CAR-MIL: Counterfactual Attention Regularization for Multiple Instance Learning

Imane Chraki[1,2], Pierre Marza[1,2], Stergios Christodoulidis[1,2], and Maria Vakalopoulou[1,2]

[1] Université Paris-Saclay, CentraleSupélec, Gustave Roussy, INSERM, IHU PRISM, Cancer Data Science Unit, France
[2] Université Paris-Saclay, CentraleSupélec, MICS Laboratory, France

Abstract. Multiple Instance Learning (MIL) is widely used for weakly supervised learning, particularly in digital pathology, where fine-grained annotations are costly. Most MIL methods aggregate instance features via attention mechanisms. However, attention weights do not always faithfully reflect instance importance and may focus on spuriously correlated regions. In this work, we propose CAR-MIL, a framework that explicitly guides attention learning through a counterfactual attention regularization objective inspired by counterfactual explanations. Built on a standard attention-based MIL architecture, our approach introduces a lightweight counterfactual attention branch trained to produce an alternative prediction while remaining close to the factual attention distribution. This encourages prediction changes to arise from minimal, structured redistributions of attention, leading to more informative evidence allocation. The resulting factual and counterfactual attention maps capture complementary evidence: the former highlights regions supporting the prediction, while the latter reveals regions whose reweighting would challenge it. We evaluate our method on synthetic MIL benchmarks with instance-level ground truth enabling controlled analysis of attention behavior and on five digital pathology datasets across four tasks. CAR-MIL maintains competitive classification performance, with the largest gains observed on more challenging tasks, while improving attention reliability, demonstrating the benefits of integrating counterfactual explainability reasoning into attention learning.

1 Introduction

Multiple Instance Learning (MIL) addresses learning from weakly-annotated data, where samples are organized into bags of instances, and labels are assigned only at the bag level [11, 25].

Attention-based MIL methods [16, 22, 37] address this challenge by learning a weighted aggregation of instance features into a bag representation, and achieve high predictive performance. Despite recent progress, two major limitations remain. First, as shown in recent studies [12, 15, 18, 40], MIL attention may highlight irrelevant or spuriously-correlated regions, both degrading performance and interpretability. Such an issue can arise from training instability,

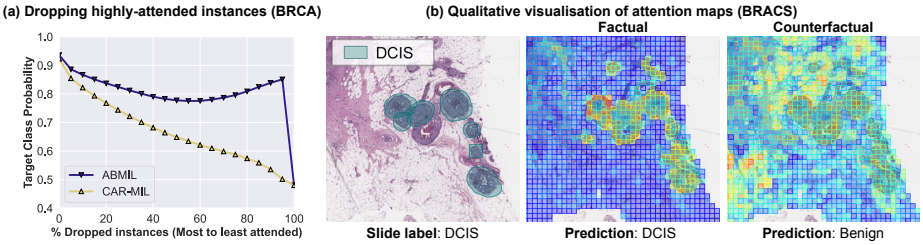

Fig. 1: CAR-MIL is a multiple instance learning method to model complex interactions between relevant regions using counterfactual explanation. **(a)** Dropping highly-attended instances for CAR-MIL leads to a higher performance drop than ABMIL on the TCGA-BRCA dataset. **(b)** Qualitative example from the BRACS dataset: the factual attention correctly focuses on the *Ductal Carcinoma In Situ* (*DCIS*) areas and predicts the correct slide-level label, whereas the counterfactual attention reallocates evidence and predicts the *Benign* class.

with the attention collapsing and highlighting only a small set of regions [33, 41]. Secondly, current methods only emphasize evidence supporting the predicted class while providing little insight into evidence that contradicts the decision, limiting both their reliability and their explanatory capacity. This is for example problematic in medical applications, where attention maps are often used to justify automated decisions. These observations suggest that attention should not be treated merely as a by-product of optimization but its training should be guided to reflect how evidence supporting or contradicting a prediction is allocated across instances.

A natural framework for reasoning about such evidence is provided by counterfactual explanations (CE) [36], which characterize predictions through minimal changes to the input of a model that would alter its output. This counterfactual perspective distinguishes evidence supporting a decision from evidence that challenges it. Importantly, existing CE methods provide post-hoc interpretability, as they are applied to a trained model to better understand its decision-making process. However, as showcased earlier, in the case of MIL, evidence reasoning should be performed at training time. We thus propose two main adaptations to this framework: (i) integrating CE at training time to guide the learning of attention, (ii) performing perturbations in attention instead of input space as it is more adapted to MIL. To the best of our knowledge, such an adaptation of CE to MIL training has not been studied in previous work. Recent methods begin to incorporate counterfactual interventions into the training of attention [8,30] but this is different from what we propose. Indeed, the latter rely on random attention perturbations while we believe learning such perturbations is beneficial.

In this work, we introduce CAR-MIL (Figure 1), a method trained to provide a factual and counterfactual attentions following CE principles. The counterfactual attention is constrained to remain close to the factual attention while producing a different prediction, encouraging changes to arise from selective re-

distributions of attention across instances. We propose a double-branch design yielding complementary attention maps: the factual attention highlights regions supporting the prediction, while the counterfactual attention reveals regions refuting it. Our framework introduces a new learning paradigm where the model learns how reallocating attention across instances affects the predicted outcome and therefore providing a contrastive training signal that guides attention learning dynamics. The counterfactual branch is implemented as a lightweight attention module sharing the encoder and classifier with the factual branch, therefore introducing minimal computational overhead and requiring no additional supervision. We evaluate CAR-MIL on synthetic MIL benchmarks designed to assess interaction modelling [15], as well as on five digital pathology datasets across four tasks. Across all settings, our method consistently maintains a competitive classification performance, with the largest performance gains observed on the more challenging tasks, while producing more reliable attention distributions, as confirmed by both quantitative metrics and qualitative analysis.

Our contributions can be summarized as follows: (i) We introduce for the first time counterfactual attention regularization for MIL, enforcing prediction divergence under minimal attention perturbations. (ii) We learn complementary factual and counterfactual attention distributions that reveal supporting and refuting evidence within attention maps. (iii) We demonstrate that explicitly guiding attention during training maintains or improves both predictive accuracy and interpretability in attention-based MIL models.

2 Related Work

Attention-based embedding-level MIL — Such MIL models aggregate instance embeddings into a global bag representation through an attention mechanism. The standard attention MIL (ABMIL) method [16] infers weights from instance features to aggregate them accordingly.

