Every Error has Its Magnitude: Asymmetric Mistake Severity Training for Multiclass Multiple Instance Learning

Supplementary Material

In-house								
	TA	TVA	TSA	HP	SSL	IP	LP	$\sum$
Train	950	343	387	602	282	181	445	3,190
Validation	205	74	82	127	59	37	94	678
Test	300(95)	81(6)	102(18)	136(8)	114(55)	38	95	866(182)
BRACS [5]								
	IC	DCIS	FEA	ADH	UDH	PB	N	$\sum$
Train	100	40	24	28	56	120	27	395
Validation	12	9	6	8	9	11	10	65
Test	20	12	11	12	9	16	7	87

Table 5. Data distribution over the classes. The values in parentheses represent 182 mixed cases, which were curated by the pathology experts.

6. Data Split

Table 5 summarizes the data splits for the In-house and BRACS datasets [5]. For the In-house dataset, all 182 complex mixed-symptom cases, which were manually curated by experts, were reserved exclusively for the test set. The remaining 4,552 cases were randomly partitioned. Note that the 4,734 In-house samples are derived from a corresponding number of independent patients, ensuring no data leakage occurs. The BRACS dataset utilizes a predefined split.

7. Details on Semantic Feature Remix

Algorithm 1 illustrates the SFR process. (lines 3-7) Following the approach in [26], we apply L-way clustering to the aggregated instances of Z_a and Z_b. This process focuses on cases where $Y_a \succ Y_b$ to identify distinctive symptoms. (lines 8-16) T refinements ensure the clustering of prototypical features. (lines 17-19) A cluster l dominated by instances from Z_a suggests that it captures typical features of the Y_a condition, which are scarce in Z_b. Therefore, we sort the L clusters by their Z_a instance ratio. The synthesized bag Z_{a+b} is formed by merging Z_b with selected Z_a instances from the top-k clusters, with Y_a serving as the definitive label.

Algorithm 1 Semantic Feature Remix

1: **Input:** Instance bags $Z_a = \{z_{a,n}\}_{n=1}^{n(X_a)}$ and $Z_b = \{z_{b,n}\}_{n=1}^{n(X_b)}$, s.t. $Y_a \succ Y_b$, the number of clusters L, refinement iterations T, top-k index $k(<L)$
2: **Output:** Semantically remixed instance bag Z_{a+b}
// **1. Cluster Initialization**
3: $p_{a+b} \leftarrow (1/(|Z_a| + |Z_b|)) \times (\Sigma_{n=1}^{|Z_a|} z_{a,n} + \Sigma_{n=1}^{|Z_b|} z_{b,n})$
4: **for** z_i in $Z_a \bigcup Z_b$ **do**
5: $s_i \leftarrow \text{sim}(p_{a+b}, z_i)$ // cosine similarity
6: Allocate z_i to l-th cluster $\varepsilon_l \in \mathcal{E} = \{\varepsilon_1, \cdots, \varepsilon_L\}$, s.t. $s_i \in [-1 + 2(l-1)/L, -1 + 2l/L]$
7: **end for**
// **2. Cluster Refinement**
8: **for** $t = 1$ to T **do**
9: **for** $l = 1$ to L **do**
10: $Z_l \leftarrow \{z | z \in \varepsilon_l\}$
11: $p_l \leftarrow (1/|Z_l|) \times \Sigma_{z \in Z_l} z$
12: **end for**
13: **for** z_i in $Z_a \bigcup Z_b$ **do**
14: Allocate z_i to ε_l, s.t. $l = \text{argmin}_l \text{ sim}(p_l, z_i)$
15: **end for**
16: **end for**
// **3. Remix**
17: $\mathcal{E}' \leftarrow$ Sort clusters by highest z proportion from Z_a
18: $Z_{a+b} \leftarrow Z_b \bigcup \{z'_a | z'_a \in \mathcal{E}[1:k]\}$
19: **return** Z_{a+b}

8. Sensitivity Analysis

We present the parameter sensitivity analysis for both datasets, BRACS and In-house. All reported results were derived using the validation set, and TransMIL [35] was leveraged for the analysis.

