Synthetic Data Reveals Generalization Gaps in Correlated Multiple Instance Learning

Ethan Harvey[1] Dennis Johan Loevlie[1] Michael C. Hughes[1]

[1]*Department of Computer Science, Tufts University, Medford, MA, USA*

Contributions

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 (MIL). 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.

Shifted Mean MIL Synthetic Data

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

$$y_i \sim \text{Bern}(q_+), \quad S_i \sim \text{Unif}(\{S_{\text{low}}, \dots, S_{\text{high}}\}). \tag{1}$$

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

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

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*:

$$u_i \mid y_i{=}1 \sim \text{Unif}(\{1, \dots, S_i - R + 1\}), \tag{3}$$

$$h_{ijk} \mid u_i, y_i{=}1 \sim \begin{cases} \mathcal{N}(\mu + \Delta, \sigma^2), & \text{if } j \in [u_i, u_i{+}R{-}1] \\ & \text{and } k \text{ is discrim.} \\ \mathcal{N}(\mu, \sigma^2), & \text{otherwise.} \end{cases} \tag{4}$$

Bayes Estimator

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:

$$p(y_i = 1 \mid h_i) = \frac{p(h_i \mid y_i = 1)p(y_i = 1)}{p(h_i)}. \tag{5}$$

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.

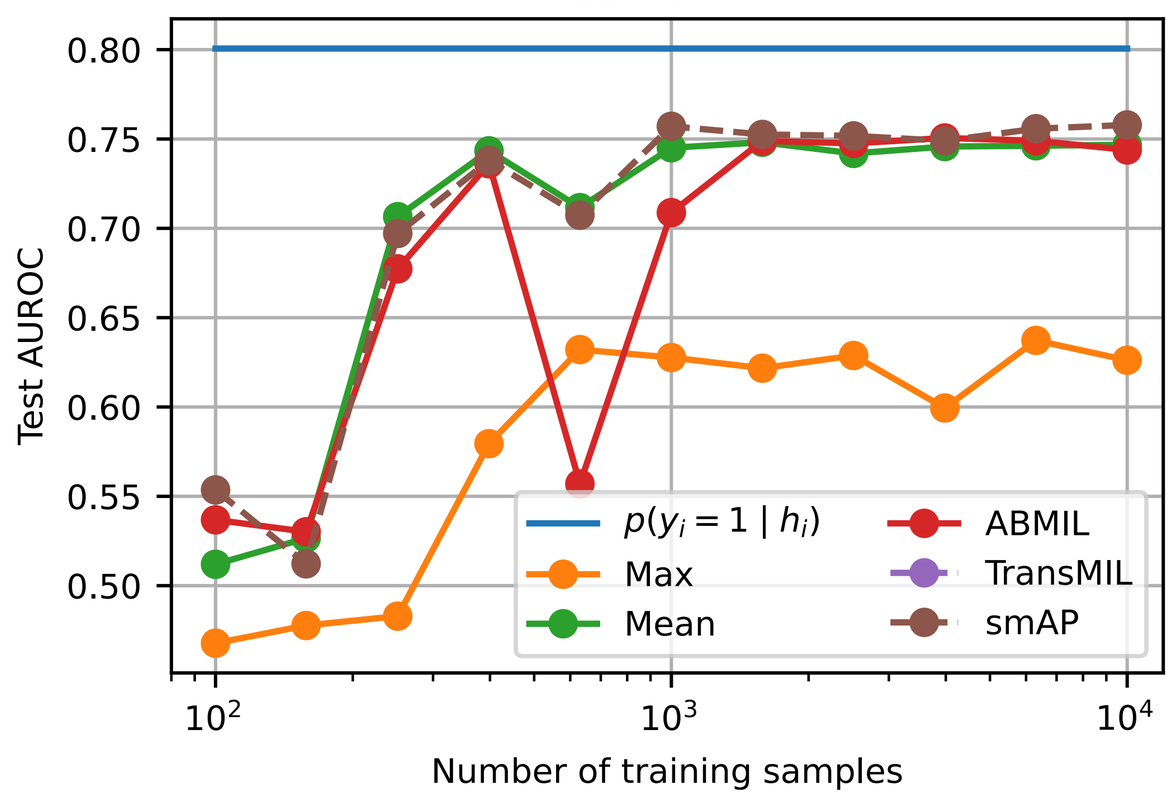

Figure 1: Test AUROC as a function of training set size N. All data is drawn from our *Shifted Mean MIL* data generating process for binary classification, with $R{=}10$, $\Delta{=}0.5$. Conventional MIL approaches (Max, Mean, ABMIL) cannot match the Bayes estimator $p(y_i = 1 \mid h_i)$ as they do not 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. **Takeaway: Our work reveals a need for data-efficient MIL that better accounts for context between instances.**

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.

References

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.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

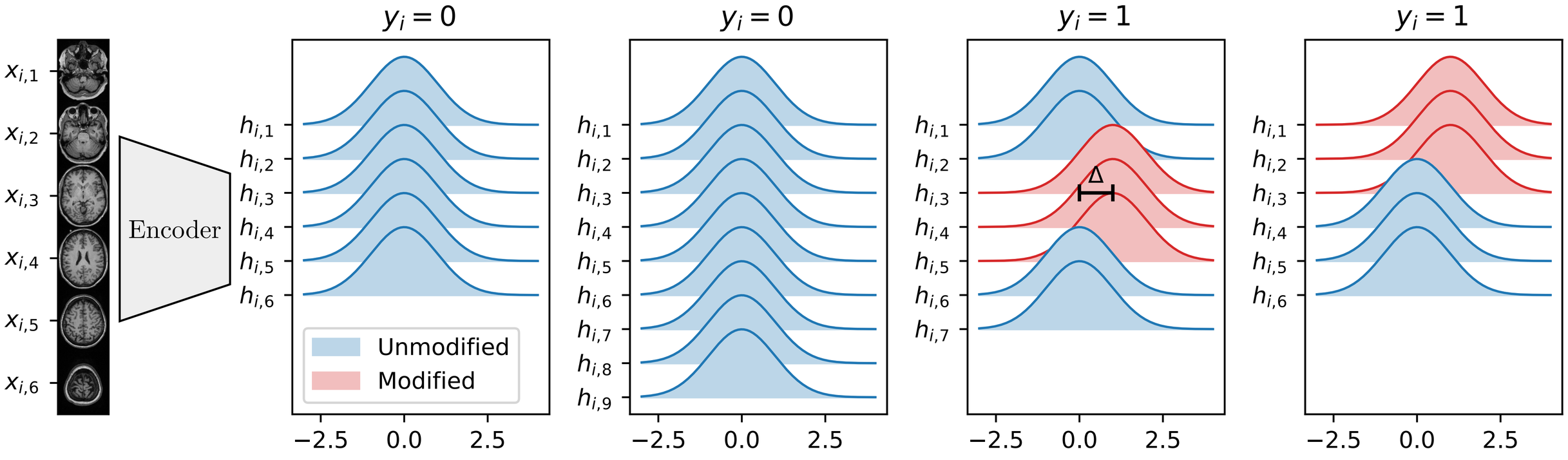

Figure 2: Example data-generating distributions for a discriminative feature for negative ($y_i{=}0$) and positive ($y_i{=}1$) "bags" of S_i instances drawn from our *Shifted Mean MIL* synthetic data. Setting $R{=}3$ means context around modified instances (in red) can help.

github.com/tufts-ml/correlated-mil Ethan.Harvey@tufts.edu