Synthetic Data Reveals Generalization Gaps in Correlated Multiple Instance Learning

Ethan Harvey[1] ETHAN.HARVEY@TUFTS.EDU
Dennis Johan Loevlie[1] DENNIS.LOEVLIE@TUFTS.EDU
Michael C. Hughes[1] MICHAEL.HUGHES@TUFTS.EDU
[1] *Department of Computer Science, Tufts University, Medford, MA, USA*

Abstract

Multiple instance learning (MIL) is often used in medical imaging to classify high-resolution 2D images by processing patches or classify 3D volumes by processing slices. However, conventional MIL approaches treat instances separately, ignoring contextual relationships such as the appearance of nearby patches or slices that can be essential in real applications. We design a synthetic classification task where accounting for adjacent instance features is crucial for accurate prediction. We demonstrate the limitations of off-the-shelf MIL approaches by quantifying their performance compared to the optimal Bayes estimator for this task, which is available in closed-form. We empirically show that newer correlated MIL methods still do not achieve the best possible performance when trained with ten thousand training samples, each containing many instances.

Data and Code Availability Synthetic data and Python code are available at `https://github.com/tufts-ml/correlated-mil` and have been integrated into torchmil (Castro-Macías et al., 2025).

Institutional Review Board (IRB) Our study uses synthetic data and does not require IRB approval.

1. Introduction

Many prediction tasks in medical imaging involve visual data with varying cardinality, resolution, or dimensionality. For example, inputs may consist of high-resolution 2D images (e.g., histopathology images) or 3D image volumes (e.g., CT or MRI scans). In these scenarios, a common approach is to divide each image into smaller 2D patches or slices known as *instances*, obtain per-instance representations, and then aggregate scores or representations across instances to make one prediction for the whole image (Ilse et al., 2018; Han et al., 2020; Shao et al., 2021; Harvey et al., 2023). Building predictors that aggregate one coherent prediction from many instance representations is known as *multiple instance learning* (MIL) (Quellec et al., 2017; Dietterich et al., 1997; Maron and Lozano-Pérez, 1997). MIL offers a practical framework for handling weakly labeled data.

Conventional MIL approaches broadly treat instances separately and independently. This assumption ignores the spatial and contextual relationships between adjacent patches or slices. Accounting for these relationships can be critical for accurate prediction in medical applications. To address this problem, recent work has proposed *correlated MIL* (Shao et al., 2021) to model dependencies between instances. Others have built upon this direction (Castro-Macías et al., 2024). Assessing the capabilities and limits of such methods remains an open problem.

In this work, we take a synthetic data approach to better understand the importance of spatial and contextual relationships between adjacent instances in multiple instance learning. Our contributions are:

- We design a novel synthetic dataset called *Shifted Mean MIL* to represent key challenges in MIL for medical imaging: (1) only some features are discriminative, (2) only a few instances in each bag signal whether it should be positive class, and (3) context from nearby instances matters, as the information from an individual instance may be statistically ambiguous.
- We derive the optimal Bayes estimator for this dataset and use its predictions as a gold standard for comparing how well MIL methods perform.
- We demonstrate that even recent correlated MIL methods designed to account for context do not achieve the best possible performance on our toy task, as shown in Fig. 1.

These contributions suggest concrete opportunities for future work to improve correlated MIL, perhaps via improved inductive biases or regularization strategies.

