MULTI-SLIDE VIEWS AND CLUSTERING-INSPIRED MEMORY RETRIEVAL FOR WEAKLY SUPERVISED WHOLE-SLIDE IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

When multiple instance learning (MIL) is applied to whole-slide image (WSI) classification, the large, variable-length instance sequences prevent multiple slides from being combined within a single batch for training. The difficulty of simultaneously incorporating information from multiple independent samples is a characteristic of this field. Moreover, the pathological significance of cross-sample view associations has not been established. To address these issues, we propose MSVCR-MIL, a two-stage weakly supervised classification framework combining multi-slide views and clustering-inspired memory retrieval. In the first stage, the model learns instance representations and attention scores under slide-level supervision. Core Memory Construction (CMC) collects high-response instances from different training slides to establish morphological references from multi-slide views and constructs a compact memory through clustering-inspired K-center coreset sampling. In the second stage, Confidence-Aware Prototype Retrieval (CAPR) generates query-conditioned prototypes and retrieval confidence scores, while Backbone-Adapted Gated Fusion (BAGF) adaptively fuses the original features with the retrieved prototypes, allowing cross-slide references to contribute to slide-level prediction. During training, Dynamic Memory Refresh (DMR) updates the instance memory and periodically reconstructs the coreset, keeping the references up to date as task representations evolve. Results on three datasets covering breast cancer, lung cancer, and skin diseases demonstrate that MSVCR-MIL consistently improves WSI classification performance (mean $\Delta\mathrm{AUC} = +1.14\%$). Retrieval-distance-based prediction risk in MSVCR-MIL is consistently lower than that obtained using methods based solely on predicted probabilities (mean $\Delta\mathrm{AURC} = -4.79\%$), indicating that it can provide complementary information for identifying prediction errors. Overall, explicit cross-slide morphological references and clustering strategies offer a feasible approach to enhancing weakly supervised WSI classification and assessing prediction risk.

1 INTRODUCTION

Histopathology supports tumor detection, typing, grading, and disease-course assessment, informing clinical care (Bera et al., 2019). Digital scanning and artificial intelligence enable gigapixel whole-slide images (WSIs) and quantitative, reproducible computational pathology (Niazi et al., 2019), with potential to improve diagnostic efficiency and consistency, alleviate specialist shortages, and extend analyses to the tumor microenvironment, biomarkers, prognosis, and drug response (Bera et al., 2019). Because detailed WSI annotation is time-consuming (Lu et al., 2021), weakly supervised multiple-instance learning (MIL) uses routine slide-level diagnoses without pixel-level delineation and is widely applied in pathology (Campanella et al., 2019; Lu et al., 2021). Each WSI forms a bag of patches whose morphology, relationships, and diagnostic relevance inform slide-level prediction (Ilse et al., 2018; Shao et al., 2021). Attention-based MIL (Ilse et al., 2018), CLAM (Lu et al., 2021), and Transformer-based MIL (Shao et al., 2021) use different instance-selection, interaction-modeling, and evidence-aggregation mechanisms, advancing weakly supervised pathology across organs and diseases.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

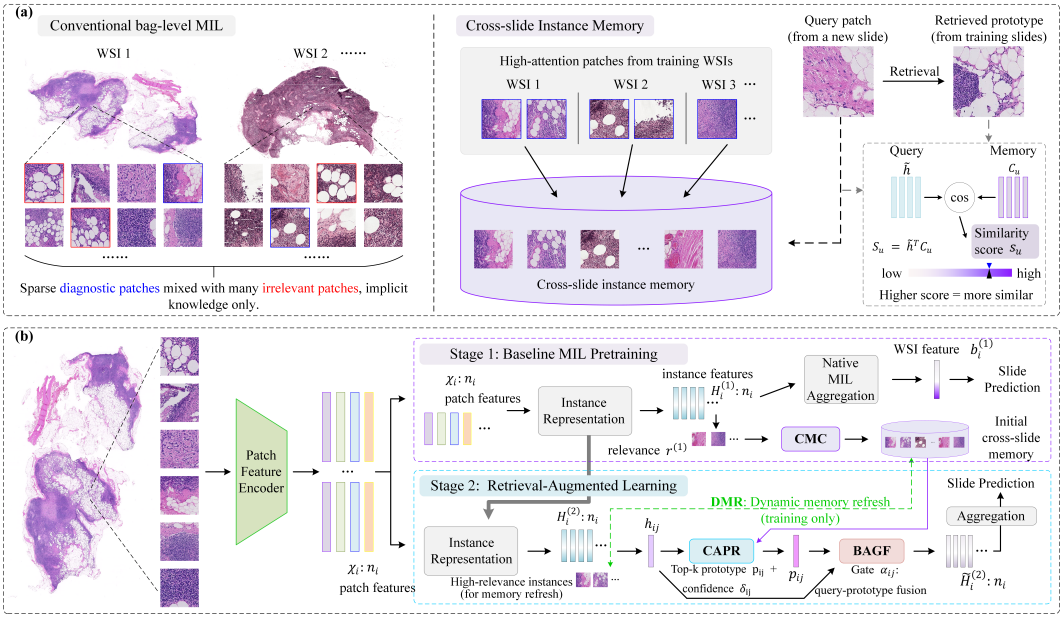

Figure 1: Motivation and overview of MSVCR-MIL. (a) Cross-slide instance memory provides explicitly retrievable morphological references beyond the current WSI. (b) Stage 1 learns task-relevant representations and constructs the memory through CMC. Stage 2 uses CAPR and BAGF to incorporate retrieved prototypes into slide-level prediction, while DMR refreshes the memory during training.

