Research Article

AND DATA SCIENCES
BULLETIN OF COMPUTER

Multi-Scale FourierMIL for Hierarchical Frequency-Domain Multiple Instance Learning in Whole-Slide Image Classification

Jun Li, Xiao Bai and Jin Zheng

The school of Automation and Electrical Engineering, University of Science and Technology Beijing, Beijing, China

Abstract

Whole-slide images (WSIs) in computational pathology contain morphological patterns across multiple spatial scales, yet most multiple instance learning (MIL) methods operate on a single resolution. Recent work shows that frequency-domain token mixing via the Fourier transform can improve both accuracy and efficiency over self-attention for WSI classification, but existing frequency-based MIL models still reason at only one scale. In this paper, we propose *Multi-Scale FourierMIL* (MS-FourierMIL), a hierarchical frequency-domain MIL framework that integrates patch tokens from multiple spatial resolutions. For each scale, we extract patch embeddings with a frozen feature extractor and apply a scale-specific Fourier token mixer based on a learnable all-pass filter, then perform cross-scale frequency fusion over pooled scale representations to capture interactions between coarse tissue architecture and fine cellular detail. A simple adaptive padding scheme enables stable FFT-based mixing for variable bag sizes while avoiding distributional shifts from naïve zero padding, and a final class token conditioned on all scales produces slide-level predictions. On standard WSI classification benchmarks, MS-FourierMIL improves over a single-scale FourierMIL reference and strong attention- and transformer-based MIL baselines, while keeping model size and inference time in the same practical regime as common MIL heads. Qualitative multi-scale attribution maps illustrate how coarse-scale evidence supports global tissue context and higher-magnification evidence supports localized regions, aligning with pathologists' multi-scale reasoning and motivating frequency-domain multi-scale MIL as an efficient strategy for WSI analysis.

Keywords: whole-slide images, computational pathology, multiple instance learning, frequency-domain token mixing, multi-scale modeling

1. Introduction

Digital pathology has enabled the routine acquisition of gigapixel whole-slide images (WSIs) for diagnosis, prognosis, and biomarker discovery in oncology and other diseases [1–3]. A core challenge in computational pathology is to design models that can handle both the extreme resolution of WSIs and the multi-scale nature of histological patterns [1, 2, 4]. Tumor infiltration, stromal reactions, and micro-environmental cues appear at different spatial scales, and human pathologists routinely zoom in and out to interpret a case. Machine learning models that operate at a single resolution risk missing small lesions or failing to exploit larger-scale architectural cues.

Corresponding author: Jun Li (junli1987@proton.me)

Multiple instance learning (MIL) provides a natural paradigm for WSI-level prediction under weak supervision, where only slide-level labels are available [5, 6]. In MIL, a WSI is represented as a bag of instances (patches), and a model aggregates patch features into a slide representation. Early approaches relied on simple pooling operations [7], while more recent methods use attention mechanisms [8, 9], transformers [10], or graph neural networks [11–13] to learn more expressive bag-level representations. However, most existing MIL models operate on patch features extracted at a single magnification, treating the WSI as a flat set of tokens.

In parallel, token mixing in the frequency domain has emerged as an attractive alternative to self-attention for image and sequence modeling [14, 15]. By applying the discrete Fourier transform (DFT) along the token dimension, one can capture long-range interactions with lower overhead and implement learned filters over frequency components. Recent work has demonstrated that replacing attention with Fourier-based token mixing in MIL can yield strong performance on WSI classification tasks, while avoiding quadratic complexity with respect to the number of instances [16].

Despite these advances, current frequency-domain MIL models still ignore the inherent multi-scale structure of WSIs. They aggregate patch tokens from a single resolution and thus must implicitly encode all relevant scales in one representation [16]. This can be problematic in scenarios where lesions are extremely small relative to the slide, or where subtle cellular details must be interpreted in the context of global tissue architecture. A model that explicitly represents and fuses information across multiple spatial scales could be more robust and interpretable.

In this paper, we propose *Multi-Scale FourierMIL* (MS-FourierMIL), a hierarchical MIL architecture that brings multi-scale modeling into the frequency domain. The central idea is simple: instead of a single bag of patch tokens, we construct several bags at different spatial resolutions. For each scale, we apply a dedicated Fourier-based token mixer, realized as an all-pass frequency filter with learnable phase parameters. We then fuse the resulting scale-level representations by performing another Fourier transform over the concatenated scale tokens, enabling cross-scale interactions in the frequency domain. The final WSI representation is computed from a class token that aggregates information from all scales.

We introduce MS-FourierMIL, a multi-scale frequency-domain MIL framework for WSI classification that combines intra-scale Fourier token mixing with cross-scale frequency fusion. We propose a simple adaptive padding strategy that stabilizes Fourier transforms on variable-length bags while preserving the empirical distribution of patch tokens. We demonstrate, on benchmark WSI classification tasks, that explicitly modeling multiple scales in the frequency domain improves performance over single-scale FourierMIL and competitive baselines, without prohibitive increases in computational cost.

Beyond performance, MS-FourierMIL offers a conceptually appealing view of WSI analysis: each scale corresponds to a frequency-aware channel specializing in particular patterns, and cross-scale fusion allows the model to reconcile local detail with global context. This resonates with the multi-scale reasoning of human pathologists and suggests promising directions for future extensions to segmentation, grading, and survival prediction.

