Journal of Pathology Informatics 19 (2025) 100514

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

Review Article

Deep learning for digital pathology: A critical overview of methodological framework

Meghdad Sabouri Rad [a], Junze (Vincent) Huang [b], Mohammad Mehdi Hosseini [a], Rakesh Choudhary [a], Harmen Siezen [c], Ratilal Akabari [a], Tamara Jamaspishvili [a], Ola El-Zammar [a], Palak G Patel [a], Saverio J. Carello [a], Michel R. Nasr [a], Bardia Rodd [a,*]

[a] SUNY Upstate Medical University, Syracuse, NY 13210, USA
[b] Columbia University, New York, NY 10027, USA
[c] University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Keywords:
Deep neural networks
Deep learning framework
Digital pathology
Machine learning framework

ABSTRACT

Deep learning frameworks have transformed the field of digital pathology by automating complex tasks and revealing intricate patterns within histopathological data. These advanced methodologies provide exceptional accuracy and scalability, facilitating the analysis of high-dimensional whole-slide images with unparalleled precision. In this article, we present a comprehensive deep learning framework highlighting recent advancements in computational pathology. We critically examine mathematical innovations and offer a comparative analysis of various models demonstrating the significant and ongoing improvements in the field.

Introduction

Digital pathology has revolutionized diagnostic workflows by digitizing histopathological slides into high-resolution whole-slide images (WSIs), significantly enhancing diagnostic accuracy, reproducibility, and efficiency. This transformation is propelled by advancements in machine learning and artificial intelligence (AI), anchored in pathology key elements such as quality control (QC), classification, segmentation, outcome prediction, and clinical decision-making. These foundational components underpin pioneering advances in deep learning frameworks such as neural networks, loss functions, attention mechanisms, and vision transformers (ViTs). Here, we present a comprehensive framework detailing these recent advancements in the field. Deep learning has emerged as a powerful tool in digital pathology revolutionizing the field by enhancing image analysis capabilities and improving diagnostic accuracy. Recent developments have shown that deep learning models can mimic human expertise in recognizing complex visual features in pathological images, often surpassing human capabilities in terms of speed and accuracy.[1,2] These AI-powered tools are being deployed for various applications, including object quantification, tissue classification, and biomarker detection allowing pathologists to automate time-consuming manual tasks and focus on critical decision-making processes.[1,3]

The integration of deep learning in digital pathology has opened up new avenues for cancer detection, diagnosis, and prognosis. Recent studies have demonstrated the potential of AI models in predicting hepatocellular carcinoma (HCC) development using WSIs of liver biopsies.[4] In addition, deep learning algorithms have shown promise in identifying early signs of hepatocarcinogenesis that may have been previously overlooked.[4] These advancements are not limited to a single type of cancer; deep learning models have been successfully applied to various oncological applications including breast cancer detection in lymph nodes and predicting tumor origin in cases of cancer of unknown primary (CUP).[5] As the field continues to evolve, researchers are focusing on developing more sophisticated algorithms to facilitate early and accurate diagnosis across a wide range of gastrointestinal and other cancers.[6,7] But the implementation of deep learning in digital pathology faces several challenges such as: data quality and quantity,[8] computational limitations,[8,9] expertise requirements,[8] transparency and trust,[10] ethical and regulatory concerns,[8,9] pervasive variability,[10] cost and affordability,[9,10] acceptance and cultural change,[9] realism of AI implementation,[10] bias in datasets.[4] Understanding and tackling these challenges are crucial for the successful integration of deep learning in digital pathology and its widespread adoption in clinical practice. This study reviews existing deep learning frameworks and recent advancements in deep learning-based digital pathology. An overview of the reviewed studies is presented in Fig. 1. Fig. 1A–C displays polar charts illustrating: (A) the number of deep learning studies with distinct ML framework contributions, (B) studies contributing to computational pathology, and (C) the number of studies focused on specific diseases. Fig. 1D–E shows 3D and 2D heatmaps

* Corresponding author.
E-mail addresses: bard.rodd@gmail.com RoddB@upstate.edu (B. Rodd).

http://dx.doi.org/10.1016/j.jpi.2025.100514
Received 19 March 2025; Received in revised form 28 July 2025; Accepted 25 August 2025
Available online 8 September 2025
2153-3539/Published by Elsevier Inc. on behalf of Association for Pathology Informatics. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

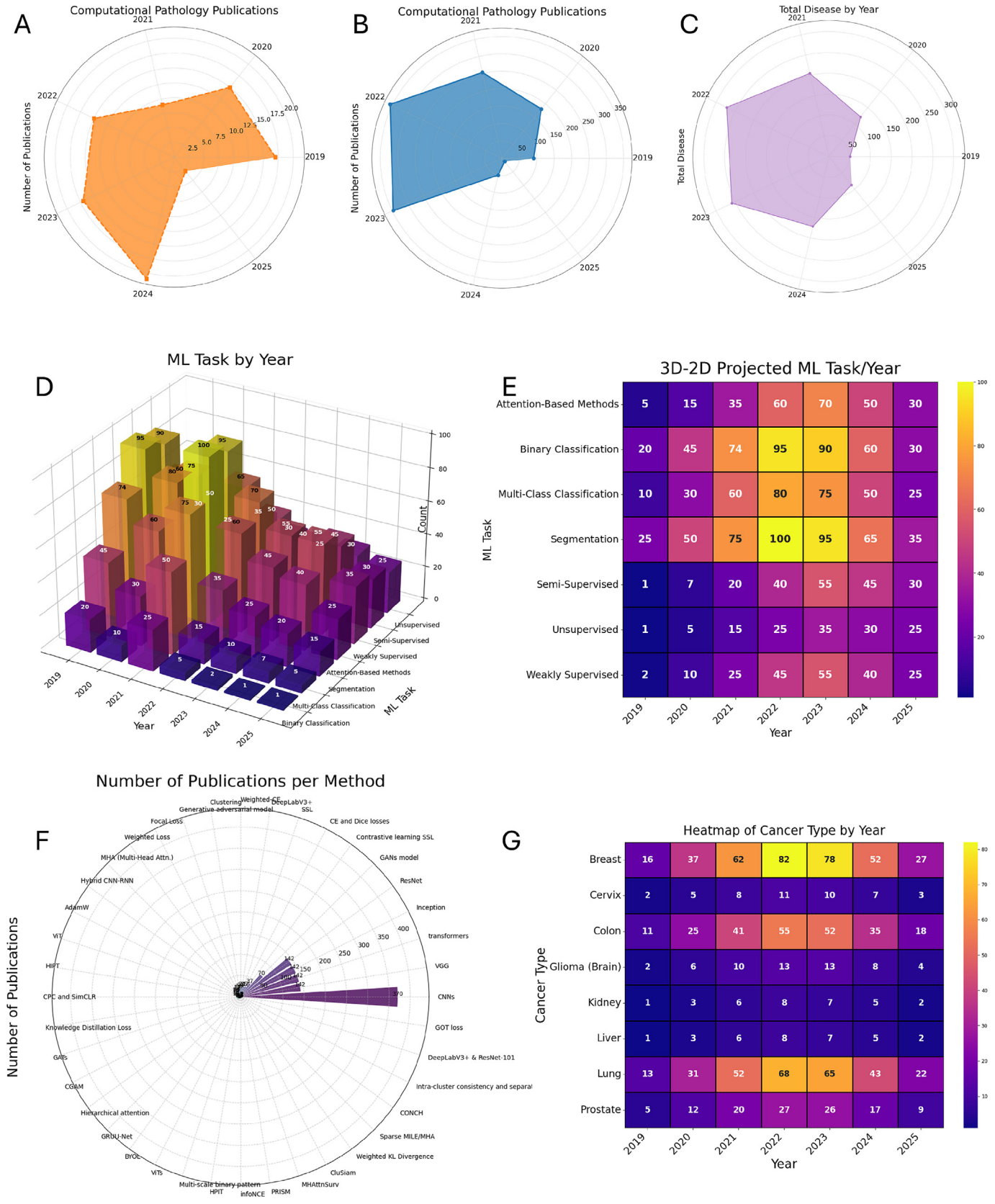


Fig. 1. Summary of reviewed computational pathology studies. A–C: Polar charts illustrating (A) the number of deep learning studies with distinct ML framework contributions, (B) studies contributing to computational pathology, and (C) the number of studies focused on specific diseases. D–E: 3D and 2D heatmaps representing the distribution of various ML tasks over the past 6 years. F: Temporal trend of deep learning methodological studies. G: Yearly heatmap of cancer types studied.

Table 1
Summary of deep learning framework 1: Key innovations in QC, segmentation, and SSL.

Summary of clinical novelty via deep learning framework				
References	Clinical application	Methods	Accuracy (%)	Novel framework proposed
[19–22]	QC-digital ink removal	GANs model	[22]: Increase: SNR:30%, Similarity: 20%, 200% visual info	$\min_G \max_D E_{x\sim p_{\text{data}}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))]$
[18,23]	QC, artifact segmentation detect air bubble, tissue folds, and ink	QC DeepLabV3 +	e.g.,[18] 99.5–100% tissue segmentation	$\text{Dice}_w = \frac{2\sum_i w_i \lvert A_i \cap B_i \rvert}{\sum_i w_i (\lvert A_i \rvert + \lvert B_i \rvert)}$
[33,34]	Histopathological tissue classification	Contrastive learning SSL	Breast, 92.59 and skin, 73.33, cancer	$\mathcal{L}\text{cnrsie} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} [k \neq i] \exp(\text{sim}(z_i, z_k)/\tau)}$
[35]	Survival prediction	PRISM, MHAttnSurv	Breast, colon, glioma, and bladder, c-index of 0.640	$\hat{y} = \sigma\left(\sum_i \frac{\exp(u^T \tanh(Vh_i))}{\sum_j \exp(u^T \tanh(Vh_j))} \cdot (w^T h_i + b)\right)$
[39]	Interactive segmentation of nuclei	CGAM	IoU 85–90	$\mathcal{L}_{CA} = \mathcal{L}_{cik} + \alpha \cdot \mathcal{L}_{sgetto}$ $\alpha_{ij}^{CGAM} = \text{softmax}\left(\frac{q_i \cdot k_j^T + \text{click}_i \cdot w}{\sqrt{d_k}}\right)$
[40]	Classifying tumor grades	SSL, clustering	Endometrial cancer: 90.5	$\mathcal{L}_C = \frac{1}{N}\sum_{i=1}^{N} \lvert z_i - c_{k(i)} \rvert^2$
[47]	Cell segmentation	GRUU-Net	Glioblastoma cell nuclei	$\mathcal{L}\text{sg} = -\frac{1}{T}\sum_{t=1}^{T} (\alpha \cdot \mathcal{L}_{Ce} + \beta \cdot \mathcal{L}_{Dc})$
[61]	Nuclei segmentation	AdamW	Colon cancer, 90.4	$\alpha_{ij}^{(L_2)} = \text{softmax}\left(\frac{q_i^{(l_2)} \left(\sum_k \alpha_{ik}^{(l_1)} v_k^{(l_1)}\right)^T}{\sqrt{d_k^{(l_2)}}}\right)$
[79]	Feature representation	BYOL, SSL	Celiac disease: 87.7, renal cancer: 60.8	$\mathcal{L}_{\text{BYOL}} = \underbrace{\lVert q_\theta(z_{u,1}) - \text{sg}(z_{u,2}) \rVert}_{\text{Loss}_u} + \underbrace{\lVert q_\theta(z_{l,1}) - \text{sg}\left(\bar{z}_{l,2}\right) \rVert}_{\text{Loss}_l}$
[39]	Segmentation	CGAM, DeepLabV3 + & ResNet-101	Liver cancer: 92.2 (IoU)	$\mathcal{L}_t(I, C_t) = \min_{\theta_t} \mathbb{E}_{i \in [1,t]} \left[max\left(l_i - \hat{f}(I, C_t; \theta_t)_{u_i, u_v}, 0\right)\right]^2 + \lambda \lVert M \odot \left(\hat{f}(I, C_t; \theta_{t-1}) - \hat{f}(I, C_t; \theta_t)\right) \rVert_2^2$

representing the distribution of various ML tasks over the past 6 years. Fig. 1F presents the temporal trend of deep learning methodological studies and Fig. 1G shows a yearly heatmap of cancer types studied. (See Table 1)

Quality control in digital pathology

This review delves into the mathematical foundations driving QC in digital pathology, particularly for WSI, where artifacts like focus errors and staining variability compromise diagnostic precision. It examines the evolution from classical methods to advanced hybrid frameworks emphasizing their role in enhancing efficiency, accuracy, and scalability. Mathematical rigor remains crucial in artifact detection, quality assessment, and adaptive integration within pathology workflows (Table 2). One of the initial works on QC is color normalization through histogram matching, which addresses staining variability by aligning the intensity histograms of an input image I with a reference R. The transformation uses cumulative distribution functions (CDFs):

$$p'(x) = H_R^{-1}(H_I(p(x))) \tag{1}$$

where H_I and H_R are the CDFs of I and R, respectively, which are defined $H_I(p) = \Sigma_{i=0}^{p} h_I(i), p \in [0, L-1]$, and $H_R(p) = \Sigma_{i=0}^{p} h_R(i)$. This method uses precomputed lookup tables for efficiency to standardize color appearance across slides. However, in practice, several real-world challenges arise. The method remains sensitive to noise, whereas it enhances minor variations in poorly stained or background regions. It requires consistent tissue coverage and stain types between input and reference images but this assumption rarely holds true in multi-institutional datasets. Computationally, the method faces challenges when applying patches across large WSIs because it needs substantial memory management and tile-based consistency. The lack of semantic understanding in histogram matching leads to overcorrection of diagnostically relevant features, particularly in regions with mixed tissue types and artifacts such as folds or debris.

To enhance local contrast in such settings, contrast-limited adaptive histogram equalization (CLAHE) has been widely adopted. CLAHE divides the image into smaller regions, applies histogram equalization with a contrast-limiting threshold to each region, and interpolates results to ensure smooth transitions across boundaries. This approach improves contrast while controlling noise amplification, key in pathology, where subtle intensity variations are diagnostically meaningful. The enhanced intensity I' is computed as:

$$I' = I + \alpha(\mu - I) \tag{2}$$

and is often used to represent global contrast stretching or simple mean-based adjustment, where μ is a local or global mean and α is a scaling factor. Although this formulation provides an intuitive sense of brightness

Table 2
Summary table: focus of key clinical studies.

Study/Review	Focus on QC in digital pathology	Application of YOLO	Combined focus (QC + YOLO)
MSKCC QMS (2023)[24]	Yes – Emphasis on standardized QC procedures for pathology workflows	No – YOLO not utilized in this study	No – No integration of YOLO in QC workflows
Effect of QC on image analysis (2021)[25]	Yes – Evaluates QC impact on downstream analysis accuracy	No – YOLO not applied	No – Separate focus on QC alone
YOLO for pressure ulcer detection (2023)[26]	Indirect – Dataset curation involved basic QC principles	Yes – YOLO used for detection tasks	No – No explicit integration of QC with YOLO
Artifact augmentation/Detection (2024)[19]	Yes – Focused on detecting artifacts affecting model performance	Indirect – Methodology supports YOLO-like architectures	No – Not combined but highly relevant for future integrations
YOLO in early EC detection (2024)[27]	Indirect – QC addressed through model performance metrics	Yes – YOLO applied to detect early-stage endometrial cancer	No – No direct QC-YOLO integration

normalization, it does not capture the histogram-driven nature of CLAHE. In contrast, CLAHE divides the image into non-overlapping tiles, clips each local histogram $H(I)$ at a threshold T to limit contrast amplification, redistributes the clipped pixels, and equalizes the resulting histogram. The enhanced tiles are then bilinearly interpolated to avoid boundary artifacts:

$$H_c(I) = \min(H(I), T). \tag{3}$$

Excess values are redistributed and adjusted CDFs map intensities for dynamic contrast enhancement. Boundary artifacts are minimized using weighted interpolation of intensities between neighboring regions. CLAHE effectively balances fine detail preservation with global contrast consistency making it ideal for heterogeneous image regions.[11–13]

Sharpness in digital pathology images is evaluated using gradient and Laplacian operators to quantify intensity variations. The gradient magnitude measuring local intensity changes, as $|\nabla I(x,y)|$. Sharper regions exhibit higher gradient values. The Laplacian operator, capturing second-order variations, calculates sharpness as $\Sigma_{(x,y)\in P}|\nabla^2 I(x,y)|$. Blur detection thresholds Laplacian values to classify regions, whereas extended metrics like Laplacian variance improve sensitivity to subtle focus inconsistencies, are shown as:

$$V_{\text{Lap}} = \frac{1}{|P|}\sum_{(x,y)\in P}\left(\nabla^2 I(x,y) - \mu_{\nabla^2 I}\right)^2 \tag{4}$$

Where $\mu_{\nabla^2 I}$ is the mean Laplacian value within region P. These formulations enable adaptive QC systems to detect and address blurred areas in pathology images efficiently.[14–16] Logistic regression is employed in HistoQC for artifact detection ensuring interpretable and reproducible QC, which achieves robust and adaptable artifact detection across diverse datasets.[12,14]

YOLO framework streamlines artifact detection by dividing images into grids, predicting bounding boxes, and classifying objects simultaneously. YOLO employs a composite loss function incorporating localization loss for bounding box accuracy, confidence loss for object presence probability, and classification loss for correct object labeling. Non-maximum suppression refines results by eliminating redundant predictions retaining the most accurate bounding boxes and reducing false positives. Adjustable parameters like grid size and the intersection of union (IoU) thresholds enable customization for specific tasks, balancing precision and recall. However, YOLO's performance can be affected by failure modes such as false positives when artifacts mimic tissue morphology or false negatives in underrepresented artifact types. These limitations highlight the importance of diverse, well-annotated training data and domain-specific tuning to improve generalization in pathology settings.With real-time performance and adaptability, YOLO effectively identifies artifacts and structural abnormalities in diverse pathology datasets supporting high-throughput QC workflows (see Fig. 2).[17–19]

For example, Hemmatirad et al.[18] demonstrated that a YOLO-based framework could successfully detect nine distinct types of artifacts in WSIs including blur, bubbles, and marker ink, achieving performance crucial for automated QC pipelines.

ResNet mitigates vanishing gradients in deep neural networks by learning residual functions instead of direct mappings. The residual transformation is defined as:

$$F(x) = H(x) - x \tag{5}$$

where $H(x)$ is the target function and x is the input. Reformulating the residual block as $H(x) = F(x) + x$, a shortcut connection bypasses deeper layers, ensuring effective gradient flow during backpropagation. For a layer with weights W_i and activation σ, $F(x)$ is:

$$F(x) = \sigma(W_2\sigma(W_1 x) + b_1) + b_2. \tag{6}$$

Gradient preservation through the shortcut is expressed as:

$$\frac{\partial H(x)}{\partial x} = \frac{\partial F(x)}{\partial x} + 1. \tag{7}$$

The residual formulation proves most beneficial for pathology-specific networks because it enables deep models to detect both low-level morphology and high-level contextual features across extensive WSIs. The residual connections enable efficient learning of deeper architectures by preventing gradient vanishing which allows for multi-scale feature extraction. The detection of tissue folds and blur artifacts requires both fine-grained texture

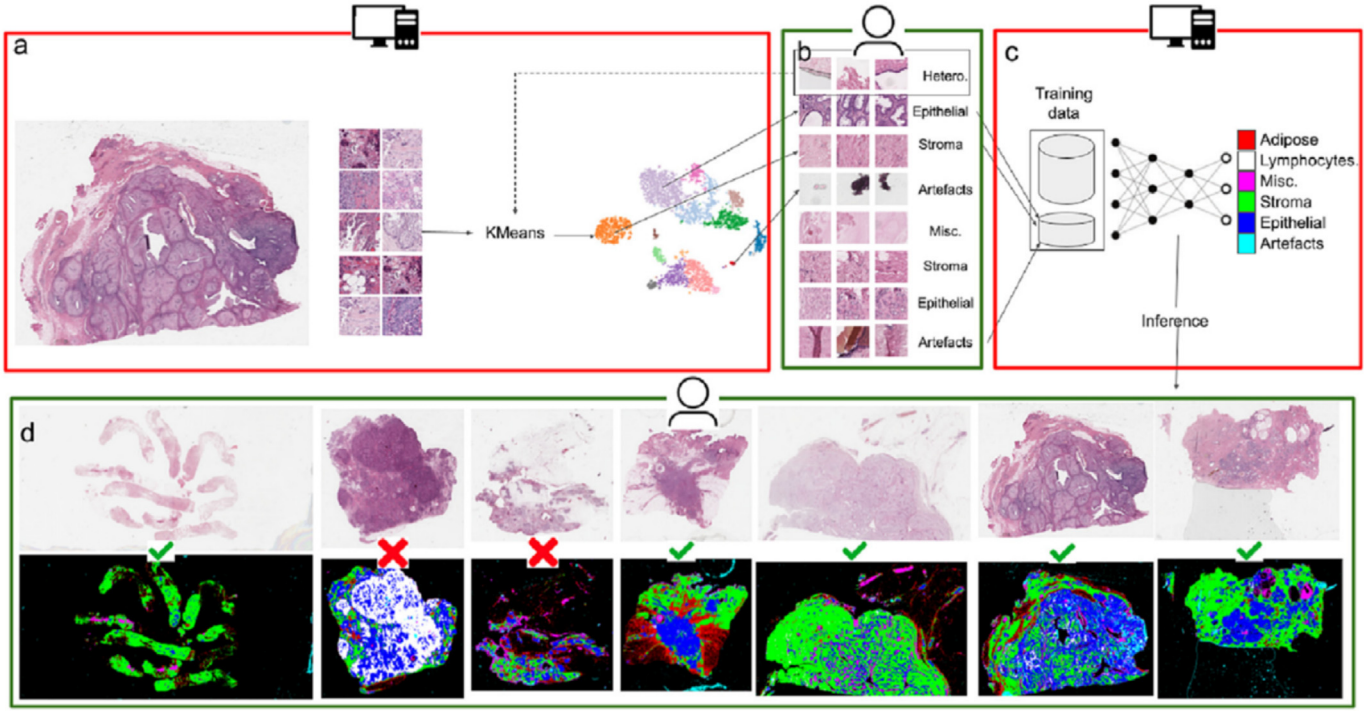


Fig. 2. HistoROI and clustering methods used for embedding patches to clusters for model training (image is taken from[17]).

modeling and slide-level spatial coherence, which deeper stable networks effectively handle. The residual blocks enable shallow-layer gradients to pass through multiple layers without loss, so WSI-based QC workflows achieve better convergence and robustness.[14,17,18,20]

Deep learning models in QC reshaped by generative adversarial networks (GANs). GANs employ adversarial training to correct artifacts and enrich datasets. Comprised of a generator G and discriminator D, GANs operate within a minimax framework, where G synthesizes artifact-free images and D differentiates real from generated data. The optimization goal is:

$$\min_G \max_D E_{x\sim p_{\text{data}}(x)}[logD(x)] + E_{z\sim p_z(z)}[log(1 - D(G(z)))] \quad (8)$$

Where $p_{\text{data}}(x)$ is the real data distribution and $p_z(z)$ is the prior distribution of the latent variables z. The generator attempts to minimize this objective, whereas the discriminator maximizes it forming a two-player adversarial game. The discriminator loss L_D is defined as:

$$L_D = -E_{x\sim p_{dt}}[\log D(x)] - E_{z\sim p_z}[\log(1 - D(G(z)))]. \quad (9)$$