8.1. λ_1, λ_2, and α

Fig 10 visualizes the validation performance with respect to the λ_1, λ_2 and variations in α. The $(\lambda_1, \lambda_2) = (1, 2)$ exhibited poor performance across both datasets. Conversely, the $(\lambda_1, \lambda_2) = (2, 1)$ demonstrated strong overall performance, suggesting that applying a stronger penalty to MSCE is highly beneficial for WSI tasks that involve class priority. Specifically, optimal performance was consistently observed at $\alpha = 1.6$.

8.2. Semantic Feature Remix

Holding the parameters found in Section 8.1 constant, we visualize the performance sensitivity to the SFR parameters in Fig. 11. In general, selecting a value for k that exceeded half of the total number of clusters (i.e., $k > L/2$) yielded acceptable performance across both datasets. Furthermore, increasing the number of clusters L consistently provided a performance advantage over using the minimum number of clusters. This suggests that higher quality clusters can be obtained by increasing L, though at the expense of computational time.

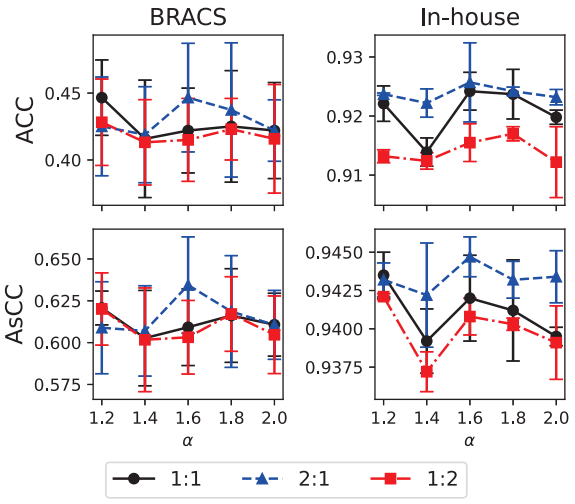

Figure 10. Performance plot according to λ_1, λ_2, and α.

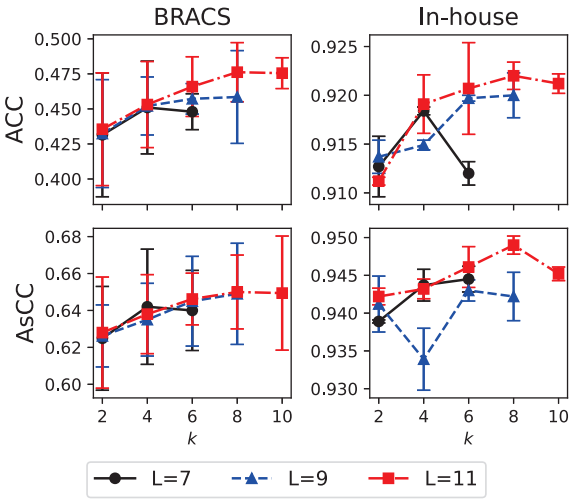

Figure 11. Performance plot according to L and k.

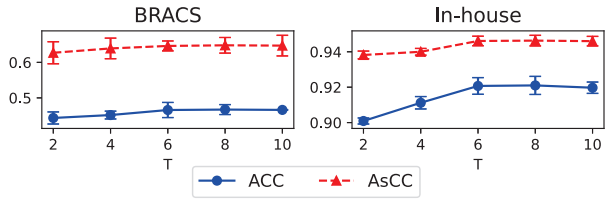

Figure 12. Performance according to the refinement iterations T.

Fig. 12 presents the performance of the two datasets as the cluster refinement iteration T with $(L, k) = (11, 6)$. When T is smaller than 6 (i.e., 2 and 4), both Acc and AsMC exhibit low performance across both datasets. Crucially, performance gains become negligible in both datasets when the T exceeds 6. Consequently, we adopted the $T = 6$ to ensure computational efficiency.

