Diagnose Like A REAL Pathologist: An Uncertainty-Focused Approach for Trustworthy Multi-Resolution Multiple Instance Learning Supplementary Material

1. Sensitivity Analysis

Fig. 1 shows the sensitivity analysis for the two datasets. In the CAMELYON16 [1] AUC, a α of 0.01 demonstrates a performance advantage across the entire range of δ values, as depicted in Fig. 1**(a)**. However, a notable performance improvement is evident when δ is a relatively large value, such as 0.49 and α is also a large value of 0.1. A similar pattern is observed in the accuracy results of the experiments where δ is 0.49. When δ is 0.25, the AUC deviation values in several experiments are significant. In terms of accuracy, the $\delta = 0.5$, $\alpha = 0.1$ setting yielded the best performance, indicating that ensuring that the model has sufficient decision space for ambiguous predictions helps to select uncertain images. Therefore, for CAMELYON16, we used $(\delta, \alpha) = (0.49, 0.1)$.

Fig. 1**(c)** and **(d)** present the experimental results for the DHMC dataset. Within the DHMC dataset, a δ value of 0.49 demonstrates a distinct advantage across the evaluation metrics. In particular, in terms of accuracy, a δ of 0.49 consistently produces superior performance compared to alternative values. This suggests that in datasets typically characterized by higher difficulty and, consequently, lower accuracy, the model benefits from a broader space for expressing uncertainty. The parameters were chosen as $(\delta, \alpha) = (0.49, 0.1)$, corresponding to the highest classification performance.

Fig. 1**(e)** and **(f)** show the sensitivity analysis results for the BCNB dataset. Unlike the other two datasets, the BCNB dataset is less sensitive to hyperparameter choices, which highlights the robustness of the proposed method to parameter selection on this dataset. Specifically, the AUC scores achieve their highest performance with minimal variance $\alpha = 0.1$ and $\delta = 0.49$. Similarly, accuracy has minimal change with varying α values, reaching its peak performance when $\delta = 0.49$ and $\alpha = 0.1$.

We identified a single set of parameters, $(\delta, \alpha) = (0.49, 0.1)$, that performs robustly across all three datasets: CAMELYON16, DHMC, and BCNB. The $\delta = 0.49$ governs the decision boundary margin of the cross-entropy, suggesting the benefit of classifying classes at a wide distance. The parameter $\alpha = 0.1$ controls the intensity of ad-

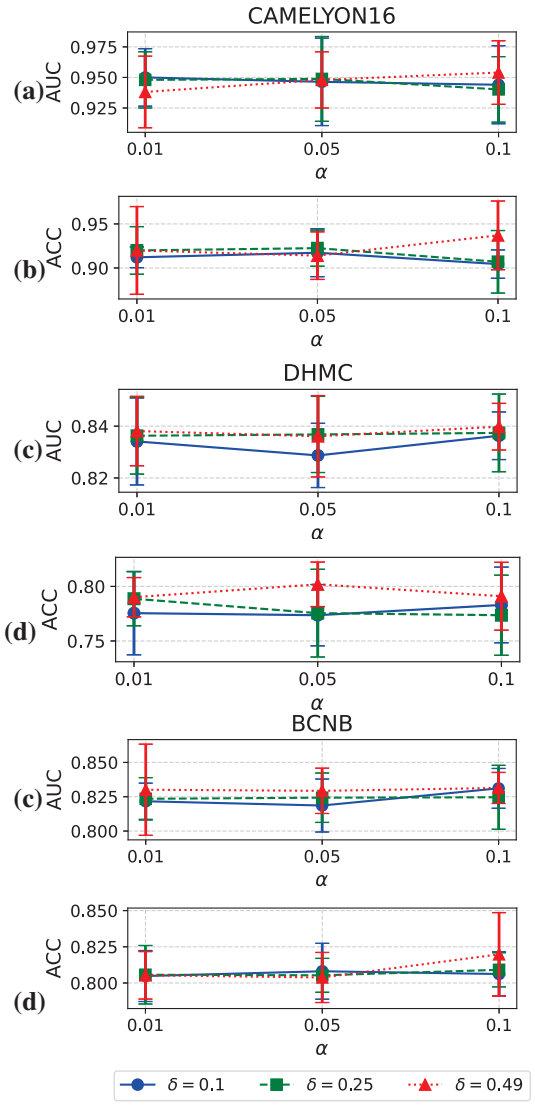

Figure 1. Sensitivity analysis results for two hyper-parameters, δ and α, on CAMELYON16 [1], DHMC [9], and BCNB [11].