Numerous variants have then been proposed to improve the attention module [21, 23, 29, 31, 33, 40, 41]. Many of these MIL variants aim to improve the attention learning process, for example, by incorporating self-attention mechanisms [39], enforcing consistency between attention weights and instance feature representations [23], constraining attention through negative bag supervision in the case of binary classification [29], or mitigating problems caused by hard instances by introducing masking modules [33]. However, modifying only the attention mechanism, without evaluating the impact of the applied techniques on the model's predictions, might not guarantee achieving the desired predictive outcome [30].

Interpretability of MIL Models — Attention-based MIL models rely on attention scores to provide a localization map of regions of interest. However, raw attention maps inherent to these models often do not guarantee reliability [15,18,40]. On the other hand, fully additive models such as [18] seek to disentangle and sum individual contributions explicitly. Post-hoc explainability strategies have also been proposed [1–3, 15, 19, 27, 28, 32, 38], including perturbation-

based methods that modify instance subsets within bags and observe prediction changes [12, 15]. Despite progress, many of these methods are constrained by computational cost, limited scalability to large bags [6, 18, 26] or limited explanatory capacity: they typically highlight instances that support the predicted class, but do not reveal evidence that contradicts or challenges the decision [15].
Counterfactual Reasoning for Explanation and Attention Learning — We can consider two main frameworks with different objectives: (i) Counterfactual Explainability (CE) tries to understand the inner mechanisms of a trained model, while (ii) Counterfactual Attention Learning (CAL) aims at improving the quality of attention at training time. CE was introduced by [36] as a post-hoc interpretability framework that explains a model's prediction by identifying minimal changes to the input that would alter its decision. The core idea is to characterize predictions through contrast: what needs to change for the outcome to differ. CE methods therefore operate after training and focus on generating alternative inputs that provide insight into decision boundaries. On the other hand, CAL [8, 30] incorporates counterfactual reasoning into training by verifying that the learned attention is meaningfully better than controlled alternatives. In particular, CIA-MIL [8] introduces a counterfactual causal intervention on attention, evaluating whether the learned attention contributes meaningfully to prediction compared to a controlled random attention. A main drawback of such approaches is the lack of control on the applied perturbations. Indeed, by comparing the learned attention with a random counterpart, they primarily use counterfactual perturbations as a diagnostic tool and do not explicitly structure how evidence should be redistributed across instances during training. To address this, we incorporate CE at training time, and we have the flexibility to learn what perturbations would impact the final prediction.

3 Methods

3.1 Preliminaries

Multiple Instance Learning — In Multiple Instance Learning (MIL), a sample is defined as a set of instances named a bag $B_i = \{x_{i,1}, \ldots, x_{i,N_i}\}$, where N_i denotes the number of instances in bag i. Importantly, in MIL, the supervision is only available at the bag level, i.e. each bag is associated with a label $y_i \in \{1, \ldots, K\}$ for a K-class classification task. Instance-level labels are unobserved. We focus on embedding-level attention-based MIL models.
Standard Attention-Based MIL — Each instance $x_{i,j}$ is independently encoded using a feature extractor $\mathcal{E}$ as : $z_{i,j} = \mathcal{E}(x_{i,j}) \in \mathbb{R}^d$. For clarity, we drop the bag index and denote a bag of embeddings as $\{z_j\}_{j=1}^N$. An attention module ψ produces an attention logit u_j for each instance:

$$u_j = \psi(z_j), \qquad a = \mathrm{softmax}(u), \tag{1}$$

where $u = (u_1, \ldots, u_N) \in \mathbb{R}^N$ denotes the vector of attention logits and $a = (a_1, \ldots, a_N) \in \mathbb{R}^N$ the corresponding normalized attention weights.

We denote the bag-level classifier by φ. The bag representation $\hat{Z}$, class logits $F(u)$ and probabilities $\hat{y}(u)$ are then obtained as:

$$\hat{Z} = \sum_{j=1}^{N} a_j z_j, \qquad F(u) = \varphi(\hat{Z}) \in \mathbb{R}^K, \qquad \hat{y}(u) = \text{softmax}(F(u)). \tag{2}$$

3.2 Counterfactual Attention Regularization

Factual and Counterfactual Branches — Our method follows a two-branch setting (Figure 2) with a factual and counterfactual branches. Both have the same ABMIL [16] architecture, but different weights that are optimized concurrently. The encoder $\mathcal{E}$ and downstream classifier φ are not part of the branches and are thus shared between both. The difference between CAR-MIL and AB-MIL lies in the additional lightweight counterfactual branch, introducing minimal computational overhead. Differences in predictions thus arise solely from attention weighting differences between the two branches. The attention modules for the factual and counterfactual branches are respectively denoted as ψ and ψ_{cf}, and similarly to eq. 1, 2, we have the following for the counterfactual branch:

$$u_j^{\text{cf}} = \psi_{\text{cf}}(z_j), \quad a^{\text{cf}} = \text{softmax}(u^{\text{cf}}), \quad \hat{Z}^{\text{cf}} = \sum_{j=1}^{N} a_j^{\text{cf}} z_j, \quad F(u^{\text{cf}}) = \varphi(\hat{Z}^{\text{cf}}). \tag{3}$$

Evidence Differential — We define the counterfactual evidence differential in logit space as:

$$\Delta F(u, u^{\text{cf}}) = F(u) - F(u^{\text{cf}}) \in \mathbb{R}^K \tag{4}$$

It quantifies the difference between the predicted logits of factual and counterfactual attention logits. For a specific class $y \in \{1, \ldots, K\}$, ΔF_y represents the element in ΔF corresponding to class y. A large ΔF_y indicates that the logit predicted from the factual attention is higher than from its counterfactual counterpart. This would suggest that the factual branch attention provides stronger class-y evidence than the counterfactual branch.