This loss penalizes the discriminator for incorrectly classifying real images x or generated images $G(z)$. The generator loss L_G is defined as:

$$L_G = -E_{z\sim p_z}[\log D(G(z))]. \quad (10)$$

The pathology field faces difficulties in stable adversarial training because of stain variability and structural heterogeneity. The training of GANs on limited distributions results in mode collapse because the generator produces restricted artifact patterns while failing to detect infrequent variations. The generator fails to detect or distorts both chromatin texture and weak staining gradients during the training process. To improve training stability and enhance convergence, a Wasserstein loss function can be employed:

$$L_W = E_{x\sim p_{dt}}[D(x)] - E_{z\sim p_z}[D(G(z))] + \lambda \cdot E_{\widehat{x}\sim p_{\widehat{x}}}\left[\left(\left|\nabla_{\widehat{x}} D(\widehat{x})\right|_2 - 1\right)^2\right]$$

where λ is the regularization coefficient and $\hat{x}$ is the interpolated data point between real and generated samples. This loss improves gradient flow and mitigates vanishing gradients during training. In histopathology, where data are high-dimensional and tissue appearance varies subtly across slides, the gradient penalty enforces Lipschitz continuity, stabilizing learning. It also reduces sensitivity to outlier patches and helps retain fine-grained detail in generated images making it preferable to standard GAN loss for QC applications. These formulations enable iterative refinement of synthetic data and robust correction of challenging artifacts. GANs, by leveraging adversarial learning principles, are highly adaptable to diverse imaging conditions making them essential for augmenting and improving the quality of training datasets in pathology.[19,21,22] GANs are also used for expanding dataset diversity by synthesizing images to augment training data. The augmented dataset is:

$$D_{\text{aug}} = D_{\text{clean}} \cup \{G(z)|z\sim p_z(z)\} \quad (11)$$

where $G(z)$ generates synthetic images from latent variable z, sampled from a prior $p_z(z)$. This enhances variability improving model generalization across diverse tissues and imaging conditions. To optimize augmentation, weighted sampling prioritizes diverse or rare artifacts, with weights defined as:

$$w(z) = \frac{1}{\|\nabla G(z)\| + \epsilon}$$

where $\|\nabla G(z)\|$ is the gradient norm and ϵ ensures stability. Whereas augmentation improves robustness,[21,22] overreliance on GAN-generated data may propagate artifacts or introduce unrealistic patterns highlighting the need for expert validation. Clinically, this underscores the importance of digitally removing ink from WSIs, which led to a 30% increase in peak signal-to-noise ratio, a 20% improvement in structural similarity index, and a 200% gain in visual information fidelity compared to previous methods. For semantic segmentation, QC DeepLabV3 + employs an encoder-decoder structure for multi-class artifact segmentation, capturing both local and global contexts (Table 2). Feature extraction at layer I is modeled as:

$$f_l = \sigma(W_l * f_{l-1} + b_l) \quad (12)$$

where W_l and b_l are layer weights and biases and σ is the activation function. Atrous convolution enhances receptive fields without extra computational cost:

$$f_{\text{atrous}}(x) = \sum_k w_k \cdot x_{r+d_k} \quad (13)$$

where d_k is the dilation rate. Segmentation accuracy is measured by the Dice score, which is also modified to a weighted Dice coefficient that balances for underrepresented classes:

$$\text{Dice}_w = \frac{2\Sigma_i w_i |A_i \cap B_i|}{\Sigma_i w_i (|A_i| + |B_i|)}.$$

This framework enables precise segmentation of artifacts in diverse pathology images,[18,23] which led to locating artifacts with 99.5% IoU accuracy.[18]

The convergence of QC and AI in digital pathology workflows

QC in digital pathology is rapidly evolving from handcrafted heuristic methods to integrated AI-enabled systems capable of operating at the clinical scale. Classical techniques like histogram matching address foundational QC tasks, whereas adaptive methods such as CLAHE enhance local contrast in stained tissue regions.[14] Modern deep learning frameworks, including YOLO for real-time artifact detection,[28] ResNet for deep residual learning,[18] GAN for data augmentation and stain normalization,[21,22] and DeepLabV3 + for semantic segmentation[23] enable precise, scalable, and context aware quality assessment across diverse datasets and imaging platforms. This automation is essential, as subtle image artifacts can significantly compromise the performance and reliability of subsequent diagnostic algorithms. For instance, clinical studies have demonstrated a critical trade-off, where filtering out poor-quality images improves algorithm–pathologist agreement but can simultaneously reduce dataset diversity, potentially harming model robustness and generalizability.[25] These advances reflect a larger shift toward scalable QC pipelines exemplified by platforms like HistoQC[14] and YOLO-based modules[28] that integrate traditional image analysis with machine learning. Such platforms serve as the technical backbone for a comprehensive quality management system (QMS), which formalizes QC across the entire preanalytical, analytical, and postanalytical workflow to ensure end-to-end reliability (Table 2). A key clinical guideline for this is the QMS implemented at Memorial Sloan Kettering Cancer Center, which systematically mapped the digital pathology workflow to establish standard operating procedures, training protocols, and continuous oversight through regular audits and performance metrics, such as slide defect rates and scan errors.[24] In practice, however, deploying these models introduces significant barriers, such as compatibility with hospital IT systems (e.g., LIS/PACS), DICOM integration, computational latency, and infrastructure demands such as GPU availability. Security, version control, and regulatory compliance (e.g., HIPAA, GDPR) further complicate clinical adoption.[12] Real-world applications of these frameworks highlight their impact. For example, YOLO modules support live scanning QC,[28] leveraging a framework already clinically validated for diagnostic tasks such as pressure ulcer classification[26] and early-stage endometrial cancer detection.[27] ResNet-based models enhance feature representation across multiple tissue scales,[18] whereas GANs enable stain robust augmentation and the digital removal of ink markings from WSIs.[21,22] By analyzing the capabilities and limitations of current QC methods, this review supports the field's progression from research prototypes to clinical-grade deployable tools. This is advanced by proposing

actionable strategies such as artifact augmentation,[19] active learning,[19] pathology-specific transfer learning,[17] and IT-aligned implementation.[12] However, despite the clear alignment of YOLO's capabilities with QC needs, explicit clinical studies validating its direct use for automating QC in pathology workflows remain limited. Bridging this gap is a critical next step to increase adoption and ensure the end-to-end integrity of digital pathology in clinical practice.

Self-supervised learning

SSL has emerged as a key solution, enabling models to extract meaningful representations from unlabeled data through innovative pretext tasks (Table 2). This paradigm shift from supervised to self-supervised approaches marks a transformative step in computational pathology. Early advancements, such as contrastive learning frameworks[29] and masked region prediction tasks,[30] introduced context-aware feature extraction without reliance on explicit labels. Recent SSL innovations address domain-specific challenges like staining variability and scanner inconsistencies.[31] These approaches, including clustering-based and hybrid frameworks,[32] leverage multi-scale histological patterns while maintaining biological relevance. SSL's theoretical foundation enables robust feature learning, optimizing models for classification, segmentation, and clustering in sparse annotation scenarios. This analysis underscores the formulation and optimization strategies driving SSL's success in pathology emphasizing its clinical utility and adaptability to complex datasets. Such models rely on composite loss functions to imply self-supervised scenarios, such as contrastive loss.[17]

Contrastive learning optimizes feature representations[33] by bringing similar samples closer in the embedding space while pushing dissimilar ones apart. This is particularly useful in pathology for capturing morphological similarities and differences in tissue patches, which has substantial application in histopathology tissue classification. The key mathematical objective of contrastive learning is the normalized temperature-scaled cross-entropy (NT-Xent) loss, defined as:

$$L\text{cnrsie} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N}[k \neq i]\exp(\text{sim}(z_i, z_k)/\tau)} \tag{14}$$

Where z_i and z_j are the normalized embeddings of positive pairs (e.g., augmented views of the same patch), $\text{sim}(z_i, z_j) = \frac{z_i \cdot z_j}{|z_i||z_j|}$ is the cosine similarity between embeddings, τ is the temperature parameter controlling the distribution sharpness, and N is the number of positive pairs in the batch, $k \neq i$. This loss encourages the model to learn invariant representations by maximizing agreement between augmented views of the same patch while minimizing similarity to other patches. In pathology, domain-specific augmentations, such as color jittering to simulate staining variability or rotation to mimic changes in tissue orientation enhance the robustness of contrastive learning.

Clustering-based approaches in SSL aim to group similar patches in the feature space without explicit annotations. Methods like DeepCluster and SwAV iteratively alternate between clustering features and updating network parameters. The clustering objective is formulated as:

$$L_{\text{cutrn}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} q_{i,k}\log p_{i,k} \tag{15}$$

Where $q_{i,k}$ is the target probability of sample i belonging to cluster k, and $p_{i,k}$ is the predicted probability of sample i belonging to cluster k, computed using the softmax of cosine similarity:

$$p_{i,k} = \frac{\exp(\text{sim}(z_i, c_k)/\tau)}{\sum_{j=1}^{K}\exp(\text{sim}(z_i, c_j)/\tau)}$$

where c_k is the centroid of cluster k. Clustering methods allow SSL models to identify and exploit latent patterns in pathology data, such as grouping similar tissue morphologies or subtypes. Incorporating multi-scale features into clustering further improves granularity, as patches from different magnifications (e.g., 20×, 10×) provide complementary information.[34] Inspired by natural language processing, masked image modeling predicts missing image regions to enforce context-aware feature learning. The loss function for this task is:

$$L_{\text{mse}} = \frac{1}{M}\sum_{i \in \text{masked}}|\widehat{z}_i - z_i|^2 \tag{16}$$

where M is the number of masked patches, $(\widehat{z}_i)$ is the predicted feature representation, and z_i is the ground truth. Masked image modeling, employed in transformer-based frameworks such as PRISM, is particularly effective for learning slide-level representations for cancer survival prediction, for breast, colon, glioma, and bladder cancers, with an average c-index of 0.640[35] (Table 2). The jigsaw puzzle pretext task involves shuffling patches from an image and training the model to predict their correct order. This helps the model learn spatial relationships, which are crucial in pathology for identifying tissue structures. Mathematically, the jigsaw puzzle task optimizes a classification loss:

$$L_{\text{jga}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{P} y_{i,j}\log\widehat{y}_{i,j} \tag{17}$$

Where $y_{i,j}$ is the true label for the j-th patch order in the i-th image and $\widehat{y}_{i,j}$ is the predicted probability for the j-th patch order. In pathology, this task enables models to capture spatial context and structural coherence in WSIs, which is vital for tasks like tumor boundary identification.[36,37] Another popular pretext task involves predicting the rotation angle applied to an image patch. By learning to distinguish rotations (e.g., 0°, 90°, 180°, and 270°), the model develops an understanding of texture and orientation. The rotation prediction task is framed as a classification problem with the loss:

$$L_{\text{rtto}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{r=1}^{R} y_{i,r}\log\widehat{y}_{i,r} \tag{18}$$

Where $y_{i,r}$ is the true label for the rotation angle r applied to the i-th image and $\widehat{y}_{i,r}$ is the predicted probability of the angle r. In pathology, rotation prediction aids in learning rotation-invariant features, which are particularly useful for processing WSIs with diverse orientations.[38] Generative tasks, such as those involving GANs, are employed to generate synthetic tissue samples or reconstruct missing features. The adversarial loss for GANs is:

$$L_{\text{GN}} = E[\log D(x)] + E[\log(1 - D(G(z)))] \tag{19}$$

where $D(x)$ is the discriminator's prediction for real data x and $G(z)$ is the generator's output for random noise z. GAN-based tasks improve the robustness of SSL frameworks by enhancing data diversity (The segmentation of nuclei from histopathology images with synthetic data). Interactive pretext tasks, such as those using click-guided attention modules (CGAM) (Table 2), integrate user interactions to refine segmentation models. In these tasks, the model learns to incorporate user-provided clicks as guidance minimizing a combined loss:

$$L_{\text{CA}} = L_{\text{cik}} + \alpha \cdot L_{\text{sgetto}} \tag{20}$$

where L_{cik} penalizes deviations from user-provided clicks and L_{sgetto} measures segmentation accuracy.[39] Representation learning is a fundamental aspect of SSL for pathology. It focuses on extracting discriminative features from raw data, such as tiles or patches from WSIs and optimizing these embeddings for downstream tasks. Representation learning typically involves three key components: feature extraction, embedding optimization, and similarity-based mathematical insights, all of which are detailed below. CNN models are widely used for feature extraction transforming raw image tiles into meaningful feature representations. The feature extraction process can be mathematically represented as:

$$h_i = f_{\text{CNN}}(x_i; \theta) \tag{21}$$

where h_i is the feature vector for the i-th tile x_i and f_{CNN} is the CNN function parameterized by trainable weights and biases θ. Frameworks like PRISM and EndoNet have successfully used CNNs for extracting multi-scale tile embeddings. To enhance feature extraction, multi-scale CNNs can integrate features from different magnifications to capture both fine-grained and global patterns simultaneously:

$$H = \text{Concat}(H_{20\times}, H_{10\times}, H_{5\times}) \tag{22}$$

where H_{kx} represents features extracted at magnification Kx.[40] In this process, embedding optimization refines extracted features into a meaningful and discriminative embedding space. In this section, we focus on clustering-based approaches. Clustering-based approaches group similar tiles into clusters, minimizing intra-cluster variability while maximizing inter-cluster separation among the endometrial cancer grades. The optimization objective is defined as:

$$L_C = \frac{1}{N}\sum_{i=1}^{N} \left| z_i - c_{k(i)} \right|^2 \tag{23}$$

Where z_i is the embedding of the i-th tile, $c_{k(i)}$ is the centroid of the cluster assigned to z_i, and $k(i)$ represents the cluster assignment representing different tumor grades. This technique has been extensively applied in pathology to identify morphological subtypes or tissue patterns. By capturing complex patterns within tissue structures, SSL features enhance the ability to link morphological variations to clinical outcomes, such as patient survival or disease progression. Frameworks like MHAttnSurv effectively combine SSL representations with attention mechanisms to prioritize prognostically relevant regions in WSIs improving prediction reliability and clinical impact.[35,41] Deep learning models have become central to the advancement of computational pathology enabling automated and highly accurate analysis of complex histopathological data. As illustrated in Fig. 3, a growing number of studies have utilized diverse datasets to train and validate deep learning models reflecting their adaptability across various pathology tasks (Fig. 3A). Moreover, deep learning methods have been widely applied to disease-specific applications with a noticeable concentration on major cancer types (Fig. 3B). The network diagrams in Fig. 3C and D highlight the evolving methodological landscape showing how innovations in deep learning frameworks are increasingly tailored to address disease-specific challenges and data availability.

Recurrent neural networks

Recurrent neural networks (RNNs) are commonly used for sequential feature integration, as they process a sequence of tile features in a temporal order. Each tile's embedding updates a hidden state, s_t, which accumulates information about the sequence up to that point. This enables the model to learn dependencies between neighboring tiles capturing morphological patterns across regions. For example, when analyzing ducts, vessels, or tumor margins, RNNs can model the gradual changes in tissue structure, which are diagnostically relevant. The update rule for RNNs is:

$$s_t = \sigma(W_h h_t + W_s s_{t-1} + b) \tag{24}$$

where the trainable weights, W_h and W_s, allow the network to adapt to the morphological variations across tiles.[32,35] To further enhance the ability to model spatial relationships, bidirectional RNNs (BiRNNs) extend standard RNNs by incorporating both forward and backward dependencies in the sequence. This is particularly useful in pathology, where tissue structures often exhibit complex spatial interactions. For example, tumor regions may influence the surrounding stroma and vice versa. The hidden states from the forward and backward passes are concatenated:

$$s_t^B = \text{Concats}_t^{\text{frad},s_t^{\text{bcwr}}}. \tag{25}$$

This bidirectional approach has been shown to improve the model's ability to capture long-range dependencies enhancing its ability to differentiate between subtle variations in tissue morphology.[35]

PHTNet, proposed by Shahtalebi et al. (2020),[42] is a BRNN designed for pathological hand tremor detection. Using gated recurrent units (GRU) cells, it processes sequential motion data to filter involuntary tremors while predicting voluntary motion. GRU mechanisms, with reset (r_t) and update (z_t) gates, regulate information flow as:

$$r_t = \sigma(W_r m_t + U_r h_{t-1} + b_r) z_t = \sigma(W_z m_t + U_z h_{t-1} + b_z).$$

The updated hidden state, h_t, is computed as:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \text{ReLU}(W_u m_t + U_u(r_t \odot h_{t-1}) + b_u). \tag{26}$$

Mathematically, PHTNet minimizes training errors, $E_i = \text{Error}\left(m_i^{GT}, \hat{m}_i(t:t)\right)$ and evaluates convergence using mean validation error, $\overline{E}_{\text{validation}} = \frac{1}{N}\Sigma_{i=1}^{N} E_i$. To improve robustness, it generates synthetic tremor data as $m_i(t:t) = \Sigma_j s_i(j)$. For real-time applications, it minimizes shifted mean square error (MSE). PHTNet's bidirectional structure enables accurate filtering of tremor noise making it a powerful tool for medical diagnostics and real-time rehabilitation. Sequential feature integration methods ultimately generate a slide-level embedding that aggregates information from all tiles. This slide-level representation, H_{slide}, is produced by applying the sequential model (e.g., RNN or transformer) to the sequence of tile embeddings, $H_{\text{slide}} = f_{\text{sequential}}(H)$, where $H = [h_1, h_2, \ldots, h_T]$ represents the sequence of tile embeddings. The slide-level embedding can then be used for downstream tasks such as classification or survival prediction. The model's parameters are optimized using a binary cross-entropy (CE) loss.[35,43,44] RNNs have become integral to digital pathology due to their ability to model sequential data capturing both short- and long-term dependencies critical for tasks like histopathological image analysis and disease detection.[45,46] Advanced architectures, such as long short-term memory (LSTM) and GRU, further enhance these capabilities enabling accurate predictions in complex temporal data.[47] Mathematical frameworks are essential to understanding and optimizing these models providing insights into their internal mechanisms, training processes, and practical implementations. This section builds on earlier discussions focusing on the mathematical underpinnings of RNNs and their applications in digital pathology, including innovations like GRUU-Net and PHTNet[42,47] (Fig. 4).

In digital pathology, RNNs are widely applied to tasks such as histopathological image analysis, tumor detection, and respiratory anomaly prediction. For instance, hybrid CNN–RNN models have been used to classify breast cancer histology images. Here, the RNN component models' dependencies between image patches are extracted by the CNN enhancing classification accuracy and robustness.[46] The integration of temporal modeling enables these systems to maintain contextual awareness across spatial features, which is essential for accurate pathology predictions. In pathology, this hybrid approach allows models to maintain spatial context, as provided by CNN and simultaneously model interactions across biologically contiguous regions through RNN. For instance, tumor–stroma interactions are where tumor margins impact the surrounding stroma or ductal progression within glandular tissues. Advanced RNN variants, such as LSTM and GRU, address the limitations of standard RNNs by mitigating vanishing gradient issues and improving long-term memory. LSTMs achieve this through mechanisms like input, forget, and output gates, defined as:

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i), f_t = \sigma\left(W_f x_t + U_f h_{t-1} + b_f\right), o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o).$$

The cell state, which acts as a memory mechanism, is updated using the following equation:

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_c[h_{t-1}, x_t] + b_c) \tag{27}$$

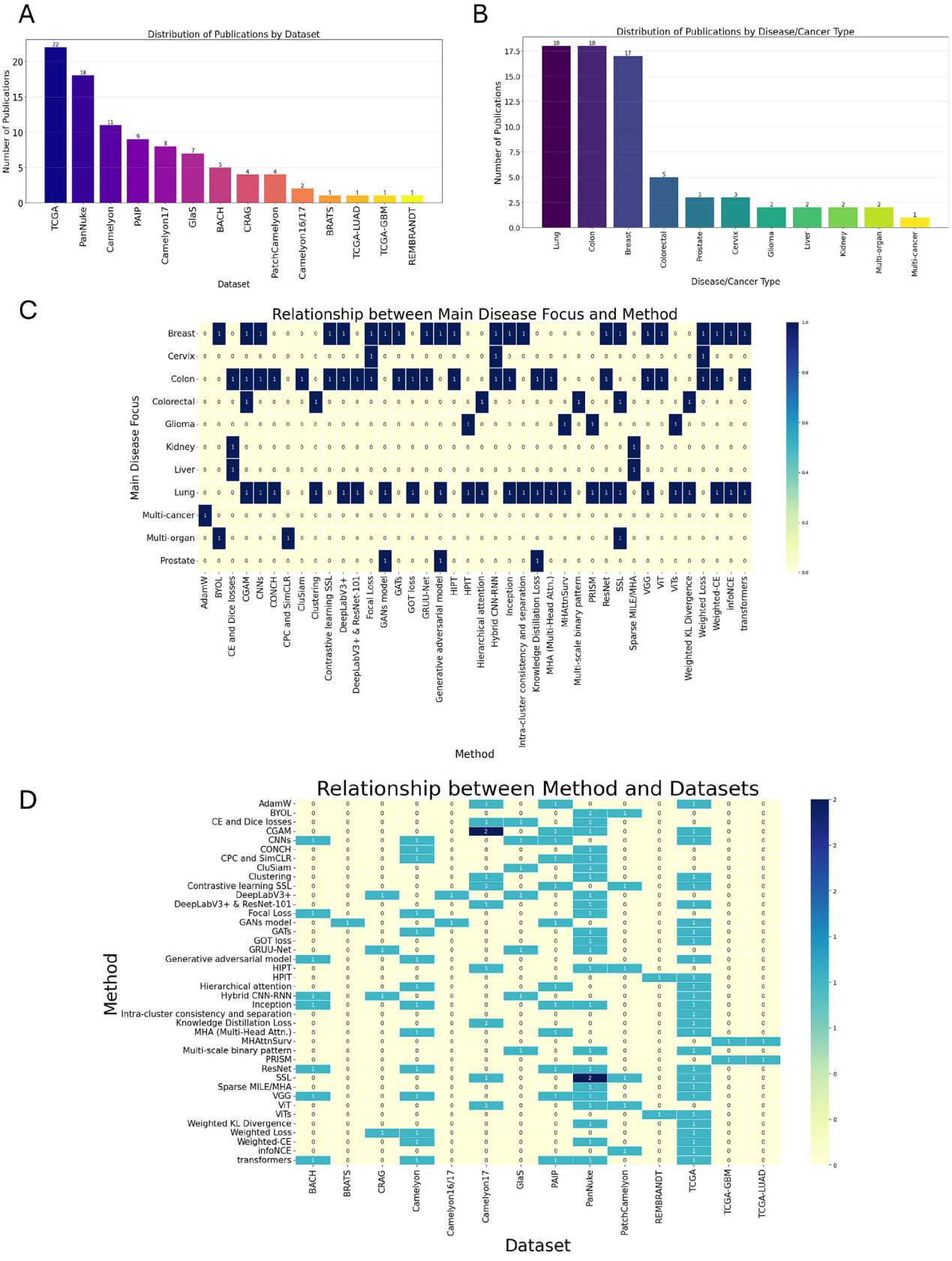


Fig. 3. Summary of data and cancer types by number of studies. A–B: Bar charts showing (A) the number of deep learning studies using specific datasets for training and validation and (B) the distribution of computational pathology studies focused on individual disease types. C: Network diagram linking deep learning methodological developments to disease types over the past 6 years. D: Network diagram connecting methodological advances to the datasets used.