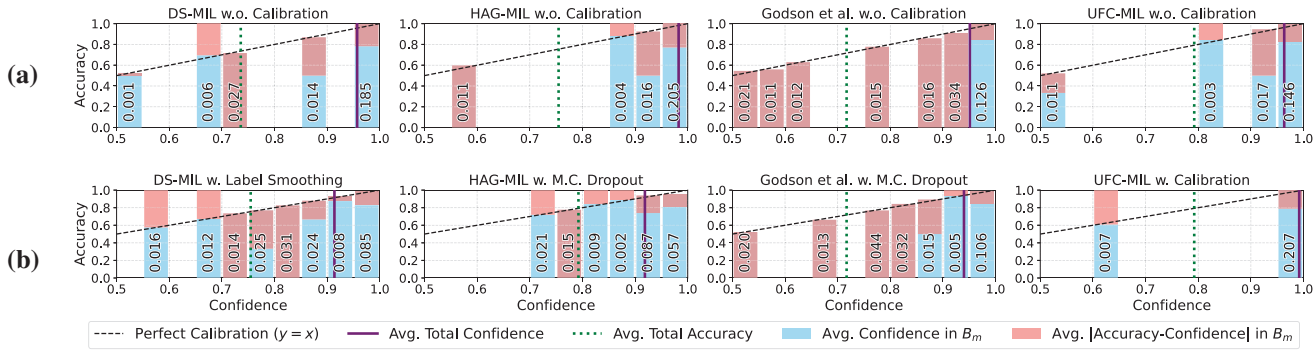

Figure 2. Reliability diagrams on DHMC. We plot histograms comparing uncalibrated models **(a)** with methods achieving the best ECE **(b)** for each.

ditional calibration, indicating that strong smoothing across resolutions and samples can yield performance benefits. Crucially, the ability of a single set of parameters to perform well on datasets from different anatomical sites supports the strong generalization capability of our proposed methodology.

2. Additional Experiment Results on DHMC

2.1. Qualitiatve Analysis

A reliability histogram in Fig. 2 qualitatively shows the ECE for various models and calibration methods. In experiments without model calibration (Fig. 2(**a**)), HAG-MIL [10] and UFC-MIL appropriately separated difficult and easy-to-predict samples, which function as a referable confidence for users. In the case of Godson et al., [3] the model confidence spans the entire range, making it difficult for users to connect numerical values with their actual belief. This also fails to align with the accuracy, thereby exhibiting a high ECE. DS-MIL [6] exhibits a more tempered prediction profile; however, the considerable gap between its confidence and accuracy can lead to untrustworthy results. UFC-MIL, on the other hand, minimizes the difference between confidence and accuracy. Its predictions do not force users to interpret ambiguous confidence intervals.

Fig. 2(**b**) presents the results of the calibration training on each architecture. For DS-MIL, label smoothing helps flatten prediction probabilities and minimize the divergence between accuracy and confidence. However, it shows high ECE in some confidence intervals. M.C. Dropout, when applied to HAG-MIL and Godson et al., demonstrates a moderating effect on model confidence. This reduced the accuracy-confidence gap for HAG-MIL, but the effect was not significant for Godson et al. Additionally, the broad distribution of confidence across all ranges makes it unclear for users to ascertain the trustworthiness of the model's decisions. Our approach demonstrates improvements over comparative methods. It clearly differentiates between predictions where the model struggles and those where it exhibits high confidence.

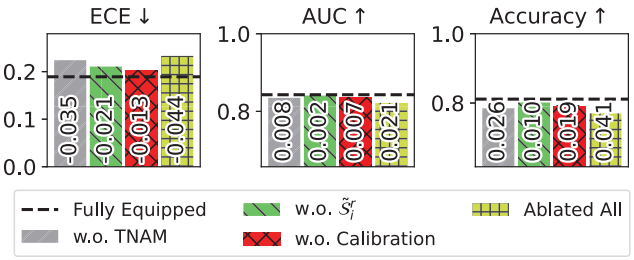

Figure 3. Ablation results on the DHMC dataset. Performance with all proposed methods is shown by a dashed line, with the difference from each ablation indicated above the bars.

