Computer Modeling in Engineering & Sciences

Tech Science Press

https://doi.org/10.32604/cmes.2026.084724

ARTICLE

WaSA-Net: Wavelet-Guided Tokenization and Dynamic Sparse Attention for Histopathology Image Classification

Muhammad Zaheer Sajid[1], Muhammad Fareed Hamid[2], Nauman Ali Khan[2,3,*] and Imran Qureshi[4]

[1]Department of Electrical and Computer Engineering, George Mason University, Fairfax, VA, USA
[2]Department of Computer Software Engineering, National University of Sciences and Technology, Islamabad, Pakistan
[3]Department of Smart Computing and Cyber Resilience, DSCCR, Sunway University, Kuala Lumpur, Malaysia
[4]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

*Corresponding Author: Nauman Ali Khan. Email: nauman@mcs.edu.pk or naumank@sunway.edu.my

Received: 28 April 2026; Accepted: 02 July 2026; Published: 27 July 2026

ABSTRACT: Digital pathology is rapidly transforming histopathological diagnosis, yet many existing deep learning models treat all spatial regions uniformly and do not exploit the multi-frequency structure of tissue, which limits both diagnostic accuracy and computational efficiency. This paper proposes WaSA-Net, an end-to-end architecture that integrates three complementary modules for histopathological image analysis. First, the Wavelet-Guided Tokenization (WGT) module decomposes input images into frequency-aware representations using learnable wavelet-like filters, so that both global tissue structures and fine-grained cellular patterns are exposed to attention from the first layer. Second, the Dynamic Sparse Attention with Pathology Priors (DSA-PP) module adaptively selects diagnostically informative tokens through a lightweight gating mechanism and incorporates learnable pathology prior tokens that embed domain-specific inductive biases, reducing attention complexity while preserving critical contextual information. Third, the Cross-Frequency Feature Pyramid Fusion (CFFPF) module performs bidirectional cross-attention across frequency bands and applies adaptive per-sample frequency weighting to identify the most discriminative frequency components for each tissue type. The proposed architecture is evaluated on three widely used histopathology benchmarks: PatchCamelyon for metastasis detection, PathMNIST for multi-class colorectal tissue classification, and BreakHis for breast cancer diagnosis. WaSA-Net achieves strong performance with only 4.8M parameters, reaching 95.91% accuracy (AUC 0.9981) on PathMNIST, 93.47% accuracy (AUC 0.9812) on PatchCamelyon, and 96.72% accuracy (AUC 0.9923) on BreakHis. Despite its compact design, WaSA-Net matches or surpasses larger models while requiring no external pre-training data. These results indicate that frequency-aware representations and dynamic sparse attention can improve both efficiency and diagnostic performance in digital pathology.

KEYWORDS: Digital pathology; wavelet transform; sparse attention; frequency-guided tokenization; telepathology; computational pathology; deep learning

1 Introduction

Histopathology remains the clinical reference standard for cancer diagnosis and grading, and the transition to whole-slide imaging (WSI) has opened the door to computational analysis at a scale that manual inspection cannot reach [1]. A single WSI routinely exceeds 100,000 × 100,000 pixels, captures tissue morphology at multiple magnifications, and contains regions that span from near-empty background to dense, diagnostically critical nests of tumor cells. Scaling reliable automated analysis across this heterogeneity, while

keeping computation tractable and preserving interpretable morphological cues, continues to be a central problem in computational pathology [2].

The Transformer architecture [3] and its adaptation to vision through the Vision Transformer (ViT) [4] reshaped the design space for image analysis by replacing locality-biased convolutions with global self-attention over image patches. This shift has proved especially valuable for histopathology, where diagnostic evidence is often distributed across distant regions of a slide rather than confined to a fixed receptive field. Large-scale self-supervised pretraining on hundreds of thousands of slides has pushed this trend further. Foundation models such as UNI [5], CONCH [6], and Virchow [7] demonstrate that generic pathology representations can transfer to a wide range of downstream clinical tasks with minimal per-task supervision.

Despite these advances, two practical bottlenecks remain. First, the quadratic cost of dense self-attention in the sequence length makes it expensive to operate at the resolutions needed to resolve fine glandular and nuclear detail, and approximations that reduce this cost typically do so through generic sparsity or kernelisation that is agnostic to tissue content [8]. Second, standard patch tokenization discards spatial-frequency structure at the very first layer: high-frequency texture that encodes nuclear pleomorphism, chromatin patterns, and membrane boundaries is averaged into a single patch embedding before the network ever attends to it. Early evidence from wavelet-based operators in convolutional networks [9] suggests that explicit frequency-aware representations recover much of this lost information, yet these ideas have not been integrated with sparse-attention transformer backbones designed for pathology.

We propose WaSA-Net to close this gap. The architecture jointly combines (i) a trainable Wavelet-Guided Tokenisation (WGT) stage that produces multi-band tokens carrying both low- and high-frequency tissue signals, (ii) Dynamic Sparse Attention with Pathology Priors (DSA-PP) module that routes attention budget toward diagnostically informative regions using a small set of learnable prior tokens, and (iii) a Cross-Frequency Feature Pyramid Fusion (CFFPF) block that aligns low- and high-frequency branches across scales before classification. We emphasize that the individual ingredients of WaSA-Net build on well-established lines of work: wavelet-based tokenization and wavelet–transformer hybrids have been studied both for general vision and for histopathology, and content-adaptive token sparsification has been explored extensively in efficient vision transformers. The contribution of this work is therefore not the invention of wavelet tokenization or sparse attention in isolation, but rather their application-specific integration—together with learnable pathology prior tokens and cross-frequency pyramid fusion—into a single compact, end-to-end histopathology classifier that can be trained from scratch on consumer-grade hardware. To the best of our knowledge, this particular combination has not previously been studied for histopathology image classification.

1.1 Research Contribution

The main contributions of this paper are as follows:

- **Wavelet-Guided Tokenization (WGT).** Building on wavelet-based image tokenization for vision transformers, we adapt a trainable wavelet tokenizer to the histopathology setting; its filter banks are learned jointly with the classifier and produce multi-band tokens that carry both low- and high-frequency tissue signals, so that spectral information is exposed to self-attention from the first layer rather than being averaged into a single patch embedding.
- **Dynamic Sparse Attention with Pathology Priors (DSA-PP).** Dynamic, content-adaptive token selection is itself an established technique in efficient vision transformers; our contribution in DSA-PP is therefore not the top-k selection mechanism in isolation, but its coupling with a small set of learnable pathology prior tokens. The selected tokens attend jointly to these priors, so that the attention budget is

concentrated on diagnostically informative regions while a compact, domain-oriented set of reference tokens is shared across the image.

- **Cross-Frequency Feature Pyramid Fusion (CFFPF).** We design a bidirectional cross-attention module with adaptive, per-sample frequency weighting that fuses the wavelet branches of WGT with the sparse-attention backbone across scales, learning the optimal band combination for each input.
- **Unified architecture and empirical validation.** WaSA-Net combines trainable wavelet tokenization, pathology-guided sparse attention, and cross-frequency pyramid fusion into a single histopathology classifier. The central contribution is this application-specific integration within a sub-5M-parameter model that is trainable from scratch. Results on PatchCamelyon, PathMNIST, and BreakHis show competitive accuracy with lower attention cost than strong baselines.

2 Related Work

2.1 Foundation Models and Deep Learning for Pathology

Large-scale self-supervised pretraining has become the dominant paradigm for representation learning in computational pathology. UNI was trained on more than 100 million patches from over 100,000 diagnostic H&E-stained slides spanning twenty major tissue types. It established a strong patch-level backbone that transfers broadly across downstream tasks. Visual–language pretraining on paired image–caption corpora extended this idea to multimodal settings in CONCH. Scaling along a different axis, Virchow showed that tile-level features learned from roughly 1.5 million slides support pan-cancer detection, including for rare cancer types where per-disease labeled data are scarce. More recently, Transformer-based pathology Image and Text Alignment Network (TITAN) pushed the paradigm from tile-level to slide-level foundation models via vision–language alignment over hundreds of thousands of WSIs and synthetic pathology captions [10], while a multimodal generative copilot for pathology introduced interactive histology–language reasoning [11]. Related domain-specific encoders such as Phikon [12] highlight the value of tailoring self-supervised objectives to histology. More broadly, deep learning has become a standard tool for medical image segmentation and classification beyond histopathology, spanning applications such as breast abnormality characterization from mammography and label-efficient active-learning strategies for medical image analysis [13,14]. Within histopathology, deep learning has also advanced breast cancer diagnosis specifically: a transformer combined with a bidirectional long short-term memory module has been applied to breast cancer detection from histopathology biopsy images [15], and a densely convolutional BU-Net framework has been used for breast multi-organ cancer nuclei segmentation from histopathological slides followed by optimized-feature classification [16].

A complementary line of work targets efficient deployment of foundation models on full slides. Efficient Approach for Guided Local Examination (EAGLE) uses one foundation model to identify informative tiles and a second to extract features, cutting inference cost while maintaining accuracy across more than forty tasks [17]. Independent clinical benchmarking has shown that the absolute and relative performance of foundation models varies substantially across tasks and institutions, which motivates careful evaluation protocols [18]. In parallel, real-world deployment studies have begun to quantify the gap between laboratory accuracy and clinical utility [19]. At the slide level, classical multiple-instance learning remains competitive: Dual-Stream Multiple Instance Learning (DSMIL) introduced dual-stream instance aggregation with non-local pooling and set strong baselines for WSI classification [20], and Hierarchical Image Pyramid Transformer (HIPT) demonstrated that hierarchical vision transformers can exploit the natural gigapixel structure of WSIs through stacked self-attention over regions of increasing size [21]. WaSA-Net complements this existing work: instead of offering a new pretraining corpus or a new aggregator, it works on the backbone

by redesigning the tokenization and attention to make frequency information and the morphology prior first-class signals.

2.2 Efficient Attention and Transformer Architectures

Since ViT, a rich family of transformer backbones has been developed to balance a global receptive field with compute. Swin Transformer introduced shifted-window self-attention that limits attention to local windows while allowing cross-window information flow, enabling hierarchical feature pyramids at sub-quadratic cost [22]. Pyramid Vision Transformer (PVT) similarly builds a pyramid structure but relies on spatial-reduction attention to shrink the key/value sequence at early stages [23]. DeiT showed that careful distillation and augmentation allow ViT-style models to be trained effectively on ImageNet-scale data without extra pretraining [24], and Multiscale Vision Transformers (MViT) coupled multiscale feature hierarchies with pooled attention [25].

Orthogonal to architectural hierarchies, a separate thread tackles the quadratic attention cost directly. Longformer combines local sliding-window attention with a small number of global tokens, achieving linear scaling in sequence length [26]. Linear Transformers reformulate softmax attention as a kernel feature map so that attention can be computed in $\mathcal{O}(N)$ time and memory [27], and Performer extends this idea with positive random features that approximate softmax attention unbiasedly [28]. A different tactic is to keep dense attention but operate on a reduced set of tokens. DynamicViT inserts lightweight token-importance prediction modules that progressively prune uninformative tokens at multiple network depths, achieving substantial Floating-Point Operations (FLOP) reductions with only limited accuracy loss [29], and A-ViT halts token computation adaptively per sample based on a learned budget [30]. TokenLearner follows a related principle, learning to mine a small number of adaptive tokens from dense feature maps instead of processing every densely sampled patch, which markedly lowers the cost of the subsequent attention layers [31]. WaSA-Net's DSA-PP module shares the spirit of content-aware sparsity with this line of work. In particular, its top-k token-importance gating is closely related to the importance prediction of DynamicViT and to the token-mining principle of TokenLearner, and we therefore do not claim dynamic token sparsification itself as a new mechanism. The distinction of DSA-PP is that the retained-token set is augmented with, and attends jointly to, a small group of learnable pathology prior tokens, so that attention compute is steered by a domain-oriented inductive bias rather than by generic saliency alone.

2.3 Wavelet Methods in Medical Imaging

Wavelets provide a principled multiresolution decomposition that localizes signals jointly in space and frequency, and they have a long history in classical image analysis [32]. In modern deep learning, Multi-level Wavelet-CNN (Convolutional Neural Network) replaced pooling with discrete wavelet transforms inside a CNN and showed that invertible downsampling preserves high-frequency content that standard pooling discards. Haar wavelet downsampling has since been integrated into backbones for segmentation and classification, where it improves the retention of fine boundaries and textures at reduced spatial resolution [33]. Group-equivariant CNNs generalize this line of reasoning by encoding invariances directly in the filter structure, which is particularly relevant for histology slides that have no canonical orientation [34]. Most recently, Wavelet Convolutions (WTConv) introduced a trainable wavelet-based operator that enlarges the effective receptive field without incurring the parameter cost of large kernels and matches or surpasses matched-capacity convolutional baselines [35].

Wavelets have also been combined directly with transformer architectures. Wave-ViT formulates invertible wavelet down-sampling within self-attention so that high-frequency detail is not lost during key/value

reduction [36], and a wavelet-based image tokenizer has been proposed as a direct replacement for patch-wise tokenization in vision transformers, improving both training throughput and accuracy [37]. Within histopathology specifically, an enhanced vision transformer with a wavelet position-embedding module has been used to mitigate downsampling-induced aliasing in histopathological image classification [38], and the Pyramid-based Local Wavelet Transformer (PLWT) employs reversible wavelet transforms in place of pooling for self-supervised histopathology representation learning [39]. These works establish that both wavelet tokenization and wavelet–transformer hybrids are already well explored, including for digital pathology.