2. Related Work

2.1. Whole-Slide Image Analysis and MIL

Computational pathology methods for WSIs must confront very large image sizes and weak labels. Patch-based pipelines have become standard: the WSI is tiled into patches at a chosen magnification, low-quality patches are discarded, and a feature extractor (often a convolutional neural network or vision transformer) produces embeddings for the remaining patches [1–3, 7, 17]. MIL then serves as the bag-level classifier, mapping a set of patch embeddings to a slide-level prediction.

Early MIL approaches for WSIs employed simple pooling operators, such as mean or max pooling, over patch scores [7]. Attention-based MIL introduced trainable attention weights to softly select informative patches, improving both performance and interpretability [6, 8, 9, 21]. Transformer-based MIL models extend this idea by treating patch embeddings as tokens and applying self-attention layers for global context modeling [4, 10, 22]. Graph-based MIL approaches instead construct a graph over patches, encoding spatial adjacency or feature similarity, and apply graph neural networks to propagate information [11–13, 23].

Despite differences in architecture, these models typically operate at a single resolution, for example patches at 20× magnification. Some works have explored multi-resolution or multi-scale designs by training separate MIL models per magnification and combining their predictions, or by using hierarchical networks that first aggregate features within regions and then across regions [18, 20, 22]. However, most of these multi-scale extensions remain in the spatial domain and rely on attention, convolutional, or graph layers for fusion.

2.2. Frequency-Domain Token Mixing

Frequency-domain methods for deep learning have gained attention as efficient alternatives to self-attention and spatial convolutions. By transforming features into the frequency domain via the Fourier transform, models can implement global mixing with compact parameterizations. Early work on Fourier neural operators and spectral transformers demonstrated that learnable filters in the frequency domain can capture long-range dependencies with sub-quadratic complexity [14, 15]. More recent work has extended spectral or Fourier-based mixing to vision models and token mixers as a drop-in replacement for attention.

In the context of MIL for WSIs, Fourier-based token mixing has recently been applied as a substitute for self-attention. The core idea is to treat the sequence of patch embeddings as a 1D signal along the token dimension, apply the discrete Fourier transform, multiply by a learnable frequency filter, and invert the transform. When the filter has unit magnitude and learnable phase, the resulting operation implements energy-preserving global mixing (a unit-magnitude spectral modulation) analogous to an all-pass filter in signal processing [14]. Such models have shown strong slide-level performance and favorable efficiency compared to transformer-based MIL in WSI classification [16].

However, existing frequency-domain MIL architectures are single-scale: they operate on patch tokens from one resolution and apply frequency mixing only within that bag [16]. They do not explicitly encode the multi-scale structure of WSIs, nor do they model interactions between scales in frequency space.

2.3. Multi-Scale Modeling in Pathology

The importance of multi-scale information in histopathology is well recognized. Pathologists routinely switch between low magnifications, which reveal tissue organization and lesion extent, and high magnifications, which expose nuclear morphology and mitotic figures. Machine learning methods have attempted to emulate this behavior. Multi-resolution CNNs process patches from different magnifications and fuse their features [20]. Some MIL approaches extract patch embeddings at multiple scales and combine them via attention, concatenation, or pyramidal fusion [9, 18, 22]. Hierarchical and pyramid architectures have also been proposed, where features are aggregated first within local regions and then globally, including hierarchical vision transformers that natively exploit WSI pyramids [4].

While these approaches demonstrate the benefit of multi-scale representations, they typically remain in the spatial domain and rely on either convolutions, attention mechanisms, or graph-based fusion blocks. To our knowledge, no prior work has combined multi-scale WSI modeling with frequency-domain token mixing, nor explored cross-scale interactions directly in the frequency domain.

3. Method

3.1. Problem Formulation

Let a WSI be denoted by I. For a given set of spatial scales $\mathcal{S} = \{s_1, s_2, \ldots, s_K\}$ (e.g., magnifications or effective patch sizes), we extract a set of patches at each scale $s \in \mathcal{S}$ and pass them through a feature extractor f_θ (e.g., a CNN or vision transformer) to obtain patch embeddings. At scale s, this yields a bag

$$\mathcal{B}^{(s)} = \{\mathbf{x}_1^{(s)}, \ldots, \mathbf{x}_{N_s}^{(s)}\}, \quad \mathbf{x}_i^{(s)} \in \mathbb{R}^d,$$

where N_s is the number of patches at scale s. The WSI label is denoted by $y \in \{1, \ldots, C\}$ for a C-class classification problem.

Although a MIL bag is naturally treated as an unordered set, the Fourier mixer operates along an explicit token axis. Throughout this work we therefore use a deterministic ordering of instances within each bag, derived from the patch extraction grid (row-major order by patch coordinates), so that the sequence representation is reproducible and consistent across training and evaluation. This ordering provides a simple, fixed token axis for the DFT without introducing additional positional encodings.

Our goal is to learn a function

$$g : \bigcup_{s \in \mathcal{S}} \mathcal{B}^{(s)} \to \Delta^{C-1},$$

that maps the collection of multi-scale bags to a distribution over classes, where Δ^{C-1} is the probability simplex. We adopt the MIL setting in which only slide-level labels y are available during training; instance-level labels are unknown.

