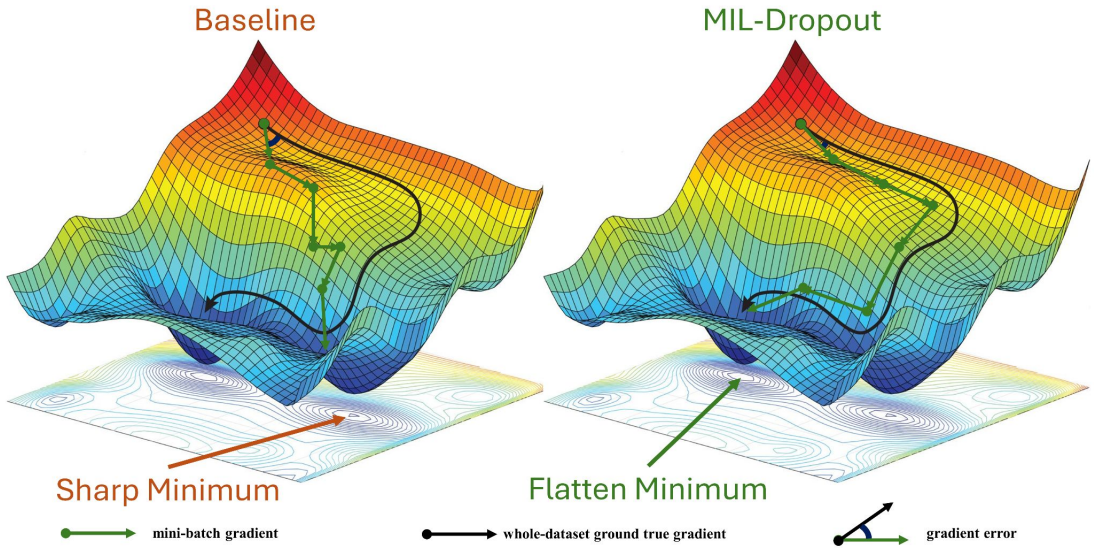
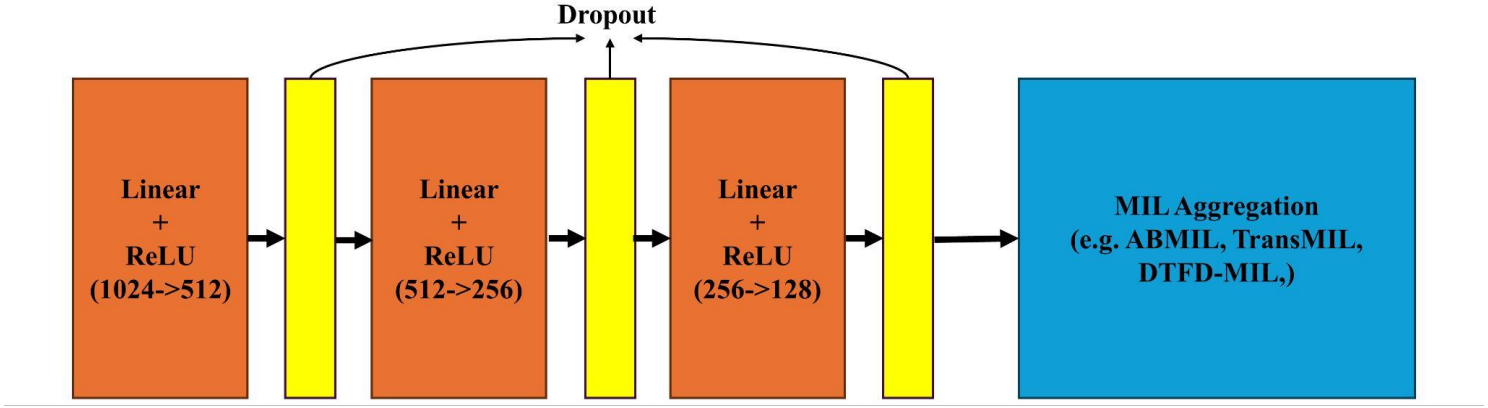
How Effective Can Dropout Be in Multiple Instance Learning?

