Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.39443

Deep Learning in Computational Pathology: From Patch-Level Benchmarks to Whole-Slide Clinical Deployment

Licheng Yang

College of Automotive Engineering, Jilin University, Changchun, China
yanglc1523@mails.jlu.edu.cn

Abstract. Cancer diagnosis depends on the visual inspection of stained tissue, a task that is slow, subjective, and under growing workforce pressure. As cancer incidence rises, the shortage of trained pathologists becomes more acute and digitized whole-slide images have opened the door to computational alternatives. Deep learning is leading this shift, but its literature is dominated by isolated accuracy records, lacking a clear route from laboratory benchmarks to clinical deployment. This paper organizes that route explicitly, following computational pathology from patch-level benchmarks to whole-slide clinical use. The survey covers three aspects of the field's development. Firstly, standardized patch-level datasets and the CNN architectures trained on them, from transfer learning to hybrid CNN-Transformer designs, can establish reproducible evaluation and strong baseline performance. Secondly, weakly supervised multiple-instance learning makes pixel-level annotation unnecessary and pathology foundation models push slide-level accuracy toward clinical-grade performance. Thirdly, applications in tissue classification, mutation prediction, and survival prediction show that histology carries molecular and prognostic information beyond visible morphology and by reading them, models can inform diagnosis and support treatment decisions. However, problems such as domain shift across institutions, privacy restrictions on data sharing, and the scarce validation of interpretability tools cause clinical adoption to lag behind. Altogether, data quality, annotation cost, computational feasibility, and clinical trust decide whether an accurate model becomes a usable one. This paper concludes that applying computational pathology to practice requires shared cross-center evaluation protocols, open-source and reproducible methods, and explanation techniques with verified faithfulness.

Keywords: Computational pathology, convolutional neural networks, whole-slide images, foundation models, clinical deployment

1. Introduction

Histopathology is undoubtedly the gold standard for cancer diagnosis. However, as cancer increasingly affects people's lives nowadays, the supply of trained pathologists is lacking and manual slide review is also time-consuming and subjective. Digitization of whole-slide images now makes computational analysis of these slides technically feasible. The early results are striking: in the CAMELYON16 challenge, deep learning algorithms detect tumor tissue with 92.4% sensitivity,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.39443

surpassing the 73.2% achieved by pathologists under the same conditions [1]. This success fueled the application of artificial intelligence from early detection to treatment planning [2].

Deep learning in computational pathology first focused on the patch level. Convolutional neural networks trained on small tiles extracted from whole sections are able to encode surprisingly rich biological information. Coudray et al. predicted non-small cell lung cancer subtypes with an AUC of 0.97 and six common driver mutations from H&E slides alone [3]. Standardized patch-level benchmarks such as those in the MedMNIST suite then became the common ground for comparing models and training strategies [4]. However, patch-level analysis depends on costly pixel-level annotation or heuristic tile selection and ignores the global tissue context pathologists rely on. In addition, it does not scale naturally to gigapixel WSIs, leaving several problems unresolved.

A second wave then shifted the unit of analysis from patches to whole slides. Multiple-instance learning treats each slide as a bag of patches and learns from slide-level labels alone, having no need for pixel-level annotation. Campanella et al. trained a weakly supervised MIL system on 44,732 WSIs and achieved AUCs above 0.98 across three cancer detection tasks [5]. More recently, pathology foundation models that have been trained with millions of slides have begun to provide transferable representations. Multimodal systems like PRISM2 are approaching clinically validated product accuracy without fine-tuning to specific tasks [6]. Built on these foundations, structure-preserving MIL frameworks may further improve both classification accuracy and the interpretability of attention maps [7].

However, obstacles still remain in applying these methods in real-world practice. Model performance degrades across institutions and scanners due to domain shift [8]. Privacy regulations also restrict how patient data can be shared across centers, which motivates federated learning [9]. Clinical practices also demand interpretability and explainability, so methods such as SHAP, LIME, and Grad-CAM have been applied to pathology CNNs, though validation remains limited [10].

This review synthesizes the methodological trajectory of computational pathology from patch-level benchmarks to whole-slide clinical deployment. Section 2 surveys patch-level datasets and CNN architectures, weakly supervised learning and pathology foundation models. Section 3 reviews applications in tissue classification, mutation prediction, and survival prediction, and then discusses the challenges of domain shift, privacy, and explainability. Section 4 gives a conclusion. This paper emphasizes not only reported accuracy in previous studies but also their data, annotation, computational, and trust constraints. These constraints truly determine whether these methods can ultimately aid clinical practice.