2.2. Ablation Study

Fig. 3 presents the ablation study results of UFC-MIL's main components applied to the DHMC dataset. In all metrics, the fully equipped UFC-MIL* demonstrated outstanding performance in the DHMC dataset. Experiments without calibration training using SRLS showed high ECE, indicating that model calibration training influences the alignment with expected accuracy. Furthermore, in this calibration training, $\tilde{S}_i^r$ affected its quantity, indicating that using both the mean and deviation, rather than just the mean as in conventional entropy measurement approaches [7], is effective. The removal of TNAM led to a decrease in overall performance. Notably, while TNAM's removal caused the smallest drop in accuracy, it resulted in a larger increase in ECE. This suggests that TNAM, which aggregates patch-level features, contributes to construct patch-level uncertainty.

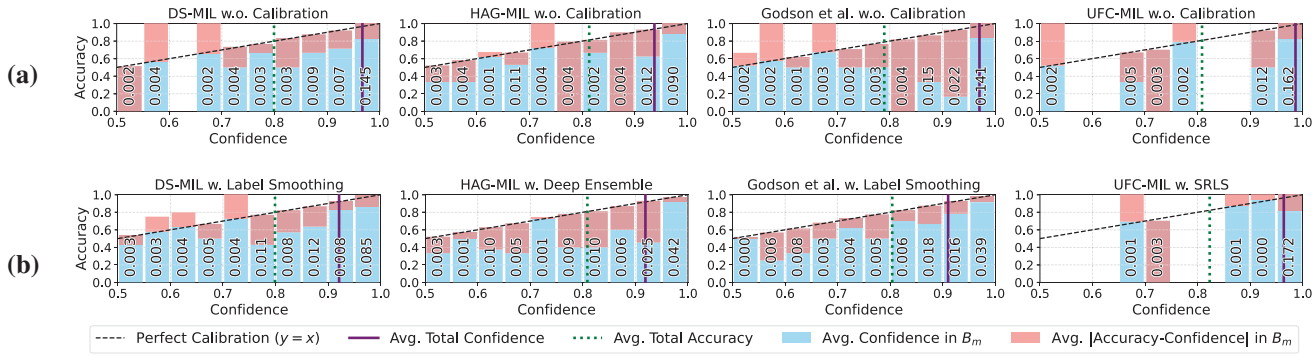

Figure 4. Reliability diagrams on BCNB. We plot histograms comparing uncalibrated models **(a)** with methods achieving the best ECE **(b)** for each.

3. Additional Experiment Results on BCNB

3.1. Qualitative Analysis

We present the reliability diagram for the BCNB dataset in Fig. 4. In uncalibrated results, DS-MIL and Godson et al. demonstrated high average confidence but with low average accuracy, indicating a notable degree of overconfidence. Specifically, the concordance between confidence and accuracy in the high-confidence region degraded substantially for Godson et al., leading to a higher ECE. HAG-MIL and UFC-MIL, however, were better calibrated than the other two models, as the difference between their average confidence and accuracy was minimal. The UFC-MIL model achieves a low ECE except for the highest confidence range. By placing classification confidence into discrete intervals, this approach proves to be highly applicable in clinical settings. Fig. 4**(b)** shows the qualitative results following calibration training. Label Smoothing significantly improved the overall ECE for DS-MIL and contributed to narrowing the confidence-accuracy gap. HAG-MIL achieved better confidence-accuracy alignment with Deep Ensemble, although it still exhibited a high ECE in the high-confidence range. While Label Smoothing also reduced the gap for Godson et al., a high ECE was still observed across the entire confidence spectrum. The SRLS calibration, when applied to UFC-MIL, contributed to a reduction in the confidence-accuracy gap and enhanced the quantization of prediction confidence. This provides a clearer basis for users to interpret and rely on the model's predictions.

3.2. Ablation Study

We present the ablation results of our proposed methodology on the BCNB dataset (Fig. 5). Our complete model achieved superior overall performance compared to its ablated variants, with the removal of any proposed component leading to significant degradation in ECE. Specifically, the absence of TNAM resulted in a high ECE and a low AUC, demonstrating its impact on the classifier's overall performance. Training without SRLS resulted in a notable increase in ECE, accompanied by only a slight decrease in AUC and accuracy, suggesting that SRLS contributes to more reliable model training. The removal of the $\tilde{\mathcal{S}}_i^r$ component also increased ECE, showing that $\tilde{\mathcal{S}}_i^r$ functions as a critical calibration component in conjunction with $\tilde{\mathcal{M}}_i^r$.