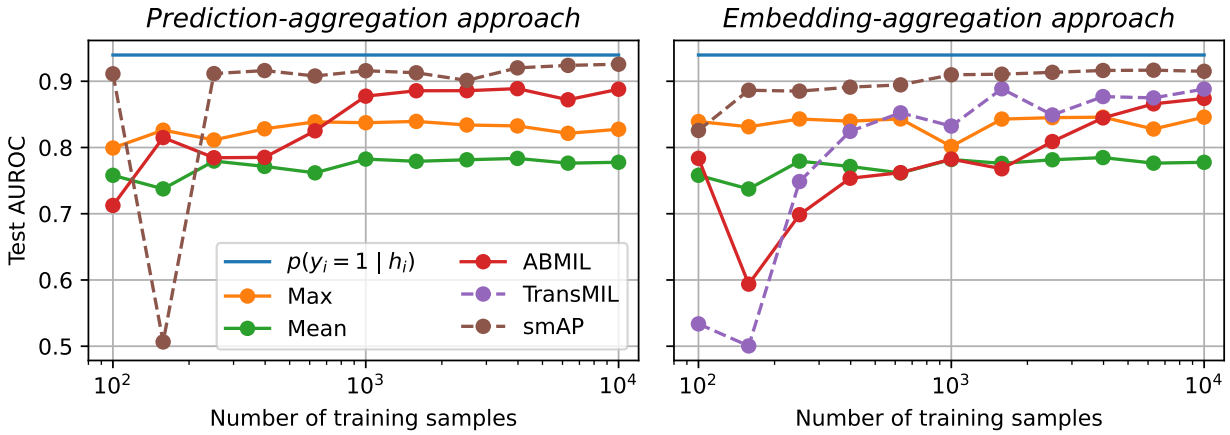

Figure 1: Test AUROC as a function of training set size N. All data is drawn from our Shifted Mean MIL data-generating process for binary classification, with R=3, Δ=2. Conventional MIL approaches (Max, Mean, ABMIL) cannot match the Bayes estimator $p(y_i = 1 \mid h_i)$ as they do not account for dependencies between instances within a bag. Surprisingly, even with N=10000, correlated MIL approaches (TransMIL (Shao et al., 2021), smAP (Castro-Macías et al., 2024)) do not reach the ceiling set by the Bayes estimator. smAP comes close, but bootstrapping reveals the Bayes estimator maintains a statistically significant advantage (mean of AUROC difference is 0.014, 95% confidence interval of [0.007, 0.022] does *not* include zero or any negative values). **Takeaway: Our work reveals a need for data-efficient MIL that better accounts for context between instances.**

2. Related Work

Multiple instance learning. MIL is a branch of weakly supervised learning where a variable-sized set of instances has a single label. Early MIL approaches used simple, non-trainable operations such as max or mean pooling to aggregate instance representations (Pinheiro and Collobert, 2015; Zhu et al., 2017; Feng and Zhou, 2017). Recent work has proposed attention-based pooling (Ilse et al., 2018). Several works have extended attention-based pooling while maintaining permutation-invariance (Li et al., 2021; Lu et al., 2021; Keshvarikhojasteh et al., 2024). Correlated MIL (Shao et al., 2021) extends traditional MIL by modeling relationships between instances within a bag, allowing the pooling operation to capture morphological and spatial information rather than treating instances as independent. Castro-Macías et al. (2024) proposed a smoothing operator to introduce local dependencies among neighbors. Shao et al. (2025) showed that transfer learning with MIL approaches improves generalization.

Most similar in spirit to our work are the *algorithmic unit tests* for MIL proposed by Raff and Holt (2023). They suggest three synthetic classification tasks designed to reveal whether learned models violate key MIL assumptions, such as a bag is positive if and only if one or more instances have a positive label. Our new data focuses on cross-instance dependency, which was not examined by Raff and Holt (2023).

Adjacent context in deep learning. Several prior works have introduced architectural modifications to better capture dependencies between adjacent patches in 2D images or slices in 3D images. Shifted windows allow for cross-window connections in vision transformers (ViT) (Liu et al., 2021). Weight inflation transfers pre-trained weights from lower- to higher-dimensional model (e.g., 2D to 3D CNNs) (Carreira and Zisserman, 2017; Zhang et al., 2022).

3. Background

3.1. Multiple Instance Learning

To train MIL models, the training dataset $\mathcal{D} = \{(x_{i,1:S_i}, y_i)\}_{i=1}^{N}$ consists of N labeled bags. Each bag is a set of S_i independent instance feature vectors $\{x_{i,1}, \ldots, x_{i,S_i}\}$ with a single label y_i. Among deep MIL pipelines, there are two broad paradigms: *embedding-aggregation* and *prediction-aggregation*. Both approaches have neural architectures consisting of three parts: an encoder, a pooling operation, and classifier. They differ in the ordering of these components.

In the *embedding-aggregation* approach, the order is encode, pool, then classify. First, each instance's B-channel raw image $x_{i,j} \in \mathbb{R}^{B \times W \times H}$ is encoded into