2. Datasets and approaches

2.1. Patch-level datasets

At 40× magnification, a typical slide contains billions of pixels. Since a whole-slide image is far too large to process directly, most pipelines cut it into small patches and treat each patch as one training sample. This patch-level unit of analysis creates a need for datasets that standardize sampling, labeling, and evaluation. Patch-level datasets therefore lay the foundation of computational pathology.

Several benchmark collections standardize this practice. MedMNIST v2 is one of the most widely used lightweight suites [4]. It standardizes 18 datasets, including 12 2D and 6 3D, into uniform 28×28 inputs, totaling 708,069 2D and 9,998 3D images, and provides fixed splits for training, validation and testing alongside a common evaluator. PathMNIST is its colorectal histology subset, containing 107,180 patches in nine tissue classes derived from the NCT-CRC-HE-100K cohort [4,

DOI: 10.54254/2755-2721/2026.39443

10]. The 28×28 format, however, trades morphological detail for accessibility. Studies that need native resolution therefore turn to disease-specific collections such as BreakHis [11] and GasHisSDB [12]. Altogether, these collections differ in scale and consist of two regimes of resolution: rapid prototyping and clinically realistic inputs (Table 1).

Table 1. Representative patch-level datasets

Dataset	Modality	Image size	Samples	Classes
MedMNIST v2 (18 subsets) [4]	2D+3D, multiple organs	28×28/ 28×28×28	708,069 + 9,998	varies
PathMNIST [4, 10]	Colorectal H&E	28×28 RGB	107,180	9
NCT-CRC-HE 100K [10]	Colorectal H&E	224×224	100,000	9
BreakHis [11]	Breast H&E	40×–400×	7,909	2
GasHisSDB [12]	Gastric H&E	varies	245,536	2

Lightweight benchmarks lower the entry barrier for method development. Mi et al. pretrained a small multi-task backbone on four MedMNIST tasks and showed that parameter-efficient tuning with LoRA updating only about 5,000 parameters matches full fine-tuning on PathMNIST [13]. These benchmarks also expose more than raw accuracy. Jin combined frozen encoders with lightweight adapters and temperature scaling on MedMNIST+. Expected calibration error was cut from 0.1043 to 0.0084, showing that small benchmarks can evaluate not only accuracy but also reliability [14].

It should also be noted that benchmark data are curated. In real clinical practice, preprocessing pitfalls remain a problem. Digital slides often carry pen markings and other artefacts, so Ali et al. proposed a CNN-based pipeline to remove ink from whole-slide images before analysis [15].

2.2. Deep learning-based models

2.2.1. Transfer learning-based CNNs

Building a deep network from scratch needs far more data than most pathology archives offer. Transfer learning addresses this by fine-tuning ImageNet-pretrained networks. Petriková and Cimrák surveyed this line of work and organized it into three families: pretrained fine-tuning, training from scratch and multi-stage pipelines [16]. Among these, pretrained fine-tuning is the default choice when data are scarce, precisely the situation in most pathology archives.

Representative studies show this recipe works. Liu et al. proposed AlexNet-BC for breast cancer histopathology, combining ImageNet pretraining, data augmentation, and a modified cross-entropy loss that penalizes over-confident predictions [11]. The model surpassed prior state of the art on BreakHis, IDC, and UCSB, and remained robust across magnifications. Beyond customizing a single network, ensembling several transferred models pushes accuracy further. Azmoodeh-Kalati et al. ensembled EfficientNetV1 and EfficientNetV2 variants for binary breast classification and reported 99.58% accuracy with Grad-CAM interpretability [17]. Transfer learning, however, is no guarantee. Vu-Xuan et al. compared five transferred architectures on an ovarian tissue microarray dataset [18]. EfficientNet-B0 performed best in training but dropped to 82.40% test accuracy against

DOI: 10.54254/2755-2721/2026.39443

97.27% in training. The gap is a reminder that a small clinical dataset can still overfit a large pretrained network.