3.3 Training Objective

we add to the main classification loss of the model an additional composite loss term constraining the two branches attentions to be close yet their predictions to be different. Therefore, prediction shifts are encouraged to arise from small but structured differences of the attention logits. The proposed CAR-MIL framework is therefore optimized using three complementary loss terms:

– **Classification loss:** Standard cross entropy loss on the factual branch prediction. This main loss ensures accurate bag-level prediction from the main (factual) branch:

$$\mathcal{L}_{\text{cls}} = \text{CE}(\hat{y}(u), y) \tag{5}$$

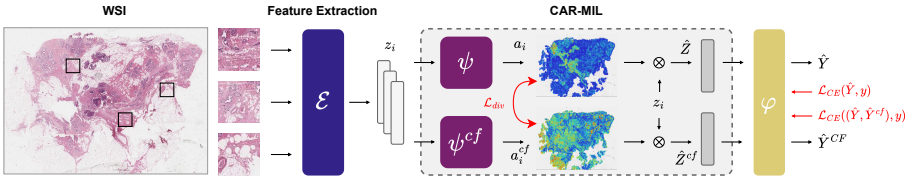

Fig. 2: CAR-MIL Overview. CAR-MIL is a two-branch attention MIL architecture trained with an explainable proximity-evidence constraint. Both branches operate on the same instance embeddings and share the downstream classifier. During training, a joint loss encourages (i) accurate bag-level predictions from the main(factual) branch, (ii) label-consistent evidence separation between main and counterfactual branches via the evidence differential, and (iii) meaningful differences between their attention vectors w.r.t the prediction shift.

– **Counterfactual evidence differential loss:** Encourages the evidence differential ΔF to favor difference between the two branches predictions on the ground-truth class, enforcing label-consistent separation between factual and counterfactual predictions:

$$\mathcal{L}_{\text{diff}} = \text{CE}\big(\text{softmax}(\Delta F), y\big) = \log\left(1 + \sum_{k \neq y} e^{\Delta F_k - \Delta F_y}\right) \tag{6}$$

In particular, if $\mathcal{L}_{\text{diff}} \leq \varepsilon$, then for all $k \neq y$,

$$\Delta F_y - \Delta F_k \geq -\log(e^{\varepsilon} - 1) \tag{7}$$

Thus, minimizing $\mathcal{L}_{\text{diff}}$ explicitly encourages the two branches to differ primarily on the ground-truth class.

– **Attention-logit proximity regularization:** Constrains the counterfactual attention logits to remain close to the factual attention logits, ensuring that prediction differences arise from minimal and controlled weighting updates of attention:

$$\mathcal{L}_{\text{div}} = D(u, u^{\text{cf}}), \tag{8}$$

where $D(\cdot, \cdot)$ denotes a distance metric in logit space. In practice, $D(\cdot, \cdot)$ can be instantiated using a normalized L_1 distance or a cosine-based dissimilarity. The L_1 formulation,

$$\mathcal{L}_{\text{div}} = D_{L_1}(u, u^{\text{cf}}) = \frac{1}{N} \sum_{j=1}^{N} \left| u_j - u_j^{\text{cf}} \right|, \tag{9}$$

penalizes absolute deviations in a sparse and interpretable manner, encouraging the counterfactual branch to modify the raw attention values of only a small subset of influential instances. This property aligns with the intuition that counterfactual evidence should emerge from *minimal* attention changes.

Alternatively, a cosine-based distance,

$$\mathcal{L}_{\text{div}} = D_{\cos}(u, u^{\text{cf}}) = 1 - \frac{\sum_{j=1}^{N} u_j \, u_j^{\text{cf}}}{\|u\|_2 \, \|u^{\text{cf}}\|_2}, \tag{10}$$

captures directional disagreement between the two attention patterns, encouraging the counterfactual head to orient its attention toward different subsets of instances while remaining scale-invariant. This is particularly useful when the magnitude of the logits is less important than their relative orientation across instances.

Overall Objective — The full training objective implements a counterfactual attention regularization principle to guide the learning dynamics of attentions. It encourages solutions where factual and counterfactual attentions remain close but induce label-consistent prediction differences. As a result, the factual attention is encouraged to highlight supporting evidence for the predicted class compared to the counterfactual branch. The final loss is given as follows,

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha \mathcal{L}_{\text{diff}} + \lambda \mathcal{L}_{\text{div}}, \qquad \alpha, \lambda \geq 0, \tag{11}$$

where α and λ are hyperparameters controlling the contribution of each loss in the final objective. In Section 4 of the supplementary material, we provide additional analytical results that explain why the CAR module increases factual attention on instances supporting the class prediction, while decreasing CF attention on those instances and increasing it on non-supportive ones.

4 Experiments and Results

We conducted experiments on both synthetic and histopathological datasets. In the following sections, we summarize the datasets used, along with their implementation details, the comparison methods, and the evaluation strategy.

4.1 Experiments on MNIST-bags Synthetic Data

Datasets — We leverage the recent MIL classification datasets from [15] derived from MNIST [10]. In this synthetic setting, instance-level annotations are available to indicate whether an instance provides supporting or refuting evidence for the bag label according to the dataset construction rule. We consider two dataset variants: *4-bags*: Class 1 if the bag contains an 8, class 2 if it contains a 9, class 3 if it contains both 8 and 9, and class 0 otherwise. This setting evaluates the model's ability to capture interactions between instances that jointly determine the bag label. *Adjacent Pairs*: a bag is class 1 if it contains any pair of consecutive digits between 0 and 4; otherwise, the bag corresponds to class 0. In this setting, the evidential role of an instance depends on the presence of other instances in the bag, making the prediction inherently contextual. This

Table 1: Comparison of Baseline and Counterfactual MIL Models on the MNIST-bags Synthetic Datasets. Results are reported for binary (*Adjacent Pairs*) and multi-class (*Four Bags*) classification tasks. *Bag AUC* measures slide-level predictive performance, while *Instance AUPRC*$^{+}$ and *AUPRC*$^{\pm}$ assess alignment between attention scores and instance-level evidence supporting or refuting the bag label. *CAR-MIL* substantially improves instance-level performance, indicating better discriminative evidence of the attention, while maintaining strong bag-level performance.

		Adjacent Pairs			Four Bags		
Model	**Evidence**	**AUC** (↑)	**AUPRC**$^{+}$ (↑)	**AUPRC**$^{\pm}$ (↑)	**AUC** (↑)	**AUPRC**$^{+}$ (↑)	**AUPRC**$^{\pm}$ (↑)
ABMIL [16]	attention	85.7 ± 5.9	76.9 ± 6.1	61.5 ± 0.9	99.1 ± 0.1	**86.8 ± 0.2**	53.1 ± 0.1
AddMIL [18]	attention	81.0 ± 2.2	64.3 ± 4.8	63.5 ± 1.2	97.0 ± 2.7	82.1 ± 12.1	53.4 ± 0.8
	patch-scores	N/A	66.7 ± 7.4	67.6 ± 6.6	N/A	76.4 ± 15.0	76.8 ± 14.6
TransMIL [31]	attention	**94.6 ± 4.4**	81.5 ± 6.9	61.7 ± 2.1	99.5 ± 0.1	81.6 ± 6.3	51.9 ± 0.8
CAR-MIL	attention	94.1 ± 5.9	**85.7 ± 6.5**	**84.6 ± 5.2**	**99.6 ± 0.1**	**86.8 ± 0.6**	**89.7 ± 0.9**

setup therefore evaluates whether models correctly capture context-dependent evidence relationships between instances.

