

SC-MIL: Supervised Contrastive Multiple Instance Learning for Imbalanced Classification in Pathology

Dinkar Juyal
 PathAI Inc
 Boston, USA

dinkar.juyal@pathai.com

Siddhant Shingi*
 University of Massachusetts
 Amherst, USA

Syed Ashar Javed
 PathAI Inc

Harshith Padigela
 PathAI Inc

Chintan Shah
 PathAI Inc

Anand Sampat
 PathAI Inc

Archit Khosla
 PathAI Inc

John Abel
 PathAI Inc

Amaro Taylor-Weiner
 PathAI Inc

Abstract

Multiple Instance learning (MIL) models have been extensively used in pathology to predict biomarkers and risk-stratify patients from gigapixel-sized images. Machine learning problems in medical imaging often deal with rare diseases, making it important for these models to work in a label-imbalanced setting. In pathology images, there is another level of imbalance, where given a positively labeled Whole Slide Image (WSI), only a fraction of pixels within it contribute to the positive label. This compounds the severity of imbalance and makes imbalanced classification in pathology challenging. Furthermore, these imbalances can occur in out-of-distribution (OOD) datasets when the models are deployed in the real-world. We leverage the idea that decoupling feature and classifier learning can lead to improved decision boundaries for label imbalanced datasets. To this end, we investigate the integration of supervised contrastive learning with multiple instance learning (SC-MIL). Specifically, we propose a joint-training MIL framework in the presence of label imbalance that progressively transitions from learning bag-level representations to optimal classifier learning. We perform experiments with different imbalance settings for two well-studied problems in cancer pathology: subtyping of non-small cell lung cancer and subtyping of renal cell carcinoma. SC-MIL provides large and consistent improvements over other techniques on both in-distribution (ID) and OOD held-out sets across multiple imbalanced settings.

1. Introduction

Pathology is the microscopic study of tissue and a key component in medical diagnosis and drug development [28]. The digitization of tissue slides, resulting in whole slide images (WSIs), has made pathology data more accessible for quantitative analysis. However, the large size (billions of pixels) and information density (hundreds of thousands of cells and heterogeneous tissue organization) of WSIs make manual analysis challenging [13, 25], highlighting the need for machine learning (ML) approaches [1–3, 5, 9, 18, 21, 29]. ML techniques have been used for predicting a patient’s clinical characteristics from a WSI. These models predict a label or score for the entire WSI, referred to as a slide-level prediction. Traditional approaches for handling large WSIs include the use of hand-engineered representations or breaking the slide into thousands of smaller patches [8]. Both of these approaches require pixel or patch level annotations which are costly. To overcome the need for patch level labels, multiple instance learning (MIL) [22] has been applied to pathology by treating patches from slides as instances that form a bag, with a slide-level label associated with each bag. The MIL framework thus provides an end-to-end learning approach for problems in pathology.

Label distribution in real-world settings can vary considerably depending on factors such as disease prevalence, population characteristics and the hospital or laboratory of origin. For example, a dataset of WSIs from a diagnostic lab may have a different class distribution compared to a dataset from a clinical trial enriched for certain disease characteristics. In fact, label imbalance in pathology datasets

*Work done during internship at PathAI

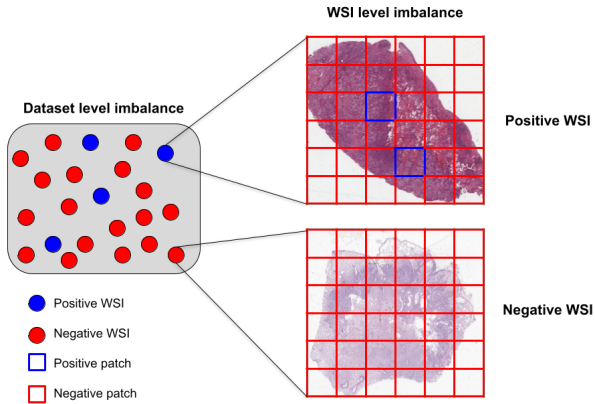


Figure 1. Label imbalance in histopathology domain is present at two levels - Dataset and Whole Slide Image (WSI). In datasets, imbalance arises from different prevalence rates of diseases. For a given WSI with positive label, only a small subset of patches contribute to that positive label. This compounds the severity of imbalance, making imbalanced classification in pathology challenging.

