Adaptive Regional Feature Fusion for Brain WSI Iron Concentration Grading in Neurodegeneration

Xiaomin Liu[1,†], Bo Yu[2,†], Daniëlle Toen[3], Louise van der Weerd[3,4], Jouke Dijkstra[3], Yunke Zhang[1], Jiuman Song[1,*], and Hechang Chen[1,*]
[1]School of Artificial Intelligence, Jilin University, Changchun 130015, China
[2]Xiaomi Laboratory of Sports and Health, Xiaomi Inc, Beijing, 100085, China
[3]Dept Radiology, Leiden University Medical Center, Leiden, Netherlands
[4]Dept Human Genetics, Leiden University Medical Center, Leiden, Netherlands

***Abstract*—Iron accumulation in cerebral gray matter is a critical research focus in neurodegenerative diseases. Whole-slide images (WSIs) offer new opportunities for iron grading, yet manual WSI scoring remains subjective and prone to bias. Developing objective deep learning models for iron concentration grading faces two key challenges. First, natural tissue heterogeneity causes stained white matter to appear darker than gray matter, leading to the misclassification of normal white matter as iron-rich gray matter. Second, non-uniform iron deposition across cortical layers makes it difficult to capture global distribution patterns beyond sparse local cues, often biasing predictions toward a few high response patches. To address these limitations, we propose an Adaptive Regional Feature Fusion (ARFF) framework. ARFF employs a discriminative region mask extraction module to isolate gray matter and suppress structural and background noise. It then leverages cluster-based adaptive feature sampling and fusion weights learned via proximal policy optimization (PPO) to construct a comprehensive WSI-level representation. Experiments show that ARFF outperforms state-of-the-art methods, delivering average AUC gains of 9.5% on the Frontotemporal Dementia dataset and 16.4% on the Alzheimer's Disease dataset, thereby helping reduce subjectivity in neuropathological assessment. Code: https://github.com/xiaominLiu2001/ARFF.**

***Index Terms*—Whole-slide images, Image classification, Iron concentration grading, Neurodegenerative disease.**

I. Introduction

Neurodegenerative diseases such as Alzheimer's disease (AD) impose a major burden on patients and healthcare systems [1]. Recent studies show that abnormal brain iron metabolism is closely associated with cognitive decline [2], making accurate assessment of brain iron deposition essential for elucidating the pathological mechanisms of neurodegenerative diseases. However, existing approaches for assessing cerebral iron levels remain limited [3], [4]. With advances in digital pathology, whole-slide images (WSIs) enable high-resolution analysis of iron-stained brain tissues. Nevertheless, manual WSI scoring is subjective and prone to observer bias, thereby necessitating the development of objective and efficient iron grading methods for WSIs.

† Xiaomin Liu and Bo Yu contributed equally to this work. Email: liuxm23@mails.jlu.edu.cn, yubo10@xiaomi.com.
* Corresponding authors: Jiuman Song and Hechang Chen. Email: songjw22@mails.jlu.edu.cn, chenhc@jlu.edu.cn.

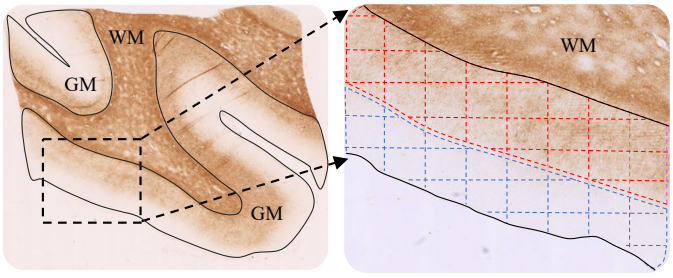


Fig. 1. WSI case of MIL overestimation. Dominant high-iron patches in inner GM lead to overlooked low-iron regions and label overestimation.

Multiple instance learning (MIL) has become a mainstream weakly supervised paradigm for WSI analysis, as it only requires slide-level labels [5]. Early MIL methods aggregate instance features or predictions into bag-level representations using pooling or fusion before performing the final classification [6], which may provide limited interpretability and weak localization of diseased regions in WSIs. To alleviate these issues, attention-based MIL models assign learnable weights to highlight discriminative instances and improve bag aggregation [7], [8]. In addition, advances such as multi-scale feature modeling and self-supervised pre-training have further advanced MIL-based WSI classification [9], [10].

Despite the progress made in MIL-based WSI analysis, using brain WSIs for iron grading presents two specific challenges. ***Challenge I*** is how to suppress interference from irrelevant regions. Due to intrinsic tissue heterogeneity, stained white matter (WM) often appears darker than gray matter (GM) [11], causing MIL models to misclassify normal WM as iron-rich GM and thus bias the grading results. ***Challenge II*** is how to effectively fuse global features. Iron deposition varies across cortical layers, leading to non-uniform iron spatial distributions in GM [12]. Many existing MIL models predominantly focus on a limited subset of high-attention patches during the aggregation stage [13], which may lead to overestimation of iron severity by neglecting widespread regions with medium- and low-concentration levels (Fig. 1).

In view of this, we propose an Adaptive Regional Feature Fusion (ARFF) framework for objective prediction of iron

concentration grades from brain WSIs. For ***Challenge I***, ARFF employs a discriminative region mask extraction module to extract GM masks while filtering structural and background noise. By narrowing the scope of the analysis from entire slides to GM regions, our approach aligns closely with manual visual scoring practices. For ***Challenge II***, ARFF implements a cluster-based adaptive sampling and a PPO-based adaptive fusion strategy. It groups GM patches by iron deposition feature similarity and samples from all clusters to preserve global spatial distribution while reducing redundancy. Subsequently, we frame feature fusion as a reinforcement learning (RL) task, employing proximal policy optimization (PPO) [14] to learn optimal fusion weights across sampled clusters. This ensures a robust WSI-level representation that accounts for varying iron accumulation levels. In summary, our main contributions are:

- We propose ARFF, a two-stage framework for predicting brain iron concentration grades from WSIs, following a "locate-then-grade" strategy aligned with expert practice.
- To suppress structural and background interference, we train a discriminative region mask extraction module that generates precise GM masks.
- We design an cluster-based adaptive feature sampling module and a PPO-based adaptive feature fusion module to capture comprehensive WSI-level global representations of iron concentration.

II. Related Works

A. Iron Assessment

Brain iron is commonly assessed via biological assays or imaging techniques [15]. Biological measurements (e.g., ferritin or transferrin) provide indirect indicators of iron metabolism but lack regional specificity [16]–[18]. Imaging approaches such as quantitative susceptibility mapping (QSM) offer superior spatial resolution for mapping iron distribution, while ultra-high-field MRI (e.g., 7T) further enhances sensitivity to subtle iron-induced magnetic changes [19]. However, MRI is vulnerable to motion artifacts, which may compromise quantification accuracy [20]. Given this, histological staining, which encompasses ferritin binding and other iron associated forms, provides a reliable way to characterize iron accumulation in tissue [11]. This motivates our use of histologically stained high-resolution WSIs for regional iron analysis.

