Topological Inductive Bias fosters Multiple Instance Learning in Data-Scarce Scenarios

Anonymous authors
Paper under double-blind review

Abstract

Multiple instance learning (MIL) is a framework for weakly supervised classification, where labels are assigned to sets of instances, i.e., bags, rather than to individual data points. This paradigm has proven effective in tasks where fine-grained annotations are unavailable or costly to obtain. However, the effectiveness of MIL drops sharply when training data are scarce, such as for rare disease classification. To address this challenge, we propose incorporating topological inductive biases into the data representation space within the MIL framework. This bias introduces a topology-preserving constraint that encourages the instance encoder to maintain the topological structure of the instance distribution within each bag when mapping them to MIL latent space. As a result, our Topology Guided MIL (TG-MIL) method enhances the performance and generalizability of MIL classifiers across different aggregation functions, especially under scarce-data regimes. Our evaluations show average performance improvement of 15.3% for synthetic MIL datasets, 2.8% for MIL benchmarks, and 5.5% for rare anemia datasets compared to current state-of-the-art MIL models, where only 17–120 samples per class are available. We make our code publicly available at `https://anonymous.4open.science/r/TGMIL-59B6`.

1 Introduction

Multiple Instance Learning (MIL) is a variant of weakly supervised learning that operates without annotations for individual data. In MIL, each 'bag,' which represents a group of instances, is assigned a single label (Lu et al., 2020). A bag is labeled positive if it contains at least one positive instance; otherwise, it is labeled negative. For example, in blood sample–based disease classification, each sample can be viewed as a bag of individual cells. The sample is labeled positive if it contains any diseased cells and negative otherwise.

Proper instance representation is critical for ensuring model reliability in clinical decision-making. However, data scarcity, a common issue in the medical diagnosis of rare diseases, hinders the ability of MIL models to learn such representations. In such cases, the need for MIL-based training approaches that operate in the scarce-data regime is paramount.

To address this gap, a potential solution is leveraging additional structure from data through inductive biases, a strategy that has shown promise in other machine learning domains Goyal & Bengio (2022). By treating each bag as a point cloud in a high-dimensional data space,we can define an inductive bias based on the topological features of the bag. These features should be preserved after embedding bag instances into the model's latent space, ensuring that the essential topological properties of the data are maintained. Being able to capture fundamental topological principles of data at multiple scales, topological algorithms[1] recently arose as a source of such inductive biases, permitting the integration into deep learning models (Hensel et al., 2021). The primary appeal of such algorithms lies in their robustness to noise and perturbations, resulting in *stable multi-scale representations*. Also, these algorithms improve generalizability, and predictive performance when a pronounced geometrical-topological signal is present in the data (Horn et al., 2022; Waibel et al., 2022),

[1] Despite their name, these algorithms also capture geometrical aspects of data, but we will refrain from writing *geometrical-topological algorithms* for brevity.

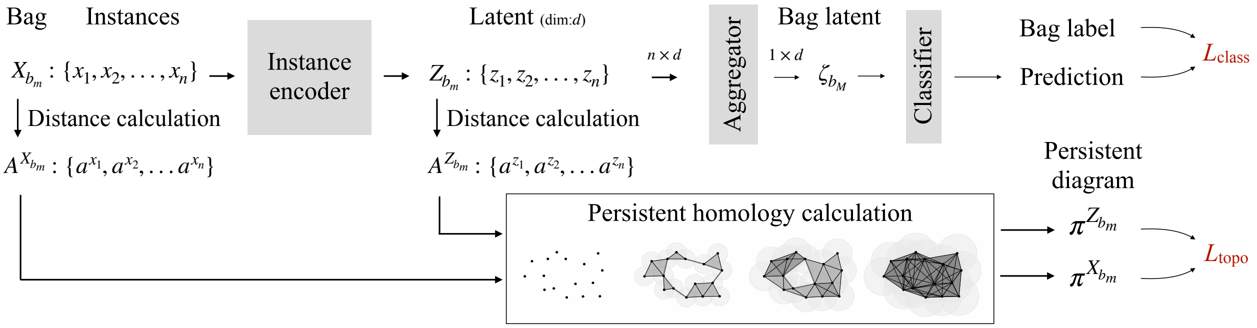


Figure 1: Topologically Guided Multiple Instance Learning (TG-MIL): We calculate the distance matrix A of input instances x_i inside each bag X_{b_m}. Subsequently, we apply persistent homology based on the Vietoris-Rips complex by treating each bag's instances as a point cloud. We employ the same process for the latent feature vectors Z of each bag. Generating shape descriptors (*persistence diagrams* π) for both the latent space and the image space representations of the bag, we calculate a topological loss (L_{topo}) and combine it with the standard MIL loss (L_{class}).

even in the presence of *singular structures*, which preclude the use of standard techniques (von Rohrscheidt & Rieck, 2023).

We introduce *Topologically Guided Multiple Instance Learning (TG-MIL)*, a data-centered solution to address the challenges of training MIL with scarce training data in an end-to-end manner. By leveraging multi-scale shape descriptors on the level of MIL bags, we develop a scheme that ensures the preservation of crucial topological information in the latent space of our model (see Figure 1 for a schematic overview). We demonstrate that maintaining the topological bias inherent in a bag's data distribution enhances the performance of MIL classifiers, even with varying amounts of data. This topological bias, reflecting each bag's natural structure and relationships, carries critical information for accurate classification. Preserving this intrinsic structure allows MIL classifiers to learn and generalize more effectively, improving performance accuracy, robustness, and adaptability across different training data. The **main contributions** of our work are:

- We develop TG-MIL, the first method to improve the generalizability of MIL to data-scarce scenarios.
- Our method can be integrated with any MIL aggregation strategy, improving performance under data scarcity in an end-to-end training setting.
- TG-MIL outperforms the state-of-the-art on MIL benchmarks and rare anemia classification.

2 Background

MIL Architectures. MIL architectures typically comprise three key components: an instance encoder, an aggregation function, and a classifier (Figure 1). Given a collection of bags $b_1, \ldots, b_M$, each bag contains a set of instances, represented as $X_{b_m} := \{x_1, \ldots x_n\}$ with n denoting the number of instances in the bag. An instance encoder f_θ with parameters θ transfers instance data into a latent space, yielding feature vectors $z_i := f_\theta(x_i)$. The aggregation function then creates a global representation of a bag ζ_{b_m} from these embedded instances, from which the classifier head predicts the overall label of the bag.

Geometry & Topology. Our work is based on recent advances in topological machine learning (Hensel et al., 2021), a nascent field that aims to leverage geometry and topology from data to elicit improved representations. We employ *persistent homology*, a technique for calculating multi-scale geometrical-topological information from data Edelsbrunner & Harer (2009). Persistent homology treats data as a point cloud and aims to define this cloud using simple geometric elements (simplices) that describe the overall geometry of the data, such as connected components, loops, higher-dimensional voids, and geometric properties like curvature or convexity (Bubenik et al., 2020; Turkes et al., 2022). This information is collected in a set of *persistence diagrams*, which serve as multi-scale topological descriptors. These descriptors are calculated by approximating the data with a simplicial complex—a generalized graph—typically based on distance functions like the Euclidean distance (Figure 1). Recent work proved that persistent homology can be integrated with deep learning models, leading to a new class of hybrid models that are capable of capturing topological aspects of data. Such models have shown exceptional performance as regularization terms in different applications (Chen et al., 2019; Vandaele et al., 2022; Waibel et al., 2022).