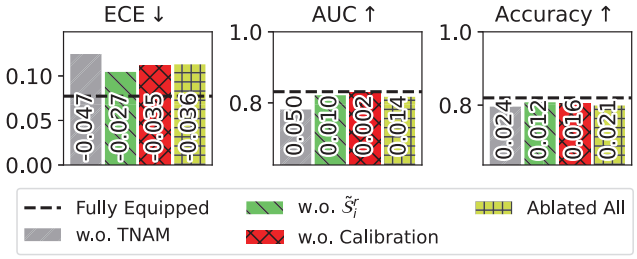

Figure 5. Ablation results on the BCNB dataset. Performance with all proposed methods is shown by a dashed line, with the difference from each ablation indicated above the bars.

4. Further Quanltitative Results

Table 1 presents additional recall scores and AUC for various MRMILs and calibration methods across three datasets. The results consistently show a lack of correlation between R@$k\%$ and AUC, indicating a misalignment between low Type II error rate and discriminative power. A comparison with the accuracy values in the main text is necessary, as recall scores can be inflated due to overestimation.

Temperature Scaling generally improves AUC, though in some cases it comes at the expense of R@$k\%$. A direct correlation between these two metrics is difficult to establish, as there are also instances where both AUC and Recall are improved. This same pattern is observed with Label Smoothing and MC Dropout. The Deep Ensemble experiments revealed a significant reduction in AUC across all datasets, which we attribute to the decision boundary

Calibration Method	MRMIL	CAMELYON16 [1]		DHMC [9]		BCNB [11]	
		R@50% ↑	AUC ↑	R@70% ↑	AUC ↑	R@50% ↑	AUC ↑
-	DS-MIL [6]	0.892 (0.012)	0.924 (0.009)	0.847 (0.056)	0.818 (0.018)	0.961 (0.028)	0.811 (0.012)
	HAG-MIL [10]	0.923 (0.021)	0.877 (0.025)	0.781 (0.059)	0.800 (0.016)	0.975 (0.022)	0.811 (0.011)
	Godson et al. [3]	0.965 (0.060)	0.952 (0.031)	0.840 (0.044)	0.856 (0.006)	0.955 (0.021)	0.807 (0.013)
	UFC-MIL	0.971 (0.028)	0.952 (0.033)	0.873 (0.046)	0.836 (0.015)	0.984 (0.014)	0.829 (0.023)
Temperature Scaling [4]	DS-MIL [6]	0.963 (0.003)	0.958 (0.020)	0.874 (0.052)	0.827 (0.011)	0.977 (0.001)	0.814 (0.035)
	HAG-MIL [10]	0.833 (0.004)	0.833 (0.007)	0.763 (0.033)	0.794 (0.019)	0.967 (0.036)	**0.831 (0.011)**
	Godson et al. [3]	0.966 (0.053)	0.954 (0.028)	0.861 (0.057)	**0.857 (0.008)**	0.947 (0.027)	0.824 (0.016)
	UFC-MIL	0.968 (0.003)	0.948 (0.008)	0.862 (0.036)	0.823 (0.018)	0.970 (0.017)	0.792 (0.028)
Label Smoothing [8]	DS-MIL [6]	0.912 (0.125)	0.939 (0.052)	0.846 (0.075)	0.797 (0.027)	0.944 (0.029)	0.779 (0.075)
	HAG-MIL [10]	0.920 (0.030)	0.873 (0.019)	0.790 (0.031)	0.806 (0.015)	0.958 (0.017)	0.817 (0.012)
	Godson et al. [3]	0.952 (0.060)	0.948 (0.033)	0.823 (0.028)	**0.857 (0.007)**	0.975 (0.008)	0.824 (0.017)
	UFC-MIL	0.965 (0.026)	0.953 (0.027)	0.885 (0.004)	0.834 (0.007)	0.978 (0.011)	0.824 (0.018)
M.C. Dropout† [2]	DS-MIL [6]	0.929 (0.041)	**0.964 (0.005)**	0.836 (0.062)	0.825 (0.021)	0.958 (0.025)	0.813 (0.008)
	HAG-MIL [10]	0.884 (0.038)	0.895 (0.003)	0.790 (0.038)	0.803 (0.015)	0.963 (0.023)	0.819 (0.024)
	Godson et al. [3]	0.950 (0.065)	0.950 (0.029)	0.833 (0.065)	0.843 (0.012)	0.959 (0.014)	0.808 (0.018)
	UFC-MIL	0.971 (0.028)	0.951 (0.031)	0.879 (0.039)	0.834 (0.019)	0.995 (0.005)	0.829 (0.022)
Deep Ensembles† [5]	DS-MIL [6]	0.892	0.908	0.866	0.754	0.925	0.665
	HAG-MIL [10]	0.878	0.848	0.769	0.754	0.984	0.698
	Godson et al. [3]	0.971	0.879	0.846	0.760	0.960	0.674
	UFC-MIL	0.966	0.912	0.785	0.813	0.951	0.718
UDLS† [7]	DS-MIL [6]	0.958 (0.036)	0.940 (0.003)	0.825 (0.117)	0.823 (0.025)	0.966 (0.012)	0.808 (0.034)
	HAG-MIL [10]	0.853 (0.072)	0.864 (0.054)	0.802 (0.093)	0.803 (0.017)	0.985 (0.012)	0.804 (0.009)
	Godson et al. [3]	0.727 (0.182)	0.802 (0.113)	0.695 (0.289)	0.813 (0.024)	0.956 (0.017)	0.825 (0.012)
	UFC-MIL	0.840 (0.081)	0.883 (0.013)	**0.884 (0.044)**	0.826 (0.017)	0.979 (0.029)	0.811 (0.024)
UFC-MIL★		**0.973 (0.027)**	**0.964 (0.024)**	0.881 (0.038)	0.843 (0.011)	**0.994 (0.010)**	**0.831 (0.011)**