B. Multiple Instance Learning

MIL enables weakly supervised training by assigning labels at the bag level while leaving instances unlabeled, a paradigm particularly suited for histopathological analysis where fine-grained annotation is impractical [21], [22]. In WSI analysis, slides are split into instance patches whose features are aggregated into bag-level representations for final slide-level predictions [23], [24]. Advanced MIL variants, such as CLAM and DSMIL, utilize instance losses or dual-stream architectures to constrain attention distributions [25]–[27]. However, these models often prioritize local regions with the highest attention scores, which may lead to under-modeling the global spatial heterogeneity. To address this, we perform cluster-based sampling to obtain all iron-level cluster-specific sampled features and then fuse them for the final prediction.

C. Proximal Policy Optimization

RL trains an agent to observe environment states, take actions under a policy, and maximize cumulative reward through interaction. PPO is an on-policy RL method that stabilizes training by limiting policy changes via a clipped surrogate objective [14]. The clipping constrains the ratio between new and old action probabilities, avoiding overly large updates during gradient ascent [28]. In ARFF, we treat fusion weight selection as the action, fusion features as the state, and prediction improvement as the reward, allowing PPO to learn adaptive fusion weights that better integrate heterogeneous regional cues for global iron grading.

III. Methods

This section details the ARFF framework (Fig. 2), introducing three core modules.

A. Discriminative Region Mask Extraction

In brain WSIs, iron distribution reflects both pathological progression and normal anatomical variations, the latter causing heterogeneous staining patterns that must be distinguished from abnormal iron accumulation. Additionally, scanner artifacts and irrelevant regions further complicate GM localization, posing a challenge for accurate iron grading.

To address this, we built a discriminative region mask extraction module based on DeepLabV3 [29], leveraging its multi-scale feature extraction through Atrous Spatial Pyramid Pooling (ASPP) to generate GM masks from WSIs, as depicted in Fig. 2(b). The input WSI is denoted as $I \in R^{H\times W\times 3}$, and binary mask M_{gray} represent the ground-truth mask label.

1) Feature Encoding and Dilated Convolutions: We adopt ResNet-50 (V1c) [30] as the feature encoder f_{enc}, which outputs four hierarchical features $\{F_{\text{I1}}, F_{\text{I2}}, F_{\text{I3}}, F_{\text{I4}}\}$ from four residual stages [31]. To enhance multi-scale perception while preserving spatial detail, we apply dilated convolutions with increasing dilation rates of 2 and 4 in the third and fourth stages, respectively. These two stage-wise features $\{F_{\text{I3}}, F_{\text{I4}}\}$ are propagated to downstream modules for further processing.

2) Multi-scale Context Aggregation via ASPP: To capture both global context and fine boundaries, we apply an ASPP module f_{ASPP} on F_{I4}. As shown in Fig. 3, f_{ASPP} consists of five parallel branches. The outputs of all branches are concatenated and then compressed by a 1×1 convolution to form feature F_{ASPP}, which is then fed into a classification layer f_{seg} to produce the final GM mask prediction $\hat{M}_{gray}$.

3) Auxiliary Supervision: To improve gradient flow and leverage mid-level semantics, we introduce an auxiliary prediction head f_{FCN} on F_{I3}, followed by the same f_{seg} to produce an auxiliary GM mask prediction $\hat{M}_{aux}$ (Fig. 3). $\hat{M}_{aux}$ is solely utilized to facilitate the training process:

$$L_{mask} = \mathcal{L}_{CE}(\hat{M}_{gray}, M_{gray}) + a\mathcal{L}_{CE}(\hat{M}_{aux}, M_{gray}) \quad (1)$$

where $\mathcal{L}_{CE}$ is cross-entropy loss and a is the auxiliary weight.

This module outputs GM masks $\hat{M}_{gray}$ that pinpoint discriminative regions, suppressing noise, isolating GM from adjacent tissues, and easing computational burden.

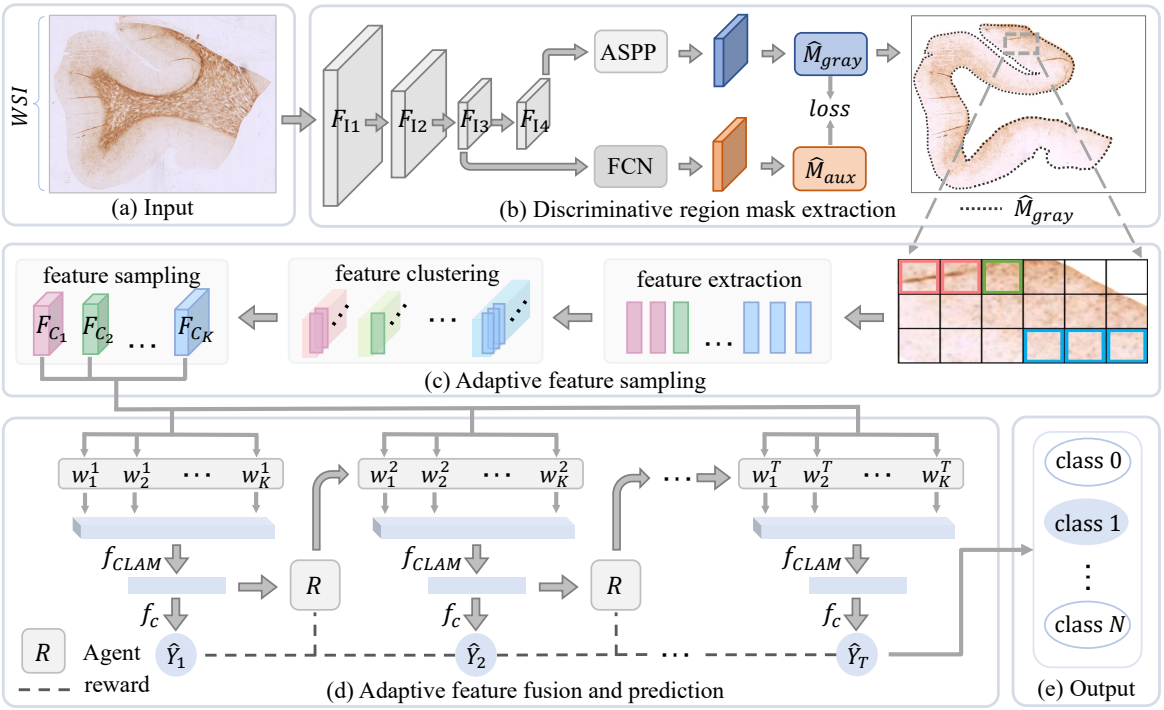


Fig. 2. Overview of the ARFF framework. (a) Receives high-resolution WSIs; (b) Extracts GM masks using DeeplabV3; (c) Cuts and processes patches via feature extraction, clustering, and adaptive sampling; (d) Over T iterative steps, generates global fused features via weighted fusion to predict iron concentration grades, guiding agent R to compute fusion weights for the next step; (e) outputs final predictions by selecting the highest probability.