exists at both dataset and WSI level as shown in Figure 1. MIL models should be robust to variations in label distribution to succeed in clinical applications and maintain physician trust. Different approaches have been proposed to deal with label imbalance, ranging from data resampling (oversampling of minority classes or undersampling of majority classes) [23,24], loss reweighting [26], selective enrichment of minority classes in image or feature space [6], decoupling representation learning from classification [34], and custom loss functions [4].

Contrastive learning aims to learn representations that maximize the agreement among positive instances, e.g., different augmentations of the same image, and minimize the agreement with negative instances, e.g., other images in the dataset [7]. In supervised contrastive learning (SCL) [16], the contrastive loss formulation incorporates label information by treating all instances within the same class as positive examples for a given image. SCL adopts a two-stage learning technique where a feature extractor is learned in the first stage using a contrastive loss, followed by learning a classifier using the cross-entropy loss in the second stage.

This work proposes SC-MIL: a novel MIL technique to tackle label imbalance in pathology, that integrates SCL into the MIL framework. We take inspiration from prior work [10, 14] which shows that a) contrastive loss learns balanced feature spaces (i.e., feature spaces with similar inter-class separation for all classes) compared to cross-entropy, and b) this balance is positively related to performance across imbalanced settings. Additionally, we use a smooth transition from feature learning to classifier learning in the course of training, which allows the model to learn a

Table 1. Training data-distribution of TCGA RCC sub-typing across imbalance ratios

Classes	Imbalance Ratio		
	1	5	10
KIRC	96	205	240
KIRP	96	41	24
KICH	96	41	24
Total	288	287	288

more discriminative latent space, aiding in imbalanced classification [31]. In the MIL setting, labels are only available for a bag (i.e., a collection of patches) and not individual patches. Applying SCL to patch features assumes assigning a bag-label to individual patches. A prior work [19] iteratively assigns pseudo-labels to patches from bag-level labels. However, a single patch might not have any information about the WSI label. For example, a malignant WSI might have many patches which contain only normal tissue. This motivates our bag-level formulation of SC-MIL where contrastive loss is applied to the bag features. Feature learning with bag-level contrastive loss tackles dataset imbalance, while the multiple instance formulation addresses imbalance within a WSI. The contributions of this work are as follows:

1. We tackle the problem of label imbalance by proposing a formulation that extends SCL to the MIL setting. We investigate two training strategies for optimal feature and classifier learning with SC-MIL.
2. We conduct an extensive study on the performance of this technique across different degrees of label imbalance on two open-source datasets: subtyping in non-small cell lung cancer (NSCLC) and renal cell carcinoma (RCC). We compare this to previous state-of-the-art methods used for label imbalance and demonstrate the effectiveness of using SC-MIL over these methods.
3. We show substantial performance improvements with SC-MIL on OOD data across multiple degrees of label imbalance, making a strong case for its utility in real-world deployment scenarios.

2. Supervised Contrastive Multiple Instance Learning

2.1. Multiple Instance Learning

MIL is a weakly supervised learning approach that allows learning and making predictions on a group of instances. Unlike supervised learning, the MIL framework