Wenhui Zhu
Arizona State University

ICML
International Conference
On Machine Learning



DropNeuron vs DropInstance In MIL



In an MIL framework, Dropout is typically applied to the shallow feature extractor f_θ. For simplicity, considering f_θ as an MLP with L layers, the feature map at the l-th layer of f_θ is a 2-dimensional matrix $\boldsymbol{f}^{(l)} \in \mathbb{R}^{N \times D^{(l)}}$, where $D^{(l)}$ represents the embedding dimension. In this scenario, we consider randomly zeroing out either entries (Hinton et al., 2012) or entire instances in $\boldsymbol{f}^{(l)}$. For brevity, we denote the former as *DropNeuron* and the latter as *DropInstance*. For a given rate $p \in [0, 1]$, both DropNeuron and DropInstance can be defined as performing an element-wise masking operation over the feature map $\boldsymbol{f}^{(l)}$ at the l-th layer of f_θ:

$$\hat{\boldsymbol{f}}^{(l)} = \boldsymbol{f}^{(l)} \odot \boldsymbol{M}^{(l)},$$

where $\odot$ denotes element-wise multiplication, and $\boldsymbol{M}^{(l)} \in \mathbb{R}^{N \times D^{(l)}}$ is the binary Dropout mask at the l-th layer. In the regime of DropNeuron, each entry $M^{(l)}_{n,d}$ in $\boldsymbol{M}^{(l)}$ are from a Bernoulli distribution: $M^{(l)}_{n,d} \sim Bernoulli(p)$. In contrast, each row in $\boldsymbol{M}^{(l)}$ has the same entry and is sampled from a Bernoulli distribution in the case of DropInstance: $M^{(l)}_{n} \sim Bernoulli(p)$.

DropNeuron: Random Dropout elements in feature map

DropInstance: Random drop out whole instance

DropNeuron vs DropInstance In MIL

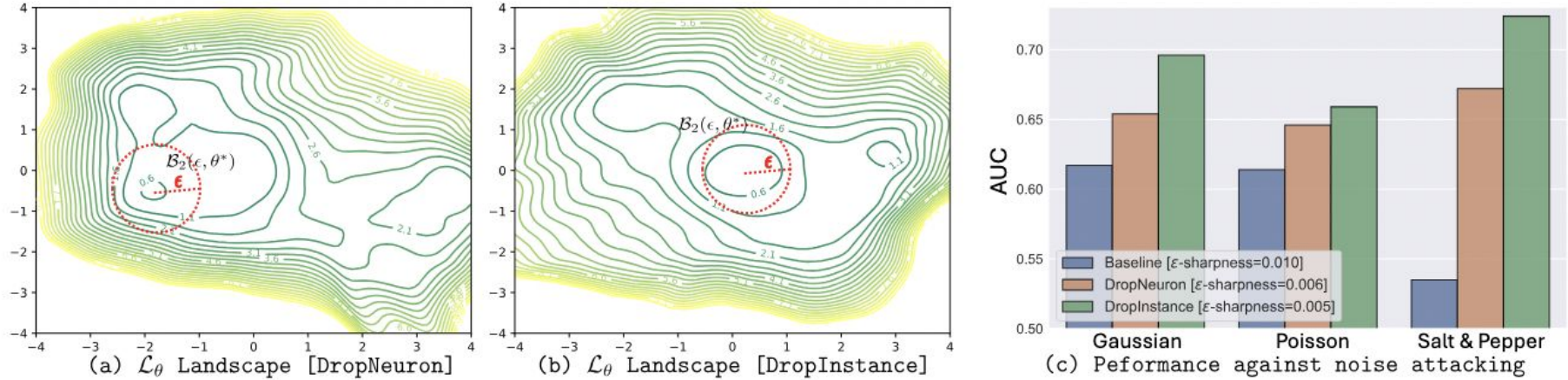


Figure 3. The landscape of the loss function $\mathcal{L}_\theta$ for two different dropout strategies (a) DropNeuron and (b) DropInstance as well as (c) the performance of MIL models against different noise attacks. We mark the Euclidean ball $\mathcal{B}_2(\epsilon, \theta^*)$ around the optimal parameter θ^* (see Eq. 5) in subpanel figure (a) and (b) with a red circle. We observe that the landscape of $\mathcal{L}_\theta$ in the DropInstance scenario leads to flatter minima compared to DropNeuron, which also results in a better performance in AUC.

Dropout based on instance leads to flatter local minima that typically have better generalizability.