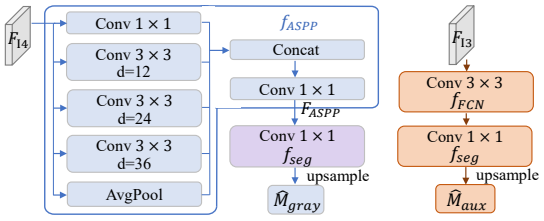


Fig. 3. ASPP module and auxiliary head in DeepLabV3.

B. Cluster-based Adaptive Feature Sampling

To mitigate information redundancy and the computational burden of processing numerous patches in large-scale WSIs, we propose a cluster-based adaptive feature sampling strategy.

1) Feature Extraction and Clustering: In Fig. 2(c), the WSI is partitioned into $P\times P$ non-overlapping patches. Given the predicted GM mask $\hat{M}_{gray}$, we construct $P_{gray}=\{p_1,\ldots,p_m\}$ by selecting patches whose GM occupancy ratio exceeds threshold τ. Each patch p_i is mapped to a feature $f_i\in\mathbb{R}^{2048}$ via a pre-trained ResNet-50, forming the feature matrix $F=[f_1,\ldots,f_m]$. Subsequently, K-means clustering [32] partitions F into K clusters $C=\{C_1,\ldots,C_K\}$, where $C_k=\{i|\text{cluster}(f_i)=k\}$ groups patches with similar iron levels to provide an unsupervised structure for subsequent fusion.

2) Adaptive Sampling: Since clusters vary in size and importance to WSI-level iron concentration grades, we employ an adaptive sampling strategy. For each cluster C_k, the sampling length l_k is dynamically adjusted based on its relative size:

$$l_k = \frac{|C_k|}{\sum_{j=1}^{K}|C_j|}\times L \tag{2}$$

where L is a predetermined global sampling length. We then uniformly sample l_k indices from C_k to form S_k and construct cluster-specific sampled features $F_{C_k}=\text{concat}(\{f_t \mid t\in S_k\})$. With $l_k\propto|C_k|$, we retain all iron-level clusters and approximately preserve their relative prevalence, while reducing patches from m to L to lower computation.

C. PPO-based Adaptive Feature Fusion

The variations in iron distribution and diagnostic relevance among clusters complicate the effective fusion of sampled features $\{F_{C_k}\}_{k=1}^{K}$. We therefore propose an adaptive fusion strategy using PPO, where the determination of cluster-specific fusion weights is modeled as a sequential decision-making process. This RL-driven framework adaptively adjusts weights through multi-step interactions, ensuring the resulting global fused features faithfully reflect the overall brain iron grades.

1) Fusion and Prediction Process: As illustrated in Fig. 2(d), for a single sample, we define an episode with length T. At the initial time step $t=1$, uniform weights $w_k^1=1/K$ are applied to obtain the initial global fused feature F_{agg}^1. In subsequent time steps, the PPO policy generates weights to iteratively update the fusion and prediction. The state s_t of the agent is defined as the global fused feature from the previous step F_{agg}^{t-1}. The action a_t is the fusion weight vector $w_t=\{w_k^t\}_{k=1}^{K}$. The weighted fusion formula is expressed as:

$$F_{agg}^t, L_{ins}^t = f_{CLAM}\left(\text{concat}(w_k^t F_{C_k})_{k=1}^{K}\right) \tag{3}$$

where $f_{CLAM}(\cdot)$ denotes the CLAM feature aggregator [25]. This allows weaker clusters to be amplified during the

weighted fusion stage before being refined by the attention mechanism, thereby reducing the attention bias. L_{ins}^{t} is instance-level loss. Subsequently, the global fused feature F_{agg}^{t} is fed into a classification head $f_c(\cdot)$ to obtain the classification logits $\hat{Y}_t$=softmax$(f_c(F_{agg}^{t}))$. At the terminal step T, the final iron grade label is predicted as $\hat{y}$=argmax$\hat{Y}_T$.

2) Agent Architecture: We use a GRU Actor-Critic architecture. Old policy $\pi_{\theta_{old}}$ collects trajectories, while target policy π_θ is updated offline. As shown in Fig. 4, at time step t, state s_t=F_{agg}^{t-1} is processed through an encoder $f_s(\cdot)$ and a GRU unit $f_{gru}(\cdot)$ to produce hidden state h_t. An actor network $f_{actor}(\cdot)$ outputs the mean action μ_t based on h_t. During *sampling*, action w_t is sampled from an isotropic Gaussian $\pi_{\theta_{old}}(\cdot|s_t)$=$\mathcal{N}(\mu_t, \sigma^2\mathbf{I})$ with fixed standard deviation σ. We store log-probability $\log \pi_{\theta_{old}}(w_t|s_t)$ for PPO updates. During *updating*, π_θ feeds h_t into the critic network $f_{critic}(\cdot)$ to obtain the value estimate $V_\phi(s_t)$, and the probability ratio is:

$$\rho_t = \frac{\pi_\theta(w_t|s_t)}{\pi_{\theta_{old}}(w_t|s_t)} = exp(\log \pi_\theta(w_t|s_t) - \log \pi_{\theta_{old}}(w_t|s_t)) \tag{4}$$

3) Reward: The reward r_t=$\hat{Y}_t[y] - \hat{Y}_{t-1}[y]$ is positive only when the agent increases the prediction probability for the ground-truth label y. After an episode concludes, the discounted return $\hat{G}_t$ and the advantage function A_t are calculated based on the reward sequence $\{r_t\}_{t=2}^{T}$:

$$\hat{G}_t = \sum_{l=0}^{T-t} \gamma^l r_{t+l}, \quad A_t = \hat{G}_t - V_\phi(s_t) \tag{5}$$

where γ is the discount factor.