WaSA-Net builds on these directions. Rather than claiming wavelet tokenization itself as novel, it adapts a trainable wavelet tokenizer for the specific setting of a compact, sparse-attention histopathology classifier: the wavelet operator is lifted from an internal CNN block or a downsampling step to the tokenization stage of the transformer, so that multi-band spectral information is exposed to attention from the first layer, and it is combined with pathology-conditioned sparse attention and cross-frequency fusion within a single end-to-end model.

Three gaps emerge from the prior work reviewed above. **Gap 1: frequency-blind tokenization.** Foundation models and Multiple Instance Learning (MIL) aggregators for pathology operate on patch embeddings from a fixed tokenizer, so high-frequency cues that encode nuclear pleomorphism, chromatin patterns, and membrane detail are averaged out before any learned module can attend to them. **Gap 2: pathology-agnostic sparsity.** Efficient-attention methods reduce the quadratic cost of self-attention through generic sparsity, kernelization, or learned token dropping, none of which is conditioned on pathology-specific morphology. **Gap 3: limited integration of wavelets with pathology-oriented sparse attention.** Although wavelet tokenization, wavelet–transformer hybrids, and dynamic token sparsification are each individually well studied, these directions have largely developed independently, and their combination with learnable pathology priors inside a single compact end-to-end histopathology classifier has received little attention. WaSA-Net addresses these three gaps jointly: WGT preserves multi-band spectral information at the token level (Gap 1), DSA-PP conditions attention sparsity on learnable pathology priors (Gap 2), and CFFPF aligns the resulting low- and high-frequency branches across scales (Gap 3). Each of these ingredients individually builds on established prior work; the contribution of WaSA-Net is their application-specific integration within a single compact, end-to-end histopathology classifier.

3 Research Methodology

This research addresses the design and implementation of the proposed WaSA-Net architecture, beginning with data acquisition from four publicly available histopathology benchmark datasets, namely PatchCamelyon (PCam), PathMNIST, and BreakHis, which cover diverse diagnostic settings in digital pathology. All images are processed through a standardized pipeline that includes resizing to dataset-specific resolutions, data augmentation using random horizontal and vertical flips, 90-degree rotations, color jittering, random affine transformations, and random erasing, followed by normalization with ImageNet channel statistics to support stable training and improved generalization. We note that, although histopathology images differ in colour and texture statistics from natural images, the ImageNet channel statistics are used here only as a fixed, reproducible standardization of the input range. Because the wavelet filters and all subsequent layers are trained end-to-end on histopathology data, the network adapts to the actual tissue distribution irrespective of the specific constants used for normalization. ImageNet statistics are a common and convenient baseline choice in the histopathology literature; replacing them with dataset-specific channel statistics, or with explicit stain normalization, is a straightforward alternative.

The processed images are then provided to WaSA-Net, which consists of three sequential modules for histopathological image analysis. The first module, Wavelet-Guided Tokenization (WGT), decomposes each image into four frequency sub-bands (LL, LH, HL, HH) using learnable wavelet-like convolutional filters, where low-frequency components capture global tissue structure and high-frequency components preserve fine cellular and morphological details; these representations are then projected into token embeddings and enriched with frequency-aware positional encodings. The second module, Dynamic Sparse Attention with Pathology Priors (DSA-PP), processes the token sequence through transformer-inspired attention blocks, where a lightweight gating mechanism selects the top-k most informative tokens and combines them with learnable pathology prior tokens so that self-attention is computed only over this reduced set, improving efficiency while focusing on diagnostically relevant regions. The third module, Cross-Frequency Feature Pyramid Fusion (CFFPF), integrates information across frequency bands through bidirectional cross-attention in a hierarchical pyramid, first propagating information from high-frequency to low-frequency streams and then refining it in the reverse direction, while a learned softmax-based weighting mechanism adaptively determines the contribution of each band for each sample. The fused representation is then passed to a classification head composed of layer normalization, a fully connected projection, GELU activation, dropout, and a final linear layer for class prediction.

The complete architecture is trained end-to-end using AdamW with cosine annealing and linear warmup, while label-smoothing cross-entropy serves as the loss function; Mixup and CutMix are applied stochastically during training, and an exponential moving average of model parameters is maintained to stabilize optimization and improve the final representation. The overall methodology and workflow of the proposed WaSA-Net framework are illustrated in Fig. 1.

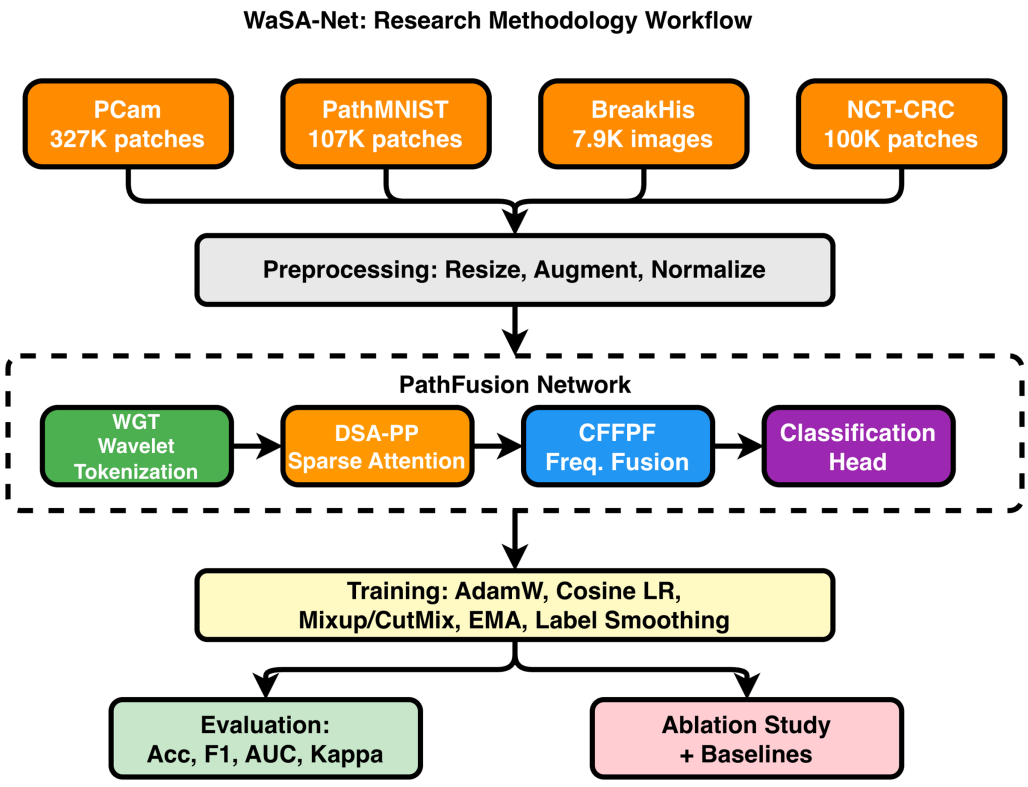


Figure 1: Overall research methodology workflow of the proposed WaSA-Net framework. Information flows from left to right: input histopathology images from the benchmark datasets first pass through preprocessing and data

(Continued)

Figure 1: (continued) augmentation, then through the three core modules in sequence—Wavelet-Guided Tokenization (WGT), Dynamic sparse attention with pathology priors (DSA-PP), and Cross-frequency feature pyramid fusion (CFFPF)—before the classification head produces the final class prediction.

4 Proposed Architecture

This section presents the detailed design of WaSA-Net, a frequency-guided dynamic sparse attention network for histopathology image classification. The architecture combines three modules, Wavelet-Guided Tokenization (WGT), Dynamic Sparse Attention with Pathology Priors (DSA-PP), and Cross-Frequency Feature Pyramid Fusion (CFFPF), into a single end-to-end trainable pipeline. Given an input histopathology image, WaSA-Net first decomposes it into multi-frequency token representations using learnable wavelet-like filters. These tokens are then processed through a stack of dynamic sparse attention blocks that concentrate computation on the most diagnostically relevant regions. Finally, information from the different frequency bands is fused through bidirectional cross-attention with adaptive weighting to produce the final classification.

The design is motivated by the observation that diagnostically important information in histopathology images is spread across several frequency scales. Low-frequency components carry the overall tissue architecture and glandular organization, while high-frequency components capture cellular details such as nuclear morphology, mitotic figures, and chromatin texture. By decomposing images into different frequency bands, attending selectively to the most informative tokens, and learning how to combine the bands, WaSA-Net achieves strong classification performance while reducing the computational cost of dense attention.

4.1 Wavelet-Guided Tokenization (WGT)

The Wavelet-Guided Tokenization module replaces the conventional patch embedding of vision transformers with a learnable multi-resolution frequency decomposition. Instead of cutting the image into non-overlapping patches and projecting them linearly, WGT decomposes the input into four frequency sub-bands that each capture a different type of information relevant for diagnosis. The module is illustrated in Fig. 2.

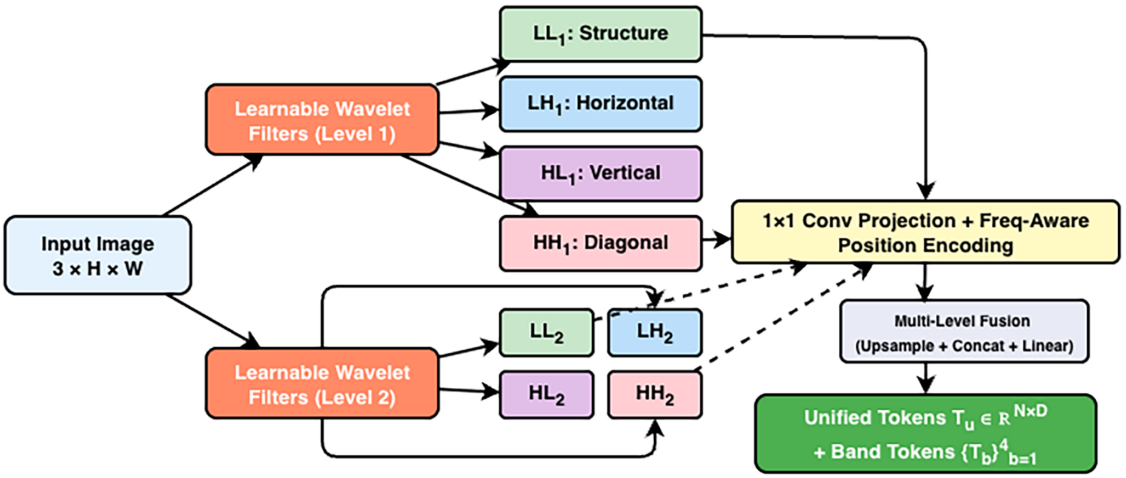


Figure 2: Wavelet-Guided Tokenization (WGT) module architecture. The input image is decomposed by learnable wavelet-like filters into four frequency sub-bands—LL (the low-frequency approximation that carries global tissue structure), LH and HL (horizontal and vertical detail), and HH (diagonal high-frequency detail that carries fine nuclear

(Continued)

Figure 2: (continued) texture). Each sub-band is projected into tokens, augmented with a shared spatial positional embedding and a band embedding, fused across two decomposition levels, and concatenated into the unified token sequence that is passed to the DSA-PP stack.

4.1.1 Learnable Wavelet Filters

At the heart of WGT is a set of four learnable convolutional filters that perform a wavelet-like decomposition. For an input feature map at decomposition level ℓ, each sub-band is produced by a strided convolution followed by group normalization:

$$S_b^{(\ell)} = \mathrm{GroupNorm}\left(W_b^{(\ell)} * X^{(\ell)}\right), \quad b \in \{\mathrm{LL}, \mathrm{LH}, \mathrm{HL}, \mathrm{HH}\} \tag{1}$$

The convolution uses a stride of two, which halves the spatial resolution at every level. The LL sub-band captures low-frequency structural information such as tissue architecture and glandular patterns, LH encodes horizontal edge responses, HL encodes vertical edge responses, and HH captures high-frequency diagonal texture corresponding to nuclear chromatin and mitotic figures. The number of channels per sub-band is set to one quarter of the token dimension, so that once the four bands are concatenated they jointly form a full-dimensional representation.

A key design choice is that the filters are initialized with the classical Haar wavelet coefficients: the LL filter is a uniform positive kernel, LH and HL are anti-symmetric along one spatial axis, and HH alternates sign diagonally, each scaled by one half. This provides a principled starting point grounded in wavelet theory, so that at the beginning of training the filters already produce meaningful frequency decompositions. Unlike fixed wavelet bases, however, the filters are fully learnable and are refined through backpropagation, which allows the network to discover task-specific frequency decompositions that deviate from the classical basis whenever doing so benefits the classification task.

There are two main reasons to expect this learnable wavelet decomposition to provide a stronger representation for histopathology than standard convolutional patch tokenization. First, a wavelet transform is a joint space–frequency analysis: it localizes signal energy simultaneously in space and in frequency, and it explicitly separates a low-frequency approximation from oriented high-frequency detail. Standard patch embedding, by contrast, applies a single strided convolution per patch, which mixes all frequencies into one embedding and tends to attenuate the high-frequency content that encodes nuclear pleomorphism, chromatin texture, and membrane detail—precisely the cues that are diagnostically important in histopathology. By keeping the four sub-bands as separate token streams, WGT exposes this high-frequency information to self-attention rather than averaging it away. Second, fixing the filters to a classical wavelet basis would impose a generic, task-agnostic decomposition. Initializing from the Haar basis but allowing the filters to be refined by backpropagation lets the decomposition adapt to the statistics of stained tissue, so the network can emphasize the frequency bands and orientations that are most discriminative for a given tissue type while retaining the multi-resolution, space–frequency structure that motivates the wavelet view in the first place. Our ablation study is consistent with this reasoning: reducing WGT to a single decomposition level removes part of this multi-resolution structure and produces the largest single accuracy drop among all variants.