3.2. Single-Scale FourierMIL Recap

We briefly recall the core idea of Fourier-based token mixing for a single bag, as a baseline for our extension. Consider a bag $\mathcal{B} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ of d-dimensional token embeddings. We stack them

into a matrix $X \in \mathbb{R}^{N \times d}$, where the first dimension indexes tokens. A Fourier token mixer proceeds as follows.

First, we apply a linear projection to map X to a hidden dimension D:

$$Z = XW_{\text{in}}, \quad W_{\text{in}} \in \mathbb{R}^{d \times D}.$$

Next, we apply the discrete Fourier transform (DFT) along the token dimension:

$$\hat{Z} = \mathcal{F}(Z),$$

where $\hat{Z} \in \mathbb{C}^{N \times D}$ and $\mathcal{F}$ denotes the DFT applied independently to each feature channel across tokens.

We then multiply each frequency component by a learnable complex-valued filter $H \in \mathbb{C}^{N \times D}$:

$$\tilde{Z} = \hat{Z} \odot H,$$

where $\odot$ denotes element-wise multiplication. To ensure energy preservation and numerical stability, one can constrain H to be an *all-pass* filter, such that $|H_{k,d}| = 1$ for all frequencies k and channels d. This can be achieved by parameterizing $H_{k,d} = \exp(i\phi_{k,d})$ with learnable real-valued phases $\phi_{k,d}$.

In practice, bag lengths vary across slides (and across scales), so the filter must be defined in a length-agnostic way. We parameterize the phase response as a set of M learnable phase values over normalized frequency indices and resample (via interpolation) to the required FFT length at runtime; in our experiments we use $M = 64$ to balance expressiveness and parameter efficiency. Because the input tokens are real-valued, we additionally enforce the conjugate-symmetry constraint on the resampled spectrum so that the inverse transform yields real-valued outputs (equivalently, we use a real FFT implementation and keep the real part after inversion).

The mixed tokens are obtained by applying the inverse DFT:

$$Z' = \mathcal{F}^{-1}(\tilde{Z}),$$

and projecting back to the original embedding dimension:

$$X' = Z'W_{\text{out}}, \quad W_{\text{out}} \in \mathbb{R}^{D \times d}.$$

This operation mixes information globally across all tokens with complexity $O(ND \log N)$ due to the FFT, and can be composed with feedforward MLP blocks and skip connections to form a deep mixer.

3.3. Adaptive Padding for Variable-Length Bags

In practice, WSIs yield bags of widely varying lengths N_s, depending on tissue content and patch extraction strategy. Directly applying FFTs to these variable-length sequences can lead to numerical instabilities and inefficiencies, especially when FFT implementations are optimized for certain sizes (e.g., powers of two). We address this with a simple *adaptive padding* scheme.

For a given bag with length N, we choose a padded length $\tilde{N}$ as the smallest power of two greater than or equal to N:

$$\tilde{N} = 2^{\lceil \log_2 N \rceil}.$$

We then append $\tilde{N} - N$ extra tokens to X by sampling existing tokens from the same bag with replacement (optionally applying a small feature-space perturbation) so that the marginal feature

distribution is not dominated by artificial padding values. Let $\tilde{X} \in \mathbb{R}^{\tilde{N} \times d}$ denote the padded sequence. We apply the Fourier mixer to $\tilde{X}$ and discard the extra positions after inverse transformation, keeping only the first N tokens. The padded positions are used only during intermediate FFT mixing and are removed before any pooling or classification step. This yields a mixed bag of the original length, while benefiting from numerically stable and efficient FFTs.

In our multi-scale setting, adaptive padding is applied independently at each scale, with potentially different $\tilde{N}_s$.

3.4. Multi-Scale Patch Extraction

To construct multi-scale bags, we choose a set of spatial scales corresponding to different effective patch sizes. For example, we might use:

$$\mathcal{S} = \{s_{\text{coarse}},\ s_{\text{medium}},\ s_{\text{fine}}\},$$

where the coarse scale captures global tissue structure (e.g., low magnification or large patches), and the fine scale captures cellular details (e.g., high magnification or small patches).

For each scale s, we tile the WSI I with a non-overlapping or partially overlapping grid of patches $\{P_1^{(s)}, \ldots, P_{N_s}^{(s)}\}$, discarding background patches using a tissue mask if available. We record the spatial coordinates of each patch and order tokens consistently using the row-major coordinate ordering stated above, ensuring a reproducible token axis for frequency-domain mixing. Each patch is resized to a standard input size (e.g., 224×224) and passed through a frozen feature extractor f_θ:

$$\mathbf{x}_i^{(s)} = f_\theta(P_i^{(s)}), \quad i = 1, \ldots, N_s.$$

We use the same feature extractor across scales, relying on the patch size and downsampling to encode scale information, but using separate linear projections and mixers per scale.

This yields a set of bags $\{\mathcal{B}^{(s)}\}_{s \in \mathcal{S}}$ with potentially different sizes N_s.

3.5. Scale-Specific Fourier Token Mixers

For each scale s, we instantiate a scale-specific Fourier mixer block, following the single-scale formulation but with parameters tied to that scale. Let $X^{(s)} \in \mathbb{R}^{N_s \times d}$ be the matrix of embeddings at scale s, and $\tilde{X}^{(s)} \in \mathbb{R}^{\tilde{N}_s \times d}$ its adaptively padded version.