However, several issues remain. First, most MIL methods treat each WSI independently, largely confining inter-instance modeling and evidence aggregation to the current slide (Fig. 1a). Morphology from other training slides is mainly encoded indirectly in shared parameters (Shu et al., 2024). Although recent studies incorporate cohort-level morphology through global prototypes (Rymarczyk et al., 2023), cross-bag contrastive learning (Wang et al., 2022), graph structures (Shu et al., 2024), or cross-slide interactions (Wang et al., 2022; Shu et al., 2024), this information may still be compressed into model parameters or learned prototypes (Wang et al., 2022; Rymarczyk et al., 2023). Cross-slide information alone does not establish retrievable instance references. Classification performance and attention heatmaps, commonly used to evaluate MIL (Lu et al., 2021; Shao et al., 2021), are insufficient to establish the pathological significance of cross-slide associations.

Second, slide-level classification performance does not readily reveal a model's familiarity with the morphology encountered at test time. Even high classification confidence does not establish whether the current instance has adequate morphological references in the training data or whether its prediction relies on morphological patterns that were sparsely represented during training (Ovadia et al., 2019). Attention weights reflect the relative importance of an instance within the current slide, but do not directly measure how closely it matches representative morphology in the training cohort (Ilse et al., 2018). Therefore, beyond evaluating classification outcomes, it is necessary to examine shifts in the feature distribution of test instances relative to the training references, identify which morphological patterns have sufficient nearest-neighbor support and which are weakly matched, and further investigate how this degree of matching relates to prediction errors and confidence.

One-class anomaly detection methods learn the feature distribution of normal samples and identify anomalies not specified in advance by measuring how far test instances deviate from this distribution (Ruff et al., 2018), offering a transferable modeling approach to MIL's limitations. Local features from normal samples form a memory bank, coreset sampling retains representative features, and distances between test patches and their nearest memory entries estimate abnormality (Roth et al., 2022). By retaining representative training instances as explicitly retrievable external references, this approach allows the current slide to directly use morphology from other training slides without relying entirely on indirect encoding in model parameters. Comparing query patches with retrieved neighbors enables examination of their morphological consistency and whether their association provides a pathological basis for prediction.

Motivated by this observation, we propose MSVCR-MIL, a two-stage framework that enhances weakly supervised slide classification through cross-slide instance memory (Fig. 1b). Stage 1 learns task-relevant instance representations and attention scores under slide-level supervision. Core Memory Construction (CMC) selects high-response instances from different training slides, providing the current slide with multi-slide morphological references and constructing compact cross-slide memory through clustering-inspired K-center coreset sampling. In Stage 2, Confidence-Aware Prototype Retrieval (CAPR) uses instance–memory similarity to generate query-conditioned prototypes and retrieval confidence scores. Backbone-Adapted Gated Fusion (BAGF) adaptively balances the contributions of original instance features and retrieved prototypes to support subsequent slide-level prediction. During training, Dynamic Memory Refresh (DMR) continually incorporates high-response instance features and periodically reconstructs the coreset, keeping memory references current as task-relevant representations evolve. Results on three datasets covering breast cancer, lung cancer, and skin diseases demonstrate that MSVCR-MIL consistently improves WSI classification performance (mean $\Delta\mathrm{AUC} = +1.14$ percentage points relative to the corresponding baselines). Retrieval-distance-based prediction risk in MSVCR-MIL is consistently lower than that obtained using methods based solely on predicted probabilities (a 4.79% reduction in the area under the risk–coverage curve), indicating that it can provide complementary information for identifying prediction errors. Overall, MSVCR-MIL enables the current slide to explicitly retrieve and use instance information from the training cohort, extending within-slide evidence aggregation to incorporate cross-slide morphological references. The instance correspondences established through retrieval and their degree of matching can be used to assess the pathological significance of the cross-slide association mechanism, analyze deviations of test morphology from the training references, and examine the relationship between such deviations and predictive reliability.

2 METHOD

2.1 WEAKLY SUPERVISED WHOLE-SLIDE CLASSIFICATION

Let

$$\mathcal{D} = \{(\mathcal{B}_i, y_i)\}_{i=1}^{N}, \qquad y_i \in \{1, \ldots, C\}, \tag{1}$$

denote the training set, where $\mathcal{B}_i = \{\boldsymbol{x}_{ij}\}_{j=1}^{n_i}$ is the bag associated with the ith whole-slide image (WSI), n_i is the number of tissue patches in that slide, and $\boldsymbol{x}_{ij} \in \mathbb{R}^{d_0}$ is a pre-extracted patch feature. Only the slide-level label y_i is observed during training; instance labels, lesion boundaries, and the fraction of diagnostically relevant tissue within each bag are unavailable. The learning objective is therefore a permutation-invariant mapping

$$F_\Theta : \{\boldsymbol{x}_{ij}\}_{j=1}^{n_i} \longrightarrow \boldsymbol{\pi}_i \in [0,1]^C, \qquad \sum_{c=1}^{C} \pi_{ic} = 1, \tag{2}$$

such that the predicted slide class $\hat{y}_i = \arg\max_c \pi_{ic}$ agrees with y_i.