4) Two-stage Training Strategy: (1) Pre-training. We train the backbone $f_{\text{CLAM}}(\cdot)$ and $f_c(\cdot)$ with uniform weights ($w_k = 1/K$) for T steps to minimize the joint loss:

$$L_{iron} = \frac{1}{T}\sum_{t=1}^{T}\left[\lambda L_{CE}(\hat{Y}_t, y) + (1-\lambda)L_{ins}^{t}\right] \tag{6}$$

where λ balances global and instance-level tasks. (2) Agent Training. After pre-training, we freeze $f_{\text{CLAM}}(\cdot)$ and $f_c(\cdot)$. The PPO agent is optimized via iterative *sampling* and *updating*. In *sampling*, $\pi_{\theta_{old}}$ executes T steps: t=1 sets a uniform weight baseline, while for $t\in[2,T]$, the agent samples w_t from s_t to collect trajectories $\{s_t, w_t, r_t, \log \pi_{\theta_{old}}\}$. In *updating*, π_θ undergoes S rounds of optimization rounds to minimize:

$$\begin{aligned} L_{\text{PPO}} = &- \min\left(\rho_t A_t, \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon)A_t\right) \\ &+ c_1\|V_\phi(s_t) - \hat{G}_t\|_2^2 - c_2\mathcal{H}(\pi_\theta) \end{aligned} \tag{7}$$

where ϵ is the clipping range, and c_1, c_2 are coefficients for value loss and entropy regularization. After each *update* phase, the parameters of $\pi_{\theta_{old}}$ are replaced by π_θ for the next episode.

IV. Experiments and Results

A. Dataset

We conducted experiments on two real-world WSI datasets from Dutch brain donation programs: FTD-LUMC and AD-LUMC. All procedures followed Dutch ethical guidelines

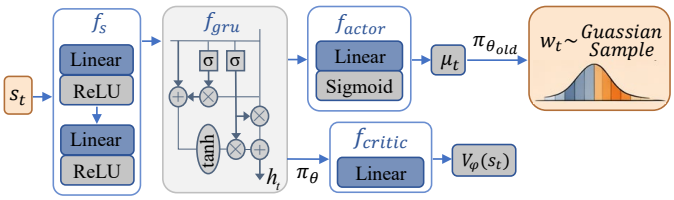


Fig. 4. Architecture of the GRU-based Actor-Critic agent.

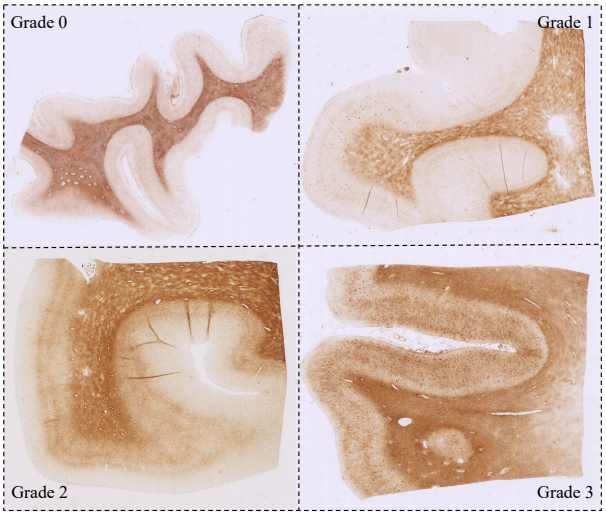


Fig. 5. Representative WSIs with four grades of cortical iron concentration.

(Code for Proper Secondary Use of Human Tissue, Dutch Federation of Medical Scientific Societies).

FTD-LUMC includes cases of Frontotemporal Dementia (FTD), where iron accumulation typically begins in the mid-cortical layers with a diffuse pattern, sometimes extending to the outer layers [33]. The dataset comprises 54 WSIs (average matrix size: 119,296×76,402 pixels), with 10, 18, 12, and 14 images for grades 0–3, respectively.

AD-LUMC contains samples from Alzheimer's Disease (AD) patients. In contrast to FTD, iron deposition in AD progresses from the mid-cortical to inner and finally outer layers, occasionally co-localizing with amyloid plaques. The dataset contains 87 WSIs (avg. matrix size: 107,302×66,189 pixels), with counts by grade of 16/21/31/19 for grades 0–3.

All tissue blocks were processed using a standardized protocol and stained with the Leiden Iron staining method, a modified DAB-enhanced Perl's technique [11]. Following prior neuropathological studies [2], [33], [34], cortical iron accumulation was scored using a four-grade scale (0–3), with representative examples shown in Fig. 5. While this grading correlates with disease severity, the relationship is not strictly linear. It serves as an objective pathological tool and should be interpreted alongside other biomarkers and clinical indicators for comprehensive disease assessment.

The four-grade scoring system are detailed as follows:

• grade 0: homogeneous gray matter structure with minimal iron deposition, typically observed in healthy controls;

• grade 1: localized iron-rich bands in the mid-cortical layers, exceeding normal anatomical variation;

TABLE I

PERFORMANCE ON FTD-LUMC AND AD-LUMC DATASETS. RESULTS ARE REPORTED AS THE MEAN OF THREE INDEPENDENT RUNS WITH DIFFERENT RANDOM SEEDS. THE BEST AND SECOND-BEST RESULTS IN EACH COLUMN ARE HIGHLIGHTED IN **BOLDFACE** AND UNDERLINED, RESPECTIVELY.

Model	FTD-LUMC						AD-LUMC					
	ACC↑	AUC↑	P↑	R↑	F1↑	Time↓	ACC↑	AUC↑	P↑	R↑	F1↑	Time↓
ABMIL	0.5112	0.8155	0.4042	0.4444	0.3735	0.0362	0.4706	0.6903	0.4872	0.4097	0.3947	0.0333
MSBP	0.6970	0.8209	0.7917	0.7570	0.7111	0.0958	0.4510	0.6395	0.3343	0.3680	0.3091	0.0766
DSMIL	0.5455	0.9167	0.7250	0.6458	0.5345	0.0209	0.5686	0.7212	0.6917	0.6250	0.5993	0.0190
CLAM-SB	0.8788	0.9514	0.9061	0.8788	0.8663	0.0471	0.5294	0.8474	0.6562	0.5294	0.5477	0.0309
CLAM-MB	**0.9091**	0.9688	**0.9495**	0.9091	**0.9111**	0.0485	0.6275	0.8162	0.6248	0.6275	0.5836	0.0329
DTFD-MIL	0.6364	0.8608	0.6597	0.6250	0.6118	**0.0124**	0.6275	0.9081	0.4770	0.5555	0.5102	**0.0108**
MS-DA-MIL	0.8485	0.9494	0.7445	0.7917	0.7537	0.4509	0.6667	0.8117	0.4724	0.5833	0.5057	0.4338
MHIM-MIL	0.7273	0.9560	0.6028	0.6250	0.5794	0.0394	0.7843	0.9157	0.7520	0.7222	0.7026	0.0202
HDMIL	**0.9091**	0.8993	0.9375	0.8750	0.8810	0.0719	0.6471	0.9204	0.5357	0.6458	0.5566	0.0482
ARFF (Ours)	**0.9091**	**0.9866**	0.9278	**0.9167**	0.9074	0.0338	**0.8824**	**0.9257**	**0.8992**	**0.8819**	**0.8731**	0.0285

• grade 2: iron deposition bands extend to the deep gray matter layers, while outer layers remain unaffected;

• grade 3: pronounced iron deposition across all layers, characterized by widened iron-rich bands in the mid-cortical layers, representing a severe form of cortical iron accumulation.