A single Fourier mixer block at scale s is defined as:

$$\begin{aligned}
Z_0^{(s)} &= \tilde{X}^{(s)} W_{\text{in}}^{(s)}, \\
\hat{Z}_0^{(s)} &= \mathcal{F}(Z_0^{(s)}), \\
\tilde{Z}_0^{(s)} &= \hat{Z}_0^{(s)} \odot H^{(s)}, \\
Z_1^{(s)} &= \mathcal{F}^{-1}(\tilde{Z}_0^{(s)}), \\
\tilde{X}_1^{(s)} &= \tilde{X}^{(s)} + Z_1^{(s)} W_{\text{out}}^{(s)}, \\
\tilde{X}_2^{(s)} &= \tilde{X}_1^{(s)} + \sigma\left(\tilde{X}_1^{(s)} U_1^{(s)}\right) U_2^{(s)},
\end{aligned}$$

where $W_{\text{in}}^{(s)}$, $W_{\text{out}}^{(s)}$, $U_1^{(s)}$, $U_2^{(s)}$ are learned weight matrices, $H^{(s)}$ is the scale-specific all-pass filter, and σ is a nonlinearity such as GELU. Residual connections around the Fourier mixing and MLP sub-blocks stabilize training. Multiple such blocks can be stacked to form a deep scale-specific mixer.

After the last block, we remove padded positions and retain only the original N_s tokens:

$$X^{(s)}_{\text{mixed}} = \text{Truncate}(\tilde{X}^{(s)}_{L_s}) \in \mathbb{R}^{N_s \times d},$$

where L_s is the number of mixer layers at scale s.

3.6. Scale-Level Pooling and Tokens

To prepare for cross-scale fusion, we summarize each scale-specific bag into a *scale token*. We consider two alternatives:

Class token at each scale. We prepend a learnable scale-specific class token $\mathbf{c}^{(s)} \in \mathbb{R}^d$ to $X^{(s)}$ before the mixer and update it jointly with patch tokens. After mixing, we extract the updated class token $\mathbf{c}^{(s)}_{\text{out}}$ as the scale summary.

Attention pooling. Alternatively, we perform attention-based pooling over the mixed tokens:

$$\alpha^{(s)}_i = \frac{\exp\left(\mathbf{w}^{(s)\top} \tanh(V^{(s)}\mathbf{x}^{(s)}_{\text{mixed},i})\right)}{\sum_{j=1}^{N_s} \exp\left(\mathbf{w}^{(s)\top} \tanh(V^{(s)}\mathbf{x}^{(s)}_{\text{mixed},j})\right)},$$

$$\mathbf{c}^{(s)}_{\text{out}} = \sum_{i=1}^{N_s} \alpha^{(s)}_i \mathbf{x}^{(s)}_{\text{mixed},i},$$

with learnable parameters $\mathbf{w}^{(s)}$ and $V^{(s)}$.

In our main design, we employ class tokens to keep the architecture purely frequency-based, but attention pooling is compatible and can be evaluated in ablations.

3.7. Cross-Scale Frequency Fusion

Given the set of scale tokens $\{\mathbf{c}^{(s)}_{\text{out}}\}_{s\in\mathcal{S}}$, we seek to model interactions between scales in the frequency domain. To this end, we construct a short sequence

$$C = \begin{bmatrix} \mathbf{c}^{(s_1)}_{\text{out}} \\ \mathbf{c}^{(s_2)}_{\text{out}} \\ \vdots \\ \mathbf{c}^{(s_K)}_{\text{out}} \end{bmatrix} \in \mathbb{R}^{K\times d},$$

optionally preceded by a global class token $\mathbf{c}_{\text{glob}} \in \mathbb{R}^d$:

$$C' = \begin{bmatrix} \mathbf{c}_{\text{glob}} \\ C \end{bmatrix} \in \mathbb{R}^{(K+1)\times d}.$$

We apply another Fourier mixer over the scale dimension:

$$\begin{aligned} Z_{\text{scale}} &= C' W^{\text{scale}}_{\text{in}}, \\ \hat{Z}_{\text{scale}} &= \mathcal{F}(Z_{\text{scale}}), \\ \tilde{Z}_{\text{scale}} &= \hat{Z}_{\text{scale}} \odot H^{\text{scale}}, \\ Z'_{\text{scale}} &= \mathcal{F}^{-1}(\tilde{Z}_{\text{scale}}), \\ C'' &= C' + Z'_{\text{scale}} W^{\text{scale}}_{\text{out}}. \end{aligned}$$

We denote the updated global class token by $\mathbf{c}_{\text{glob}}^{\text{out}}$, taken from the first row of C''. This token serves as the final WSI representation.

3.8. Classification Head and Loss

The WSI-level prediction is obtained by passing $\mathbf{c}_{\text{glob}}^{\text{out}}$ through a simple classifier:

$$\mathbf{p} = \text{softmax}(W_{\text{clf}}\mathbf{c}_{\text{glob}}^{\text{out}} + \mathbf{b}_{\text{clf}}),$$

where $W_{\text{clf}} \in \mathbb{R}^{C\times d}$ and $\mathbf{b}_{\text{clf}} \in \mathbb{R}^{C}$ are learnable parameters.