The model is optimized with slide-level cross-entropy,

$$\mathcal{L}_{\mathrm{bag}} = -\frac{1}{N} \sum_{i=1}^{N} \log \pi_{i,y_i}. \tag{3}$$

This formulation does not assign the slide label to every patch. Instead, the model must identify and aggregate an unknown, typically sparse set of diagnostically relevant instances from a large and heterogeneous bag.

2.2 TWO-STAGE LEARNING PROCEDURE OF MSVCR-MIL

MSVCR-MIL uses two consecutive training stages so that cross-slide retrieval is established in a task-discriminative representation space. Stage 1 first optimizes the bag-level multiple-instance learning (MIL) path. Let Ψ_{pre} denote the transformation from pre-extracted patch features to the selected query layer. Its output is

$$\boldsymbol{H}_i = \Psi_{\mathrm{pre}}(\mathcal{B}_i) = [\boldsymbol{h}_{i1}, \ldots, \boldsymbol{h}_{in_i}]^\top \in \mathbb{R}^{n_i \times d}. \tag{4}$$

At this point, $\boldsymbol{H}_i$ still contains n_i patch-level representations and is not a WSI-level representation. We denote the subsequent instance interaction operator by Ψ_{post} and the permutation-invariant aggregation operator by Γ. A forward pass in Stage 1 is written as

$$\boldsymbol{Z}_i = \Psi_{\mathrm{post}}(\boldsymbol{H}_i), \qquad \boldsymbol{b}_i = \Gamma(\boldsymbol{Z}_i), \qquad \boldsymbol{\pi}_i = \mathrm{softmax}(c_\omega(\boldsymbol{b}_i)) . \tag{5}$$

Here, $\boldsymbol{Z}_i$ remains an instance sequence, whereas $\boldsymbol{b}_i \in \mathbb{R}^d$ is the single bag representation associated with the WSI. Depending on the MIL formulation, Ψ_{post} and Γ may correspond to token interaction followed by class-token readout, attention-weighted pooling, or hierarchical instance aggregation. All of these forms map multiple patch representations to one slide prediction. Stage 1 is trained with Eq. (3). Once training has converged, Core Memory Construction (CMC) reads the query-layer representations $\boldsymbol{H}_i^{(1)}$ and their downstream relevance scores to form the initial memory bank $\mathcal{M}_0$ and coreset $\mathcal{C}_0$ over the training set.

Stage 2 continues from the Stage 1 parameters. For each query $\boldsymbol{h}_{ij}$ produced by Eq. (4), Confidence-Aware Prototype Retrieval (CAPR) retrieves a query-specific prototype $\boldsymbol{p}_{ij}$ from the current coreset $\mathcal{C}_t$ and estimates the corresponding neighborhood confidence δ_{ij}. Backbone-Adapted Gated Fusion (BAGF) then computes a dimension-preserving representation from $\boldsymbol{h}_{ij}$, $\boldsymbol{p}_{ij}$, and δ_{ij}:

$$(\boldsymbol{p}_{ij}, \delta_{ij}) = \mathrm{CAPR}(\boldsymbol{h}_{ij}; \mathcal{C}_t), \qquad \widetilde{\boldsymbol{h}}_{ij} = \mathrm{BAGF}(\boldsymbol{h}_{ij}, \boldsymbol{p}_{ij}, \delta_{ij}). \tag{6}$$

The resulting sequence $\widetilde{\boldsymbol{H}}_i$ replaces $\boldsymbol{H}_i$ as the input to Ψ_{post}, after which Γ produces the single bag representation $\boldsymbol{b}_i$. If $\boldsymbol{b}_i$ lies in the same task-specific space as the coreset and bag-level querying is enabled, CAPR and BAGF perform one additional query and gated update before classification:

$$(\boldsymbol{p}_i^b, \delta_i^b) = \mathrm{CAPR}(\boldsymbol{b}_i; \mathcal{C}_t), \qquad \widehat{\boldsymbol{b}}_i = \mathrm{BAGF}(\boldsymbol{b}_i, \boldsymbol{p}_i^b, \delta_i^b). \tag{7}$$

Otherwise, $\widehat{\boldsymbol{b}}_i = \boldsymbol{b}_i$. The final slide-level probabilities are

$$\boldsymbol{\pi}_i = \mathrm{softmax}\Big(c_\omega(\widehat{\boldsymbol{b}}_i)\Big) . \tag{8}$$

At the end of each training epoch, Dynamic Memory Refresh (DMR) selects high-response instances according to downstream relevance scores. The features written to memory are the pre-retrieval queries $\boldsymbol{h}_{ij}$ associated with Eq. (6), rather than their updated representations. The memory bank is updated across epochs, and the coreset is reconstructed at a prescribed interval.