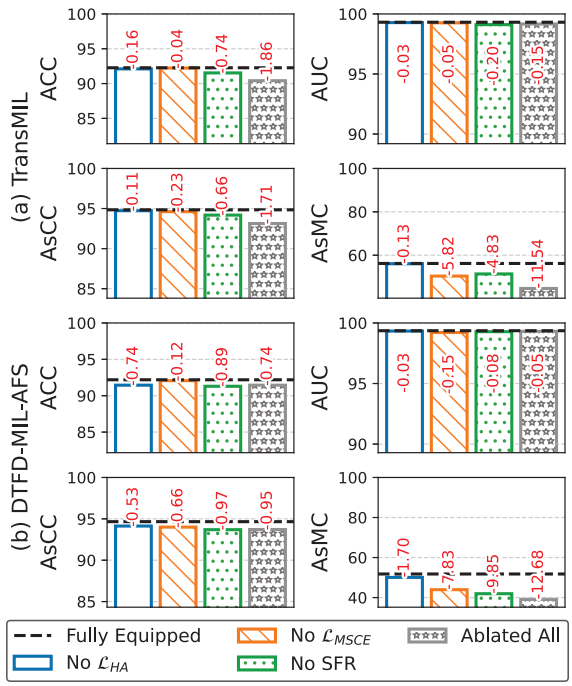

Figure 13. Ablation results for both models on the In-house dataset. The red values denote the difference in metrics between the fully-equipped model and the ablated results.

9. Ablation Study on In-house Dataset

We present the ablation study results for the In-house dataset in Fig. 13. In both model architectures, the ablation of $\mathcal{L}_{HA}$ resulted in a drop in ACC, underscoring the utility of leveraging a hierarchical approach in multiclass settings. Removing $\mathcal{L}_{MSCE}$ resulted in a more severe degradation of the severity metrics compared to removing $\mathcal{L}_{HA}$. We confirm that $\mathcal{L}_{MSCE}$ effectively mitigates severe errors in the inference stage by strongly regularizing high-severity errors during training. The SFR facilitated mistake severity ability in the MIL without imposing direct regulation on the model parameters. Its removal caused a significant drop in both AsCC and AsMC. The ablation of all components resulted in the largest performance decline, proving that the proposed PAMS is an effective solution for the multiclass WSI classification MIL task.

10. Time Complexity Analysis on Remix

Table 6 summarizes the time complexity and the actual implementation time per sample for the remix strategies. ReMix [41] exhibits the largest time complexity, derived from the time required for prototyping plus the additional overhead D for distributed instance generation. This results in an implementation time that exceeds the second-best method by a factor of 10 or more. PseudoMix [26] re-

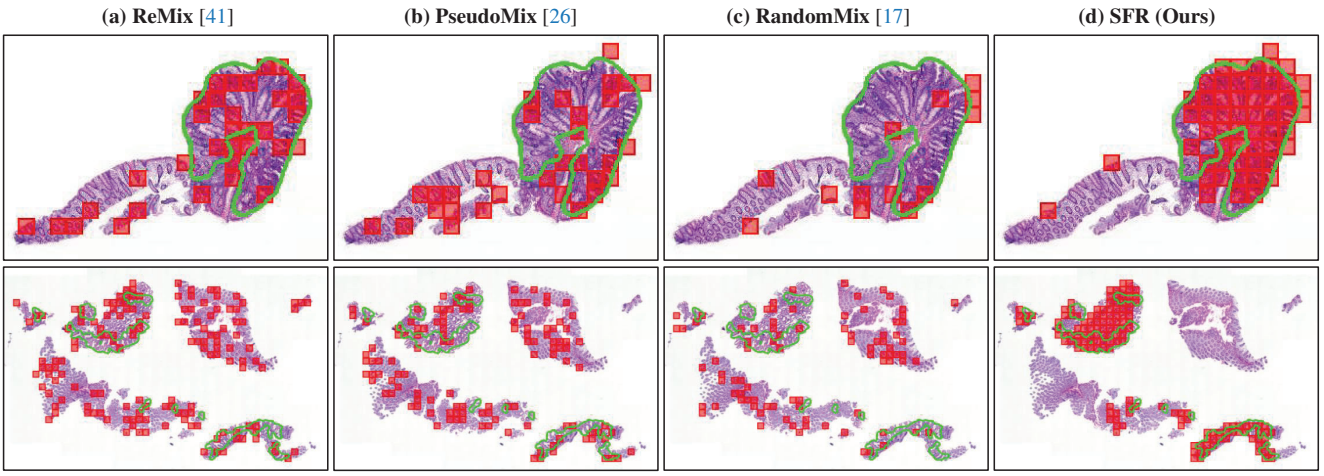

Figure 14. Additional visualization results on various remix strategies. The green polygon indicates the most severe diagnosis labeled by the experts. Red boxes highlight the patches selected by the remix methodology to synthesize the sample without pixel-level annotation.