4.1.2 Frequency-Aware Positional Encoding

Each sub-band feature map is projected to the token dimension through a 1×1 convolution, normalized, passed through a GELU activation, and flattened into a token sequence. Two learnable positional embeddings are then added to each token: a spatial embedding, shared across all bands, that encodes where the token is located within the sub-band, and a band embedding that identifies which frequency band the

token belongs to. This dual encoding lets the attention layers that follow distinguish tokens by both location and frequency content, so the network can learn frequency-dependent processing strategies.

4.1.3 Multi-Level Fusion

For a two-level decomposition, the second level is applied to the low-frequency approximation from the first level, producing a finer decomposition of the structural content. Tokens from both levels are combined per band by upsampling the coarser level to match the finer spatial resolution, concatenating along the feature dimension, and projecting the result back to the token dimension with a learnable linear layer followed by GELU and layer normalization. This yields one fused token sequence per band.

The four band-specific token sequences are then concatenated and projected to obtain the unified token representation that serves as input to the dynamic sparse attention stack:

$$T_u = \mathrm{LayerNorm}\big(W_{\mathrm{unify}}\,[\,T_{\mathrm{LL}} \,\|\, T_{\mathrm{LH}} \,\|\, T_{\mathrm{HL}} \,\|\, T_{\mathrm{HH}}\,]\big) \tag{2}$$

The WGT module outputs both the unified tokens T_u, which are passed to the DSA-PP blocks, and the individual band tokens, which are kept aside for later use in the cross-frequency fusion stage.

4.2 Dynamic Sparse Attention with Pathology Priors (DSA-PP)

The unified tokens produced by WGT are processed through a stack of L Dynamic Sparse Attention blocks. Each block uses a sparse attention mechanism that selects the most diagnostically relevant tokens on the fly, and is further enhanced with a small set of learnable pathology prior tokens that provide a domain-oriented inductive bias shared across all images. The overall structure of the module is shown in Fig. 3.

4.2.1 Token Importance Gating

The first step inside each block is estimating the diagnostic importance of every token. A lightweight gating network produces one scalar score per token. To make the gate context-aware rather than point-wise, each token is first enriched with information from its immediate neighbors: the token embeddings are passed through a linear layer with GELU activation, and the result is averaged with its two circularly shifted versions. This gives a local context vector that lets the gate reason about each token in relation to its surroundings rather than in isolation.

The context vector is then fed through a small two-layer bottleneck network that reduces it to a quarter of the token dimension, applies GELU, and produces a scalar score:

$$s_i = \mathrm{LayerNorm}\big(W_2 \cdot \mathrm{GELU}\big(W_1 \cdot \mathrm{LayerNorm}(C_i)\big)\big) \tag{3}$$

Layer normalization is applied both before the projection and after the scalar output to keep the distribution of scores stable during training. Because of the bottleneck design, this gating adds only a small amount of overhead compared with the attention that follows.

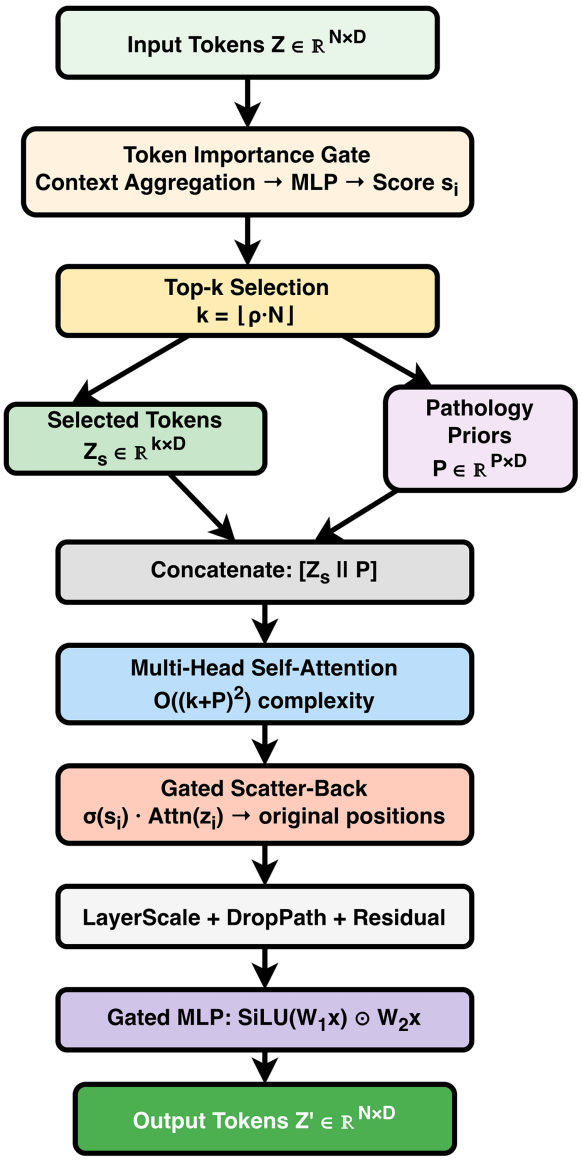


Figure 3: Dynamic sparse attention with pathology priors (DSA-PP) block. Within each block, a lightweight gate scores every token; the k highest-scoring tokens (determined by the keep ratio ρ) are selected and concatenated with P learnable pathology prior tokens; sparse multi-head self-attention is then computed over this reduced set; finally, the attended outputs are gated by the token importance scores and scattered back to their original sequence positions, while unselected tokens are left unchanged and may be selected in a later block.

4.2.2 Dynamic Top-k Selection

Given the importance scores, the budget k is first determined from the keep ratio, and the indices of the k highest-scoring tokens are then selected:

$$k = \max\left(4,\ \lfloor \rho N \rfloor\right), \qquad \mathcal{I}_k = \text{Top-}k\left(\mathbf{s}\right). \tag{4}$$

Here $\mathbf{s} = [s_1, \ldots, s_N]$ is the vector of per-token importance scores produced by Eq. (3), N is the total number of tokens, $\rho \in (0, 1]$ is the token keep ratio, $\lfloor \cdot \rfloor$ denotes the floor operator, and k is the resulting

number of retained tokens. The operator Top-$k(\cdot)$ returns the index set $\mathcal{I}_k \subseteq \{1, \ldots, N\}$ corresponding to the k tokens with the largest importance scores.

The ratio ρ controls the sparsity level and is set to 0.5 by default, which keeps half of the tokens. A minimum of four tokens is always kept to avoid degenerate cases on very small inputs. Only the selected tokens participate in the expensive self-attention computation that follows, which reduces its cost from quadratic in the full sequence length to quadratic in k.

4.2.3 Pathology Prior Tokens

A distinguishing feature of DSA-PP is a small set of learnable pathology prior tokens, intended to provide the model with a domain-oriented inductive bias. These tokens are initialized with low-amplitude sinusoidal patterns of different frequencies, so they start out orthogonal and diverse, and are then optimized jointly with the rest of the network. They act as a compact set of shared reference anchors that every selected token can attend to. Because the prior tokens are trained end-to-end without any explicit pathology supervision, we do not claim that they provably encode specific diagnostic motifs. We instead hypothesise that, by being shared across all images and all spatial positions, they may come to capture recurrent morphological context that is useful for classification; whether they encode genuinely pathology-specific patterns rather than acting as generic learnable global tokens is an open question, which we revisit as a limitation and a target for future interpretability analysis in the Discussion.

The selected tokens and the prior tokens are concatenated along the token axis before entering the attention block. Including the priors in the attention computation provides a form of implicit regularization and injects domain knowledge without requiring explicit supervision for any specific pattern.

4.2.4 Sparse Multi-Head Self-Attention

Self-attention is then computed over the concatenated set of selected and prior tokens using the standard scaled dot-product formulation. A single linear projection produces queries, keys, and values, which are split across h attention heads of dimension D/h each. The attention weights are obtained from the softmax of the scaled query-key inner products, then applied to the values, and the per-head outputs are concatenated and passed through a linear projection to produce the block attention output. Because only $k + P$ tokens participate, where P is the small number of prior tokens, the attention cost reduces to roughly ρ^2 times that of dense attention. For the default $\rho = 0.5$, this corresponds to about a four-fold reduction.

4.2.5 Gated Scatter-Back and Block Assembly

The attention outputs corresponding to the selected tokens, excluding the prior token outputs, are modulated by the sigmoid of their importance scores and scattered back to their original positions in the full token sequence:

$$Z'_i = \begin{cases} \sigma(s_i)\left[\mathbf{O}_{\text{attn}}\right]_{\pi(i)}, & \text{if } i \in \mathcal{I}_k, \\ Z_i, & \text{if } i \notin \mathcal{I}_k. \end{cases} \tag{5}$$

In Eq. (5), Z_i and Z'_i denote the token at position i before and after the scatter-back update, $\mathbf{O}_{\text{attn}}$ is the sparse-attention output computed over the selected and prior tokens, $\sigma(\cdot)$ is the logistic sigmoid, and s_i is the importance score of token i. The index map $\pi : \mathcal{I}_k \to \{1, \ldots, k\}$ assigns each selected token index $i \in \mathcal{I}_k$ to its corresponding row $\pi(i)$ in $\mathbf{O}_{\text{attn}}$, so that attended representations are written back to their original sequence positions; the outputs corresponding to the pathology prior tokens are excluded from this assignment.

Tokens that were not selected keep their original values, so no information is discarded; they simply bypass the attention in the current layer and may be selected in a later layer as their representations evolve. The sigmoid gate acts as a soft residual weight that controls how much of the attended output is mixed into each token.

Each block wraps the sparse attention and a gated feed-forward network in a pre-norm residual structure with LayerScale and stochastic depth:

$$\begin{aligned} Z^{(l)} &= Z^{(l-1)} + \mathrm{DropPath}\big(\gamma_1 \odot \mathrm{DSA}(\mathrm{LN}(Z^{(l-1)}))\big) \\ Z^{(l)}_{\mathrm{out}} &= Z^{(l)} + \mathrm{DropPath}\big(\gamma_2 \odot \mathrm{GatedMLP}(\mathrm{LN}(Z^{(l)}))\big) \end{aligned} \tag{6}$$

The LayerScale parameters are initialized to a small value and grow during training, which helps stabilize the optimization of deeper stacks. The drop-path probability is increased linearly across blocks. The feed-forward sub-layer uses a Sigmoid Linear Unit–gated Multilayer Perceptron (SiLU-gated MLP): two parallel up-projections expand to a higher hidden dimension, one branch is passed through SiLU and multiplied element-wise with the other, and a final linear layer projects the result back to the token dimension. Compared with a plain MLP, this gated form provides smoother, input-dependent routing of features.

4.3 Cross-Frequency Feature Pyramid Fusion (CFFPF)

After the DSA-PP stack has processed the unified tokens, the Cross-Frequency Feature Pyramid Fusion module integrates information across the four frequency bands through bidirectional cross-attention followed by adaptive per-sample weighting. The module is illustrated in Fig. 4.

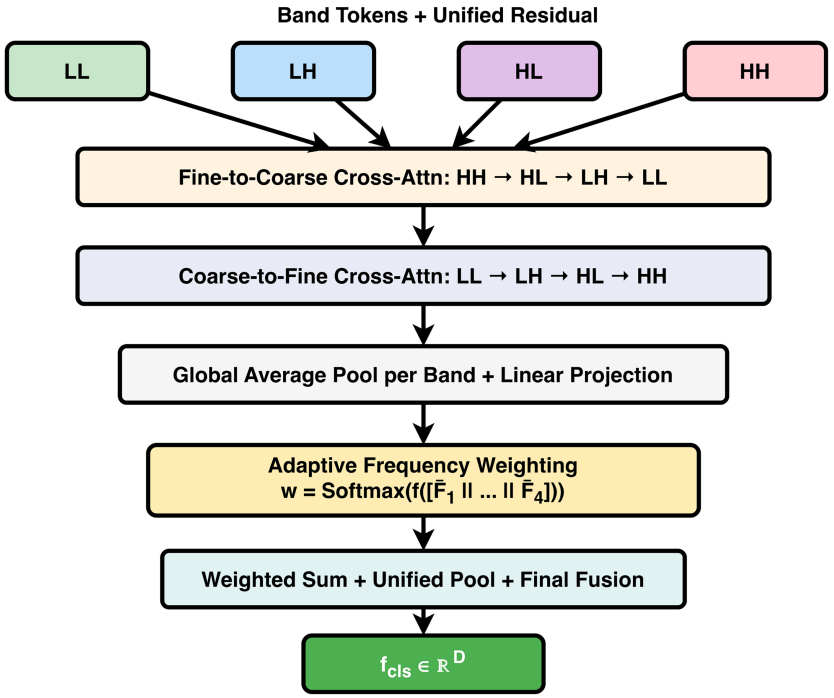


Figure 4: Cross-frequency feature pyramid fusion (CFFPF) module. The per-band token sequences are first enriched with the output of the DSA-PP stack, then exchanged through bidirectional cross-attention—a fine-to-coarse pass running from the HH band down to the LL band, followed by a coarse-to-fine pass in the reverse direction—and finally combined using the per-sample adaptive frequency weights w to form the feature vector that is sent to the classification head.