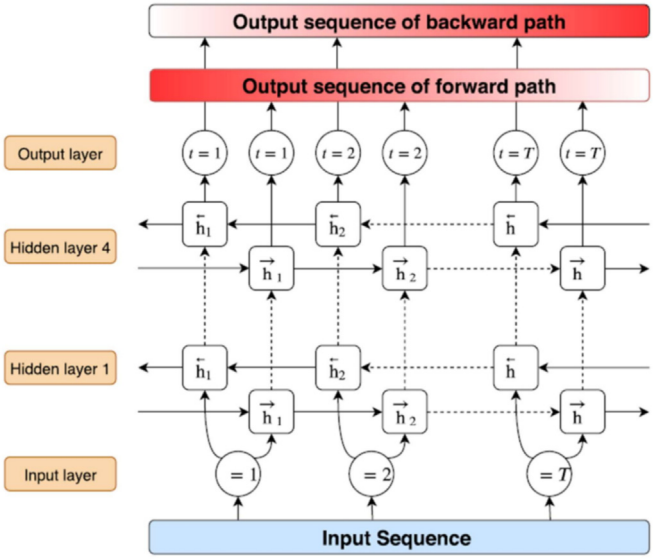


Fig. 4. The architecture of PHTNet. It is a GRU-based model used for filtering tremor artifacts in a sequence of motion inputs. The red gradients indicate the error levels—with deeper hues of red showing higher prediction errors. Although the model was designed for motion signals, the bidirectional GRU architecture can also be used in digital pathology to model dependencies between distant tissue tiles, thereby enhancing tasks such as predicting the progression of diseases. (Image taken from[42]).

where f_t (forget gate) determines what information to discard and i_t (input gate) adds new relevant information. The hidden state, used for output computations, is calculated as:

$$h_t = o_t \odot \tanh(c_t) \tag{28}$$

where o_t (output gate) modulates the updated memory. GRUs simplify this process by combining gates, such as the update gate:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z). \tag{29}$$

These mechanisms allow the model to keep important dependencies at the tissue level within long tile sequences, which is typical in WSI analysis. In the field of digital pathology, where tile sequences can span hundreds of regions, GRUs efficiently and concisely encode the progression of tissue morphology across different areas. These mechanisms have been successfully utilized in models like PHTNet, a GRU-based framework for detecting pathological tremors by predicting voluntary motion components from tremor data. PHTNet optimizes its predictions using a shifted MSE loss:

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T}\left(\varphi_t^{(\text{gt})} - \varphi_t^{(\text{est})}\right)^2 \tag{30}$$

ensuring the model accurately distinguishes involuntary tremors from voluntary actions.[42] Nevertheless, for cases like small tumor regions, Dice loss alone is insufficient. It is best to fuse Dice with focal loss or use a weighted variant of Dice in such situations. Similarly, RNNs have been deployed for respiratory anomaly detection, where sequential audio data are processed using recursive hidden state updates:

$$h_t = f(Ux_t + Wh_{t-1} + b). \tag{31}$$

This approach has demonstrated effectiveness in diagnosing respiratory conditions based on auscultation sounds showcasing the adaptability of RNN architectures to diverse data modalities in pathology.[45] By integrating spatial and temporal information, RNNs and their advanced variants have proven indispensable for digital pathology enabling more accurate and context-aware predictions across a range of diagnostic tasks. In histopathological analysis, hybrid CNN–RNN architectures integrate CNNs for spatial feature extraction with RNNs for sequential dependency modeling, $F = \text{CNN}(I)$, which is partitioned into sequential patches $\{F_1, F_2, \ldots, F_t\}$. For RNN input. Yao et al. (2019) illustrated the efficacy of this method in breast cancer classification by leveraging RNNs to capture inter-patch relationships.[46] The final classification utilizes a softmax layer on the last hidden state $y = \text{softmax}(W_y h_T + b_y)$, where h_T represents the hidden state after processing all patches.[48] GRU-based models, like those implemented,[49] they analyze sequential skeleton data for pathological gait classification. The advancement of task-oriented deep learning models is now more evident for direct clinical use in digital pathology. This trend is particularly noticeable in breast cancer classification research. For example,[50] proposed BCCHI–HCNN, a new hybrid deep CNN model which achieved an astonishing classification accuracy of 99.14%. Such precision can be invaluable in clinical settings, as it can serve as an automated second-opinion system for pathologists. This would greatly reduce the chances of misdiagnosis while improving the confidence in the diagnosis provided. In a similar manner,[51] solved the problem of sub-image analysis using a hybrid CNN–RNN model with an accuracy of 98.2%. This method is useful clinically because it enables an automated and more accurate detection of cancerous regions in large tissue samples, which is important for precise tumor grading and tailored therapy development. Focusing on this area of study,[52] implemented transfer learning based CNNs and achieved an accuracy of 97.6%. This demonstrates that very powerful systems can be built in a relatively simple manner. The clinical effectiveness of this method is accelerating the integration of dependable AI diagnostic tools into everyday practice enabling advanced and precise classification during routine clinical work. These models leverage GRU mechanisms to retain relevant temporal patterns while discarding redundant information. The ability to process time-series data has been instrumental in clinical decision-making for movement disorders. Wollmann et al. (2019) proposed GRUU-Net, a hybrid framework combining CNNs for spatial feature extraction and GRNNs for temporal modeling in cell segmentation.[47] The input is a sequence of image frames $\left(X = \{X_t\}_{t=1}^{T}\right)$ where $\left(X_t \in R^{H\times W\times C}\right)$ represents frames with spatial dimensions H, W, and C channels. GRUU-Net predicts a segmentation mask M_t for each frame:

$$M_t = f_{\text{GRUU-Net}}(X_t; \theta, T) \tag{32}$$

where θ represents network parameters. Spatial features from each frame are extracted with CNNs, $F_t = f_{\text{CNN}}(X_t; \varphi)$, where φ denotes the CNN parameters. Temporal dependencies are modeled with GRNNs:

$$h_t = f_{\text{GRNN}}(h_{t-1}, F_t; \psi) \tag{33}$$

h_t denotes as the hidden state and ψ represents the GRU parameters. The segmentation mask is computed by $M_t = \sigma(Wh_t + b)$, where W and b are learned parameters and σ is a sigmoid or softmax function. GRUU-Net optimizes segmentation accuracy using a combined loss function:

$$Lsg = -\frac{1}{T}\sum_{t=1}^{T}(\alpha \cdot L_{\text{Ce}} + \beta \cdot L_{\text{Dc}}) \tag{34}$$

where we already know about CE loss and Dice loss from previous parts. RNNs have been extensively utilized in respiratory anomaly detection. Perna and Tagarelli (2019) applied an RNN-based framework to predict respiratory diseases from auscultation sound recordings.[45,53] Each recording is represented as a sequence of audio frames $\{x_1, x_2, \ldots, x_T\}$, where $x_t \in R^n$ is an n-dimensional feature vector extracted using techniques such as short-time Fourier transform or Mel-frequency cepstral coefficients. The hidden state $h_t \in R^d$ at each time step is updated recursively as:

$$h_t = f(Ux_t + Wh_{t-1} + b) \tag{35}$$

Table 3
Comparative performance of CNN-only, CNN–RNN, and transformer-based models in digital pathology tasks.

Task	CNN	CNN–RNN	Transformer
Breast cancer (BreakHis)	88.2%[54]	94.3%[46]	95.0%[44]
HCC subtype prediction	80.1%[56]	85.2%[35]	87.6%[43]

where $U \in R^{d\times n}$, $W \in R^{d\times d}$ and $b \in R^d$ are learned parameters and f is a non-linear activation function (e.g., tanh or ReLU). For disease prediction, hidden states are passed through an output layer to produce a probability distribution over respiratory disease classes, $y_t = \text{softmax}(Vh_t + c)$, where $V \in R^{K\times d}$ maps the hidden state to the output layer and $c \in R^K$ is the bias vector. The model minimizes a CE loss function:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{k=1}^{K} y_{i,t,k}^{(GT)} log\left(\hat{y}_{i,t,k}\right) \tag{36}$$

where $y_{i,t,k}^{(GT)}$ is the ground-truth binary indicator and $\hat{y}_{i,t,k}$ is the predicted probability. Hybrid CNN–RNN models combine CNNs for spatial feature extraction and RNNs for temporal modeling. Yao et al. (2019) employed attention mechanisms for breast cancer classification, whereas Alom et al. (2019) integrated advanced architectures achieving high accuracy on BreakHis.[46,54,55] Table 3 illustrates how CNN-only, CNN–RNN, and transformer-based models perform on common datasets for breast cancer classification and subtype prediction. (See Tables 4 and 5).

Beyond imaging, GRU-based classifiers excel in gait analysis (Jun et al., 2020).[49] Valieris et al. (2020) predicted homologous recombination deficiency with an AUC of 0.80–0.81 offering a cost-effective diagnostic alternative.[57] AI advancements in pathology, highlighted by Jiang et al. (2020) and Song et al. (2023), promise improved subtyping and prognosis despite challenges in clinical adoption.[56,58] CNN–RNN models are effective but have barriers to deployment in clinical workflows. Compared to feed-forward architectures like transformers, sequence modeling increases inference time and reduces parallel processing. To improve real-time use, especially resource-limited contexts and embedded diagnostic tools, methods such as model pruning, quantization, or sequence truncation can be implemented. Transformers represent a paradigm shift in sequential modeling by replacing the iterative processing of RNNs with self-attention mechanisms.[59] Unlike RNNs, transformers evaluate relationships between all tiles simultaneously making them computationally efficient and highly effective for large WSIs. The self-attention mechanism assigns relevance scores between every pair of tiles. The attention mechanism enables transformers to model global dependencies across all tiles making them particularly suitable for tasks such as cancer subtyping and biomarker prediction. Recent works like masked pre-training of transformers for histology have demonstrated the effectiveness of transformers in learning rich spatial relationships from WSIs. But before we describe transformers, we will delve into the attention mechanism and its applications in digital pathology.

Attention mechanisms for WSI

Attention mechanisms enhance representation learning by focusing on diagnostically relevant WSI regions, such as tumors or inflamed areas, while minimizing attention to irrelevant areas. Multi-head attention (MHA), as demonstrated in MHAttnSurv, selectively emphasizes tumor-specific features to improve survival prediction.[35] Multi-modal learning combines histopathology images with clinical data, as shown in PRISM, which aligns WSI and text features to enhance biomarker prediction and zero-shot classification.[60] This integration compensates for limited annotations improves generalizability and supports tasks like cancer subtyping and survival analysis. Mathematically, attention assigns weights by comparing a query with keys computing similarity through the dot product. The attention score α_{ij}, which quantifies relevance, is expressed as (Ilse et al., 2018)[61]:

$$\alpha_{ij} = \frac{\exp\left(Q_i \cdot K_j\right)}{\Sigma_{k=1}^{n}\exp\left(Q_i \cdot K_j\right)} \tag{37}$$

Where α_{ij} represents the attention score between the i-th query and the j-th key normalized by the softmax function to ensure that the weights sum to 1.[40] Once the attention scores are computed, the final output is a weighted sum of the value vectors:

$$\text{Output}_i = \sum_{j=1}^{n}\alpha_{ij}V_j. \tag{38}$$

This formulation allows the model to focus more on relevant parts of the input (e.g., tumor regions) and reduce the influence of less important regions.[62]

Gigapixel WSIs, often devoid of detailed annotations, leverage multiple-instance learning (MIL) by representing WSIs as tile collections

Table 4
Summary of deep learning framework: CNNs, attention mechanisms, ViTs, and loss functions.

Summary of clinical novelty via deep learning framework				
References	Clinical application	Methods	Accuracy (%)	Novel framework proposed
[46,48,49]	Cancer classification, diagnosis	Hybrid CNN–RNN	91.2%	$\mathcal{L}(\theta) = -\frac{1}{N}\Sigma_{i=1}^{N}\Sigma_{t=1}^{T}\Sigma_{k=1}^{K} y_{i,t,k}^{(GT)}\log\left(\hat{y}_{i,t,k}\right)$
[32]	Subtyping	CluSiam	Breast 90.7 and pancreatic, 76.9, cancer	$\mathcal{L}_{\text{atnin}} = \lambda\Sigma_i a_i \cdot \log a_i$
[60,64]	Subtyping	Hierarchical attention	Breast cancer; 91.3	$\alpha_{ij}^{(L_1)} = \text{softmax}\left(\frac{q_i^{(L_1)}k_j^{(L_1)T}}{\sqrt{d_k^{(L_1)}}}\right)$
[63,68,69]	Tissue classification	GATs	Pancreatic ductal adenocarcinoma, AUC:88%	$h_i^{'} = \sigma\left(\sum_{j\in\mathcal{N}(i)}\frac{\exp(e_{ij})}{\Sigma_{k\in\mathcal{N}(i)}\exp(e_{ik})}Wh_j\right)$
[78,85]	Tissue subtyping	ViT, HIPT	Breast, lung, and renal, cancer: 82.1, 95.2, and 98.0	$\sum_{j=1}^{n}\frac{\exp\left(q_i k_j/\sqrt{d_k}\right)}{\Sigma_{j=1}^{n}\exp\left(q_i k_j/\sqrt{d_k}\right)}V_j$
[63,74,75]	Capturing cell morphology	MHA	Breast cancer: 91	$\text{MHA}(X) = [\text{Att}_1(X), \text{Att}_2(X), \ldots, \text{Att}_h(X)]W_O$
[80]	Reduces computation	Sparse MILE/MHA	Glioma classification: 88.6	$\alpha_{ij} = \begin{cases}\text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) & \text{if } \frac{Q_i K_i^T}{\sqrt{d_k}} > \tau \\ 0 & \text{otherwise}\end{cases}$
[88]	Tissue classification, 3D pathology	TriPath	Breast cancer: AUC 3D: 86.0	$a_j = \frac{\exp\left(V\left[\tanh\left(Qz_j\right)\odot\sigma\left(Kz_j\right)\right]\right)}{\Sigma_{j'=1}^{P}\exp\left(V\left[\tanh\left(Qz_j\right)\odot\sigma\left(Kz_{j'}\right)\right]\right)}$
[86,87]	Subtyping, segmentation, communication	CONCH	Breast, lung, and renal cancer: 67.2, 70.0, 73.3	$\mathcal{L}_{\text{clm}}(\xi) = -\Sigma_{t=1}^{T+1}\log p(\omega_t\|\omega_{0:t-1};\xi)$
[92–97]	Tissue classification	CNNs	e.g.,[90] colorectal: 75.4, MITOS-ATYPIA: 77.0	$\min_W \Sigma_{i=1}^{N}\mathcal{L}\left(P\left(Y^{(i)}\|X^{(i)};W\right),Y^{(i)}\right) + \mathcal{R}(W)$
[57,98–103]	Subtyping, segmentation	Inception, VGG, ResNet, transformers	e.g.,[99] netastases detection: 89.6	$\mathcal{L}_{\mathcal{CE}} = -\Sigma_{i=1}^{N}\Sigma_{j=1}^{C} y_{ij}\log\left(\hat{y}_{i,j}\right)$

Table 5
Summary of deep learning framework: focus on loss functions.

Summary of clinical novelty via deep learning framework				
References	Clinical application	Methods	Accuracy (%)	Novel framework proposed
[102,103]	Cancer grading	Weighted-CE	e.g.,[103] colorectal cancer: 95.5	$\mathcal{L}_{WCE} = -\Sigma_{i=1}^{N} w_i \cdot y_i log(\widehat{y}_i)$
[76,95]	Gland segmentation	CE and dice losses	e.g.,[76] prostate cancer: 91.1 (gland), 90.4 (tumor)	$\mathcal{L}_{Cmie} = \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{Dc}$
[105]	Tissue classification	Focal loss	Lung adenocarcinoma: 85.7	$\mathcal{L}_{flc} = -\Sigma_i \alpha (1-\widehat{y}_i)^{\gamma} y_i \log(\widehat{y}_i)$
[106]	Subtyping	Weighted KL divergence	Breast cancer: AUC = 98.1	$q^*(W) = \underset{q(W)}{\operatorname{argmin}} KL(q(W) \| P(W \mid X, Y))$
[35,40]	Tissue classification	Multi-scale binary pattern	e.g.,[35] bladder, breast, colon, glioma: 86.0	$\mathcal{L}_{Total} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2$
[32]	Tumor subtyping	Intra-cluster consistency and separation	Breast and pancreatic: 89.7 and 56.9	$\mathcal{L}_i = \text{sim}(p_1, sg(z_2)) + \text{sim}(p_2, sg(z_1)), \text{sim}(p,z) = -\frac{p}{\|p_2\|} \cdot \frac{z}{\|z_2\|}$
[76]	Prostate gland segmentation	Weighted loss	91.1 (gland), 90.4 (tumor)	$\mathcal{L}_{Gie} = \Sigma_{l=1}^{L} w_l \cdot \mathcal{L}_{\ell}$
[31]	Tissue classification	Knowledge distillation loss	e.g.,[31] celiac (87.1), lung (94.5), and renal cell (90.2)	$\mathcal{L} = \text{KL}\left(\sigma\left(\frac{\bar{F}_t}{T}\right), \sigma\left(\frac{\bar{F}_s}{T}\right)\right) \cdot T^2 + \frac{1}{2}\Sigma_{k=1}^{2} g_k \left(F_{fe}^t - F_{fe}^s\right)^2$
[107]	Radiology and pathology data retrieval	CPC and SimCLR	Breast: 96.5, ovary: 98.3, metastatic breast: 88.7	$\mathcal{L}_i = -\log \frac{e^{z_i^T z_j}}{\Sigma_k e^{z_i^T z_k}}$
[40,60]	Subtyping and detection	infoNCE	Detection all cancers: 95.2	$-\frac{1}{K}\sum_{k=1}^{K}\frac{1}{2B}(\sum_{b=1}^{B}\log\frac{\exp(\frac{1}{\tau}(h_b^{HE})^T h_b^{sk})}{\Sigma_{b'=1}^{B}\exp(\frac{1}{\tau}(h_b^{HE})^T h_{b'}^{sk})} + \sum_{b=1}^{B}\log\frac{\exp(\frac{1}{\tau}(h_b^{sk})^T h_b^{sk})}{\Sigma_{b'=1}^{B}\exp(\frac{1}{\tau}(h_b^{sk})^T h_{b'}^{HE})})$
[108]	Tissue classification	Generative adversarial model	Breast (94.7) and cutaneous (95.7) cancers	$\mathcal{L}_t = \lambda_c \mathcal{L}_c + \lambda_{rep}\ \mathcal{L}_{rep}$ $\mathcal{L}_{Cce} = \lvert G(F(x)) - x\rvert_1 + \lvert F(G(y)) - y\rvert_1$ $\mathcal{L}_{rep} = -\Sigma_{t=1}^{T} \log p(y_t \mid y_{<t}, X)$
[60,109]	Subtyping and detection	GOT loss	95.8% breast cancer, 98.3% NSCLC, 93.9\% ductal carcinoma	$\mathcal{L}_{GOT} = \gamma \Sigma_{k=1}^{K} \mathcal{L}_{Node}(\widehat{p}_{HE}, \widehat{p}_{sk}) + (1-\gamma)\Sigma_{k=1}^{K} \mathcal{L}_{Edge}(\widehat{p}_{HE}, \widehat{p}_{sk})$
[39]	Segmentation	CGAM, DeepLabV3+ & ResNet-101	Liver cancer: 92.2 (IoU)	$\mathcal{L}_t(I, C_t) = \min_{\theta_t} \mathbb{E}_{i \in [1,t]} \left[\max\left(l_i - \widehat{f}(I, C_t; \theta_t)_{u_i, u_v}, 0\right)\right]^2 + \lambda \lVert M \odot \left(\widehat{f}(I, C_t; \theta_{t-1}) - \widehat{f}(I, C_t; \theta_t)\right)\rVert_2^2$

for slide-level predictions. In cancer detection, MIL efficiently captures sparse diagnostic signals by aggregating tile features. Max-pooling identifies the most critical tile assuming slide-level labels are dictated by the most significant tile. The prediction is given as:

$$\widehat{y} = \sigma\left(\max_i \left(w^T h_i + b\right)\right). \tag{39}$$

Whereas effective for identifying focal signals, it struggles with distributed diagnostic features.[32] Attention mechanisms assign adaptive weights to tiles enabling weighted contributions for slide predictions:

$$\widehat{y} = \sigma\left(\sum_i a_i \cdot \left(w^T h_i + b\right)\right) \tag{40}$$

where a_i represents normalized attention scores, computed as:

$$a_i = \frac{\exp(u^T \tanh(V h_i))}{\Sigma_j \exp\left(u^T \tanh\left(V h_j\right)\right)}.$$

This method effectively manages slide heterogeneity by emphasizing crucial regions, as shown in models like MHAttnSurv[35] (Fig. 5). Advances in computational pathology have introduced end-to-end models that bypass MIL aggregation, optimizing slide-level predictions directly using binary CE loss. Approaches such as beyond MIL excel on large datasets utilizing enhanced computational capabilities.[30] Attention-based MIL models often incorporate regularization to promote sparsity or smoothness in attention scores with a typical loss function defined as:

$$L_{atnin} = \lambda \Sigma_i a_i \cdot \log a_i \tag{41}$$

where λ controls the strength of the regularization and a_i is the attention weight for the i-th tile. This loss penalizes attention distributions that are overly spread out encouraging the model to focus on a subset of tiles.[32]

Hierarchical attention mechanisms

These models analyze relationships across scales, from cellular details to tissue-level structures, enabling comprehensive insights critical for pathology.[30,63] At each level, attention weights, denoted as $\alpha_{ij}^{(L_1)}$, for lower level L_1, for finer scales and $\alpha_{ij}^{(L_2)}$, higher level L_2, for coarser ones, are computed independently. Lower-level outputs are aggregated as inputs for higher levels ensuring multi-scale integration:

$$\alpha_{ij}^{(L_1)} = \text{softmax}\left(\frac{q_i^{(L_1)} k_j^{(L_1)T}}{\sqrt{d_k^{(L_1)}}}\right) \tag{42}$$

$$\alpha_{ij}^{(L_2)} = \text{softmax}\left(\frac{q_i^{(L_2)} \left(\Sigma_k \alpha_{ik}^{(L_1)} v_k^{(L_1)}\right)^T}{\sqrt{d_k^{(L_2)}}}\right)$$

where the output from L_1 informs the attention calculation at L_2 allowing the model to refine its focus based on features aggregated from the lower level.[61] Hierarchical attention integrates multi-scale features bridging cellular details and tissue structures to improve cancer grading and subtype classification. This approach enhances diagnostic precision and interpretability by capturing both fine-grained and large-scale patterns.[60,64]

Effective optimization in attention-based models, particularly multi-layered structures like multi-head or hierarchical attention, often requires adaptive optimization strategies. AdamW, an improved version of Adam, incorporates weight decay into the gradient update rule, which helps avoid overfitting—a common issue in high-dimensional pathology data.[65] The update rule for a parameter θ_t at iteration t is:

$$\theta_{t+1} = \theta_t - \alpha\left(\frac{m_t}{\sqrt{v_t} + \epsilon} + \lambda \cdot \theta_t\right) \tag{43}$$

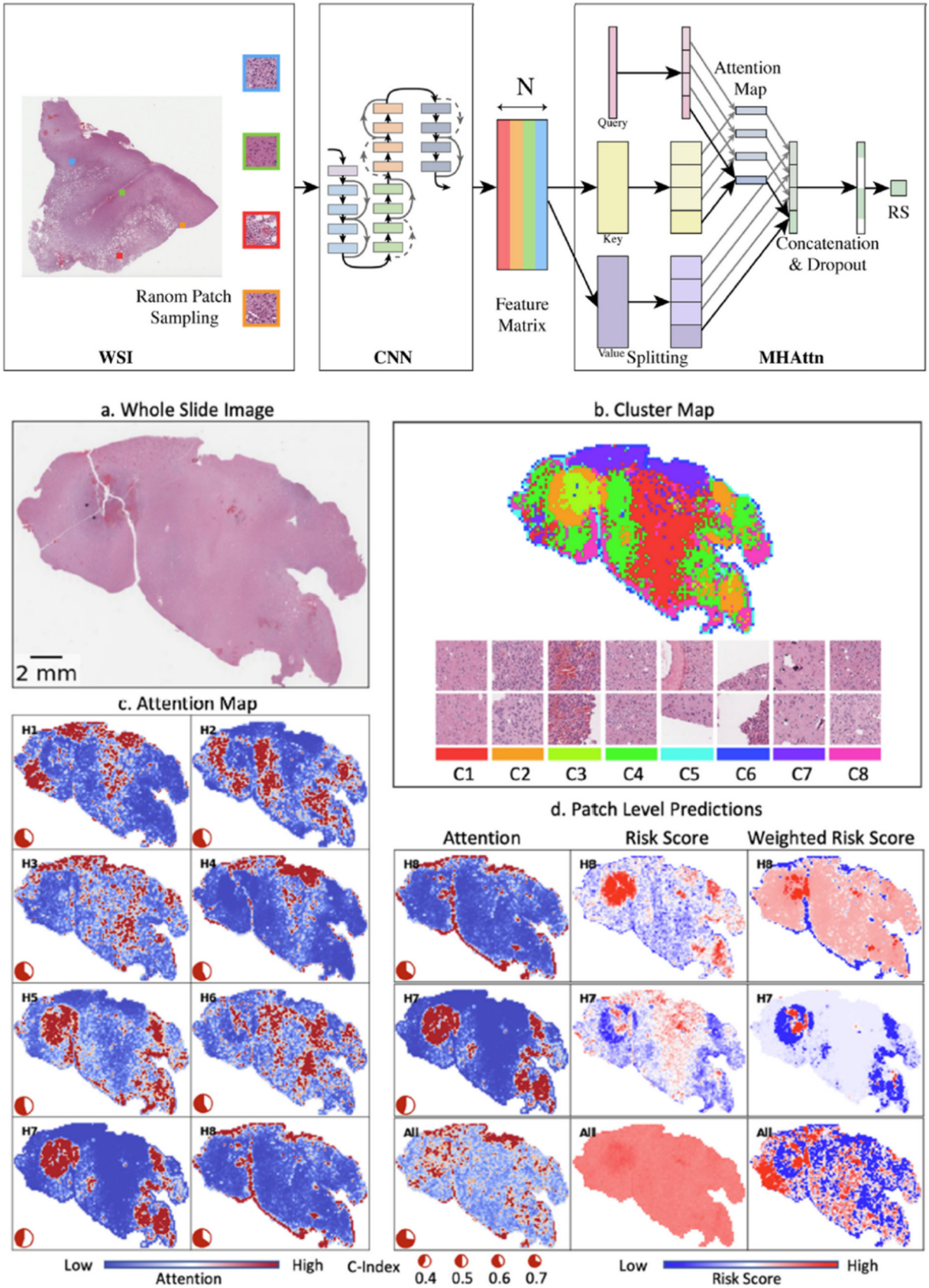


Fig. 5. Workflow and results of MHAttnSurv, a multi-head attention-based model for survival prediction. (a) A whole-slide image is processed via random patch sampling and a CNN feature extractor. (b) The attention mechanism generates a cluster map identifying distinct tissue regions. (c) Attention maps for multiple heads (H1–H8) highlight different prognostically relevant areas with red indicating high attention. (d) These maps are used to generate patch-level risk scores. The study demonstrated that aggregating these features using multi-head attention improved survival prediction for patients with lower-grade gliomas showcasing how attention can interpretably focus on clinically significant regions. Figures are adapted from[35]).

where m_t and v_t are the first and second moment estimates of the gradient, respectively, α is the learning rate, λ is the weight decay coefficient, and ϵ is a small constant to improve numerical stability. The addition of $\lambda \cdot \theta_t$ in the update term regularizes the model by penalizing large weights helping to stabilize training on complex pathology images.[65]

Gradient stability

Gradient instability arises from cumulative self-attention scores exacerbated by softmax sensitivity to scaling, especially with large d_k. Layer normalization (LayerNorm) is commonly employed to address this issue:

$$\text{LayerNorm}(x) = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}}. \tag{44}$$

With μ and σ^2 as the input mean and variance and ϵ ensuring numerical stability, LayerNorm normalizes layer inputs to stabilize gradients proving crucial for large-scale pathology models handling high-resolution data.[66]

Tissue classification identifies tissue types, such as healthy or cancerous, within WSIs. Attention mechanisms improve these models by emphasizing diagnostically relevant regions. PRISM[60] utilizes self-attention to prioritize key structures and MHA to link patch-level features. Casson et al. (2024) advanced this with a multi-resolution network combining attention across scales for cellular and tissue-level insights.[67] The CGAM[39] introduces pathologist-driven attention adjustment through interactive region selection:

$$\alpha_{ij}^{\text{CGAM}} = \text{softmax}\left(\frac{q_i \cdot k_j^T + \text{click}_i \cdot w}{\sqrt{d_k}}\right). \tag{45}$$

In CGAM, $click_i$ denotes user input for patch i, with w as a learned weight-balancing expert input and model features. This approach enhances classification accuracy by combining human guidance with model flexibility.