How to impose Dropout

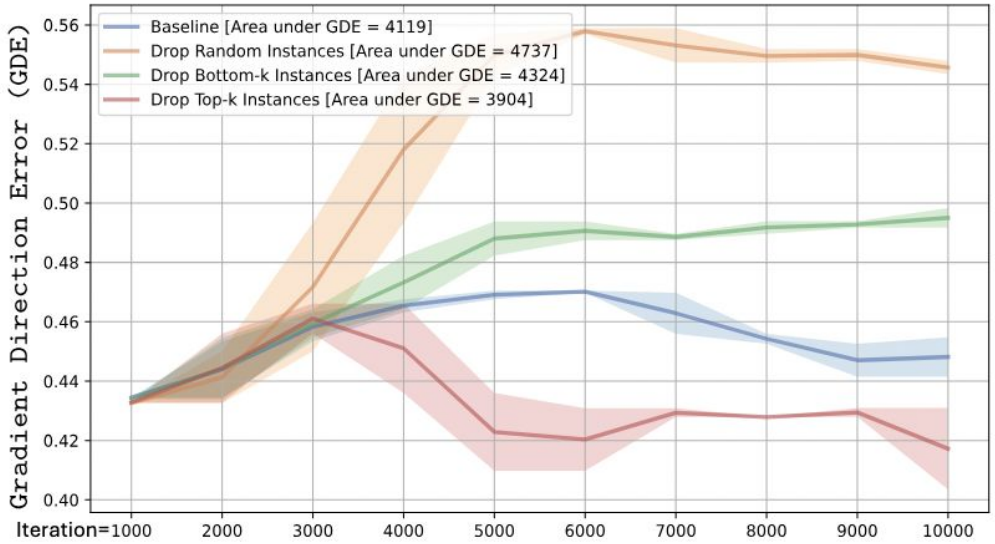
Here, we further investigate how to apply DropInstance. Previous studies have revealed that the effectiveness of algorithms or modules (e.g.Dropout) can be reflected by the gradient direction error (GDE) or gradient variance during model optimization. The gradient direction error quantifies the dissimilarity between the mini-batch gradient g_{step} and whole dataset gradient $\hat{g}$:

$$\mathrm{GDE} = \frac{1}{|G|} \sum_{g_{\mathrm{step}} \in G} \frac{1}{2} \left(1 - \frac{\langle g_{\mathrm{step}}, \hat{g} \rangle}{\|g_{\mathrm{step}}\|_2 \cdot \|\hat{g}\|_2} \right),$$

where G is a set of mini-batch gradients. Leveraging GDE, we investigate the impact of three different DropInstance strategies, including dropping (i) top-k instances, (ii) bottom-k instances, and (iii) random instances.

Liu, Zhuang, et al. "Dropout reduces underfitting." *International Conference on Machine Learning*. PMLR, 2023.

How to impose Dropout



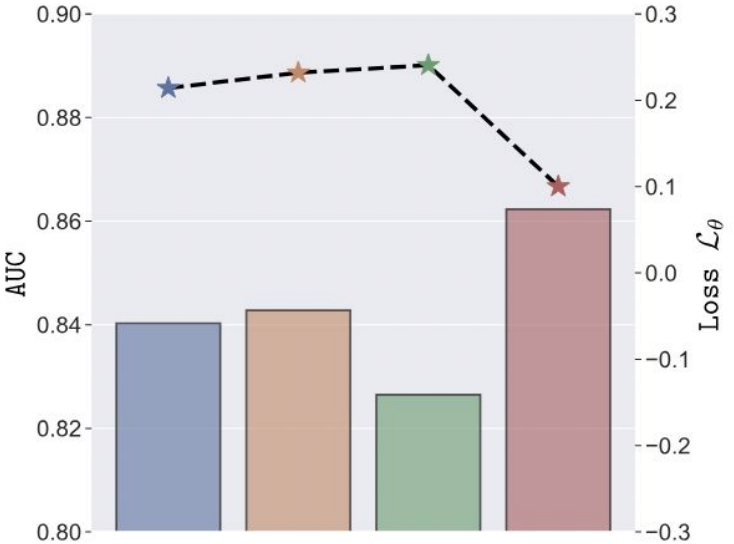


Figure 4. The comparison of change of GDE (**Left**) over the first 10,000 iterations as well as performance and loss (line plot) and AUC (bar plot) when using different instance dropout strategies (**Right**), where the area under GDE is the area enclosed by GDE and the x-axis. Dropping the top-k instances shows the smallest GDE, training loss , and highest AUC among all four strategies.