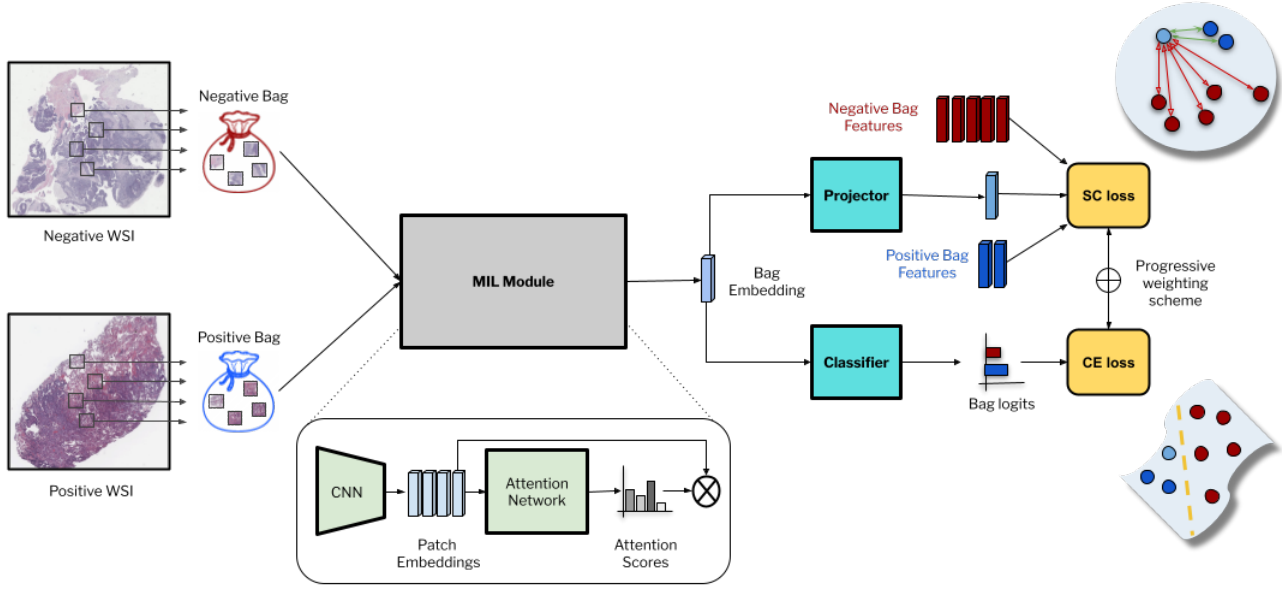


Figure 2. SC-MIL integrates supervised contrastive learning into the MIL framework. The model performs joint feature and classifier learning on bag representations computed using an attention-based aggregation on patches. The training objective transitions progressively from a contrastive to a classification loss.

Table 2. Training data-distribution of TCGA NSCLC sub-typing across imbalance ratios

Classes	Imbalance Ratio		
	1	5	10
LUAD	158	265	290
LUSC	158	53	29
Total	316	318	319

only requires labels for the group of instances, called a bag, but not the individual instances. This is valuable in the context of pathology, where a collection of patches from a WSI can be treated as a bag and this allows learning of slide-level predictors without the need for fine-grained patch-level annotations. For pooling of patches, a learnt attention based aggregation scheme [11] has been shown to be effective and is commonly used in end-to-end pathology models.

In the binary case, a bag is considered positive if it has at least one positive instance and negative if there are none. Given a set of instances $X = \{x_0, x_1, \dots, x_n\}$, the MIL prediction $p(X)$ is

$$p(X) = a(f(x_0), f(x_1), \dots, f(x_n)) \quad (1)$$

where f is an encoder for instances, a is a permutation-invariant aggregator, mapping from feature space to the

prediction space. Learnt aggregation functions like AttentionMIL and its variants DSMIL [17], CLAM [20], TransMIL [27], AdditiveMIL [12] have shown significant improvements over heuristic aggregators like Max or Mean in various tasks [11]. We will focus on the AttentionMIL (also referred to as ABMIL) formulation for our discussion.