3 Related Work

Recent advances in MIL can be categorized into two main methods and use cases: i) Methods using *transfer learning* for feature extraction and only focusing on training aggregation mechanisms to address the challenge of capturing class-relevant instances within bags. Given the constraints of limited computational memory, for scenarios like histopathology images (where bags contain thousands of tiled patches, i.e., instances), these approaches keep the instance-level representations fixed and do not support end-to-end training (Shao et al., 2021; Zhang et al., 2022; Liu et al., 2023; Tang et al., 2023). ii) *End-to-end* approaches that optimize both bag aggregation and instance representation during training. Ilse et al. (2018) introduced the Attention-based MIL (APMIL), which learns instance-level attention weights, and the Gated Attention MIL (GAPMIL) variant, which incorporates a gating mechanism to modulate those weights. Recently, APMILwD and GAPMILwD (Zhu et al., 2025) extended the classical attention-based frameworks by introducing instance-level dropout, which regularizes attention learning and significantly improves generalization, achieving state-of-the-art results on classic MIL benchmarks. The attention-based pooling (see appendix equation 13 and 14) approach may lack the necessary reliability, as it does not uniformly enhance representation across all instances. However, the attention-based pooling mechanism (see Appendix, equations 13–14) may still lack reliability, as it does not uniformly enhance representation across all instances For example, in medical diagnosis and drug discovery, where accurate identification of individual components (e.g., cells in microscopic blood sample images, chemical compounds) is crucial for making accurate predictions at the sample level. Other end-to-end methods have focused on designing alternative aggregation mechanisms to better capture complex instance relationships and improve MIL classification. BDRMIL (Huang et al., 2022) modeled pairwise relations between instances within a bag, capturing inter-instance dependencies. DistNet (Oner et al., 2023) proposed a distribution-aware formulation that represents each bag via statistical properties of its instance embeddings, enabling aggregation through learned distributional representations rather than attention scores. RGMIL (Du et al., 2023) further enhanced instance-level representation through a regressor-guided aggregator that aligns instance- and bag-level predictions, improving the signal flow through the encoder. Although this approach refines instance-level representation and enhances overall MIL performance, it struggles to accurately capture and represent the nuanced variations within the data when dealing with scarce training data.

In real-world medical applications, like classifying rare anemia disorders, key factors such as the severity of cell deformation and the ratio of deformed cells in a blood sample are paramount. Thus, it is essential that instance-level representations capture the distribution of cell classes while maintaining an interpretable bias toward deformation severity, as this ensures both the reliability and interpretability of the MIL model's predictions. The state-of-the-art approach by Kazeminia et al. (2022) in this domain refines aggregation techniques—anomaly-aware pooling—to encourage the encoder to push healthy cells toward a normal distribution. However, this approach's dependency on the quantity of training data available constrains its potential to achieve higher accuracy, as optimal performance of its learning-based aggregation still requires substantial data.

To overcome the challenges posed by data scarcity, we introduce a novel inductive bias that incentivizes models to preserve geometrical-topological information in latent representations. Our method thus falls into the second category of end-to-end techniques and emphasizes the crucial role of data representations in MIL.

4 Methods

Our approach treats each bag as a point cloud in a high-dimensional space whose geometrical-topological features should be adequately captured by the model. Each instance influences the bag's 'shape,' with positive instances notably altering its shape in comparison to the distribution of negative samples. We thus need a descriptor that captures the characteristics of a point cloud while remaining stable to perturbations and invariant under transformations like translations and rotations that are irrelevant for determining the overall shape. Persistent homology provides a suitable descriptor as Sheehy (2014) demonstrates that critical topological-geometrical features captured by persistent homology are approximately even under projections or embeddings of the data, making it highly robust. The calculation of persistent homology requires a choice of distance metric to measure the interaction of each data point with others. While our framework remains agnostic to the specific choice of distance metric, we have chosen to utilize the per-pixel Euclidean distance. This decision stems from its ease of calculation and empirical evidence from Moor et al. (2020a), demonstrating the adequacy of Euclidean distances in capturing topological features. This enables us to transform the point cloud of a bag X_{b_m} into a distance matrix $A^{X_{b_m}}$ (see Figure 1). Next, we use the distance matrix A representing the bag's point cloud and calculate its associated Vietoris–Rips complex $\mathrm{VR}(A, \epsilon)$, a simplicial complex defined by connecting points if they lie within a distance ϵ of each other, i.e.,

$$\mathrm{VR}(A, \epsilon) = \{\sigma \subseteq A \mid \forall a_i, a_j \in \sigma, d(a_i, a_j) \leq \epsilon\}. \tag{1}$$

Here, $d(a_i, a_j)$ denotes the distance between points a_i and a_j in A. The topology of VR changes as we vary ϵ. Formally, this leads to a filtration of simplicial complexes $\{\mathrm{VR}(A, \epsilon_0), \mathrm{VR}(A, \epsilon_1), \ldots, \mathrm{VR}(A, \epsilon_m)\}$, with an ordered sequence of distance thresholds $0 = \epsilon_0 < \epsilon_1 < \ldots < \epsilon_m$ (Figure 1). Persistent homology tracks the 'birth' and 'death' of topological features (e.g., connected components (0D), loops (1D), and voids (2D)), across this sequence, represented in a persistence diagram by points (β_1, β_2), where $\beta_1 = \epsilon_i$ and $\beta_2 = \epsilon_j$. This diagram constitutes a summary of the shape of each bag, measured by multi-scale topological features and yields a set of d-dimensional persistence diagrams, described in the form of *persistence pairings* $\pi^{X_{b_m}}$ of points in the input space, representing the bag's topological signature.

Our objective is to inject this signature as an inductive bias into the model to enhance its robustness and generalizable instance representation. Thus, we capture the distance matrix $A^{Z_{b_m}}$ and signature of the bag's point cloud in the latent space $\pi^{Z_{b_m}}$ and define a loss term to penalize the encoder f_θ for any inconsistency in preserving the bag's signature during projection from input to latent space. To this end, we utilize the topological loss proposed by Moor et al. (2020b), which addresses the challenge of backpropagating through topological descriptors. First we compare the distances of the input persistence pairs with the distances of the latent persistence pairs in the input space $L_{X_{b_m} \to Z_{b_m}}$. Then, we compare the distances of the input persistence pairs with the distances of the latent persistence pairs in latent space $L_{Z_{b_m} \to X_{b_m}}$. We need to account for both losses because we want to align both mappings (data → latent and latent → data). The final topological loss is defined as

$$L_{\text{topo}} := L_{X_{b_m} \to Z_{b_m}} + L_{Z_{b_m} \to X_{b_m}}, \tag{2}$$

where

$$L_{X_{b_m} \to Z_{b_m}} := \frac{1}{2} \left\| A^{X_{b_m}} \left[\pi^{X_{b_m}}\right] - A^{Z_{b_m}} \left[\pi^{X_{b_m}}\right] \right\|^2, \tag{3}$$

and

$$L_{Z_{b_m} \to X_{b_m}} := \frac{1}{2} \left\| A^{Z_{b_m}} \left[\pi^{Z_{b_m}}\right] - A^{X_{b_m}} \left[\pi^{Z_{b_m}}\right] \right\|^2. \tag{4}$$