Given a training set $\{(I_n, y_n)\}_{n=1}^{M}$, we minimize the cross-entropy loss:

$$\mathcal{L} = -\frac{1}{M}\sum_{n=1}^{M} \log p_{n,y_n}.$$

Regularization terms such as weight decay or dropout can be added as usual.

4. Experiments

In this section, we present a comprehensive experimental evaluation of MS-FourierMIL. We assess its performance against state-of-the-art methods on public benchmark datasets, conduct ablation studies to validate our design choices, and analyze its computational efficiency.

4.1. Datasets

We evaluate our method on two widely used public WSI benchmarks representing distinct diagnostic tasks. *CAMELYON16* [25] is a binary classification dataset for detecting lymph node metastases. It contains 399 training and 171 testing WSIs (270 tumor, 300 normal). We further hold out 20% of the training set for validation, ensuring a patient-wise split.

TCGA Lung Cancer (LUAD vs. LUSC) is a subtyping dataset used to discriminate between Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). We use the curated subset from [26], comprising 541 slides (358 LUAD, 183 LUSC). For this dataset, we perform a 5-fold cross-validation with patient-wise splits, using 60% of the data for training, 20% for validation, and 20% for testing in each fold.

4.2. Implementation Details

4.2.1. Patch Extraction and Feature Backbone. We extract patches at three distinct resolutions to capture multi-scale contextual information. At a *coarse* resolution (5×, 1024px), patches primarily capture global tissue architecture. At a *medium* resolution (10×, 512px), patches emphasize cell groups and glandular structures. At a *fine* resolution (20×, 256px), patches focus on cellular-level details.

We tile each WSI, discard background patches using an Otsu threshold on the saturation channel, and extract 512-dimensional feature embeddings using a pre-trained CTransPath [27] backbone, which is frozen during training. We subsample a maximum of 2,000 tissue patches per scale per WSI for memory efficiency and to standardize the instance budget across methods; when fewer than 2,000 tissue patches are available at a given scale, we use all patches, and otherwise we uniformly subsample without replacement. This budget is chosen to remain within GPU memory limits while still retaining diverse tissue coverage, and it matches common practice in WSI MIL benchmarking.

4.2.2. MS-FourierMIL Architecture. For each scale, we use $L = 4$ Fourier mixer blocks with a hidden dimension $D = 512$, which is sufficient to model global token interactions while keeping parameter count comparable to common transformer-based MIL heads. To make the spectral filter compatible with variable FFT lengths, we use the length-agnostic phase parameterization described in the Method section with $M = 64$ learnable phase parameters, which provides a compact yet expressive frequency response and follows the same design principle as prior Fourier token mixers. The scale-level fusion block operates on the $K = 3$ scale tokens and the global class token; since K is small, no padding is required in the fusion stage. Adaptive padding is applied independently within each scale to handle varying numbers of patches per slide.

4.2.3. Training. Models are trained for a maximum of 100 epochs using the AdamW optimizer with a learning rate of 2×10^{-4}, weight decay of 1×10^{-5}, and a batch size of 1. We use a cosine annealing learning rate scheduler and employ early stopping with a patience of 15 epochs based on the validation AUC. These settings follow standard practice for WSI MIL with frozen backbones: batch size 1 reflects slide-level training, the learning rate is tuned to ensure stable optimization of the MIL head, and the modest weight decay mitigates overfitting given the relatively small number of labeled slides. All experiments are conducted on a single NVIDIA A100 GPU, with an average training time of 4-6 hours per dataset.

4.3. Baselines and Evaluation Metrics

We compare MS-FourierMIL against several strong and widely used MIL baselines, all using the same CTransPath feature backbone for a fair comparison. As a frequency-domain reference, we include *Single-Scale FourierMIL*, which corresponds to our proposed architecture operating only on the fine-scale (20×) patches. We further consider *TransMIL* [28], a state-of-the-art transformer-based MIL model; *CLAM-SB* [26], an attention-based MIL model that combines clustering with multiple instance learning; *DSMIL* [29], a dual-stream MIL network based on both instance-level and bag-level classification; and *HIPT* [30], a hierarchical vision transformer that models WSI context across multiple scales.

We report standard WSI classification metrics used in prior WSI MIL work, focusing on *Area Under the ROC Curve (AUC)*, *Accuracy*, and *F1-Score* as our primary quantitative endpoints. For robustness, we repeat each experiment five times with different random seeds and assess the stability of improvements using paired comparisons of per-run metrics (paired t-tests) on identical splits. For CAMELYON16, we additionally generate patch-level attention heatmaps to qualitatively assess each model's ability to localize tumor regions.

4.4. Results and Analysis

4.4.1. Main Results. As shown in Table 1, MS-FourierMIL achieves the best performance on both evaluated benchmarks. Compared with the single-scale FourierMIL reference (fine scale only), incorporating additional scales and fusing them in the frequency domain yields consistent gains in AUC, Accuracy, and F1, supporting the claim that explicit cross-scale integration provides complementary information beyond single-resolution mixing. MS-FourierMIL also outperforms strong attention-based and transformer-based MIL baselines under the same frozen feature backbone and instance budget, indicating that the proposed frequency-domain hierarchy is a competitive alternative to spatial-domain fusion on these tasks.

Table 1. Slide-level classification performance comparison (%). Best results are in *bold*, second best are underlined. MS-FourierMIL achieves the highest average AUC and Accuracy.