Implementation Details — Feature extraction is performed for the MNIST images using ResNet18 [14] model pre-trained on ImageNet [9] from the TorchVision library [24]. Experiments were repeated 10 times with a learning rate of 0.0002 and the SGD optimizer. We compare our proposed strategy against MIL pooling baselines, reporting means and standard deviations across repetitions of performance in terms of area under the curve (AUC) at the bag level, as well as area under the precision-recall curve of attention prediction of instances' positive evidence label (AUPRC^{+}) and a two-class (positive and negative evidence classes) averaged area under the precision-recall curve (AUPRC$^{\pm}$). More details are to be found in sections 5 and 6 of our supplementary material. We evaluate our method using the $L1$ distance as the metric to measure the similarity between factual and counterfactual attentions. We present the considered baselines along with the method used to extract instance scores in parentheses: ABMIL [16] (raw attention scores), TransMIL [31] (attention rollout [17]), AddMIL [18] (patch-level scores).

In all models, the attention is interpreted as representing *positive evidence* (instances that support the predicted label) while reverse attention is used to represent *negative evidence*. In our proposed approach, counterfactual attention serves as an estimation of negative evidence, explicitly modelling instance-level factors that contradict the predicted outcome.

Counterfactual Reasoning Improves Attention — The synthetic benchmark provides a controlled setting to evaluate both downstream classification performance and the alignment between attention distributions and instance-level evidence. For the *Adjacent Pairs* binary classification task, the model must be aware of the contextual influence of individual instances to achieve accurate predictions. As reported in Table 1, CAR-MIL outperforms the ABMIL and AddMIL, and reaches performance comparable to TransMIL at the bag level (94.1 vs. 94.6 AUC). More importantly, our method yields substantially

higher instance-level performance (AUPRC^{+}=85.7 and AUPRC$^{\pm}$=84.6), indicating that the learned attention more faithfully reflects discriminative evidence. As for the *Four Bags* multi-class task, our method improves both performance at the bag-level and attention reliability to reflect the evidence provided by instances that support or refute the bag label. These synthetic experiments suggest that counterfactual reasoning applied at the attention level promotes a tighter correspondence between attention allocation and instance-level evidence. As a result, more faithful attention leads to improved predictive performance.

4.2 Experiments on Histopathology Data

Datasets — We conduct our experiments on multiple public whole-slide image (WSI) datasets from TCGA [35]: *TCGA-NSCLC* (Non–Small Cell Lung Cancer) and *TCGA-BRCA*(Breast Invasive Carcinoma) for binary cancer subtyping, *TCGA-LUAD*(Lung Adenocarcinoma) for TP53 mutation prediction, *CAMELYON16* [4] for binary metastasis detection in breast lymph node, and *BRACS* [5](BreAst Cancer Subtyping), a breast cancer dataset for multiclass tissue classification between *normal*, *benign*, and *malignant* categories.. *CAMELYON16* and *BRACS* contain instance-level annotations, with the latter having only sparse labels.

Implementation Details — Slides were processed using the standard approach as in [23] to obtain patches of size 256x256 at 20X magnification. We use pre-extracted features with UNI-V1 [7] foundation model. Please note that any other encoder could be integrated into our method. Experiments were conducted under five cross-validation settings for *TCGA-NSCLC*, *TCGA-BRCA*, and *TCGA-LUAD* using a learning rate of 0.0002. We used the originally published train-val-test split for *BRACS* in three runs with a learning rate of 0.0001, and similarly the train-test split for *CAMELYON16* in three runs with a learning rate of 0.0002. All experiments were conducted using the Adam [20] optimizer. Additional details to be found in sections 5 and 6 of our supplementary material.

We challenge CAR-MIL against several popular MIL frameworks including: baseline approaches that do not use attention (Mean-Max MIL), classical attention-based MIL (ABMIL [16]) and attention-focused architectures such as CLAM [23], ACMIL [41], DSMIL [21], TransMIL [31], AddMIL [18], and CIA-MIL [8]. We report performance metrics in terms of AUC or balanced accuracy and F1 score. For *CAMELYON16*, we report the AUPRC^{+}, the area under the precision recall curve for attention as a prediction of instance labels. Scores are reported in terms of the average $\pm$ standard deviation.

In the proposed framework, the proximity regularization constrains the counterfactual attention logits to remain close to the factual attention logits. We experiment with both L_1 and cosine distances as attention similarity measures for our method. Evaluating both variants allows us to study how different proximity geometries affect the redistribution of evidence and predictive performance.

CAR-MIL Achieves Competitive Downstream Performance — Table 2 summarizes the performance of CAR-MIL across five histopathology tasks. Overall, CAR-MIL achieves performance comparable to or better than strong

Table 2: Performance Comparison of CAR-MIL Against State-of-the-Art MIL Models on Histopathology Classification Tasks. Results are reported for BRCA and NSCLC cancer subtyping, TP53 mutation prediction in LUAD, BRACS tissue subtyping, and CAMELYON16 metastasis detection. Metrics include AUC, F1-score, and balanced accuracy. CAR-MIL consistently improves or matches the best baseline performance across datasets using UNI features.