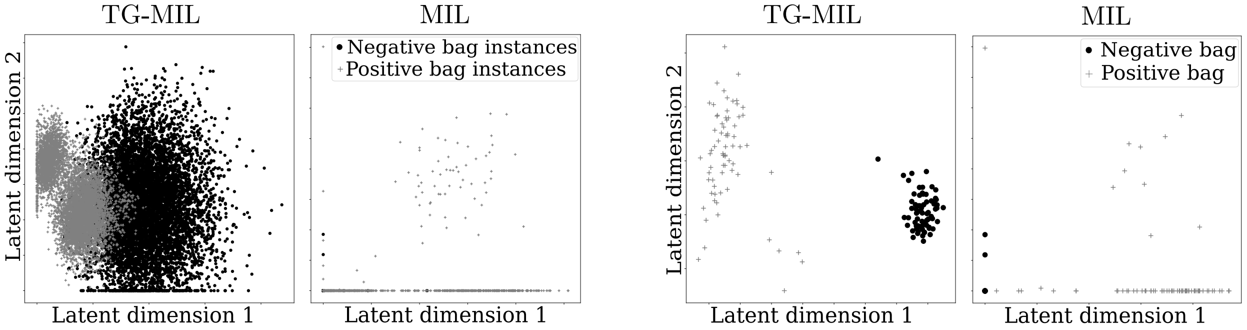
Our framework is flexible to integrate any aggregation function for representing the whole bag ζ_{b_M}. With this, the classifier head, incorporating a linear regressor and softmax functions, predicts the bag's label. Similar to standard MIL models, we train the MIL classification head using cross-entropy loss. Our formulation also gives rise to a variant of a multi-classifier head approach like the auxiliary loss that Kazeminia et al. (2022) proposed. The final loss of our topologically guided MIL (TG-MIL) framework, L_{total}, is the weighted sum of the MIL classification loss L_{class} and topological term L_{topo}:

$$L_{\text{total}} = L_{\text{class}} + \lambda L_{\text{topo}}, \tag{5}$$

where λ is a hyperparameter to adjust the influence of the topological loss. Appendix A.1 presents more details of MIL architecture examples.

Complexity and Parameters. The computational complexity involved in calculating certain topological features is comparable to the rate of the inverse Ackermann function (Cormen et al., 2022), which increases significantly slower compared to the rate of increasing n. Therefore, the computational complexity of the topological signature calculation of a bag containing n instances is dominated by the calculation of pairwise distances, i.e., $\mathcal{O}(n^2)$, considering that we only capture 0-dimensional topological features. The topological signature calculation does not introduce any additional learnable parameters, thereby keeping the model's parameter size unchanged. It merely introduces one topological loss and one hyperparameter, denoted as λ.

Limitations. The primary limitation of our approach is that the calculation of topological features does *not* exhibit favorable scaling parameters in case that higher-order topological features are required. While our implementation supports topological features of arbitrary dimension, their calculation scales progressively worse; connected components, i.e., 0-dimensional features, can still be efficiently calculated (see previous paragraph), but higher-order features may prove limiting. We plan on investigating mitigation strategies in future work, using, e.g., approximate filtrations (Sheehy, 2013) or distributed computations (Wagner et al., 2021).



(a) 2D representation of instances.

(b) 2D representation of bags.

Figure 2: TG-MIL achieves a more distinguished representation of negative and positive bags. For a toy dataset, instances are sampled from a hypersphere when projecting them to the 2D latent space (a), leading to a more distinguished latent representation of bags (b).

Toy dataset. To demonstrate the importance of preserving the topology of bags in MIL, we utilize a toy dataset. In this dataset, negative instances are sampled randomly from a 100-dimensional space, while positive instances are sampled from the surface of a 100-dimensional sphere—an established geometric object. To satisfy the positive bag definition of MIL, the sphere overlaps with the space of random negative instances. We consider a simple 2-layer encoder projecting instances from the 100-dimensional space to a visualizable representation (see Figure 2). We apply our framework utilizing regressor-guided aggregation (Du et al., 2023) as the baseline MIL model. Figure 2 illustrates the resulting instance and bag representations, contrasting the scenarios with and without topological guidance. TG-MIL preserves the topology of positive instances, resembling a circle, as the expected 2D projection of a hypersphere. As a result, the aggregated bag's latent is more distinct, leading to a higher classification accuracy (MIL and TG-MIL yield 0.55 ± 0.05 and 0.80 ± 0.22 accuracy, respectively, averaged over 5 runs).

5 Experiments

We evaluate the performance of TR-MIL across different datasets. Synthetic MIL datasets provide a controlled environment to investigate TG-MIL's performance under varying training data quantities. MIL benchmarks offer insight into its general effectiveness across heterogeneous data modalities while facilitating a fair comparison with state-of-the-art methodologies. Furthermore, our evaluation contains a real-world application, i.e., the classification of *rare anemia*, that presents additional challenges inherent to the MIL problem domain, such as the contribution of severity and the ratio of positive instances within the bag class.

5.1 Synthetic Datasets

To evaluate the robustness of the TG-MIL framework across varied MIL data properties, including the number of training bags, instance image complexity, and bag sizes, we draw on the methods outlined by Ilse et al. (2018). We create two series of synthetic datasets: the first comprises bags of MNIST images as instances, and the second comprises bags of Fashion MNIST (Xiao et al., 2017) images a more challenging scenario with complex visual data. In constructing the MIL synthetic datasets, a bag is labeled positive if it contains at least one instance of the digit '9' in MNIST or the label 'Dress' in Fashion MNIST; otherwise, it is labeled negative. We construct distinct training datasets containing a total number of 10, 14, 20, 50, 100, and 200 bags to evaluate the influence of the quantity of training data. Additionally, we explore different amounts of instances per bag, sampling them from Gaussian distributions with mean and standard deviations defined as $(10, 2)$, $(50, 10)$, and $(100, 20)$, respectively. Positive bags are defined as those containing at least one positive instance, accounting for up to 20% of the instances within the bag.

Models. We use a deep instance encoder architecture introduced by Ilse et al. (2018) (see Table 6 in the appendix for details). It consists of two convolutional layers with a kernel size of 5, a stride of 1, and ReLU activation functions. These layers generate 20 and 500 feature maps, respectively, followed by a fully-connected layer. The attention network comprises two linear layers, resulting in a final output dimension of 128 followed by 1. The topological signature of input instances is calculated on image space and latent space, applying pixel-wise Euclidean distance of instance images and latent feature vectors (Figure 1).

Results. We evaluate the effectiveness of topological guidance utilizing three aggregation functions in MIL: max pooling, average pooling, and attention-based pooling (Ilse et al., 2018), which serves as the baseline for numerous studies in the field, in addition to the regressor guided pooling technique (Du et al., 2023). We analyze the average F1-score and its standard deviation for different numbers of training bags (Figure 3) and bag sizes (Figure 6 in appendix) over five runs. Without topological guidance, models trained with few training bags perform poorly, akin to random guessing, while adding topological guidance provides a reasonable complexity for the encoder to resolve overfitting (Figure 7 in Appendix shows learning curves of MIL and TG-MIL for MNIST synthetic data) and improves the MIL model performance across both datasets. Notably, topological guidance narrows the performance gap between basic aggregations of max pooling and average pooling compared to advanced attention and regressor guided pooling techniques.