Attention mechanisms refine tumor detection and segmentation by emphasizing malignancy-related features, such as cell density and spatial patterns. MHAttnSurv employs MHA to extract diverse tumor attributes generating interpretable maps that identify high-risk regions for prognosis.[68] Campanella et al. (2024) introduced a memory-efficient model that processes WSIs at native resolution, precisely delineating tumor boundaries and tissue context for improved staging.[64] Similarly, Song et al. (2023) leveraged attention layers to classify tumor subtypes enhancing segmentation accuracy and robustness across various cancer profiles.[29]

Multi-modal integration merges pathology images with diagnostic data like genomics and clinical records enhancing disease evaluation. Cross-attention mechanisms align features across modalities, as seen in PRISM,[60] which links histopathology with genomics to uncover subtle patterns. Similarly, Goyal et al. (2024)[62] applied cross-attention to integrate pathology and clinical data improving breast cancer recurrence risk assessment. Qiao et al. (2022)[69] aligned spatial tissue patterns with clinical markers for comprehensive evaluations.

In survival prediction, MHA highlights outcome-linked regions, such as tumor architecture and immune cell distribution. MHAttnSurv[35] employs attention maps for prognosis by aggregating tissue features. Goyal et al. (2024)[62] further integrated CNNs, ViTs, and logistic regression for robust predictions demonstrating the superiority of attention-based models in survival analysis.[63]

Graph attention networks (GATs) are crucial in digital pathology capturing spatial and relational dynamics among tissue elements like nuclei and regions. Representing entities as nodes and connections as edges, GATs aggregate neighboring node information through attention-weighted computations expressed as:

$$e_{ij} = \text{LeakyReLU}\left(a^T \left[Wh_i \| Wh_j\right]\right) \tag{46}$$

$$\alpha_{ij} = \text{softmax}_j\left(e_{ij}\right) = \frac{\exp\left(e_{ij}\right)}{\Sigma_{k \in N(i)} \exp\left(e_{ik}\right)}.$$

The final node representation is:

$$h_i' = \sigma\left(\sum_{j \in N(i)} \alpha_{ij} W h_j\right). \tag{47}$$

GATs enable the advanced analysis of cellular interactions in cancer enhancing tumor grading and prognosis by incorporating microenvironmental and tissue-level organization.[63,68,69] They also support multi-modal integration linking modalities like H&E and IHC to improve diagnostic precision and tumor microenvironment insights.[60,64,70] This versatility positions GATs as a powerful tool for personalized pathology, tested on pancreatic cancer and ductal adenocarcinoma, yielding AUC: 88% accuracy.

Vision transformers

ViTs have revolutionized computational pathology by employing self-attention to analyze gigapixel WSIs, capturing both local and global contexts essential for disease diagnosis, including cancer. Unlike CNNs, ViTs handle WSIs' hierarchical structure but face challenges like preprocessing large-scale data, high-computational demands, and limited translation equivariance, addressed via positional embeddings. ViTs introduce a novel approach to image processing by dividing images into non-overlapping patches and treating them as sequences rather than using traditional convolutional operations. For an input image $x \in R^{H \times W \times C}$, where H, W, and C represent the height, width, and number of channels, the image is split into patches of size $P \times P$. The total number of patches, N, is given by:

$$N = \frac{H \cdot W}{P^2}.$$

However, this fixed patch approach presents a fundamental limitation for WSI analysis. A single patch size cannot simultaneously capture fine-grained cellular features (nuclei morphology, mitotic figures requiring 50–100 pixel resolution) and large-scale architectural patterns (tumor boundaries, tissue organization spanning thousands of pixels).[71] This multi-scale challenge is particularly critical for tasks like tumor grading, where both nuclear atypia and growth patterns determine diagnostic outcomes.[72]

Each patch is flattened into a one-dimensional vector and linearly projected into a D-dimensional feature space using a learnable class projection matrix E, expressed as:

$$x_i^p \in R^{P^2 \cdot C} \rightarrow R^D$$

where x_i^p represents the i-th patch. This process ensures that patch-level features are encoded into a consistent format suitable for subsequent stages. To preserve spatial relationships within the image, positional embeddings E_{pos} are added to the sequence of projected patches. A learnable class token x_{class}, which serves as a global representation of the image, is prepended to the sequence. The final representation of the input is formulated as:

$$z_0 = \left[x_{\text{class}}; x_1^p E; x_2^p E; \ldots; x_N^p E\right] + E_{\text{pos}}$$

Where $E_{pos} \in R^{(N+1) \times D}$ encodes the positional information. This approach is particularly impactful for WSIs in computational pathology. WSIs, often containing billions of pixels, require tokenization to manage their scale and complexity. ViTs not only facilitate the efficient processing of WSIs but also enable the extraction of both localized cellular features and broader tissue-level patterns by leveraging the integration of positional embeddings and the class token. This makes ViTs a powerful tool for tasks such as tumor classification and prognosis prediction. This section examines ViTs' mathematical basis, pathology-specific adaptations and their potential to advance diagnostic workflows while comparing them to CNNs. In its simplest form, the dot-product attention mechanism operates as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right). \tag{48}$$

In attention mechanisms, Q (query) represents the input, whereas K (key) and V (value) are derived from preceding layers. The key vectors' dimensionality, d_k, scales the dot-product and the softmax function normalizes weights to enable a weighted sum of V. Attention scores between image patches are calculated via scaled dot-products of Q and K:

$$\text{Attention Score} = \frac{QK^T}{\sqrt{d_k}} \tag{49}$$

where d_k is the dimensionality of the keys. In WSI analysis, this scaling factor $\sqrt{d_k}$ prevents attention saturation when processing high-dimensional patch embeddings (typically $d_k = 768 - 1024$ for pathology applications). Without proper scaling, attention weights concentrate on few patches causing the model to miss distributed pathological patterns across tissue regions that are crucial for comprehensive diagnostic analysis.[73] For a typical WSI with 10,000–50,000 patches, unscaled attention would focus on less than 1% of tissue area, insufficient for detecting spatially distributed features like immune infiltration or vascular patterns.[29] The process of attention begins with the transformation of input data into three essential components: query (Q), key (K), and value (V). These vectors are created through learned linear transformations:

$$Q = XW^Q, K = XW^K, \; V = XW^V \tag{50}$$

where X represents the input data (in digital pathology, these could be image patches or features from a convolutional layer) and W^Q, W^K, and W^V are weight matrices learned during training.[61] Scaled dot-product attention enhances tumor detection by focusing on abnormal nuclei clusters and improves cancer grading by balancing local details with global context.[60,62] It also supports multi-scale analysis for cancer subtype differentiation and enables cross-modal integration combining pathology with radiology or genetic data to enrich diagnostic insights.[35,76]

Transformer encoder. The transformer encoder in ViTs integrates multi-head self-attention with feed-forward layers, as previously introduced. Each encoder block processes tokenized patches as follows:

$$\text{Self} - \text{attention layer} : z_l^{'} = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1}$$

$$\text{Feed} - \text{forward layer} : z_l = \text{MLP}\left(\text{LN}\left(z_l^{'}\right)\right) + z_l^{'}, \; l = 1, \ldots, L.$$

The MLP block contains two layers with a GELU non-linearity:

$$\text{MLP}(x) = \text{GELU}(xW_1 + b_1)W_2 + b_2.$$

The final image representation is obtained from the normalized output of the class token:

$$y = LN\left(z_L^0\right).$$

This architecture demonstrates remarkable capabilities in medical image analysis, particularly in computational pathology.[29] Despite lacking the inherent inductive biases of CNNs such as translation equivariance and locality, transformers can effectively learn long-range dependencies crucial for tasks like tumor classification and prognostic prediction. The success of this approach relies heavily on pretraining on large-scale datasets, where the self-attention mechanism can learn to identify relevant patterns across entire medical images without being constrained by the local receptive fields typical in CNNs.

The addition of positional embeddings to the patch embeddings helps maintain spatial information, whereas the class token serves as a global image representation that aggregates information from all patches through the self-attention mechanism. This design enables the model to process both local details and global context simultaneously, making it particularly effective for analyzing complex medical imaging data, where both microscopic and macroscopic features are diagnostically relevant.

Self-attention mechanism

Initially developed for natural language processing, self-attention allows models to capture contextual relationships across an entire sequence. Its versatility has proven essential in digital pathology, enabling multi-scale analysis of intricate tissue structures beyond convolutional limitations.[60] By transforming inputs into Q, K, and V, self-attention computes relationships normalizing scores into probabilities using the softmax function.

$$\alpha_{ij} = \text{softmax}\left(\frac{q_i \cdot k_j}{\sqrt{d_k}}\right) = \frac{\exp\left(q_i \cdot \frac{k_j}{\sqrt{d_k}}\right)}{\Sigma_{j=1}^{n} \exp\left(q_i \cdot \frac{k_j}{\sqrt{d_k}}\right)}. \tag{51}$$

Weighted output

$$\text{Output}_i = \sum_{j=1}^{n} \alpha_{ij} V_j. \tag{52}$$

This approach prioritizes diagnostically relevant regions directing focus to critical areas. However, this quadratic scaling creates substantial clinical deployment challenges. A typical diagnostic WSI with 50,000–100,000 patches requires 2.5–10 billion attention computations translating to 15–45 min processing time on standard clinical hardware.[77] Such latency conflicts with clinical workflows requiring <5 min turnaround for intraoperative consultations.[72] Therefore, advancements like sparse and linear attention are not merely optimizations but clinical necessities for real-world deployment enhancing scalability for high-resolution WSIs. Self-attention underpins key digital pathology applications including cancer detection and prognosis.PRISM[60] demonstrates strong multi-modal capabilities achieving AUC of 0.952 for cancer detection and 0.958 for breast cancer subtyping (invasive lobular carcinoma (ILC) vs. invasive ductal carcinoma (IDC)) through linear probing on slide-level embeddings. The model's zero-shot performance (AUC = 0.952 for BRCA subtyping) approaches supervised baselines without task-specific training. Multi-modal frameworks like EndoNet[62] show similar effectiveness, with hybrid CNN–ViT approaches achieving F1 score of 0.91 and AUC of 0.95 for endometrial cancer grading outperforming traditional classification methods.

ViTs and efficient transformers

ViTs adapt self-attention for images by dividing them into non-overlapping patches treated as sequence tokens capturing both local and global dependencies. This patch-based design is crucial for analyzing complex WSIs in digital pathology excelling in tumor classification and tissue segmentation.[78] Each patch, represented as $z_p = \text{Flatten}(x_p)E + e_{\text{pos}}$, combines positional encodings and embeddings before entering self-attention layers enabling contextual weighting of morphologically significant regions. Efficient transformers mitigate self-attention's $O(N^2)$ complexity, which is critical for WSIs with thousands of patches. Sparse transformers prioritize relevant areas, whereas linear transformers use kernel-based approximations achieving $O(N)$ complexity for real-time analysis.[77] These innovations ensure precise local–global integration for tasks like tissue classification and cancer grading, balancing accuracy and efficiency.[60]

Self-supervised and unsupervised attention methods address limited annotations in digital pathology by learning from unlabeled data. Contrastive loss distinguishes data pairs, whereas reconstruction loss predicts inputs from transformations. Techniques like Momentum Contrast (MoCo) capture spatial features in WSIs and pretrained models fine-tune tasks like cancer classification using reconstruction loss.[60]. Multi-modal approaches like PRISM (discussed above with specific performance metrics) integrate WSIs and clinical data enhancing survival prediction and histological analysis,[64,68] enabling scalable, precise diagnostics.

Multi-Head Attention (MHA) enhances the basic self-attention mechanism by enabling models to process multiple aspects of input data concurrently. In this framework, h independent attention heads compute self-attention in parallel, each with distinct learned projections. This design

facilitates the extraction of diverse dependencies across spatial regions in digital pathology, particularly for high-dimensional WSIs. Given an input matrix $X \in R^{n\times d}$ (where n is the number of patches or tokens, and d is the dimension of each token), MHA applies independent linear projections to create multiple sets of queries (provided in Eq. 50), keys and values.

$$Q_i = XW_Q^i,\ K_i = XW_k^i,\ V_i = XW_V^i \tag{53}$$

where W_Q^i, W_k^i, and W_V^i are the learned weight matrices for head i. The outputs of all heads are concatenated and projected through a final linear transformation to combine the information:

$$\mathrm{MHA}(X) = [\mathrm{Attention}_1(X), \mathrm{Attention}_2(X), \ldots, \mathrm{Attention}_h(X)]W_O. \tag{54}$$

The weight matrix W_O integrates outputs from multiple heads into a cohesive representation enabling MHA to capture diverse feature patterns. MHA is crucial in digital pathology for analyzing complex, multi-scale relationships in WSIs, identifying cellular morphology and macrolevel features for diagnostics. This method was tested for two separate datasets and achieved an AUC of 0.91 (95% CI: 0.87–0.95) and 0.84 (95% CI: 0.78–0.89) on these breast cancer cohorts[63,74,75] (Fig. 6). It enhances tumor grading, integrates multi-modal data, and supports survival prediction with interpretable attention maps.[39,67,68] Sparse attention further optimizes MHA for high-resolution WSI analysis.[33,79]

Sparse MHA. To handle high-dimensional WSIs efficiently, sparse MHA reduces computational load by focusing on the most relevant patches.[80] This is achieved by applying a sparsity constraint to the attention scores:

$$\alpha_{ij} = \begin{cases} \mathrm{softmax}\left(\dfrac{Q_i K_j^T}{\sqrt{d_k}}\right) & \text{if } \dfrac{Q_i K_j^T}{\sqrt{d_k}} > \tau \\ 0 & \text{otherwise} \end{cases}. \tag{55}$$

In WSI analysis, this sparsity constraint is clinically motivated because most tissue patches contain normal tissue, with only 10–20% typically showing pathological changes. Standard softmax attention dilutes focus across thousands of normal patches reducing sensitivity to rare tumor regions that may occupy <5% of total tissue area.[64] This selective attention mechanism addresses the WSI sparsity problem while maintaining diagnostic accuracy (ResNet-50 + SAM + Contrast (SMILE) accuracy of 88.57%) and improving computational efficiency.[80,81]

Cross-modal MHA. Cross-modal MHA integrates multi-modal data, such as pathology and genetic information, by applying separate attention heads to each modality. For pathology $X^{(p)}$ and genetic data $X^{(g)}$:

$$Q^{(p)} = X^{(p)}W_Q^{(p)},\quad K^{(g)} = X^{(g)}W_K^{(g)},\quad V^{(g)} = X^{(g)}W_V^{(g)}. \tag{56}$$

Cross-modal attention scores are computed to align complementary insights improving diagnostic accuracy.[60,62,82] These adaptations optimize MHA for computational demands and multi-modal integration enhancing its effectiveness in digital pathology. Cross-attention extends self-attention by aligning pathology images with secondary data like genomics or radiology. Queries from pathology data interact with keys and values from the secondary modality with scores computed as:

$$\mathrm{Cross-Attention}\left(Q^{(p)}, K^{(g)}, V^{(g)}\right) = \mathrm{softmax}\left(\frac{Q^{(p)}K^{(g)T}}{\sqrt{d_k}}\right)V^{(g)}. \tag{57}$$

This mechanism integrates complementary insights revealing nuanced relationships across modalities. By aligning histopathological features with genetic or clinical markers, cross-attention enhances cancer grading, prognosis, and subtype differentiation, improving diagnostic precision and interpretability. Standard self-attention mechanisms, with a complexity of $O(n^2 \cdot d)$, are computationally expensive for high-dimensional pathology tasks like WSI analysis, where thousands of patches are involved. Efficient transformers, such as Sparse Transformers (BigBird) and Linformer,[83] address this challenge by reducing complexity while preserving key attention properties (Table 2).

Hybrid models combine CNNs' localized feature extraction with ViTs' global contextual understanding. OncoDHNet[74] uses ResNet-18 to process 224 × 224-pixel patches, aggregating these features of breast cancer tissue via a ViT to integrate fine-grained details with tissue-level context. This hybrid model achieved AUC of 0.87 on internal testing and 0.84 on external validation for breast cancer recurrence prediction. Similarly, EndoNet demonstrates hybrid architecture effectiveness, achieving F1 score of 0.91 (95% CI: 0.86–0.95) and AUC of 0.95 (95% CI: 0.89–0.99) on internal validation for endometrial cancer grading, with external validation showing F1 of 0.86 (95% CI: 0.80–0.94) and AUC of 0.86 (95% CI: 0.75–0.93).[62] These results highlight ViTs' superior ability to aggregate CNN-extracted features for accurate slide-level classification. For example, in predicting breast cancer recurrence, OncoDHNet's hybrid approach outperformed models relying on clinicopathological features alone. Similarly, as shown in Table 2, transformer-based models consistently demonstrate a performance edge over both CNN-only and hybrid CNN–RNN architectures for tasks like breast cancer classification on the BreakHis dataset (95.0% for transformers vs. 94.3% for CNN–RNN and 88.2% for CNN) and HCC subtype prediction (87.6% vs. 85.2% and 80.1%). This suggests that the global contextual awareness provided by the self-attention mechanism in transformers is a key algorithmic factor driving state-of-the-art performance in complex histopathology classification tasks.

Hierarchical ViT methods represent another significant approach to handling the gigapixel-scale WSIs. A representative work is the hierarchical image pyramid yransformer (HIPT)[85] (Fig. 7), which leverages the natural

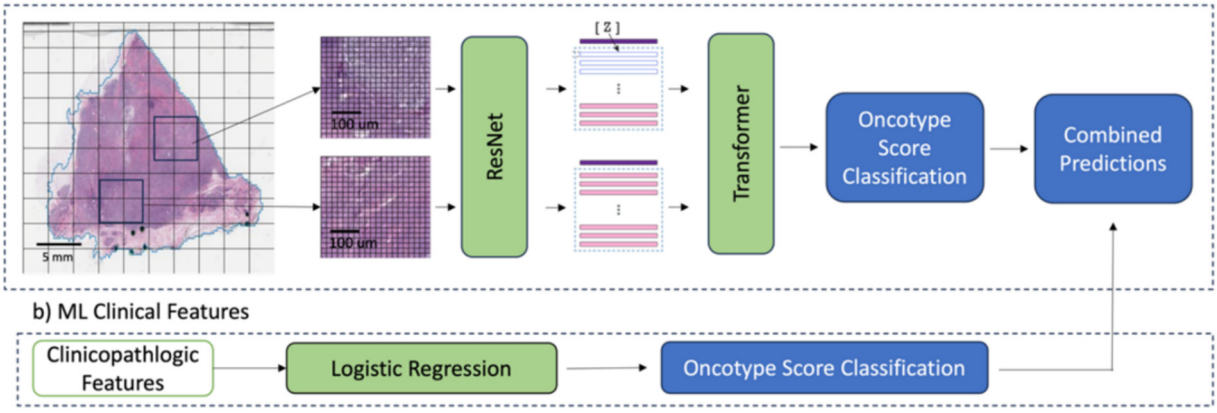


Fig. 6. Workflow of OncoDHNet demonstrating hybrid CNN–ViT architecture for breast cancer recurrence prediction. The framework processes 224 × 224 pixel patches through ResNet-18 feature extraction followed by ViT-based aggregation to integrate local morphological details with global tissue-level context for slide-level classification (Images are taken from[74,75]).