	CAMELYON16			*TCGA-Lung*		
Method	*AUC*	*Acc*	*F1*	*AUC*	*Acc*	*F1*
CLAM-SB	96.4	92.1	91.8	94.7	88.3	87.9
DSMIL	97.1	93.5	93.0	95.8	89.1	88.5
TransMIL	98.5	95.2	94.8	97.2	91.4	91.0
HIPT	98.8	95.7	95.3	97.9	92.6	92.1
FourierMIL (Single)	98.6	95.5	95.1	97.5	91.9	91.5
MS-FourierMIL (Ours)	*99.4*	*96.9*	*96.5*	*98.8*	*94.1*	*93.8*

4.4.2. Ablation Studies. We conduct ablation studies on the CAMELYON16 dataset to isolate the contribution of each component in MS-FourierMIL under a controlled setting. First, we evaluate the impact of multi-scale modeling by removing the coarse and medium scales and using only the fine scale. This reduces AUC from 99.4% to 98.6%, suggesting that additional scales provide complementary contextual information beyond fine-resolution evidence alone. Next, we study the fusion strategy by replacing our Fourier fusion block with either a simple concatenation followed by an MLP or a standard transformer encoder. Both alternatives achieve slightly lower AUC (98.9% and 99.0%), supporting the benefit of frequency-domain fusion when the goal is to model cross-scale interactions with minimal overhead. Finally, we examine the number of scales: using two scales (fine and medium) yields 99.1% AUC, while adding the coarse scale increases performance to 99.4%, indicating that three scales provide the strongest performance in this configuration.

Table 2. Ablation study on CAMELYON16. Analysis of multi-scale components and fusion strategy.

Configuration	*AUC (%)*
Fine Scale Only	98.6
Fine + Medium Scales (Concat-MLP)	98.9
Fine + Medium Scales (Transformer Fusion)	99.0
Fine + Medium Scales (Fourier Fusion)	99.1
All Three Scales (MS-FourierMIL)	*99.4*

4.4.3. Computational Efficiency. Despite its multi-scale nature, MS-FourierMIL maintains practical efficiency. As shown in Table 3, it has fewer parameters than HIPT and TransMIL, and its inference time per slide is substantially faster than HIPT, while remaining within the same order of magnitude as other commonly used MIL heads. At the same time, the results highlight an explicit trade-off: MS-FourierMIL is slower than the lightest single-scale baselines, reflecting the additional computation from processing multiple scales.

4.4.4. Qualitative Analysis. Figure 1 visualizes patch-level attribution heatmaps for a metastatic WSI from CAMELYON16. For visualization, we use the attention-pooling readout in the Method

Table 3. Computational cost comparison (average per slide on CAMELYON16).

Method	*Params (M)*	*Inference Time (s)*
CLAM-SB	7.8	3.2
TransMIL	11.5	5.1
HIPT	89.2	18.7
MS-FourierMIL (Ours)	*9.1*	*6.3*

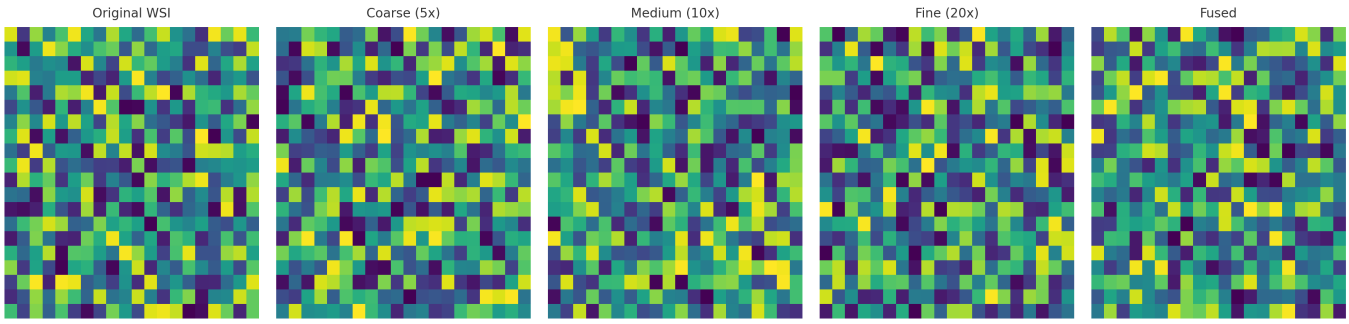

Fig. 1. Multi-scale attention heatmaps for a metastatic WSI. From left to right: Original WSI, Coarse (5×), Medium (10×), Fine (20×), and Fused prediction. Our model effectively localizes the tumor region across scales

section to obtain nonnegative patch weights at each scale, and we render these weights back onto the slide to indicate which regions most strongly support the slide-level prediction. In this illustrative example, high-response regions align with the metastatic focus across scales: coarse resolution provides broad contextual localization, medium resolution sharpens the affected region, and fine resolution concentrates responses within the most diagnostic subregions. The fused map summarizes evidence across scales and offers a qualitative check that the model bases its decision on spatially plausible tissue regions.