B. Experiment Settings

All experiments ran on an NVIDIA L40 GPU using PyTorch 1.13.1 and Python 3.7. For all datasets, a random split at a 4:1 ratio was applied at the patient level for training and testing to ensure no data leakage. Evaluation used five standard metrics: accuracy (ACC), macro one-vs-rest area under the curve (AUC), precision (P), recall (R), and F1-score (F1).

In the discriminative region mask extraction module, we employed the AdamW optimizer with a linear warm-up schedule. The learning rate was warmed up from 1.2×10^{-10} to a base value of 1.2×10^{-6}. The batch size was set to 1, and α=0.4.

For the adaptive feature sampling and fusion module, we used the Adam optimizer with a batch size of 1. The learning rates were set to $5{\times}10^{-5}$ for $f_{\mathrm{CLAM}}(\cdot)$ and $2{\times}10^{-5}$ for $f_c(\cdot)$. We used cosine annealing for learning rate scheduling. Hyperparameters were set to P=256, τ=50%, K=6, L=1024, T=6, σ=0.5, γ=0.2, λ=0.7, S=3, ϵ=0.2, c_1=0.5, and c_2=0.01.

C. Performance Comparison

Baselines include attention-based MIL models (ABMIL [13], DSMIL [26], CLAM [25]), multi-scale models (MSBP [35], HDMIL [36]), and robustness-oriented approaches (DTFD-MIL [37], MS-DA-MIL [38], MHIM-MIL [39]). Given the scarcity of brain WSIs, experiments were run three times with different random seeds to ensure reliability.

From Table I, the benchmarks exhibit performance gaps. Note that ACC reflects strict top-1 grade correctness, whereas AUC evaluates the ranking quality of predicted probabilities in a threshold-free manner; thus, models can show similar AUC yet differ in ACC when predictions are close between adjacent grades. Most methods achieve under 80% ACC on both datasets, with notable variance between FTD-LUMC and AD-LUMC. This is largely due to two factors: (1) iron concentration grading depends on assessing global GM distribution rather than isolated discriminative regions; (2) FTD and AD show distinct deposition patterns, as FTD shows layer-wise diffusion while AD may have amyloid plaque co-localization, making generalization across datasets challenging.

ABMIL and DSMIL underperform on both datasets, and MSBP's binary processing may oversimplify the task, leading to low performance on AD-LUMC. DTFD-MIL and MHIM-MIL rely on pseudo-bag sampling and instance masking, yet their 60–80% accuracy suggests that randomness adds robustness but may discard critical information. CLAM-SB/MB, MS-DA-MIL, and HDMIL drop sharply on AD-LUMC, reflecting poor adaptability to irregular spot distributions. In contrast, by using adaptive regional feature fusion across concentration gradients, ARFF reduces intermediate uncertainty and preserves critical cues, achieving 90.91% and 88.24% accuracy on FTD-LUMC and AD-LUMC.

For fairness, all methods use the same predicted GM masks, and Table I reports the per-WSI inference time excluding GM mask extraction (see Table II). DTFD-MIL is consistently the fastest and DSMIL/MHIM-MIL are also efficient, likely because they rely on lightweight instance selection/masking. ARFF does not achieve the minimum latency, however, its runtime remains comparable to mainstream attention baselines (ABMIL/CLAM) and is substantially lower than heavier structure-aware designs (MS-DA-MIL). Overall, ARFF offers the best accuracy with a moderate and practical inference cost.

D. Ablation Studies

1) Discriminative Region Mask Extraction Module: This module primarily aims to restrict analysis to GM regions to match manual scoring and reduce computation by narrowing down the analysis area. To evaluate whether it meets these two objectives, we report segmentation accuracy using the mean Intersection-over-Union (mIoU) and the GM area ratio. The module achieves 96.09% and 95.96% mIoU on FTD-LUMC and AD-LUMC, respectively. We also compare the average GM proportion in WSIs between manual annotations (R_T) and predictions (R_P), which are highly consistent (R_T/R_P: 0.2844/0.2851 for FTD-LUMC and 0.2660/0.2695 for AD-LUMC). Moreover, the exclusion of non-GM regions ($1{-}R_P$) translates to a substantial reduction in downstream computational workload by 71.49% and 73.05% for the two cohorts.

We further ablate receptive-field design (dilation rates in f_{enc} and f_{ASPP}) and auxiliary supervision (presence and

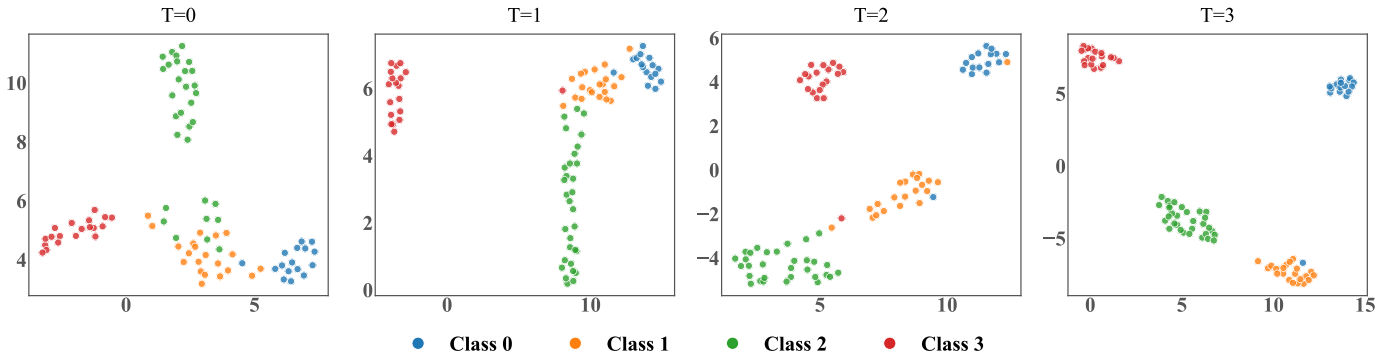


Fig. 6. Umap of global fused features over four time steps in PPO on AD-LUMC. The x- and y-axes represent the first and second principal components, respectively. Each point corresponds to an individual sample, with colors indicating different grades.