	BRCA		NSCLC		LUAD		BRACS		CAMELYON16	
Model	**AUC** (↑)	**F1** (↑)	**AUC** (↑)	**F1** (↑)	**AUC** (↑)	**F1** (↑)	**BACC** (↑)	**F1** (↑)	**AUC** (↑)	**AUPRC$^+$** (↑)
Meanmil	93.2±2.4	85.1±2.4	96.9±1.3	94.0±1.0	74.5±6.3	71.1±7.4	33.2±1.8	28.2±2.1	62.5±4.8	N/A
MaxMIL	95.4±1.5	88.0±1.1	97.5±1.0	94.6±1.9	76.0±5.4	71.8±4.9	32.3±7.1	29.4±8.6	98.3±0.4	N/A
CLAM [23]	94.5±2.3	87.6±1.7	97.9±0.8	94.1±1.7	74.0±3.0	70.6±1.8	40.4±5.5	38.0±4.9	99.7±0.3	94.4±0.3
AddMIL [18]	93.7±2.6	86.0±3.3	94.6±3.0	91.9±2.7	73.3±3.9	71.5±4.0	38.3±9.6	35.7±11.6	98.2±1.4	93.4±1.2
DSMIL [21]	94.1±1.6	85.3±2.3	97.4±1.1	94.0±2.6	66.9±4.7	64.8±5.1	42.3±2.0	40.0±2.0	98.9±1.1	80.6±11.2
TransMIL [31]	93.2±2.7	87.7±0.8	97.8±0.6	94.8±2.0	71.7±5.4	68.8±4.4	40.0±3.6	37.6±3.8	**99.9±0.1**	28.2±3.6
ACMIL [41]	94.4±2.9	88.7±2.2	97.9±0.8	93.9±2.5	75.5±7.4	72.0±6.6	41.2±1.4	38.5±2.4	99.4±0.5	**95.4±0.5**
RRT-MIL [34]	93.7±1.2	84.6±3.4	97.3±1.8	93.7±2.9	75.1±5.2	71.8±4.0	39.3± 4.8	34.5±5.2	99.6±0.5	94.4±1.1
CIA-MIL [8]	94.5±1.2	87.3±2.7	96.5±1.6	93.0±2.5	77.5±2.5	71.1±5.6	37.6±2.1	35.0±1.6	99.2±0.5	92.9±1.5
ABMIL [16]	95.3±1.7	88.9±1.3	97.6±1.0	94.2±1.2	74.0±4.3	72.1±5.4	38.2±4.1	36.0±4.5	98.7±0.3	93.2±1.2
CAR-MIL (L1)	**95.5±2.2**	87.9±1.5	**98.0±0.5**	94.4±1.7	76.4±4.9	72.4±3.8	40.8±2.5	38.1±2.5	**99.9±0.1**	94.8±1.0
CAR-MIL (COS)	95.0±1.7	**89.6±1.4**	97.5±0.9	**95.0±1.8**	**78.1±4.1**	**72.6±4.5**	**43.1±2.0**	**40.4±2.4**	99.3±0.7	91.7±0.6

MIL baselines, while paired statistical t-tests show that no attention-based MIL method significantly outperforms CAR-MIL on any dataset. Performance gains are more pronounced on the more challenging LUAD and BRACS tasks, suggesting that explicitly guiding attention during training is particularly beneficial in more challenging settings. In particular, the cosine variant of CAR-MIL tends to perform better on the more challenging LUAD and BRACS tasks, while the L_1 formulation remains competitive on binary subtyping tasks such as BRCA and NSCLC. On CAMELYON16, where patch-level tumor annotations are well defined and available, CAR-MIL maintains competitive patch-level performance (AUPRC$^+$) indicating that redistributing attention through the counterfactual branch does not cause the model to deviate from diagnostically-relevant regions.

5 Analysis of Counterfactual Attention Regularization

Hyperparameter Sensitivity — Fig. 3 analyzes the sensitivity to hyperparameters α controlling the evidence differential loss $\mathcal{L}_{\text{diff}}$, and λ controlling the attention logit proximity loss $\mathcal{L}_{\text{div}}$. We report the AUC obtained for both the cosine and L_1 variants of a run of CAR-MIL on the BRCA and LUAD datasets for different combinations of these weights. The explored configurations include values $\{0.2, 0.8, 1.0\}$, together with the baseline ABMIL setting without counterfactual regularization ($\alpha = \lambda = 0$). On BRCA, the baseline AUC of 94.5 improves to values between 95.3 and 96.0 depending on the configuration, with the best performance obtained for moderate counterfactual regularization. On LUAD, where the task is more challenging, the baseline AUC of 81.3 increases to values up to 84.2–84.4. While improvements are observed on both datasets, the effect of the hyperparameter choices is more visible on LUAD than on BRCA.

Table 3: Ablation with Respect to MIL Settings. Integrating the proposed counterfactual attention regularization (CAR) into more attention-based MIL architectures, including DSMIL, and TransMIL. Performance comparaison is assessed between the original baselines versus their CAR-augmented variants across five datasets. CAR generally improves or maintains performance across most metrics, showing that the proposed regularization can be integrated into diverse MIL attention architectures.

	BRCA		NSCLC		LUAD		BRACS		CAMELYON16	
Model	**AUC** (↑)	**F1** (↑)	**AUC** (↑)	**F1** (↑)	**AUC** (↑)	**F1** (↑)	**BACC** (↑)	**F1** (↑)	**AUC** (↑)	**AUPRC**$^{+}$ (↑)
ABMIL [16]	**95.3±1.7**	88.9±1.3	97.6±1.0	94.2±1.2	74.0±4.3	72.1±5.4	38.2±4.1	36.0±4.5	98.7±0.3	93.2±1.2
CAR-MIL	95.0±1.7	**89.6±1.4**	**98.0±0.5**	94.4±1.7	**78.1±4.1**	**72.6±4.5**	**43.1±2.0**	**40.4±2.4**	**99.9±0.1**	**94.8±1.0**
DSMIL [21]	94.1±1.6	85.3±2.3	97.4±1.1	94.0±2.6	66.9±4.7	64.8±5.1	42.3±2.0	40.0±2.0	98.9±1.1	80.6±11.2
CAR-DSMIL	94.5±2.2	85.6±2.7	97.8±1.0	94.0±2.3	71.2±4.1	67.8±3.4	42.5±5.1	39.8±2.0	99.0±0.6	87.8±2.9
TransMIL [31]	93.2±2.7	87.7±0.8	97.8±0.6	**94.8±2.0**	71.7±5.4	68.8±4.4	40.0±3.6	37.6±3.8	**99.9±0.1**	28.2±3.6
CAR-TransMIL	94.0±1.3	87.2±4.3	97.6±1.4	94.3±2.0	73.6±4.4	69.8±4.0	42.6±6.1	39.4±6.2	99.7±0.1	32.6±0.3

This difference likely reflects the relative difficulty of the tasks: when baseline performance is already high, as on BRCA, performance differences across configurations remain relatively small, whereas on the more challenging LUAD task the influence of the regularization weights becomes more visible. Across settings, the cosine variant exhibits smoother trends across configurations, while the L1 variant shows slightly larger variability depending on the choice of weights.