The aggregator function a in AttentionMIL has two components. An attention module m induces a soft-attention α_i over the instances and computes an attention weighted aggregation of instance features to generate the bag embedding $b(X)$. A classifier h maps the bag feature to the bag prediction.

$$p(X) = h(b(X)) \quad (2)$$

$$b(X) = m(f(x_0), f(x_1), \dots, f(x_n)) = \sum_{i=0}^{i=n} \alpha_i f(x_i) \quad (3)$$

$$\alpha_i = \text{softmax}(\phi_m(x_i)) \quad (4)$$

where ϕ_m is a neural network with a non-linear activation.

Table 3. Comparison of SC-MIL with other label imbalance techniques on TCGA-RCC test set for RCC subtyping (RS - Random Sampling, CB - Class Balanced)

Dataset	TCGA-RCC					
	1		5		10	
	F1	AUC	F1	AUC	F1	AUC
ABMIL-RS	87.17±2.03	96.13±0.96	83.40±2.52	95.6±0.99	78.23±2.82	93.26±1.39
ABMIL-CB	89.49±2.12	97.42±0.85	84.61±2.84	93.93±1.60	73.63±2.83	95.05±0.84
LDAM-DRW	89.35±1.92	97.65±0.66	83.50±2.45	94.96±1.23	80.66±2.93	93.08±1.54
SC-MIL-RS	88.67±2.21	98.13±0.67	86.30±2.03	96.83±0.66	87.42±2.07	96.40±0.94
SC-MIL-CB	90.13±2.17	97.98±0.65	85.53±2.55	96.69±0.82	81.34±2.66	96.35±1.11

Table 4. Comparison of SC-MIL with other label imbalance techniques on OOD-RCC test set for RCC subtyping (RS - Random Sampling, CB - Class Balanced)

Dataset	OOD-RCC					
	1		5		10	
	F1	AUC	F1	AUC	F1	AUC
ABMIL-RS	74.20±1.91	93.88±1.09	73.69±2.55	91.15±1.56	70.23±3.42	87.74±2.10
ABMIL-CB	77.33±2.55	92.79±1.48	72.31±2.33	89.49±1.65	71.38±2.88	91.82±1.47
LDAM-DRW	78.97±2.45	93.69±1.38	73.47±2.52	88.62±1.97	72.42±2.51	91.94±1.52
SC-MIL-RS	81.94±2.39	94.84±1.24	81.87±2.54	93.42±1.43	80.91±2.24	92.57±1.41
SC-MIL-CB	76.81±2.31	93.78±7.83	79.04±2.34	92.56±1.51	79.04±2.34	92.56±1.51

2.2. SC-MIL: Supervised Contrastive Multiple Instance Learning

SCL [16] proposes a way to leverage contrastive learning and incorporate supervision. It learns instance representations by pulling instances from same class together and those from different classes apart in the representation space. In MIL, we can use SCL for learning either instance or bag representations. Considering we only have labels for bags and not individual instances, using SCL to learn instance representations needs using bag labels as instance labels, thus introducing label noise and breaking the MIL assumption. Instead, we use SCL to learn bag representations.

Specifically, given a set of instances for a bag $X_i = \{x_0, x_1, \dots, x_n\}$, we compute the bag representation $b(X_i)$ using the MIL formulation, where i denotes the index of a bag in a given batch. We now use a non-linear multi-layer perceptron g to generate the projection z_i for the bag representation. We then compute the SCL loss for MIL \mathcal{L}_{SCL} as follows:

$$z_i = g(b(X_i)) \quad (5)$$

$$\mathcal{L}_{SCL} = \sum_i -\frac{1}{|\mathbf{P}_i^+|} \sum_{z_j \in \mathbf{P}_i^+} \log \frac{\exp(z_i \cdot z_j / \tau)}{\sum_{z_k \in \mathbf{B}_i} \exp(z_i \cdot z_k / \tau)} \quad (6)$$

where \mathbf{P}_i^+ denotes the positive bags sharing the same class label as bag z_i and \mathbf{B}_i is the set of all bags in the batch excluding bag z_i .

