Scilight

Journal of Artificial Intelligence for Automation
https://www.sciltp.com/journals/jaia



Article

Check for updates

Spatial Prototype and Multiscale Expert Multiple-Instance Learning for Breast Cancer Subtype Prediction

Zhixin Li [1], Huina Wang [2], Hans W. Guesgen [3] and Bo Liu [1,*]

[1] School of Mathematical and Computational Sciences, Massey University, Auckland 0745, New Zealand
[2] School of Science and Technology, Beijing Open University, Beijing 100081, China
[3] School of Mathematical and Computational Sciences, Massey University, Palmerston North 4410, New Zealand
[*] Correspondence: b.liu@massey.ac.nz

How To Cite: Li, Z.; Wang, H.; Guesgen, H.W.; et al. Spatial Prototype and Multiscale Expert Multiple-Instance Learning for Breast Cancer Subtype Prediction. *Journal of Artificial Intelligence for Automation* **2026**, *1*(3), 19. https://doi.org/10.53941/jaia.2026.100019

Received: 25 September 2026
Revised: 3 November 2026
Accepted: 11 November 2026
Published: 23 November 2026

Abstract: Breast cancer subtype prediction from whole-slide images (WSIs) remains difficult because subtype morphology overlaps, class imbalance can bias training, and conventional multiple-instance learning (MIL) may discard relationships between neighboring patches. Evidence captured at one patch-aggregation scale may also be insufficient for every patient. We propose Spatial Prototype and Multiscale Expert Multiple-Instance Learning (SPME-MIL) for patient-level prediction of four PAM50 subtypes. SPME-MIL combines prototype-based patch aggregation, a global K-means cluster representation, and features that describe spatial relationships between neighboring patches. A WSI-level adaptation of the Synthetic Minority Over-sampling Technique (SMOTE) augments the less frequent classes. Using patch coordinates, we average neighboring patch features to form three aggregation scales. The Hybrid classifier conditionally refines Luminal A/B predictions from a main model. The Cascade classifier first separates Luminal from Non-Luminal tumors, then predicts the subtype within the selected group. A patient-level prediction rule selects one of their subtype probability vectors. On the held-out test set, SPME-MIL achieved an accuracy of 0.6933 and a macro-F1 score of 0.6035. The prediction rule increased macro-F1 by 0.0651 over the stronger individual classifier. SPME-MIL also achieved the highest patient-level metrics among eight adapted baseline pipelines evaluated with identical features and the same patient split; the highest baseline macro-F1 was 0.4202.

Keywords: breast cancer subtyping; whole-slide imaging; multiple-instance learning

1. Introduction

Histopathological examination of hematoxylin and eosin (H&E)-stained tissue is part of routine breast cancer diagnosis and informs subsequent clinical assessment. Digitization preserves this morphology in whole-slide images (WSIs) for computational analysis. Breast cancer molecular subtypes differ in biological behavior, prognosis, and treatment response. The PAM50 classifier assigns intrinsic subtypes from a 50-gene expression profile [1]. PAM50 assignment therefore requires gene-expression profiling, which is costly and not universally accessible [2,3]. WSI-based prediction could provide a lower-cost, scalable estimate of molecular subtype alongside molecular testing.

Breast cancer subtypes have unequal prevalence, and this difference carries into WSI datasets used for subtype prediction. In a population-based study of 1319 women, Luminal A represented 53.1% of tumors [4]. Luminal B, HER2-enriched, and Basal-like tumors represented 20.5%, 13.0%, and 9.8%, respectively. During training, this distribution can bias a classifier toward Luminal A. Overall accuracy may then conceal poor performance in the less frequent subtypes. Evaluation and model development must therefore account for class-specific performance and unequal class sizes [5].

Conventional sample-level oversampling does not directly account for the structure of a WSI bag. The Synthetic Minority Over-sampling Technique (SMOTE) uses a fixed neighborhood and may interpolate through locally ambiguous regions [6]. A WSI, however, contains a variable set of patch morphologies. Slide neighborhoods used for rebalancing

Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

should therefore reflect both overall appearance and the morphological patterns represented within each bag.

Subtype-associated morphology overlaps and varies with the field of view [7–9]. Lower magnification shows broader tissue architecture, while higher magnification resolves cellular detail; pathologists move between these views during slide examination [10]. Computational multiscale models similarly combine coarse tissue context with finer morphology [11–13]. A single patch-aggregation scale may therefore omit evidence needed for either broad subtype grouping or finer subtype separation.

Multiple-instance learning (MIL) can aggregate thousands of patch embeddings without region-level annotations [14–16], but standard attention pooling treats patches as an unordered bag. This representation does not preserve the coordinates of the tissue patterns or describe their prevalence against a common slide-level reference. Patch extraction divides the tissue for computation, while histological structures often continue across adjacent patches. Breast tumor microenvironments also contain spatially organized multicellular structures associated with molecular subtype and genomic context [17,18]. Patch importance or prototype abundance alone cannot distinguish a continuous region from separated occurrences of the same morphology. These differences across tissue composition, spatial organization, and scale also suggest that one classifier may not be equally informative for every patient.

To address these problems, we propose SPME-MIL with four principal contributions:

(1) A prototype-based aggregation module assigns patches to trainable prototypes through entropy-regularized optimal transport and pools the assigned patch features into a fixed-length WSI representation.
(2) A global cluster representation fits shared K-means centers from class-balanced training patches and describes each WSI by its cluster proportions and residual deviations from those centers.
(3) Spatial-neighborhood modeling uses patch coordinates to measure the continuity and interfaces of neighboring prototype assignments. The formulation adapts multivariate Moran statistics to WSI bags and retains both within-prototype and cross-prototype spatial relationships.
(4) A WSI-level adaptive SMOTE method selects same-class neighbors from whole-slide and within-slide morphology, adjusts the neighborhood to the local class distribution, and retains generated bags only when they pass margin and radius checks.

The four contributions appear in the training-bag augmentation and multi-branch MIL panels of Figure 1. Coordinate-aligned aggregation assigns related subtype decisions to different scales in the Cascade classifier, while the Hybrid classifier applies conditional refinement between two models. A patient-level prediction rule chooses one four-class probability vector from these classifiers. Contribution-aligned ablations and controlled baseline comparisons examine the components using one patient-level split and a common feature representation.

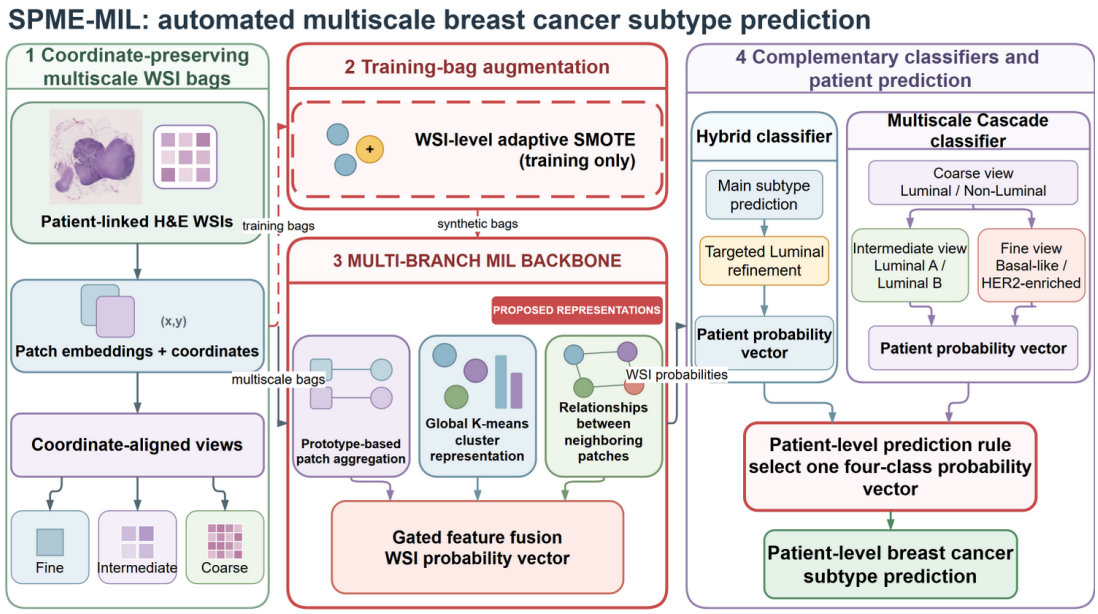


Figure 1. Overview of SPME-MIL. Coordinate-preserving aggregation forms fine, intermediate, and coarse WSI bags. WSI-level adaptive SMOTE augments the less frequent classes during training. Each scale-specific model combines prototype-based patch aggregation, a global K-means cluster representation, and spatial features from neighboring patches. The Hybrid and Cascade classifiers produce subtype probability vectors, and a patient-level prediction rule selects the final vector.

2. Related Work

2.1. Multiscale Multiple-Instance Learning for Whole-Slide Images

Attention-based MIL learns permutation-invariant aggregation from slide-level labels [14]. Clustering-constrained Attention Multiple-Instance Learning (CLAM, accessed on 14 April 2025) adds instance-clustering constraints, dual-stream MIL (DSMIL) relates a critical instance to the remaining bag, and Transformer-based MIL (TransMIL) models dependencies among instances with a transformer [15,16,19]. These approaches provide general mechanisms for learning from variable-size WSI bags.

Histological evidence changes with the field of view. Lower magnification preserves broad tissue architecture, whereas higher magnification resolves cellular morphology. Pathologists therefore move between magnifications when examining a specimen [10]. Computational models have followed this workflow in several ways. A multi-resolution MIL model used suspicious regions identified at one resolution for fine-grained grade prediction at another [11], and Navigator searches at low resolution before progressively localizing finer tissue classes [13]. HIPT models nested image resolutions, while MILBooster and ViLa-MIL combine information across scales or magnifications [12,20,21].