TABLE II
ABLATION RESULTS FOR THE DISCRIMINATIVE REGION MASK EXTRACTION MODULE.

Method	FTD-LUMC		AD-LUMC		Avg	
	Dice↑	Time↓	Dice↑	Time↓	Dice↑	Time↓
M_{ND}	0.9526	2.9752	0.9647	2.5833	0.9587	2.7793
$M_{A\text{-}Std}$	0.9500	3.0668	0.9634	2.6525	0.9567	2.8597
M_{NA}	0.9517	3.0724	0.9595	2.6167	0.9556	2.8446
$M_{AW0.1}$	0.9549	3.0305	**0.9677**	2.6100	0.9613	2.8203
$M_{AW0.7}$	0.9554	3.0260	0.9645	2.6619	0.9599	2.8440
M_{Ours}	**0.9558**	**2.9409**	**0.9677**	**2.5492**	**0.9618**	**2.7451**

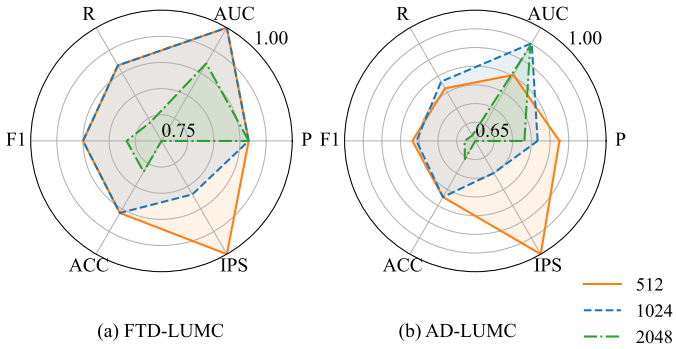
aux loss weight α) in Table II. Disabling dilation in f_{enc} (M_{ND}: (1,1,1,1)) or using narrower ASPP rates ($M_{A\text{-}Std}$: (1,6,12,18)) reduces the Dice similarity coefficient (Dice) score, indicating the need for multi-scale context. Removing the auxiliary branch (M_{NA}) also reduces performance, while overly small/large auxiliary weights ($M_{AW0.1}$: α=0.1 / $M_{AW0.7}$: α=0.7) are suboptimal. Our default configuration (dilation rates=(1,1,2,4) in f_{enc} and (1,12,24,36) in f_{ASPP}, α=0.4) yields the best Dice with competitive runtime.

2) Adaptive Feature Sampling Module and Adaptive Feature Fusion Module: To assess the contributions of these two modules, we evaluate four variants: baseline (w/o SP & PPO), adaptive sampling only (w/o PPO, K=6), PPO-based adaptive fusion only (w/o SP, K=1), and the full ARFF. For variants without PPO, we use only the first time step. According to Table III, sampling alone (w/o PPO) improves FTD-LUMC ACC from 0.6364 to 0.7273 with negligible time overhead (0.0320 vs. 0.0323), while AD-LUMC shows slight gains (AUC/R/F1) with similarly marginal latency change (0.0267 vs. 0.0274). In contrast, PPO-based feature fusion brings larger performance gains (0.8182/0.7059 ACC) but also accounts for most of the runtime increase (to 0.0334/0.0280), indicating that the PPO agent dominates the extra computation. Nevertheless, this module is essential given its strong accuracy improvement, and combining it with SP yields the best results (0.9091/0.8235 ACC) at small extra time cost (0.0338/0.0285), suggesting a synergistic effect where cluster-based feature sampling captures prior structural information, while PPO-based feature fusion integrates it more effectively.

To further evaluate the PPO-based adaptive feature fusion module, we visualized the global fused features at t=0-3 using UMAP (Fig. 6). At t=0, different grade classes partially overlap. As PPO optimization proceeds (t=1-2), each grade forms increasingly distinct clusters. By t=3, four clearly separated clusters emerge, indicating that this module enhances the discriminability of global fused representations for classification.

E. Parameter Sensitivity Analysis

We study the effect of global sampling length L using $L\in\{512, 1024, 2048\}$, and report inference efficiency via inferences per second (IPS), defined as IPS=$1/t$. For radar plots, IPS is min–max normalized and linearly rescaled to the radius range. From Fig. 7, L=512 and L=1024 outperform L=2048, indicating that longer global sampling length does not necessarily improve performance. An appropriate L can maintain accuracy while reducing computational overhead.



(a) FTD-LUMC (b) AD-LUMC

Fig. 7. Effect of the global sampling length on ARFF performance.

We investigate the impact of the number of clusters by setting $K\in\{1, 2, 4, 6, 8\}$, where K=1 denotes no clustering. As illustrated in Fig. 8, too few clusters (K=1) yield the weakest performance, while K=6 provides the most balanced results across datasets. The optimal K also depends on pathology and data characteristics. FTD-LUMC is less sensitive to K, likely due to its smaller size and simpler iron deposition patterns. In contrast, AD-LUMC improves from K=4 to K=6, consistent with its more complex deposition patterns. However, overly fine clustering (K=8) may over-segment features and slightly degrade performance. Inference efficiency is largely unaffected

TABLE III
ABLATION STUDY RESULTS OF ADAPTIVE FEATURE SAMPLING MODULE (SP) AND ADAPTIVE FEATURE FUSION MODULE (PPO) IN ARFF.

Methods	FTD-LUMC						AD-LUMC					
	ACC↑	AUC↑	P↑	R↑	F1↑	Time↓	ACC↑	AUC↑	P↑	R↑	F1↑	Time↓
(w/o) SP & PPO	0.6364	0.9479	0.6250	0.5417	0.5333	**0.0320**	0.6471	0.7918	0.6813	0.6458	0.6161	**0.0267**
(w/o) PPO	0.7273	0.9162	0.7917	0.6667	0.6917	0.0323	0.6471	0.8015	0.6500	0.6667	0.6399	0.0274
(w/o) SP	0.8182	0.9653	0.8333	0.7917	0.7833	0.0334	0.7059	**0.8596**	**0.8864**	0.6667	0.6908	0.0280
ARFF (Ours)	**0.9091**	**1.0000**	**0.9167**	**0.9167**	**0.9000**	0.0338	**0.8235**	0.8540	0.8167	**0.8333**	**0.8076**	0.0285

by K: IPS remains at 35.09–35.71 s^{-1} on FTD-LUMC and 29.50–29.94 s^{-1} on AD-LUMC (within 2% variation), indicating that increasing K mainly impacts discriminability with negligible runtime cost.