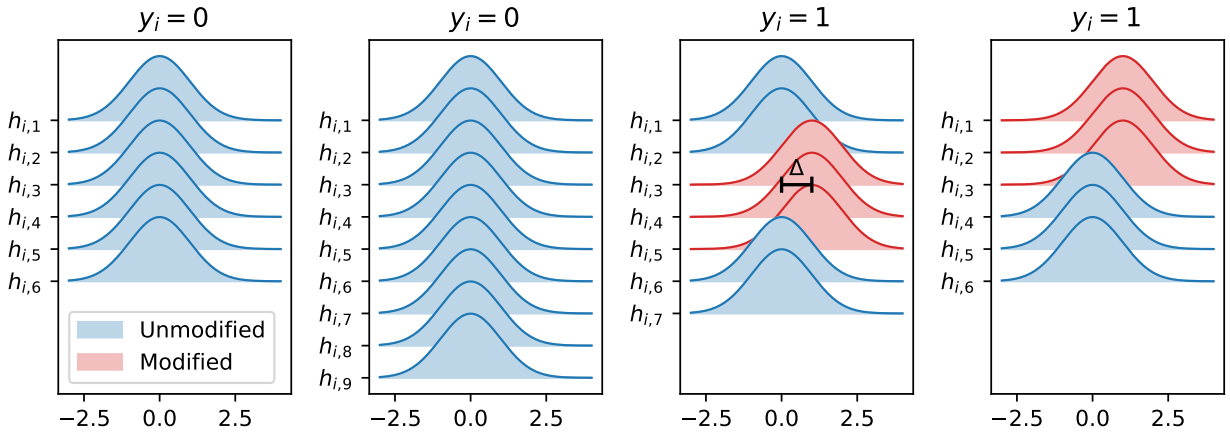

Figure 2: Example data-generating distributions for a discriminative feature for negative (y_i=0) and positive (y_i=1) "bags" of S_i instances drawn from our Shifted Mean MIL synthetic data. Setting R=3 means context around modified instances (in red) can help. In our experiments, we set R=3, S_{low}=15, S_{high}=45, K=1, M=768, μ=0, and σ=1. We study how prediction quality changes as we vary training set size N (Fig. 1) and class separation Δ (Fig. A.1).

a representation vector $h_{i,j} = f(x_{i,j}) \in \mathbb{R}^M$. Second, a pooling operation σ (e.g., max, mean, or attention-based pooling) aggregates all S_i instance representations $\{h_{i,1}, \ldots, h_{i,S_i}\}$ into a single representation vector $z_i = \sigma(h_{i,1:S_i}) \in \mathbb{R}^M$. Usually, this pooling is *permutation-invariant*. Finally, the bag level representation vector z_i is classified into a predicted probability vector over C classes, $c(z_i) \in \Delta^C \subset \mathbb{R}^C$. We can denote the ultimate prediction as $\hat{y}_i = c(\sigma(f(x_{i,1:S_i}))$. In this notation, applying f to a set yields another set containing a mapping of each instance.

In the *prediction-aggregation* approach, the ordering of c and σ is swapped. A separate prediction score (e.g., a logit or probability vector) is produced for each of the S_i instances separately, and then pooling determines the final prediction, $\hat{y}_i = \sigma(c(f(x_{i,1:S_i}))))$.

Ultimately, in either approach, model parameters for all three parts (encoder, pooling, and classifier) are trained to minimize binary or multi-class cross entropy averaged across all data: $\frac{1}{N}\sum_{i=1}^{N} \ell^{\text{CE}}(y_i, \hat{y}_i)$, where $\hat{y}_i$ is a function of input features and parameters.

3.2. Pooling Methods

The design of the pooling layer σ, which aggregates across instances, is generally most important for understanding how spatial context is incorporated. We describe several architectures below. We focus on *embedding-aggregation* for concreteness; a translation to *prediction-aggregation* is straightforward. Here, we take as input a set of embeddings $h_i := h_{i,1:S_i}$ for bag i. Each instance j in the bag is embedded as vector $h_{ij} \in \mathbb{R}^M$.

Max and Mean. Two simple poolings find the maximum or mean ***element-wise*** for M-dim. vectors:

$$z_i = \max_{j=1,\ldots,S_i} h_{ij}, \quad \text{or} \quad z_i = \operatorname{mean}_{j=1,\ldots,S_i} h_{ij}. \tag{1}$$

Attention-based pooling. Attention-based pooling (ABMIL) (Ilse et al., 2018) assigns an attention weight a_{ij} to each instance, then forms bag-level embedding vector z_i via a weighted average:

$$z_i = \sum_{j=1}^{S_i} a_{ij} h_{ij}, \quad a_{ij} = \frac{\exp\left(u^\top \tanh\left(U h_{ij}\right)\right)}{\sum_{k=1}^{S_i} \exp\left(u^\top \tanh\left(U h_{ik}\right)\right)},$$

where the weights a_{ij} are non-negative and sum to one: $\sum_j a_{ij} = 1$; $a_{ij} \geq 0$ for all j. Here, vector $u \in \mathbb{R}^L$ and matrix $U \in \mathbb{R}^{L \times M}$ are trainable parameters.

Smooth attention pooling. Smooth attention pooling (smAP) (Castro-Macías et al., 2024) uses a smoothing operation to add local interactions between instance embeddings. The smoothed embeddings $g_i \in \mathbb{R}^{S_i \times M}$ for all S_i instances are obtained by solving an optimization problem

$$\texttt{Sm}(h_i) = \operatorname*{argmin}_{g_i} \alpha \mathcal{E}_D(g_i) + (1-\alpha)\|h_i - g_i\|_F^2, \tag{2}$$

where $\alpha \in [0, 1)$ controls the amount of smoothness, $\|\cdot\|_F$ denotes the Frobenius norm, and

$$\mathcal{E}_D(g_i) = \frac{1}{2}\sum_{j=1}^{S_i}\sum_{k=1}^{S_i} A_{ijk}\|g_{ij} - g_{ik}\|_2^2. \tag{3}$$

Here, $A_i \in \mathbb{R}^{S_i \times S_i}$ is an adjacency matrix defining local relationships between instances and $\|\cdot\|_2^2$ denotes the squared Euclidean norm aka "sum of squares".