Table 1. Additional recall score and AUC on CAMELYON16 [1], DHMC [9], and BCNB [11] dataset. We report the mean and standard deviation, with the latter indicated in parentheses. In each metric, the highest value is bolded. A dagger † indicates that the calibration methods require extra inference steps for model calibration training.

becoming less distinct during the aggregation of multiple model outputs. UDLS experienced a decline in both recall and AUC on the CAMELYON16 and DHMC datasets. This suggests that the single-resolution-based UDLS requires a more refined methodology to operate effectively in a multi-resolution context. Conversely, its application to the BCNB dataset resulted in an improvement in both metrics, highlighting the utility of UDLS calibration for datasets with class imbalance. The SRLS-trained UFC-MIL★ demonstrated competitive performance on the CAMELYON16 and BCNB datasets. While UFC-MIL achieved the second-best R@70% on DHMC, it attained the highest R@70% among uncalibrated MILs. Although the highest AUC was achieved by Godson et al., it was accompanied by a poor ECE, which is not aligned with our goal of developing a well-calibrated MIL.

5. Additional Diagnostic Process Visualization

Fig. 6 shows visualized examples of traditional attention map and uncertainty-based map from the coarsest (2 MPP) to the finest (0.5 MPP) resolutions. Although HAG-MIL focused on the coarsest-resolution lesions, it was designed to drop patches during its iterative zooming process because of multiple discontinuous models. This limits its observation to narrow areas at higher, information-rich resolutions.

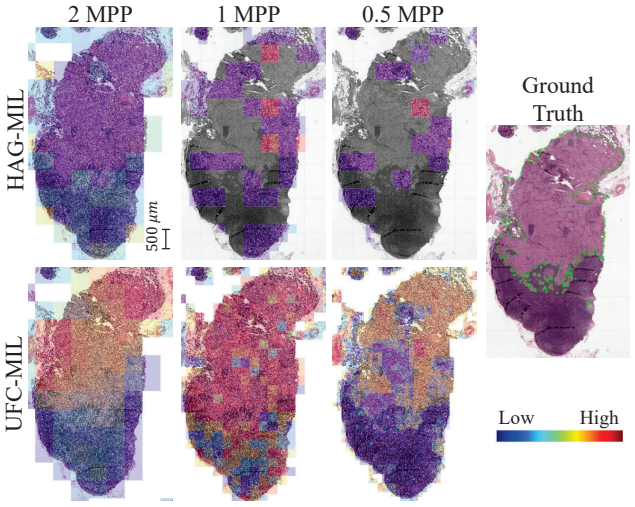

Figure 6. Illustration of attention map versus uncertainty map. In the attention map of HAG-MIL [10], patches with low attention scores that were dropped during the zooming process are shown in grayscale, which the fine-grained model had no opportunity to observe.