Breast cancer subtype studies have also reported that discriminative morphology varies within slides and across patch scales [7]. Existing multiscale designs commonly fuse scale features into one prediction or use a coarse view to guide a finer model [11–13, 20, 21]. SPME-MIL instead assigns related subtype decisions to different patch-aggregation scales. Coordinate-aligned aggregation supplies a coarse scale for Luminal/Non-Luminal separation, an intermediate scale for Luminal A/Luminal B separation, and the original patches for Basal-like/HER2-enriched separation. These scales are constructed from one patch-feature resolution rather than from separate optical magnifications. The scale-specific and Cascade ablations test whether this coarse-to-fine division contributes to patient-level classification.

2.2. SMOTE and Data Augmentation for Imbalanced WSI Classification

Weighted losses, balanced sampling, and synthetic oversampling are established approaches to class imbalance [5,22]. Balanced patch selection and resampling have also been studied for breast cancer subtype prediction [23, 24]. SMOTE interpolates samples from less frequent classes within a fixed neighborhood [6]. Adaptive Synthetic Sampling (ADASYN) allocates more synthetic samples to difficult regions, while Safe-Level-SMOTE adjusts interpolation according to the local minority density [25, 26]. BO-SMOTE guides minority-sample generation through Bayesian optimization [27]. Dou et al. combine missing-value imputation with a triple-weight oversampling strategy in an empirical feature space [28]. AW-SMOTE uses clustering and adaptive boundary and density weights to guide minority-sample generation [29]. Recent work has also extended pairwise interpolation to several same-class neighbors by sampling barycentric weights from a Dirichlet distribution [30].

Related methods have adapted augmentation to bags and WSIs. Mera et al. studied oversampling at the MIL-bag level [31]. MixUp-MIL interpolates features between WSIs [32]. ReMix uses K-means-derived prototypes to mix latent WSI content, whereas Attention-Guided Prototype Mixing selects attentive prototype instances from minority bags [33,34]. Together, these studies show that resampling can operate at the bag, feature, or prototype level. For subtype prediction, slide neighborhoods must account for both overall appearance and the variable set of morphologies within each WSI, while interpolation should remain within a locally safe same-class region.

This gap motivates the WSI-level adaptive SMOTE method in Section 3.2, which uses whole-bag morphology to select neighbors and screens generated bags against the local class structure.

2.3. Prototype Learning and Cluster-Based WSI Representations

Prototype learning reduces a large patch bag to a smaller set of recurring feature patterns. ProtoMIL compares patches with trainable prototypical parts to support case-based explanations [35]. TPMIL learns prototypes for WSI classification, while PANTHER estimates morphological prototypes and their distributions with a Gaussian mixture model [36,37]. ReMix uses cluster-centroid prototypes to construct augmented WSI representations [33]. Across these methods, prototypes support explanation, bag summarization, distribution modeling, or augmentation.

Cluster-based image representations provide a related but distinct summary. In the conventional terminology, K-means centers form a visual dictionary, the centers are called visual words, and an image is described by the proportion of local descriptors assigned to each center. The Vector of Locally Aggregated Descriptors (VLAD) additionally records the residual directions of local descriptors relative to their assigned centers [38]. Because the present study operates on histopathology patch features rather than language tokens, we use the direct terms K-means cluster centers, cluster proportions, and VLAD-style residuals below. PANTHER likewise represents a slide through

mixture probabilities and component statistics, while ReMix summarizes slide content with feature-space centers before mixing representations [33,37].

Prototype similarity alone does not control how patches are distributed across trainable prototype slots. Entropy-regularized optimal transport is an established method for matching two discrete distributions under prescribed marginal constraints [39]. SPME-MIL uses this formulation to assign patches to trainable prototypes before aggregating their features. The fixed K-means centers remain separate from the trainable prototype slots: the trainable slots adapt to subtype classification, whereas the fixed centers provide the same cluster reference for every WSI.

2.4. Spatial Context Modeling in Whole-Slide Images

Standard attention-based MIL treats a WSI as an unordered bag [14]. Its permutation-invariant pooling can identify influential patches, but it does not distinguish adjacent patches from similar patches located far apart. CLAM and DSMIL improve instance learning and feature-based interaction without using the original patch coordinates in their pooling operations [15,16]. Once spatial context is discarded, the bag representation cannot determine whether repeated morphologies form one continuous region, several separated regions, or an interface between tissue patterns.

Several WSI models restore spatial context between patches. Patch graphs connect tissue regions through spatial or morphological proximity [17]. GLAMIL learns local regional correlations before constructing a slide-level representation [40], while PEGTB-MIL encodes normalized two-dimensional patch positions and uses position reconstruction to preserve spatial-semantic consistency [41]. These approaches establish the value of patch-to-patch or region-to-region relationships, but they do not directly summarize spatial autocorrelation and transitions between the prototype-assignment patterns of neighboring patches.

This distinction matters because patch extraction partitions a histological structure for computation rather than marking a biological boundary. Neighboring patches may belong to the same tumor, stromal, or mixed tissue region, and breast tumor microenvironments contain multicellular structures associated with molecular subtype [18]. Prototype abundance alone cannot recover this organization when the same frequencies arise from different spatial layouts.

Moran's I measures spatial autocorrelation for a graph signal, and its multivariate extension summarizes association across several variables [42]. SPME-MIL applies this formulation to the prototype-assignment weights at each patch location, complementing prototype abundance with within-prototype continuity, cross-prototype association, interface mass, and neighborhood diversity.

3. Methodology

3.1. Dataset, Preprocessing, and Multiscale WSI Representation

3.1.1. Dataset and Patient-Level Split

We used diagnostic H&E WSIs from The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA) collection for patient-level PAM50 subtype classification. Each patient had a matched label from the four-class target taxonomy: Basal-like, HER2-enriched, Luminal A, or Luminal B. The analysis set and predefined patient-level partition are summarized in Table 1. The development set contained 692 WSIs from 649 patients, and the fixed test set contained 171 WSIs from 163 patients. Each original bag comprised the patch features from one WSI. For final model fitting, 65 development patients with 71 WSIs were reserved for internal validation, leaving 584 patients with 621 WSIs for fitting and oversampling.

Table 1. Composition of the TCGA-BRCA dataset and patient-level partition.

Subtype	WSIs	Patients	Patients in Development Set	Patients in Test Set
Basal-like	161	152	121	31
HER2-enriched	72	71	57	14
Luminal A	432	413	330	83
Luminal B	198	176	141	35
Total	863	812	649	163

Figure 1 summarizes the proposed method. Patch coordinates group neighboring patches at three aggregation scales. Features are averaged within each group. WSI-level adaptive SMOTE augments the less frequent classes during training. The MIL model combines prototype-based patch aggregation, a global K-means cluster representation, and spatial features from neighboring patches. The Hybrid classifier conditionally refines Luminal

A/B predictions from a main model. The Cascade classifier uses a Luminal/Non-Luminal decision followed by subtype classification. A patient-level prediction rule selects the final probability vector.

3.1.2. WSI Preprocessing and Feature Extraction

We used the WSI segmentation, patch extraction, and feature-export workflow released with CLAM [15]. CLAM preprocessing generated the inputs, and SPME-MIL performed slide classification. Tissue-containing regions were divided into patches on a coordinate grid, and the original WSI coordinates were retained. A fixed ImageNet-pretrained ResNet-50 encoded each patch. For a WSI with T retained patches, the Hierarchical Data Format version 5 (HDF5) input contained a feature matrix and its paired coordinates:

$$\mathbf{F} \in \mathbb{R}^{T \times d_x}, \qquad \mathbf{C} \in \mathbb{R}^{T \times 2}. \tag{1}$$

These paired arrays supplied both multiscale aggregation and the coordinate neighborhood graph.

3.1.3. Multiscale Patch Aggregation

The unaggregated 1×1 representation retains the original patch features. The intermediate and coarse representations average features within occupied coordinate blocks. For an $s \times s$ block,

$$\mathbf{x}_b^{(s)} = \frac{1}{|\mathcal{I}_b^{(s)}|} \sum_{j \in \mathcal{I}_b^{(s)}} \mathbf{x}_j, \qquad s \in \{2, 4\}, \tag{2}$$

where $\mathcal{I}_b^{(s)}$ contains the original patches in occupied block b. Empty blocks were excluded, and the block origin was retained as the new coordinate. The coarse representation supports Luminal versus Non-Luminal classification, the intermediate representation separates Luminal A from Luminal B, and the original patches separate Basal-like from HER2-enriched tumors and also support the full subtype model.