MIL Setting — Table 3 evaluates the effect of integrating our counterfactual attention regularization (CAR) into different MIL settings: ABMIL [16] (already presented in Table 2, but reported for comparison with other MIL methods), DSMIL [21] and TransMIL [31]. We provide in section 7 of our supplementary material more technical details on the integration of CAR following the different attention formulations. For each setting, we report the best-performing distance variant (either L_1 or cosine) in order to focus on the contribution of the proposed CAR mechanism rather than the choice of distance metric. Across architectures, the proposed regularization consistently improves or maintains performance compared to the original models. In particular, the gains are most visible for the ABMIL backbone. Similar trends are observed when integrating the method into DSMIL and TransMIL, where CAR variants improve mutation prediction LUAD and BRACS classification performance while preserving strong results on easier tasks. Overall, these results indicate that counterfactual attention regularization can be effectively combined with different MIL aggregation mechanisms to improve slide-level prediction.

Feature Extractor — Table 4 reports an analysis with respect to the feature encoder. Using pretrained ResNet50 [14] on ImageNet [9,24] features, CAR-MIL consistently improves performance over the ABMIL baseline across all datasets. As expected, absolute performance remains lower than with UNI features reported in the main experiments, reflecting the stronger representational capacity of pathology-specific foundation models. Nevertheless, the consistent gains demonstrate that the proposed counterfactual attention regularization is not tied to a specific encoder and can improve MIL aggregation even with stan-

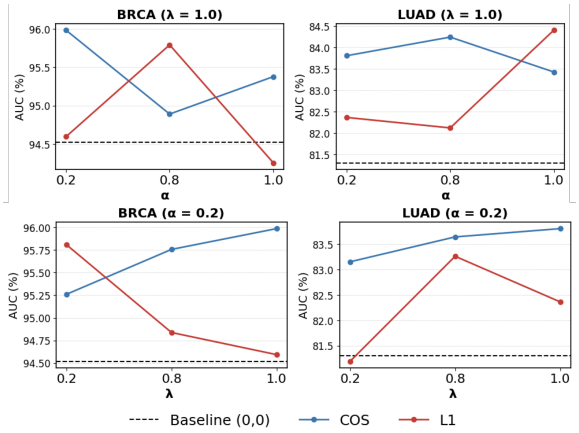

Fig. 3: Hyperparameters Sensitivity. Sensitivity to hyperparameters α and λ is evaluated on BRCA and LUAD for the Cosine and L_1 divergence variants on a single run. Cosine variant tends to have more consistent trends than the L1 variant.

Table 4: Ablation on the Feature Encoder Using ResNet50 Features. Comparison between ABMIL and CAR-MIL across three TCGA datasets – BR.: BRCA, NS.: NSCLC, LU.: LUAD. CAR-MIL consistently improves AUC and F1 over the baseline ABMIL indicating that the proposed CAR remains effective even when using standard convolutional features.

	Model	AUC (↑)	F1 (↑)
BR.	ABMIL [16]	89.2±2.6	79.7±3.9
	CAR-MIL	**90.3±2.1**	**81.9±5.0**
NS.	ABMIL [16]	93.4±1.7	88.2±1.6
	CAR-MIL	**94.3±1.2**	**88.9±1.3**
LU.	ABMIL [16]	68.2±5.6	66.3±4.8
	CAR-MIL	**70.4±6.1**	**68.4±5.1**

dard convolutional features. In fact the gains are even more pronounced with out-of-domain features as the ones from ResNet50 compared to UNI features, highlighting the positive impact of guiding the learning dynamics of attention under generic features representations.

Multi-class Setting — We investigate the counterfactual branch predicted class logit under the multi-class setting. We find that the class whose CF logit increases most varies with the ground-truth class. On Four Bags, class 1 shifts toward class 3 (100% of bags) while class 3 shifts toward class 1 (62%), reflecting the dataset evidence structure. On BRACS, redistribution is diffuse across all 7 classes with no dominant competitor, consistent with higher semantic ambiguity.

6 Attention Analysis

Counterfactual Attention Regularization for Interpretability — We assess how much the proposed counterfactual attention regularization contributes to bridging the gap between predictive performance and interpretability relevance in MIL frameworks. Using the synthetic benchmark, we compare the attention from our method with the raw attention from *ABMIL* and attention maps generated from two baselines: (i) a *Random* assignment of attention weights to instances, serving as a lower bound, and (ii) a post-hoc explainability method, *Perturbation Single* [13, 15], which estimates instance importance by measuring prediction sensitivity to perturbations, by passing bags of single instances through the model ("single"). We then assess the interpretability of the inherent learned attention in CAR-MIL counterfactual variant. As shown

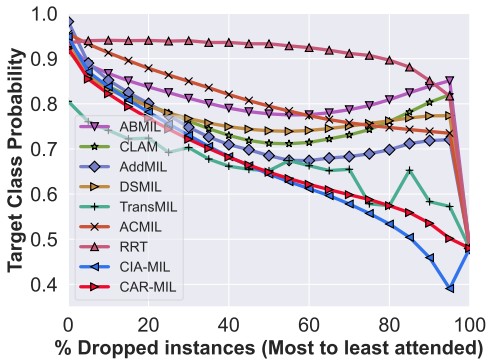

Fig. 4: Effect of Removing Highly Attended Instances on the Prediction Confidence. Removing attention according to top-ranked instances shows that CAR-MIL yields a smooth, monotonic drop in confidence, indicating faithful instance importance, whereas ABMIL exhibits non-monotonic behavior.

Table 5: Attention as Interpretability Proxy on MNIST-bags. Evaluation of attention as a proxy for instance-level evidence on synthetic MIL datasets with ground-truth evidence labels. CAR-MIL attention is reliable as an interpretability signal, outperforming ABMIL raw attention and perturbation evidence attribution for explaining ABMIL. – Rand.: Random, Att.: Attention, Pert.: Perturbation, Adj. P.: Adjacent Pairs, Four B: Four Bags.

	Model	Evid.	AUPRC^{+})	AUPRC$^{\pm}$
Adj. P.	ABMIL [16]	Rand.	54.2 ± 0.9	54.2 ± 0.9
		Att.	76.9 ± 6.1	61.5 ± 0.9
		Pert.	78.5 ± 1.0	78.5 ± 1.0
	CAR-MIL	Att.	**85.7 ± 6.5**	**84.6 ± 5.2**
Four B.	ABMIL [16]	Rand.	30.6 ± 0.2	31.2 ± 0.2
		Att.	**86.8 ± 0.2**	53.1 ± 0.1
		Pert.	87.9 ± 1.6	87.7 ± 2.1
	CAR-MIL	Att.	**86.8 ± 0.6**	**89.7 ± 0.9**

in Table 5, counterfactual-guided attention learning substantially improves the alignment between attention and instance-level evidence compared to both ABMIL and the perturbation-based explanation. In particular, CAR-MIL achieves the highest AUPRC^{+} and AUPRC$^{\pm}$ scores on both tasks.