UFC-MIL showed high uncertainty around the lesions and their boundaries at the 2 MPP. When the model observes these uncertain areas, we see that the model's uncertainty

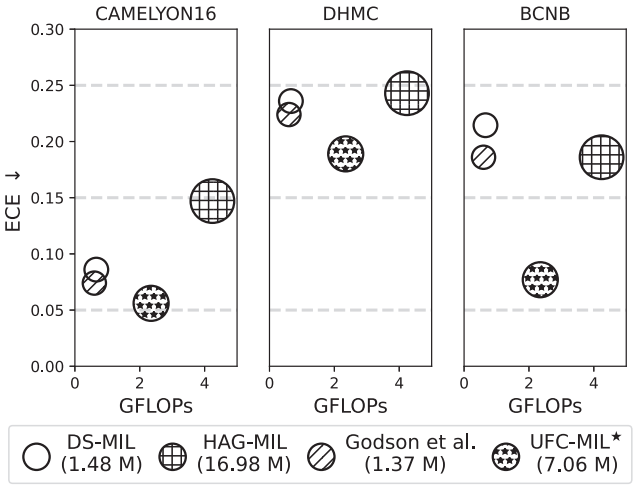

Figure 7. We plot the GFLOPs and ECE of the MRMILs across three datasets that were leveraged in the experiments. The area of each plotted dot is proportional to the number of trainable parameters for each model, with the exact count provided in parentheses.

is significantly reduced at the next resolution. This behavior mirrors a human expert's process of resolving diagnostic uncertainty through magnification, which is made possible by our single model's continuous multi-resolution analysis.

6. Model Complexity

Fig. 7 plots the computational complexity (GFLOPs) and ECE for various MRMILs across three datasets used in the study. DS-MIL and Godson et al. have the fewest parameters and lowest GFLOPs (≈ 0.65). In contrast, HAG-MIL uses a relatively large number of learnable parameters (16.98 M) and 4.24 GFLOPs due to its independent models for each resolution and their serial operations. UFC-MIL sits between these extremes with 7.06 M learnable parameters and 2.35 GFLOPs, a computational footprint that is still manageable on a single 8 GiB GPU. The results for the CAMELYON16 and DHMC datasets show that UFC-MIL effectively leverages its additional parameter count, achieving a lower ECE compared to the Godson et al. and DS-MIL. Conversely, HAG-MIL fails to benefit from a larger parameter count, reporting the highest ECE on these datasets. In the BCNB result, UFC-MIL significantly reduces ECE, demonstrating a clear advantage in producing more reliable MIL models. HAG-MIL achieves a lower ECE than DS-MIL by using a substantially larger number of parameters, but its ECE is on par with that of Godson et al. These findings demonstrate that our proposed methodology effectively improves ECE with a modest and computationally feasible increase in parameters, even when using a GPU with limited memory.