3.2. WSI-Level Adaptive SMOTE

We extend SMOTE to WSI feature bags through the training-only procedure in Figure 2. The procedure expands the less frequent classes within the current training subset. Each WSI i is summarized by its mean patch embedding $\boldsymbol{\mu}_i$ and a set $\mathbf{P}_i$ of within-slide cluster centers. Inter-slide distance combines the mean-feature distance with a symmetric Chamfer distance between the two center sets:

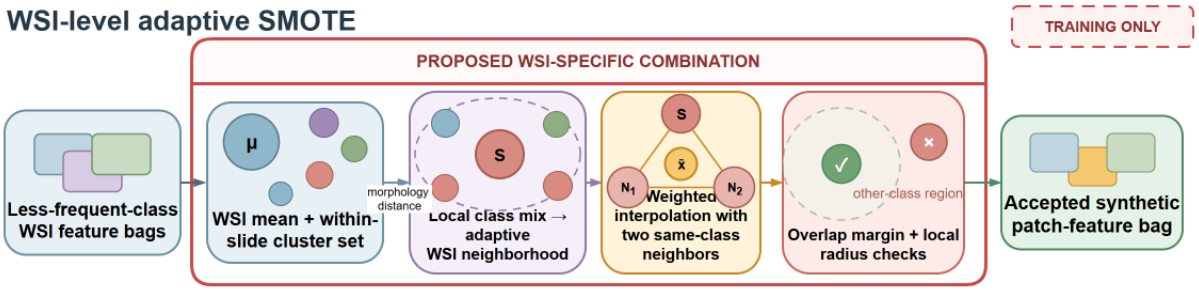


Figure 2. WSI-level adaptive SMOTE. Whole-slide and within-slide morphology define an adaptive same-class neighborhood for each source WSI. Weighted interpolation between the source and two same-class neighbors generates a candidate feature bag, and margin and radius checks determine whether it is retained. The procedure is used only during training.

$$d(i, j) = \lambda \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|_2 + (1 - \lambda) d_{\mathrm{Ch}}(\mathbf{P}_i, \mathbf{P}_j). \tag{3}$$

The coefficient $\lambda \in [0, 1]$ balances the distance between slide means and the symmetric Chamfer distance d_{Ch} between the two within-slide center sets. A mixed-class neighborhood then measures the same-class proportion $s_i \in [0, 1]$ and the largest proportion $m_i \in [0, 1]$ contributed by any other class around source WSI i. The selected same-class neighborhood size adapts to this local class structure:

$$k_i = \lfloor \mathrm{clip}(k_{\min} + \alpha s_i - \beta m_i, k_{\min}, k_{\max}) \rfloor. \tag{4}$$

The lower and upper bounds are $k_{\min}$ and $k_{\max}$. Coefficients α and β control the contributions of same-class

support and other-class intrusion, and $\mathrm{clip}(\cdot)$ restricts the result to the specified range. The clipped value is rounded down to an integer.

The source WSI and two selected same-class neighbors define a local triangular interpolation region, or simplex. Their patch sequences are matched to a common length and combined with Dirichlet weights with concentration 2 for the source WSI and total concentration 8 for the two neighbors, allocated according to normalized inverse-distance weights. Margin and local-radius checks reject candidates that approach another class or leave the local neighborhood.

Synthetic bags augment the patch-feature aggregation path, including prototype-based aggregation. Global cluster and spatial-neighborhood descriptors are computed from observed WSIs and are not generated for synthetic bags.

3.3. Prototype-Based Patch Aggregation

Figure 3 shows how learnable prototypes aggregate a variable-length WSI bag into a fixed number of feature vectors. The MIL head therefore receives the same number of aggregated feature vectors from every WSI. Let $\mathbf{X} \in \mathbb{R}^{T \times d_x}$ denote a WSI bag. A linear projection maps each patch to $\mathbf{z}_t \in \mathbb{R}^{d_h}$. The projected features and K_p prototype vectors are ℓ_2-normalized, and their temperature-scaled cosine similarity is

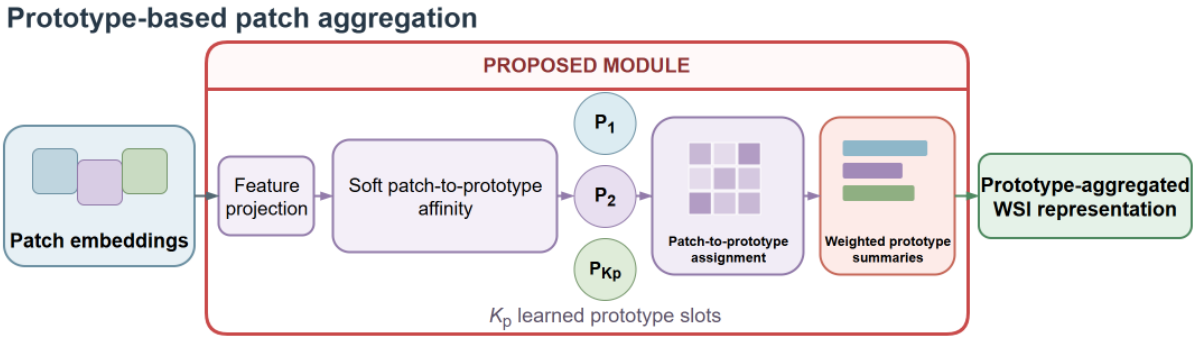


Figure 3. Prototype-based patch aggregation. Similarity scores relate projected patch embeddings to trainable prototype slots. A Sinkhorn-computed assignment with a uniform prototype marginal prevents a small number of slots from absorbing most of the patch mass. The resulting summaries convert a variable-length WSI into a fixed set of prototype-indexed features.

$$s_{tk} = \frac{\langle \mathbf{z}_t, \mathbf{c}_k \rangle}{\tau \|\mathbf{z}_t\|_2 \|\mathbf{c}_k\|_2}, \tag{5}$$

where t indexes patches, k indexes learned prototype slots, $\mathbf{c}_k$ is prototype k, and $\tau > 0$ is the similarity temperature. These trainable slots are shared across WSIs and are distinct from the fixed K-means centers described in the next subsection. We formulate the patch-to-prototype assignment as entropy-regularized optimal transport. With cost $C_{tk} = -s_{tk}$, normalized patch mass $\boldsymbol{\mu}$, and a uniform prototype marginal $\boldsymbol{\nu}$, the assignment matrix is

$$\boldsymbol{\pi}^\star = \underset{\substack{\boldsymbol{\pi} \geq 0,\ \boldsymbol{\pi}\mathbf{1} = \boldsymbol{\mu} \\ \boldsymbol{\pi}^\top \mathbf{1} = \boldsymbol{\nu}}}{\arg\min} \ \langle \boldsymbol{\pi}, \mathbf{C} \rangle + \epsilon \sum_{t,k} \pi_{tk} (\log \pi_{tk} - 1). \tag{6}$$

The nonnegative matrix $\boldsymbol{\pi}$ contains the assignment weights between patches and prototypes, $\mathbf{1}$ is an all-ones vector of compatible size, $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product, and $\epsilon > 0$ controls entropy regularization. A Sinkhorn procedure solves Equation (6) [39]. The uniform prototype marginal prevents a small number of slots from absorbing most of the patch mass. Prototype k then receives the weighted summary

$$\mathbf{v}_k = \sum_{t=1}^{T} \frac{\pi_{tk}}{\sum_{r=1}^{T} \pi_{rk} + 10^{-8}} \mathbf{z}_t, \tag{7}$$

where π_{tk} is the assignment weight from patch t to prototype k, and 10^{-8} prevents division by zero. The summaries retain feature variation within each slot rather than reducing the bag to prototype counts. A DSMIL head compares and pools this fixed-length set into a WSI representation [16].

The Sinkhorn assignment is recomputed for each forward pass without differentiating through its iterations. The classification branch updates projected patch representations through Equation (7); the spatial branch also updates the shared prototype slots through the similarities in Equation (5). The resulting prototype summaries support both WSI classification and the spatial analysis described below.

3.4. Global K-Means Cluster and Residual Representation

The global branch describes cluster proportions and mean residuals using a fixed set of K-means centers shared across WSIs, as summarized in Figure 4. To prevent the most frequent subtype from dominating center estimation, training patches are sampled by class after applying the per-WSI sampling limit. Mini-batch K-means is then fitted to the normalized training embeddings. The resulting K_g centers $\{\mathbf{c}_g\}_{g=1}^{K_g}$ are fixed during model training and test inference. The cluster proportions preserve the relative frequency of feature patterns against this common reference. These proportions complement the aggregated feature vectors produced by the learnable prototypes in Figure 3. The uniform prototype marginal assigns equal total mass to each prototype.

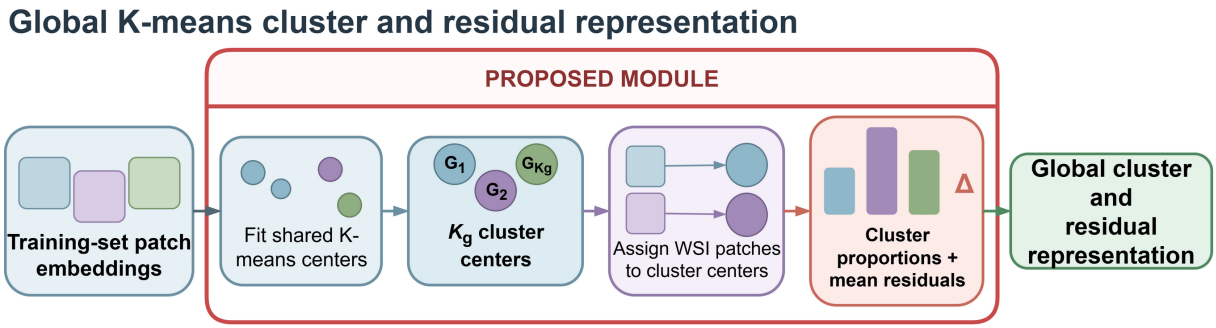


Figure 4. Global K-means cluster and residual representation. Shared cluster centers are fitted from training-set patch embeddings and then held fixed. Each WSI is described by its cluster proportions and its mean residual from each assigned center.

For normalized patch embedding $\bar{\mathbf{x}}_t$, the nearest-center assignment, cluster proportion, and projected mean residual are

$$a_t = \arg\max_g \bar{\mathbf{x}}_t^{\mathsf{T}} \mathbf{c}_g, \qquad h_g = \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}[a_t = g], \tag{8}$$

$$\mathbf{r}_g = \frac{1}{n_g} \sum_{t:a_t=g} \mathbf{U}(\bar{\mathbf{x}}_t - \mathbf{c}_g), \tag{9}$$

The index a_t identifies the cluster center assigned to patch t, $\mathbb{I}[\cdot]$ is the indicator function, h_g is the proportion of patches assigned to center g, and n_g is their number. A center with no assigned patches receives a zero residual. The matrix $\mathbf{U}$ is a training-fitted dimensionality-reduction transform shared across centers. The vector $\mathbf{h}$ records cluster proportions, while the concatenated residuals record how the assigned patches differ from their centers. This is a Vector of Locally Aggregated Descriptors (VLAD)-style encoding [38]; the implementation uses the mean residual in Equation (9). Two WSIs can therefore have similar cluster proportions but different residual morphology.