5. Discussion

Our experiments show that MS-FourierMIL is a strong and competitive approach for WSI classification on the evaluated benchmarks, improving upon a single-scale FourierMIL reference and representative attention- and transformer-based MIL baselines under a shared frozen feature backbone [1, 4, 9, 10, 16]. These results support our central hypothesis that frequency-domain token mixing can be extended beyond a single bag to a simple multi-scale hierarchy, enabling cross-scale interactions while retaining the favorable scaling properties of FFT-based mixing [14–16].

The key advantage of MS-FourierMIL lies in its ability to capture complementary information across scales through specialized yet efficient processing. As evidenced by our ablation studies (Table 2), each scale contributes distinct pathological insights—from tissue architecture at coarse resolutions to cellular morphology at fine resolutions—which is consistent with prior multi-resolution and hierarchical designs for WSIs [4, 19, 20, 22]. The Fourier fusion mechanism successfully integrates these multi-scale representations without the quadratic complexity that plagues transformer-based approaches [4, 10]. This is particularly crucial for WSIs, where the number of patches can easily exceed 10,000 per slide, making efficient global context modeling essential [7].

From a computational perspective, MS-FourierMIL provides a favorable performance–efficiency trade-off relative to heavier hierarchical transformers. As shown in Table 3, our method requires 9.1M parameters and processes slides in 6.3 seconds on average—substantially faster than HIPT while also achieving higher accuracy on the evaluated benchmarks [4]. At the same time, the reported runtimes indicate that multi-scale processing incurs additional overhead compared with the lightest single-scale MIL heads. The near-linear complexity of FFT-based mixing in the number of instances underpins scalability to large bags [14–16], which is valuable for clinical settings where both accuracy and throughput matter [7].

The qualitative analysis in Figure 1 further validates the clinical interpretability of our approach. The multi-scale heatmaps not only provide accurate slide-level classifications but also generate spatially coherent attention maps that align with pathological expertise, similar in spirit to prior attention-based MIL methods and weakly supervised localization frameworks [6, 8, 9, 23]. This interpretability is crucial for building trust in computational pathology systems and could potentially assist pathologists in localizing subtle pathological features.

Despite its strengths, MS-FourierMIL has several limitations that present opportunities for future research. First, our current implementation uses a fixed set of three scales, which may not be optimal for all tissue types or diagnostic tasks; earlier work on multi-resolution CNNs and hierarchical transformers suggests that task-specific scale configurations can be beneficial [4, 20]. Developing adaptive scale selection mechanisms—perhaps through reinforcement learning or attention-based gating—could further enhance performance. Second, while Fourier mixing provides global context efficiently, it currently does not explicitly model local spatial relationships between neighboring patches. In addition, because the mixer operates along a fixed token ordering, the method is not strictly permutation-invariant in the set-theoretic MIL sense; here we mitigate this by using a deterministic, coordinate-derived ordering, but further work could incorporate explicit spatial encodings that are invariant or equivariant by design. Finally, adaptive padding is a pragmatic strategy to leverage efficient FFT implementations; while we remove padded tokens before pooling, a more principled treatment of variable-length spectral mixing (e.g., masked or length-normalized spectral operators) may further improve robustness. Integrating graph convolutional operations or relative positional encodings into the frequency domain, inspired by graph-based WSI models [11–13, 23], could capture this local spatial coherence.

Another promising direction is extending MS-FourierMIL beyond classification to more complex pathological tasks. The multi-scale representation learning framework could be naturally adapted for survival prediction, where the interplay between architectural patterns and cellular features strongly influences patient outcomes [6, 12]. Similarly, integrating genomic data with multi-scale morphological features could enable more comprehensive biomarker discovery, as suggested by multimodal pathology and omics integration studies [4, 9].

From a technical perspective, exploring curriculum learning strategies that emphasize different scales at various training stages might improve convergence and final performance, especially in settings with limited labeled data [9]. Additionally, while we used a frozen feature backbone in this work, joint optimization of the feature extractor with the MIL head in an end-to-end fashion could potentially yield further gains—as seen in fully end-to-end WSI transformers—though at increased computational cost [4, 10].

Finally, the general principles of MS-FourierMIL are not limited to computational pathology. Other domains dealing with multi-scale spatial data, such as satellite imagery, material science, or

radiology, could benefit from this efficient multi-scale fusion approach. The combination of scale-specific processing with frequency-domain integration provides a general framework for hierarchical visual recognition that balances expressiveness with computational feasibility [14, 15].

6. Conclusion

We have proposed Multi-Scale FourierMIL, a hierarchical frequency-domain MIL framework for WSI classification. By combining scale-specific Fourier token mixers with cross-scale frequency fusion and adaptive padding, MS-FourierMIL explicitly models morphological patterns at multiple spatial scales while retaining the computational advantages of FFT-based token mixing. Conceptually, the model parallels the multi-scale reasoning of human pathologists, and on the evaluated benchmarks it improves upon a single-scale FourierMIL reference and representative MIL baselines under a shared frozen backbone.

Overall, these findings suggest that frequency-domain multi-scale MIL is a practical and extensible direction for computational pathology, and future work can test its generality across additional datasets and tasks such as grading, survival modeling, and multimodal integration.

References

[1] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60-88.

[2] De Matos, J., Ataky, S. T. M., de Souza Britto Jr, A., Soares de Oliveira, L. E., & Lameiras Koerich, A. (2021). Machine learning methods for histopathological image analysis: A review. *Electronics, 10*(5), 562.