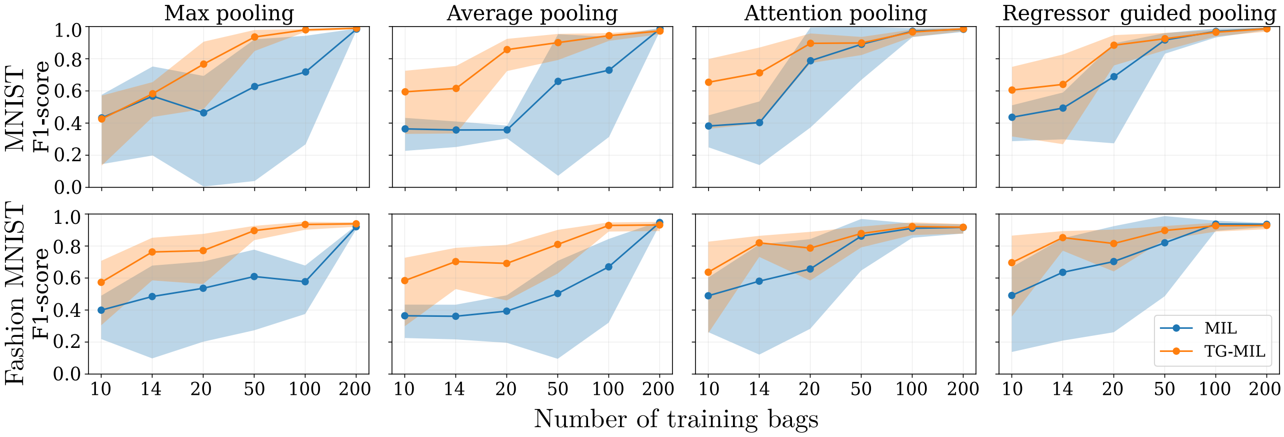


Figure 3: TG-MIL outperforms MIL across different pooling strategies (max, average, attention, and regressor guided) on both MNIST and FashionMNIST datasets in the data-scarce regime. For each number of training bags, the plots show the mean and standard deviation of F1-scores over 5 runs for different bag sizes containing 10, 50, and 100 instances (15 runs in total), highlighting consistent gains especially for smaller training sets.

We conducted a statistical analysis using the Wilcoxon rank-sum test (Haynes et al., 2013) to assess the significance of the improvements resulting from topological guidance (Table 1). This non-parametric test compares the p-values of results between MIL experiments with and without topological guidance. Given

Table 1: The Wilcoxon rank-sum test shows the significance of topological guidance enhancement over MIL models. Results are reported as p-values for each amount of training data and aggregation technique. P-values less than 0.002 indicate significant enhancement after multiple testing correction(*).

		Number of training bags					
	Pooling	10	14	20	50	100	200
MNIST	Max	6.6×10^{-1}	7.7×10^{-1}	3.7×10^{-3}	1.4×10^{-2}	3.8×10^{-2}	4.8×10^{-1}
	Average	1.3×10^{-3}*	7.5×10^{-5}*	3.1×10^{-6}*	3.1×10^{-1}	3.7×10^{-1}	1.7×10^{-2}
	Attention	7.9×10^{-3}	2.3×10^{-3}	8.0×10^{-1}	3.8×10^{-1}	6.0×10^{-1}	6.5×10^{-1}
	Regressor	7.1×10^{-2}	7.8×10^{-2}	6.6×10^{-3}	7.1×10^{-1}	5.1×10^{-1}	3.9×10^{-1}
Fashion	Max	1.1×10^{-1}	1.5×10^{-3}*	5.8×10^{-3}	1.3×10^{-2}	1.1×10^{-2}	9.7×10^{-2}
	Average	1.2×10^{-1}	2.8×10^{-4}*	9.7×10^{-5}*	2.9×10^{-2}	1.1×10^{-1}	5.6×10^{-1}
	Attention	3.8×10^{-1}	5.1×10^{-3}	1.1×10^{-1}	6.5×10^{-1}	2.3×10^{-1}	4.3×10^{-1}
	Regressor	9.3×10^{-2}	6.2×10^{-3}	2.8×10^{-1}	3.5×10^{-1}	2.9×10^{-1}	4.1×10^{-1}

the 24 tests for each dataset, we applied the Bonferroni correction (Weisstein, 2004) to account for multiple comparisons, reducing the significance level from 0.05 to approximately ($\frac{0.05}{24} \approx$) 0.002. Topological guidance significantly enhances classification in 7 out of 48 cases at this significance level, particularly when using average pooling (Table 1.

5.2 MIL Benchmarks

Dataset. We assess the performance of TG-MIL using five benchmark MIL datasets commonly employed in evaluating end-to-end MIL methods in the literature. These include three image-based datasets (FOX, TIGER, and ELEPHANT) (Dietterich et al., 1997), each comprising 200 bags. In these datasets, only tabular features of (instances) are available. Although these classic benchmarks are relatively small and of limited quality, we include them to ensure a fair and consistent comparison with previous MIL approaches.

Additionally, we evaluate our method on the MUSK1 and MUSK2 datasets (Andrews et al., 2002), which contain data of 92 and 102 molecules, respectively. In these datasets, each molecule is represented by a bag of instances, with each instance corresponding to a different molecular conformation. The number of instances per bag ranges from as few as 1 to as many as 1044, providing a comprehensive assessment of our model's adaptability and robustness across different scales of data representation.

Models. We used an identical encoder architecture in the literature to clarify and ensure a fair comparison with existing MIL methods. This architecture includes 2 linear layers with a ReLU activation, projecting input features into a 512-dimensional space for both layers. The only modification in our setup is integrating a topological signature calculator into instances' input and latent space.

Results. The original RGMIL model used 231 features for FOX, TIGER, and ELEPHANT datasets and 167 features for MUSK1 and MUSK2 datasets, including a last feature representing the repeated label of the bag for each instance. However, in our re-implementation, we followed the standard benchmark settings of 230 and 166 features for the respective datasets to align with previous works and provide a comprehensive comparison (see Table 2). We run both the MIL and TG-MIL models (using regressor guided aggregation) 5 times, applying 10-fold cross-validation and reporting the average optimal performance of the model during the training.

When using this instance feature vector, we observed a decline in the performance of our reimplemented RGMIL baseline, which aligns with the standard benchmark configuration. Among all prior methods listed in Table 2, we specifically selected RGMIL for reimplementation, as it provides a simpler MIL framework without additional instance-discrimination modules introduced in more recent models such as DistNet (Oner et al., 2023) and APMILwD/GAPMILwD (Zhu et al., 2025).

While TG-MIL trails the dropout-based variants of Zhu et al. (2025) on MUSK1 and FOX, it outperforms them on MUSK2 and achieves the top results on TIGER and ELEPHANT, suggesting that topological guidance confers broader gains in generalization beyond small molecular or classical image feature settings.

Table 2: Topological guidance improves the classification accuracy (%) of SOTA approaches in most of MIL benchmarks. Among previous methods, we specifically reimplemented RGMIL for our analysis. Other results (gray) are collected from papers proposed by Ilse et al. (2018) (APMIL and GAPMIL), Yan et al. (2018) (DPMIL), Li et al. (2021) (DSMIL), Huang et al. (2022) (BDRMIL), Oner et al. (2023) (DistNet), Zhu et al. (2025) (APMILwD and GAPMILwD), and Du et al. (2023) (RGMIL).