Curriculum-based feature and classifier learning using both contrastive and cross entropy losses has been shown to be effective in long-tailed image classification [15]. We apply the same approach to the MIL setting at a bag level. For classifier learning, we use the cross-entropy loss. The classifier branch projects the bag embedding $b(X)$ to the prediction $p(X)$ as shown in Equation 2 and uses cross entropy \mathcal{L}_{CE} to learn the classifier:

$$\mathcal{L}_{SC-MIL} = \beta_t \mathcal{L}_{SCL} + (1 - \beta_t) \mathcal{L}_{CE} \quad (7)$$

where the weight $\beta_t \in [0, 1]$ is decayed through the course of training iterations t using a curriculum to gradually transition from feature to classifier learning.

Table 5. Comparison of SC-MIL with other label imbalance techniques on TCGA-NSCLC test set for NSCLC subtyping (RS - Random Sampling, CB - Class Balanced)

Dataset	TCGA-NSCLC					
	1		5		10	
	F1	AUC	F1	AUC	F1	AUC
ABMIL-RS	82.07±2.5	91.36±1.76	81.62±2.62	89.84±2.03	77.35±3.18	89.68±2.14
ABMIL-CB	83.23±2.45	91.81±1.76	82.61±2.72	90.13±1.99	78.0±2.95	88.57±2.18
LDAM-DRW	85.8±2.28	91.9±1.78	80.91±2.91	89.89±2.13	81.56±2.79	89.14±2.15
SC-MIL-RS	87.65±2.21	94.81±1.34	86.66±2.34	92.45±1.78	84.05±2.65	91.64±1.81
SC-MIL-CB	84.85±2.52	93.14±1.65	87.02±2.29	92.21±1.87	80.37±3.02	90.96±1.93

Table 6. Comparison of SC-MIL with other label imbalance techniques on OOD-NSCLC test set for NSCLC subtyping (RS - Random Sampling, CB - Class Balanced)

Dataset	OOD-NSCLC					
	1		5		10	
	F1	AUC	F1	AUC	F1	AUC
ABMIL-RS	71.03±5.34	92.42±2.33	19.82±7.91	75.95±4.24	12.34±6.5	88.61±2.64
ABMIL-CB	36.3±7.97	92.54±1.93	15.58±7.04	77.25±4.41	8.14±5.57	76.71±4.01
LDAM-DRW	61.63±6.81	90.31±2.5	26.28±8.11	89.57±2.56	28.85±8.02	88.36±2.89
SC-MIL-RS	76.46±5.27	93.64±2.45	37.68±7.75	91.58±2.02	41.5±8.23	92.97±1.95
SC-MIL-CB	58.82±6.68	84.83±3.08	49.65±8.26	94.23±1.95	29.06±8.2	79.04±4.4

3. Experiments and Results

We first introduce the datasets used for experimentation. We describe the mechanism of simulating different degrees of imbalance in these datasets while ensuring that the total number of samples remains consistent. We then discuss results on all datasets using SC-MIL and other baselines. Finally, we present ablation studies to understand the trade-offs made in terms of training supervised contrastive loss with cross-entropy jointly vs sequentially, and the impact of hyperparameters.

3.1. Datasets and Setup

We considered two datasets from The Cancer Genome Atlas (TCGA) [32] - prediction of cancer subtypes in non-small cell lung carcinoma (NSCLC) and renal cell carcinoma (RCC). TCGA-NSCLC contains a total of 1002 WSIs stained with H&E, 538 of which were collected from patients with the adenocarcinoma histologic subtype (LUAD) and 464 from squamous cell carcinoma (LUSC). TCGA-RCC contains 948 WSIs with three histologic subtypes: 158

WSIs with the label chromophobe RCC (KICH), 504 WSIs belonging to clear cell RCC (KIRC), and 286 to papillary RCC (KIRP).