The residual blocks undergo signed square-root and ℓ_2 normalization before concatenation with $\mathbf{h}$. A learned projection ϕ_g maps the complete descriptor $\mathbf{d}_{\text{glob}}$ into the MIL hidden space. The projected descriptor is fused with the prototype-aggregated representation through a scalar residual gate, and an auxiliary subtype head provides direct supervision:

$$\mathbf{h}_{\text{glob}} = \mathbf{h}_{\text{tr}} + \tanh(g_g)\, \phi_g(\mathbf{d}_{\text{glob}}). \tag{10}$$

This branch supplies a common WSI-level cluster reference and complements the trainable summaries produced by prototype-based patch aggregation.

3.5. Spatial Relationships between Neighboring Patches

Histological structures often extend across adjacent patches, so prototype abundance alone does not describe how tissue patterns are arranged [17, 18]. Figure 5 uses the patch coordinates to connect nearby patches. At each patch, a softmax converts its prototype similarities into assignment weights across the learned prototypes. These weights form a multivariate spatial signal on the WSI patch graph. The construction adapts the graph formulation of multivariate Moran's I [42]: patches are graph vertices, coordinate neighbors define spatial edges, and prototype-assignment channels are the variables observed at each vertex.

Patch coordinates define a directed k_s-nearest-neighbor graph with binary spatial-weight matrix $\mathbf{W}$. Let $q = \sum_{t,u} W_{tu}$ and $q_{tk} = \text{softmax}_k(s_{tk})$ be the soft assignment of patch t to prototype k. After standardizing each assignment channel over valid patches, the matrices $\mathbf{Q} = [q_{tk}]$ and $\mathbf{Z} = [z_{tk}]$ yield directed prototype adjacency and

Moran-style spatial association:

$$f_a = \frac{1}{T}\sum_t q_{ta}, \tag{11}$$

$$A_{ab} = \frac{1}{q}\sum_{t,u} W_{tu} q_{ta} q_{ub}, \tag{12}$$

$$M_{ab} = \frac{1}{q}\sum_{t,u} W_{tu} z_{ta} z_{ub}, \tag{13}$$

The indices a and b denote prototype channels. The frequency f_a records the mean assignment to prototype a, A_{ab} records directed adjacency mass from channel a to channel b, and M_{ab} records their standardized spatial association. Equivalently, $\mathbf{M} = q^{-1}\mathbf{Z}^{\mathsf{T}}\mathbf{W}\mathbf{Z}$. Each diagonal element M_{aa} measures whether neighboring patches show similar standardized assignment to prototype a. The off-diagonal elements retain directional association between different prototype channels. Following the multivariate global statistic of Yamada [42], their aggregate within-prototype spatial autocorrelation is

$$I_{\text{multi}} = \frac{1}{K_p}\operatorname{tr}(\mathbf{M}) = \frac{1}{K_p}\sum_{a=1}^{K_p} M_{aa}. \tag{14}$$

SPME-MIL retains I_{multi}, the individual diagonal values, and the full cross-prototype matrix. The scalar follows the cited multivariate Moran formulation, while retaining the matrix preserves relations that would be lost by trace aggregation alone.

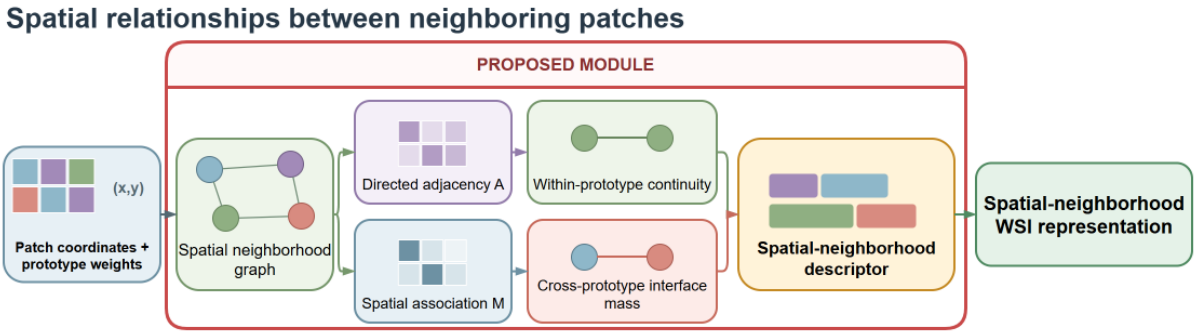


Figure 5. Spatial relationships between neighboring patches. Patch coordinates connect nearby prototype-assignment patterns. Directed adjacency and Moran-style spatial association describe within-prototype continuity and cross-prototype interfaces, which are combined into a spatial WSI representation.

The adjacency matrix supplies complementary information because it is computed from the unstandardized assignment mass. Its row-normalized form $R_{ab} = A_{ab}/\sum_b A_{ab}$ describes the neighboring prototype distribution conditional on source prototype a. Cross-prototype interface mass and frequency-weighted adjacency entropy are

$$\eta_{\text{int}} = \frac{\sum_{a\neq b} A_{ab}}{\sum_{a,b} A_{ab}}, \tag{15}$$

$$H_{\text{adj}} = -\sum_a f_a \sum_b R_{ab}\log R_{ab}. \tag{16}$$

The interface term η_{int} is the fraction of adjacency mass linking different prototype channels. The entropy H_{adj} summarizes the diversity of the row-normalized neighboring-channel distributions, weighted by prototype frequency. The spatial-neighborhood descriptor concatenates $\mathbf{f}$, normalized versions of $\mathbf{A}$ and $\mathbf{R}$, the signed normalized association matrix $\mathbf{M}$, its diagonal, I_{multi}, η_{int}, and H_{adj}. Together, these quantities distinguish WSIs with similar prototype abundance but different local continuity, interfaces, or neighborhood diversity. The descriptor is projected into the MIL hidden space before residual-gated fusion.

Residual-Gated Integration

Each scale-specific model in Figure 1 combines the three representations. The global cluster and residual descriptor is first added to the prototype-aggregated representation through a learned gate. A second gate adds the spatial-neighborhood descriptor:

$$\mathbf{h}_{\text{out}} = \mathbf{h}_{\text{in}} + \tanh(g_s)\,\phi_s(\mathbf{d}_s), \tag{17}$$

where $\mathbf{h}_{\text{in}}$ is the representation after fusion with the global cluster branch, $\mathbf{d}_s$ is the spatial-neighborhood descriptor, ϕ_s projects that descriptor into the MIL hidden space, and the scalar g_s controls its residual contribution. The classifier maps $\mathbf{h}_{\text{out}}$ to a scale-specific WSI probability vector.

3.6. Progressive Training and Patient-Level Prediction

3.6.1. Progressive Model Training

Each scale-specific model follows the same progressive schedule. Backbone pretraining first establishes a four-class representation without prototype-based patch aggregation. Training then separates Luminal from Non-Luminal tumors. The next stage classifies tumors into Basal-like, HER2-enriched, and combined Luminal groups. Four-class training then separates Luminal A from Luminal B before fine-tuning.

3.6.2. Hybrid Classifier

The Hybrid classifier in Figure 1 uses two models on the original patches. The main model produces a complete four-class probability vector $\mathbf{p}^m$. When its total Luminal probability reaches the refinement threshold τ_{Lum}, the refiner redistributes that Luminal mass between Luminal A and Luminal B:

$$m^m_{\text{Lum}} = p^m_{\text{LA}} + p^m_{\text{LB}}, \tag{18}$$

$$p^H_k = \begin{cases} p^m_k, & k \in \{\text{Basal}, \text{HER2}\}, \\ m^m_{\text{Lum}} \dfrac{p^r_k}{p^r_{\text{LA}} + p^r_{\text{LB}}}, & k \in \{\text{LA}, \text{LB}\}, \end{cases} \tag{19}$$

where superscripts m and r denote the main and refinement models. This operation preserves the total Luminal and Non-Luminal probability assigned by the main model. For patients with more than one WSI, entropy-weighted averaging gives greater weight to more concentrated WSI probability vectors:

$$w_j = \frac{\max\{1 - H(\mathbf{p}_j)/\log 4, \epsilon_w\}}{\sum_r \max\{1 - H(\mathbf{p}_r)/\log 4, \epsilon_w\}}, \qquad \mathbf{p}^H = \sum_j w_j \mathbf{p}_j. \tag{20}$$

The function $H(\mathbf{p}_j)$ is the Shannon entropy of WSI probability vector $\mathbf{p}_j$, j indexes the WSIs from a patient, and $\epsilon_w > 0$ prevents any slide from receiving zero weight.

3.6.3. Multiscale Cascade Classifier

The Cascade classifier uses the coarse representation for the first decision. Its Luminal and Non-Luminal probabilities are

$$g_N = p^4_{\text{Basal}} + p^4_{\text{HER2}}, \qquad g_L = p^4_{\text{LA}} + p^4_{\text{LB}}. \tag{21}$$

If $g_N \geq g_L$, the original-patch model separates Basal-like from HER2-enriched tumors. Otherwise, the intermediate aggregation model separates Luminal A from Luminal B. Superscripts 4, 2, and 1 denote the coarse, intermediate, and original-patch representations, respectively. Normalized branch probabilities form the complete Cascade vector

$$\mathbf{p}^C = \left[g_N \widetilde{p}^1_{\text{Basal}}, g_N \widetilde{p}^1_{\text{HER2}}, g_L \widetilde{p}^2_{\text{LA}}, g_L \widetilde{p}^2_{\text{LB}}\right]. \tag{22}$$

Patient-level stage probabilities are obtained by averaging the corresponding WSI components. Cascade confidence is the product of the first-stage confidence and the selected second-stage confidence.