2.3 Core Memory Construction (CMC)

After Stage 1 converges, CMC receives the task-specific instance representations and relevance scores of the training WSIs and produces the initial memory bank $\mathcal{M}_0$ and coreset $\mathcal{C}_0$ required in Stage 2. Each training bag is first passed through the fitted MIL model. Let $\boldsymbol{h}_{ij}^{(1)}$ denote the patch representation at the layer selected as the retrieval space, and let $r_{ij}^{(1)}$ denote its corresponding instance relevance score. CMC retains the q highest-scoring instances from each bag:

$$\mathcal{S}_i = \mathrm{TopK}_{j\in\{1,\dots,n_i\}}\left(r_{ij}^{(1)}, q\right), \qquad \mathcal{M}_0 = \bigcup_{i\in\mathcal{I}_{\mathrm{train}}} \left\{\boldsymbol{h}_{ij}^{(1)} : j \in \mathcal{S}_i\right\}. \tag{9}$$

Retrieval from the complete memory would scale with the number of candidate instances. CMC therefore applies K-Center Greedy (Sener & Savarese, 2018; Roth et al., 2022) to retain a representative and dispersed subset. All memory vectors are first ℓ_2-normalized. Let $\overline{\boldsymbol{m}} = \boldsymbol{m}/\|\boldsymbol{m}\|_2$, and let $\overline{\mathcal{M}}_0$ denote the normalized memory. Starting from an initial center, K-Center Greedy iteratively selects the item whose distance to its nearest existing center is maximal:

$$\boldsymbol{c}_t = \arg\max_{\overline{\boldsymbol{m}}\in\overline{\mathcal{M}}_0} \min_{\boldsymbol{c}\in\mathcal{C}_{t-1}} \|\overline{\boldsymbol{m}} - \boldsymbol{c}\|_2^2 , \qquad \mathcal{C}_t = \mathcal{C}_{t-1} \cup \{\boldsymbol{c}_t\}. \tag{10}$$

Selection stops when the target coreset capacity is reached. Because all vectors lie on the unit hypersphere, the squared Euclidean distance in Eq. (10) is monotonically related to cosine distance. The resulting coreset $\mathcal{C}_0$ maintains broad coverage of the original memory while reducing the number of similarity evaluations per query. It also provides the initial search space for CAPR.

2.4 CONFIDENCE-AWARE PROTOTYPE RETRIEVAL (CAPR)

CAPR receives a query representation $\boldsymbol{h}$ and the current coreset $\mathcal{C}_t$, and returns a query-specific prototype $\boldsymbol{p}$ and neighborhood confidence δ. Initially, $\mathcal{C}_t$ is the coreset $\mathcal{C}_0$ produced by CMC. After a memory refresh, CAPR uses the most recent coreset reconstructed by DMR. We omit the bag index below and write the query as $\boldsymbol{h} \in \mathbb{R}^d$. CAPR first normalizes a search copy of the query:

$$\overline{\boldsymbol{h}} = \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}. \tag{11}$$

Its cosine similarity to each normalized coreset entry $\{\boldsymbol{c}_u\}_{u=1}^{M}$ is

$$s_u = \overline{\boldsymbol{h}}^{\top} \boldsymbol{c}_u. \tag{12}$$

Let $\pi_1, \ldots, \pi_k$ index the k largest similarities in descending order. CAPR assigns temperature-scaled softmax weights within this neighborhood,

$$w_l = \frac{\exp(s_{\pi_l}/\tau)}{\sum_{r=1}^{k} \exp(s_{\pi_r}/\tau)}, \qquad l = 1, \ldots, k, \tag{13}$$

and constructs the query-specific prototype

$$\boldsymbol{p} = \sum_{l=1}^{k} w_l \boldsymbol{c}_{\pi_l}. \tag{14}$$

Compared with single-neighbor retrieval, Eq. (14) preserves local morphological agreement while assigning greater weight to coreset entries that are closer to the query.

CAPR also measures how clearly the nearest neighbor is separated from the other candidates in the retrieved neighborhood. For $k > 1$, confidence is defined as the margin between the largest and second-largest similarities:

$$\delta = \mathrm{sg}(s_{\pi_1} - s_{\pi_2}), \tag{15}$$

where $\mathrm{sg}(\cdot)$ blocks gradient propagation through the confidence value. When $k = 1$, δ is defined as the top-one similarity. A larger margin indicates a more clearly separated nearest coreset entry.

2.5 BACKBONE-ADAPTED GATED FUSION (BAGF)

BAGF takes the original query $\boldsymbol{h}$, the CAPR prototype $\boldsymbol{p}$, and the neighborhood confidence δ, and returns a dimension-preserving representation $\widetilde{\boldsymbol{h}}$. Because the utility of a retrieved prototype may vary across patches and training stages, BAGF learns a query-dependent coefficient from these three inputs:

$$\alpha = \sigma(\boldsymbol{W}_2\, \rho(\boldsymbol{W}_1[\boldsymbol{h}; \boldsymbol{p}; \delta] + \boldsymbol{b}_1) + \boldsymbol{b}_2), \tag{16}$$

$$\widetilde{\boldsymbol{h}} = \alpha \boldsymbol{h} + (1 - \alpha)\boldsymbol{p}, \tag{17}$$

where $[\cdot;\cdot]$ denotes feature concatenation, ρ is a rectified linear unit, and σ is the sigmoid function. The scalar $\alpha \in (0, 1)$ is computed independently for each query. The gate can retain more of the input representation when the external reference is uninformative, or assign greater weight to the prototype when the retrieved context is useful.

Equation (12) normalizes only the search copy of the query. The original $\boldsymbol{h}$ is retained in Eqs. (16)–(17); consequently, the retrieval metric does not alter the magnitude of the query supplied to the gating branch.

Because $\widetilde{\boldsymbol{h}}$ and $\boldsymbol{h}$ have the same dimensionality, the updated instance sequence proceeds directly to the subsequent token interaction, attention estimation, or hierarchical aggregation operators. Aggregation yields one bag representation per WSI. If this representation has the same dimensionality as the coreset, the computation in Eqs. (16)–(17) can also update a bag-level query before classification.