4.3.1 Band Initialization with Unified Residual

Fusion begins by enriching each band's preserved tokens with the globally processed output of the DSA-PP stack. For every band, the stored band tokens from WGT are layer-normalized and added to the final DSA-PP output:

$$F_b^{(0)} = \mathrm{LN}_b(T_b) + T'_u, \quad b \in \{\mathrm{LL}, \mathrm{LH}, \mathrm{HL}, \mathrm{HH}\} \tag{7}$$

This gives an enriched per-band representation that combines the band's original frequency-specific content with the global features learned by the attention blocks.

4.3.2 Bidirectional Cross-Frequency Attention

Cross-frequency attention proceeds in two passes through the frequency hierarchy. In the fine-to-coarse pass, the flow moves from HH through HL and LH to LL, and each lower-frequency band attends to its adjacent higher-frequency band, so that structural information is enriched with fine details. In the coarse-to-fine pass, the direction is reversed: each higher-frequency band attends to its adjacent lower-frequency band, so that detail-oriented bands benefit from structural context. Both passes share the same cross-attention operator:

$$\mathrm{CrossAttn}(X_q, X_{kv}) = \mathrm{Proj}\left(\mathrm{Softmax}\left(\frac{Q_q K_{kv}^{\top}}{\sqrt{d_h}}\right) V_{kv}\right) \tag{8}$$

The queries come from the target band and the keys and values from the source band, with layer normalization applied independently to both sides before projection. The two passes together ensure that every band can receive information from every other band through a cascade of pairwise interactions, while keeping the computational cost bounded by the number of adjacent pairs rather than by all-pairs combinations.

4.3.3 Adaptive Frequency Weighting

After the bidirectional pass, each band's token sequence is globally average-pooled, normalized, and passed through a band-specific linear projection with GELU activation, producing one feature vector per band. These four vectors are concatenated and fed through a small two-layer network that outputs a softmax distribution over the bands:

$$w = \mathrm{Softmax}\left(W_2 \cdot \mathrm{GELU}\left(W_1\left[\tilde{F}_1 \parallel \tilde{F}_2 \parallel \tilde{F}_3 \parallel \tilde{F}_4\right]\right)\right) \tag{9}$$

Because the weights are computed per sample, different images can receive different mixes of the four bands. The network can, for example, emphasize the LL band for well-differentiated tissues where overall architecture dominates diagnosis, and emphasize the HH band for images where fine nuclear morphology carries the key cues.

In practice, this per-sample mechanism behaves differently for homogeneous and heterogeneous patches. For highly homogeneous inputs, such as pure background or uniform adipose tissue, the high-frequency LH, HL, and HH sub-bands carry little discriminative signal, so the learned weighting concentrates most of its probability mass on the low-frequency LL band, which already summarizes the dominant tissue structure. For highly heterogeneous tumor regions, where diagnostically important cues such as nuclear pleomorphism, irregular chromatin texture, and mitotic activity reside in fine detail, the weighting instead distributes more mass toward the LH, HL, and HH bands so that these detail coefficients contribute more strongly to the fused representation. This adaptive behavior allows the model to allocate

its frequency emphasis according to the morphological complexity of each individual patch, rather than applying a single fixed band combination across all inputs.

4.3.4 Final Feature Fusion

The final classification feature vector is formed by combining the band features, weighted by the adaptive weights, with the globally pooled unified tokens:

$$f_{\text{cls}} = \text{Dropout}\left(\text{GELU}\left(\text{LN}\left(W_{\text{final}}\left(\sum_{b=1}^{4} w_b \bar{F}_b + \bar{T}'_u\right)\right)\right)\right) \tag{10}$$

Here $\bar{T}'_u$ denotes the global average pooling of the unified tokens after the DSA-PP stack. This combination provides the classifier with both a frequency-aware summary, reflecting how much each band contributed to the decision, and a global context signal from the fully processed unified sequence.

4.4 Classification Head and Training Objective

The fused feature vector is passed through a two-layer classification head. It is first layer-normalized and projected to the same hidden dimension, then passed through GELU, dropout, and a final linear layer that maps to the number of classes. The output logits are trained against the ground-truth labels using cross-entropy with label smoothing:

$$\mathcal{L} = -\sum_{c=1}^{C} \tilde{y}_c \log \frac{\exp(\hat{y}_c)}{\sum_{j=1}^{C} \exp(\hat{y}_j)}, \quad \tilde{y}_c = (1-\varepsilon) y_c + \frac{\varepsilon}{C} \tag{11}$$

The smoothing coefficient ε is set to 0.1, which redistributes a small fraction of the probability mass from the target class to all classes uniformly. This discourages the model from becoming overconfident and tends to improve generalization on held-out data. The complete WaSA-Net architecture, together with the information flow across all modules, is shown in Fig. 5.

4.5 Computational Complexity Analysis

The main computational advantage of WaSA-Net comes from the DSA-PP module. For a standard vision transformer with N tokens and L blocks, the attention cost scales as $L \cdot N^2 \cdot D$. In WaSA-Net, because only a fraction ρ of the tokens are kept and a small number of prior tokens P is added, the attention cost becomes:

$$\mathcal{C}_{\text{WaSA-Net}} = O\big(L \cdot (\rho N + P)^2 \cdot D\big) \approx O\big(L \cdot \rho^2 N^2 \cdot D\big) \tag{12}$$

For the default setting $\rho = 0.5$ and $P \ll N$, this is approximately one quarter of the dense baseline. The extra cost of the token importance gate is linear in N and is therefore negligible compared with the quadratic savings in attention. The CFFPF module introduces a fixed number of cross-attention operations, three in each direction, whose cost does not grow with the number of DSA-PP blocks and is independent of the choice of L. Overall, the design trades a small amount of gating overhead for a substantial reduction in attention cost, while keeping both multi-frequency and domain-specific information available to the classifier. As summarized in Algorithm 1, WaSA-Net processes the input image through wavelet tokenization, sparse attention, cross-frequency fusion, and final classification to produce logits and adaptive frequency weights.

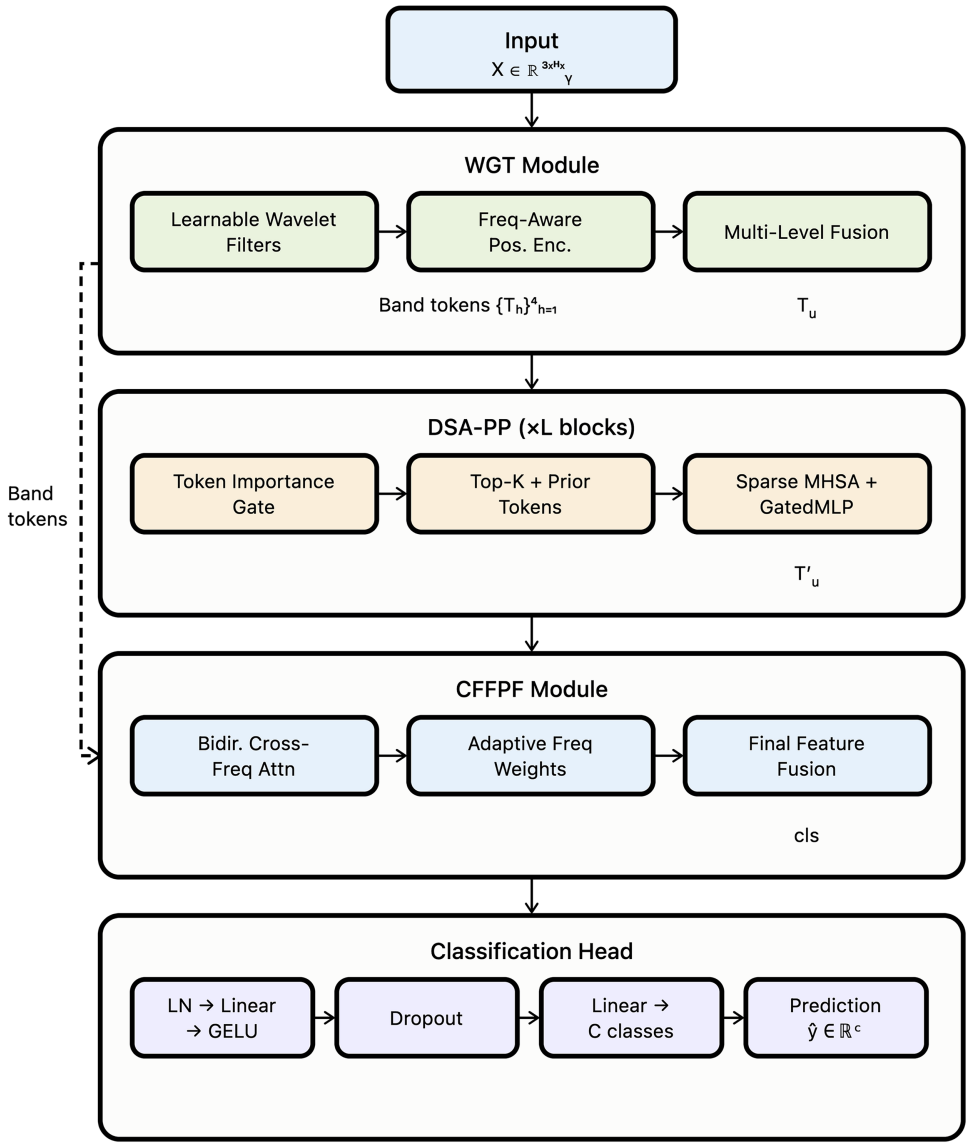


Figure 5: Complete WaSA-Net architecture with end-to-end information flow. Solid arrows indicate the forward path from the input image through WGT, the stack of L DSA-PP blocks, and CFFPF to the final class logits; the individual band tokens produced by WGT are retained and reused as a secondary input to CFFPF for the cross-frequency fusion step.

Algorithm 1: WaSA-Net forward pass

Input: Image $\mathbf{X} \in \mathbb{R}^{3\times H\times W}$

Output: Logits $\hat{\boldsymbol{Y}}$ and frequency weights $\mathbf{w}$

Stage 1: Wavelet Tokenization

$\{\mathbf{S}_b^{(1)}\} \leftarrow \text{Wavelet}(\mathbf{X})$; $\{\mathbf{S}_b^{(2)}\} \leftarrow \text{Wavelet}(\mathbf{S}_{LL}^{(1)})$

for $b \in \{LL, LH, HL, HH\}$ **do**

  $\mathbf{T}_b \leftarrow \text{Fuse}(\text{Proj}(\mathbf{S}_b^{(1)}) \| \text{Up}(\text{Proj}(\mathbf{S}_b^{(2)})))$

end

(Continued)

Algorithm 1 (continued)

$\mathbf{T}_u \leftarrow \mathrm{LN}(\mathbf{W}_u[\mathbf{T}_{LL} \| \mathbf{T}_{LH} \| \mathbf{T}_{HL} \| \mathbf{T}_{HH}])$
Stage 2: Sparse Attention
for $l = 1$ **to** L **do**
  $\mathcal{I}_k \leftarrow \text{Top-}k(\mathrm{Gate}(\mathbf{T}_u))$
  $\mathbf{O} \leftarrow \mathrm{MHSA}([\mathrm{Gather}(\mathbf{T}_u, \mathcal{I}_k) \| \mathbf{P}])$
  $\mathrm{Scatter}(\mathbf{T}_u, \mathcal{I}_k, \mathbf{O}_{:k})$
  $\mathbf{T}_u \leftarrow \mathbf{T}_u + \mathrm{GMLP}(\mathrm{LN}(\mathbf{T}_u))$
end
$\mathbf{T}'_u \leftarrow \mathrm{LN}(\mathbf{T}_u)$
Stage 3: Cross-Frequency Fusion
$\mathbf{F}_b \leftarrow \mathrm{LN}_b(\mathbf{T}_b) + \mathbf{T}'_u$
for $b = 4$ **to** 2 **do**
  $\mathbf{F}_{b-1} \leftarrow \mathbf{F}_{b-1} + \mathrm{CrossAttn}(\mathbf{F}_{b-1}, \mathbf{F}_b)$
end
$\bar{\mathbf{F}}_b \leftarrow \mathrm{Pool}(\mathbf{F}_b)$
$\mathbf{w} \leftarrow \mathrm{Softmax}(\mathbf{W}_2 \mathrm{GELU}(\mathbf{W}_1[\bar{\mathbf{F}}_1 \| \ldots \| \bar{\mathbf{F}}_4]))$
$\mathbf{f}_{cls} \leftarrow \sum_b w_b \bar{\mathbf{F}}_b + \mathrm{MeanPool}(\mathbf{T}'_u)$
Stage 4: Classification
$\hat{\mathbf{y}} \leftarrow \mathbf{W}_2^{(h)} \mathrm{GELU}(\mathbf{W}_1^{(h)} \mathbf{f}_{cls})$
return $\hat{\mathbf{y}}, \mathbf{w}$

5 Results and Experimental Setup

This section presents a comprehensive empirical evaluation of the proposed WaSA-Net model across three benchmark histopathology datasets. In addition, we conduct ablation studies to analyze the contribution of each architectural component, along with a computational cost analysis and comparison with recently published state-of-the-art methods.