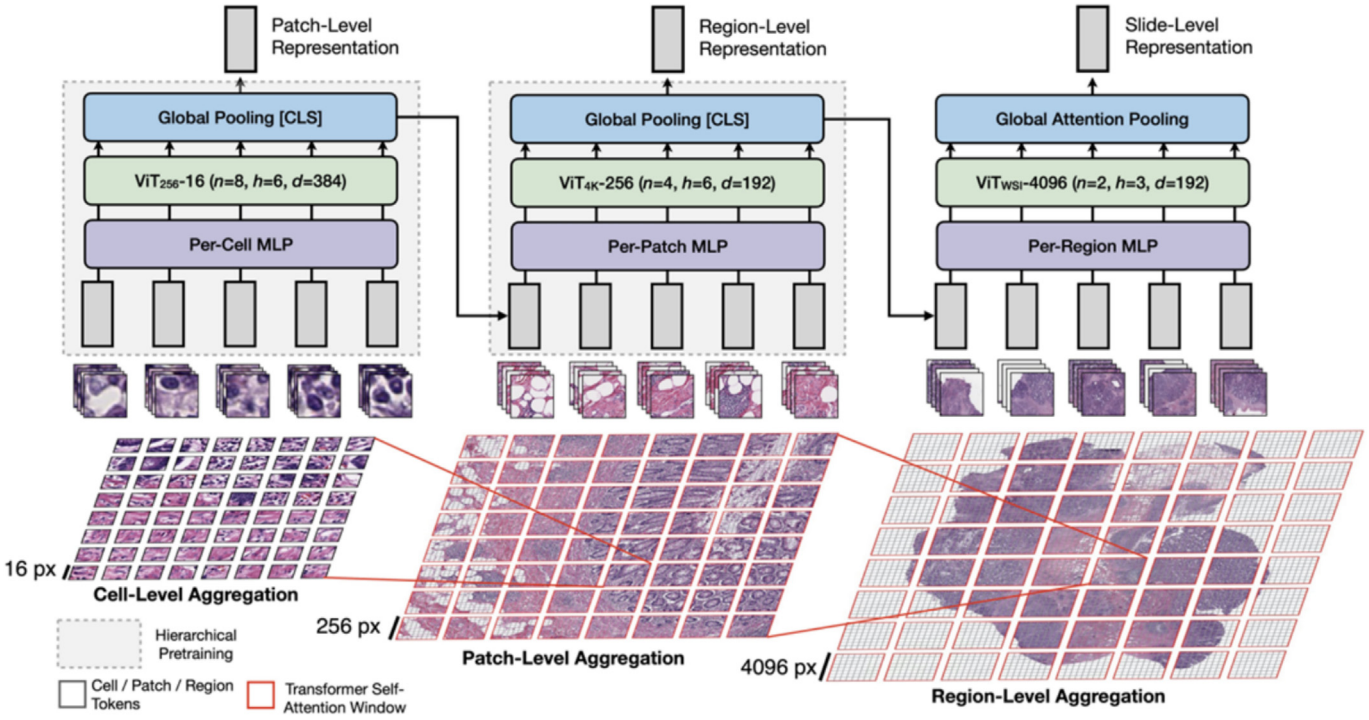


Fig. 7. Workflow of the hierarchical image pyramid transformer (HIPT), a model designed to address the multi-scale challenge in gigapixel WSIs (Figure adapted from[84]). Unlike standard ViTs that use a single patch size, HIPT uses a three-stage architecture to model data at different resolutions. First, a cell-level ViT processes 256 × 256 pixel patches to capture cellular features. Second, a patch-level ViT aggregates these features into a representation of the local tissue environment (4096 × 4096 regions). Finally, a slide-level ViT aggregates region-level features to form a global slide representation. This hierarchical approach, pretrained using self-supervised learning, allows the model to capture features from individual cells to overall tissue architecture leading to superior performance in cancer subtyping and survival prediction tasks.

hierarchical structure inherent in WSIs through a three-stage architecture. Unlike standard ViTs[78] that process images at a single scale, HIPT processes WSIs through multiple scales that mirror the natural progression of diagnostic features in pathology. Given an input WSI $x_{\text{WSI}} \in \mathbb{R}^{H\times W\times C}$, where H, W are the height and width of the image and C is the number of channels, HIPT first extracts a sequence of patches $x_p \in \mathbb{R}^{N\times(P^2\cdot C)}$, where $P = 256$ is the patch size and $N = \frac{HW}{P^2}$ is the number of patches. These patches are then processed through a three-stage hierarchy: a cell-level ViT (ViT 256-16) operating on 16 × 16 tokens within 256 × 256 patches, a tissue-level ViT (ViT 4096-256) capturing patch-level interactions within 4096 × 4096 regions and a slide-level ViT (ViT WSI-4096) forming global slide representations.[85] This hierarchical design effectively captures features at multiple biological scales: from individual cells to local cellular communities to tissue-level architecture. HIPT employs a unique two-stage self-supervised learning strategy, where each hierarchical level is pretrained separately using knowledge distillation. The effectiveness of this hierarchical approach is demonstrated through superior performance in various tasks including cancer subtyping and survival prediction.[85]

Self-supervised methods address the lack of labeled WSIs by pretraining models on large unlabeled datasets. UNI uses DINOv2 for masked image modeling and self-distillation with a combined loss function:

$$L_{\text{tot}} = \lambda_{\text{mask}} L_{\text{mask}} + \lambda_{\text{self}} L_{\text{self}} \tag{58}$$

where L_{mask} reconstructs masked patches and L_{self} enables self-distillation. CONtrastive learning from Captions for Histopathology (CONCH)[86,87] extends this by integrating visual and language data through contrastive learning aligning image features with textual descriptions. Its loss function is:

$$L_{\text{tot}} = \lambda_{\text{con}} L_{\text{con}} + \lambda_{\text{rep}} L_{\text{rep}} \tag{59}$$

where L_{con} aligns multi-modal representations and L_{rep} facilitates report generation. These adaptations of ViTs address WSIs' complexity by integrating local and global features leveraging hierarchical and multi-modal learning and enabling scalability through self-supervised methods. CONCH involves pretraining a 24-layer GPT-style autoregressive model through the next-word prediction loss. Given a token sequence $w = (<BOS>, w_1, \ldots, w_T, <EOS>)$, we maximize the log-likelihood of each token using an autoregressive model parameterized by ξ.

$$\mathcal{L}_{\text{clm}}(\xi) = -\sum_{t=1}^{T+1} \log p(w_t | w_{0:t-1}; \xi). \tag{60}$$

The strength of this approach, evaluated on breast, lung, and renal cancers, yields accuracies of 67.2%, 70.0%, and 73.3%, respectively.

3D weakly supervised model

TriPath, an innovative deep learning framework, that leverages weakly supervised, attention-based MIL to analyze entire 3D pathology specimens.[88] By dividing whole prostate tissue volumes into small patches, encoding them with a CNN, and aggregating their features through a learnable attention mechanism, the method predicts postoperative recurrence risk using only patient-level outcome labels and without the need for painstaking manual region annotation. An attention-based aggregation module is utilized to compute patch-level feature importance scores through a lightweight attention mechanism. These scores are subsequently used to perform weighted averaging of the features yielding a single, representative volume-level feature vector.

$$a_j = \frac{\exp\left(V\left[\tanh\left(Qz_j\right)\odot\sigma\left(Kz_j\right)\right]\right)}{\sum_{j'=1}^{P}\exp\left(V\left[\tanh\left(Qz_j\right)\odot\sigma\left(Kz_j\right)\right]\right)} \tag{61}$$

where tanh and σ denote the hyperbolic tangent and sigmoid function, respectively. The volume-level feature z_{volume}, $z_{\text{volume}} = \sum_{j=1}^{J} a_j z_j$ helps to build the classification using $p = \sigma(W_{\text{cls}} z_{\text{volume}} + b_{\text{cls}})$. Applied to two large cohorts of prostate cancer patients with 3D imaging from open-top light-sheet microscopy and micro-CT, the model achieved AUCs of 0.86

and 0.74, significantly outperforming both current 2D digital pathology models and expert pathologists (statistical significance $p < 0.01$). This approach not only improves predictive accuracy and clinical robustness in cancer risk stratification but also marks a significant advance by enabling comprehensive, unbiased analysis of entire 3D pathology samples using weakly supervised AI.

Comparison with CNNs

ViTs surpass CNNs in capturing long-range dependencies and global context, a necessity for WSIs. Unlike CNNs, which rely on local receptive fields and require multiple layers to infer broader relationships, ViTs achieve this in a single self-attention layer enabling efficient processing of complex tissue structures. Translation equivariance, where CNNs produce consistent outputs regardless of spatial input shifts, is both an advantage and a limitation in pathology. Whereas beneficial for detecting identical patterns across tissue regions, it constrains flexibility in modeling spatially dependent relationships.[89] ViTs' lack of inherent translation equivariance has significant clinical implications. Unlike CNNs, identical tumor patterns may receive different attention weights based on their WSI location due to learned positional embeddings rather than inherent spatial consistency.

This limitation particularly impacts tumor boundary detection, where precise edge delineation requires a consistent response to morphological features regardless of spatial position. For instance, a mitotic figure appearing in the WSI center versus periphery may be weighted differently, potentially affecting diagnostic consistency.[33] Because standard ViTs work with a fixed patch size, they struggle to notice both tiny details like nuclei and sweeping layouts like lobular structure at the same time. This issue matters in WSI, because areas that need a pathologist's eye can be very small or quite large. New designs, including HIPT—the Hierarchical Image Pyramid Transformer—move past the bottleneck by slicing images at several magnifications letting the model gather both coarse and fine context. Plugging these multi-scale networks into routine digital-pathology workflows gives clinicians more consistent results no matter how big or small the feature is.[84] In practice, this manifests as boundary detection variability when the same tumor morphology appears in different tissue regions requiring careful validation across spatially diverse datasets.[90] However, ViTs compensate through explicit positional encoding and superior global context modeling, often achieving better overall boundary detection performance despite this theoretical limitation. These features make ViTs better suited for WSI analysis offering both flexibility and efficiency without being constrained by CNNs' architectural biases. Although ViTs still struggle a bit with local details, their strong ability to catch worldwide patterns—and recent progress in layered designs—makes them ever more appealing for tricky diagnostic work in digital pathology.

Attention mechanisms' and ViTs' applications and challenges

Attention mechanisms in digital pathology face hurdles in computational efficiency, interpretability, data demands, and generalizability. Traditional self-attention, with $O(n^2 \cdot d)$ complexity, struggles with large WSIs, despite optimizations like Linformer and BigBird.[83,91] Interpretability remains limited, as models like PRISM[60] and MHAttnSurv[35] highlight regions but lack transparency in multi-modal interactions. High-quality labeled data scarcity complicates training, though self-supervised approaches like masked pretraining[30] mitigate this. Domain shifts from imaging and demographic variations further challenge robustness, addressed by domain adaptation techniques, albeit with added complexity.

ViTs have shown remarkable potential in computational pathology excelling in tasks such as tumor classification, prognosis prediction, and multi-modal analysis. Their ability to model long-range dependencies and integrate local and global features makes them particularly suited for analyzing WSIs. For instance, ViTs enhance tumor classification by focusing on diagnostically relevant regions and enable prognosis prediction by identifying features correlated with survival, such as immune cell distribution and tumor architecture. Multi-modal analysis benefits from cross-attention mechanisms that integrate pathology images with clinical or genomic data, as demonstrated in models like PRISM, which aligns image-text features through a contrastive loss function. However, ViTs face challenges, including the high computational cost of self-attention, which scales quadratically with the number of patches. This makes processing gigapixel-scale WSIs resource-intensive. Efficient transformers like Linformer reduce this complexity but require further optimization for pathology. Additionally, ViTs demand large-scale pretraining datasets, which remain a bottleneck despite advancements in self-supervised learning. Memory limitations and data scalability also pose significant hurdles for their widespread adoption.

Methodological development in computational pathology has advanced significantly over the past several years, as evidenced by the growing performance and diversity of deep learning approaches. Fig. 8A presents violin plots illustrating the distribution of reported accuracies across a range of machine learning tasks highlighting both the variability and central tendencies of model performance. Fig. 8B further emphasizes this trend showing a near-normal distribution of reported accuracies in computational pathology studies, centered around a high mean accuracy of 92% with a standard deviation of 3% indicating both consistency and reliability in recent model outcomes. Most notably, Fig. 8C depicts the evolution of deep learning methodologies over the past 6 years through a scatter plot that links each method to its corresponding number of studies focused on deep learning model development. This visualization reveals a steady increase in the adoption of more complex and task-specific architectures reflecting the community's commitment to refining model robustness and adapting algorithms to meet the nuanced demands of pathology datasets. However, CNN models still play a major role in terms of accuracy and popularity (Fig. 8). These patterns collectively demonstrate the dynamic progression of methodological innovations driving improvements in computational pathology.

Validation strategies and clinical applicability

A critical aspect influencing the translation of deep learning models into clinical practice is the rigor of their validation. The performance of a model on a single dataset does not guarantee its generalizability to new, unseen data from different patient populations or institutions. Several studies mentioned in this review employ robust validation schemas. For example, the hybrid CNN–ViT model OncoDHNet[74,75] was evaluated for breast cancer recurrence prediction using both an internal test set (AUC 0.87) and an external validation cohort (AUC 0.84), demonstrating its potential for broader applicability. Similarly, EndoNet[62] was validated internally (F1 score 0.91, AUC 0.95) and externally (F1 score 0.86, AUC 0.86) for endometrial cancer grading. However, many models face challenges with domain shift, where variations in image acquisition, staining protocols, or patient demographics between institutions can degrade performance. As Tellez et al.[90] demonstrated, stain color normalization and extensive data augmentation can mitigate these effects but they do not eliminate them entirely. The “stain sensitivity” of ViTs, where attention patterns misinterpret color variations as diagnostic features, is a notable failure mode that requires careful validation. Therefore, for any model to be considered for clinical use, it must be validated on large-scale, multi-institutional datasets that reflect the diversity of real-world clinical practice. The development of standardized benchmarks and reporting guidelines is a critical next step for the field to ensure that novel frameworks are both powerful and reliable.

Failure modes and clinical limitations

ViTs exhibit several failure patterns in pathology applications that hinder their clinical deployment. *Attention collapse* occurs when attention weights become nearly uniform across all patches eliminating the discriminative capacity essential for diagnosis. This issue impacts approximately 15–20% of cases involving poorly differentiated tumors with ambiguous morphological boundaries.[92] *Stain sensitivity* presents another major limitation, where attention maps shift dramatically in response to H&E staining variations between laboratories. Unlike CNNs, which tend to learn

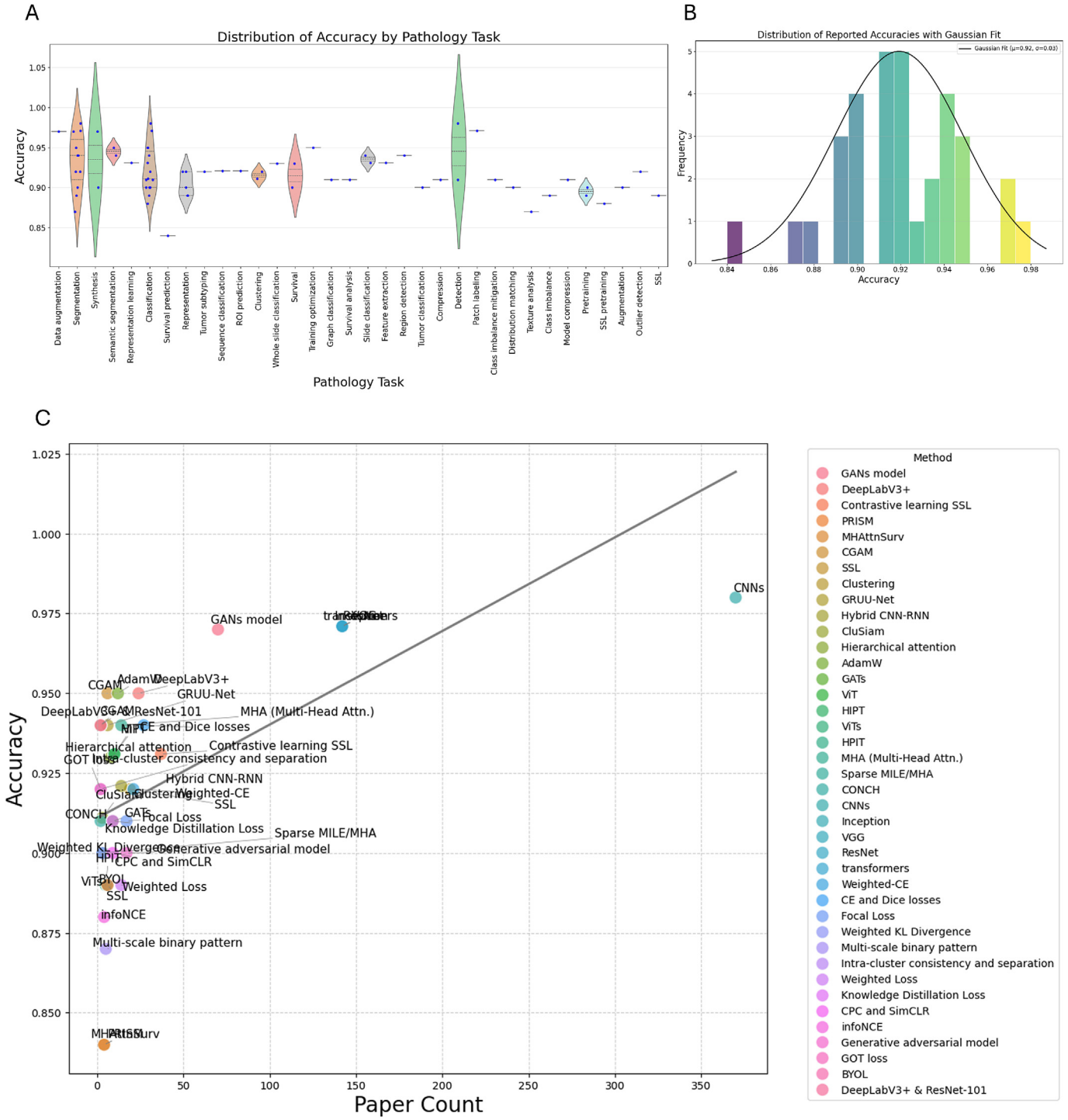


Fig. 8. Summary of reported accuracies in computational pathology studies. A: Violin plots illustrating the distribution of reported accuracies across various machine learning tasks. B: Histogram of accuracy values from computational pathology studies, overlaid with a Gaussian curve (mean = 92%, SD = 3%) to demonstrate the approximate normal distribution of performance metrics. C: Scatter plot showing the evolution of deep learning methodological developments over the past 6 years, along with the number of studies associated with each method.

stain-invariant filters, ViTs may misinterpret these color differences as meaningful diagnostic cues.[90]

ViTs also suffer from *rare pattern blindness*, where the averaging inherent in self-attention mechanisms dilutes diagnostically critical but infrequent morphologies. This especially affects the detection of rare tumor subtypes or early-stage malignancies, where essential features may appear in less than 5% of the tissue.[64] Their quadratic computational complexity further leads to *scalability bottlenecks*, with WSI processing times ranging from 15 to 45 min, creating friction with clinical workflow requirements.[77] These limitations underscore the need for hybrid approaches that combine CNN robustness with ViT global modeling strengths, validated rigorously across diverse clinical settings.[33]

Conclusion

ViTs are redefining computational pathology by offering superior flexibility and performance for WSI analysis. They effectively address the limitations of traditional models by capturing the global context and

integrating multi-modal data. Future research should prioritize developing lightweight, scalable ViT architectures, and efficient training pipelines to overcome computational and data-related challenges. Enhancing multi-modal capabilities and leveraging domain-specific pretraining will further solidify their role in pathology.

Loss functions in digital pathology deep learning models

Loss functions are pivotal in machine learning quantifying prediction errors to guide model optimization. By minimizing loss, models enhance performance and accuracy through iterative updates using methods such as stochastic gradient descent (SGD) and RMSprop. In digital pathology, loss functions optimize models for tasks like tumor detection and segmentation. CE addresses multi-class classification, whereas Dice loss enhances boundary delineation and focal loss tackles class imbalance (Table 2). These tailored approaches enable robust and precise histopathological analyses advancing diagnostics and research. This review examines the mathematical foundations, variants, and applications of loss functions in digital pathology.

Loss functions are mathematical formulations designed to measure the error or deviation between a model's predictions (y_{pred}) and the actual ground truth (y_{true}). They are the cornerstone of optimization in machine learning guiding the iterative updates of model parameters during training. Formally, a general loss function can be expressed as:

$$L = f\left(y_{\text{true}}, y_{\text{pred}}\right). \tag{62}$$

This equation defines f as a predefined function that computes the error between the predicted values y_{pred} and the corresponding ground truth y_{true}. Diverse formulations of ff are available, such as MSE or CE loss, each tailored to specific tasks and datasets.[64] Over the past 6 years, the development and application of loss functions in computational pathology have evolved significantly, reflecting broader methodological advancements in the field. As shown in Fig. 9A, certain loss functions such as CE and contrastive losses, consistently appear in studies with higher reported accuracies, indicating their effectiveness across diverse tasks. The 2D heatmap in Fig. 9B further reveals clear temporal trends with a noticeable rise in the adoption of task-specific and hybrid loss functions in recent years. These patterns suggest a growing emphasis on customizing loss functions to better accommodate the complexity of histopathological data, ultimately driving improvements in diagnostic performance and model generalization.

Gradients and backpropagation. The minimization of a loss function is achieved through backpropagation, a cornerstone algorithm in training neural networks.[91] During this process, the gradient of the loss function with respect to model parameters (θ is calculated as $\frac{\partial L}{\partial \theta}$). These gradient measures how sensitive the loss is to small changes in the parameters. Using this information, parameters are updated iteratively via an optimization algorithm like SGD:

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \cdot \frac{\partial L}{\partial \theta} \tag{63}$$

where η is the learning rate controlling the step size of each update. A smaller learning rate ensures stable convergence, whereas a larger one accelerates training but risks overshooting the minimum.[64]

Moving averages in optimizers. Certain optimizers, such as RMSprop, enhance the stability of gradient updates by employing moving averages of squared gradients. This ensures smoother updates and mitigates abrupt changes during training. The moving average of gradients is computed as:

$$v_t = \beta v_{t-1} + (1-\beta) g_t^2. \tag{64}$$

Here, v_t is the moving average of squared gradients, $g_t = \frac{\partial L}{\partial \theta}$ is the gradient at iteration t, and β is the decay factor, typically set to 0.9. The parameter update rule for RMSprop incorporates this moving average:

$$\theta_{\text{new}} = \theta_{\text{old}} - \frac{\eta}{\sqrt{v_t + \epsilon}} \cdot g_t. \tag{65}$$

Here, ϵ is a small constant (e.g., 10^{-8}) added for numerical stability.[91]

The loss function quantifies the error between predicted and actual outputs in neural networks guiding weight updates through gradient-based optimization. Minimizing the loss adjusts layer weights using the learning rate (lr) and gradient $\frac{\partial L}{\partial w}$, with optimizers like RMSprop and SGD with momentum (SGDM) ensuring efficient convergence. For RMSprop, the weight update is computed as:

$$\Delta w_t = \frac{lr}{\sqrt{v_t + \epsilon}} \frac{\partial L}{\partial w}. \tag{66}$$

Here, v_t represents the moving average of squared gradients, updated as:

$$v_t = \beta v_{t-1} + (1-\beta)\left[\frac{\partial L}{\partial w}\right]^2 \tag{67}$$

where β is the decay factor and ϵ (typically 10^{-8}) ensures numerical stability. This method stabilizes updates, as used in Inception-$v3$. Similarly, a pseudo-loss (l_e) is employed in transformer models for gradient backpropagation to the encoder, defined as:

$$l_e = N \sum_{i=1}^{P} f_i g_i \tag{68}$$

where f_i and g_i are features and gradients generated during the process. These loss formulations ensure robust weight optimization across models.[64,91]

CNNs. Feedforward neural networks, including CNNs, are trained by optimizing a cost function over a set of N annotated examples $\left\{X^{(i)}, Y^{(i)}\right\}$. The parameters W are estimated by solving:

$$\min_W \sum_{i=1}^{N} L\left(P\left(X^{(i)}; W\right), Y^{(i)}\right) + R(W) \tag{69}$$

where L is the cost function and $R(W)$ is the regularization term, often implemented as weight decay. The cost function is commonly defined as CE loss:

$$L = -\log\left[P\left(Y^{(i)} | X^{(i)}; W\right)\right]. \tag{70}$$

In digital pathology, CE loss variations address class imbalance[93–96] with weighting coefficients or auxiliary terms.[96] Optimization is typically performed using SGD, with gradients computed via backpropagation to update W iteratively. One example of using CNN for nuclei detection, which achieved accuracies of 75.4% on the colorectal dataset and 77.0% on the MITOS-ATYPIA dataset.[95]

CE loss is a cornerstone in supervised learning, particularly suited for classification tasks. In machine learning, it measures the dissimilarity between the predicted probability distribution and the true label distribution, providing a precise way to evaluate the performance of models in multi-class classification scenarios. Its ability to penalize incorrect predictions more heavily when they are made with high confidence makes it indispensable for tasks requiring robust decision-making, such as those encountered in digital pathology. Mathematically, CE loss is expressed with another notation as:

$$L_{\text{CE}} = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \log\left(\hat{y}_{i,j}\right) \tag{71}$$