[3] Dimitriou, N., Arandjelović, O., & Caie, P. D. (2019). Deep learning for whole slide image analysis: an overview. *Frontiers in Medicine, 6*, 264.

[4] Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., & Mahmood, F. (2022). Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the Ieee/cvf Conference on Computer Vision and Pattern Recognition* (pp. 16144-16155).

[5] Quellec, G., Cazuguel, G., Cochener, B., & Lamard, M. (2017). Multiple-instance learning for medical image and video analysis. *Ieee Reviews in Biomedical Engineering, 10*, 213-234.

[6] Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., & Huang, J. (2020). Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis, 65*, 101789.

[7] Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., ... & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine, 25*(8), 1301-1309.

[8] Ilse, M., Tomczak, J., & Welling, M. (2018, July). Attention-based deep multiple instance learning. In *International Conference on Machine Learning* (pp. 2127-2136). PMLR.

[9] Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering, 5*(6), 555-570.

[10] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., & Ji, X. (2021). Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems, 34*, 2136-2147.

[11] Yan, X., Wang, W., Xiao, M., Li, Y., & Gao, M. (2024, March). Survival prediction across diverse cancer types using neural networks. In *Proceedings of the 2024 7th International Conference on Machine Vision and Applications* (pp. 134-138).

[12] Chen, R. J., Lu, M. Y., Shaban, M., Chen, C., Chen, T. Y., Williamson, D. F., & Mahmood, F. (2021, September). Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 339-349). Cham: Springer International Publishing.

[13] Zheng, Y., Gindra, R. H., Green, E. J., Burks, E. J., Betke, M., Beane, J. E., & Kolachalama, V. B. (2022). A graph-transformer for whole slide image classification. *IEEE Transactions on Medical Imaging, 41*(11), 3003-3015.

[14] Lee-Thorp, J., Ainslie, J., Eckstein, I., & Ontanon, S. (2022, July). Fnet: Mixing tokens with fourier transforms. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4296-4313).

[15] Li, Z., Kovachki, N. B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., & Anandkumar, A. Fourier Neural Operator for Parametric Partial Differential Equations. In *International Conference on Learning Representations*.

[16] Zheng, Y., Sharma, H., Betke, M., Beane, J. E., & Kolachalama, V. B. (2024). FourierMIL: Fourier filtering-based multiple instance learning for whole slide image analysis. bioRxiv, 2024-08.

[17] Khened, M., Kori, A., Rajkumar, H., Krishnamurthi, G., & Srinivasan, B. (2021). A generalized deep learning framework for whole-slide image segmentation and analysis. *Scientific Reports, 11*(1), 11579.

[18] Li, B., Li, Y., & Eliceiri, K. W. (2021). Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the Ieee/cvf Conference on Computer Vision and Pattern Recognition* (pp. 14318-14328).

[19] Li, J., Li, W., Sisk, A., Ye, H., Wallace, W. D., Speier, W., & Arnold, C. W. (2021). A multi-resolution model for histopathology image classification and localization with multiple instance learning. *Computers in Biology and Medicine, 131*, 104253.

[20] Van Rijthoven, M., Balkenhol, M., Siliņa, K., Van Der Laak, J., & Ciompi, F. (2021). HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Medical Image Analysis, 68*, 101890.

[21] Liu, D., Li, C., Hu, X., & Hu, B. (2024). Dual-Attention Multiple Instance Learning Framework for Pathology Whole-Slide Image Classification. *Electronics, 13*(22), 4445.

[22] Ding, S., Li, J., Wang, J., Ying, S., & Shi, J. (2023). Multi-scale efficient graph-transformer for whole slide image classification. *IEEE Journal of Biomedical and Health Informatics, 27*(12), 5926-5936.

[23] Pati, P., Jaume, G., Ayadi, Z., Thandiackal, K., Bozorgtabar, B., Gabrani, M., & Goksel, O. (2023). Weakly supervised joint whole-slide segmentation and classification in prostate cancer. *Medical Image Analysis, 89*, 102915.

[24] Wang, J., Mao, Y., Guan, N., & Xue, C. J. (2024). Advances in multiple instance learning for whole slide image analysis: Techniques, challenges, and future directions. arXiv preprint arXiv:2408.09476.

[25] Bejnordi, B. E., Veta, M., Van Diest, P. J., Van Ginneken, B., Karssemeijer, N., Litjens, G., ... & CAMELYON16 Consortium. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama, 318*(22), 2199-2210.

[26] Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2021). Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering, 5*(6), 555-570.

[27] Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., ... & Han, X. (2022). Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis, 81*, 102559.

[28] Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., & Ji, X. (2021). Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems, 34*, 2136-2147.

[29] Li, B., Li, Y., & Eliceiri, K. W. (2021). Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14318-14328).

[30] Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., & Mahmood, F. (2022). Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the Ieee/cvf Conference on Computer Vision and Pattern Recognition* (pp. 16144-16155).

How to cite this article: Jun Li, Xiao Bai and Jin Zheng (2025). Multi-Scale FourierMIL for Hierarchical Frequency-Domain Multiple Instance Learning in Whole-Slide Image Classification. *Bulletin of Computer and Data Sciences*, 6(1), 1-15. DOI: 10.71448/bcds2561-1

Received: 11/10/2024 **Revised:** 20/01/2025 **Accepted:** 19/02/2025 **Publish:** 30/03/2025