3.6.4. OOF-Derived Patient-Level Prediction Rule

Hybrid is the default patient-level classifier. Cascade is selected if it predicts Luminal B and Hybrid predicts Basal-like. It is also selected if it predicts Luminal B, Hybrid predicts HER2-enriched, and the Hybrid mean WSI

entropy exceeds 0.33. When Cascade predicts Basal-like and Hybrid predicts Luminal A, Cascade is selected if its confidence is at least 0.5 and $p^H_{\mathrm{HER2}} \geq p^H_{\mathrm{LB}}$.

Hybrid labels are the maximum-probability classes of the patient-level vectors, while Cascade labels follow the hierarchical decision above. The Hybrid mean WSI entropy is the arithmetic mean of its WSI prediction entropies, computed with natural logarithms. The probabilities p^H_{HER2} and p^H_{LB} are patient-level Hybrid probabilities; Cascade confidence is defined above.

Patient-stratified training folds generate out-of-fold (OOF) predictions for the Hybrid and Cascade classifiers, following the training-only construction of stacked generalization [43]. Every fold refits trainable components and constructs synthetic samples without its held-out patients. The Luminal refinement and Cascade confidence thresholds are selected by patient-level OOF macro-F1. The entropy threshold of 0.33 and the Hybrid probability comparison are fixed settings.

Let $g_i \in \{0, 1\}$ denote which classifier is selected for patient i. The final probability vector is

$$\mathbf{p}_i^{\mathrm{final}} = \begin{cases} \mathbf{p}_i^C, & g_i = 1, \\ \mathbf{p}_i^H, & g_i = 0. \end{cases} \tag{23}$$

The same selection is applied to all WSIs from that patient for secondary WSI-level evaluation. The conditions and thresholds are fixed before test inference.

3.6.5. Comparison of Patient-Level Fusion Methods

We compared the conditional rule with probability averaging, an entropy-weighted ensemble, and an OOF-trained logistic regression meta-classifier. All three methods used the same patient-level Hybrid and Cascade probability vectors. Probability averaging assigned equal weight to the two vectors. The entropy-weighted ensemble used $w_m = \max\{1 - H(\mathbf{p}^m)/\log 4,\ 0.1\}$ for $m \in \{H, C\}$, where $H(\mathbf{p}) = -\sum_{c=1}^{4} p_c \log p_c$. The fused vector was $(w_H \mathbf{p}^H + w_C \mathbf{p}^C)/(w_H + w_C)$. Here, entropy was computed from each classifier's patient-level probability vector. Both fusion methods predicted the class with the highest fused probability.

The third alternative was an L_2-regularized logistic regression meta-classifier trained on the concatenated eight probabilities from the 649 development patients' OOF predictions. It used inverse regularization strength $C = 1$, the `lbfgs` solver, a maximum of 2000 iterations, and no class weighting. These settings were fixed without test-set tuning.

3.7. Experimental Setup

3.7.1. Implementation Settings

Patches were extracted on a 256-pixel coordinate grid, resized to 224 × 224 pixels, and encoded as 1024-dimensional ResNet-50 features. Training sampled at most 2000 patch instances per WSI and projected them to a 256-dimensional hidden space. The Hybrid main model and Luminal refiner used 32 and 64 learned prototype slots, respectively. The fixed global representation used 24 K-means centers, and the spatial graph connected each patch to as many as eight coordinate neighbors. WSI-level adaptive SMOTE summarized each slide by its mean feature and 16 within-slide centers; neighbor selection, weighted multi-neighbor interpolation, and candidate checks were applied only within the current training partition.

Final training of each Hybrid and Cascade component model generated 194, 259, 0, and 170 synthetic bags for Basal-like, HER2-enriched, Luminal A, and Luminal B, respectively. After oversampling, the training set contained 311 bags per class. During OOF construction, synthetic bags were generated separately within each fold's training subset.

Five patient-stratified folds generated the OOF predictions used to select the Luminal refinement and Cascade confidence thresholds. The selected refinement threshold was $\tau_{\mathrm{Lum}} = 0.99$; all conditions were fixed before test evaluation. Models were trained with AdamW, focal cross-entropy, cosine learning-rate decay, and early stopping on validation macro-F1. A fixed random seed of 42 controlled data partitioning, initialization, oversampling, and OOF construction. Training used an NVIDIA GeForce RTX 5070 Ti with 16 GB of memory. Table 2 lists the shared training and augmentation parameters.

3.7.2. Processing Time

Runtime was measured on a workstation with an Intel Core i5-14600KF CPU, 64 GB RAM, and an NVIDIA GeForce RTX 5070 Ti GPU with 16 GB VRAM. With the models and WSI-specific global descriptors preloaded, processing the 109,434 × 90,504-pixel WSI from TCGA-A1-A0SP through CLAM feature extraction, multiscale input preparation, and final classification took 946.98 s (15.78 min). Classification from the prepared H5 features took 0.776 s. Most processing time was spent extracting features from the raw WSI, and the time required for this

stage increases with the number of tissue-containing patches retained after background exclusion. This WSI yielded 92,116 patches; only 32 of the 863 study WSIs (3.71%) contained more patches. The patch-count distribution suggests that most WSIs could be processed within approximately 15 min under the same conditions.

Table 2. Training and augmentation parameters shared by the Hybrid and Cascade component models. Stage values follow the order of four-class pretraining without prototype aggregation, binary classification, three-class classification, four-class classification, and fine-tuning.

Parameter	Value	Parameter	Value
Optimization and loss			
Batch size	1 WSI bag	Minimum learning rate	10^{-6}
Weight decay/base dropout probability	0.01/0.3	Minimum epochs/early-stopping patience	10/15 epochs per stage
Internal validation fraction for final training	0.1	Focal loss focusing parameter	2 for every class
Maximum epochs per stage	60, 25, 70, 90, 100		
Initial learning rates	$10^{-3}, 5 \times 10^{-4}, 5 \times 10^{-4}, 3 \times 10^{-4}, 2 \times 10^{-4}$		
Label smoothing by stage	0, 0, 0, 0, 0.1		
Class weights	Inverse square roots of stage-specific training class counts, normalized to mean 1		
Prototype aggregation and global representation			
Similarity temperature τ	0.15	Sinkhorn regularization ϵ/iterations	0.3/80
Patches per WSI for global K-means fitting	At most 600	Residual projection dimension	64
K-means batch size/initializations/iterations	4096/5/at most 150	Global auxiliary classification loss weight	0.3
Initial global/spatial fusion gate parameters	0.5/0		
WSI-level adaptive SMOTE and feature augmentation			
Inter-slide distance coefficient λ	0.5	Adaptive neighbor bounds $k_{\min}, k_{\max}$	5, 10
Neighborhood coefficients α, β	5, 4	Mixed-class neighborhood size	20
Same-class neighbors used for synthesis	2	Acceptance margin/radius multiplier	0.04/1.25
Dirichlet parameters	Source: 2; neighbors: total concentration 8, distributed by normalized inverse-distance weights		
Maximum synthesis attempts per class	15 times the target synthetic bag count		
Gaussian feature noise	Standard deviation 0.01; applied with probability 0.3		
Random patch removal	10% of patches; applied with probability 0.2 to bags with more than 10 patches		
MixUp	Disabled		

3.7.3. Evaluation Metrics

Let C denote the number of subtypes and N the number of patients. For subtype c, precision, recall, and F1 were defined as

$$P_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \qquad R_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \qquad F1_c = \frac{2P_cR_c}{P_c + R_c}. \tag{24}$$

Patient-level macro-F1 was the primary metric because it gives each subtype equal weight. Weighted F1 instead reflects the observed subtype frequencies:

$$\mathrm{MacroF1} = \frac{1}{C}\sum_{c=1}^{C} F1_c, \qquad \mathrm{WeightedF1} = \sum_{c=1}^{C} \frac{n_c}{N} F1_c, \tag{25}$$

where n_c is the number of patients with subtype c. Accuracy and balanced accuracy were calculated as

$$\mathrm{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i), \qquad \mathrm{BalancedAccuracy} = \frac{1}{C}\sum_{c=1}^{C} R_c. \tag{26}$$

Area under the receiver operating characteristic curve (AUROC) and average precision (AP) were computed using one-versus-rest class probabilities. Their macro averages were

$$\mathrm{MacroAUROC} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{AUROC}_c, \qquad \mathrm{MacroAP} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{AP}_c\,. \tag{27}$$

Class-specific precision, recall, and F1 supplied additional views of performance. WSI-level metrics were secondary because labels and the final prediction rule were defined at patient level.

Patient-level bootstrap resampling quantified uncertainty in macro-F1 and its paired differences between SPME-MIL and Hybrid, Cascade, or the three alternative fusion methods on the 163-patient test set. Using fixed model predictions, we generated 10,000 resamples with replacement within each subtype, preserving the observed class counts (random seed 42). The same patient indices were used for all compared methods in each resample. We recalculated macro-F1 and subtracted each comparator's score from the SPME-MIL score. Separate 95% confidence

intervals were defined by the 2.5th and 97.5th percentiles of the resampled scores or paired differences.

3.7.4. Data Partitioning and Leakage Control

All partitions were defined at patient level, with no patient shared between training and test sets. Model fitting, class rebalancing, early stopping, K-means center estimation, and OOF parameter selection used only the training patients or patient-stratified folds within that set. Validation and test patients were excluded from SMOTE neighbor search and synthetic-sample generation. The trained models and patient-level selection settings were fixed before test evaluation.

3.7.5. Ablation Protocol

The model using the original patches and 32 learned prototype slots served as the common backbone. Each ablation retained the patient partition and training protocol while removing or replacing the indicated component. Direct patch pooling also removed the spatial descriptor that depends on prototype assignments, so that comparison evaluates their combined removal. The direct four-class control used the same total maximum epoch budget as progressive training.