Correlated MIL pooling. Shao et al. (2021)'s transformer-based correlated MIL (TransMIL) allows instance interactions to inform pooling. First, TransMIL uses convolutions over instances in a pyramidal position encoding generator to model dependencies. Second, interactions between *all pairs* of instances are captured via multi-head self-attention. For layer ℓ and head h, there's a $S_i{+}1 \times S_i{+}1$ attention matrix, where rows sum to one and weight j, k is:

$$a_{i,j,k}^{(\ell,h)} \propto \exp\left(\left(q_{i,j}^{(\ell,h)}\right)^\top k_{i,k}^{(\ell,h)} / \sqrt{d}\right). \tag{4}$$

Here, each instance j has L-dim. embeddings for query $q_{i,j}^{(\ell,h)}{=}W_Q^{(\ell,h)} h_{i,j}^{\ell-1}$, key $k_{i,j}^{(\ell,h)}{=}W_K^{(\ell,h)} h_{i,j}^{\ell-1}$, and value $v_{i,j}^{(\ell,h)}{=}W_V^{(\ell,h)} h_{i,j}^{\ell-1}$. Propagating embeddings via attention-weighted value averages over several layers and heads allows instance features to interact flexibly to inform the ultimate bag-level embedding.

4. Shifted Mean MIL Synthetic Data

We propose a new data-generating process designed to mimic several key challenges in real-world multiple-instance medical imaging tasks:

- Across the whole dataset, only a few features of many are discriminative (K of M).
- For each positively-labeled bag, only a few instances are relevant (R of S_i) and they are adjacent in a known 1D listing of all S_i instances.
- Context matters. Adjacent instances together provide stronger statistical signal than any one relevant instance's discriminative feature value alone.

The generative process for bag i first draws the bag's binary label and the number of instances in the bag

$$y_i \sim \text{Bern}(q_+), \quad S_i \sim \text{Unif}(\{S_{\text{low}}, \ldots, S_{\text{high}}\}). \tag{5}$$

Next, for negative bags we sample all features m for all instances j independently from a common Gaussian:

$$h_{i,j,k} \mid y_i{=}0 \sim \mathcal{N}(\mu, \sigma^2). \tag{6}$$

For positive bags, most instances and features are sampled from this same Gaussian. However, for the K discriminative features, we select R adjacent instances (using u_i to denote the starting index) and sample these from a Gaussian with *shifted mean*:

$$u_i \mid y_i{=}1 \sim \text{Unif}(\{1, \ldots, S_i - R + 1\}), \tag{7}$$

$$h_{ijk} \mid u_i, y_i{=}1 \sim \begin{cases} \mathcal{N}(\mu + \Delta, \sigma^2), & \text{if } j \in [u_i, u_i{+}R{-}1] \\ & \text{and } k \text{ is discrim.} \\ \mathcal{N}(\mu, \sigma^2), & \text{otherwise.} \end{cases}$$

Here $\Delta{>}0$ indicates the magnitude of shift for discriminative features. Setting $R{>}1$ indicates that context helps. Given a fixed μ, bags drawn from this process are more challenging to classify (even with knowledge of the true process) when Δ is smaller, R is smaller, $\frac{K}{M}$ is smaller, and σ is larger.

This data-generating process is illustrated in Fig. 2, depicting only one feature that is discriminative. In each positive bag, a different contiguous block of $R{=}3$ instances draw from the shifted mean Gaussian. If future work wanted to model correlations between features within an instance, the sampling of vector h_{ij} in Eq. (7) could be modified to draw from a multivariate Gaussian with a non-diagonal covariance matrix.

5. Bayes Estimator

Given a data-generating process, a *Bayes estimator* is a decision rule that minimizes the posterior expected loss with respect to the data-generating distribution (DeGroot, 1970; Murphy, 2022). It is an oracle upper bound on performance. By comparing conventional or recent MIL methods to the Bayes estimator for our synthetic dataset, we can quantify how close they come to the best possible performance.

Given a new bag h_i containing S_i instances and assuming our data-generating process defined above, a Bayes estimator for class label probability is:

$$p(y_i = 1 \mid h_i) = \frac{p(h_i \mid y_i = 1) p(y_i = 1)}{p(h_i)}. \tag{8}$$

Each term on the right-hand side can be computed in closed-form. For brevity we omit how random variable S_i cancels out here; see App. D for details. The denominator term is given by the sum rule:

$$p(h_i) = p(h_i \mid y_i{=}0) p(y_i{=}0) + p(h_i \mid y_i{=}1) p(y_i{=}1).$$

where we recall $p(y_i{=}1)$ is q_+. The class-conditional likelihood for the negative class factors over instances:

$$p(h_i \mid y_i = 0) = \prod_{j=1}^{S_i} \prod_{k=1}^{M} \text{NormPDF}(h_{ijk} \mid \mu, \sigma^2).$$

Each positive bag has a latent segment of R consecutive relevant instances. The class-conditional likelihood marginalizes out the unknown index u:

$$p(h_i \mid y_i{=}1) = \sum_{u=1}^{S_i - R + 1} \left[p(u \mid y_i{=}1) \prod_{j=1}^{S_i} \prod_{k=1}^{M} p(h_{ijk} \mid u, y_i{=}1) \right].$$

The two lines of Eq. (7) provide the necessary PDF values to evaluate the right hand side.