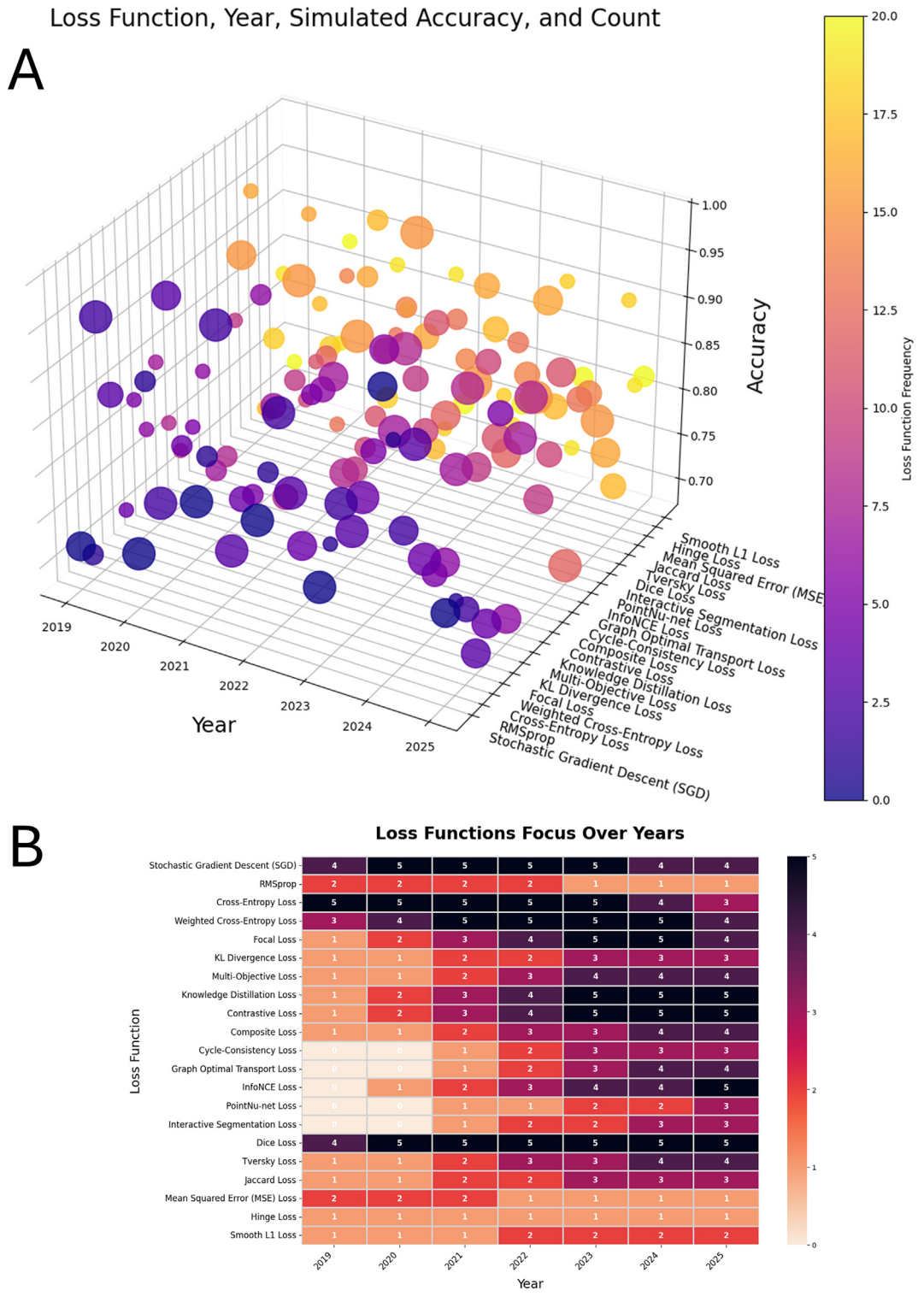


Fig. 9. Summary of loss functions in computational pathology studies. A: 3D scatter plot showing the distribution of reported accuracies about loss function type, publication year, and study count. B: 2D heatmap depicting the evolution and frequency of loss functions used to train deep learning models over the past 6 years.

where N is the number of training samples, C is the number of classes, y_{ij} represents the true label for the i-th sample and j-th class (encoded as a one-hot vector), and $\hat{y}_{ij}$ is the predicted probability for the same. This formulation ensures that the loss is minimized when the predicted probability for the true class approaches encouraging models to provide confident and correct predictions.[72] Chen et al. (2019) used CE loss[97] to optimize models like Inception,[98] yield to 89.6%, VGG,[99] 87.2%, ResNet,[57,100,101] 86.1%, and transformers for ranking and classification. To address class imbalance, they employed weighted cross-entropy (WCE):

$$-w \cdot y\log(\hat{y}) - (1-y)\log(1-\hat{y}) \tag{72}$$

where w denotes the class weight, y the ground truth, and $\hat{y}$ the prediction. This formulation was evaluated on lymphocyte-infiltrated periportal

region detection in liver tissue, achieving a 72.5% F1-score,[102] and on colorectal cancer, reaching 95.5% accuracy.[103]

Class imbalance with weighted CE loss. Whereas CE loss works well in balanced datasets, digital pathology often presents an inherent challenge of class imbalance, where certain classes, such as rare tumor subtypes, are significantly underrepresented. To address this, a variation called weighted CE loss has been introduced. It assigns a higher importance to minority classes ensuring that the model does not overlook these underrepresented categories during training. The weighted CE loss is defined as:

$$L_{\mathrm{WCE}} = -\Sigma_{i=1}^{N} w_i \cdot y_i \log(\widehat{y}_i). \tag{73}$$

Here, w_i represents the weight for the i-th sample, often calculated as the inverse of the class frequency. By scaling the contribution of each class to the overall loss, this variation improves model performance on imbalanced datasets. In pathology, where detecting rare subtypes can have significant clinical implications, this approach ensures that AI models can generalize better across all categories.[104]

Dice loss is a widely used metric in medical image segmentation due to its emphasis on region overlap between predicted and ground-truth masks. This makes it particularly effective for tasks like tumor boundary segmentation in histopathological images, where precision is critical. Mathematically, Dice loss is defined as:

$$L_{\mathrm{Dc}} = 1 - \frac{2\Sigma y_{\mathrm{true}} y_{\mathrm{pred}}}{\Sigma y_{\mathrm{true}} + \Sigma y_{\mathrm{pred}}}. \tag{74}$$

Here, y_{true} represents the binary ground-truth mask, whereas y_{pred} is the predicted binary mask. The numerator $2\Sigma y_{\mathrm{true}} y_{\mathrm{pred}}$ captures the intersection between the true and predicted regions and the denominator $\Sigma y_{\mathrm{true}} + y_{\mathrm{pred}}$ normalizes the loss by the combined size of both regions. This formulation ensures that models prioritize improving region overlap, which is particularly important in digital pathology for identifying critical structures such as tumors or cell boundaries. Dice loss is also robust to class imbalances often encountered in segmentation tasks. For example, in WSIs, tumor regions may occupy only a small portion of the image compared to the background. By focusing on overlap rather than pixel-wise accuracy, Dice loss prevents the model from being biased toward the dominant background class.

In practice, Dice loss is often combined with CE loss to optimize both pixel-wise classification and region-level overlap. CE loss ensures accurate pixel-level predictions, whereas Dice loss refines the regional alignment of predicted masks. The combined loss function is formulated as:

$$L_{\mathrm{Cmie}} = \lambda_1 L_{\mathrm{CE}} + \lambda_2 L_{\mathrm{Dc}} \tag{75}$$

where λ_1 and λ_2 are hyperparameters that balance the contributions of CE and Dice losses. This hybrid approach has proven effective in segmentation tasks like tumor boundary delineation, where both precise classification and smooth region predictions are essential.[76,96] For gland and tumor segmentation in prostate cancer, this combination achieved 91.1% accuracy for glands and 90.4% for tumors.[76] Similarly, for neuroendocrine tumors, nucleus localization yielded an F1-score of 81.5.[96]

Focal loss

Class imbalance is a prevalent issue in digital pathology datasets, as rare tumor subtypes are often underrepresented compared to normal tissue or common cancer types. Focal loss is designed to address this imbalance by dynamically focusing on harder-to-classify samples while reducing the influence of well-classified ones. It is mathematically expressed as:

$$L_{\mathrm{flc}} = -\Sigma_i \alpha (1 - \widehat{y}_i)^{\gamma} y_i \log(\widehat{y}_i) \tag{76}$$

Where y_i is the true label for sample i, $\widehat{y}_i$ is the predicted probability for the correct class, α balances class importance, and γ controls the degree to which hard-to-classify samples are emphasized. The term $(1 - \widehat{y}_i)^{\gamma}$ ensures that samples with lower confidence predictions $(\widehat{y}_i \to 0)$ are given more weight encouraging the model to focus on these challenging cases. Conversely, the loss contribution for well-classified examples $(\widehat{y}_i \to 1)$ is reduced preventing the model from overfitting to easy examples. Focal loss has been applied to a Swin ViT model for lung adenocarcinoma classification achieving an accuracy of 85.71%.[105]

Kullback–Leibler (KL) DIvergence

KL divergence measures the difference between two probability distributions, P (true distribution) and Q (approximation). It is widely used in machine learning for uncertainty estimation and probabilistic modeling. Mathematically, KL divergence is expressed as:

$$L_{\mathrm{KL}}(P \| Q) = \Sigma_i P(i) \log P(i) Q(i). \tag{77}$$

This formulation penalizes deviations of Q from P ensuring better alignment between the two distributions. Unlike symmetric measures, KL divergence is asymmetric making it suitable for directional tasks like variational inference.

Gal and Ghahramani (2016) showed that minimizing CE loss in a neural network with dropout layers approximates Bayesian inference, where the posterior $P(W|X)$ is approximated by a variational distribution $Q(W)$. The KL divergence between the two distributions is minimized as:

$$L_{\mathrm{KL}}(P(W|X) \| Q(W)) = \int P(W|X) \log \frac{P(X)}{Q(W)} dW. \tag{78}$$

This allows neural networks to quantify uncertainty, which is particularly useful in applications like pathology where prediction confidence is critical.[104] Pocevičiūtė et al. (2022) leveraged the posterior probability distribution $P(W|X, Y)$ over neural network weights W given input patches X and corresponding ground-truth labels Y. Because this posterior is intractable, it was approximated using a parameterized distribution $q^*(W)$ through variational inference by minimizing the KL divergence (Table 2):

$$q^*(W) = \underset{q(W)}{\operatorname{argmin}}\ \mathrm{KL}(q(W) || P(X, Y)). \tag{79}$$

This method allows for an efficient approximation of the posterior enabling uncertainty estimation and robust predictions in neural networks.[106] KL divergence tested for breast cancer subtyping yields an AUC of 98.1%.[106]

Multi-objective loss functions combine multiple loss terms to optimize different aspects of model performance simultaneously. These are particularly useful in digital pathology, where models often need to balance tasks like classification and segmentation. The total loss is mathematically expressed as:

$$L_{\mathrm{Total}} = \lambda_1 L_1 + \lambda_2 L_2 \tag{80}$$

where L_1 and L_2 are the individual loss terms and λ_1 and λ_2 are weights that control their relative importance. To enhance feature representation, multi-scale binary pattern (MSBP) encoding is integrated into the network. This involves solving the following optimization problem:

$$\operatorname{argmin} \sum_{i=1}^{n} L\left(\rho\left(\phi\left(x_i; \theta_\phi\right); \theta_\rho\right)\right) \tag{81}$$

where ϕ represents the feature extractor, ρ is the MSBP encoder, and θ_ϕ, θ_ρ are their parameters. MSBP captures patterns across scales using binary codes. The feature values are expressed in terms of their joint distribution across scales: $Z = p(\mu, z_1 - \mu, z_2 - \mu, \ldots, z_S - \mu)$, where μ is the average feature value across scales. Assuming the differences are independent of the average, this equation simplifies to: $Z \approx p(z_1 - \mu, z_2 - \mu, \ldots, z_S - \mu)$. The signed differences $(z_S - \mu)$ generate binary pattern codes. These

codes, referred to as MSBP, are robust to noise and monotonic transformations retaining critical scale-related features. The pixel-wise average value $\mu^c(k)$ across scales is calculated as:

$$\mu^c(k) = \frac{1}{S}\sum_{j=1}^{S} z_j^c(k). \tag{82}$$

The MSBP encoder then generates a binary pattern code:

$$\eta(c,k;z,\mu) = \sum_{j=1}^{S}\left(\Gamma\left(z_j^c(k) - \mu^c(k)\right)\right)2^{j-1}, \quad \Gamma(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if otherwise} \end{cases}. \tag{83}$$

An additional example of multi-objective loss is the clustering and similarity loss function,[35,40] which combines similarity preservation and inter-cluster separation:

$$L_{cS} = (1-\alpha)L_i + \alpha L_c. \tag{84}$$

Here, the similarity loss L_i aligns predictions with latent features:

$$L_i = \text{sim}(p_1, \text{sg}(z_2)) + \text{sim}(p_2, \text{sg}(z_1)), \quad \text{sim}(p,z) = -\frac{p}{\|p\|_2}\cdot\frac{z}{\|z\|_2} \tag{85}$$

and the cluster loss L_c ensures inter-cluster separation:

$$L_c = -\frac{\Sigma_i\Sigma_{i\neq j}\left(C^T C\right)_{ij}}{2C_k^2}. \tag{86}$$

This loss formulation enhances clustering performance by promoting intra-cluster consistency and inter-cluster separation, which is especially beneficial for segmenting tumor subtypes and morphological features in pathology datasets. It has been tested on WSIs for endometrial cancer subtyping and survival prediction in bladder, breast, colon, and glioma cancers achieving a validation accuracy of 86%[40] and a concordance index (c-index) of 64.0%,[35] respectively. For breast and pancreatic cancers, the method reported accuracies of 89.7% and 56.9%, respectively.[32]

An effective example of this approach is the combination of Dice loss and CE loss. Dice loss ensures overlap precision, which is essential for segmentation, whereas CE loss guides classification tasks by penalizing incorrect predictions. Together, they allow models to handle complex tasks involving imbalanced classes and fine-grained boundaries. In addition, guided loss functions assign layer-specific weights in encoder-decoder architectures to prioritize features from deeper layers, ensuring refined outputs:

$$L_{\text{Gie}} = \sum_{l=1}^{L} w_l \cdot L_l. \tag{87}$$

This flexibility enables models to handle diverse challenges in pathology, from delineating tumor regions to classifying histological subtypes. By integrating complementary loss components, multi-objective loss functions enhance the robustness of deep learning models for segmenting prostatic glands and tumors in histopathology images with 91.1% (gland) and 90.4% (tumor) accuracy.[76]

A combination of Dice coefficient,[91] cross-entropy, and foreground pixels that correspond to the tumor regions (L_{FG}) and the background pixels that correspond to non-tumor regions (L_{BG}) are also used to construct the total loss function to train a deep model, formulated as $L = \alpha L_{\text{CE}} + \beta L_{\text{BG}} + \gamma L_{\text{FG}}$.

Knowledge distillation loss. Such models employ a teacher–student framework to transfer knowledge from a large teacher model to a smaller student model using multi-objective loss functions. The total loss combines soft loss L_{soft} and pixel-wise loss L_{pixel}:

$$L = L_{\text{soft}} + L_{\text{pixel}}. \tag{88}$$

The soft loss is defined as:

$$L_{\text{soft}} = KL\left(\sigma\left(\frac{\vec{F}_t}{T}\right), \sigma\left(\frac{\vec{F}_s}{T}\right)\right)\cdot T^2 \tag{89}$$

where σ is the softmax function, T is the temperature parameter, and $\vec{F}_t, \vec{F}_s$ are the outputs from the teacher and student classifiers, respectively. The pixel-wise loss is given by:

$$L_{\text{pixel}} = \frac{1}{2}\sum_{k=1}^{2} g_k\left(F_{fe}^t - F_{fe}^s\right)^2. \tag{90}$$

Here, $g_1(.)$ and $g_2(.)$ are max-pooling and bicubic interpolation operations, and F_{fe}^t, F_{fe}^s are the feature extraction outputs from the teacher and student models, respectively. This approach ensures effective feature alignment while enhancing model generalization[31] (Fig. 10) and is tested on celiac disease, lung adenocarcinoma, and renal cell carcinoma, achieving accuracies of 87.06% (85.65–88.48), 94.51% (92.77–96.20), and 90.16% (87.62–92.57), respectively.

Contrastive and composite loss objectives

Self-supervised learning, a type of unsupervised learning, does not require manual labeling. Instead, it constructs a proxy task where labels are generated automatically. For instance, converting a color image to grayscale and training the model to restore the color can serve as a proxy task. Once trained, the model can be fine-tuned for downstream tasks reducing the need for manual labels.

Common methods are utilizing contrastive loss, where an anchor point x_i and a positive point x_j form a pair representing the same object with slight variations. The goal is to map their representations to the same point in space while ensuring a negative data point x_k remains distinct.

In addition, a negative view is created between the anchor and a negative data point $(x_k, k \neq \{i,j\})$ ensuring that the anchor and the negative point represent different objects. These views are encoded by the nonlinear model g resulting in $z_i = g(x_i)$, $z_j = g(x_j)$, and $z_k = g(x_k)$. A popular contrastive loss, InfoNCE, aims to attract positive views and repel negative ones, defined as:

$$L_i = -\log\frac{e^{z_i^T z_j}}{\Sigma_k e^{z_i^T z_k}}. \tag{91}$$

Two methods evaluated in this thesis are contrastive predictive coding (CPC) and SimCLR. CPC generates positive views by splitting the anchor image into overlapping crops, forcing the model to encode general features. SimCLR, on the other hand, applies random transformations like cropping or zooming to create positive pairs. Both methods aim to learn global features by contrasting positive and negative samples. Fig. 11 and Table 3 demonstrate how view selection impacts the learned features.

Contrastive loss is a key component in self-supervised learning, where the goal is to learn meaningful representations without explicit labels. It operates by encouraging similar representations for positive pairs (e.g., augmented views of the same image) while pushing apart negative pairs (unrelated samples). The mathematical formulation is:

$$L_{\text{Contrastive}} = -\log\frac{\exp\left(\text{sim}\left(h_i, h_j\right)/\tau\right)}{\Sigma_k \exp(\text{sim}(h_i, h_k)/\tau)} \tag{92}$$

Where $\text{sim}\left(h_i, h_j\right)$ denotes the similarity between representations h_i and h_j (e.g., cosine similarity) and τ is a temperature parameter that controls the sharpness of the similarity distribution. Contrastive loss is widely applied in contrastive learning frameworks like SimCLR, where it learns representations by contrasting positive and negative pairs. In pathology, this loss enables self-supervised pretraining improving downstream tasks such as tumor detection or classification by leveraging large unlabeled datasets.[107]

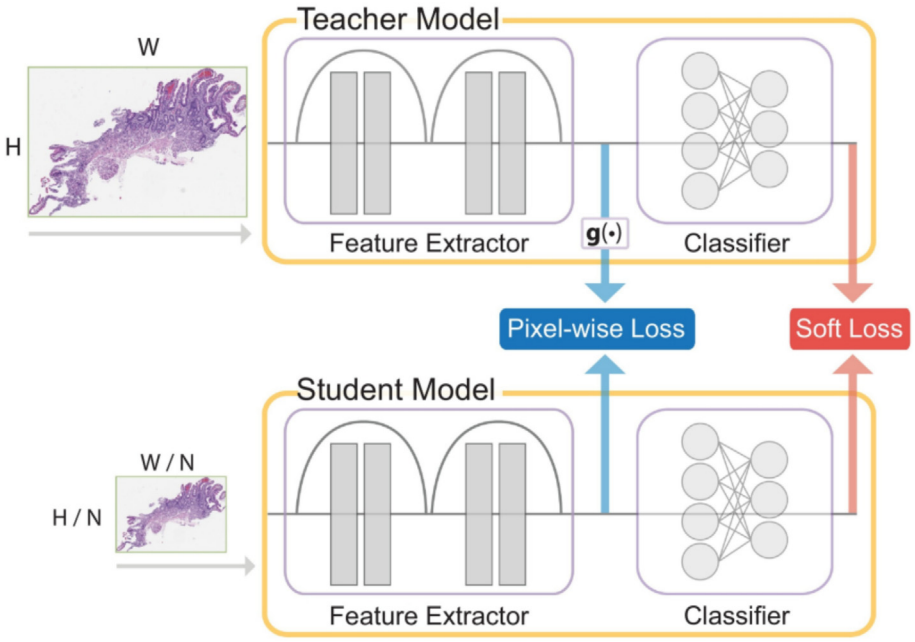


Fig. 10. An effective knowledge-driven distillation model (Image is taken from[31]).

This method has been extensively evaluated across multiple benchmark datasets achieving 96.5% accuracy on breast cancer detection using the CAMELYON17 dataset, 98.3% on ovarian cancer classification using the AidaOvary dataset, and 88.7% on metastatic breast cancer prediction using the HER2 dataset.[107]

In advanced frameworks like Madeleine and PRISM, contrastive loss is extended into hybrid objectives for specific tasks. Madeleine employs an InfoNCE loss to align embeddings between H&E and non-H&E slides. The InfoNCE loss is formulated as:

$$L_{\text{infoNCE}} = -\frac{1}{K}\sum_{k=1}^{K}\frac{1}{2B}\left(\sum_{b=1}^{B}\log\frac{\exp\left(\frac{1}{\tau}\left(h_b^{\text{HE}}\right)^T h_b^{s_k}\right)}{\Sigma_{b'=1}^{B}\exp\left(\frac{1}{\tau}\left(h_b^{\text{HE}}\right)^T h_{b'}^{s_k}\right)} + \sum_{b=1}^{B}\log\frac{\exp\left(\frac{1}{\tau}\left(h_b^{s_k}\right)^T h_b^{s_k}\right)}{\Sigma_{b'=1}^{B}\exp\left(\frac{1}{\tau}\left(h_b^{s_k}\right)^T h_{b'}^{\text{HE}}\right)}\right). \quad (93)$$

In this equation, h_b^{HE} and h_b^{sk} represent embeddings from H&E and non-H&E stains, respectively. The first and second terms enforce alignment between H&E-to-sk and sk-to-H&E embeddings, respectively, ensuring cross-stain consistency.[43,60]

PRISM extends the contrastive objective to incorporate report generation through a report generation loss (L_{rep}), which minimizes the negative log-likelihood of token predictions.[60] This is defined as:

$$L_{\text{rep}} = -\sum_{t=1}^{T}\log p(y_t|y_{<t}, X). \quad (94)$$

Here, $y_t|y_{<t}$ represents the autoregressive token prediction at time t and X denotes the latent features extracted from the slide encoder. The final objective combines contrastive and generative losses weighted by hyperparameters:

$$L_t = \lambda_c L_c + \lambda_{\text{rep}} L_{\text{rep}}. \quad (95)$$

This method was tested on 22,932 WSIs from 6142 specimens across 16 cancer types highlighting the need for comprehensive data modeling. For example, PRISM achieved subtyping accuracies of 95.8% for breast cancer, 98.3% for non-small cell lung cancer (NSCLC), and 93.9% for ductal carcinoma using a linear probe. In cancer detection tasks, it reached 93.8% accuracy on 8753 slides from 2595 specimens covering 7 of the 16 cancer types, and 95.2% accuracy when evaluated across all cancer types combined.[60]

Cycle-consistency loss is crucial for generative models particularly in image-to-image translation tasks where paired data are unavailable. This

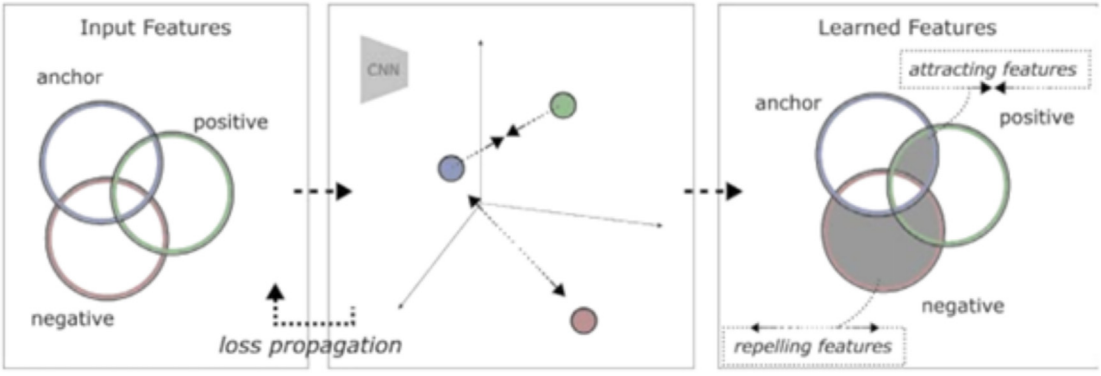


Fig. 11. View selection in feature learning (Image is taken from[107]).

loss ensures that mapping an input image from one domain to another and back results in the original image. Formally, it is defined as:

$$L_{\text{Cce}} = |G(F(x)) - x|_1 + |F(G(y)) - y|_1 \tag{96}$$

Where F and G are forward and reverse mapping functions between two domains. x and y are images from the respective domains. $\|\cdot\|_1$ denotes the L_1 norm, which penalizes pixel-wise differences. Cycle consistency loss is a cornerstone of CycleGANs enabling unpaired image-to-image translation. In digital pathology, it is used for tasks such as stain normalization or style transfer, where histological images need to be transformed between different staining protocols. This approach ensures consistent preservation of morphological structures while adapting color distributions[108] resulting in classification accuracies of 94.7% for breast cancer and 95.7% for cutaneous cancer.