3.7.6. Baseline Comparison Protocol

Eight comparison pipelines used the same 1024-dimensional ResNet-50 features and fixed patient split. DTFD-MIL and PANTHER followed their public implementations where available; the remaining pipelines adapted the published aggregation or classification procedures to the shared inputs. The Tafavvoghi OvR pipeline used XGBoost version 3.0.2 in both the original and SMOTE conditions. Each baseline was evaluated with and without training-only SMOTE while retaining its other settings. All reported values were generated in this study; citations identify the source methods.

4. Results

4.1. Ablation Studies

The ablations examine WSI-level adaptive SMOTE, the three backbone representations, progressive training, and the patient-level prediction rule.

4.1.1. WSI-Level Adaptive SMOTE

Seven configurations tested the proposed oversampling method and six controlled changes (Table 3). The changes replaced the adaptive neighborhood, multi-neighbor interpolation, or acceptance checks, alone or in combination, and then removed synthetic oversampling entirely.

Table 3. Component ablation of WSI-level adaptive SMOTE.

Configuration	Design Change	Patient Macro-F1	WSI Macro-F1
Adaptive neighborhood, multi-neighbor interpolation, and acceptance checks	Full oversampling configuration	0.5280	0.5096
Fixed neighborhood	Remove sample-specific neighborhood sizing	0.4878	0.4867
Fixed-neighborhood pairwise SMOTE	Remove adaptive sizing and multi-neighbor interpolation	0.5007	0.4987
Without multi-neighbor interpolation	Use pairwise interpolation	0.4927	0.4935
Without overlap or radius checks	Disable both candidate-acceptance checks	0.4774	0.4863
Without multi-neighbor interpolation, overlap, or radius checks	Use pairwise interpolation without acceptance checks	0.5046	0.4989
Without synthetic oversampling	Disable feature-bag synthesis	0.4503	0.4425

The proposed SMOTE method achieved patient and WSI macro-F1 scores of 0.5280 and 0.5096. Each single-component change reduced patient macro-F1: 0.4878 with a fixed neighborhood, 0.4927 with pairwise interpolation, and 0.4774 without the acceptance checks. Disabling synthetic oversampling produced the lowest scores, with patient macro-F1 of 0.4503 and WSI macro-F1 of 0.4425. The combined procedure therefore performed better than its simplified variants in this ablation.

4.1.2. Backbone Components and Training Strategy

Table 4 reports the backbone comparisons. Removing Sinkhorn balance changes the patch-to-prototype assignment, whereas direct patch pooling also disables the spatial descriptor that depends on prototype-assignment weights. The latter therefore evaluates the two changes together.

Table 4. Controlled ablation of the $K = 32$ backbone components and training strategy.

Configuration	Controlled Change	Patient Accuracy	Patient Macro-F1	WSI Accuracy	WSI Macro-F1
Full $K = 32$ backbone	None	0.6503	0.5280	0.6316	0.5096
Without spatial-neighborhood modeling	Disable the descriptor of neighboring prototype assignments	0.6319	0.4877	0.6257	0.4882
Without the global cluster representation	Disable the cluster-proportion and residual branch and its auxiliary head	0.5951	0.5125	0.5789	0.5003
Without Sinkhorn balance	Use unbalanced similarity-based soft assignment	0.6380	0.5178	0.6316	0.5172
Without prototype-based patch aggregation	Pool projected patch features; disable the dependent spatial-neighborhood descriptor	0.6319	0.5268	0.6199	0.5195
Direct four-class training	Replace the five-stage curriculum with one four-class stage trained from random initialization	0.5706	0.4619	0.5614	0.4601

The representation ablations showed different effects on accuracy and macro-F1. Removing the global cluster proportions and residuals produced the largest decrease in accuracy, from 0.6503 to 0.5951 at patient level and from 0.6316 to 0.5789 at WSI level. Removing spatial-neighborhood modeling reduced patient macro-F1 from 0.5280 to 0.4877 and WSI macro-F1 from 0.5096 to 0.4882. Without Sinkhorn balance, patient macro-F1 was 0.5178; direct patch pooling, which also removes the dependent spatial descriptor, yielded 0.5268. These comparisons support retaining the global cluster representation alongside prototype-based aggregation and spatial modeling. The two prototype-assignment variants had different effects on accuracy and macro-F1.

The training control replaced the progressive schedule with direct four-class training from random initialization under the same patient split, backbone, and maximum epoch budget. Patient accuracy decreased from 0.6503 to 0.5706, and patient macro-F1 decreased from 0.5280 to 0.4619. The corresponding WSI metrics fell from 0.6316 and 0.5096 to 0.5614 and 0.4601.

4.1.3. Multiscale Classifiers and Patient-Level Prediction

This analysis compared the scale and prototype configurations used before the final patient-level prediction rule. The model based on the original 1×1 patches with $K = 32$ gave the strongest individual patient results, with 0.6503 accuracy and 0.5280 macro-F1 (Table 5). In the test-set comparison, the 2×2 patch-aggregation model with $K = 32$ exceeded its $K = 64$ counterpart by 0.0409 macro-F1. The 4×4 patch-aggregation model had lower standalone four-class performance; within Cascade, this scale is used for the broader Luminal/Non-Luminal decision rather than the final subtype prediction.

Table 5. Test-set comparison of the individual scale-specific models. K denotes the number of learned prototypes.

Configuration	Scale/Role	Patient Accuracy	Patient Macro-F1	WSI Accuracy	WSI Macro-F1
Single model, $K = 32$	4×4 Patch aggregation	0.6012	0.4925	0.5965	0.4911
Single model, $K = 32$	1×1 Original patches	0.6503	0.5280	0.6316	0.5096
Single model, $K = 32$	2×2 Patch aggregation	0.6258	0.5099	0.6140	0.5045
Single model, $K = 64$	1×1 Original patches	0.6380	0.5044	0.6316	0.5065
Single model, $K = 64$	2×2 Patch aggregation	0.6074	0.4690	0.5965	0.4627

Training-set OOF predictions selected $\tau_{\mathrm{Lum}} = 0.99$ by patient macro-F1. At this value, Hybrid reached 0.6348

patient accuracy, 0.5519 macro-F1, and 0.6594 Luminal A/B macro-F1 in OOF evaluation. Disabling the $K = 64$ refinement reduced the corresponding values to 0.6225, 0.5375, and 0.6271.

The OOF results supported conditional $K = 64$ refinement within Hybrid.

The final experiment asked whether the OOF-derived patient-level rule improved performance beyond using Cascade or Hybrid alone. Table 6 compares the two classifiers with SPME-MIL on the same fixed test set.

Table 6. Test-set comparison of the Hybrid and Cascade classifiers with the patient-level prediction rule. CI denotes confidence interval; intervals were calculated from 10,000 patient-level bootstrap resamples stratified by subtype.

Configuration	Classifier/Role	Patient Accuracy	Patient Macro-F1 (95% CI)	WSI Accuracy	WSI Macro-F1
Cascade classifier	Multiscale hierarchy	0.6258	0.5357 [0.4451, 0.6219]	0.6140	0.5281
Hybrid classifier	$\tau_{\mathrm{Lum}} = 0.99$	0.6503	0.5384 [0.4478, 0.6244]	0.6316	0.5238
SPME-MIL	**OOF-derived prediction rule**	**0.6933**	**0.6035** [0.5114, 0.6888]	**0.6725**	**0.5824**
Paired patient macro-F1 comparison			Difference (95% CI)		
SPME-MIL − Hybrid			0.0651 [0.0251, 0.1121]		
SPME-MIL − Cascade			0.0678 [−0.0082, 0.1440]		

Bold indicates the highest accuracy and macro-F1 values.

SPME-MIL achieved a higher observed patient-level macro-F1 than either classifier alone. The paired confidence interval for the difference from Hybrid excluded zero, whereas the interval for the difference from Cascade included zero.

The conditional rule achieved higher patient-level accuracy and macro-F1 than the three alternative fusion methods evaluated here (Table 7). The paired confidence intervals for its macro-F1 differences from all three alternatives excluded zero.

Table 7. Patient-level comparison of fusion methods on the same test set. Δ macro-F1 is the SPME-MIL score minus the score of the method in that row. CI denotes the 95% paired bootstrap confidence interval.

Fusion Method	Patient Accuracy	Patient Macro-F1	Δ Macro-F1 (95% CI)
Probability averaging	0.6196	0.5035	0.1000 [0.0482, 0.1588]
Entropy-weighted ensemble	0.6258	0.5133	0.0902 [0.0395, 0.1477]
OOF logistic regression	0.6319	0.5225	0.0810 [0.0269, 0.1443]
SPME-MIL conditional rule	**0.6933**	**0.6035**	—

Bold indicates the highest accuracy and macro-F1 values.

The two classifiers were correct on different patients. Both were correct for 89 of 163 patients, while Cascade alone was correct for 13 and Hybrid alone for 17. An oracle that always selected a correct classifier when either was correct would classify 119 patients correctly, corresponding to 0.7301 accuracy. The fixed prediction rule classified 113 patients correctly. It selected Cascade for eight patients, of whom seven were correctly classified by Cascade and none by Hybrid. Relative to Hybrid alone, this increased the number of correct patients from 106 to 113.

The improvement was distributed across all four subtypes. Final class-specific F1 scores were 0.6538 for Basal-like, 0.4444 for HER2-enriched, 0.8156 for Luminal A, and 0.5000 for Luminal B. The corresponding Hybrid-only values were 0.5818, 0.4000, 0.8111, and 0.3607. The largest gain occurred for Luminal B, which was the main target of the subtype-specific selection conditions.

4.2. Comparison with State-of-the-Art Methods

The final controlled experiment compared SPME-MIL with alternative WSI and breast-subtyping pipelines using the same feature representation, four-class taxonomy, and patient split. It included three general WSI models, CLAM-SB [15], DTFD-MIL [44], and PANTHER [37], together with five breast-subtyping pipelines adapted to the shared inputs from Tafavvoghi et al. [3], Liu et al. [23], Phan et al. [2], Kiraz et al. [9], and Jaber et al. [7]. Every entry in Table 8 was generated under the shared protocol described in Sections 3 and 3.7; none is a performance value transcribed from a source paper. The table reports each baseline both without and with the added SMOTE step.