6. Experimental Results

Setup. In our experiments, we sample bag labels uniformly ($q_+{=}0.5$) and the number of instances per bag uniformly between $S_{\text{low}}{=}15$ and $S_{\text{high}}{=}45$. We use $M{=}768$ features to match the size of ViT-B/16

embeddings (Dosovitskiy et al., 2021) and use only a single discriminative feature (K=1). We set R=3 so context matters, and fix μ=0, σ=1. For main results in Fig. 1, we fix Δ=2. For results varying Δ, see the Appendix.

For each train set size N, we draw a training dataset of N bags from our data-generating process. We further draw a separate dataset of $\frac{1}{4}N$ bags for a validation set used to select hyperparameters. After selecting a final model at each N, we report AUROC performance on a common test set of 1000 bags.

For each MIL method, we try both *prediction-aggregation* and *embedding-aggregation* approaches when possible. We train models for 1000 epochs, with potential for early stopping. We explore a range of possible hyperparameters for each method, including learning rate in $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ and weight decay in $\{10^{0}, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 0\}$. Both early stopping and hyperparameter selection seek to maximize validation set AUROC.

Examining the results of our synthetic dataset experiments in Fig. 1, key findings are:

- **Conventional MIL cannot match the Bayes estimator when context matters** (R=3). Even given N=10000 bags, conventional deep MIL approaches (Max, Mean, ABMIL (Ilse et al., 2018)) deliver test AUROC at least 0.04 below the Bayes estimator in Fig. 1. Trends over N do not suggest this gap will close with more data.

- **TransMIL cannot match the Bayes estimator when R=3.** This is surprisingly, since TransMIL purports to handle context between instances. TransMIL's AUROC is at least 0.03 below the Bayes estimator at any tested N value.

- **smAP cannot quite match the Bayes estimator when R=3, though it is consistently the best method at almost all N.** At almost all tested N above 500, smAP (Castro-Macías et al., 2024) scored within 0.02 AUROC of the Bayes estimator. At the largest training set size of N=10000 in Fig. 1, the *prediction-aggregation* variant of smAP reached its highest test set AUROC. To determine whether the AUROC difference from our Bayes estimator to this best smAP model was statistically significant, we performed bootstrap analysis (see App. C). This analysis revealed that the 95% confidence interval of the AUROC difference did not include zero or any negative values, so we conclude that smAP remains inferior to the Bayes estimator even at N=10000. We expect this performance gap to increase for R>3 because the chain graph adjacency matrix used in smAP limits instance interaction to immediate neighbors.

- **The ability to capture adjacent instance context is a primary reason for the gap.** To verify that the context-dependent R=3 setting is what is important, we verified that for data drawn from a context-free R=1 version of our process, many MIL methods including max pooling, ABMIL, and TransMIL can all match the Bayes estimator with handcrafted parameters (see App. B).

- **The key limitation appears to be in the training process, though architecture changes may help too.** We were able to *handcraft* neural network parameters for TransMIL that closely match the Bayes estimator when R=3, as in Fig. B.3 in the Appendix. However, the training of pooling and classification layers yields consistently suboptimal predictions, even when N=10000. We conjecture that better regularization beyond simple weight decay would help. Some architecture changes are likely also beneficial if they introduce useful inductive bias. To demonstrate this latter point, in App. B, we show how adding context to ABMIL via convolutions over instances can help when R=3.

7. Conclusion

We designed an easy-to-implement synthetic dataset designed to mimic key challenges in MIL for medical imaging, especially the need for context from nearby instances. We then demonstrated the limited ability of conventional MIL on such data, by quantifying the performance gap between the optimal Bayes estimator. Correlated MIL methods like TransMIL are still notably worse than optimal even with a labeled dataset of 10000 bags. Very recent methods like smAP come closer, but still fall short.

Outlook. Our work reveals a key gap that must be overcome for MIL to succeed on real medical data. Many popular 3D brain scan datasets where MIL could help contain labeled datasets far smaller than the largest N tested here. For example, RSNA (Flanders et al., 2020), as used in Castro-Macías et al. (2024), has only 1150 scans for training and evaluation. We hope our synthetic dataset enables the development of correlated MIL methods that can be trained effectively with limited labeled data and make a difference in disease detection and treatment.

Acknowledgments

The authors gratefully acknowledge support from the Alzheimer's Drug Discovery Foundation and the U.S. National Institutes of Health (grant # 5R01NS134859-02). MCH is also supported in part by the U.S. National Science Foundation (NSF) via grant IIS # 2338962.