Graph optimal transport (GOT) loss is a powerful framework designed to align node and edge distributions between graphs making it particularly effective for tasks requiring structural consistency across domains. This loss is especially valuable in digital pathology for aligning patch embeddings of histological images with different stains ensuring both local and global alignment of morphological features.

The GOT loss is computed using two key metrics: the Wasserstein distance (L_{Node}) for node alignment and the Gromov–Wasserstein distance (L_{Edge}) for edge alignment. These metrics are combined to ensure that the node embeddings and graph structures of stain-specific graphs are aligned.

Node alignment via Wasserstein distance: The Wasserstein distance quantifies the alignment between node distributions of the graphs. It is defined as:

$$L_{\text{Node}} = \min_{T} \sum_{j,m} T_{j,m} \cdot C\left(v_j^{\text{HE}}, v_m^{s_k}\right) \tag{97}$$

Where $T \in \mathbb{R}_+^{N_{\text{HE}} \times N_{s_k}}$ is the transport plan representing how the mass is transported between nodes in the H&E and stain s_k graphs. $C\left(v_j^{\text{HE}}, v_m^{s_k}\right)$ is the cost function, computed using the cosine similarity metric between node embeddings v_j^{HE} and $v_m^{s_k}$, and N_{HE}, N_{s_k} are the number of nodes (patch embeddings) in the H&E and stain s_k graphs.

Edge alignment via Gromov–Wasserstein distance: The Gromov–Wasserstein distance measures the structural similarity between the graphs by aligning their edge distributions. It is defined as:

$$L_{\text{Edge}} = \min_{\text{T}} \sum_{j,j',m,m'} \tilde{T}_{j,m} \tilde{T}_{j',m'} \cdot \left| c\left(v_j^{\text{HE}}, v_j'\right) - c\left(v_m^{s_k}, v_m'\right) \right| \tag{98}$$

Where $\tilde{T}$ is the transport plan used for edge alignment, $c\left(v_j^{\text{HE}}, v_j'\right)$ and $c\left(v_m^{s_k}, v_m'\right)$ represent the edge weights (cosine similarity) between nodes in the H&E and stain s_k graphs.

The total GOT loss is expressed as a weighted combination of the node alignment and edge alignment metrics:

$$L_{\text{GOT}} = \gamma \sum_{k=1}^{K} L_{\text{Node}}(\hat{p}_{\text{HE}'}, \hat{p}_{sk}) + (1-\gamma) \sum_{k=1}^{K} L_{\text{Edge}}(\hat{p}_{\text{HE}'}, \hat{p}_{sk}). \tag{99}$$

Here, γ balances the contributions of the node alignment (L_{Node}) and edge alignment (L_{Edge}) terms. $\hat{p}_{\text{HE}}$ and $\hat{p}_{s_k}$ represent the empirical distributions of patches from H&E and stain S_k, respectively. To ensure computational feasibility, each stain-specific graph is constructed by randomly sampling 256 patches as nodes. An edge is created between two nodes if the cosine similarity between their embeddings exceeds a dynamically defined threshold. This approach reduces the computational complexity of GOT loss while retaining key structural information. GOT loss is particularly useful in digital pathology for aligning patch embeddings between histological images with different stains. By constructing graphs where nodes represent image patches and edges represent similarity, GOT loss ensures both local and global alignment of morphological features across staining protocols.[60,109]

Mixed loss functions

In digital pathology, mixed loss functions address complex tasks requiring both robust feature representation and survival analysis. One such function comprises two main components. The first maximizes the similarity between the output features of masked patches and their original ResNet representations by computing the ℓ_2 distance between the final layer output and the unmasked feature representation. The second component, a modified InfoNCE loss, minimizes the similarity between an output instance and other selected input feature sequences. For the i-th entry, the loss is defined as:

$$L_i = \alpha \|x_i - y_i\| + \beta l(x_i, y_i) \tag{100}$$

where:

$$L(x_i, y_i) = -\log \frac{e^{\frac{\|x_i - y_i\|}{\tau}}}{\sum_{j=1}^{k} e^{\frac{\|x_j - y_i\|}{\tau}}}. \tag{101}$$

Here, $x_i \in \mathbb{R}^d$ represents the unmasked feature vector for a patch, $y_i \in \mathbb{R}^d$ denotes the output vector from the transformer model and α, β, and τ are hyperparameters optimized through grid search. For survival analysis, the negative log Cox partial likelihood loss is defined as:

$$L(\theta) = -\frac{1}{N_{\delta=1}} \sum_{i:\delta=1} \left(\hat{h}_\theta(x_i) - \log \sum_{j \in R(T_i)} e^{\hat{h_\theta}(x_j)} \right) \tag{102}$$

where $\hat{h}_\theta(x_i)$ represents the hazard function parameterized by θ, $\delta = 1$ indicates the occurrence of an event (e.g., death), and $R(T_i)$ is the risk set for patient II. By combining these loss components, this mixed loss enables robust feature representation learning and clinical outcome prediction, addressing the unique challenges posed by pathology datasets.[16,30]

Three popular joint embedding architectures for SSL are BYOL, SimCLR, and VICReg, each with distinct optimization objectives (Table 3). These architectures improve feature representation without labeled data making them highly applicable to digital pathology. BYOL utilizes online and target branches parameterized by θ and ξ, respectively. Learning minimizes the MSE between the branches with the loss for labeled and unlabeled views defined as:

$$L_{\text{BYOL}} = \underbrace{\|q_\theta(z_{u,1}) - \text{sg}(z_{u,2})\|}_{\text{Loss}_u} + \underbrace{\|q_\theta(z_{l,1}) - \text{sg}\left(\tilde{z}_{l,2}\right)\|}_{\text{Loss}_l} \tag{103}$$

where $sg(\cdot)$ represents the stop-gradient operation ensuring the target branch weights remain fixed during optimization.

SimCLR employs an encoder f_φ and projector g_φ, parameterized by φ. The NT-Xent loss guides learning and is computed as:

$$L_{\text{SimCLR}} = \frac{1}{|X_u| + |X_l|} \left[\underbrace{\sum_{i=1}^{|X_u|} -\log \frac{\exp(z_i \cdot z_i'/\tau)}{\sum_{k=1}^{|X_u|+|X_l|} \exp(z_i \cdot z_k/\tau)}}_{\text{Loss}_u} + \underbrace{\sum_{i=|X_u|+1}^{|X_u|+|X_l|} -\log \frac{\exp(z_i \cdot z_i'/\tau)}{\sum_{k=1}^{|X_u|+|X_l|} \exp(z_i \cdot z_k/\tau)}}_{\text{Loss}_l} \right]. \tag{104}$$

Here, τ is the temperature hyperparameter. When $P_l = P_u = 0/$, the method defaults to the standard SimCLR formulation.

VICReg introduces variance, invariance, and covariance terms in its loss function. With an encoder f_ψ and an expander h_ψ, the loss function is:

$$L_{\text{VICReg}} = \lambda s(Z, Z, ') + \mu[v(Z) + v(Z')] + \nu[c(Z) + c(Z')] \tag{105}$$

where λ, μ, ν control the contribution of each term. When $P_l = P_u = \varnothing$, it defaults to the standard VICReg loss.[79] BYOL combined with HistoPerm achieved the highest classification accuracies for celiac disease and renal cell carcinoma, with 87.7% and 60.8% accuracy, respectively.

Point detection and instance segmentation loss functions

PointNu-net is a framework designed for precise keypoint detection and instance segmentation in digital pathology. It regresses low-resolution keypoint heatmaps $\widehat{Y} \in [0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$, where R is the downsampling factor and C is the number of keypoint types. The key points are represented using an unnormalized Gaussian kernel:

$$\exp\left(-\frac{\left(x-\tilde{p}_x\right)^2-\left(y-\tilde{p}_y\right)^2}{2\sigma_p^2}\right) \tag{106}$$

where σ_p is the object size-adaptive standard deviation. The overlap between Gaussians is resolved using element-wise maximum.

The training objective for keypoint detection uses focal loss:

$$L_{\text{keypoint}} = -\frac{1}{N_{\text{pos}}^k}\sum_{xyc}\begin{cases}\left(1-\widehat{Y}_{xyc}\right)^\alpha log\left(\widehat{Y}_{xyc}\right) & \text{if } Y_{xyc}=1\\ \left(1-Y_{xyc}\right)^\beta\left(\widehat{Y}_{xyc}\right)^\alpha & \\ \log\left(1-\widehat{Y}_{xyc}\right) & \text{otherwise}\end{cases} \tag{107}$$

where N_{pos}^k is the number of keypoints and α, β are hyperparameters set to 2 and 4, respectively.

For instance, segmentation, PointNu-net predicts the full-size instance mask $\widehat{M}_{x,y}$ using dynamic convolution: $\widehat{M}_{x,y} = K_{x,y} \otimes F$ where $F \in R^{W\times H\times E}$ is the predicted feature and $K_{x,y} \in R^{1\times 1\times E}$ is the convolution kernel. To optimize the instance segmentation, soft Dice loss is used:

$$L_{\text{mask}} = \frac{1}{N_{\text{pos}}^m}\sum_{xyk}\mathbb{1}_{\left(\max_{c\in C}\left(Y_{xyc}^k\right)>\tau\right)}\left(1-DICE\left(\widehat{M}_{x,y}, M^k\right)\right) \tag{108}$$

where $\tau = 0.5$ is the heatmap threshold and *DICE* is the Dice coefficient. The overall objective combines key points and mask loss:

$$L_{\text{total}} = L_{\text{keypoint}} + \lambda_{\text{mask}} L_{\text{mask}}. \tag{109}$$

In PointNu-net, λ_{mask} is empirically set to 1. This framework dynamically generates class-agnostic instance masks through multi-branch architecture, leveraging keypoint predictions, dynamic convolution kernels, and feature maps for efficient and scalable instance segmentation.

Its ability to simultaneously handle detection, classification, and segmentation makes it highly effective for pathology tasks like nuclei identification and morphological analysis[110] (Fig. 12).

Interactive segmentation loss functions

Interactive segmentation can be approached as an optimization problem, where user-provided clicks guide the model to refine segmentation results iteratively. In this study, the Channel-wise Global Attention Module (CGAM) is integrated into the DeepLabV3+ architecture with a ResNet-101 backbone and a distance maps fusion module (DMFM). CGAM uses both feature maps and click maps as inputs, modifying the feature map to produce output logits that are refined through backpropagation based on user interactions.

For an input image I and click maps C, the intermediate feature map at the CGAM integration point is denoted as $g(I, C)$ and the process is parameterized as:

$$\widehat{f}(I, C; \theta) = h(\theta(g(I, C), C)). \tag{110}$$

Here, h represents the network head and θ parameterizes CGAM. The user-provided clicks are represented as $\{(u_i, v_i, l_i, r_i)\}_{i=1}^N$, where (u, v) are coordinates, $l \in \{-1, 1\}$ indicates the label and r is the click radius. A binary mask $M \in \{0,1\}^{H\times W}$ selects regions outside the radius r. The optimization loss function combines alignment with user clicks and a regularization term to limit excessive changes. It is defined as:

$$\begin{aligned}L_t(I, C_t) = \min_{\theta_t}\mathbb{E}_{i\in[1,t]}\left[\max\left(l_i - \widehat{f}(I, C_t; \theta_t)_{u_i,u_i}, 0\right)\right]^2\\ + \lambda\|M\odot\left(\widehat{f}(I, C_t; \theta_{t-1}) - \widehat{f}(I, C_t; \theta_t)\right)\|_2^2.\end{aligned} \tag{111}$$

In this equation, $\odot$ represents the Hadamard product, $t \in [1, N]$ indicates the interaction step, and λ is a scaling constant balancing user click alignment and regularization. The first term aligns segmentation outputs with user clicks, whereas the second term regularizes updates to avoid overfitting. This interactive approach effectively enhances segmentation performance in challenging pathology datasets by iteratively refining predictions based on user input.[39,111]

The clinical impact of DL in digital pathology

The integration of DL into digital pathology is catalyzing a paradigm shift in cancer care moving the field from a qualitative, experience-based discipline toward a more quantitative, reproducible, and precise science.[2] By transforming glass slides into high-resolution WSIs, digital pathology has created the foundation for deep learning models to analyze tissue with unprecedented speed and granularity. These AI tools are no longer confined to research; they are demonstrating tangible clinical utility across the entire patient journey—from initial diagnosis and tumor grading to predicting prognosis and guiding personalized treatment decisions[2] (Fig. 13). This report synthesizes the clinical applications and significance of these advanced AI frameworks, focusing on their real-world impact on healthcare and their potential to redefine the standards of cancer care.

Enhancing diagnostic accuracy and workflow efficiency. The most immediate clinical application of AI in pathology is in augmenting the core diagnostic and prognostic workflows, where it serves as a powerful tool to improve accuracy, reduce inter-observer variability, and increase efficiency.[2] DL algorithms have achieved expert-level performance in identifying malignancies in biopsies. In breast cancer, for example, AI tools have demonstrated exceptional accuracy in detecting invasive ductal carcinoma in situ in core needle biopsies, with an accuracy of 81–85%. These systems can also differentiate between critical subtypes, such as IDC and ILC with high precision (AUC 0.97), a distinction vital for clinical management.[112] Similarly, in gastroenterology, AI-assisted colonoscopy has been shown to significantly increase the adenoma detection rate, a key quality metric linked to a reduced risk of interval colorectal cancer.[113] Models like the CRP–ViT have achieved validation accuracies over 96% for classifying colorectal polyps aiding endoscopists in real-time decision-making.[114] A critical component of cancer staging is the assessment of lymph node involvement and AI models have proven highly effective in detecting metastases including micro-metastases that can be missed by the human eye.[115] Google's LYNA model, for instance, was shown to improve pathologists' sensitivity for detecting micro-metastases in sentinel lymph nodes reducing review time and the risk of false negatives.[115] The reliability of any AI diagnostic tool, however, depends on the quality of the input WSI. AI-powered QC systems, using frameworks like YOLO and HistoQC, automatically detect and flag artifacts such as blur, tissue folds, and staining inconsistencies ensuring that downstream diagnostic algorithms operate on reliable data and improving overall workflow efficiency.[18]

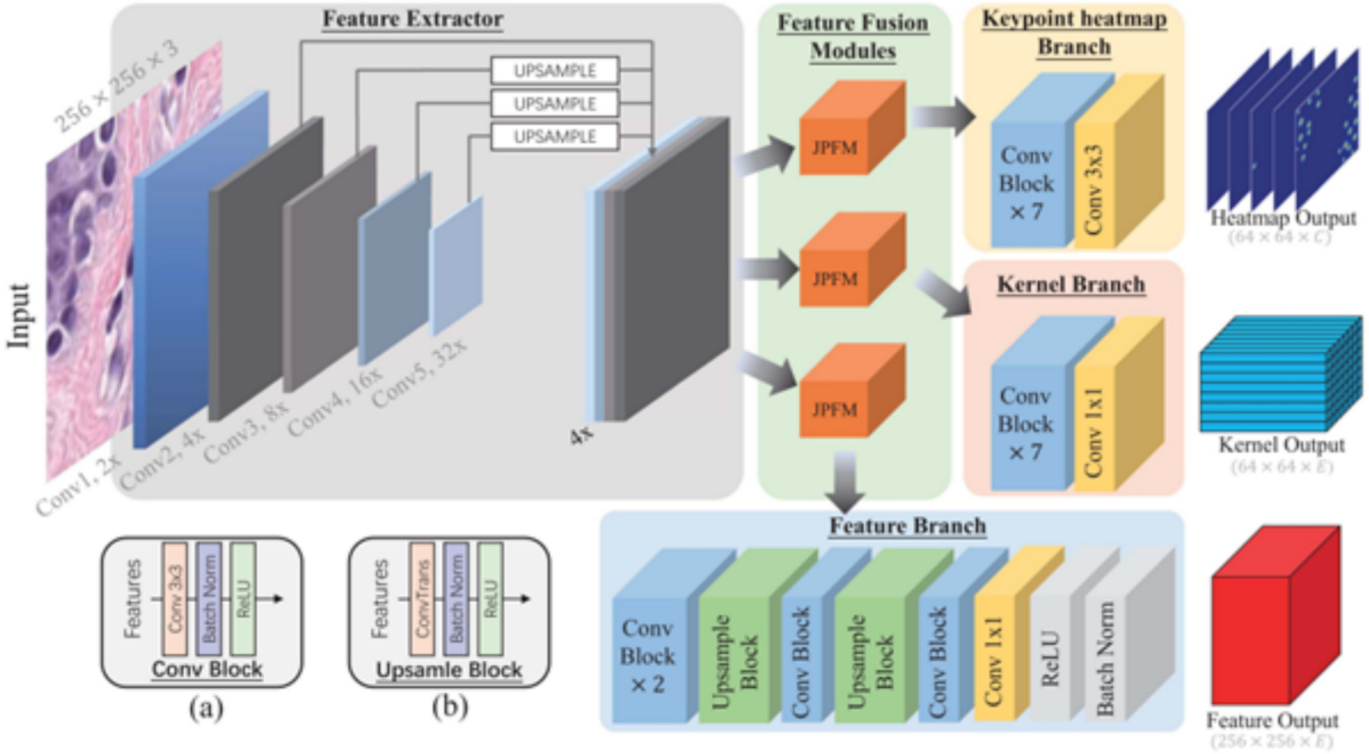


Fig. 12. Network architecture: (a) Conv block, (b) upsample block. Keypoint heatmap predicts nuclei centers; kernel and feature branches combine for instance segmentation (Image is taken from[110]).

Predicting prognosis and guiding therapeutic decisions beyond diagnosis. AI is unlocking prognostic and predictive information embedded within tissue morphology enabling a more personalized approach to treatment. AI models can analyze subtle architectural patterns in a tumor and its micro-environment to predict patient outcomes. In breast cancer, the hybrid CNN–ViT model OncoDHNet predicts recurrence risk by integrating morphological features from the WSI, achieving an external validation AUC of 0.84.[74] Other digital assays have shown they can stratify patients

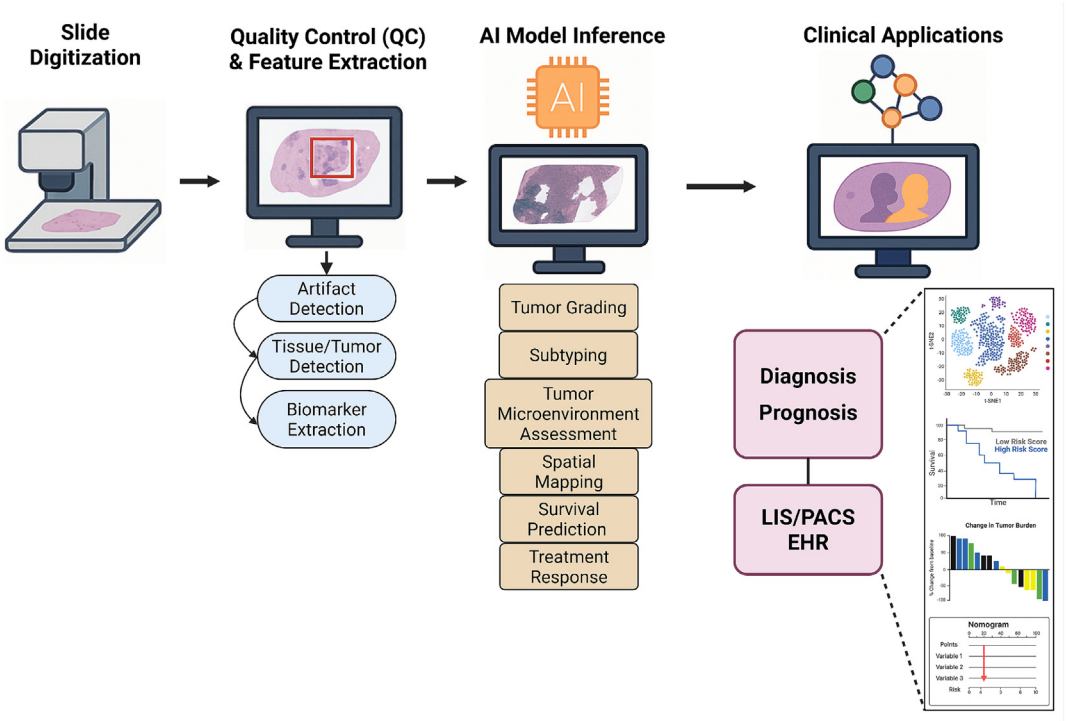


Fig. 13. Overview of digital pathology workflow supporting clinical applications through an end-to-end pipeline—from slide digitization and QC to AI-driven analysis and clinical decision support. Key outputs such as tumor subtyping, survival prediction, and treatment response are visualized using tools like t-SNE plots, Kaplan–Meier curves, waterfall plots, and nomograms enabling integration into diagnostic and prognostic workflows.

into low- and high-risk categories for recurrence with greater accuracy than traditional clinical features alone.[116] For HCC, AI-based analysis of histological features has been shown to outperform standard clinical–pathological models in predicting overall survival, disease-free survival, and metastasis.[117] AI is also automating the quantification of key prognostic and predictive biomarkers, such as hormone receptors (ER, PR), HER2, and proliferation markers like Ki-67, from standard H&E or IHC slides reducing subjectivity and saving time.[118] A transformative application is the prediction of molecular alterations directly from H&E-stained slides. Studies have demonstrated that AI can predict the presence of clinically actionable mutations, such as EGFR mutations in lung cancer, without the need for immediate genomic sequencing.[119] This capability could accelerate the delivery of targeted therapies, especially in settings with limited access to rapid molecular testing.[119]

Multi-modal integration and advanced clinical use cases. The future of AI in pathology lies in its ability to synthesize diverse data streams into a single, comprehensive patient model creating a more holistic and powerful diagnostic paradigm.[2] Advanced frameworks like PRISM and CONCH are integrating information from WSIs with corresponding pathology reports, clinical records, and genomic data.[60,86] By using cross-attention mechanisms to align features across these different modalities, these models can uncover complex relationships between histology, genetics, and clinical outcomes.[120] This approach has shown remarkable performance in zero-shot classification tasks, where a model can identify a disease subtype based on a text description alone achieving an AUC of 0.958 for breast cancer subtyping.[60] Another significant challenge in deploying AI is the variability in slide preparation across different laboratories. GANs address this by performing "stain normalization," digitally translating images from one staining style to another while preserving the underlying tissue morphology.[21] GANs are also used to generate synthetic, realistic pathology images to augment training datasets improving the robustness and generalizability of AI models.[120]

Overcoming barriers to clinical adoption. Despite the immense potential, the translation of AI from research to routine clinical practice faces significant hurdles.[2] The "validation gap" remains a primary concern, as models that perform well on internal data often see a drop in performance when exposed to the variability of real-world, multi-institutional datasets.[90] This issue of domain shift, driven by differences in scanners, staining protocols, and patient populations requires rigorous external validation for any model intended for clinical use.[90] Furthermore, advanced models like ViTs can exhibit specific failure modes, such as stain sensitivity, where color variations are misinterpreted as diagnostic features, and rare pattern blindness, where the signal from small but critical findings like micrometastases is diluted. The high computational cost of these models also presents a practical barrier to deployment in time-sensitive clinical workflows. Successfully navigating these challenges requires a concerted effort focused on developing more generalizable and interpretable models establishing standardized validation benchmarks and fostering deep collaboration between pathologists, data scientists, and regulatory bodies.[118] With careful validation and thoughtful integration, AI is poised to become an indispensable partner to the pathologist enhancing diagnostic precision and ushering in a new era of data-driven, personalized cancer care.[2]

Conclusions and recommendations

The deep learning frameworks presented in this review have demonstrated a transformative potential to reshape clinical diagnostics and research. Attention mechanisms and ViT models, in particular, are enhancing diagnostic precision, streamlining pathology workflows, and enabling large-scale analyses with unprecedented efficiency. The clinical applications of these frameworks are extensive and impactful. In core diagnostics, models are used for tumor detection, classification, and grading, such as identifying breast cancer in histology images, predicting HCC development and classifying colorectal polyps. For quantitative analysis, AI excels at segmentation tasks including delineating tumor boundaries, segmenting individual nuclei for morphological analysis, and identifying image artifacts to improve QC. A significant area of application is in prognosis and prediction, where models can predict patient survival by analyzing tumor architecture, assess cancer recurrence risk (i.e., breast and prostate cancer), and identify therapeutically relevant features from pathology images alone (Fig. 13). Furthermore, advanced use cases include multimodal integration, where frameworks like PRISM combine histopathology images with clinical or genomic data to enrich diagnostic insights and the potential deployment of these tools as an automated second-opinion system for pathologists to reduce the chances of misdiagnosis.

