly supervised attention-based MIL framework, Lu et al. further achieved prediction of the origin of cancers of unknown primary (CUP) from H&E slides, providing an important reference for clinical decision-making [23].

Nevertheless, clinical translation still faces multiple bottlenecks. First, insufficient generalization: models perform excellently on the training data distribution but degrade significantly in cross-institutional validation, and differences in scanners, staining, and tissue processing can all cause marked performance decay [24]. Second, inadequate clinical validation: most studies are retrospective single-center designs lacking prospective multi-center validation and randomized controlled trial (RCT)-level evidence [3]. Third, regulatory pathways remain under exploration: taking the U.S. FDA's publicly maintained list of AI-enabled medical devices as an example, the list snapshot retrieved on September 17, 2026 (covering authorizations through March 2026) included 1,524 AI-enabled devices, of which 1,164 (approximately 76%) belonged to radiology, while pathology accounted for only 9 devices—well below 1% of the list [25]; Paige Prostate Detect became the first FDA-authorized pathology AI product for assisting primary diagnosis of prostate cancer in 2021 [26]. Compared with radiology AI, pathology AI has substantially fewer authorizations and less regulatory experience, and its regulatory pathway awaits further clarification.

5 Explainability and Trustworthiness

Explainability is one of the important foundations for building clinical trust, but it is not the only condition. Explainability methods for pathology AI can be broadly divided into three categories: post-hoc explanations (e.g., Grad-CAM, attention weight visualization), intrinsically interpretable models (e.g., the class-specific attention mechanism of CLAM), and concept-level explanations that associate model decisions with high-level pathological concepts. The pathologist-like explainable AI framework proposed by Mittmann et al. demonstrates an implementation path for mapping model decisions to pathological concepts [27], while related spatial interpretation research systematically examined the correspondence between localization results of weakly supervised models and true lesion regions [7].

Current practice focuses mainly on attention heatmap visualization. Attention maps generated by models such as CLAM can localize diagnostic regions familiar to pathologists, providing clinical users with intuitive decision references [4]. However, attention maps have multiple limitations: attention weights reflect "regions the model considers important" rather than "regions that are truly important pathologically," and the two may be inconsistent; moreover, attention distributions may concentrate excessively on a few regions while overlooking dispersed yet important morphological clues [7]. Attention maps should therefore be regarded as auxiliary cues rather than causal evidence. Through attention weights, CLAM can identify tissue regions with high predictive contribution in WSIs, thereby providing a visual explanation for model predictions [4].

Trustworthiness assessment should not be fixed into a set of dimensions, but should be treated as a multi-faceted problem covering at least: confidence calibration and uncertainty quantification, robustness to domain shifts such as staining and scanner variations, population-level fairness, out-of-distribution (OOD) behavior, and the practical effectiveness of human-AI collaboration. General trustworthy-AI literature suggests that these facets are interrelated and indispensable in AI systems at large [28], and pathology-specific work has begun to confirm the same concerns: domain shift caused by differences in scanners, staining, and tissue preparation is a well-recognized threat to the generalization of computational pathology models [29]; the systematic application of uncertainty quantification remains limited, although clinically oriented studies have shown that uncertainty-informed thresholding can improve reliability under domain shift [30,31]; and performance disparities across demographic subgroups (e.g., differential misdiagnosis rates across population groups) have been empirically documented [32]. Most current pathology AI models still lack systematic confidence calibration and behave unpredictably on out-of-distribution data. Establishing a comprehensive assessment system that balances calibration, robustness, fairness, and human-AI collaboration effectiveness is important for moving pathology AI from "usable" to "trustworthy."

6 Frontier Trends

Multimodal fusion is driving pathology AI from morphology toward integrated morpho-molecular diagnosis. The multimodal whole-slide foundation model proposed by Chen et al. integrates pathology images with genomics data, demonstrating performance beyond unimodal models in cancer diagnosis and prognosis prediction [33]. The rise of spatial transcriptomics has opened a new dimension for multimodal fusion, enabling the prediction of spatial gene expression from H&E slides [34].