Dropping the top-k most important instances typically leads to better performance and gradient error direction !

MIL Dropout

Algorithm 1 MIL-Dropout Mechanism

Input: Input feature map $\boldsymbol{f}^{(l)} = \{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N\}$, K and G

Output: Processed Bag $\hat{\boldsymbol{f}^{(l)}}$ with MIL-Dropout

Initial: $M \leftarrow \mathbf{1}_{K \times D^{(l)}}$ (*# Initial mask*)

Select top-k important instances: $(\boldsymbol{f}_T^{(l)}, \boldsymbol{f}_R^{(l)}) \leftarrow \texttt{split}(\boldsymbol{f}^{(l)}), \quad \mathcal{A} \leftarrow [K]$ *(Eq. 7)*

Compute the similarity matrix between rest instances $\boldsymbol{f}_R^{(l)}$ and top-K instances $\boldsymbol{f}_T^{(l)}$
(*# Obtain G instances from $\boldsymbol{f}_R^{(l)}$ that are most similar to every top-K instance*)

for $i = 1$ to K **do**

  $A_i = \arg\max_{S \subseteq R, |S|=G} \sum_{j \in S} S_{i,j}$ *(Eq. 8)*

  $\mathcal{A} \leftarrow \mathcal{A} \cup A_i$

end for

$M[\mathcal{A}, :] = 0$

$\hat{\boldsymbol{f}^{(l)}} \leftarrow \gamma(\boldsymbol{f}^{(l)} \odot M)$ *(Eq. 9)*

return $\hat{\boldsymbol{f}^{(l)}}$ (*# Masking and normalization*)

Experiments

Table 1. Performance comparison on MIL benchmark datasets. Each experiment is performed five times with 10-fold cross-validation. We reported the mean of the classification accuracy (± the standard deviation of the mean).

Methods	MUSK1	MUSK2	FOX	TIGER	ELEPHANT
mi-Net	0.889 ± 0.039	0.858 ± 0.049	0.613 ± 0.035	0.824 ± 0.034	0.858 ± 0.037
MI-Net	0.887 ± 0.041	0.859 ± 0.046	0.622 ± 0.038	0.830 ± 0.032	0.862 ± 0.034
MI-Net with DS	0.894 ± 0.042	0.874 ± 0.043	0.630 ± 0.037	0.845 ± 0.039	0.872 ± 0.032
MI-Net with RC	0.898 ± 0.043	0.873 ± 0.044	0.619 ± 0.047	0.836 ± 0.037	0.857 ± 0.040
ABMIL	0.892 ± 0.040	0.858 ± 0.048	0.615 ± 0.043	0.839 ± 0.022	0.868 ± 0.022
ABMIL-Gated	0.900 ± 0.050	0.863 ± 0.042	0.603 ± 0.029	0.845 ± 0.018	0.857 ± 0.027
GNN-MIL	0.917 ± 0.048	0.892 ± 0.011	0.679 ± 0.007	0.876 ± 0.015	0.903 ± 0.010
DP-MINN	0.907 ± 0.036	0.926 ± 0.043	0.655 ± 0.052	0.897 ± 0.028	0.894 ± 0.030
NLMIL	0.921 ± 0.017	0.910 ± 0.009	0.703 ± 0.035	0.857 ± 0.013	0.876 ± 0.011
ANLMIL	0.912 ± 0.009	0.822 ± 0.084	0.643 ± 0.012	0.733 ± 0.068	0.883 ± 0.014
DSMIL	0.932 ± 0.023	0.930 ± 0.020	0.729 ± 0.018	0.869 ± 0.008	0.925 ± 0.007
ABMIL + MIL-Dropout	0.964 ± 0.033	0.954 ± 0.019	**0.789 ± 0.043**	0.917 ± 0.036	**0.934 ± 0.046**
ABMIL-Gated + MIL-Dropout	**0.967 ± 0.019**	**0.958 ± 0.021**	0.788 ± 0.016	**0.919 ± 0.033**	0.927 ± 0.033

Experiments

Table 2. Comparison of performance before and after plugging MIL-Dropout into four different types of MIL aggregators and their variants on CAMELOYON16 and TCGA-NSCLC datasets. Δ denotes the performance gains after the integration of MIL-Dropout. The classification accuracy (%), F1 score (%), and AUC (%) are reported (± the standard deviation of the mean) by running each experiment five times.