2.2.2. Hybrid CNN-Transformer architectures

Pure CNNs sacrifice global context for local features. Transformers are the opposite, and hybrid designs try to combine both. The case for hybrids rests on a broader trend. Hu et al. traced the progression from handcrafted features to CNNs and visual transformers on gastric histopathology sub-images and found the models are complementary [12]. Each architecture captures something the other misses so combining them is . Sirisha et al. took exactly this path with DualPath-CRC for colorectal histology [10]. One branch uses attention-augmented InceptionNeXt blocks, the other adopts an adaptive-window Swin Transformer, and a gated fusion merges the two paths. The model reached 99.98% accuracy on NCT-CRC-HE-100K with Grad-CAM and LIME maps pointing to morphologically relevant regions. Given that baselines on this benchmark already exceed 99%, the margin should be read as evidence of robustness rather than remaining headroom.

2.2.3. Slide-level learning: MIL and foundation models

Patch-level labels are exceptions in clinical archives. Slide-level weak supervision changes the unit of analysis. Campanella et al. proposed a weakly supervised MIL system that trained only on reported diagnoses and evaluated it on 44,732 whole-slide images [5]. All three cancer-detection tasks reached AUC above 0.98, and the system could exclude 65–75% of negative slides at 100% sensitivity. Weak supervision is not synonymous with MIL. Tellez et al. compressed each gigapixel slide into a compact tensor with unsupervised neural image compression and trained a CNN on that representation instead [19]. Nor is the paradigm limited to cancer: Bussola et al. applied the same approach to rat liver toxicopathology [20].

Foundation models now supply the patch features that MIL aggregates. On the capability side, Vorontsov et al. introduced PRISM2, a multimodal slide-level model trained on 2.3 million slides and 14 million question–answer pairs [6]. With prompt-based inference it matches or exceeds clinically validated products on prostate, breast and breast-lymph-node cancer detection, without task-specific fine-tuning. Bridging the two lines, Waqas et al. combined three foundation backbones with EAGLE-Net, a structure-preserving MIL architecture, and reported up to 3% higher classification accuracy and the best concordance index in six of seven survival tasks [7]. As capability rises, so does model size, and deployment outside large data centers therefore demands efficiency. Cai et al. distilled major foundation models into LitePath, which processes 208 slides per hour on an edge device while preserving diagnostic utility across 45 cohorts [21].

Taken together, these methods trace a clear progression: patch-level benchmarks standardized evaluation, transfer learning and hybrid architectures pushed patch-level accuracy, and slide-level weak supervision with foundation models brought whole-slide analysis within reach.

3. Applications and discussion

3.1. Applications

3.1.1. Tissue classification

Tissue and subtype classification is by far the most mature application. Coudray et al. trained an Inception v3 network on lung whole-slide images and distinguished LUAD from LUSC with an

DOI: 10.54254/2755-2721/2026.39443

AUC of 0.97, comparable to pathologists [3]. Stroma carries signal as well. Bejnordi et al. showed that tumor-associated stroma alone discriminates breast cancer with an AUC of 0.92 [22]. Detection tasks have moved furthest toward practice. A recent review of axillary lymph node assessment reports that deep learning detectors achieve near-perfect sensitivity and are already reducing immunohistochemistry workload in clinical use. Predicting nodal status from primary-tumour morphology, in contrast, remains open, with external AUCs of 0.50–0.83 [23].

3.1.2. Mutation prediction

Histology images carry molecular information. Coudray et al. showed that six of ten recurrent lung-cancer mutations are predictable from H&E slides, with AUCs between 0.733 and 0.856 [3]. The pattern generalizes across cancers with variable strength. Jang et al. predicted five actionable mutations in colorectal cancer and reached AUCs of 0.693–0.809 on frozen TCGA sections [24]. Liao et al. combined diagnostic classification and mutation prediction for hepatocellular carcinoma [25]. Liu et al. predicted IDH status in gliomas and showed that GAN-based augmentation plus clinical covariates raised accuracy from 0.794 to 0.882 [26]. Kosaraju et al. moved toward screening with a deep learning model diagnosing ALK-rearranged lung cancer from H&E with an accuracy over 95%, outputting calibrated uncertainty alongside each prediction [27].

3.1.3. Survival prediction

Prognosis is where histology meets treatment decisions. Loeffler et al. built HIBRID, a deep learning model that predicts recurrence risk in colorectal cancer from H&E slides and combines it with circulating tumour DNA status [28]. The combination stratified patients better than either signal alone. Most importantly, patients negative for molecular residual disease but at high deep-learning risk derived a significant chemotherapy benefit (HR 0.49). This is a concrete case in which a pathology AI influences treatment choices.