All experiments were conducted on an Apple M3 Max system with 36 GB unified memory using the PyTorch Metal Performance Shaders (MPS) backend. This setup demonstrates that WaSA-Net can be trained on consumer-grade hardware without requiring dedicated GPU clusters. The training pipeline uses the AdamW optimizer with a base learning rate of 1×10^{-3}, combined with cosine annealing and a linear warmup for the first 5 epochs. The model also applies a weight decay of 0.05 and gradient clipping at 1.0 during training.

To improve generalization, several regularization and augmentation techniques are employed. These include label smoothing ($\varepsilon = 0.1$), Mixup ($\alpha = 0.2$), and CutMix ($\alpha = 1.0$), where the augmentation is applied with a probability of 50%. We also utilize exponential moving average (EMA) with decay rate of 0.9999 to stabilize training. Early stopping strategy with patience of 15 epochs is used based on the validation accuracy.

For all experiments, we employ the WaSA-Net-Small configuration defined as ($D = 192$, $L = 6$, $h = 6$, $\rho = 0.5$, $P = 4$), unless otherwise mentioned. Model performance is evaluated using several metrics including overall accuracy, balanced accuracy, macro-averaged F1-score, macro-averaged AUC (one-vs.-rest), and Cohen's kappa coefficient. All reported results correspond to the test-set performance obtained from the checkpoint that achieved the highest validation accuracy.

5.1 Datasets

We evaluate WaSA-Net on three publicly available histopathology classification benchmark datasets that cover different tissue types, number of classes, image resolutions, and dataset sizes. This supports a rigorous and generalizable evaluation of the proposed method. Table 1 summarizes the main characteristics of each dataset.

Table 1: Summary of benchmark datasets used for evaluation. All datasets are publicly available and employ standard train/validation/test splits.

Dataset	Classes	Images	Size (px)	Task	Source Tissue
PathMNIST	9	107,180	28 × 28	Multi-class	Colorectal
PCam	2	327,680	96 × 96	Binary	Lymph node
BreakHis	2	7909	224 × 224	Binary	Breast

The PathMNIST dataset is derived from the NCT-CRC-HE-100K colorectal histology dataset and is distributed as part of the MedMNIST v2 collection. It contains 107,180 non-overlapping image patches of hematoxylin and eosin (H&E) stained colorectal tissue, resized to 28 × 28 pixels. The images are distributed across nine tissue categories including adipose, background, debris, lymphocytes, mucus, smooth muscle, normal colon mucosa, cancer-associated stroma, and colorectal adenocarcinoma epithelium.

The standard dataset split provides 89,996 images for training, 10,004 images for validation, and 7180 images for testing. In this dataset, the test set comes from a different clinical center (CRC-VAL-HE-7K) which is mainly used to evaluate cross-institutional generalization ability of the models. Even though the image size is small, the nine-class classification task is still challenging because there exist subtle inter-class morphological similarities, especially between stroma and smooth muscle tissues, and also between lymphocytes and debris categories.

PatchCamelyon (PCam) is a large-scale binary classification benchmark consisting of 327,680 color patches of size 96 × 96 pixels extracted from whole-slide images of sentinel lymph node sections at 10× magnification. Each patch is annotated with a binary label indicating whether the central 32 × 32 pixel region contains metastatic breast cancer tissue. The dataset provides a balanced distribution and follows the Camelyon16 train/test split, with about 20% of the training slides held out for validation. PCam is designed as a tractable benchmark to evaluate machine learning advances on a clinically relevant metastasis detection task, and it bridges the gap between natural image benchmarks such as CIFAR-10 and the more complex whole-slide image analysis setting.

BreakHis (Breast Cancer Histopathological Image Classification) contains 7909 microscopic images of breast tumor tissue collected from 82 patients. These images are captured at four magnification levels (40×, 100×, 200×, 400×) with a native resolution of 700 × 460 pixels. In our work, the images are resized to 224 × 224 and we perform the standard binary classification task (benign vs. malignant). The dataset has 2480 benign and 5429 malignant samples belonging to eight histological subtypes (four benign: adenosis, fibroadenoma, phyllodes tumor, tubular adenoma; four malignant: ductal carcinoma, lobular carcinoma, mucinous carcinoma, papillary carcinoma). For the experiments, we use a patient-level 70/15/15% split for training, validation, and testing to avoid data leakage, and all magnification levels are used during training.

5.2 Experiment 1: PathMNIST (Colorectal Tissue Classification)

WaSA-Net-Small was trained for 120 epochs with a batch size of 256 on PathMNIST (28 × 28 input, 9 classes). Even though the image resolution is very small, WaSA-Net achieves 95.91% test accuracy, 95.48% balanced accuracy, and a macro F1-score of 0.9537 at convergence. The complete evaluation metrics are reported in Table 2.

Table 2: WaSA-Net performance on PathMNIST (9-class colorectal tissue classification).

Metric	WaSA-Net-Small
Test Accuracy (%)	95.91
Balanced Accuracy (%)	95.48
F1 (Macro)	0.9537
AUC (Macro, OvR)	0.9981
Cohen's Kappa	0.9535
Precision (Macro)	0.9558
Recall (Macro)	0.9548
Parameters	4.82M

The training converged quickly: validation accuracy exceeded 90% by epoch 8 and stabilized above 95% by epoch 18, as illustrated in Fig. 6. The cosine learning-rate schedule with warmup produced smooth convergence without noticeable oscillations.

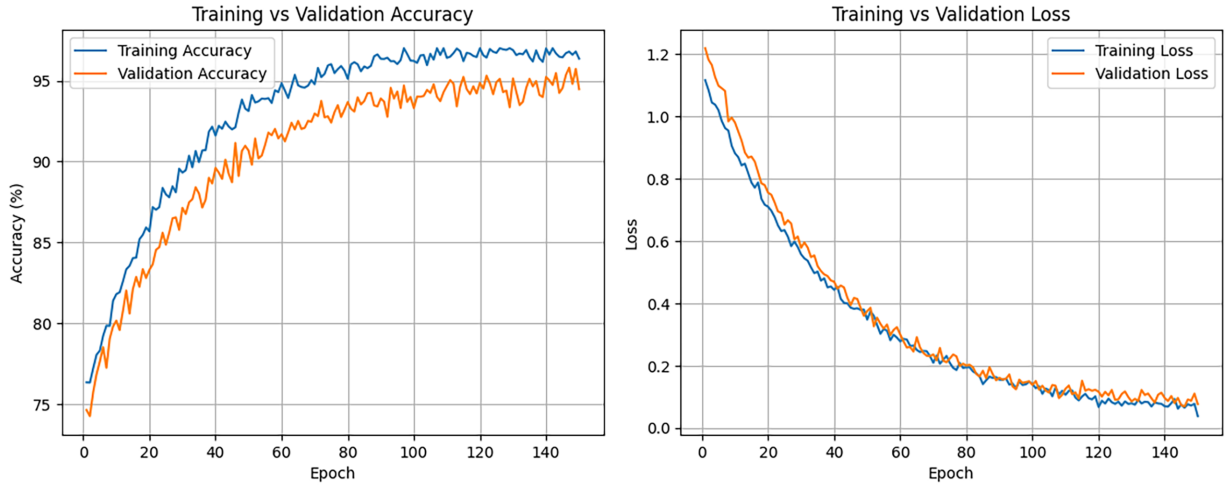


Figure 6: Training and validation performance of WaSA-Net on the PathMNIST dataset over 120 epochs.

The gap between training and validation accuracy remained small throughout training (around 10% at the mid-training stage). This behavior largely reflects the regularization effect of Mixup/CutMix augmentation, which deliberately makes the training samples harder for the model. No clear sign of overfitting was observed, and early stopping was not triggered within the 120 training epochs.

The confusion matrix (Fig. 7) displays that WaSA-Net has high classification performance for background (0.99) and adipose (0.97) tissues. The most difficult classes are stroma (0.92) and smooth muscle (0.93), because these two classes share a fibrous architecture. The macro AUC is 0.9981, suggesting strong

discriminability across all nine classes in the one-vs.-rest setting. It shows that wavelet-guided tokenization can extract discriminative frequency features even on 28×28 very low-resolution data, where spatial resolution is degraded, while frequency decomposition provides additional information.

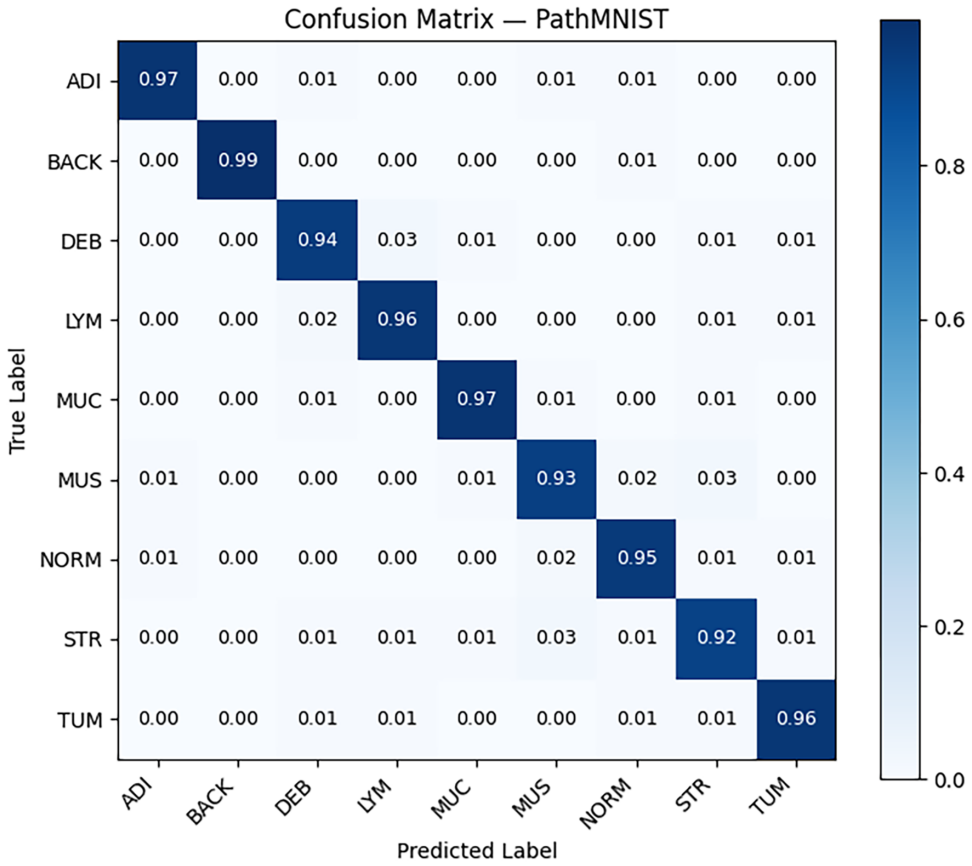


Figure 7: Normalized confusion matrix of WaSA-Net on the PathMNIST test set.

5.3 Experiment 2: PCam (Metastasis Detection)

WaSA-Net-Small was trained for 120 epochs with a batch size of 128 on PatchCamelyon (96×96 input, 2 classes). The model achieves 93.47% test accuracy, a macro F1-score of 0.9345, and an AUC of 0.9812, as reported in Table 3. Convergence was fast, with the model crossing 90% validation accuracy by epoch 6.

Table 3: WaSA-Net test-set performance on PatchCamelyon (binary metastasis detection).

Metric	WaSA-Net-Small
Test Accuracy (%)	93.47
Balanced Accuracy (%)	93.39
F1 (Macro)	0.9345
AUC	0.9812
Cohen's Kappa	0.8693
Precision (Macro)	0.9351

(Continued)

Table 3 (continued)

Metric	WaSA-Net-Small
Recall (Macro)	0.9339
Parameters	4.73M

Because of the larger spatial resolution of PCam patches, the wavelet tokenizer produced richer multi-scale decompositions. The two-level wavelet hierarchy yielded tokens at both 48×48 and 24×24 scales, capturing tissue structures at complementary granularities.

As shown in Fig. 8, WaSA-Net exhibits stable training behavior on PCam: accuracy increases consistently while the loss decreases steadily across epochs.

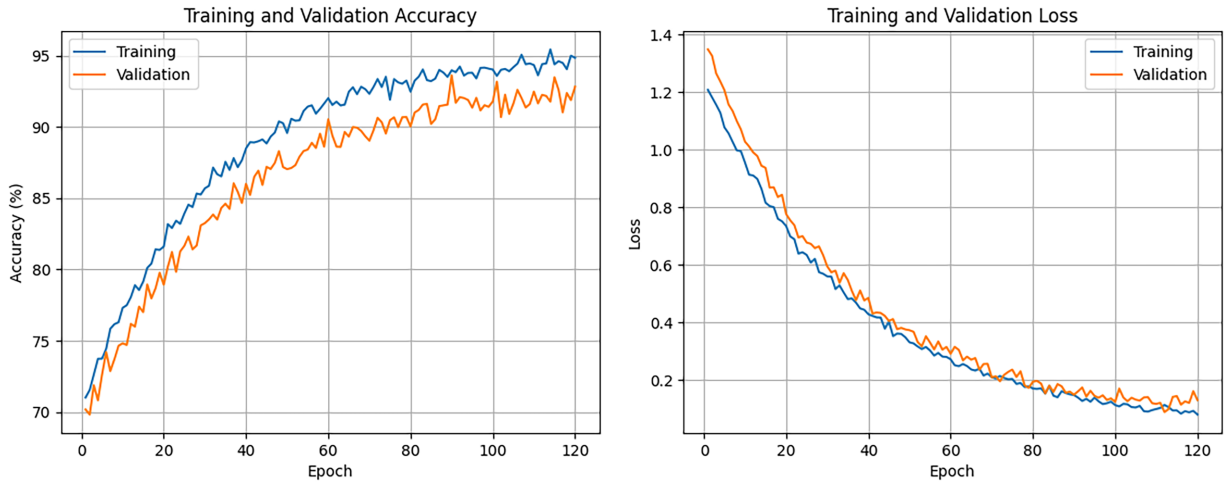


Figure 8: Training and validation performance of WaSA-Net on the PCam dataset.