2.6 Dynamic Memory Refresh (DMR)

During Stage 2, DMR updates the memory bank $\mathcal{M}_t$ and coreset $\mathcal{C}_t$ between training epochs. This update reduces the discrepancy between the fixed Stage 1 memory and the evolving instance representation while keeping the memory capacity constant. At epoch t, the downstream relevance operator selects high-scoring instances from each bag, and their *pre-retrieval features* form the candidate set

$$\mathcal{U}_t = \bigcup_{i \in \mathcal{I}_{\text{train}}} \left\{ \boldsymbol{h}_{ij}^{(t)} : j \in \text{TopK}(r_{ij}^{(t)}, q) \right\}. \tag{18}$$

Using pre-retrieval features prevents the memory from being populated directly with representations that it has itself modified. Candidate features are detached from the computation graph, optionally mapped to the common retrieval space, and ℓ_2-normalized.

Let $L = |\mathcal{M}_{t-1}|$ be the fixed memory capacity and $m_t = \min(\lfloor \rho L \rfloor, |\mathcal{U}_t|)$, where ρ is the update ratio. DMR randomly samples m_t candidate features and replaces the same number of existing entries:

$$\mathcal{M}_t = \text{Keep}(\mathcal{M}_{t-1}, L - m_t) \cup \text{Sample}(\mathcal{U}_t, m_t). \tag{19}$$

The memory bank is normalized again after replacement. At a fixed refresh interval, K-Center Greedy is rerun on $\mathcal{M}_t$ to obtain

$$\mathcal{C}_t = \text{KCenter}(\mathcal{M}_t, M), \tag{20}$$

where M is the coreset capacity. CAPR uses the most recently constructed coreset between refresh points. Memory updates and coreset reconstruction are performed outside the forward pass and introduce no gradient path through stored instances. They are disabled during validation and testing, for which the memory state associated with the evaluated checkpoint remains fixed.

2.7 Learning Objective and Inference

When the MIL computation includes auxiliary supervision, the complete objective can be written as

$$\mathcal{L} = \mathcal{L}_{\text{bag}} + \sum_r \lambda_r \mathcal{L}_{\text{native}}^{(r)}, \tag{21}$$

where $\mathcal{L}_{\text{native}}^{(r)}$ denotes an existing auxiliary bag- or instance-level term in the MIL path. No additional loss explicitly forces the query representation to match the retrieved prototype. During the Stage 2 forward pass, CAPR obtains $\boldsymbol{p}$ and δ from the non-parametric coreset, whereas the BAGF gating parameters are updated by the downstream classification objective. DMR changes the memory state across epochs according to the non-parametric update in Eq. (19).

3 Experiments

3.1 Datasets and Experimental Settings

Camelyon16 (Ehteshami Bejnordi et al., 2017) is used for binary classification of breast cancer lymph node metastasis. The training set comprises 270 whole-slide images (WSIs), including 159 normal slides and 111 tumor slides; the test set comprises 129 slides, including 80 normal slides and 49 tumor slides.

Skin-PEAN (Nan et al., 2025) is used for five-class skin histopathology classification, covering melanocytic nevus, basal cell carcinoma, melanoma, squamous cell carcinoma, and seborrheic keratosis. The available dataset comprises 1,468 training slides, 2,431 internal validation slides, and 1,982 independent external test slides. The internal validation set is used for model selection and hyperparameter determination, whereas the external test set is used only for the final performance evaluation.

Lung-HSFLA (Zhao et al., 2026) is used for three-class classification of lung adenocarcinoma according to the degree of invasion, comprising adenocarcinoma in situ, minimally invasive adenocarcinoma, and invasive adenocarcinoma. The available training set contains 195 slides, with 50, 60, and 85 slides in the three classes, respectively; the independent test set contains 64 slides, with 21, 21, and 22 slides, respectively.

Table 1: WSI classification performance (%).

	CAMELYON16				Skin-PEAN				Lung-HSFLA			
Method	ACC	AUC	Recall	Precision	ACC	AUC	Recall	Precision	ACC	AUC	Recall	Precision
MaxMIL	81.40	79.29	65.31	82.05	67.51	84.30	67.49	57.66	65.63	63.94	65.66	66.81
RNN-MIL	82.95	83.88	63.27	88.57	70.48	88.41	70.45	60.73	67.19	68.32	67.17	67.98
DTFD-MIL	82.95	82.77	79.59	76.747	73.71	90.11	73.78	64.74	70.31	76.89	70.35	70.78
IBMIL-AB	88.37	90.43	87.76	82.69	74.22	91.54	74.18	64.88	67.19	72.68	67.17	67.98
CaMIL-AB	86.05	83.21	71.43	89.74	74.92	90.30	74.95	65.58	71.88	78.00	71.86	72.22
CaMIL-Trans	<u>89.15</u>	88.09	83.67	87.23	75.28	91.20	75.26	65.83	73.44	78.30	73.45	73.64
ABMIL	87.60	83.06	67.35	100.00	78.81	<u>94.93</u>	71.24	74.73	75.00	83.52	74.89	75.66
+ MSVCR-MIL	88.37	84.21	69.39	100.00	**82.80**	**95.67**	72.21	82.03	78.13	84.46	77.99	79.89
CLAM	84.50	84.36	59.18	100.00	72.10	93.62	65.81	68.46	<u>79.69</u>	<u>90.24</u>	79.51	83.08
+ MSVCR-MIL	85.27	85.46	63.27	96.88	76.84	93.88	68.55	71.70	**81.25**	**90.32**	81.31	83.46
TransMIL	<u>89.15</u>	<u>94.89</u>	75.51	94.87	77.45	93.74	73.44	68.62	75.00	84.39	74.68	77.19
+ MSVCR-MIL	**90.70**	**94.94**	79.59	95.12	<u>78.96</u>	94.48	74.71	70.48	76.56	89.57	76.19	80.01