SPME-MIL achieved the highest value for every patient-level metric in Table 8. Among the general WSI MIL baselines, CLAM-SB with SMOTE had the highest macro-F1 at 0.4102. Among the breast-subtyping pipelines, the Phan patch-mean model with SMOTE reached 0.4202, followed by the original Jaber PCA-RBF-SVM at 0.4188.

Relative to the strongest baseline, SPME-MIL increased macro-F1 by 0.1833 and balanced accuracy by 0.1652. It also retained non-zero F1 for all four subtypes.

Table 8. Patient-level controlled comparison on the fixed test set ($n = 163$). Bal. acc. denotes balanced accuracy, macro-AUROC denotes the macro-average area under the receiver operating characteristic curve, and macro-AP denotes macro-average precision. SMOTE denotes the additional training-only balancing condition. All values were generated in this study; citations identify the source method. Bold indicates the highest value in each column.

Model	Accuracy	Bal. Acc.	Macro-F1	Weighted F1	Macro-AUROC	Macro-AP
CLAM-SB [15]	0.5092	0.2500	0.1687	0.3436	0.6636	0.4098
CLAM-SB + SMOTE	0.5153	0.4218	0.4102	0.5086	0.6857	0.4297
DTFD-MIL [44]	0.5092	0.2500	0.1687	0.3436	0.5562	0.3173
DTFD-MIL + SMOTE	0.4479	0.3707	0.3365	0.4383	0.5853	0.3443
PANTHER [37]	0.5153	0.2714	0.2236	0.3922	0.6297	0.3823
PANTHER + SMOTE	0.3988	0.3934	0.3541	0.4246	0.6239	0.3456
Tafavvoghi OvR-XGBoost [3]	0.4847	0.3829	0.3796	0.4961	0.6520	0.4028
Tafavvoghi OvR-XGBoost + SMOTE	0.1840	0.2869	0.1819	0.1749	0.5239	0.2769
DPMIL [23]	0.5092	0.2500	0.1687	0.3436	0.5455	0.3041
DPMIL + SMOTE	0.5521	0.3236	0.3086	0.4547	0.6614	0.3977
Phan patch-mean model [2]	0.5092	0.2500	0.1687	0.3436	0.6400	0.3785
Phan patch-mean model + SMOTE	0.5153	0.4203	0.4202	0.5140	0.6883	0.4437
Kiraz ABMIL [9]	0.5092	0.2500	0.1687	0.3436	0.6378	0.3747
Kiraz ABMIL + SMOTE	0.4908	0.3731	0.3571	0.4650	0.6270	0.3802
Jaber PCA-RBF-SVM [7]	0.5092	0.4349	0.4188	0.5076	0.6768	0.4352
Jaber PCA-RBF-SVM + SMOTE	0.5583	0.3939	0.4040	0.5193	0.6809	0.4514
SPME-MIL	**0.6933**	**0.5855**	**0.6035**	**0.6852**	**0.8107**	**0.5806**

Without SMOTE, CLAM-SB, DTFD-MIL, DPMIL, the Phan patch-mean model, and Kiraz ABMIL predicted Luminal A for every test patient. The small, imbalanced training set may have contributed to this pattern. We therefore compared each baseline with and without SMOTE applied only to the training set.

SMOTE improved macro-F1 for six of the eight adapted baselines but reduced it for the Tafavvoghi and Jaber pipelines, showing that oversampling was not uniformly beneficial across classifiers. Because all entries use the same ResNet-50 features and patient split, the table compares the adapted pipelines under this experimental setting rather than reproducing the results reported in their source studies.

5. Discussion

The experiments support combining WSI-level class rebalancing, prototype-based patch aggregation, spatial context, and scale-specific prediction within one patient-level framework. The common feature representation and patient split allow a direct comparison among the adapted pipelines in this study; the source methods used different encoders, preprocessing procedures, and datasets.

Errors in less frequent subtypes remain a practical constraint. The patient-level rule can use complementary predictions from the Hybrid and Cascade classifiers, but it cannot correct a case that both classifiers misclassify. Future work will investigate larger and more diverse training sets together with pathology-specific encoders such as UNI, CONCH, and Prov-GigaPath [45–47].

This study used one TCGA dataset and one shared feature representation, so performance across institutions and scanning conditions remains unknown. Evaluation on independent datasets and pathologist review are needed to assess the spatial-neighborhood descriptors and prototype assignments. Future work will also examine pan-cancer WSI classification with cancer-specific retraining and external validation. Larger training datasets may support a learned patient-level classifier in place of the current OOF-derived rule set.

6. Conclusions

SPME-MIL combines prototype-based patch aggregation, a global K-means cluster representation, and spatial relationships between neighboring patches within a multiscale WSI framework. Coordinate-aligned aggregation provides three scales for coarse-to-fine subtype decisions, while the patient-level prediction rule selects between the Hybrid and Cascade classifiers. WSI-level adaptive SMOTE augments the training bags from less frequent classes. On the fixed TCGA test set, SPME-MIL achieved 0.6933 accuracy and 0.6035 macro-F1. The component ablations linked the corresponding representations and training procedure to patient-level performance, and the prediction rule improved macro-F1 over either classifier alone. Evaluation on independent datasets is needed to determine transferability across clinical sites.

Author Contributions

Z.L.: Conceptualization; Data curation; Methodology; Software; Validation; Visualization; Writing (original draft); Funding acquisition. H.W.: Writing (review and editing). H.W.G.: Supervision; Writing (review and editing). B.L.: Supervision; Funding acquisition; Writing (review and editing). All authors have read and agreed to the submitted version of the manuscript.

Funding

This work was supported in part by the China Scholarship Council (Grant No. 202500560017) and the Massey University Research Fund.

Institutional Review Board Statement

This secondary study used only publicly available, de-identified diagnostic H&E WSIs and associated PAM50 subtype data. It involved no direct interaction with human participants, access to identifiable personal information, or collection of new biological specimens. Therefore, no additional institutional ethical approval was required for this study.

Informed Consent Statement

Not applicable.

Data Availability Statement

The diagnostic H&E WSIs and associated PAM50 subtype data used in this study are available from the National Cancer Institute Genomic Data Commons under the TCGA Breast Invasive Carcinoma project (TCGA-BRCA). The code and processed analysis files are available from the corresponding author upon reasonable request.

Acknowledgments

The results published here are in whole or part based upon data generated by the TCGA Research Network.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During manuscript preparation, the authors used DeepL Translator and OpenAI Codex (GPT-5.6 Sol) for translation and language editing. The authors reviewed and edited all AI-assisted output and take full responsibility for the content of the published article.

References

1. Parker, J.S.; Mullins, M.; Cheang, M.C.U.; et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J. Clin. Oncol.* **2009**, *27*, 1160–1167. https://doi.org/10.1200/jco.2008.18.1370.
2. Phan, N.N.; Huang, C.C.; Tseng, L.M.; et al. Predicting Breast Cancer Gene Expression Signature by Applying Deep Convolutional Neural Networks from Unannotated Pathological Images. *Front. Oncol.* **2021**, *11*, 769447. https://doi.org/10.3389/fonc.2021.769447.
3. Tafavvoghi, M.; Sildnes, A.; Rakaee, M.; et al. Deep Learning-Based Classification of Breast Cancer Molecular Subtypes from H&E Whole-Slide Images. *J. Pathol. Inform.* **2025**, *16*, 100410. https://doi.org/10.1016/j.jpi.2024.100410.