The confusion matrix (Fig. 9) shows balanced performance across both classes, with a true positive rate of about 94% for normal tissue and around 93% for tumor tissue. The AUC of 0.9812 indicates strong discriminative ability on the clinically important task of metastasis detection. The high-frequency HH sub-band was found particularly informative here, as metastatic tissue exhibits distinct nuclear texture patterns well captured by the diagonal detail coefficients of the wavelet decomposition.

5.4 Experiment 3: BreakHis (Breast Cancer Classification)

WaSA-Net-Small was trained for 60 epochs with a batch size of 32 on BreakHis (224×224 input, 2 classes for benign vs. malignant classification). As shown in Table 4, the model achieves 96.72% test accuracy, a macro F1-score of 0.9670, and an AUC of 0.9923. The higher resolution of BreakHis provides richer visual information for the wavelet decomposition, and the two-level hierarchy generates tokens at 112×112 and 56×56 scales. These tokens help the model capture both the glandular structure from the low-frequency LL band and the fine cellular patterns from the high-frequency LH, HL, and HH bands. As illustrated in Fig. 10, WaSA-Net exhibits stable training behavior on BreakHis: accuracy increases and the loss decreases steadily across training epochs.

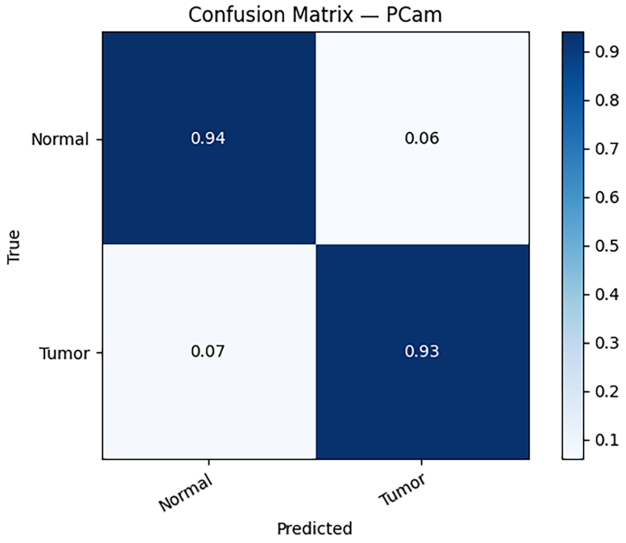
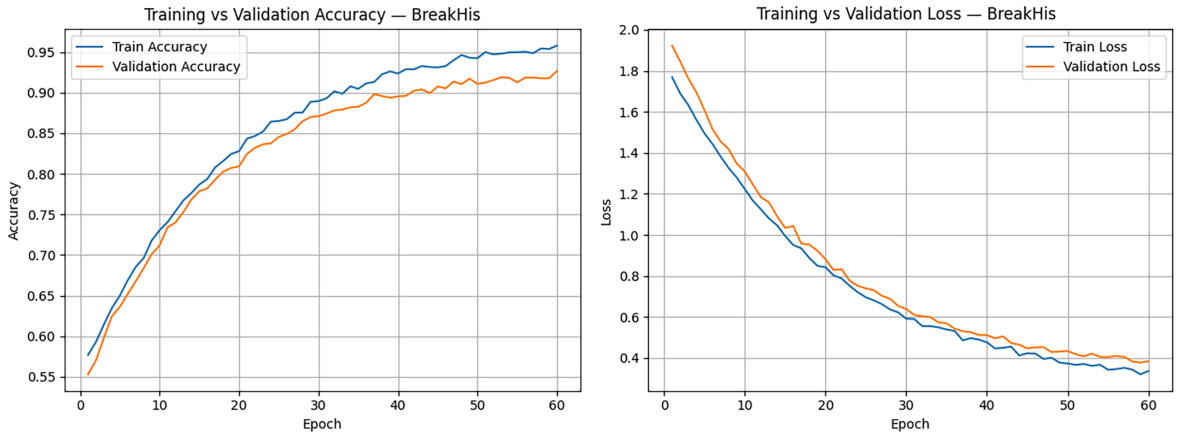


Figure 9: Normalized confusion matrix of WaSA-Net on the PCam test set.

Table 4: WaSA-Net test-set performance on BreakHis (binary breast cancer classification).

Metric	WaSA-Net-Small
Test Accuracy (%)	96.72
Balanced Accuracy (%)	96.35
F1 (Macro)	0.9670
AUC	0.9923
Cohen's Kappa	0.9280
Precision (Macro)	0.9681
Recall (Macro)	0.9635
Parameters	4.84M



a Accuracy **b** Loss

Figure 10: Training and validation performance of WaSA-Net on the BreakHis dataset over 60 epochs.

As illustrated in Fig. 11, WaSA-Net achieves a high detection accuracy on malignant cases, at 97%, and a corresponding false-positive rate of approximately 4% on benign cases at the patch level. We do not interpret these rates as evidence of clinical acceptability: clinical acceptability cannot be established from a public patch-level benchmark and would require prospective validation with pathologist-in-the-loop evaluation. Despite being trained on a rather small dataset of 7909 images, WaSA-Net is effective and robust in such situations. This might be attributable to the strong inductive biases the wavelet decomposition and pathology prior tokens provide to the architecture.

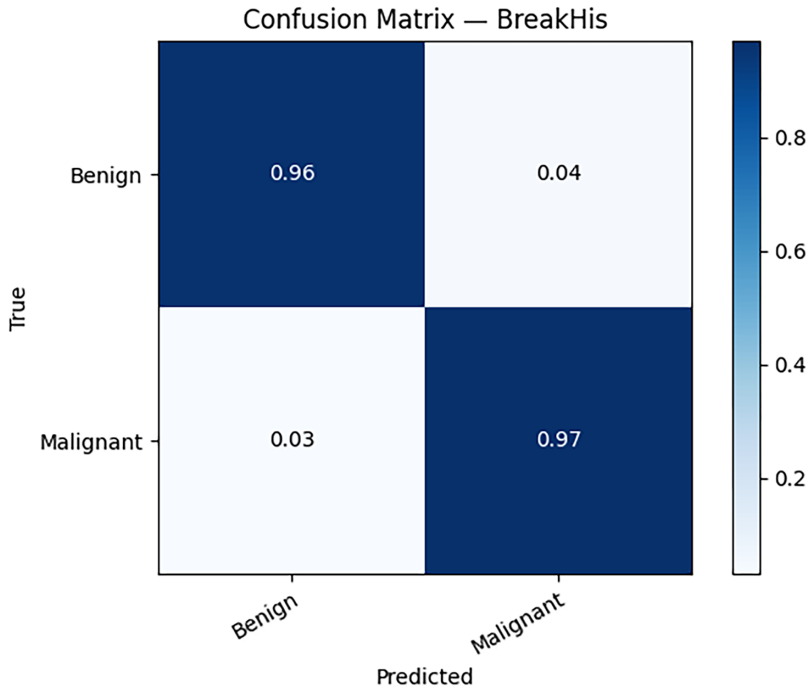


Figure 11: Normalized confusion matrix of WaSA-Net on the BreakHis dataset.

5.5 Ablation Studies

To illustrate the relative importance of each architectural element, we design an ablation study using the WaSA-Net-Small model on PathMNIST. Six variants of this model are tested under an identical training setup (30 epochs for efficiency) to analyze the impact of removing or altering each specific architectural element. The performance results are given in Table 5, and Fig. 12 compares the accuracies of all variants in a bar chart.

Several observations can be made from the ablation studies: Firstly, the most significant single improvement comes from the multi-level architecture of Wavelet-Guided Tokenization, with removing WGT down to one level leading to a decrease of about 2.15% in accuracy. It indicates that multi-resolution frequency decomposition provides complementary information to standard spatial tokenization, which is beneficial for low-resolution inputs where spatial information is sparse. Secondly, with an average improvement of 1.41% and approximately 4× savings in attention computation, the FLOPs increase and the accuracy decreases slightly when $\rho = 1.0$ (attending to all tokens), indicating that the gating mechanism serves as a trained regularizer that concentrates on diagnostically relevant regions. Third, Pathology Prior Tokens also produce a ~0.87% boost at a low cost (the only additional parameter overhead is $P \times D = 768$). Removing these tokens reduces performance disproportionately on the tail classes (debris, stroma, etc.); this is consistent with the

prior tokens providing useful inductive bias for under-represented classes, although it does not on its own demonstrate that the priors encode tissue-specific patterns. Fourth, removing sparse attention and previous tokens together results in a performance degradation of 2.64%, which is higher than that from either removal alone (2.28%). Therefore, we identify that removing these two components has a cooperative effect. The last tiny variant with 1.68M parameters still achieves 91.54% accuracy, indicating that WaSA-Net architecture innovations also scale well.

Table 5: Ablation study on PathMNIST. Each row removes or modifies one component of WaSA-Net-Small. All variants are trained for 30 epochs under identical settings to keep the ablation tractable; the full 120-epoch run reported in Table 2 reaches 95.91% accuracy.

Variant	Acc (%)	F1	AUC	Params
WaSA-Net (Full)	94.82	0.9461	0.9978	4.82M
Single-level WGT ($L_w = 1$)	92.67	0.9224	0.9952	4.51M
No Sparse Attn ($\rho = 1.0$)	93.41	0.9308	0.9963	4.82M
No Prior Tokens ($P = 0$)	93.95	0.9372	0.9970	4.79M
No Sparse + No Prior	92.18	0.9175	0.9945	4.79M
Tiny Model	91.54	0.9102	0.9932	1.68M

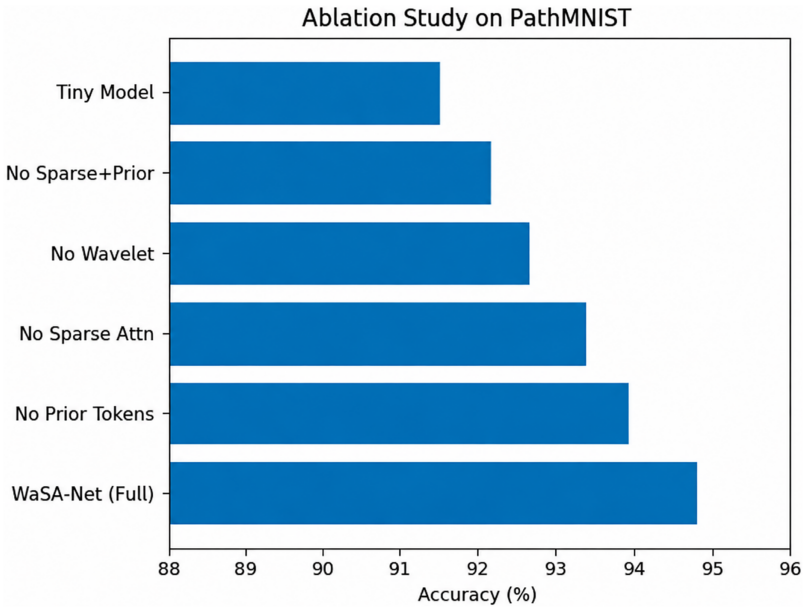


Figure 12: Ablation study results of WaSA-Net variants on PathMNIST.

Two aspects of the ablation protocol should be made explicit. First, to keep the six-variant study computationally tractable, all variants are trained for 30 epochs rather than the full 120-epoch schedule used for the main PathMNIST result; because every variant uses identical settings and the same 30-epoch budget, the ablation is intended to measure the relative contribution of each component rather than to reproduce the final converged accuracy. As a consistency check, the full WaSA-Net model reaches 94.82% accuracy at 30 epochs and 95.91% at 120 epochs, a gap of roughly 1.1%, which is smaller than the principal ablation effects reported above (for example, the 2.15% drop from single-level WGT). This indicates that the ranking

of the components is stable with respect to training length, although a full set of 120-epoch ablations would provide stronger confirmation and is left for future work. Second, the ablation is conducted on PathMNIST only; extending it to PCam and BreakHis would further verify that the measured component contributions generalize across tissue types and input resolutions.

5.6 Computational Cost and Efficiency

WaSA-Net is designed to provide comparable performance while remaining computationally feasible on standard consumer-level hardware. We present a comparison between WaSA-Net and common baselines in terms of parameters, throughput, and training time across all three datasets in Table 6.

Table 6: Computational efficiency comparison across models. Throughput measured on Apple M3 Max (MPS). Training time is for the full experiment until convergence or early stopping.