Foreground–background segmentation is first performed on the WSIs, after which the tissue regions are divided into non-overlapping 256×256-pixel patches at $20\times$ magnification. An ImageNet-pretrained (Deng et al., 2009) ResNet-50 (He et al., 2016) is used as the image encoder to extract patch features, and the features from each slide are organized into an instance bag. The proposed method is adapted to three widely used bag-level MIL frameworks, namely ABMIL (Ilse et al., 2018), CLAM (Lu et al., 2021), and TransMIL (Shao et al., 2021), with MaxMIL (Ilse et al., 2018), RNN-MIL (Campanella et al., 2019), DTFD-MIL (Zhang et al., 2022), IBMIL-AB (Lin et al., 2023), CaMIL-AB (Chen et al., 2024), and CaMIL-Trans (Chen et al., 2024) additionally included as comparison methods. Official implementations and parameter settings are used for the baselines where available. The code has been uploaded as supplementary information. AI was used to polish writing in this manuscript.

3.2 WSI Classification Results

To evaluate whether MSVCR-MIL can consistently improve WSI classification performance across different datasets and bag-level aggregation mechanisms, comparative experiments were conducted on CAMELYON16, Skin-PEAN, and Lung-HSFLA. As shown in Table 1, consistent improvements in both AUC and accuracy were achieved with MSVCR-MIL across all three backbone architectures. Specifically, on CAMELYON16, the AUC of ABMIL was increased from 83.06% to 84.21% with MSVCR-MIL, while the accuracy of TransMIL was increased from 89.15% to 90.70%. On Skin-PEAN, AUC gains of 0.74, 0.26, and 0.74 percentage points were obtained over ABMIL, CLAM, and TransMIL, respectively, with corresponding accuracy gains of 3.99, 4.74, and 1.51 percentage points. In particular, the highest accuracy and AUC on this dataset were achieved by ABMIL combined with MSVCR-MIL, reaching 82.80% and 95.67%, respectively. On Lung-HSFLA, the largest AUC gain was observed for TransMIL, whose AUC increased from 84.39% to 89.57%, with accuracy also improving from 75.00% to 76.56%. Overall, the consistent improvements in AUC and accuracy across datasets and backbone architectures support the applicability of MSVCR-MIL to different bag-level aggregation mechanisms and indicate its ability to improve discriminative performance in WSI classification.

3.3 Patch-level Heatmap Visualization

To compare patch-level evidence localization results across different models, Figure 2 presents selected WSIs from the test set. The heatmaps use the signed tumor-vs.-normal logit-margin contribution to represent the contribution of each patch to the slide-level prediction and are normalized using a shared 99th-percentile scale within each slide. As shown in Figure 2, in the left-hand example, MSVCR-MIL produces more pronounced red responses within the tumor regions marked by Ground Truth, particularly in the upper and lower lesion regions indicated by the red arrows; the responses of the Baseline are more diffuse and less clear. Meanwhile, the non-diagnostic regions indicated by the blue arrow exhibit light-blue responses in the MSVCR-MIL heatmap, in contrast to the white response in the Baseline. In the right-hand example, MSVCR-MIL similarly retains concentrated positive responses near the lesion contours, while displaying more extensive blue regions around

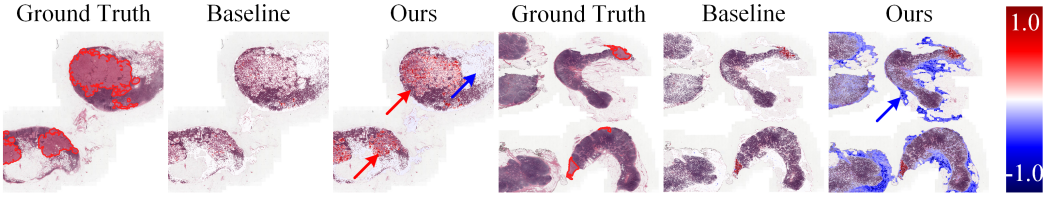

Figure 2: Patch-level heatmaps on selected test WSIs.

the lesions and along the tissue margins. The color distribution of the Baseline is comparatively uniform, and its identification of non-lesion regions is less apparent. Overall, MSVCR-MIL enhances responses in lesion regions while suppressing non-diagnostic regions.