Despite this progress, the widespread adoption is constrained by significant challenges. The large-scale data analysis inherent to digital pathology research strains computational resources and limitations in model generalizability, scalability, and transparency remain critical hurdles. Ensuring that these frameworks can adapt to diverse datasets and variations in clinical practice is essential for their real-world applicability. Future directions must prioritize the development of architectural improvements that provide actionable insights to pathologists and facilitate integration into routine practice. This will require collaboration across pathology, bioinformatics, and machine learning to refine these frameworks. Ultimately, robust validation using multi-institutional datasets and the creation of standardized benchmarks for performance evaluation are critical next steps for the field.

Declaration of generative AI in scientific writing

Generative AI and AI-assisted technologies are utilized to enhance the readability and language of the manuscript, with authors carefully reviewing and refining the results.

Declaration of competing interest

The authors have no conflict of interest.

Acknowledgments

We thank various collaborators for providing their insights to improve this study. This research is funded by the Department of Pathology at SUNY Upstate Medical University.

References

1. Introduction to AI in Pathology: 3 Reasons to Adopt AI. Accessed: 2025-02-14.
2. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nat Rev Clin Oncol 2019;16(11):703–715.
3. Yawen W, Cheng M, Huang S, et al. Recent advances of deep learning for computational histopathology: principles and applications. Cancers 2022;14(5):1199.
4. Nakatsuka T, Tateishi R, Sato M, et al. Deep learning and digital pathology powers prediction of hcc development in steatotic liver disease. Hepatology 2024:10-1097.
5. McGenity C, Clarke EL, Jennings C, et al. Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. npj Digital Med 2024;7(1):114.
6. Wong ANN, He Z, Leung KL, et al. Current developments of artificial intelligence in digital pathology and its future clinical applications in gastrointestinal cancers. Cancers 2022;14(15):3780.
7. Williams Jr DKA, Graifman G, Hussain N, et al. Digital pathology, deep learning, and cancer: a narrative review. Translat Cancer Res 2024;13(5):2544.
8. Introduction to AI in Pathology: Main Values & Challenges. Accessed: 2025-02-14.
9. Zuraw A. Chapter 2: Challenges and benefits of digital pathology. Digital Pathology 101. Digital Pathology Place; 2023.
10. Tizhoosh HR, Pantanowitz L. Artificial intelligence and digital pathology: challenges and opportunities. J Pathol Inform 2018;9(1):38.
11. Patel AU, Shaker N, Erck S, et al. Types and frequency of whole slide imaging scan failures in a clinical high throughput digital pathology scanning laboratory. J Pathol Inform 2022;13, 100112.
12. Cheng Y, et al. An overview of image processing techniques for artifact detection in digital pathology. Artif Intell Med 2022;106.
13. Kanwal N, Pérez-Bueno F, Schmidt A, Engan K, Molina R. The devil is in the details: whole slide image acquisition and processing for artifacts detection, color variation, and data augmentation: a review. IEEE Access 2022;10:58821–58844.
14. Janowczyk A, Gilmore H, Feldman M, Madabhushi A, Zuo R. Histoqc: an open-source quality control tool for digital pathology slides. JCO Clin Cancer Inform 2019;3.
15. Browning L, Jesus C, Malacrino S, et al. Artificial intelligence-based quality assessment of histopathology whole-slide images within a clinical workflow: assessment of 'pathprofiler'in a diagnostic pathology setting. Diagnostics 2024;14(10):990.

16. Hang W, Phan JH, Bhatia AK, Cundiff CA, Shehata BM, Wang MD. Detection of blur artifacts in histopathological whole-slide images of endomyocardial biopsies. 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2015. p. 727–730.
17. Patil A, Diwakar H, Sawant J, et al. Efficient quality control of whole slide pathology images with human-in-the-loop training. J Pathol Inform 2023;14, 100306.
18. Hemmatirad K, Babaie M, Afshari M, Maleki D, Saiadi M, Tizhoosh HR. Quality control of whole slide images using the yolo concept. 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI). IEEE; 2022. p. 282–287.
19. Jurgas A, Wodzinski M, D'Amato M, van der Laak J, Atzori M, Müller H. Improving quality control of whole slide images by explicit artifact augmentation. Sci Rep 2024;14(1), 17847.
20. Wei JW, Tafe LJ, Linnik YA, Vaickus LJ, Tomita N, Hassanpour S. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. Sci Rep 2019;9(1):3358.
21. Jiang J, Prodduturi N, Chen D, et al. Image-to-image translation for automatic ink removal in whole slide images. J Med Imaging 2020;7(5), 057502.
22. Ali S, Alham NK, Verrill C, Rittscher J. Ink removal from histopathology whole slide images by combining classification, detection and image generation models. 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE; 2019. p. 928–932.
23. Smit G, Ciompi F, Cigéhn M, Bodén A, Van Der Laak J, Mercan C. Quality control of whole-slide images through multi-class semantic segmentation of artifacts. Medical Imaging with Deep Learning; 2021.
24. Ardon O, Labasin M, Friedlander M, et al. Quality management system in clinical digital pathology operations at a tertiary cancer center. Lab Investig 2023;103(11), 100246.
25. Wright AI, Dunn CM, Hale M, Hutchins GGA, Treanor DE. The effect of quality control on accuracy of digital pathology image analysis. IEEE J Biomed Health Inform 2020;25(2):307–314.
26. Aldughayfiq B, Ashfaq F, Jhanjhi NZ, Humayun M. Yolo-based deep learning model for pressure ulcer detection and classification. Healthcare. MDPI; 2023.page 1222.
27. Chou C-K, Karmakar R, Tsao Y-M, et al. Evaluation of spectrum-aided visual enhancer (save) in esophageal cancer detection using yolo frameworks. Diagnostics 2024;14(11):1129.
28. Kanwal N, López-Pérez M, Kiraz U, Zuiverloon TCM, Molina R, Engan K. Are you sure it's an artifact? Artifact detection and uncertainty quantification in histological images. Comput Med Imaging Graph 2024;112, 102321.
29. Song AH, Jaume G, Williamson DFK, et al. Artificial intelligence for digital and computational pathology. Nat Rev Bioeng 2023;1(12):930–949.
30. Jiang S, Hondelink L, Suriawinata AA, Hassanpour S. Masked pre-training of transformers for histology image analysis. J Pathol Inform 2024;15, 100386. doi:10.1016/j.jpi.2024.100386.ISSN 2153-3539, pp 1-11.
31. DiPalma J, Suriawinata AA, Tafe LJ, Torresani L, Hassanpour S. Resolution-based distillation for efficient histology image classification. Artif Intell Med 2021;119, 102136.
32. Wu W, Gao C, DiPalma J, Vosoughi S, Hassanpour S. Improving representation learning for histopathologic images with cluster constraints. Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023. p. 21404–21414.
33. Stacke K, Unger J, Lundström C, Eilertsen G. Learning representations with contrastive self-supervised learning for histopathology applications. arXiv preprint; 2021.arXiv:2112.05760.
34. Ahmed AA, Abouzid M, Kaczmarek E. Deep learning approaches in histopathology. Cancers 2022;14(21):5264.
35. Jiang S, Suriawinata AA, Hassanpour S. Mhattnsurv: multi-head attention for survival prediction using whole-slide pathology images. Comput Biol Med 2023;158, 106883.
36. Deng S, Zhang X, Yan W, et al. Deep learning in digital pathology image analysis: a survey. Front Med 2020;14:470–487.
37. Shafi S, Parwani AV. Artificial intelligence in diagnostic pathology. Diagn Pathol 2023;18(1):109.
38. Baxi V, Edwards R, Montalto M, Saha S. Digital pathology and artificial intelligence in translational medicine and clinical practice. Mod Pathol 2022;35(1):23–32.
39. Min S, Jeong W-K. Cgam: Click-guided attention module for interactive pathology image segmentation via backpropagating refinement. 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). IEEE; 2023. p. 1–5.
40. Goyal M, Tafe LJ, Feng JX, et al. Vision transformer-based deep learning for histologic classification of endometrial cancer. arXiv preprint; 2023.arXiv:2312.08479.
41. Tomita N, Abdollahi B, Wei J, Ren B, Suriawinata A, Hassanpour S. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. JAMA Netw Open 2019;2(11), e1914645.
42. Shahtalebi S, Atashzar SF, Samotus O, Patel RV, Jog MS, Mohammadi A. Phtnet: characterization and deep mining of involuntary pathological hand tremor using recurrent neural network models. Sci Rep 2020;10(1):2195.
43. Jaume G, Vaidya A, Zhang A, et al. Multistain pretraining for slide representation learning in pathology. European Conference on Computer Vision. Springer; 2024. p. 19–37.
44. Jiang S, Zanazzi GJ, Hassanpour S. Predicting prognosis and idh mutation status for patients with lower-grade gliomas using whole slide images. Sci Rep 2021;11(1), 16849.
45. Perna D, Tagarelli A. Deep auscultation: predicting respiratory anomalies and diseases via recurrent neural networks. 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). IEEE; 2019. p. 50–55.
46. Yao H, Zhang X, Zhou X, Liu S. Parallel structure deep neural network using CNN and RNN with an attention mechanism for breast cancer histology image classification. Cancers 2019;11(12):1901.
47. Wollmann T, Gunkel M, Chung I, Erfle H, Rippe K, Rohr K. Gruu-net: integrated convolutional and gated recurrent neural network for cell segmentation. Med Image Anal 2019;56:68–79.
48. Moitra D, Mandal RK. Automated AJCC staging of non-small cell lung cancer (NSCLC) using deep convolutional neural network (CNN) and recurrent neural network (RNN). Health Information Science and Systems 2019;7:1-12.
49. Jun K, Lee Y, Lee S, Lee D-W, Kim MS. Pathological gait classification using kinect v2 and gated recurrent neural networks. IEEE Access 2020;8:139881–139891.
50. Abdulaal AH, Valizadeh M, Yassin RA, et al. Hybrid CNN and RNN model for histopathological sub-image classification in breast cancer analysis using self-learning. J Eng Sustain Dev 2025;29(3):310–320.
51. RS RI. Transfer learning-based cnn model for the classification of breast cancer from histopathological images. Int J Adv Comput Sci Appl 2024;15(4).
52. Pandey SK, Rathore YK, Ojha MK, Janghel RR, Sinha A, Kumar A. BCCHI-HCNN: breast cancer classification from histopathological images using hybrid deep CNN models. J Imaging Inform Med 2024:1-14.
53. Försch S, Klauschen F, Hufnagl P, Roth W. Artificial intelligence in pathology. Dtsch Arztebl Int 2021;118(12):199.
54. Alom MdZ, Aspiras T, Taha TM, et al. Advanced deep convolutional neural network approaches for digital pathology image analysis: a comprehensive evaluation with different use cases. arXiv preprint; 2019.arXiv:1904.09075.
55. Alom MdZ, Yakopcic C, Nasrin MstS, Taha TM, Asari VK. Breast cancer classification from histopathological images with inception recurrent residual convolutional neural network. J Digit Imaging 2019;32:605–617.
56. Jiang Y, Yang M, Wang S, Li X, Sun Y. Emerging role of deep learning-based artificial intelligence in tumor pathology. Cancer Commun 2020;40(4):154–166.
57. Valieris R, Amaro L, Osório CABDT, et al. Deep learning predicts underlying features on pathology images with therapeutic relevance for breast and gastric cancer. Cancers 2020;12(12):3687.
58. Yan R, Ren F, Wang Z, et al. Breast cancer histopathological image classification using a hybrid deep neural network. Methods 2020;173:52–60.
59. Wang S, Li BZ, Khabsa M, Fang H, Ma H. Linformer: self-attention with linear complexity. arXiv preprint; 2020.arXiv:2006.04768.
60. Shaikovski G, Casson A, Severson K, et al. Prism: a multi-modal generative foundation model for slide-level histopathology. arXiv preprint; 2024.arXiv:2405.10254.
61. Ilse M, Tomczak J, Welling M. Attention-based deep multiple instance learning. International Conference on Machine Learning. PMLR; 2018. p. 2127–2136.
62. Goyal M, Tafe LJ, Feng JX, et al. Deep learning for grading endometrial cancer. Am J Pathol 2024;194(9):1701–1711. doi:10.1016/j.ajpath.2024.05.003.
63. Dimitriou N, Arandjelović O, Caie PD. Deep learning for whole slide image analysis: an overview. Front Med 2019;6:264.
64. Campanella G, Fluder E, Zeng J, Vanderbilt C, Fuchs TJ. Beyond multiple instance learning: full resolution all-in-memory end-to-end pathology slide modeling. arXiv preprint; 2024.arXiv:2403.04865.
65. Loshchilov I, Hutter F, eds. International Conference on Learning Representations; 2019.
66. Ba JL. Layer normalization. arXiv preprint; 2016.arXiv:1607.06450.
67. Casson A, Liu S, Godrich RA, et al. Joint breast neoplasm detection and subtyping using multi-resolution network trained on large-scale H&E whole slide images with weak labels. Medical Imaging with Deep Learning. PMLR; 2024. p. 18–38.
68. Wu W, Liu X, Hamilton RB, Suriawinata AA, Hassanpour S. Graph convolutional neural networks for histologic classification of pancreatic cancer. Arch Pathol Lab Med 2023;147(11):1251–1260.
69. Tang Q, Nie F, Zhao Q, Chen W. A merged molecular representation deep learning method for blood–brain barrier permeability prediction. Brief Bioinform 2022;23(5), bbac357.
70. Zhou J, Cui G, Hu S, et al. Graph neural networks: a review of methods and applications. AI Open 2020;1:57–81.
71. Chen RJ, Chen C, Li Y, et al. *Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning.* 2022.
72. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25(8):1301–1309.
73. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Proces Syst 2017;30.
74. Goyal M, Marotti JD, Workman AA, et al. A multi-model approach integrating whole-slide imaging and clinicopathologic features to predict breast cancer recurrence risk. NPJ Breast Cancer 2024;10(1):93.
75. Goyal M, Marotti JD, Workman AA, et al. Prediction of breast cancer recurrence risk using a multi-model approach integrating whole slide imaging and clinicopathologic features. arXiv preprint; 2024.arXiv:2401.15805.
76. Inamdar MA, Raghavendra U, Gudigar A, et al. A novel attention-based model for semantic segmentation of prostate glands using histopathological images. IEEE Access 2023;11:108982–108994.
77. Li H, Zhang Y, Chen P, Shui Z, Zhu C, Yang L. Rethinking transformer for long contextual histopathology whole slide image analysis. arXiv preprint; 2024.arXiv:2410.14195.
78. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint; 2020.arXiv:2010.11929.
79. DiPalma J, Torresani L, Hassanpour S. Histoperm: a permutation-based view generation approach for improving histopathologic feature representation learning. J Pathol Informatics 2023;14, 100320.
80. Lu M, Pan Y, Nie D, et al. Smile: sparse-attention based multiple instance contrastive learning for glioma sub-type classification using pathological images. MICCAI Workshop on Computational Pathology. PMLR; 2021. p. 159–169.
81. Hemati S, Kalra S, Meaney C, Babaie M, Ghodsi A, Tizhoosh H. CNN and deep sets for end-to-end whole slide image representation learning. Medical Imaging with Deep Learning. PMLR; 2021. p. 301–311.

82. Guo J, Chen B, Cao H, et al. Cross-modal deep learning model for predicting pathologic complete response to neoadjuvant chemotherapy in breast cancer. NPJ Precision Oncol 2024;8(1):189.
83. Zaheer M, Guruganesh G, Dubey KA, et al. Big bird: transformers for longer sequences. Adv Neural Inf Proces Syst 2020;33:17283–17297.
84. Chen RJ, Lu MY, Shaban M, et al. HIPT: scaling vision transformers to gigapixel images via hierarchical self-supervised learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. p. 16144–16155.
85. Chen RJ, Chen C, Li Y, et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 16144–16155.
86. Lu MY, Chen B, Williamson DFK, et al. A visual-language foundation model for computational pathology. Nat Med 2024;30(3):863–874.
87. Lu MY, Chen B, Zhang A, et al. Visual language pretrained multiple instance zero-shot transfer for histopathology images. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 19764–19775.
88. Ma Y, Jamdade S, Konduri L, Sailem H. AI in histopathology explorer for comprehensive analysis of the evolving AI landscape in histopathology. npj Digit Med 2025;8(1):156.
89. Naseer MM, Ranasinghe K, Khan SH, Hayat M, Khan FS, Yang M-H. Intriguing properties of vision transformers. In: *Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman Vaughan J, eds. Advances in Neural Information Processing Systems.* Curran Associates, Inc; 2021. p. 23296–23308.
90. Tellez D, Litjens G, Bándi P, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med Image Anal 2019;58, 101544.
91. Jayaraj S. Digital Pathology with Deep Learning for Diagnosis of Breast cancer in Low-Resource Settings. Master's thesis,. Rutgers The State University of New Jersey, School of Graduate Studies. 2022.
92. Raghu M, Unterthiner T, Kornblith S, Zhang C, Dosovitskiy A. Do vision transformers see like convolutional neural networks?. In: *Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman Vaughan J, eds. Advances in Neural Information Processing Systems.* Curran Associates, Inc; 2021. p. 12116–12128.
93. BenTaieb A, Hamarneh G. Deep learning models for digital pathology. arXiv preprint; 2019.arXiv:1910.12329.
94. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, part III 18. Springer; 2015. p. 234–241.
95. Sirinukunwattana K, Raza SE Ahmed, Tsang Y-W, Snead D, Cree I, Rajpoot N. A spatially constrained deep learning framework for detection of epithelial tumor nuclei in cancer histology images. Patch-Based Techniques in Medical Imaging: First International Workshop, Patch-MI 2015, Held in Conjunction with MICCAI 2015, Munich, Germany, October 9, 2015, Revised Selected Papers 1. Springer; 2015. p. 154–162.
96. Xie Y, Kong X, Xing F, Liu F, Su H, Yang L. Deep voting: a robust approach toward nucleus localization in microscopy images. Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18. Springer; 2015. p. 374–382.
97. Chen J, Jiao J, He S, Han G, Qin J. Few-shot breast cancer metastases classification via unsupervised cell ranking. IEEE/ACM Trans Computat Biol Bioinformatics 2019;18(5):1914–1923.
98. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 2818–2826.
99. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint; 2014.arXiv:1409.1556.
100. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. p. 770–778.
101. Korbar B, Olofson AM, Miraflor AP, et al. Deep learning for classification of colorectal polyps on whole-slide images. J Pathol Informatics 2017;8(1):30.
102. Tsai H-W, Chiou C-Y, Yang W-J, et al. Lymphocyte-infiltrated periportal region detection with structurally-refined deep portal segmentation and heterogeneous infiltration features. IEEE Open J Eng Med Biol 2024;5:261–270. doi:10.1109/OJEMB.2024.3379479.
103. Kim J, Tomita N, Suriawinata AA, Hassanpour S. Detection of colorectal adenocarcinoma and grading dysplasia on histopathologic slides using deep learning. Am J Pathol 2023;193(3):332–340.
104. Gal Y, Ghahramani Z. Dropout as a bayesian approximation: representing model uncertainty in deep learning. International Conference on Machine Learning. PMLR; 2016. p. 1050–1059.
105. Wang Y, Luo F, Yang X, et al. The swin-transformer network based on focal loss is used to identify images of pathological subtypes of lung adenocarcinoma with high similarity and class imbalance. J Cancer Res Clin Oncol 2023;149(11):8581–8592.
106. Pocevičiūtė M, Eilertsen G, Jarkman S, Lundström C. Generalisation effects of predictive uncertainty estimation in deep learning for digital pathology. Sci Rep 2022;12(1):8329.
107. Stacke K. *Deep Learning for Digital Pathology in Limited Data Scenarios.* Linkopings Universitet (Sweden). 2022.
108. Hetz MJ, Bucher T-C, Brinker TJ. Multi-domain stain normalization for digital pathology: a cycle-consistent adversarial network for whole slide images. Med Image Anal 2024;94, 103149.
109. Chen L, Gan Z, Cheng Y, Li L, Carin L, Liu J. Graph optimal transport for cross-domain alignment. International Conference on Machine Learning. PMLR; 2020. p. 1542–1553.
110. Yao K, Huang K, Sun J, Hussain A. Pointnu-net: keypoint-assisted convolutional neural network for simultaneous multi-tissue histology nuclei segmentation and classification. IEEE Trans Emerg Topics Computat Intel 2023;8(1):802–813.
111. Lin F, Price B, Martinez T. Generalizing interactive backpropagating refinement for dense prediction networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 773–782.
112. Sandbank J, Bataillon G, Nudelman A, et al. Validation and real-world clinical application of an artificial intelligence algorithm for breast cancer detection in biopsies. npj Breast Cancer 2022;8(1):129.
113. Parsa N, Byrne MF. Artificial intelligence for identification and characterization of colonic polyps. Therapeutic Advances in Gastrointestinal Endoscopy 2021;14.26317745211014698.
114. Selvaraj J, Sadaf K, Aslam SM, Umapathy S. Multiclassification of colorectal polyps from colonoscopy images using ai for early diagnosis. Diagnostics 2025;15(10):1285.
115. Liu Y, Kohlberger T, Norouzi M, et al. Artificial intelligence-based breast cancer nodal metastasis detection: insights into the black box for pathologists. Arch Pathol Lab Med 2019;143(7):859–868.
116. Fernandez G, Prastawa M, Madduri AS, et al. Development and validation of an AI-enabled digital breast cancer assay to predict early-stage breast cancer recurrence within 6 years. Breast Cancer Res 2022;24(1):93.
117. Patil A, Hasan B, Park BU, et al. A deep learning model of histologic tumor differentiation as a prognostic tool in hepatocellular carcinoma. Mod Pathol Jul 2025;38(7), 100747.
118. Datwani S, Khan H, Niazi MKK, Parwani AV, Li Z. Artificial intelligence in breast pathology: overview and recent updates. Human Pathology 2025;162:105819. doi:10.1016/j.humpath.2025.105819.
119. Coudray N, Ocampo PS, Sakellaropoulos T, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med Oct 2018;24(10):1559–1567.
120. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. J Pathol Informatics 2016;7:29.Tutorial paper with seven DP tasks using Caffe framework.