Algorithm 1 UFC-MIL Calibration Training

Input: R resolutions of WSI instances dataset $\mathcal{D}$ which are yield by pre-trained feature extractor, Initialized UFC-MIL $f_\theta = \{f_{\theta_r}\}_{r=1:R}$, PW Loss margin δ, Gumbel softmax τ, Smoothing factor α, Training Epoch $\mathbf{E}$
Output: Calibration-trained UFC-MIL* f_θ
\\ Training
for $\mathbf{e} = 1$ to $\lfloor 0.9 \times \mathbf{E} \rfloor$ **do**
  for $\{z_i^{r=1:R}, \mathcal{A}_i^{r=1:R}, Y_i\} \in \mathcal{D}$ **do**
    $\hat{p}_i^{r=1:R}, \hat{\mathbf{p}}_i^{r=1:R} = f_\theta(z_i^{r=1:R}, \mathcal{A}_i^{r=1:R}, \tau)$
    Cross Entropy Loss: $\mathcal{L}_i^{r=1:R}$
    Patch-Wise Loss with δ: $\mathcal{L}_i^{r=1:R}$
  end for
  $\mathcal{L} = \sum_i \sum_r \left(\mathcal{L}_{i,CE}^r + \mathcal{L}_{i,PW}^r\right)$
  $f_\theta \leftarrow Optimizer(f_\theta, \mathcal{L})$
  if $\mathbf{e} = \lfloor 0.9 \times \mathbf{E} \rfloor$ **then**
    $\mathcal{M}^{r=1:R} \leftarrow \bigcup_{i \in \mathcal{D}} \text{mean}\left(H(\hat{\mathbf{p}}_i^{r=1:R})\right)$
    $\mathcal{S}^{r=1:R} \leftarrow \bigcup_{i \in \mathcal{D}} \text{std}\left(H(\hat{\mathbf{p}}_i^{r=1:R})\right)$
    Log $\hat{\mathbf{p}}_i^{r=1:R}$
  end if
end for
\\ Model Calibration Training
for $\mathbf{e} = 1$ to $\mathbf{E} - \lfloor 0.9 \times \mathbf{E} \rfloor$ **do**
  for $\{z_i^{r=1:R}, \mathcal{A}_i^{r=1:R}, Y_i\} \in \mathcal{D}$ **do**
    if $\mathbf{e} = 1$ **then**
      $\tilde{\mathcal{M}}_i^{r=1:R} = \frac{\text{mean}(H(\hat{\mathbf{p}}_i^{r=1:R})) - \min(\mathcal{M}^{r=1:R})}{\max(\mathcal{M}^{r=1:R}) - \min(\mathcal{M}^{r=1:R})}$
      $\tilde{\mathcal{S}}_i^{r=1:R} = \frac{\text{std}(H(\hat{\mathbf{p}}_i^{r=1:R})) - \min(\mathcal{S}^{r=1:R})}{\max(\mathcal{S}^{r=1:R}) - \min(\mathcal{S}^{r=1:R})}$
      $\epsilon_i^{r=1:R} = \frac{1}{2}(\tilde{\mathcal{M}}_i^{r=1:R} + \tilde{\mathcal{S}}_i^{r=1:R}) \times \alpha$
      $\tilde{Y}_i^{r=1:R} \leftarrow (1 - \varepsilon_i^{r=1:R}) Y_i + \varepsilon_i^{r=1:R}/C$
      Log $\tilde{Y}_i^{r=1:R}$
    end if
    $\hat{p}_i^{r=1:R}, \hat{\mathbf{p}}_i^{r=1:R} = f_\theta(z_i^{r=1:R}, \mathcal{A}_i^{r=1:R}, \tau)$
    Cross Entropy Loss using $\tilde{Y}_i^{r=1:R}$: $\mathcal{L}_i^{r=1:R}$
  end for
  $\mathcal{L} = \sum_i \sum_r \left(\mathcal{L}_{i,CE}^r\right)$
  $f_\theta \leftarrow Optimizer(f_\theta, \mathcal{L})$
end for

7. Pseudo Training Algorithm

We present the detailed training process of UFC-MIL* and the additional calibration procedure in Algorithm 1.

References

[1] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, 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.

[2] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[3] Lucy Godson, Navid Alemi, Jérémie Nsengimana, Graham P Cook, Emily L Clarke, Darren Treanor, D Timothy Bishop, Julia Newton-Bishop, and Derek Magee. Multi-level graph representations of melanoma whole slide images for identifying immune subgroups. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 85–96. Springer, 2023.

[4] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[5] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[6] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021.

[7] Hyeongmin Park, Sungrae Hong, Chanjae Song, Jongwoo Kim, and Mun Yong Yi. Uncertainty-based data-wise label smoothing for calibrating multiple instance learning in histopathology image classification. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 599–608. IEEE, 2025.

[8] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[9] Jason W Wei, Laura J Tafe, Yevgeniy A Linnik, Louis J Vaickus, Naofumi Tomita, and Saeed Hassanpour. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific reports*, 9(1):3358, 2019.

[10] Conghao Xiong, Hao Chen, Joseph JY Sung, and Irwin King. Diagnose like a pathologist: Transformer-enabled hierarchical attention-guided multiple instance learning for whole slide image classification. *arXiv preprint arXiv:2301.08125*, 2023.

[11] Feng Xu, Chuang Zhu, Wenqi Tang, Ying Wang, Yu Zhang, Jie Li, Hongchuan Jiang, Zhongyue Shi, Jun Liu, and Mulan Jin. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in Oncology*, page 4133, 2021.