3.4 Retrieval-based Risk Ranking

Table 2: Risk Ranking Performance

Backbone	$1-\text{MSP}\downarrow$	Entropy↓	Distance↓	ΔAURC
		Lung-HSFLA		
TransMIL	0.1094	0.1065	0.1075	−0.0019
CLAM	0.1518	0.1204	0.1239	−0.0279
ABMIL	0.1859	0.1740	0.1320	−0.0539
		CAMELYON16		
CLAM	0.0887	0.0748	0.0748	−0.0139
ABMIL	0.0948	0.0950	0.0826	−0.0122
TransMIL	0.0691	0.0280	0.0282	−0.0409
		Skin-PEAN		
TransMIL	0.1460	0.0826	0.0857	−0.0603
CLAM	0.2394	0.1013	0.1007	−0.1387
ABMIL	0.1396	0.0592	0.0584	−0.0812

To examine whether the degree of matching between test slides and training references can indicate prediction errors, a retrieval-distance-based risk score was evaluated against classification confidence and predictive entropy. With the weights and classification predictions of each MSVCR-MIL model held fixed, three scores were computed: the complement of the maximum class probability ($1-\text{MSP}$) (Hendrycks & Gimpel, 2017), predictive entropy (Entropy) (Geifman et al., 2019), and the attention-weighted nearest-neighbor reference distance among the top 10% of instances ranked by attention (Distance); higher values indicate greater risk for all three scores. Different proportions of slides were retained in ascending order of each score, and the classification error rate among the retained slides was calculated. Risk–coverage curves were thereby constructed, and the area under each risk–coverage curve (AURC) was calculated (Geifman et al., 2019), with lower values indicating lower overall prediction risk.

As shown in Table 2, the mean AURC of Distance for MSVCR-MIL was 0.0882, lower than those of $1-\text{MSP}$ (0.1361) and Entropy (0.0935). Compared with $1-\text{MSP}$, lower AURC values were obtained with Distance across all combinations, with absolute differences ranging from -0.0019 to -0.1387 and a mean reduction of 0.0479. These results indicate that the degree of matching to training references can provide supplementary information for identifying high-risk predictions and consistently outperforms risk ranking based solely on the maximum class probability across the evaluated combinations.

3.5 Memory Bank Evaluation

This experiment was conducted from two perspectives, memory intervention and reference coverage, to assess the contribution of cross-slide memory during prediction and the ability of coreset sampling to retain representative features. With the model weights held fixed, three settings—original memory, memory disabled, and random memory of equal capacity—were compared on the test sets using AUROC as the performance metric. Subsequently, two training WSIs were selected from each class, and coresets containing 1,200 entries were constructed from the same candidate bank using K-center and random selection. With the full candidate bank serving as a reference, feature coverage was measured by the attention-weighted nearest-neighbor cosine distance for the top 10% of patches ranked by attention, where lower distances indicate more complete matching.

As shown in Table 3(a), the AUC values with the original memory were 84.46%, 84.21%, and 95.67% on the three datasets, respectively. Disabling memory resulted in decreases of approximately 0.7–1.2 percentage points, indicating the effectiveness of the memory mechanism. In Table 3(b), at the same

Table 3: Memory Bank Evaluation

(a) Memory removal

Dataset	Original memory AUROC (%)↑	Memory disabled AUROC (%)↑	ΔAUROC [95% CI]
Lung-HSFLA	84.46	83.56	$-0.90\,[-2.78,\ 0.76]$
CAMELYON16	84.21	83.06	$-1.15\,[-3.12,\ 0.13]$
Skin-PEAN	95.67	94.93	$-0.75\,[-0.96,\ -0.54]$

(b) Memory bank selection

Dataset	K-center 1,200 entries	Random selection 1,200 entries	Full candidate bank
Lung-HSFLA	0.406 ± 0.014	0.421 ± 0.021	0.370
CAMELYON16	0.294 ± 0.000	0.382 ± 0.047	0.294
Skin-PEAN	0.705 ± 0.003	0.802 ± 0.010	0.661

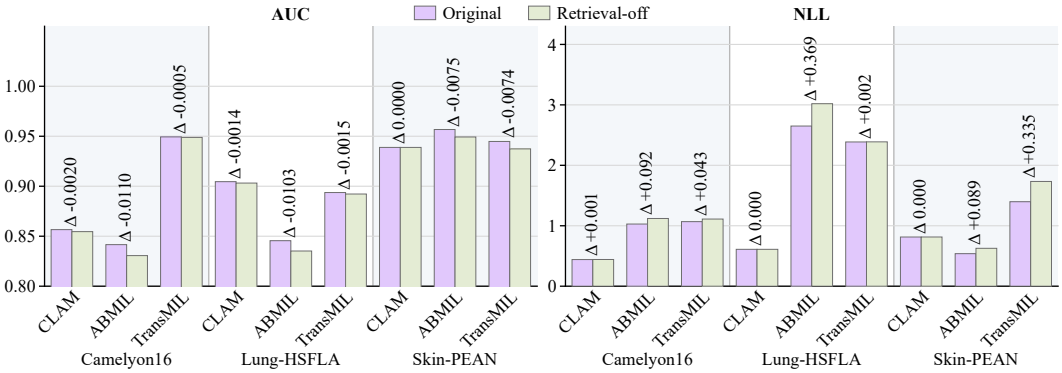

Figure 3: Fixed-checkpoint comparison of Original and Retrieval-off. Higher AUC and lower NLL indicate better performance. Δ denotes Retrieval-off minus Original.

capacity, the mean nearest-neighbor distances obtained with K-center were 0.406, 0.294, and 0.705 on the three datasets, respectively, corresponding to reductions of approximately 3.6%, 23.0%, and 12.1% relative to random selection. These results indicate that enabling the memory branch enhances the classification performance of the current model and that K-center provides more complete feature coverage than random selection under a limited capacity.