		CAMELOYON16						TCGA-NSCLC					
		ImageNet Pretrained			SimCLR Pretrained			ImageNet Pretrained			SimCLR Pretrained		
		Accuracy	F1	AUC	Accuracy	F1	AUC	Accuracy	F1	AUC	Accuracy	F1	AUC
ABMIL		86.3 ± 1.1	85.0±1.0	86.0 ± 0.5	85.6 ± 0.9	84.2 ± 1.3	86.6 ± 1.4	87.5 ± 0.8	87.5 ± 0.8	92.4 ± 0.5	87.9 ± 0.8	88.1 ± 0.6	93.8 ± 0.8
	+MIL Dropout	87.2 ± 1.0	86.4±0.8	90.1± 0.8	88.6 ± 1.1	87.4 ± 1.0	88.3 ± 1.2	91.1 ± 1.3	91.1 ± 1.3	95.6 ± 0.4	91.4 ± 0.6	91.5 ± 0.5	95.9 ± 0.1
	Δ	**+0.9**	**+1.4**	**+4.1**	**+3.0**	**+3.2**	**+1.7**	**+3.6**	**+3.6**	**+3.2**	**+3.5**	**+3.4**	**+2.1**
ABMIL-Gated		86.9 ± 1.1	85.7±1.2	86.2 ± 1.2	84.3 ± 1.1	83.4 ± 1.0	85.9 ± 1.6	87.9 ± 0.9	87.9± 0.9	92.8 ± 0.9	89.0 ± 1.2	89.0 ± 1.2	94.4 ± 0.7
	+MIL Dropout	90.4 ± 1.3	89.6±1.2	90.7 ± 0.9	87.7 ± 1.3	86.7 ± 1.3	87.4 ± 0.9	90.0 ± 0.6	90.0 ± 0.6	95.3 ± 0.3	90.8 ± 0.7	90.8 ± 0.7	95.8 ± 0.2
	Δ	**+3.5**	**+3.9**	**+4.6**	**+3.4**	**+4.2**	**+1.5**	**+2.1**	**+2.1**	**+2.5**	**+1.8**	**+1.8**	**+1.4**
DSMIL		85.5± 0.8	84.3± 1.1	89.0± 1.8	83.3 ± 1.0	82.0 ± 1.4	85.9 ± 1.6	89.3 ± 0.7	89.4± 0.7	94.2 ± 0.3	84.1 ± 1.8	86.2 ± 1.5	92.0 ± 1.6
	+MIL Dropout	87.9± 1.5	86.8±1.6	90.6± 1.2	85.6 ± 0.9	84.8 ± 0.5	87.6 ± 0.8	89.9 ± 0.6	90.0± 0.5	95.3 ± 0.6	86.9 ± 0.4	88.3 ± 0.2	93.9 ± 0.3
	Δ	**+2.4**	**+2.6**	**+1.6**	**+2.3**	**+2.8**	**+1.7**	**+0.6**	**+0.6**	**+1.1**	**+2.8**	**+2.1**	**+1.9**
TransMIL		84.7 ± 2.1	83.3±2.9	86.5 ± 2.4	86.8 ± 1.0	85.9 ± 1.2	89.7 ± 0.6	86.9 ± 0.6	87.0 ± 0.6	93.3 ± 0.7	88.2 ± 2.1	88.3 ± 2.1	94.6 ± 1.1
	+ MIL Dropout	86.0 ± 1.5	84.9±1.5	89.4 ± 0.9	89.7 ± 1.3	88.7 ± 1.4	90.3 ± 1.2	88.0 ± 0.5	88.5 ± 1.1	94.3 ± 0.4	91.6 ± 0.9	92.0 ± 0.7	96.2 ± 0.6
	Δ	**+1.3**	**+1.3**	**+2.9**	**+2.9**	**+2.8**	**+0.6**	**+1.1**	**+1.6**	**+1.0**	**+2.8**	**+3.7**	**+1.6**
DTFD-MIL(AFS)		84.1 ± 0.6	75.5 ± 0.6	88.2 ± 0.3	87.4 ± 0.9	81.8 ± 1.2	89.6 ± 0.9	88.5 ± 0.5	88.0 ± 0.3	94.4 ± 0.2	87.6 ± 0.3	87.8 ± 0.4	93.1 ± 0.2
	+MIL Dropout	85.7 ± 1.4	79.1 ± 2.2	89.9 ± 0.6	88.5 ± 0.7	84.2 ± 0.6	92.5 ± 1.0	90.3 ± 0.4	90.0 ± 0.4	94.8 ± 0.1	91.5 ± 0.4	91.8 ± 0.4	96.1 ± 0.2
	Δ	**+1.6**	**+3.6**	**+1.6**	**+1.5**	**+2.4**	**+2.9**	**+1.8**	**+2.0**	**+0.4**	**+3.9**	**+4.0**	**+3.0**
DTFD-MIL(MaxS)		84.7 ± 1.8	78.3 ± 2.4	87.8 ± 0.8	87.7 ± 1.5	82.0 ± 2.3	88.4 ± 0.9	87.4 ± 1.0	87.3± 1.0	93.8 ± 0.1	85.1 ± 1.2	84.9 ± 2.2	91.0 ± 1.0
	+MIL Dropout	86.5 ± 0.9	81.0 ± 1.3	89.8 ± 0.9	89.5 ± 0.4	84.4 ± 0.4	91.6 ± 0.5	88.8 ± 0.5	88.4 ± 0.5	95.0 ± 0.4	87.5 ± 2.5	88.2 ± 2.2	93.2 ± 1.0
	Δ	**+1.8**	**+2.7**	**+2.0**	**+1.8**	**+2.4**	**+3.2**	**+1.4**	**+1.1**	**+1.2**	**+2.4**	**+3.3**	**+2.2**