We performed a label-stratified split of both datasets while ensuring there is no leakage of case information (i.e., combination of tissue source site and study participant) across splits. The splitting ratio was 60:15:25 (train:val:test); other clinical or sample characteristics were not used during splitting. To simulate varying degrees of label imbalance, we sampled WSIs from the available classes to generate imbalance in the train set, while the heldout sets were kept the same. In line with previous works [4, 31], we used imbalance ratio $\rho = \frac{\max_i \{n_i\}}{\min_i \{n_i\}}$ which denotes the ratio of number of examples of the majority class to the minority class. We experimented with imbalance ratios of 1, 5 and 10. We ensured that the number of training examples remained consistent across different imbalance ratios to remove any confounding effect of the number of data points and to enable comparison of model performance across imbalance ratios. Since there were three classes in TCGA-RCC, the two classes with least number of samples (KIRP

Algorithm 1 SC-MIL Pseudocode, PyTorch-like

```
# f: patch level feature extractor
# m: attention module
# g: projector MLP
# h: classifier
# X: bag of patches
# Y: bag label

# load a bag X=[x_1, ..., x_n] with n patches
for X in loader:
    # patch level embeddings, n-by-d
    E = f(X)
    # attention weights, n-by-1
    attn_wts = m(E)
    # bag level embedding, d-by-1
    B = Sum(attn_wts*X)
    # projected bag embedding
    Z = norm(g(B))
    # bag level predictions
    P = h(B)

    # t_i: the current iteration
    # t: total number of iterations
    beta_t = 1 - t_i/t
    L_scl = scl_loss(Z, other Z_i in minibatch)
    L_ce = cross_entropy_loss(P, Y)
    Loss = beta_t*L_scl + (1-beta_t)*L_ce

    Loss.backward()
```

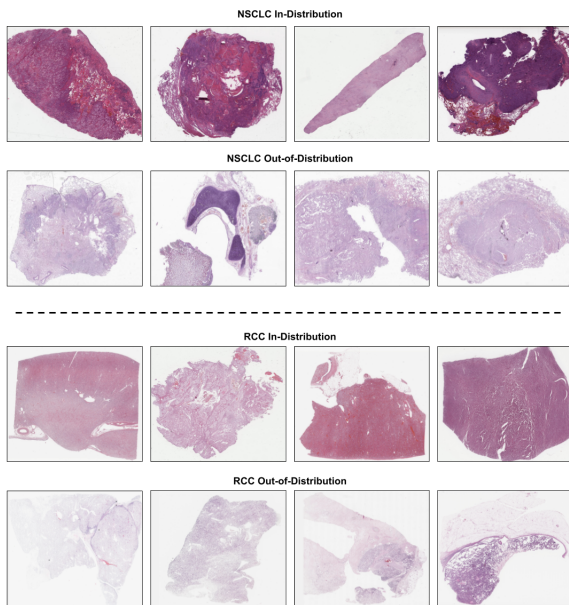


Figure 3. Visual comparison of in-distribution (ID) and out-of-distribution (OOD) WSIs from the cancer subtyping datasets. The first and third rows show in-distribution TCGA WSIs from NSCLC and RCC respectively. The second and fourth rows show WSIs procured from a different lab site and scanner. We can see the variations in tissue preparation and scanning which lead to significant drops in performance.

and KICH) were treated as minority classes. The details of the resulting dataset composition is shared in Table 1 and 2.

We also deployed all models on two OOD datasets col-

lected from different patient populations and having different sample characteristics for NSCLC and RCC. These OOD datasets are acquired from other laboratories using varying image acquisition and processing steps resulting in visual differences from their TCGA counterparts. OOD NSCLC has 162 LUAD and 45 LUSC WSIs, while OOD RCC has 254 KIRC, 134 KIRP and 46 KICH WSIs. Example images comparing ID and OOD datasets are shared in Figure 3.