Method	MUSK1	MUSK2	FOX	TIGER	ELEPHANT
APMIL (2018)	89.2±4.0	85.8±4.8	61.5±4.3	83.9±2.2	86.8±2.2
GAPMIL (2018)	90.0±5.0	86.3±4.2	60.3±2.9	84.5±1.8	85.7±2.7
DPMIL (2018)	90.7±3.6	92.6±4.3	65.5±5.2	89.7±2.8	89.4±3.0
DSMIL (2021)	93.2±2.3	93.0±2.0	72.9±1.8	86.9±0.8	92.5±0.7
BDRMIL (2022)	92.6±7.9	90.5±9.2	62.9±11.0	86.9±6.6	90.8±5.4
DistNet (2023)	92.3±7.1	93.2±6.7	68.0±7.5	86.4±5.4	90.0±7.7
APMILwD (2025)	96.4±3.3	95.4±1.9	**78.9±4.3**	91.7±3.6	93.4±4.6
GAPMILwD (2025)	**96.7±1.9**	95.8±2.1	78.8±1.6	91.9±3.3	92.7±3.3
RGMIL (2023)	94.0±7.0	92.0±10.6	71.4±10.7	84.2±8.8	91.5±4.2
TG-MIL (ours)	94.6±7.8	**97.0±4.2**	74.7±5.4	**96.1±4.0**	**94.1±5.4**

More analysis on these experiments unveil a tendency of the RGMIL model to overfit, which is mitigated by including topological guidance (Figure 8 in appendix). This phenomenon is characterized by the MIL classifier performing best during the initial training epochs.

5.3 Anemia Classification

The diagnosis of anemia relies on the presence of a minority of red blood cells in a patient's blood sample that shows morphological deformations associated with the disease. Anemia disorders lead to various aberrant shapes such as sickle-shaped (SCD), crumpled or perforated (thalassemia), star-shaped (Xero), or even spherical (HS) cells. These deformations can manifest with varying degrees of severity and in different proportions, while it is also possible for other cell types unrelated to anemia conditions to coexist.

Dataset. Our dataset consists of 521 microscopy images of blood samples obtained from patients who underwent various treatments at different times. Each sample comprises 4 to 12 images, each containing 12 to 45 cells. The data is distributed among five classes: (i) SCD with 13 patients and 170 samples, (ii) Thalassemia with 3 patients and 25 samples, (iii) Xero with 9 patients and 56 samples, (iv) HS with 13 patients and 89, as well as (v) healthy control group consisting of 33 individuals and 181 samples. Similar to previous works, we implement a patient-centric approach by dividing the dataset into three equivalent folds. This division allocates two folds for training and reserves one for test. This dataset is generated using a pre-trained segmentation method to crop single-cell images (instances) from the whole sample (bag). Cropped images are zero-padded to be the same size, and their encoded extracted features ($4 \times 4 \times 256$ each cell) in the segmentation model are used as the input of the MIL instance encoder (see Table 7 in appendix). We posit that features extracted by the segmentation model, irrespective of the cell type, may lack crucial shape information and thus potentially manipulate the topology of the bag and capture the topological signature of each bag directly from the image instances and the 500-dimensional latent space.

Models. For a fair comparison, we apply topological guidance to the previous MIL architecture used for this application, where the instance encoder contains 3 convolutional layers followed by 2 ReLu and Tanh activation functions and a 2 linear layers to obtain a latent representation of instances in a 500-dimensional space.

Table 3: Topological guidance improves classification accuracy (%) for all pooling strategies in Anemia classification. We apply it to different MIL methods with various pooling techniques containing Average, Anomaly-aware (Kazeminia et al., 2022), Attention (Sadafi et al., 2020), and Max pooling. Numbers depict average classification performance and standard deviation from 3 cross-validation runs. Best performance is indicated by bold text. We compare classification performance without (✗) and with (✓) topological guidance, with the superior result underlined.

Pooling	Average		Anomaly-aware		Attention		Max	
L_{topo}	✗	✓	✗	✓	✗	✓	✗	✓
Accuracy	72.3±7.0	**81.3±2.5**	77.9±3.7	79.5±1.2	73.7±3.8	77.8±1.6	64.3±5.8	71.4±5.6
F1-Score	70.5±7.4	**80.3±3.1**	76.7±4.0	77.0±1.8	72.4±3.8	74.7±1.6	63.0±5.0	68.8±5.4
AUROC	89.9±2.7	**93.7±4.4**	89.1±4.3	90.9±2.5	91.6±3.0	91.9±2.5	84.8±2.8	89.7±3.1
Recall	59.7±7.8	**65.1±5.0**	63.2±3.2	66.0±3.3	59.3±6.4	60.4±2.3	52.8±8.6	53.8±4.7
Precision	61.8±7.1	**79.1±12.0**	67.4±4.5	69.7±4.6	64.9±4.9	73.2±7.9	52.9±9.6	63.4±7.5

Results. We employ five standard evaluation metrics: Accuracy, F1-Score, Area Under the Receiver Operating Characteristic Curve (AUROC), Precision, and Recall. All metrics are macro-weighted because of the dataset imbalance. Table 3 shows that topological guidance improves the performance of MIL models using *all* aggregation functions and results in higher average performance, often resulting in reduced variance. Topologically guided MIL with *average pooling* surpasses other aggregation schemes. This aligns with our findings from experiments on synthetic datasets, where we observe that topological guidance particularly narrows the gap between the performance of the MIL employing different aggregation functions. The inherent ambiguity in the anemia dataset for MIL suggests that enhancing instance projection in latent space via average pooling is more effective than attention pooling, as it better captures the ratio of positive instances. Without topological guidance, scarce training data impede the instance encoder from generating meaningful, generalizable latent representations. However, integrating topological inductive bias into the latent space mitigates these challenges, significantly improving MIL performance.

Instance-level analysis. Figure 4 shows anomaly scores with and without topological guidance. Without topological guidance, the anomaly detector assigns different scores to visually similar instances, indicating inconsistency. This inconsistency is mitigated with topological guidance, highlighting a challenge in the model's ability to evaluate similar data points uniformly and demonstrating the effectiveness of topological guidance in enhancing model explainability. Further illustrating this point, we visualize the distance matrix of instances within a bag in the input space and compare them with their corresponding matrices in the latent space, both with and without topological guidance. Figure 5 shows that MIL with topological guidance better preserves the distances between bag instances in the latent space projection. In contrast, MIL without topological guidance selects a limited number of deformed cells and pushes them far away from other instances in the bag. This indicates that, without topological guidance, only a few instances are projected far from the majority, explaining the observed inconsistency in anomaly scores for deformed shapes.

In addressing potential inquiries regarding our choice of topological guidance over a distance-preservation-based loss, it's noteworthy to emphasize the distinct advantages of our approach. Topological loss is particularly robust against noise and highly effective in high-dimensional spaces. It exhibits scale invariance, a critical feature that enables preserving the distance pattern of instances within a bag. This aspect underscores the strategic advantage of our chosen method, confirming the efficacy of topological guidance in maintaining the integrity of instance relationships in the latent space, thereby enhancing both the model's performance and its explainability.