Ablation Study and Visualization

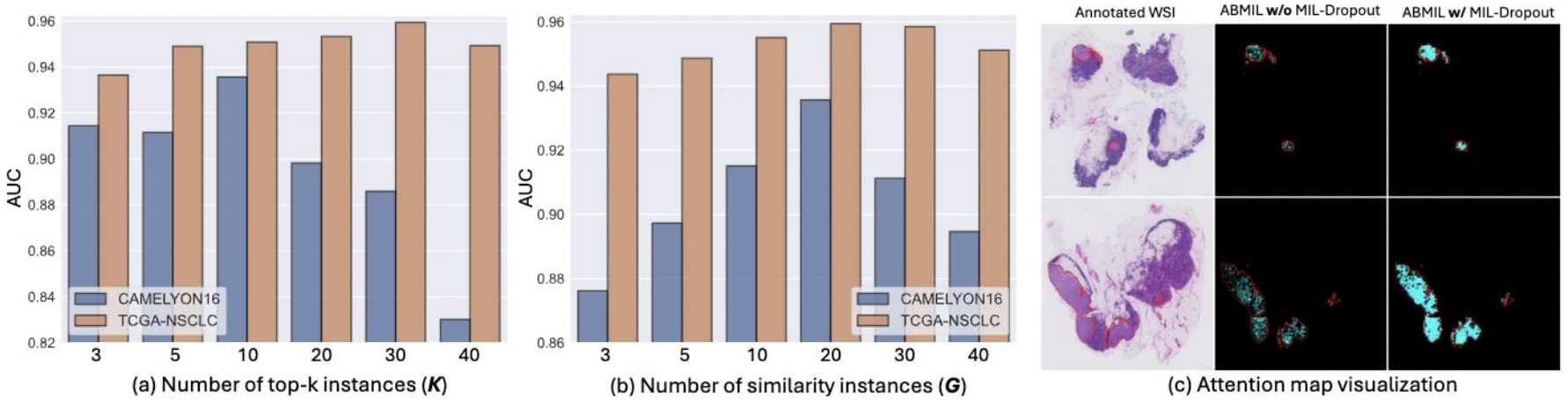


Figure 5. Ablation studies on the number of top-k instances K (a) and similarity instance S (b) using CAMELYON16 and TCGA-NSCLC datasets. (c) Attention map from ABMIL without and with MIL-Dropout, with tumor regions outlined in red. Brighter cyan in columns two and three indicates higher tumor probability (higher attention score) for corresponding locations.