3.2. Implementation Details

We trained five models: a baseline AttentionMIL model with random sampling (ABMIL-RS) and class balanced sampling (ABMIL-CB), a version using label-distribution-aware margin loss with deferred reweighting (LDAM-DRW [4], previously shown to be successful for addressing label imbalance in single instance classification), and our proposed SC-MIL with random (SC-MIL-RS) and class balanced sampling (SC-MIL-CB). Non-overlapping patches of size 224×224 pixels were selected from tissue regions (using a separate model which masks background and artifacts) at a resolution of 1 micron per pixel. We extracted 1.45 million patches from TCGA-NSCLC and 768k patches from TCGA-RCC. Bag sizes (number of patches in a bag) varied from 24 to 1500 patches and batch sizes (number of bags in a batch) varied from 8 to 32. Augmentations applied included color-based augmentations (random grayscaling, HSV transforms), gaussian blur and sharpening, horizontal and vertical flips, center crops. Augmentation related parameters were kept consistent across all techniques. An ImageNet-pretrained ShuffleNet [33] was used to extract features from input patches. All models were trained end-to-end with the Adam optimizer and a learning rate of $1e-4$. SC-MIL models were trained with a temperature $\tau = 1$, and the training was performed jointly with cross entropy with a linear curriculum as described in Section 2.2, with $\beta_t = 1$ at the start of training. For inference, patches were exhaustively sampled from a WSI and the majority prediction across bags was selected as the WSI-level prediction. For RCC, macro-averaged F1 score and macro-average of 1-vs-rest AUROC was computed. Standard deviation for the metrics was computed over 1000 bootstrapped runs on the test set. Training and inference was performed on Quadro RTX 8000 GPUs using PyTorch v1.11 and CUDA 10.2. The training time for SC-MIL was comparable with other techniques (10-14 GPU hours). The code, along with the trained SC-MIL models for the three imbalance ratios on the two datasets, is shared in the Supplementary material.

3.3. Experimental Results and Ablation Studies

3.3.1 Comparison of SC-MIL with other techniques

We compared the predictive performance of SC-MIL with other techniques across different imbalance ratios. Table 3