At the levels of slide parsing and interaction, hierarchical, multi-object, and multi-scale WSI navigation and parsing methods provide new ideas for efficient information extraction from gigapixel images [35]. Large language models (LLMs) and multimodal generative models bring new capabilities for pathology report processing and interactive diagnostic assistance: pathology vision-language assistants represented by PathChat have reached or exceeded the level of general-purpose multimodal models on multiple diagnostic question-answering tasks [36]; however, the hallucination risks, output reliability, and regulatory compliance of LLMs require careful assessment, and routine clinical use remains distant.

Federated learning provides a solution for privacy-preserving multi-institutional collaboration, promoting multi-center data cooperation through the paradigm of "models move while data stays" [37]. These frontier trends, however, also impose higher demands on WSL, including new challenges such as multimodal weak supervision, federated weak supervision, and real-time weakly supervised inference.

7 Discussion and Conclusion

The core view of this review is that weakly supervised learning has substantially lowered the annotation cost barrier for pathology AI, but the ultimate realization of clinical value still depends on model generalization, trustworthiness, and real-world validation. It should be noted that this review does not construct a novel assessment framework; rather, on the basis of synthesizing existing research, it summarizes assessment priorities at three levels: at the algorithm level, focusing on annotation efficiency, task adaptability, and explainability; at the model level, on representational generality, domain robustness, and scaling efficiency; and at the clinical level, on validation quality, generalization evidence, human-AI collaboration effectiveness, and regulatory status.

For researchers, the priorities follow naturally: design studies around clinical need rather than benchmark convenience; test generalization across institutions instead of assuming it; build and adopt standardized evaluation benchmarks; treat explainability and trustworthiness as design requirements, not afterthoughts; push on few-shot and zero-shot learning; and practice open science. For policymakers we restrict ourselves to what this literature directly supports: clarify the regulatory pathway under which pathology AI software is evaluated and approved as a medical device; set up national or regional evaluation infrastructure with mandatory standardized external validation; and enable multi-institutional data sharing without sacrificing privacy, federated learning being one proven mechanism [1,3,26]. Larger questions—who pays for AI-assisted diagnosis, who is liable when AI contributes to a medical incident, how to train a pathology workforce that uses AI—matter greatly, but they exceed the evidence

ISSN 3106-7018 (Print) ISSN 3106-5791 (Online)

base of this review and deserve separate treatment.

Where does this leave the field? Scaling foundation models further, maturing multimodal fusion, and building multi-institutional and regulatory frameworks all favor the clinical adoption of weakly supervised pathology AI. None of it guarantees adoption. The evidence assembled here points to conditions rather than destiny: prospective multi-center validation, demonstrable cross-institutional generalization, and a growing body of regulatory evidence. Only when these accumulate will the promise of computational pathology be more than a promise.

References

[1] Song AH, Jaume G, Williamson DFK, et al. Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering, 2023, 1(12): 930-949.

[2] Bejnordi BE, Veta M, van Diest PJ, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA, 2017, 318(22): 2199-2210.

[3] Shmatko A, Ghaffari Laleh N, Gerstung M, Kather JN. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. Nature Cancer, 2022, 3(9): 1026-1038.

[4] Lu MY, Williamson DFK, Chen TY, et al. Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering, 2021, 5(6): 555-570.

[5] Ilse M, Tomczak JM, Welling M. Attention-based deep multiple instance learning. ICML, 2018: 2127-2136.

[6] Shao Z, Bian H, Chen Y, et al. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. NeurIPS, 2021, 34: 2136-2147.

[7] Sharma A, et al. A method for spatial interpretation of weakly supervised deep learning models in computational pathology. Scientific Reports, 2025, 15: 1-9.

[8] Kang M, Song H, Park S, et al. Benchmarking self-supervised learning on diverse pathology datasets. CVPR, 2023.

[9] Li B, Li Y, Eliceiri KW. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. CVPR, 2021: 14318-14328.

[10] Quiros AC, Murray-Smith R, Yuan K. PathologyGAN: Learning deep representations of cancer tissue. arXiv:1907.02644, 2019.