CO2 Emission Related to Experiments. All experiments were conducted for 228 hours on institute's infrastructure with 0.432 kgCO$_2$eq/kWh carbon efficiency and A100 PCIe 40/80GB hardware (TDP 250W),

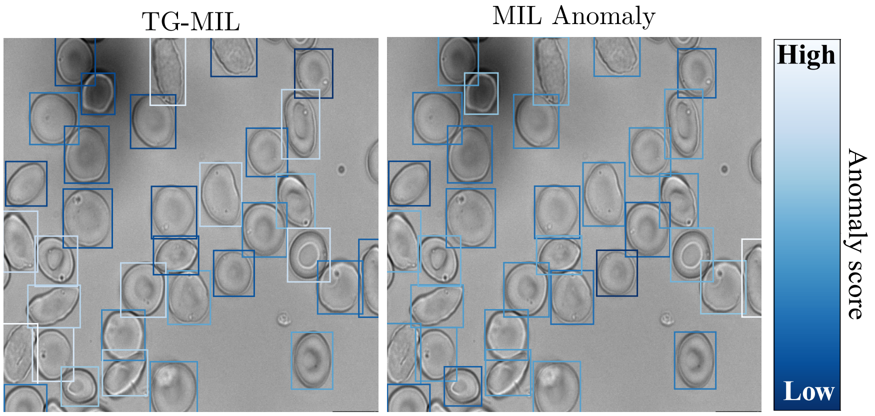


Figure 4: Topological guidance enhances the model's ability to identify disease-relevant (sickle) cells more effectively. TG-MIL results in more uniform anomaly scores for deformed cells, in contrast to the varied scores resulting from MIL Anomaly.

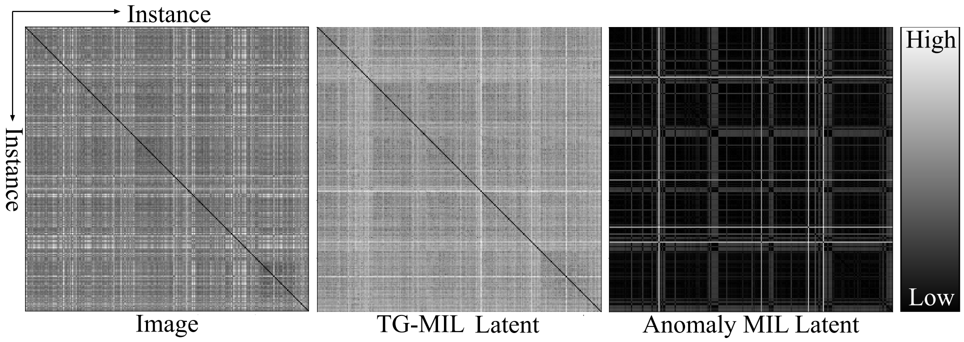


Figure 5: Heatmaps of distances between instance images (left), TG-MIL latent vectors (middle), and anomaly-aware MIL latent vectors (right) for the same bag. Topological guidance preserves the topological characteristics of the image space, making the latent space more closely resemble the original high-dimensional space.

resulting in 24.62 kgCO_2eq total emissions with no direct offset. Calculations utilized the MachineLearning Impact calculator Lacoste et al. (2019).

6 Conclusion

We present TG-MIL, a novel approach that leverages geometrical-topological properties of bag inputs as an inductive bias to the latent of the latent of MIL. This is achieved by employing a topological loss term within the MIL objective function. Our method ensures that intrinsic geometrical-topological characteristics of bags in the input space are consistently maintained in their latent projections. Testing our approach on different datasets, we show that the preservation of data topology in the latent MIL models leads to substantial improvements in terms of predictive performance and generalization performance, especially when dealing with scarce training data. While the datasets used in our experiments involve non-trivial MIL problems, the individual instances (images) exhibit relatively simple visual structures, e.g., centered, grayscale, background-free. In such settings, pixel-level differences provide an effective reference for defining topology in the latent space. However, in more visually complex domains, where instances are more visually complex, e.g., diversity of textures, color, or higher background noise, pixel-based topology may become less reliable, as it can transfer irrelevant visual noise into the latent representation. The robustness of TG-MIL on benchmark datasets, where the original images were not available and only pre-extracted image features served as inputs, further suggests that defining topology over higher-level feature representations is a viable alternative. This observation does not imply a direct comparison with image-based inputs but rather highlights the potential of feature-level topology as a more general inductive bias for visually complex image instances.

As for future research directions, we plan to explore alternative methods for describing images geometry and topology, with a particular focus on cubical complexes that can directly operate on images. Additionally, we aim to investigate the geometrical and topological properties of bag spaces, leveraging recent advances in metric geometry, including the Gromov–Hausdorff distance, which has previously been utilized to characterize shapes in related studies Chazal et al. (2009).

References

Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. *Advances in neural information processing systems*, 15, 2002.

Peter Bubenik, Michael Hull, Dhruv Patel, and Benjamin Whittle. Persistent homology detects curvature. *Inverse Problems*, 36(2):025008, 2020.

Frédéric Chazal, David Cohen-Steiner, Leonidas J. Guibas, Facundo Mémoli, and Steve Y. Oudot. Gromov–Hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum*, 28(5):1393–1403, 2009.

Chao Chen, Xiuyan Ni, Qinxun Bai, and Yusu Wang. A topological regularizer for classifiers via persistent homology. In *International Conference on Artificial Intelligence and Statistics*, pp. 2573–2582, 2019.

Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2022.

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.

Zhaolong Du, Shasha Mao, Yimeng Zhang, Shuiping Gou, Licheng Jiao, and Lin Xiong. Rgmil: Guide your multiple-instance learning model with regressor. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Herbert Edelsbrunner and John Harer. Computational topology: An introduction. *American Mathematical Society*, 9(2):117–138, 2009.

Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266):20210068, 2022.

Winston Haynes et al. Wilcoxon rank sum test. *Encyclopedia of systems biology*, 3(1):2354–2355, 2013.

Felix Hensel, Michael Moor, and Bastian Rieck. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4, 2021.

Max Horn, Edward De Brouwer, Michael Moor, Yves Moreau, Bastian Rieck, and Karsten Borgwardt. Topological graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2022. URL `https://openreview.net/forum?id=oxxUMeFwEHd`.

Shiluo Huang, Zheng Liu, Wei Jin, and Ying Mu. Bag dissimilarity regularized multi-instance learning. *Pattern Recognition*, 126:108583, 2022.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

Salome Kazeminia, Ario Sadafi, Asya Makhro, Anna Bogdanova, Shadi Albarqouni, and Carsten Marr. Anomaly-aware multiple instance learning for rare anemia disorder classification. In *25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 341–350. Springer, 2022.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

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*, pp. 14318–14328, 2021.

Kangning Liu, Weicheng Zhu, Yiqiu Shen, Sheng Liu, Narges Razavian, Krzysztof J Geras, and Carlos Fernandez-Granda. Multiple instance learning via iterative self-paced supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3355–3365, 2023.

Cheng Lu, Bing Han, Yiqun Liu, Gang Niu, Rui Zhang, En Li, Yizhou Zhou, and Shaoting Zhang. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 26(9):1301–1309, 2020.

Michael Moor, Max Horn, Karsten Borgwardt, and Bastian Rieck. Challenging euclidean topological autoencoders. In *TDA {\&} Beyond*, 2020a.

Michael Moor, Max Horn, Bastian Rieck, and Karsten Borgwardt. Topological autoencoders. In *International Conference on Machine Learning*, pp. 7045–7054, 2020b.

Mustafa Umit Oner, Jared Marc Song Kye-Jet, Hwee Kuan Lee, and Wing-Kin Sung. Distribution based mil pooling filters: Experiments on a lymph node metastases dataset. *Medical Image Analysis*, 87:102813, 2023.