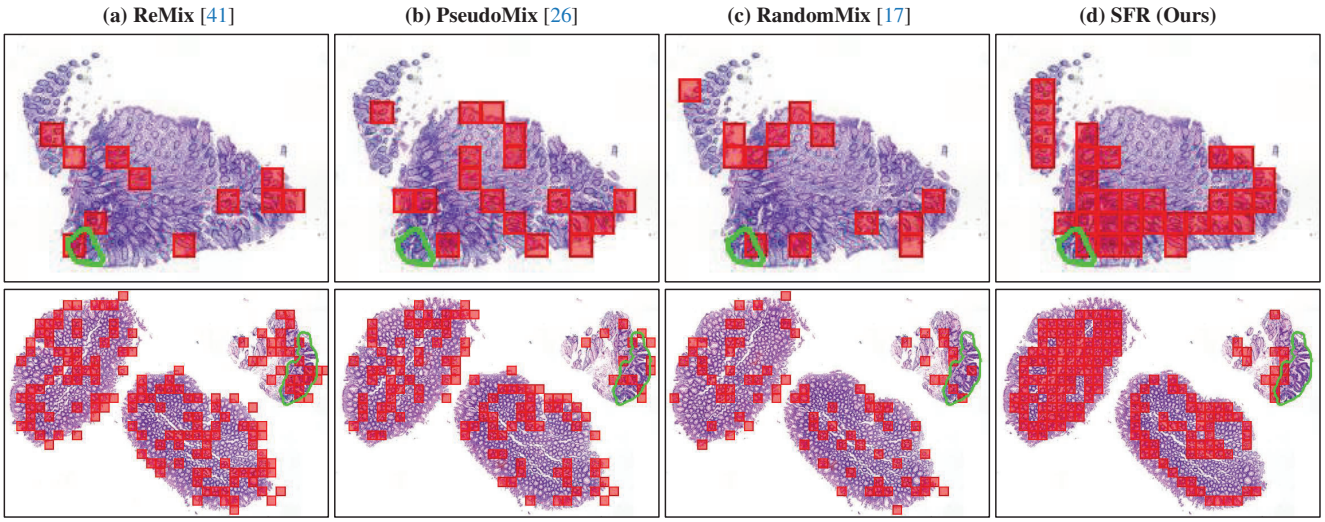

Figure 15. Visualization of failure cases. The green polygon indicates the most severe diagnosis labeled by the experts. Red boxes highlight the patches selected by the remix methodology to synthesize the sample without pixel-level annotation.

Method	Time Complexity	Mean Implementation Time (s) / Sample	
		BRACS [5]	In-house
ReMix [41]	$\mathcal{O}(L{\cdot}k{\cdot}T{\cdot}D{\cdot}n(X_i))$	0.233	0.061
PseudoMix [26]	$\mathcal{O}(L{\cdot}k{\cdot}T{\cdot}n(X_i))$	0.015	0.004
RandomMix [17]	$\mathcal{O}(1)$	0.005	0.003
SFR (Ours)	$\mathcal{O}(L{\cdot}k{\cdot}T{\cdot}n(X_i)\log n(X_i))$	0.023	0.005

Table 6. Time complexity of various remix methods and their practical implementation time on the training set of the two datasets.

quires only prototype clustering, giving it the smallest time complexity among computationally involved remix strategies. Its implementation time was shorter than both ReMix and SFR across both datasets. RandomMix [17] has the shortest, constant time complexity because it relies only on random sampling for selection. SFR requires the time complexity for prototyping, with an added overhead for the sorting operation.

Despite the theoretical computational complexity of SFR and competing methods, practical implementation using FAISS [10, 21] significantly minimized time constraints for their utilization. The empirical runtime of both SFR and PseudoMix was comparable to the $\mathcal{O}(1)$ complexity of RandomMix. Even ReMix, which has a significantly larger time complexity, yields an acceptable training time when executed on the datasets.