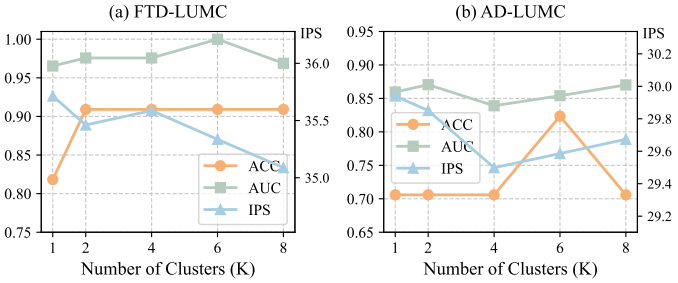


Fig. 8. Effect of the number of clusters on ARFF performance.

F. Visual Analysis

Qualitative analysis of three representative cases (Fig. 9) highlights ARFF's strengths and limitations. In Case A (Grade 0), iron deposition within the GM is overall low and relatively homogeneous. Although the attention heatmap exhibits higher responses in locally iron-enriched regions, when the visual differences are subtle, the attention mechanism may amplify the contribution of these high-response areas, causing baselines (e.g., CLAM-MB, HDMIL and ABMIL) to misclassify the case as Grade 2-3. In contrast, ARFF integrates all clusters and employs an agent to produce relatively balanced fusion weights based on similar sampled features. This design allows low attention regions to effectively contribute to the global prediction, thereby alleviating attention bias.

Case B emphasizes the impact of image quality. This slide contains very thin WM regions that appear markedly white after staining (Fig. 9-B2, black dashed box). Although ARFF does not misclassify these areas as GM, such tissue preparation artifacts can distort tissue appearance and perturb the estimation of iron distribution. Case C illustrates the performance boundary of the discriminative region mask extraction module. When severe iron accumulation obscures the GM-WM boundary, or when low iron GM in healthy controls resembles the background, GM mask extraction becomes challenging. In the future, incorporating the assessment of specific iron distribution patterns, such as diffuse mid-cortical iron bands [33], may provide additional diagnostic cues, reduce reliance on intact GM, and enhance diagnostic robustness.

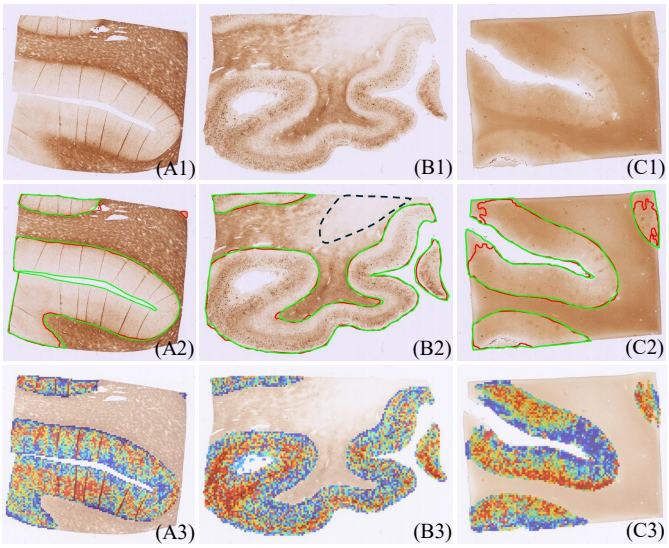


Fig. 9. Visualization of the ARFF applied to three representative cases. Row 1 shows original WSIs. Row 2 shows segmentation from the discriminative region mask extraction module (green: ground truth; red: prediction). Row 3 shows attention heatmaps from f_{CLAM} (red: high attention; blue: low).

V. CONCLUSION

In this paper, we propose an Adaptive Regional Feature Fusion (ARFF) framework for objective grading of iron concentration from brain WSIs. ARFF first extracts discriminative GM masks to suppress interference from WM and background, then performs cluster-based adaptive sampling and autonomously explores fusion weights via a PPO agent to model global iron distribution patterns. Comprehensive experiments on FTD-LUMC and AD-LUMC demonstrate that ARFF yields robust improvements across datasets with different pathological characteristics. The proposed framework may facilitate research on neurodegenerative diseases and provide auxiliary evidence for clinical decision-making. Nevertheless, ARFF remains sensitive to segmentation quality and GM integrity. In future work, we plan to assess specific iron distribution patterns, such as diffuse mid-cortical iron bands [33]. Integrating such features could offer additional diagnostic markers, reduce reliance on GM integrity, and improve diagnostic robustness.

VI. ACKNOWLEDGEMENTS

This work is supported in part by the Natural Science Foundation of Jilin Province (No. YDZJ202401184ZYTS).

REFERENCES

[1] M. A. Better, "Alzheimer's disease facts and figures," *Alzheimers Dement*, vol. 19, no. 4, pp. 1598–1695, 2023.

[2] M. Bulk, B. Kenkhuis, L. M. van der Graaf, J. J. Goeman, R. Natté, and L. van der Weerd, "Postmortem t 2*-weighted mri imaging of cortical iron reflects severity of alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 65, no. 4, pp. 1125–1137, 2018.

[3] S. Ayton, S. Janelidze, P. Kalinowski, S. Palmqvist, A. A. Belaidi, E. Stomrud, A. Roberts, B. Roberts, O. Hansson, and A. I. Bush, "Csf ferritin in the clinicopathological progression of alzheimer's disease and associations with apoe and inflammation biomarkers," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 94, no. 3, pp. 211–219, 2023.

[4] A. Yang, L. Du, W. Gao, B. Liu, Y. Chen, Y. Wang, X. Liu, K. Lv, W. Zhang, H. Xia *et al.*, "Associations of cortical iron accumulation with cognition and cerebral atrophy in alzheimer's disease," *Quantitative Imaging in Medicine and Surgery*, vol. 12, no. 9, p. 4570, 2022.

[5] S. Kapse, P. Pati, S. Das, J. Zhang, C. Chen, M. Vakalopoulou, J. Saltz, D. Samaras, R. R. Gupta, and P. Prasanna, "Si-mil: Taming deep mil for self-interpretability in gigapixel histopathology," in *Proc. of the 41st IEEE/CVF CVPR*, 2024, pp. 11 226–11 237.

[6] A. Obeid, S. Boumaraf, A. Sohail, T. Hassan, S. Javed, J. Dias, M. Bennamoun, and N. Werghi, "Advancing histopathology with deep learning under data scarcity: A decade in review," *arXiv preprint arXiv:2410.19820*, 2024.

[7] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji *et al.*, "Transmil: Transformer based correlated multiple instance learning for whole slide image classification," *Advances in neural information processing systems*, vol. 34, pp. 2136–2147, 2021.

[8] M. Waqas, M. A. Tahir, S. Al-Maadeed, A. Bouridane, and J. Wu, "Simultaneous instance pooling and bag representation selection approach for multiple-instance learning (mil) using vision transformer," *Neural Computing and Applications*, vol. 36, no. 12, pp. 6659–6680, 2024.