Ario Sadafi, Asya Makhro, Anna Bogdanova, Nassir Navab, Tingying Peng, Shadi Albarqouni, and Carsten Marr. Attention based multiple instance learning for classification of blood cell disorders. In *23rd International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 246–256. Springer, 2020.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

Donald R. Sheehy. Linear-size approximations to the vietoris–rips filtration. *Discrete & Computational Geometry*, 49(4):778–796, 2013. doi: 10.1007/s00454-013-9513-1.

Donald R Sheehy. The persistent homology of distance functions under random projection. In *Proceedings of the thirtieth annual symposium on Computational geometry*, pp. 328–334, 2014.

Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4078–4087, 2023.

Renata Turkes, Guido Montufar, and Nina Otter. On the effectiveness of persistent homology. In *Advances in Neural Information Processing Systems*, 2022.

Robin Vandaele, Bo Kang, Jefrey Lijffijt, Tijl De Bie, and Yvan Saeys. Topologically regularized data embeddings. In *International Conference on Learning Representations*, 2022.

Julius von Rohrscheidt and Bastian Rieck. Topological singularity detection at multiple scales. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning (ICML)*, number 202 in Proceedings of Machine Learning Research, pp. 35175–35197. PMLR, 2023.

Alexander Wagner, Elchanan Solomon, and Paul Bendich. Improving metric dimensionality reduction with distributed topology, 2021. arXiv:2106.07613.

Dominik J. E. Waibel, Scott Atwell, Matthias Meier, Carsten Marr, and Bastian Rieck. Capturing shape information with multi-scale topological loss terms for 3D reconstruction. In *25th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 150–159, 2022.

Eric W Weisstein. Bonferroni correction. *https://mathworld. wolfram. com/*, 2004.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yongluan Yan, Xinggang Wang, Xiaojie Guo, Jiemin Fang, Wenyu Liu, and Junzhou Huang. Deep multi-instance learning with dynamic pooling. In *Asian Conference on Machine Learning*, pp. 662–677. PMLR, 2018.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18802–18812, 2022.

Wenhui Zhu, Peijie Qiu, Xiwen Chen, Zhangsihao Yang, Aristeidis Sotiras, Abolfazl Razi, and Yalin Wang. How effective can dropout be in multiple instance learning ? In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=qsYHqLFCH5.

A Appendix

A.1 Experimental MIL Architectures

Component	**Configuration**
Input channels	100
Instance encoder	Linear(100 → 64), ReLU, Linear(64 → 2), ReLU
Latent dimension	2
Pooling	Regressor guided pooling
Classifier head	Linear (100 → 2)
Loss	BCEWithLogitsLoss and TopoRegLoss
Optimizer	Adam (lr: 0.0005)
λ	0.005

Table 4: Architecture of the MIL model for the toy experiment.

The RGMIL (Du et al., 2023) model uses regressor guided pooling technique. In this model, the regressor (with parameters W and B) calculates the binary probability value of instance latent representation z_i as

$$(p_i^+, p_i^-) := W^T z_i + B, \tag{6}$$

where p_i^+ and p_i^- show the probability of instance z_i to belong to the positive or negative class. Then it gets the difference between these two achieved probabilities,

$$p_i = p_i^+ - p_i^-, \tag{7}$$

normalizes the result

$$\omega_i = \frac{p_i - \mathbb{E}[p_i]}{\sqrt{Var(p_i)}}, \tag{8}$$

and applies softmax on it

$$\alpha_i = \frac{\exp(\omega_i)}{\sum_{j=1}^{n} \exp(\omega_j)}. \tag{9}$$

where, α_i specifies the pooling weight of z_i. Then the latent of bag is

$$\zeta_{b_m} = \sum_{i=1}^{n} \exp(\alpha_i z_i). \tag{10}$$

Table 4 specifies more details of the MIL architecture we developed to classify toy dataset.

Components	**Elephant, Fox, Tiger**	**Musk1, Musk**
Input channels	230	166
Instance encoder	Linear(231, 512), ReLU, Linear(512, 512), ReLU	
Latent Dimension	512	
Pooling	Regressor guided pooling	
Classifier head	Linear (512 → 2)	
Loss	BCEWithLogitsLoss and TopoRegLoss	
Optimizer	Adam (lr: 0.00005, Betas: [0.9, 0.999])	
Max Epochs	40	
λ	0.05	

Table 5: Architecture of the MIL Model for Benchmarks

Table 5 shows the settings of this architecture.

Parameter	**Max Pooling**	**Average Pooling**	**Attention Pooling**	**RGP Pooling**
Pooling	Max	Average	Attention	Regressor guided
In dimension	28 × 28			
Input channel	1			
Instance encoder	Linear(1→20), ReLU, Linear(20→50), ReLU, Linear(50→500), ReLU			
Latent dimension	500			
Attention latent dimension	128			
Linear layer	500 × 2			
Loss	BCEWithLogitsLoss and TopoRegLoss			
Optimizer	Adam (LR: 0.005)		Adam (LR: 0.0005)	
Batch size	1			
Max epochs	100			

Table 6: General architecture and configurations of TG-MIL Model for synthetic datasets over different pooling strategies. The value λ is not reported here as it differs between datasets, training budgets, and bag sizes. Please find its relevant value in the source code.

For synthetic data we employed same architecture for both MIL-MNIST and MIL-FashionMNIST datasets. We explored different aggregation functions containing max pooling

$$\zeta_{b_m} = \max_{i \leq n} z_i, \tag{11}$$

average pooling

$$\zeta_{b_m} = \frac{\sum_{i=1}^{n} z_i}{n}, \tag{12}$$

and attention pooling with parameters W and V

$$\zeta_{b_m} = \sum_{i=1}^{n} a_i z_i, \tag{13}$$

where

$$a_i = \frac{\exp(W^T \tanh(V z_i^T))}{\sum_{i=1}^{n} \exp(W^T \tanh(V z_i^T))}. \tag{14}$$

Parameter	**Max Pooling**	**Average Pooling**	**Aux Attention**	**Anomaly Detection**
Image dimentions	64 × 64			
Image channels	1			
Features channels	256			
Instance encoder	Conv2D(256→301), ReLU, Conv2D(301→500), ReLU, Conv2D(500→650), Tanh, Linear(650→500))			
Latent dimension	500			
Instance classifier head	-		Linear(500 → 500), Linear(500 → 5)	
Pooling	Max	Average	Attention	Anomaly
Attention layer	-		Linear(500→128), Tanh, Linear(128→1)	
Bag classifier head	Linear(500 → 2)			
Loss	bag CrossEntropyLoss and TopoRegLoss		CrossEntropyLoss (bag and instance) and TopoRegLoss	
Optimization	Adam (lr=0.0005			
Learning rate	0.0005			
Max epochs	300			
Early stopping	patience: 50			
Image input channels	1			
λ	0.005			

Table 7: Configuration of MIL model for anemia classification with different pooling strategies

Table 6 shows the settings of this architecture.