Method	Equation	Distance Penalty	Asymmetric Penalty
Cross Entropy (CE)	$-\sum_c \tilde{Y}[c] \log \hat{p}[c]$	✗	✗
Weighted CE	$-\sum_c \tilde{Y}[c] w_c \log \hat{p}[c]$	✗	✗
HXE [4]	$-\sum_h \exp(-\alpha h(N^h)) \log p(N^h \mid N^{h+1})$, where $h(N)$ is depth of node N , $p(N^h \mid N^{h+1}) = \frac{\sum_{A \in \text{Leaves}(N^h)} p(A)}{\sum_{B \in \text{Leaves}(N^{h+1})} p(B)}$, and $p(N) = \prod_h p(N^h \mid N^{h+1})$	△	✗
CO2 [2]	$-\sum_c \tilde{Y}[c] \log \hat{p}[c] +$ $\lambda \sum_c \mathbb{1}(c \geq c') \times \text{ReLU}(\delta + \hat{p}[c+1] - \hat{p}[c]) +$ $\lambda \sum_c \mathbb{1}(c \leq c') \times \text{ReLU}(\delta + \hat{p}[c] - \hat{p}[c+1])$, where $c' = \text{argmax}_c \tilde{Y}$	✓	✗
CDW-CE [32]	$-\sum_c (\log(1 - \hat{p}[c]) \times \lvert c - c' \rvert^{\alpha})$, , where $c' = \text{argmax}_c \tilde{Y}$	✓	✗
MSCE (Ours)	$-\hat{p} W Y^\top \sum_c \tilde{Y}[c] \log \hat{p}[c]$	✓	✓

Table 7. The Cross Entropy and its associated severity loss terms. $\tilde{Y} \in \mathbb{R}^{1 \times C}$ and $\hat{p} \in \mathbb{R}^{1 \times C}$ indicate one-hot vectorized true label and predicted probability, respectively.

11. Additional Qualitative Analysis on Remix

We provide additional visualizations of various remix strategies in Fig 14. **(Top)** SFR successfully select patches that have symptoms, even in relatively narrow regions. By basing these semantically selected patches, SFR synthesizes a new instance bag containing only pertinent pathological characteristics. While ReMix and PseudoMix selected patches based on prototyping and clustering, this often resulted in the inclusion of some irrelevant patches. RandomMix provides the fastest execution time but may sometimes fail to capture pathological features entirely. **(Bottom)** SFR consistently selects higher-priority finding patches, even in complex-shaped tissue. Notably, SFR included even tiny areas, smaller than a single patch size, in the mixing. The absence of any sampling in the non-symptom tissue, which is located at top right side, demonstrates the precision of the pathological feature-based sorting and selecting. In contrast, comparison methods uniformly selected patches across all tissues, inevitably including numerous irrelevant patches.

Fig. 15 presents failure cases. **(Top)** In instances where the pathology is extremely small and tissue presence is sparse, SFR selected patches irrelevant to the intended mixing source. This indicates that SFR struggles with too small regions of pathology because the clustering stage heavily incorporates low-level features, such as shape. Conversely, the comparison method PseudoMix completely failed to include a pathology patch. RandomMix did include pathology patches, but its probabilistic nature can make inclusion difficult when the lesion is small. **(Bottom)** SFR struggled to identify the correct features when the source WSI contained unique patterns specific to the finding but not representative of the tumor.

12. Discussion on Mistake Severity Metrics

12.1. Risk Calculation

$$\mathbb{E}(Risk) = \frac{1}{\sum_{i \neq j} S_{i,j}} \sum_i \sum_j \left(0.5^{\mathbb{1}(i \succ j)} \times S_{i,j} W_{i,j} \right) \tag{11}$$

Equation 11 details the calculation of Risk used in Fig. 8. We incorporate an increase in weight 2 for severe errors and define $\mathbb{E}(\text{Risk})$ as the expected risk over the set of all misclassified samples.