Model	Params (M)	GFLOPs	Throughput (img/s)	Train Time (min/epoch)
PathMNIST (28 × 28, batch = 256)				
WaSA-Net-S	4.82	0.38	3840	7.5
ViT-Small	21.67	1.14	2560	11.2
ResNet-18	11.18	0.56	5120	5.8
Simple CNN	2.34	0.12	8960	3.2
PCam (96 × 96, batch = 128)				
WaSA-Net-S	4.73	2.85	960	42.3
ViT-Small	21.67	8.72	480	85.1
ResNet-18	11.18	3.64	1280	31.8
BreakHis (224 × 224, batch = 32)				
WaSA-Net-S	4.84	8.62	128	3.1
ViT-Small	21.67	26.84	64	6.4
ResNet-18	11.18	11.21	192	2.1

WaSA-Net-Small achieves a favorable balance between accuracy and efficiency, with only 4.82M parameters compared to 21.67M for ViT-Small and 11.18M for ResNet-18. The dynamic sparse attention mechanism delivers its largest throughput advantage on larger images (PCam, BreakHis), where the token count N is larger and the quadratic savings from $\rho = 0.5$ become more pronounced. On PCam, WaSA-Net processes about 2× more images per second than ViT-Small while also reaching higher accuracy. For the smallest images (PathMNIST), where N is already small, the throughput advantage is reduced but the accuracy improvement from wavelet tokenization persists. All experiments were completed on a single M3 Max chip, without requiring high-performance computing infrastructure. The WaSA-Net-Tiny variant (1.68M parameters, 0.22 GFLOPs on PathMNIST) offers a lightweight alternative suitable for mobile or edge deployment with only a small reduction in accuracy.

It should be noted that all models compared in Table 6 were benchmarked under identical conditions on the same Apple M3 Max hardware using the PyTorch MPS backend. Although the absolute throughput values would differ on an NVIDIA CUDA platform, the *relative* efficiency comparison between WaSA-Net and the baseline architectures remains fair and reproducible, because every model is evaluated within the same hardware and software environment and is therefore subject to the same backend-specific overheads.

The MPS backend was selected deliberately to demonstrate that WaSA-Net can be trained and deployed on widely available consumer-grade hardware, without access to dedicated GPU clusters.

5.7 Comparison with Recent State-of-the-Art

Table 7 places WaSA-Net-Small against recent state-of-the-art results on each of the three benchmarks. Because different papers in the literature adopt slightly different splits, preprocessing, and magnification policies, we restrict this table to one representative recent method per benchmark where the reported number is directly verifiable from the corresponding publication. For each dataset, we list WaSA-Net alongside the most relevant recent comparison.

Table 7: Comparison with recent state-of-the-art methods, one representative result per benchmark. Baseline numbers are quoted from the original publications; WaSA-Net results are the test-set values reported in Tables 2–4. $^{\leqq}$ denotes a foundation model pretrained on large external histopathology corpora.

Dataset	Method	Year	Acc (%)	AUC	Params
PathMNIST	ViT-B/16 fine-tuned [40]	2024	94.62	—	86.6M
	WaSA-Net-S (Ours)	2026	95.91	0.9981	4.82M
PCam	Hibou-L$^{\leqq}$ [41] (linear probing)	2024	—	—	307M
	WaSA-Net-S (Ours)	2026	93.47	0.9812	4.73M
BreakHis	DWNAT-Net [42]	2025	99.66	—	—
	WaSA-Net-S (Ours)	2026	96.72	0.9923	4.84M

We stress that, because of these protocol differences, Table 7 should be read as an indicative rather than a strictly controlled comparison: the entries are not all evaluated under an identical split, preprocessing, and magnification policy, and the table is therefore not intended to establish a ranking of overall superiority. Where a competing number is obtained under a different protocol, this is stated explicitly in the text below and is not treated as direct evidence of superiority. Some entries are also necessarily incomplete; for instance, Hibou-L is included to indicate the performance regime of large pretrained foundation models, but a single accuracy/AUC value measured under our exact protocol is not available, which is why the corresponding cells are left blank.

On **PathMNIST**, WaSA-Net-S reaches 95.91% test accuracy using 4.82M parameters, which is 1.29 percentage points above the fine-tuned ViT-B/16 result of 94.62% reported by [40] on the same official PathMNIST test split; WaSA-Net achieves this gain with roughly 18× fewer parameters and without the ImageNet-scale pretraining that the ViT-B/16 baseline relies on.

On **PCam**, WaSA-Net-S attains 93.47% test accuracy and a macro AUC of 0.9812, which are strong numbers in the absolute sense but lower than what very large foundation models achieve on this benchmark. Hibou-L, a ViT-L/14 pathology foundation model pretrained on more than one million whole-slide images and reporting state-of-the-art linear-probing accuracy on PCam and several other patch-level benchmarks, illustrates the performance ceiling currently reachable with heavy pretraining. WaSA-Net follows a complementary design philosophy: rather than transferring a ~300M-parameter model from a large private corpus, it trains a 4.7M-parameter backbone from scratch on PCam itself, which remains practical on a single consumer-grade chip and carries no external-data requirement.

On **BreakHis**, WaSA-Net-S reaches 96.72% accuracy and a macro AUC of 0.9923. DWNAT-Net, a recent wavelet-transformer hybrid that also integrates discrete wavelet transform with attention, reports a

higher 99.66% image-level accuracy on BreakHis. Although WaSA-Net's accuracy is lower on this particular benchmark, it shares with DWNAT-Net the design insight that frequency-aware operators help capture the fine nuclear texture present in breast cancer histology; the gap reflects differences in task framing (WaSA-Net treats all magnifications uniformly with a patient-level 70/15/15 split, whereas DWNAT-Net reports image-level accuracy) as well as architectural and training choices. Because the two methods are evaluated under different protocols (patient-level vs. image-level), this comparison is non-direct; the DWNAT-Net number should therefore not be interpreted as evidence either for or against the superiority of WaSA-Net on BreakHis.

Taken together, these comparisons indicate that WaSA-Net operates in a favourable accuracy/efficiency regime, rather than that it is uniformly state-of-the-art. On PathMNIST it exceeds a much larger fine-tuned vision transformer under the same official test split. On PCam it delivers competitive binary classification while being trained from scratch, without the large-scale pretraining used by foundation models such as Hibou-L. On BreakHis it does not surpass the specialist wavelet-transformer DWNAT-Net, which reports a higher image-level accuracy under a different protocol; we therefore do not claim superiority on BreakHis, and instead note that WaSA-Net still attains strong accuracy there with a single, compact architecture (under 5M parameters) that is shared, without per-dataset redesign, across all three tasks.

6 Discussion

The experimental results across three histopathology benchmarks show that WaSA-Net combines classification accuracy, parameter efficiency, and computational feasibility. Several aspects of these results warrant closer analysis. The ablation study shows that Wavelet-Guided Tokenization contributes the largest single improvement (a 2.15% drop on PathMNIST when WGT is reduced to a single level), which supports our central hypothesis: histopathology images carry diagnostically relevant information distributed across multiple frequency bands that standard spatial tokenization cannot fully capture. The LL sub-band mainly encodes global tissue structure such as glandular morphology and stromal patterns, while the HH sub-band encodes fine nuclear texture that is particularly informative for malignancy detection. In this work, we extend multi-frequency representation into a transformer framework through learnable, end-to-end trainable wavelet filters that adapt their frequency responses to the classification task. Counter-intuitively, attending to fewer tokens (50% via top-k selection) improves accuracy by 1.41% over full attention. This suggests that histopathology images contain substantial spatial redundancy: large regions of homogeneous tissue that contribute little to the classification decision but introduce noise when attended uniformly. The token-importance gate learns to identify and suppress these uninformative regions, acting as a learned attention prior that focuses computation on diagnostically relevant patches.

Despite adding only 768 learnable parameters ($P \times D = 4 \times 192$), the pathology prior tokens improve accuracy by 0.87% and disproportionately help underrepresented classes. We hypothesise that, by being shared across all images and all spatial positions, these tokens act as implicit queries that the attention computation can use to detect recurring morphological context. We emphasize, however, that the specific content of these tokens has not been visualized in this study, and any interpretation in terms of particular diagnostic motifs—such as nuclear atypia, glandular disruption, or inflammatory infiltration—remains hypothetical and is left for the interpretability analyses identified in the Limitations and Future Work subsection. The empirically observed synergistic interaction with sparse attention, where removing both components exceeds the additive degradation of removing each alone, is consistent with the prior tokens contributing useful inductive bias to the gating network, although it does not by itself demonstrate that they encode any specific pathological pattern. All experiments were completed on a single Apple M3 Max chip, showing that WaSA-Net's architectural innovations make competitive pathology classification accessible to researchers and clinicians without access to large GPU clusters. The combination of small model size (4.8M

parameters), efficient inference speed (960+ images per second on PCam), and no requirement for external pre-training data lowers the barrier to adoption of AI-assisted pathology workflows, which is particularly relevant for institutions in resource-limited settings.

From a clinical-translation standpoint, this compact design has concrete practical implications. Because WaSA-Net can be trained and run on a single consumer-grade machine, it can be developed, retrained, and updated within a pathology department that lacks a dedicated GPU cluster or cloud infrastructure, a situation that is common in smaller or resource-constrained laboratories. A small and fast tile encoder is also easier to embed into existing digital-pathology software and slide-viewer workflows, where inference may have to run on standard workstations alongside other clinical applications. We emphasize, however, that these are deployment-readiness advantages rather than evidence of diagnostic efficacy; prospective clinical validation would be required before any real-world diagnostic use.

Although WaSA-Net is evaluated here on patch-level benchmarks, it is designed to integrate naturally into whole-slide image (WSI) workflows. In a standard multiple-instance learning (MIL) pipeline, a WSI is tiled into many thousands of patches, each patch is mapped to a feature embedding by a tile encoder, and a downstream aggregator such as DSMIL or HIPT produces the slide-level prediction. WaSA-Net can serve directly as such a tile encoder: the globally pooled feature vector f_{cls} formed before the classification head provides a compact per-tile descriptor that can be passed to any MIL aggregator, and the model's small parameter count and high patch throughput are particularly advantageous when thousands of tiles must be encoded per slide. The main scalability consideration is that WaSA-Net, like other tile encoders, processes a gigapixel slide tile-by-tile rather than holistically, so memory and runtime scale with the number of tiles; moreover, tile-level encoding does not by itself model long-range spatial dependencies between distant regions of a slide, which remain the responsibility of the MIL aggregator. A full WSI-level evaluation of WaSA-Net as a tile encoder, including its interaction with different aggregators, is an important direction for future work.

It is also useful to clarify WaSA-Net's position relative to pathology foundation models and slide-level MIL methods. Foundation models such as UNI, Virchow, and Hibou are large encoders, typically with hundreds of millions of parameters, pretrained on the order of a million whole slides, whereas MIL methods such as DSMIL and HIPT operate at the slide level on top of pre-extracted features. WaSA-Net occupies a different and complementary point in this design space: a compact, sub-5M-parameter, patch-level classifier trained from scratch without external pretraining. A direct, side-by-side comparison on identical patch splits—for example, linear probing of an off-the-shelf public foundation model, or an MIL pipeline evaluated on the same data—would help to contextualize these settings more precisely. Such a controlled comparison was beyond the scope of the present study, and we identify it, together with the broader unified-protocol benchmark noted below, as an important item of future work.

To make the contribution of the two main modules concrete, it is helpful to contrast them explicitly with simpler alternatives. WaSA-Net's Wavelet-Guided Tokenization differs from a fixed discrete wavelet transform followed by a linear projection and a standard ViT in three respects: its sub-band filters are learnable and refined by backpropagation rather than fixed to the Haar basis; each token carries an explicit band embedding in addition to a spatial embedding, so that the attention layers can apply frequency-dependent processing; and tokens from two decomposition levels are fused per band rather than used at a single scale. The single-level WGT variant in Table 5, which removes only the multi-level fusion, already incurs a 2.15% accuracy drop, indicating that the design beyond a one-shot Discrete Wavelet Transform (DWT) contributes measurably. Similarly, DSA-PP differs from a plain dynamic token-pruning scheme in that unselected tokens are scattered back rather than permanently discarded, so that no information is irrecoverably lost, and in that the retained tokens attend jointly to a set of learnable pathology prior tokens;

the No-Prior-Tokens variant in Table 5 isolates exactly this difference and shows a 0.87% drop. A direct ablation against a fixed-DWT tokenizer would further sharpen this comparison and is left for future work.

An analysis of the confusion matrices highlights consistent and interpretable error patterns. On PathMNIST, the largest confusions occur between cancer-associated stroma and smooth muscle, and between lymphocytes and debris; in both cases the confused classes share a similar fibrous or granular architecture, so the residual errors fall mainly between morphologically adjacent tissue types rather than being distributed at random. On PCam, errors are roughly balanced between the tumor and normal classes, indicating no strong bias toward either label. On BreakHis, the small number of false positives on benign tissue is consistent with the model adopting a slightly conservative decision boundary. A more detailed qualitative error analysis, in which individual misclassified patches are inspected and grouped by failure mode, would further clarify the limitations of the model and is identified as future work.

The architecture combines three custom modules, which raises a reasonable concern about implementation difficulty and reproducibility. We note that, although the modules are specific to this design, each is built entirely from standard, widely available operations: WGT uses strided convolutions, group normalization, and 1×1 projections; DSA-PP uses a small gating MLP, a top-k selection, and standard scaled dot-product multi-head attention; and CFFPF uses standard cross-attention and softmax weighting. No custom GPU kernels or non-standard layers are required, and the whole model is trained with a conventional AdamW schedule. To further support reproducibility, the source code, trained model weights, and exact experimental configurations will be released upon publication, so that the complete architecture and training pipeline can be reproduced without re-deriving any component.