References

Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Francisco M. Castro-Macías, Pablo Morales-Álvarez, Yunan Wu, Rafael Molina, and Aggelos K. Katsaggelos. `Sm`: enhanced localization in Multiple Instance Learning for medical imaging classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Francisco M Castro-Macías, Francisco J Sáez-Maldonado, Pablo Morales-Álvarez, and Rafael Molina. torchmil: A pytorch-based library for deep multiple instance learning. *arXiv preprint arXiv:2509.08129*, 2025. URL https://github.com/Franblueee/torchmil.

Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, 1970.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Ji Feng and Zhi-Hua Zhou. Deep MIML Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

Adam E. Flanders, Luciano M. Prevedello, George Shih, Safwan S. Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T. Mongan, Anouk Stein, Felipe C. Kitamura, Matthew P. Lungren, Geetika Choudhary, Luciano Cala, Luís Coelho, Mads Mogensen, Fátima Morón, Eric Miller, Ichiro Ikuta, Vahe Zohrabian, Oran McDonnell, Christoph Lincoln, Luciano Shah, Devon Joyner, Ashish Agarwal, Richard K. Lee, and Jayashree Nath. Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. *Radiology: Artificial Intelligence*, 2(3): e190211, 2020.

Giles M. Foody. Classification accuracy comparison: Hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority. *Remote Sensing of Environment*, 113(8):1658–1663, 2009. ISSN 0034-4257. URL https://www.sciencedirect.com/science/article/pii/S0034425709000923.

Zhongyi Han, Benzheng Wei, Yanfei Hong, Tianyang Li, Jinyu Cong, Xue Zhu, Haifeng Wei, and Wei Zhang. Accurate Screening of COVID-19 Using Attention-Based Deep 3D Multiple Instance Learning. *IEEE Transactions on Medical Imaging*, 39(8): 2584–2594, 2020.

Ethan Harvey, Wansu Chen, David M. Kent, and Michael C. Hughes. A Probabilistic Method to Predict Classifier Accuracy on Larger Datasets given Small Pilot Data. In *Machine Learning for Health (ML4H)*, 2023.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. In *International Conference on Machine Learning (ICML)*, 2018.

Hassan Keshvarikhojasteh, Josien P. W. Pluim, and Mitko Veta. Multi-head Attention-based Deep Multiple Instance Learning. In *Proceedings of the MICCAI Workshop on Computational Pathology*, 2024.

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 (CVPR)*, 2021.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

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.

Oded Maron and Tomás Lozano-Pérez. A Framework for Multiple-Instance Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1997.

Kevin S. Murphy. *Probabilistic Machine Learning: An Introduction*, chapter 5.1 Bayesian decision theory. MIT Press, 2022.

Pedro O. Pinheiro and Ronan Collobert. From Image-Level to Pixel-Level Labeling With Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Gwenolé Quellec, Guy Cazuguel, Béatrice Cochener, and Mathieu Lamard. Multiple-Instance Learning for Medical Image and Video Analysis. *IEEE Reviews in Biomedical Engineering*, 10:213–234, 2017.

Edward Raff and James Holt. Reproducibility in Multiple Instance Learning: A Case For Algorithmic Unit Tests. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Daniel Shao, Richard J. Chen, Andrew H. Song, Joel Runevic, Ming Y. Lu, Tong Ding, and Faisal Mahmood. Do Multiple Instance Learning Models Transfer? In *International Conference on Machine Learning (ICML)*, 2025.

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. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yuhui Zhang, Shih-Cheng Huang, Zhengping Zhou, Matthew P. Lungren, and Serena Yeung. Adapting Pre-trained Vision Transformers from 2D to 3D through Weight Inflation Improves Medical Image Segmentation. In *Machine Learning for Health (ML4H)*, 2022.

Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. Deep Multi-instance Networks with Sparse Label Assignment for Whole Mammogram Classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2017.

Appendix A. Experiments Varying the Class Separation Δ

Here, we study how the prediction quality of various MIL methods varies with the class separation parameter Δ, which controls how similar the discriminative feature values are in the data-generating process for the positive class (mean $\mu + \Delta$, variance σ^2) and negative class (mean μ, variance σ^2).

Experiments here use $R = 3$, $N = 400$, and $\sigma = 1$. Given these settings, the chance of a positive discriminative feature value landing closer to the negative mean (μ) than the positive mean ($\mu + \Delta$) is $Pr(h_{ijm} \leq \mu + \frac{1}{2}\Delta) = \Phi(\frac{(\mu+\frac{1}{2}\Delta)-(\mu+\Delta)}{\sigma}) = \Phi(-\frac{1}{2}\Delta)$, where Φ is the standard Normal CDF. When $\Delta = 2$ as in the main paper, this is 0.1587; when $\Delta = 3$, this is 0.0068; when $\Delta = 4$, this is 0.0228; when $\Delta = 5$, this is 0.0062. For $\Delta \geq 4$, any individual discriminative feature has over 97% chance to correctly indicate the class label without much need for its surrounding context.

Indeed, in the actual experimental results below, as Δ increases, we see all methods improve and approach near perfect classification, except for the Mean pooling baseline.