[11] Yellapragada A, et al. PathLDM: Text conditioned latent diffusion model for histopathology. IEEE Winter Conference on Applications of Computer Vision (WACV), 2024: 5170-5179.

[12] Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. Nature Medicine, 2024, 30(3): 850-862.

[13] Vorontsov E, Bozkurt A, Casson A, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. Nature Medicine, 2024, 30(10): 2924-2935.

[14] Xu H, Usuyama N, Bagga J, et al. A whole-slide foundation model for digital pathology from real-world data. Nature, 2024, 630(8015): 181-188.

[15] Lu MY, Chen B, Williamson DFK, et al. A visual-language foundation model for computational pathology. Nature Medicine, 2024, 30(3): 863-874.

[16] Xiang J, Zhang J, et al. MUSK: A vision-language foundation model for precision oncology. Nature, 2025, 638: 769-778.

[17] Campanella G, Chen S, Singh M, et al. A clinical benchmark of public self-supervised pathology foundation models. Nature Communications, 2025, 16: 3640.

[18] Neidlinger P, El Nahhas OSM, Muti HS, et al. Benchmarking foundation models as feature extractors for weakly supervised computational pathology. Nature Biomedical Engineering, 2025.

[19] Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine, 2019, 25(8): 1301-1309.

[20] Bulten W, Kartasalo K, Chen PHC, et al. Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. Nature Medicine, 2022, 28(1): 154-163.

[21] Pantanowitz L, Quiroga-Garza GM, Bien L, et al. An artificial intelligence algorithm for prostate cancer diagnosis in whole slide images of core needle biopsies: a blinded clinical validation and deployment study. Lancet Digital Health, 2020, 2(9): e407-e416.

[22] Kather JN, Pearson AT, Halama N, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature Medicine, 2019, 25(7): 1054-1056.

[23] Lu MY, Chen RJ, Kong D, et al. AI-based pathology predicts origins for cancers of unknown primary. Nature, 2021, 594: 106-110.

[24] Bulten W, Pinckaers H, van Boven H, et al. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. Lancet Oncology, 2020, 21(2): 233-241.

[25] U.S. Food and Drug Administration. Artificial Intelligence-Enabled Medical Devices List. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices (data retrieved on September 17, 2026).

[26] van der Laak J, Litjens G, Ciompi F. Deep learning in histopathology: the path to the clinic. Nature Medicine, 2021, 27: 775-784.

[27] Mittmann G, Tschuchnig M, et al. Pathologist-like explainable AI for interpretable Gleason grading. Nature Communications, 2025.

[28] Liu L, et al. Trustworthy AI: A computational perspective. ACM Transactions on Intelligent Systems and Technology, 2022, 13(6): 1-59.

[29] Jahanifar M, Raza M, Xu K, et al. Domain Generalization in Computational Pathology: Survey and Guidelines. ACM Computing Surveys, 2025. DOI: 10.1145/3724391.

[30] Abdar M, Pourpanah F, Hussain S, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion, 2021, 76: 243-297.

[31] Dolezal JM, et al. Uncertainty-informed deep learning models enable high-confidence predictions for digital

histopathology. Nature Communications, 2022, 13: 7252.
[32] Vaidya A, Zhang A, Jaume G, et al. Demographic bias in misdiagnosis by computational pathology models. Nature Medicine, 2024, 30: 1174-1190.
[33] Ding T, Wagner SJ, Song AH, et al. A multimodal whole-slide foundation model for pathology. Nature Medicine, 2025, 31: 3749-3761.
[34] Chen W, Zhang P, Tran TN, et al. A visual-omics foundation model to bridge histopathology with spatial transcriptomics. Nature Methods, 2025.
[35] Tran M, Wagner S, Weichert W, et al. Navigating through whole slide images with hierarchy, multi-object, and multi-scale data. IEEE Transactions on Medical Imaging, 2025.
[36] Lu MY, Chen B, Williamson DFK, et al. A multimodal generative AI copilot for human pathology. Nature, 2024, 634(8033): 466-473.
[37] Talebi N, Azimi M. Federated learning for multi-institutional AI in healthcare via digital pathology. Preprints, 2025.