3.6 Inference Dependency Analysis

To examine whether the retrieval branch continues to make an immediate contribution to predictions after training, an inference dependency analysis was conducted with a fixed checkpoint. During inference on the test set, two conditions were compared: retaining the retrieval branch (Original) and disabling it (Retrieval-off), with all other conditions held identical. Higher AUC and lower NLL indicate better performance. As shown in Figure 3, disabling retrieval reduced AUC in eight of the nine groups by 0.05–1.10 percentage points, while the remaining group was unchanged. NLL also generally increased, with increases ranging from 0.001 to 0.369. The largest changes were observed for ABMIL on Lung-HSFLA and TransMIL on Skin-PEAN, with AUC decreases of 1.03 and 0.74 percentage points and corresponding NLL increases of 0.369 and 0.335, respectively. These results indicate that, although compatibility with the backbone architecture varies, explicit cross-slide references directly contribute to accurate inference in MSVCR-MIL through the retrieval and use of representative instances from the training cohort at inference time.

4 Conclusion

This study introduced MSVCR-MIL to enhance weakly supervised WSI classification through multi-slide views and clustering-inspired memory retrieval. The framework combines coreset sampling, confidence-aware retrieval, gated fusion, and dynamic memory refresh to establish representative instances from the training cohort as cross-slide morphological references that can be explicitly retrieved and used. Extensive dataset evaluations demonstrated improvements in classification performance, lesion evidence localization, and prediction risk identification, and verified the direct value of cross-slide references during inference. Future work will evaluate its generalization on larger-scale multicenter datasets and further validate cross-slide instance associations through pathological assessment.

REFERENCES

Kaustav Bera, Kurt A. Schalper, David L. Rimm, Vamsidhar Velcheti, and Anant Madabhushi. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature Reviews Clinical Oncology*, 16(11):703–715, 2019. doi: 10.1038/s41571-019-0252-y.

Gabriele Campanella, Matthew G. Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J. Busam, Edi Brogi, Victor E. Reuter, David S. Klimstra, and Thomas J. Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019. doi: 10.1038/s41591-019-0508-1.

Kaitao Chen, Shiliang Sun, and Jing Zhao. CaMIL: Causal multiple instance learning for whole slide image classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2): 1120–1128, 2024. doi: 10.1609/aaai.v38i2.27873.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22): 2199–2210, 2017. doi: 10.1001/jama.2017.14585.

Yonatan Geifman, Guy Uziel, and Ran El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJfb5jCqKm.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19830–19839, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Lin_Interventional_Bag_Multi-Instance_Learning_on_Whole-Slide_Pathological_Images_CVPR_2023_paper.html.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. doi: 10.1038/s41551-020-00682-w.

Tianhang Nan, Song Zheng, Siyuan Qiao, Hao Quan, Xin Gao, Jun Niu, Bin Zheng, Chunfang Guo, Yue Zhang, Xiaoqin Wang, Liping Zhao, Ze Wu, Yaoxing Guo, Xingyu Li, Mingchen Zou, Shuangdi Ning, Yue Zhao, Wei Qian, Hongduo Chen, Ruiqun Qi, Xinghua Gao, and Xiaoyu Cui. Deep learning quantifies pathologists' visual patterns for whole slide image diagnosis. *Nature Communications*, 16:5493, 2025. doi: 10.1038/s41467-025-60307-1.

Muhammad Khalid Khan Niazi, Anil V. Parwani, and Metin N. Gurcan. Digital pathology and artificial intelligence. *The Lancet Oncology*, 20(5):e253–e261, 2019. doi: 10.1016/S1470-2045(19)30154-8.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper_files/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Roth_Towards_Total_Recall_in_Industrial_Anomaly_Detection_CVPR_2022_paper.html.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4393–4402. PMLR, 2018. URL https://proceedings.mlr.press/v80/ruff18a.html.

Dawid Rymarczyk, Adam Pardyl, Jarosław Kraus, Aneta Kaczyńska, Marek Skomorowski, and Bartosz Zieliński. ProtoMIL: Multiple instance learning with prototypical parts for whole-slide image classification. In *Machine Learning and Knowledge Discovery in Databases*, volume 13713 of *Lecture Notes in Computer Science*, pp. 421–436. Springer, 2023. doi: 10.1007/978-3-031-26387-3_26.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2136–2147, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/10c272d06794d3e5785d5e7c5356e9ff-Abstract.html.

Tong Shu, Jun Shi, Dongdong Sun, Zhiguo Jiang, and Yushan Zheng. SlideGCD: Slide-based graph collaborative training with knowledge distillation for whole slide image classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15004 of *Lecture Notes in Computer Science*, pp. 470–480. Springer, 2024. URL https://papers.miccai.org/miccai-2024/712-Paper2479.html.

Xiyue Wang, Jinxi Xiang, Jun Zhang, Sen Yang, Zhongyi Yang, Ming-Hui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. SCL-WC: Cross-slide contrastive learning for weakly-supervised whole-slide image classification. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18009–18021, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/726204cea3ec27790a644e5b379175e3-Abstract-Conference.html.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18802–18812, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhang_DTFD-MIL_Double-Tier_Feature_Distillation_Multiple_Instance_Learning_for_Histopathology_Whole_CVPR_2022_paper.html.

Jianwei Zhao, Junyan Zhang, Yihao Wang, Xinwen Zhong, and Xiaojiao Guan. Hybrid supervised deep learning for lung adenocarcinoma diagnosis to optimize surgical strategies. *npj Precision Oncology*, 10:238, 2026. doi: 10.1038/s41698-026-01441-x.