However, for modest Δ values like 2 or below, there's a noticeable gap between the ideal Bayes estimator and the actual method performance. We expect the relatively small gap in AUROC between smAP at the largest N and the Bayes estimator shown in Fig. 1 for $\Delta = 2$ would notably increase as Δ gets smaller: note below a gap of roughly 0.1 between smAP and the Bayes estimator at $\Delta = 1$.

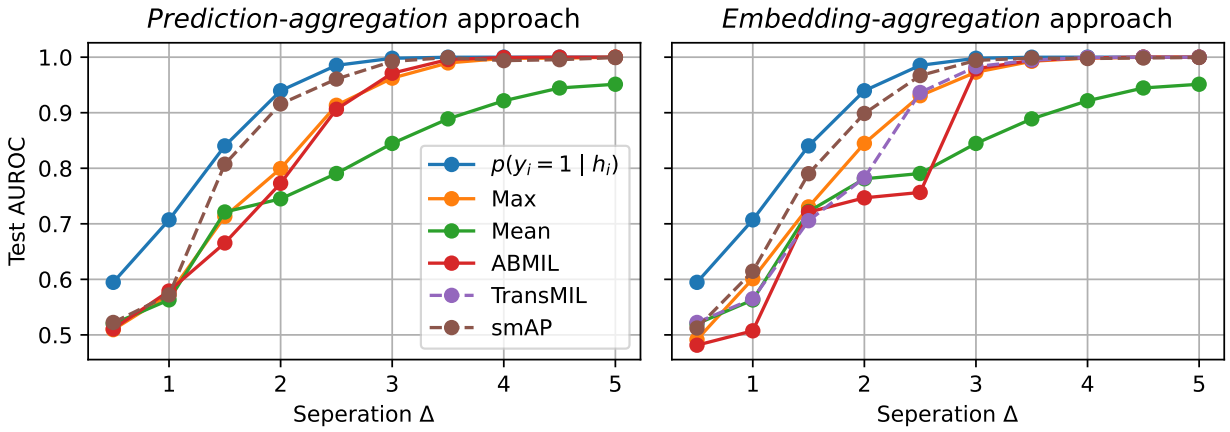

Figure A.1: Test AUROC as a function of separation Δ. All data drawn from our Shifted Mean MIL data-generating process fro binary classification, with R=3 and N=400.

Appendix B. Handcrafted Parameters

We set classifier weights corresponding to each discriminative feature to one, all other weights to zero, and the bias parameter to $-\frac{\Delta}{2}$. For attention pooling, we use the linear part of the tanh function to create attention weights proportional to each feature's linear score. For instance convolutions, we set the center R weights to one and all other weights and biases to zero to sum up each feature's linear score over R instances.

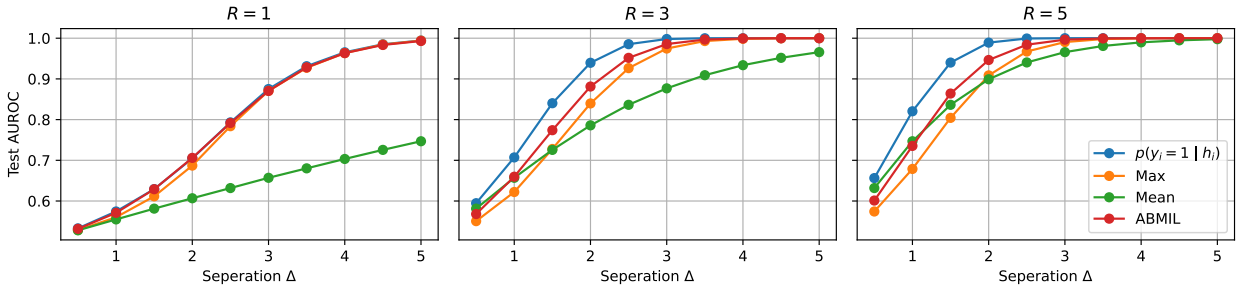

Figure B.1: Handcrafted parameters (*prediction-aggregation* approach).

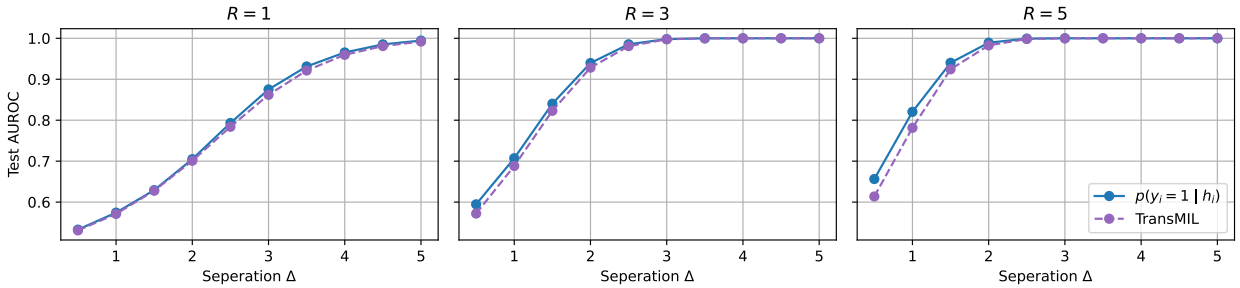

Figure B.2: Handcrafted parameters (*embedding-aggregation* approach).