[9] G. Bontempo, A. Porrello, F. Bolelli, S. Calderara, and E. Ficarra, "Dasmil: distilling across scales for mil classification of histological wsis," in *Proc. of 26th MICCAI*, 2023, pp. 248–258.

[10] Y. Schirris, E. Gavves, I. Nederlof, H. M. Horlings, and J. Teuwen, "Deepsmile: Contrastive self-supervised pre-training benefits msi and hrd classification directly from h&e whole-slide images in colorectal and breast cancer," *Medical Image Analysis*, vol. 79, p. 102464, 2022.

[11] S. van Duijn, R. J. Nabuurs, S. G. van Duinen, and R. Natté, "Comparison of histological techniques to visualize iron in paraffin-embedded brain tissue of patients with alzheimer's disease," *Journal of Histochemistry & Cytochemistry*, vol. 61, no. 11, pp. 785–792, 2013.

[12] M. Fukunaga, T.-Q. Li, P. Van Gelderen, J. A. De Zwart, K. Shmueli, B. Yao, J. Lee, D. Maric, M. A. Aronova, G. Zhang *et al.*, "Layer-specific variation of iron content in cerebral cortex as a source of mri contrast," *Proceedings of the National Academy of Sciences*, vol. 107, no. 8, pp. 3834–3839, 2010.

[13] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *Proc. of the 35th ICML*, 2018, pp. 2127–2136.

[14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[15] D. Tran, P. DiGiacomo, D. E. Born, M. Georgiadis, and M. Zeineh, "Iron and alzheimer's disease: from pathology to imaging," *Frontiers in Human Neuroscience*, vol. 16, p. 838692, 2022.

[16] S. Ayton, N. G. Faux, and A. I. Bush, "Ferritin levels in the cerebrospinal fluid predict alzheimer's disease outcomes and are regulated by apoe," *Nature communications*, vol. 6, no. 1, p. 6760, 2015.

[17] E. Ficiarà, S. Boschi, S. Ansari, F. D'Agata, O. Abollino, P. Caroppo, G. Di Fede, A. Indaco, I. Rainero, and C. Guiot, "Machine learning profiling of alzheimer's disease patients based on current cerebrospinal fluid markers and iron content in biofluids," *Frontiers in Aging Neuroscience*, vol. 13, p. 607858, 2021.

[18] A. M. Daugherty, B. Larsen, and N. Schröder, "The impacts of iron accumulation on cognitive impairments," p. 1198276, 2023.

[19] A. G. Murley, C. Rua, H. Biggs, C. T. Rodgers, T. Matys, J. van den Ameele, R. Horvath, and P. F. Chinnery, "7t mri detects widespread brain iron deposition in neuroferritinopathy," *Annals of Clinical and Translational Neurology*, vol. 11, no. 5, pp. 1359–1364, 2024.

[20] E. M. Haacke, S. Liu, S. Buch, W. Zheng, D. Wu, and Y. Ye, "Quantitative susceptibility mapping: current status and future directions," *Magnetic resonance imaging*, vol. 33, no. 1, pp. 1–25, 2015.

[21] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *The knowledge engineering review*, vol. 25, no. 1, pp. 1–25, 2010.

[22] Z.-H. Zhou, "Multi-instance learning: A survey," *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, vol. 1, p. 4, 2004.

[23] M. Waqas, S. U. Ahmed, M. A. Tahir, J. Wu, and R. Qureshi, "Exploring multiple instance learning (mil): A brief survey," *Expert Systems with Applications*, p. 123893, 2024.

[24] H. Xiang, J. Shen, Q. Yan, M. Xu, X. Shi, and X. Zhu, "Multi-scale representation attention based deep multiple instance learning for gigapixel whole slide image analysis," *Medical Image Analysis*, vol. 89, p. 102890, 2023.

[25] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature biomedical engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[26] B. Li, Y. Li, and K. W. Eliceiri, "Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning," in *Proc. of the 38th IEEE/CVF CVPR*, 2021, pp. 14 318–14 328.

[27] H. Li, F. Yang, Y. Zhao, X. Xing, J. Zhang, M. Gao, J. Huang, L. Wang, and J. Yao, "Dt-mil: deformable transformer for multi-instance learning on histopathological image," in *Proc. of the 24th MICCAI*, 2021, pp. 206–216.

[28] S. Huang, A. Kanervisto, A. Raffin, W. Wang, S. Ontañón, and R. F. J. Dossa, "A2c is a special case of ppo," *arXiv preprint arXiv:2205.09123*, 2022.

[29] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[30] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 558–567.

[31] M. Contributors, "Openmmlab's pre-training toolbox and benchmark," https://github.com/open-mmlab/mmpretrain, 2023.

[32] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[33] L. A. Giannini, M. Bulk, B. Kenkhuis, A. Rajicic, S. Melhem, I. Hegeman-Kleinn, L. Bossoni, E. Suidgeest, E. G. Dopper, J. C. van Swieten *et al.*, "Cortical iron accumulation in mapt-and c9orf 72-associated frontotemporal lobar degeneration," *Brain Pathology*, vol. 33, no. 4, p. e13158, 2023.

[34] S. van Duijn, M. Bulk, S. G. van Duinen, R. J. Nabuurs, M. A. van Buchem, L. van der Weerd, and R. Natté, "Cortical iron reflects severity of alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 60, no. 4, pp. 1533–1545, 2017.

[35] T. T. Vuong, B. Song, K. Kim, Y. M. Cho, and J. T. Kwak, "Multi-scale binary pattern encoding network for cancer classification in pathology images," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1152–1163, 2021.

[36] J. Dong, J. Jiang, K. Jiang, J. Li, and Y. Zhang, "Fast and accurate gigapixel pathological image classification with hierarchical distillation multi-instance learning," in *Proc. of the 42nd IEEE/CVF CVPR*, 2025, pp. 30 818–30 828.

[37] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, and Y. Zheng, "Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proc. of the 39th IEEE/CVF CVPR*, 2022, pp. 18 802–18 812.

[38] N. Hashimoto, D. Fukushima, R. Koga, Y. Takagi, K. Ko, K. Kohno, M. Nakaguro, S. Nakamura, H. Hontani, and I. Takeuchi, "Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images," in *Proc. of the 37th IEEE/CVF CVPR*, 2020, pp. 3852–3861.

[39] W. Tang, S. Huang, X. Zhang, F. Zhou, Y. Zhang, and B. Liu, "Multiple instance learning framework with masked hard instance mining for whole slide image classification," in *Proc. of the 40th IEEE/CVF CVPR*, 2023, pp. 4078–4087.