4. Sweeney, C.; Bernard, P.S.; Factor, R.E.; et al. Intrinsic Subtypes from PAM50 Gene Expression Assay in a Population-Based Breast Cancer Cohort: Differences by Age, Race, and Tumor Characteristics. *Cancer Epidemiol. Biomarkers Prev.* **2014**, *23*, 714–724. https://doi.org/10.1158/1055-9965.epi-13-1023.
5. Zhang, Y.; Kang, B.; Hooi, B.; et al. Deep Long-Tailed Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10795–10816. https://doi.org/10.1109/tpami.2023.3268118.
6. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; et al. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. https://doi.org/10.1613/jair.953.
7. Jaber, M.I.; Song, B.; Taylor, C.; et al. A Deep Learning Image-Based Intrinsic Molecular Subtype Classifier of Breast Tumors Reveals Tumor Heterogeneity that May Affect Survival. *Breast Cancer Res.* **2020**, *22*, 12. https://doi.org/10.1186/s13058-020-1248-3.
8. Jang, W.; Lee, J.; Park, K.H.; et al. Molecular Classification of Breast Cancer Using Weakly Supervised Learning. *Cancer Res. Treat.* **2025**, *57*, 116–125. https://doi.org/10.4143/crt.2024.113.
9. Kiraz, U.; Fernandez-Martin, C.; Rewcastle, E.; et al. Validation of an Artificial Intelligence Model for Breast Cancer Molecular Subtyping Using Hematoxylin and Eosin-Stained Whole-Slide Images in a Population-Based Cohort. *Cancers* **2025**, *17*, 3234. https://doi.org/10.3390/cancers17193234.
10. Rasoolijaberi, M.; Babaei, M.; Riasatian, A.; et al. Multi-Magnification Image Search in Digital Pathology. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 4611–4622. https://doi.org/10.1109/jbhi.2022.3181531.
11. Li, J.; Li, W.; Sisk, A.; et al. A Multi-Resolution Model for Histopathology Image Classification and Localization with Multiple Instance Learning. *Comput. Biol. Med.* **2021**, *131*, 104253. https://doi.org/10.1016/j.compbiomed.2021.104253.
12. Chen, R.J.; Chen, C.; Li, Y.; 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), New Orleans, LA, USA, 19–24 June 2022; pp. 16123–16134. https://doi.org/10.1109/cvpr52688.2022.01567.
13. Tran, M.; Wagner, S.; Weichert, W.; et al. Navigating through Whole Slide Images with Hierarchy, Multi-Object, and Multi-Scale Data. *IEEE Trans. Med. Imaging* **2025**, *44*, 2002–2015. https://doi.org/10.1109/tmi.2025.3532728.
14. Ilse, M.; Tomczak, J.M.; Welling, M. Attention-Based Deep Multiple Instance Learning. In Proceedings of the 35th International Conference on Machine Learning (ICML 2018), Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 2127–2136.
15. Lu, M.Y.; Williamson, D.F.K.; Chen, T.Y.; et al. Data-Efficient and Weakly Supervised Computational Pathology on Whole-Slide Images. *Nat. Biomed. Eng.* **2021**, *5*, 555–570. https://doi.org/10.1038/s41551-020-00682-w.
16. Li, B.; Li, Y.; Eliceiri, K.W. 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), Nashville, TN, USA, 20–25 June 2021; pp. 14313–14323. https://doi.org/10.1109/cvpr46437.2021.01409.
17. Levy, J.; Haudenschild, C.; Barwick, C.; et al. Topological Feature Extraction and Visualization of Whole Slide Images using Graph Neural Networks. In Proceedings of the Pacific Symposium on Biocomputing 2021, Kohala Coast, HI, USA, 3–7 January 2021; Volume 26, pp. 285–296. https://doi.org/10.1142/9789811232701_0027.
18. Danenberg, E.; Bardwell, H.; Zanotelli, V.R.T.; et al. Breast Tumor Microenvironment Structures Are Associated with Genomic Features and Clinical Outcome. *Nat. Genet.* **2022**, *54*, 660–669. https://doi.org/10.1038/s41588-022-01041-y.
19. Shao, Z.; Bian, H.; Chen, Y.; et al. TransMIL: Transformer-Based Correlated Multiple Instance Learning for Whole Slide Image Classification. In Proceedings of the Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, 5–10 December 2013; Volume 26, pp. 2292–2300.
20. Qu, L.; Yang, Z.; Duan, M.; et al. Boosting Whole Slide Image Classification from the Perspectives of Distribution, Correlation and Magnification. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 21406–21416. https://doi.org/10.1109/iccv51070.2023.01962.
21. Shi, J.; Li, C.; Gong, T.; et al. ViLa-MIL: Dual-Scale Vision-Language Multiple Instance Learning for Whole Slide Image Classification. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 11248–11258. https://doi.org/10.1109/cvpr52733.2024.01069.
22. Kang, B.; Xie, S.; Rohrbach, M.; et al. Decoupling Representation and Classifier for Long-Tailed Recognition. In Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia, 26–30 April 2020.
23. Liu, H.; Xu, W.D.; Shang, Z.H.; et al. Breast Cancer Molecular Subtype Prediction on Pathological Images with Discriminative Patch Selection and Multi-Instance Learning. *Front. Oncol.* **2022**, *12*, 858453. https://doi.org/10.3389/fonc.2022.858453.
24. Mota, A.M.; Mendes, J.; Matela, N. Breast Cancer Molecular Subtype Prediction: A Mammography-Based AI Approach. *Biomedicines* **2024**, *12*, 1371. https://doi.org/10.3390/biomedicines12061371.
25. He, H.; Bai, Y.; Garcia, E.A.; et al. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IJCNN 2008—Hong Kong), Hong Kong, China, 1–8 June 2008; pp. 1322–1328. https://doi.org/10.1109/ijcnn.2008.4633969.
26. Bunkhumpornpat, C.; Sinapiromsaran, K.; Lursinsap, C. Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling TEchnique for Handling the Class Imbalanced Problem. In *Advances in Knowledge Discovery and Data Mining*; Springer Berlin Heidelberg: Berlin, Heidelberg, Germany, 2009; Volume 5476, pp. 475–482. https://doi.org/10.1007/978-3-642-01307-2_43.

27. Yan, S.; Zhao, Z.; Liu, S.; et al. BO-SMOTE: A Novel Bayesian-Optimization-Based Synthetic Minority Oversampling Technique. *IEEE Trans. Syst. Man Cybern. Syst.* **2024**, *54*, 2079–2091. https://doi.org/10.1109/tsmc.2023.3335241.
28. Dou, J.; Wei, G.; Song, Y.; et al. Switching Triple-Weight-SMOTE in Empirical Feature Space for Imbalanced and Incomplete Data. *IEEE Trans. Autom. Sci. Eng.* **2024**, *21*, 1850–1866. https://doi.org/10.1109/tase.2023.3240759.
29. Yan, S.; Guo, H.; Su, X. Adaptive Weighting–Synthetic Minority Oversampling Technique. *Mathematics* **2026**, *14*, 3238. https://doi.org/10.3390/math14173238.
30. Kachan, O.; Savchenko, A.; Gusev, G. Simplicial SMOTE: Oversampling Solution to the Imbalanced Learning Problem. In Proceedings of the KDD '25: The 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Toronto, ON, Canada, 3–7 August 2025; pp. 625–635. https://doi.org/10.1145/3690624.3709268.
31. Mera, C.; Arrieta, J.; Orozco-Alzate, M.; et al. A Bag Oversampling Approach for Class Imbalance in Multiple Instance Learning. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*; Springer International Publishing: Cham, Switzerland, 2015; Volume 9423, pp. 724–731. https://doi.org/10.1007/978-3-319-25751-8_87.
32. Gadermayr, M.; Koller, L.; Tschuchnig, M.; et al. MixUp-MIL: Novel Data Augmentation for Multiple Instance Learning and a Study on Thyroid Cancer Diagnosis. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2023*; Springer Nature Switzerland: Cham, Switzerland, 2023; Volume 14225, pp. 477–486. https://doi.org/10.1007/978-3-031-43987-2_46.
33. Yang, J.; Chen, H.; Zhao, Y.; et al. ReMix: A General and Efficient Framework for Multiple Instance Learning Based Whole Slide Image Classification. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022*; Springer Nature Switzerland: Cham, Switzerland, 2022; Volume 13432, pp. 35–45. https://doi.org/10.1007/978-3-031-16434-7_4.
34. Raswa, F.H.; Lu, C.S.; Wang, J.C. Attention-Guided Prototype Mixing: Diversifying Minority Context on Imbalanced Whole Slide Images Classification Learning. In Proceedings of the 2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2024; pp. 7609–7618. https://doi.org/10.1109/wacv57701.2024.00745.
35. Rymarczyk, D.; Pardyl, A.; Kraus, J.; et al. ProtoMIL: Multiple Instance Learning with Prototypical Parts for Whole-Slide Image Classification. In *Machine Learning and Knowledge Discovery in Databases*; Springer International Publishing: Cham, Switzerland, 2023; Volume 13713, pp. 421–436. https://doi.org/10.1007/978-3-031-26387-3_26.
36. Yang, L.; Mehta, D.; Liu, S.; et al. Trainable Prototype Enhanced Multiple Instance Learning for Whole Slide Image Classification. In Proceedings of the 6th International Conference on Medical Imaging with Deep Learning (MIDL 2023), Nashville, TN, USA, 10–12 July 2023; Volume 227, pp. 1655–1665.
37. Song, A.H.; Chen, R.J.; Ding, T.; et al. Morphological Prototyping for Unsupervised Slide Representation Learning in Computational Pathology. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 11566–11578. https://doi.org/10.1109/cvpr52733.2024.01099.
38. Jegou, H.; Douze, M.; Schmid, C.; et al. Aggregating Local Descriptors into a Compact Image Representation. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 3304–3311. https://doi.org/10.1109/cvpr.2010.5540039.
39. Cuturi, M. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In Proceedings of the Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013, Lake Tahoe, NV, USA, 5–10 December 2013; Volume 26, pp. 2292–2300.
40. Yu, J.; Ma, T.; Fu, Y.; et al. Local-to-Global Spatial Learning for Whole-Slide Image Representation and Classification. *Comput. Med. Imaging Graph.* **2023**, *107*, 102230. https://doi.org/10.1016/j.compmedimag.2023.102230.
41. Shi, J.; Sun, D.; Wu, K.; et al. Positional Encoding-Guided Transformer-Based Multiple Instance Learning for Histopathology Whole Slide Images Classification. *Comput. Methods Programs Biomed.* **2025**, *258*, 108491. https://doi.org/10.1016/j.cmpb.2024.108491.
42. Yamada, H. Moran's I for Multivariate Spatial Data. *Mathematics* **2024**, *12*, 2746. https://doi.org/10.3390/math12172746.
43. Wolpert, D.H. Stacked Generalization. *Neural Netw.* **1992**, *5*, 241–259. https://doi.org/10.1016/s0893-6080(05)80023-1.
44. Zhang, H.; Meng, Y.; Zhao, Y.; et al. DTFD-MIL: Double-Tier Feature Distillation Multiple Instance Learning for Histopathology Whole Slide Image Classification. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 18780–18790. https://doi.org/10.1109/cvpr52688.2022.01824.
45. Chen, R.J.; Ding, T.; Lu, M.Y.; et al. Towards a General-Purpose Foundation Model for Computational Pathology. *Nat. Med.* **2024**, *30*, 850–862. https://doi.org/10.1038/s41591-024-02857-3.
46. Lu, M.Y.; Chen, B.; Williamson, D.F.K.; et al. A Visual-Language Foundation Model for Computational Pathology. *Nat. Med.* **2024**, *30*, 863–874. https://doi.org/10.1038/s41591-024-02856-4.
47. Xu, H.; Usuyama, N.; Bagga, J.; et al. A Whole-Slide Foundation Model for Digital Pathology from Real-World Data. *Nature* **2024**, *630*, 181–188. https://doi.org/10.1038/s41586-024-07441-w.