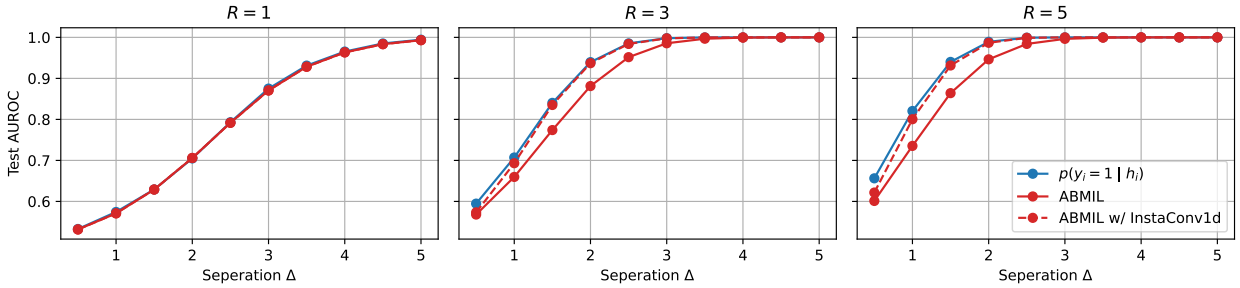

Figure B.3: Handcrafted parameters (*prediction-aggregation* approach).

Appendix C. Bootstrap Analysis

We use bootstrapping (Foody, 2009) to access the statistical significance of the AUROC difference between the Bayes estimator and smAP for the *prediction-aggregation* approach trained with N=10000. We report the mean and 95% confidence interval over 500 subsamples of the test set.

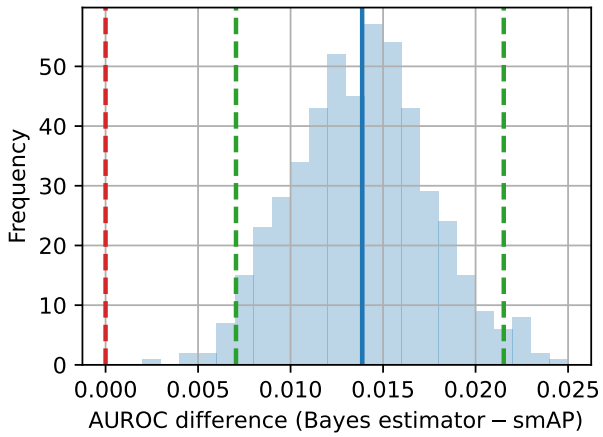

Figure C.1: Bootstrap analysis comparing the AUROC difference between the Bayes estimator and smAP for the *prediction-aggregation* approach trained with N=10000.

Appendix D. Details Needed for Derivation of Bayes Estimator

One reviewer asked whether the generative process for the number of instances per bag S_i needs to be accounted for in the Bayes estimator derivation. We can show that because this distribution is uniform and not dependent on class label, that is $p(S_i, y_i) = p(y_i)p(S_i)$, the relevant term can be factored out of the numerator and denominator in the label-given-features posterior and canceled

$$p(y_i = 1 \mid h_i, S_i) = \frac{p(h_i \mid y_i = 1, S_i)p(y_i = 1)\cancel{p(S_i)}}{p(h_i \mid y_i = 0, S_i)p(y_i = 0)\cancel{p(S_i)} + p(h_i \mid y_i = 1, S_i)p(y_i = 1)\cancel{p(S_i)}}. \tag{9}$$

For brevity, we omitted the dependence on S_i in many statements in the main paper. The complete class-conditional likelihood PDFs needed to evaluate the right-hand-side above are:

$$p(h_i|y_i = 0, S_i) = \prod_{j=1}^{S_i} \prod_{k=1}^{M} \text{NormPDF}(h_{ijk}|\mu, \sigma^2) \tag{10}$$

$$p(h_i|y_i = 1, S_i) = \sum_{u=1}^{S_i - R + 1} \left[\underbrace{\frac{1}{S_i - R + 1}}_{p(u|y_i=1,S_i)} \prod_{j=1}^{S_i} \prod_{k=1}^{M} \underbrace{\text{NormPDF}(h_{ijk}|\mu + \Delta\delta_{u,j,k}, \sigma^2)}_{p(h_{ijk}|u,y_i=1,S_i)} \right] \tag{11}$$

where $\delta_{u,j,k}$ is 1 if $j \in [u, u + R - 1]$ and k is a discriminative feature, and 0 otherwise.