3.2. Challenges

3.2.1. Domain shift

Models trained at one institution may degrade at another. Staining, preparation, and scanner differences shift the input distribution. Asadi-Aghbolaghi et al. proposed AIDA, a Fourier-based adversarial domain adaptation method that improved subtype classification across four cancers and multiple centers [8]. Domain shift is thus a statistical barrier between centers. Privacy is a legal one.

3.2.2. Privacy

Histopathology data are personal sensitive data and privacy regulations restrict sharing across institutions. This limits the pooling of large cohorts that deep learning normally requires. Federated learning offers a solution. Instead of moving data, it moves model updates, so each center keeps its slides in place. Andreux et al. showed that SiloBN, which keeps batch-normalization statistics local during federated training, not only transfers better across centers than standard federated averaging but also reduces information leakage [9]. Privacy and generalization can thus be achieved together.

DOI: 10.54254/2755-2721/2026.39443

3.2.3. Interpretability

Clinical trust demands explanations, so saliency methods are required. Sirisha et al. used Grad-CAM and LIME to verify that their colorectal model attends to diagnostically relevant regions [10]. Azmoodeh-Kalati et al. likewise paired their breast-cancer ensemble with Grad-CAM to show that their predictions rely on tissue morphology rather than artefacts [17]. Beyond giving explanations, calibrated uncertainty also supports safer screening. Kosaraju et al. detected ALK-rearranged lung cancer with evidential deep learning, which outputs a confidence estimate alongside each prediction [27]. Such post-hoc explanations are a useful starting point, but their faithfulness should also be verified before informing clinical use.

4. Conclusion

This review provides a comprehensive account of deep learning in computational pathology. It covers patch-level datasets and CNN architectures, weakly supervised multiple-instance learning, and foundation models, together with their applications in tissue classification, mutation prediction, and survival prediction. It can be seen from these studies that the field is advancing quickly. Weak supervision removed the need for pixel annotation and foundation models pushed diagnostic accuracy toward clinical levels. Nevertheless, accuracy alone is not enough. Domain shift still degrades cross-institution performance, privacy restricts the pooling of patient data, and interpretability methods remain rarely validated. This paper therefore argues that data quality, annotation cost, computational feasibility and clinical trust play a major part in whether these methods can finally reach the clinic. Future work should prioritize privacy-preserving federated learning, standardized cross-center evaluation and faithful explanations.

References

[1] Cui, M., & Zhang, D. Y. (2021). Artificial intelligence and computational pathology. *Laboratory Investigation, 101*(4), 412–422.
[2] Kuchana, S. K., Alahari, N. V., Kode, R., Jalapu, K. R., & Alahari, S. K. (2026). Artificial intelligence in oncology: Linking biological discovery to clinical utility. *Molecular Cancer*, in press.
[3] Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., et al. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine, 24*(10), 1559–1567.
[4] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., et al. (2023). MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. *Scientific Data, 10*, 41.
[5] Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., et al. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine, 25*(8), 1301–1309.
[6] Vorontsov, E., Shaikovski, G., Casson, A., Viret, J., Zimmermann, E., Tenenholtz, N., et al. (2026). End-to-end multimodal pathology foundation model with clinical dialogue. *Nature Medicine*.
[7] Waqas, M., Bandyopadhyay, R., Showkatian, E., Muneer, A., Zafar, A., Alvarez, F. R., et al. (2026). The next layer: Augmenting foundation models with structure-preserving and attention-guided learning for local patches to global context awareness in computational pathology. *npj Precision Oncology, 10*, 109.
[8] Asadi-Aghbolaghi, M., Darbandsari, A., Zhang, A., Contreras-Sanz, A., Boschman, J., Ahmadvand, P., et al. (2024). Learning generalizable AI models for multi-center histopathology image classification. *npj Precision Oncology, 8*, 151.
[9] Andreux, M., Ogier du Terrail, J., Beguier, C., & Tramel, E. W. (2020). Siloed federated learning for multi-centric histopathology datasets. *arXiv Preprint*, arXiv:2008.07424.
[10] Sirisha, U., Grandhi, A., Padmavathi, P., Desireddy, J., & Chattu, K. (2026). DualPath-CRC: Adaptive Swin Transformer blocks and attention-augmented InceptionNeXt for classification of colorectal cancer histopathology.