Faithfulness Evaluation via Perturbation Analysis — To further assess the faithfulness of attention, we perform a MORF (Most Relevant First) perturbation test [12, 15], where instances in correctly-predicted bags are removed in order of decreasing attention importance. For CAR-MIL, importance is computed from the discrepancy between factual and counterfactual attention scores. A faithful attention mechanism is expected to produce a *monotonic decrease* in the model's confidence, since removing the most relevant instances should consistently degrade predictive performance. In Figure 4, our proposed CAR-MIL model exhibits a smooth and monotonic drop in prediction probability as a function of the percentage of removed patches, suggesting that learned attentions align closely with true predictive importance. In contrast, models such as ABMIL show non-monotonic or convex response curves, suggesting that their attention weights may not reliably capture relevant evidence among instances. Interestingly, CAR-MIL showcases a similar trend as the recent CIA-MIL method that focuses on causally aligning attention with predictions at the cost of a trade-off between performance and explainability. As shown in Table 2, CAR-MIL outperforms CIA-MIL in terms of downstream performance. Indeed, by improving solely explainability through counterfactual intervention, CIA-MIL can lead to a drop in performance compared with CAR-MIL.

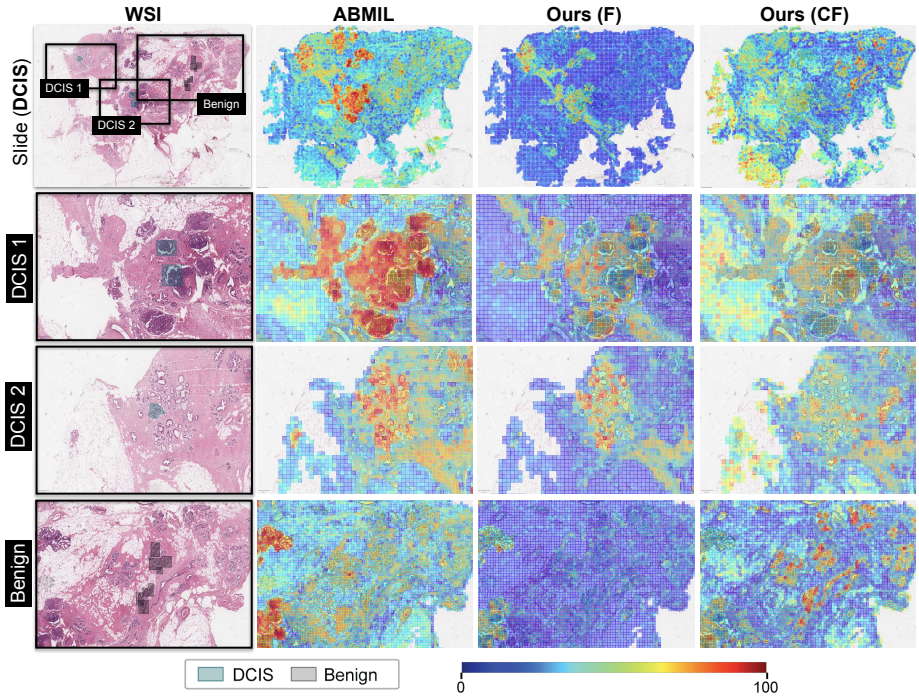

Fig. 5: Comparisons Between the Attentions of ABMIL and the Two CAR-MIL Attention Branches (F: Factual, CF: Counterfactual). Qualitative comparison of attention maps on a DCIS WSI from the BRACS dataset. ABMIL highlights broad regions within the annotated DCIS area but produces a diffuse attention pattern. Our model (factual attention) concentrates attention on more precise subregions that correspond closely to the annotated malignant ducts. The counterfactual branch highlights benign or non-malignant areas, providing a complementary view of "negative evidence". Together, the two branches provide a more structured and interpretable separation between supporting and refuting regions.

WSI Heatmaps — To qualitatively illustrate the differences between attentions, we visualize attention maps for a representative WSI from the *BRACS* dataset labeled as *DCIS* (*Ductal Carcinoma In Situ*). Figure 5 displays the attentions obtained with ABMIL, CAR-MIL factual attention (F), and CAR-MIL counterfactual (CF) branch. Although ABMIL broadly highlights areas containing annotated DCIS regions, its attention is spatially diffuse. In contrast, CAR-MIL produces more localized and concentrated attention regions. The counterfactual head provides further informative signal: its highest value occurs in benign or non-neoplastic areas, *i.e.,* regions whose removal would weaken the model's confidence in the positive class. This complementary pattern aligns with our design goal: factual attention captures supporting evidence, while counterfactual attention highlights refuting or neutral evidence. More quantitatively, the factual branch does not simply reproduce ABMIL localization (Spearman r=0.67

on the representative example in Figure 5 between attention vectors), while the factual and counterfactual branches attend to distinct regions (r=0.29). Across BRACS slides, factual attention remains only moderately correlated with ABMIL (L1: r=0.69±0.05; Cosine: r=0.69±0.03), whereas factual and counterfactual attentions are strongly anti-correlated under L1 (r=−0.67±0.02) and weakly correlated under cosine (r=0.20±0.63). Our supplementary material Figures 1,2 further illustrate this behavior and show that CAR-MIL concentrates attention on DCIS regions while suppressing benign tissue compared with ABMIL, DSMIL, AddMIL, and CLAM.

7 Conclusion

In this work, we introduce CAR-MIL, a counterfactual attention regularization framework designed to jointly improve the performance and reliability of attention-based MIL models. Motivated by the observation that attention plays a dual role, as both the mechanism that aggregates instance information and the most commonly used proxy for MIL interpretability, we propose a double-branch (factual-counterfactual) architecture and a training strategy that explicitly guides attention using learned counterfactual perturbations. By encouraging the counterfactual branch to induce prediction changes while remaining close to the main attention distribution, CAR-MIL reveals informative regions that standard attention may overlook. Through experiments on synthetic datasets and multiple whole-slide image benchmarks, we demonstrate that CAR-MIL maintains or improves downstream performance while producing attention maps that better align with instance-level evidence. These results suggest that integrating counterfactual explainability reasoning into the attention optimization process offers a promising direction for designing MIL models that are not only more accurate but also more interpretable. Future work could explore how the learned counterfactual attentions can be leveraged beyond training, for instance as auxiliary supervisory signals or for downstream tasks such as improving weak supervision, guiding active learning strategies, supporting model debugging, and facilitating deeper analysis of model behavior and decision-making processes.