For anemia classification we followed the architecture of the state of the art in this application (Kazeminia et al., 2022) (Table 7). This method introduced anomaly score to be considered in addition to attention values to estimate the importance of each instance. To this end the distribution of negative instances is estimated from negative bags by fitting a gaussian mixture model on their latent representation. The anomaly score of each instance latent z_i is calculated as

$$d_i = \sqrt{(z_i - \mu)^T \Sigma^{-1} (z_i - \mu)}, \tag{15}$$

where μ and Σ are mean and covariance of the fitted GMM on negative distribution. Then the pooling weight of the instance is calculated as a linear combination of attention score a_i and anomaly score d_i. With this the bag latent is

$$\zeta_{b_m} = \sum_{i=1}^{n} (W_{D_i} d_i + W_{A_i} a_i) z_n. \tag{16}$$

The other consideration of this approach is the formulation of $Loss_{class}$ with a dual classifier head (Sadafi et al., 2020) that comprises a bag classifier head and an instance classifier head. The bag classifier head is trained using a cross-entropy loss function L_{bag}, calculated as the difference between the predicted bag label and the corresponding ground truth label for the bag. The instance classifier head is trained using a cross-entropy loss function L_{Instance} that utilizes the noisy labels of instances as the repeated labels of the bag for all instances. The final MIL classification loss is calculated as

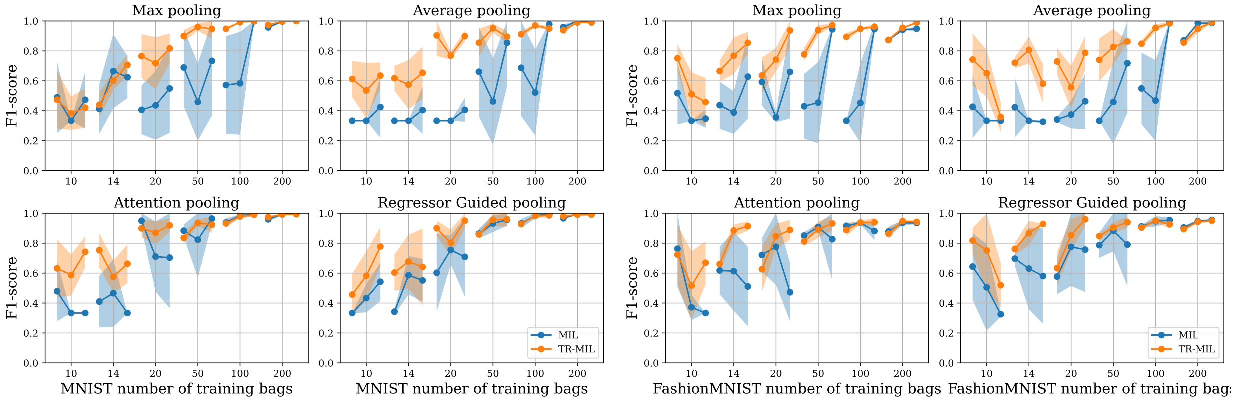
$$L_{class} = (1 - \gamma) L_{\text{bag}} + \gamma L_{Instance}, \tag{17}$$

where γ is a coefficient that decreases as with epoch number increasing.

A.2 Detailed performance on synthetic dataset

Table 6 distinctly illustrates how topological regularization addresses this issue in MIL employing both average and attention aggregation functions. This effect is particularly pronounced with smaller bags, leading to improved robustness (lower variance across multiple runs) and higher accuracy on average.

Furthermore, the figure effectively highlights the advantages of the attention mechanism over average pooling in enhancing MIL performance, especially when trained with limited data. However, for a small amount of training bags, topological regularization improves performance by generating more accurate and robust results.



(a) MIL MNIST

(b) MIL FashionMNIST

Figure 6: TG-MIL outperforms MIL models irrespective of the aggregation function when subjected to a limited amount of training bags. For each number of training bags, the average and standard deviation for the F1-score of the model's performance in 5 runs for each bag size of 10, 50, and 100 (in total 15 runs) is shown.

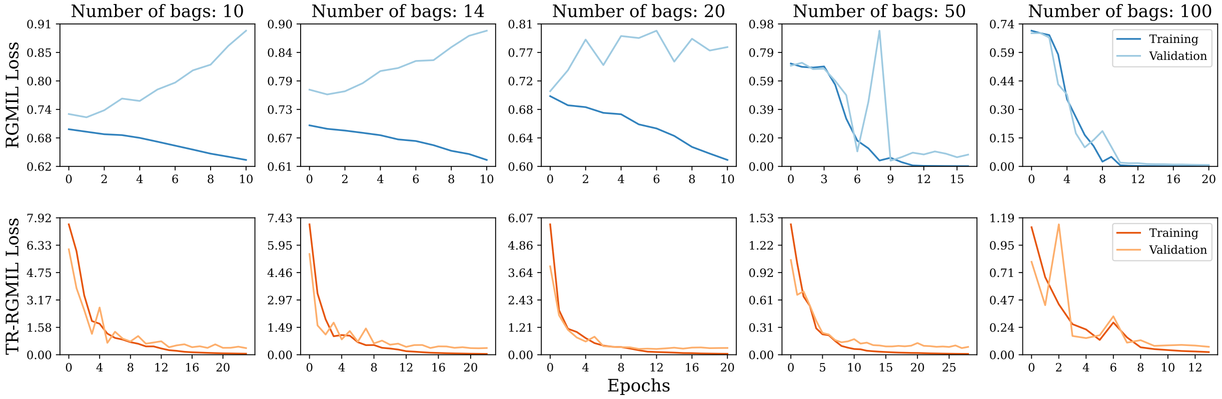


Figure 7: Topological guidance enhances the RGMIL model generalizability for scarce training data. Each column shows learning curves of models trained with 10 bags, each containing 10 instances on average.

A.3 Learrning curves on Benchmarks

The RGMIL model tends to overfit when performing on benchmark datasets, given their limited data size. Figure 8 displays the learning curves of training RGMIL alongside TR-RGMIL. The introduction of topological regularization addresses overfitting in RGMIL and results in a significant improvement in its classification performance.

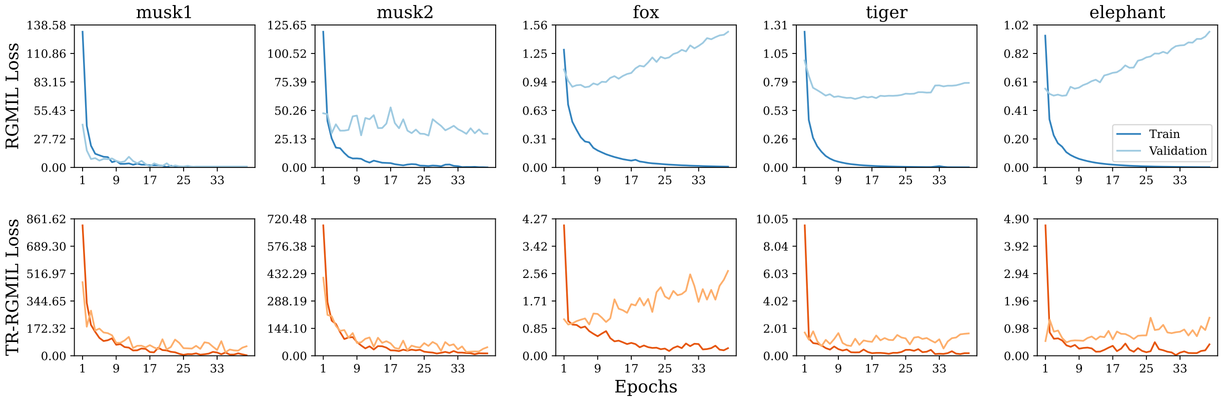


Figure 8: Topological regularization enhances RGMIL model generalizability for Benchmarks. Each column shows learning curves achieved by a MIL benchmark dataset.