DOI: 10.54254/2755-2721/2026.39443

Artificial Intelligence in Health.
[11] Liu, M., Hu, L., Tang, Y., Wang, C., He, Y., Zeng, C., et al. (2022). A deep learning method for breast cancer classification in the pathology images. *IEEE Journal of Biomedical and Health Informatics, 26*(10), 5025–5033.
[12] Hu, W., Chen, H., Liu, W., Li, X., Sun, H., Huang, X., et al. (2022). A comparative study of gastric histopathology sub-size image classification: From linear regression to visual transformer. *Frontiers in Medicine, 9*, 1072109.
[13] Mi, G., Ye, T., & Wood, D. (2025). A lightweight medical multi-task backbone for cross-modal pretraining and parameter-efficient few-shot transfer on MedMNIST. *Journal of Technology Informatics and Engineering, 4*(3), 572-589.
[14] Jin, A. (2026). Lightweight medical foundation model adapters with probability calibration on MedMNIST+. *Exploration and Research, 1*(1), 14.
[15] Ali, S., Alham, N. K., Verrill, C., & Rittscher, J. (2019). Ink removal from histopathology whole slide images by combining classification, detection and image generation models. *arXiv Preprint*, arXiv:1905.04385.
[16] Petriková, D., & Cimrák, I. (2023). Survey of recent deep neural networks with strong annotated supervision in histopathology. *Computation, 11*(4), 81.
[17] Azmoodeh-Kalati, M., Shabani, H., Maghareh, M. S., Barzegar, Z., & Lashgari, R. (2025). Leveraging an ensemble of EfficientNetV1 and EfficientNetV2 models for classification and interpretation of breast cancer histopathology images. *Scientific Reports, 15*, 21541.
[18] Vu-Xuan, A. T., Nguyen, T. T. N., Tran, L. T., & Nguyen-Tat, T. B. (2024). Deep learning-based classification of ovarian cancer subtypes in histopathology. *Springer Conference Paper*.
[19] Tellez, D., Litjens, G., van der Laak, J., & Ciompi, F. (2020). Neural image compression for gigapixel histopathology image analysis. *IEEE Transactions on Medical Imaging, 39*(2), 590–601.
[20] Bussola, N., Xu, J., Wu, L., Gorini, L., Zhang, Y., Furlanello, C., et al. (2023). A weakly supervised deep learning framework for whole slide classification to facilitate digital pathology in animal study. *Chemical Research in Toxicology, 36*(8), 1321–1331.
[21] Cai, Y., Jin, C., Zhang, Z., Ma, J., Zhou, F., Xu, Y., et al. (2026). A deployment-friendly foundational framework for efficient computational pathology. *arXiv Preprint*, arXiv:2602.14010.
[22] Bejnordi, B. E., Lin, J., Glass, B., Mullooly, M., Gierach, G. L., Sherman, M. E., et al. (2017). Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images. *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, 929–932.
[23] Dedic, V., Kadric, M., Selak, N., & Ceric, T. (2026). Deep learning in breast cancer histopathology: Predicting and detecting axillary lymph node metastasis. *Virchows Archiv*.
[24] Jang, H. J., Lee, A., Kang, J., Song, I. H., & Lee, S. H. (2020). Prediction of clinically actionable genetic alterations from colorectal cancer histopathology images using deep learning. *World Journal of Gastroenterology, 26*(40), 6207–6223.
[25] Liao, H., Long, Y., Han, R., Wang, W., Xu, L., Liao, M., et al. (2020). Deep learning-based classification and mutation prediction from histopathological images of hepatocellular carcinoma. *Clinical and Translational Medicine, 10*(2), e102.
[26] Liu, S., Shah, Z., Sav, A., Russo, C., Berkovsky, S., Qian, Y., et al. (2020). Isocitrate dehydrogenase (IDH) status prediction in histopathology images of gliomas using deep learning. *Scientific Reports, 10*, 7733.
[27] Kosaraju, S. C., Parsa, S. P., Song, D. H., An, H. J., Choi, Y. L., Han, K., et al. (2025). Evidential deep learning-based ALK-expression screening using H&E-stained histopathological images. *npj Digital Medicine, 8*, 610.
[28] Loeffler, C. M. L., Bando, H., Sainath, S., Muti, H. S., Jiang, X., van Treeck, M., et al. (2025). HIBRID: Histology-based risk stratification with deep learning and ctDNA in colorectal cancer. *Nature Communications, 16*, 7561.