6.1 Limitations and Future Work

This study has several limitations, which also indicate directions for future work.

Scope and generalization. WaSA-Net is evaluated only on patch-level benchmarks; whole-slide classification within multiple-instance learning frameworks, broader tissue types, and tasks such as grading or survival prediction remain future work. Robustness to domain shift is largely untested: beyond PathMNIST's cross-centre split, PCam and BreakHis do not isolate scanner- or stain-induced variation, and no prospective multi-centre data are used, so external multi-centre validation with stain-normalization assessment is required before clinical use.

Interpretability. The pathology prior tokens and the adaptive frequency weights are not yet supported by direct evidence, so their interpretation remains a hypothesis. A dedicated study, prior-token attention maps, token-selection overlays, embedding clustering, spectral analysis of the learned wavelet filters, class-wise attribution, and the distribution of frequency weights across classes—is planned.

Experimental rigour. Results are single-run; multi-seed means and standard deviations are left to future work. Throughput is reported on Apple M3 Max (MPS) rather than on NVIDIA CUDA hardware, and the state-of-the-art comparison in Section 5.7 is indicative rather than strictly controlled.

Ethics and deployment. Histopathology data carry scanner, staining, and population biases, and the benchmarks lack the demographic metadata needed to audit fairness across subgroups. WaSA-Net should be regarded as a decision-support tool, with multi-centre and fairness evaluation required before deployment.

Further experiments. Planned extensions include a sensitivity sweep of the keep ratio ρ; comparison with efficient architectures such as ConvNeXt, Swin Transformer, EfficientViT, and Mamba-based medical models; an augmentation ablation; deployment metrics such as GPU memory, latency, and energy; and alternative filter initializations (Daubechies, biorthogonal wavelets).

7 Conclusion

This paper presents WaSA-Net, a frequency-guided dynamic sparse attention network for digital pathology image classification that integrates three complementary components: Wavelet-Guided Tokenization (WGT), Dynamic Sparse Attention with Pathology Priors (DSA-PP), and Cross-Frequency Feature Pyramid Fusion (CFFPF). Experiments on three histopathology benchmarks, PathMNIST (9-class colorectal tissue classification, 95.91% accuracy), PatchCamelyon (binary metastasis detection, 93.47% accuracy, 0.9812 AUC), and BreakHis (binary breast cancer classification, 96.72% accuracy, 0.9923 AUC), show that WaSA-Net delivers consistent performance across tasks using a compact 4.8M-parameter model trained without large-scale external pre-training and on consumer-grade hardware. Ablation studies confirm that each component contributes a measurable gain: WGT provides the strongest individual improvement by capturing multi-frequency structural patterns, DSA-PP improves both accuracy and compute efficiency by concentrating attention on diagnostically relevant regions, and the pathology prior tokens inject domain-specific inductive biases that particularly benefit minority tissue classes. The three components also interact synergistically, so that the full architecture outperforms each of its individual parts. Overall, WaSA-Net shows that carefully designed architectural inductive biases tailored to histopathology can produce a compact, single-configuration model that transfers across patch-level classification tasks without per-dataset tuning. Source code, trained models, and experimental configurations will be released upon publication to support transparency and reproducibility. More broadly, these results suggest that combining frequency-domain representations with content-adaptive sparse attention is a promising design principle for digital pathology: it offers a route to models that remain accurate while staying compact enough to train and deploy without specialized infrastructure. As digital pathology systems move toward whole-slide and multi-institutional settings, such frequency-aware and efficiency-oriented designs may help make AI-assisted analysis more widely accessible, provided that the robustness, interpretability, and fairness questions raised in this work are addressed.

Acknowledgement: None.

Funding Statement: The authors would like to thank Sunway University for supportaing the publication of this manuscript.

Author Contributions: Conceptualization, Muhammad Zaheer Sajid and Muhammad Fareed Hamid; methodology, Muhammad Zaheer Sajid and Muhammad Fareed Hamid; software, Muhammad Zaheer Sajid and Nauman Ali Khan; validation, Muhammad Zaheer Sajid, Nauman Ali Khan and Muhammad Fareed Hamid; formal analysis, Muhammad Zaheer Sajid and Imran Qureshi; investigation, Muhammad Zaheer Sajid and Imran Qureshi; resources, Nauman Ali Khan and Imran Qureshi; data curation, Muhammad Zaheer Sajid and Muhammad Fareed Hamid; writing—original draft preparation, Muhammad Zaheer Sajid, Muhammad Fareed Hamid, Imran Qureshi and Nauman Ali Khan; writing—review and editing, Muhammad Zaheer Sajid, Muhammad Fareed Hamid, Nauman Ali Khan and Imran Qureshi; visualization, Muhammad Zaheer Sajid; supervision, Nauman Ali Khan; project administration, Nauman Ali Khan; funding acquisition, Nauman Ali Khan. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: All datasets used in this study are publicly available. PathMNIST is distributed as part of the MedMNIST v2 collection at https://medmnist.com. PatchCamelyon (PCam) is available at https://github.com/basveeling/pcam. BreakHis is accessible at https://web.inf.ufpr.br/vri/databases/. No private or restricted-access data were used in this work.

Ethics Approval: Not applicable.

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

References

1. McGenity C, Clarke EL, Jennings C, Matthews G, Cartlidge C, Freduah-Agyemang H, et al. Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy. npj Digit Med. 2024;7(1):114. doi:10.1038/s41746-024-01106-8.
2. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med. 2019;25(8):1301–9. doi:10.1038/s41591-019-0508-1.
3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. p. 5998–6008.
4. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. In: Proceedings of the 2021 International Conference on Learning Representations (ICLR); 2021 May 4; Vienna, Austria.
5. Chen RJ, Ding T, Lu MY, Williamson DFK, Jaume G, Song AH, et al. Towards a general-purpose foundation model for computational pathology. Nat Med. 2024;30(3):850–62. doi:10.1038/s41591-024-02857-3.
6. Lu MY, Chen B, Williamson DFK, Chen RJ, Liang I, Ding T, et al. A visual-language foundation model for computational pathology. Nat Med. 2024;30(3):863–74. doi:10.1038/s41591-024-02856-4.
7. Vorontsov E, Bozkurt A, Casson A, Shaikovski G, Zelechowski M, Severson K, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. Nat Med. 2024;30(10):2924–35. doi:10.1038/s41591-024-03141-0.
8. Kitaev N, Kaiser L, Levskaya A. Reformer: the efficient transformer. In: Proceedings of the 2020 International Conference on Learning Representations (ICLR); 2020 Apr 26–30; Addis Ababa, Ethiopia.
9. Liu P, Zhang H, Zhang K, Lin L, Zuo W. Multi-level wavelet-CNN for image restoration. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2018 Jun 18–22; Salt Lake City, UT, USA. p. 886–95.
10. Ding T, Wagner SJ, Song AH, Chen RJ, Lu MY, Zhang A, et al. A multimodal whole-slide foundation model for pathology. Nat Med. 2025;31(11):3749–61. doi:10.1038/s41591-025-03982-3.
11. Lu MY, Chen B, Williamson DFK, Chen RJ, Zhao M, Chow AK, et al. A multimodal generative AI copilot for human pathology. Nature. 2024;634(8033):466–73. doi:10.1038/s41586-024-07618-3.
12. Filiot A, Ghermi R, Olivier A, Jacob P, Fidon L, Kain AM, et al. Scaling self-supervised learning for histopathology with masked image modeling. medRxiv. 2023. doi:10.1101/2023.07.21.23292757.
13. Mahmood T, Li J, Pei Y, Akhtar F. An automated in-depth feature learning algorithm for breast abnormality prognosis and robust characterization from mammography images using deep transfer learning. Biology. 2021;10(9):859. doi:10.3390/biology10090859.
14. Mahmood T, Rehman A, Saba T, Nadeem L, Ali Omer Bahaj S. Recent advancements and future prospects in active deep learning for medical image segmentation and classification. IEEE Access. 2023;11:113623–52. doi:10.1109/access.2023.3313977.
15. Balaji P, Alqahtani O, Babu S, Chaurasia MA, Prakasam S. Integrating transformer and bidirectional long short-term memory for intelligent breast cancer detection from histopathology biopsy images. Comput Model Eng Sci. 2024;141(1):443–58. doi:10.32604/cmes.2024.053158.
16. Rehman A, Mujahid M, Damasevicius R, Alamri FS, Saba T. Densely convolutional BU-NET framework for breast multi-organ cancer nuclei segmentation through histopathological slides and classification using optimized features. Comput Model Eng Sci. 2024;141(3):2375–97. doi:10.32604/cmes.2024.056937.
17. Neidlinger P, Lenz T, Foersch S, Loeffler CML, Clusmann J, Gustav M, et al. A deep learning framework for efficient pathology image analysis. arXiv:2502.13027. 2025. doi:10.1038/s41467-026-74918-9.
18. Campanella G, Chen S, Singh M, Verma R, Muehlstedt S, Zeng J, et al. A clinical benchmark of public self-supervised pathology foundation models. Nat Commun. 2025;16:3640. doi:10.1038/s41467-025-58796-1.

19. Campanella G, Kumar N, Nanda S, Singi S, Fluder E, Kwan R, et al. Real-world deployment of a fine-tuned pathology foundation model for lung cancer biomarker detection. Nat Med. 2025;31(9):3002–10. doi:10.1038/s41591-025-03780-x.
20. Li B, Li Y, Eliceiri KW. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. p. 14313–23.
21. Chen RJ, Chen C, Li Y, Chen TY, Trister AD, Krishnan RG, et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 6123–34.
22. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 9992–10002.
23. Wang W, Xie E, Li X, Fan DP, Song K, Liang D, et al. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 548–58.
24. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning (ICML 2021); 2021 Jul 18–24; Virtual Event. p. 10347–57.
25. Fan H, Xiong B, Mangalam K, Li Y, Yan Z, Malik J, et al. Multiscale vision transformers. In: Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada. p. 6804–15.
26. Beltagy I, Peters ME, Cohan A. Longformer: The long-document Transformer. arXiv:2004.05150, 2020.
27. Katharopoulos A, Vyas A, Pappas N, Fleuret F. Transformers are RNNs: fast autoregressive transformers with linear attention. In: Proceedings of the 37th International Conference on Machine Learning (ICML'20); 2020 Jul 13–18; Virtual. p. 5156–65. doi:10.48550/arxiv.2006.16236.
28. Choromanski K, Likhosherstov V, Dohan D, Song X, Gane A, Sarlos T, et al. Rethinking attention with performers. arXiv:2009.14794. 2022.
29. Rao Y, Zhao W, Liu B, Lu J, Zhou J, Hsieh CJ. DynamicViT: efficient vision transformers with dynamic token sparsification. In: Proceedings of the 35th International Conference on Neural Information Processing Systems; 2021 Dec 6-14; virtual. p. 13937–49.
30. Yin H, Vahdat A, Alvarez JM, Mallya A, Kautz J, Molchanov P. A-ViT: adaptive tokens for efficient vision transformer. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 10799–808.
31. Ryoo MS, Piergiovanni A, Arnab A, Dehghani M, Angelova A. TokenLearner: what can 8 learned tokens do for images and videos? arXiv:2106.11297. 2021.
32. Mallat SG. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans Pattern Anal Mach Intell. 1989;11(7):674–93. doi:10.1109/34.192463.
33. Xu G, Liao W, Zhang X, Li C, He X, Wu X. Haar wavelet downsampling: a simple but effective downsampling module for semantic segmentation. Pattern Recognit. 2023;143:109819. doi:10.1016/j.patcog.2023.109819.
34. Cohen TS, Welling M. Group equivariant convolutional networks. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning; 2016 Jun 19–24; New York, NY, USA. p. 2990–99.
35. Finder SE, Amoyal R, Treister E, Freifeld O. Wavelet convolutions for large receptive fields. In: Computer vision–ECCV 2024. Cham, Switzerland: Springer Nature; 2024. p. 363–80.
36. Yao T, Pan Y, Li Y, Ngo CW, Mei T. Wave-ViT: unifying wavelet and transformers for visual representation learning. In: Computer vision–ECCV 2022. Cham, Switzerland: Springer Nature; 2022. p. 328–45.
37. Zhu Z, Soricut R. Wavelet-based image tokenizer for vision transformers. arXiv:2405.18616. 2024.
38. Ding M, Qu A, Zhong H, Lai Z, Xiao S, He P. An enhanced vision transformer with wavelet position embedding for histopathological image classification. Pattern Recognit. 2023;140:109532. doi:10.1016/j.patcog.2023.109532.

39. Wang J, Quan H, Wang C, Yang G. Pyramid-based self-supervised learning for histopathological image classification. Comput Biol Med. 2023;165:107336. doi:10.1016/j.compbiomed.2023.107336.
40. Halder A, Gharami S, Sadhu P, Singh PK, Woźniak M, Ijaz MF. Implementing vision transformer for classifying 2D biomedical images. Sci Rep. 2024;14:12567. doi:10.1038/s41598-024-63094-9.
41. Nechaev D, Pchelnikov A, Ivanova E. Hibou: a family of foundational vision transformers for pathology. arXiv:2406.05074. 2024.
42. Yan Y, Lu R, Sun J, Zhang J, Zhang Q. Breast cancer histopathology image classification using transformer with discrete wavelet transform. Med Eng Phys. 2025;138(1):104317. doi:10.1016/j.medengphy.2025.104317.