Acknowledgements

This work has benefited from state financial aid, managed by the Agence Nationale de Recherche under the investment program integrated into France 2030, project references ANR-21-RHUS-0003, ANR-21-CE45-0007, ANR-23-CE45-0029, ANR-23-IAHU-0002, and ANR-23-IACL-0003 – DATAIA CLUSTER (as part of IA CLUSTER program). Experiments have been conducted using HPC resources from the Mésocentre computing center of CentraleSupélec and École Normale Supérieure Paris-Saclay, supported by CNRS and Région Île-de-France, and resources from GENCI–IDRIS (Grant 2025-AD011015828, 2026-AD011015828R1). The results shown in this paper are part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. Advances in neural information processing systems **31** (2018)
2. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)
3. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.R.: How to explain individual classification decisions. The Journal of Machine Learning Research **11**, 1803–1831 (2010)
4. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017)
5. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., Pietro, G.D., Bonito, M.D., Foncubierta, A., Botti, G., Gabrani, M., Feroce, F., Frucci, M.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images (2021)
6. Van den Broeck, G., Lykov, A., Schleich, M., Suciu, D.: On the tractability of shap explanations. Journal of Artificial Intelligence Research **74**, 851–886 (2022)
7. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature medicine **30**(3), 850–862 (2024)
8. Chraki, I., Marza, P., Christodoulidis, S., Vakalopoulou, M.: Counterfactual intervention in attention multiple instance learning for digital pathology. In: Medical Imaging with Deep Learning (2026)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
10. Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Processing Magazine **29**(6), 141–142 (2012)
11. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence **89**(1-2), 31–71 (1997)
12. Early, J., Cheung, G.K., Cutajar, K., Xie, H., Kandola, J., Twomey, N.: Inherently interpretable time series classification via multiple instance learning. arXiv preprint arXiv:2311.10049 (2023)
13. Early, J., Evers, C., Ramchurn, S.: Model agnostic interpretability for multiple instance learning. arXiv preprint arXiv:2201.11701 (2022)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Hense, J., Jamshidi Idaji, M., Eberle, O., Schnake, T., Dippel, J., Ciernik, L., Buchstab, O., Mock, A., Klauschen, F., Müller, K.R.: Xmil: Insightful explanations for multiple instance learning in histopathology. Advances in Neural Information Processing Systems **37**, 8300–8328 (2024)
16. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
17. Jacovi, A., Goldberg, Y.: Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? arXiv preprint arXiv:2004.03685 (2020)

18. Javed, S.A., Juyal, D., Padigela, H., Taylor-Weiner, A., Yu, L., Prakash, A.: Additive mil: Intrinsically interpretable multiple instance learning for pathology. Advances in Neural Information Processing Systems **35**, 20689–20702 (2022)
19. Kindermans, P.J., Hooker, S., Adebayo, J., Alber, M., Schütt, K.T., Dähne, S., Erhan, D., Kim, B.: The (un) reliability of saliency methods. In: Explainable AI: Interpreting, explaining and visualizing deep learning, pp. 267–280. Springer (2019)
20. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Li, B., Li, Y., Eliceiri, K.W.: 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)
22. Lu, M.Y., Chen, T.Y., Williamson, D.F., Zhao, M., Shady, M., Lipkova, J., Mahmood, F.: Ai-based pathology predicts origins for cancers of unknown primary. Nature **594**(7861), 106–110 (2021)
23. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
24. Maintainers, T., contributors, A.: Torchvision: Pytorch's computer vision library. GitHub repository (2016)
25. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. Advances in neural information processing systems **10** (1997)
26. Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., Bischl, B.: General pitfalls of model-agnostic interpretation methods for machine learning models. In: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers. pp. 39–68. Springer (2020)
27. Montavon, G., Binder, A., Lapuschkin, S., Samek, W., Müller, K.R.: Layer-wise relevance propagation: an overview. Explainable AI: interpreting, explaining and visualizing deep learning pp. 193–209 (2019)
28. Pirovano, A., Heuberger, H., Berlemont, S., Ladjal, S., Bloch, I.: Improving interpretability for computer-aided diagnosis tools on whole slide imaging with multiple instance learning and gradient-based explanations. In: International Workshop on Interpretability of Machine Intelligence in Medical Image Computing. pp. 43–53. Springer (2020)
29. Qu, L., Wang, M., Song, Z., et al.: Bi-directional weakly supervised knowledge distillation for whole slide image classification. Advances in Neural Information Processing Systems **35**, 15368–15381 (2022)
30. Rao, Y., Chen, G., Lu, J., Zhou, J.: Counterfactual attention learning for fine-grained visual categorization and re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1025–1034 (2021)
31. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
32. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International conference on machine learning. pp. 3145–3153. PMLR (2017)
33. Tang, W., Huang, S., Zhang, X., Zhou, F., Zhang, Y., Liu, B.: 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)

34. Tang, W., Zhou, F., Huang, S., Zhu, X., Zhang, Y., Liu, B.: Feature re-embedding: Towards foundation model-level performance in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11343–11352 (June 2024)
35. Tomczak, K., Czerwińska, P., Wiznerowicz, M.: Review the cancer genome atlas (tcga): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia **2015**(1), 68–77 (2015)
36. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)
37. Wagner, S.J., Reisenbüchler, D., West, N.P., Niehues, J.M., Zhu, J., Foersch, S., Veldhuizen, G.P., Quirke, P., Grabsch, H.I., van den Brandt, P.A., et al.: Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study. Cancer cell **41**(9), 1650–1661 (2023)
38. Wang, X., Wang, D., Yao, Z., Xin, B., Wang, B., Lan, C., Qin, Y., Xu, S., He, D., Liu, Y.: Machine learning models for multiparametric glioma grading with quantitative result interpretations. Frontiers in neuroscience **12**, 1046 (2019)
39. Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., Singh, V.: Nyströmformer: A nyström-based algorithm for approximating self-attention. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 14138–14148 (2021)
40. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: 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)
41. Zhang, Y., Li, H., Sun, Y., Zheng, S., Zhu, C., Yang, L.: Attention-challenging multiple instance learning for whole slide image classification. In: European conference on computer vision. pp. 125–143. Springer (2024)