12.2. Accuracy, AsCC, and AsMC

As Zhao et al. [43] argued, severity metrics possess fundamentally different characteristics compared to conventional evaluation approaches. Accuracy is a metric that only counts samples that are classified correctly. While this can be interpreted as penalizing incorrect samples based on a perfect score 1, it remains mathematically equivalent to simply counting the number of correct classifications. Consequently, most existing metrics exhibit a positive correlation with Accuracy (e.g., Recall and AUC). In contrast, AsMC has no relationship with Accuracy (i.e., the number of correct/incorrect samples is irrelevant). It is only related in the singular case where all test samples are correctly classified, resulting in AsMC $= \infty$. AsMC is determined solely by the severity pattern exhibited by the misclassified samples. Therefore, a model with high Accuracy is not guaranteed to achieve the high AsMC. The metric balancing these two perspectives is AsCC. AsCC can be approximated as the sum of Accuracy (representing the reward for correct samples) and AsMC (quantifying the appropriate severity of incorrect samples).

Consequently, a critical question emerges regarding the appropriate metric for model selection in safety-critical applications such as clinical diagnostics. The traditional approaches adopt the models with the highest Accuracy. How-

ever, MS studies have emphasized that in domains where the severity of incorrect predictions is critical, prioritization must shift to models exhibiting high AsMC. This suggestion requires a comprehensive selection standard that bridges accuracy and AsMC. We therefore recommend that model selection be guided by the AsCC metric when both overall Accuracy and mistake severity must be simultaneously optimized.

13. Comparison on Entropy Terms

We perform a comparative analysis of the standard Cross-Entropy (CE) term against several severity-aware CE variants (Table 7). CE only penalizes the prediction probability of the ground truth label and fails to account for severe error patterns and the model's overall prediction distribution across other classes. Weighted CE applies weighted regulation only to the true label index and neglects the global prediction pattern, similar to CE. HXE [4] formulates classes into coarse-to-fine levels and implements hierarchical distance. This enforces a constraint s.t. the coarse prediction for a class node N within a specific hierarchy h matches the prediction probability of its fine-grained classes. However, the hierarchy enforced by HXE is fundamentally focused on probability alignment rather than class priority. CO2 [2] defines an ordinal order among C classes, which can be interpreted as severity. This approach, however, requires the strong assumption that all classes possess priority and cannot incorporate equivalence $\equiv$ relations. CDW-CE [32] is limited in that it uses only the distance between classes and treats the ground truth class as the most severe target. In contrast, the proposed MSCE can explicitly assign both dominance and equivalence relations for severity across all classes. Furthermore, it leverages the entire model prediction $\hat{p}$ to regulate the error pattern towards the less severe mistake.

14. Limitations and Future Works

Contributions Our proposed PAMS framework introduces the crucial problem of MS to the MIL community, which previously focused solely on Accuracy, and highlights the risk of this to clinical deployment. We propose a novel remix method that robustly accounts for severity, specifically tailored to the unique labeling characteristics of multiclass WSIs. The two metrics we introduced, AsCC and AsMC, quantify the model's mistake severity performance from a comprehensive view and solely from the perspective of misclassifications, respectively. By explicitly enforcing priority-aware training via MSCE, PAMS demonstrates superiority over existing severity comparison approaches across multiple metrics. Experiments conducted on challenging public and In-house datasets enable objective evaluation. Finally, we performed additional experiments on natural images to confirm the working mechanism and general scalability of MSCE.

Limitations During the course of this study, we identified several avenues for future work. Although MSCE enforces severity-aware training explicitly, it requires predefining the relationship between entire classes. This makes it difficult to assign precise severity values when class relationships are ambiguous or when the number of classes is extremely large. Furthermore, loss terms that involve the distance of the class can inherently lead to decreased training stability in tasks that involve numerous classes. This limitation, which is also shared by existing distance-based regularization MS approaches, requires a dedicated solution. Furthermore, we observed that MS research remains confined to multiclass classification settings. Although multi-label classification is a more generalized task, all existing MS approaches are currently specialized for multiclass problems.

Future Works To mitigate the identified limitations, we aim to develop a severity term that leverages both explicit and implicit signals. The objective is to autonomously discover patterns that are implicitly judged to be severe, while consistently respecting the explicitly defined severity. Furthermore, we seek to expand the scope from multiclass classification to multi-label classification. This is a more generalized task that is expected to contribute to the broader adoption of safety-centric deep learning ultimately.