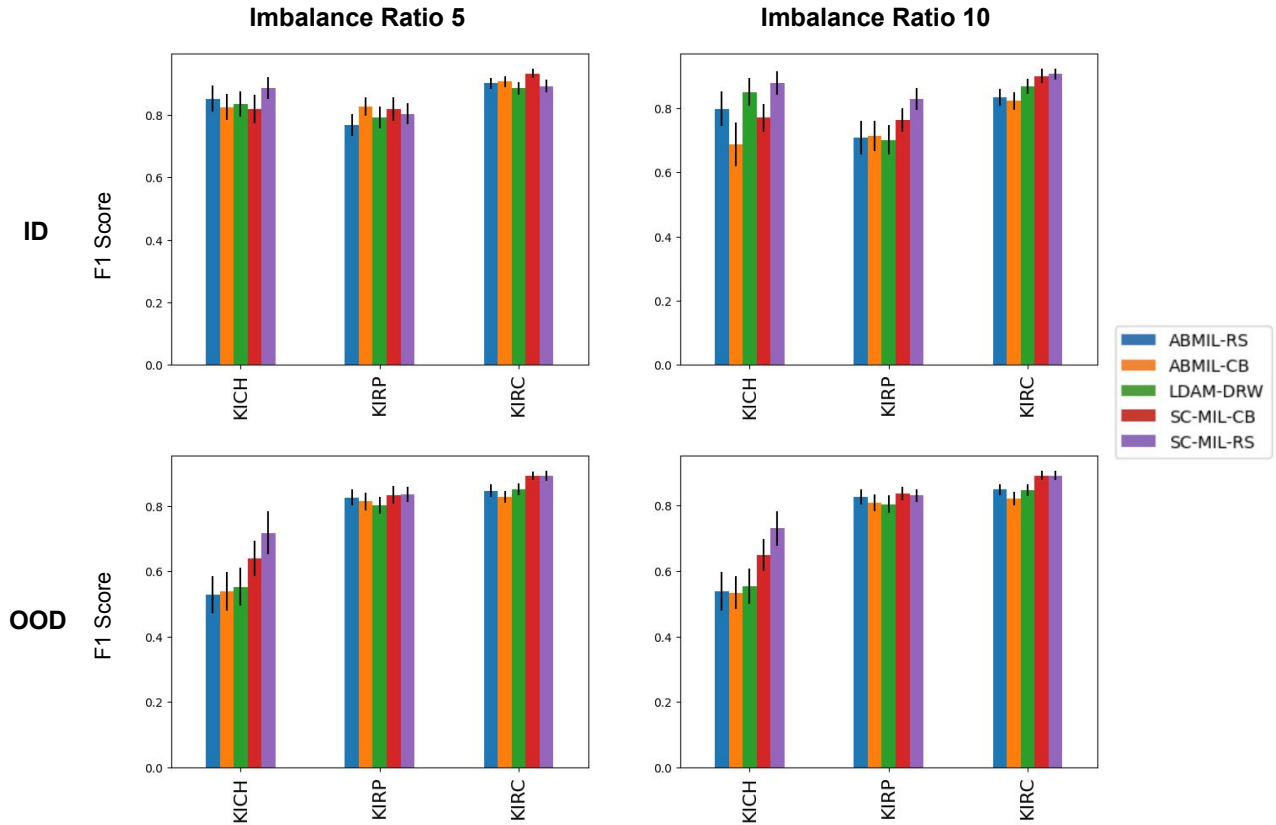


Figure 4. Class-wise F1 score comparison for RCC subtyping: SC-MIL outperforms other methods across different imbalance ratios. The performance gains are higher on minority classes, and they increase on moving from ID to OOD datasets.

Table 7. Comparison of Patch vs Bag-level SCL on TCGA-RCC subtyping. All comparisons used SC-MIL-RS training and $\tau = 1$

Ratio	Method	F1	AUC
1	SCL-RS	88.67±2.21	98.13±0.67
	Patch SCL	82.87±2.81	97.33±0.82
5	SCL-RS	86.30±2.03	96.83±0.66
	Patch SCL	79.68±2.87	94.71±1.05
10	SCL-RS	87.42±2.07	96.40±0.94
	Patch SCL	79.80±2.57	94.90±1.09

and 5 show results on the NSCLC and RCC test sets respectively. SC-MIL outperforms other techniques across all imbalance ratios, and the difference is more pronounced at higher imbalance ratios. To further stress test these methods, we also deployed these models on independent OOD

test datasets described above and the results are shown in Table 4 and 6. We found that baseline model performance dropped notably across imbalance ratios, highlighting the difficulty in generalization, and the tendency of these models to overfit in an imbalanced setting. Performance improvements using SC-MIL persist in this OOD setting. In Figure 4 we show the performance of different techniques across all classes in RCC in both ID and OOD setting, demonstrating the relative performance gain in each class.

3.3.2 Patch vs bag based SC-MIL

We conduct an experiment with a modification of SC-MIL architecture, where the supervised contrastive loss is applied on patch level embeddings instead of bag level embeddings. In Section 2.2, we theorized that naively assigning the bag level label to instances and then applying supervised contrastive loss will result in incorrect label assignment. We

Table 8. Impact of temperature (τ) on single-stage training SC-MIL-RS on TCGA-RCC subtyping.

Imb. Ratio	Temp	F1	AUC
1	0.1	86.24±2.35	97.85±0.60
	0.5	87.88±2.37	97.87±0.72
	1.0	88.67±2.21	98.13±0.67
5	0.1	86.14±2.61	97.04±0.96
	0.5	88.35±2.34	97.49±0.68
	1.0	86.30±2.03	96.83±0.66
10	0.1	85.60±2.37	96.09±0.97
	0.5	84.85±2.30	96.06±0.78
	1.0	87.42±2.07	96.40±0.94

show the results of training with such a scheme in Table 7. Patch level SC-MIL has inferior performance and higher variance as compared to our formulation. We also observe that the performance gap between the two models increases with increasing imbalance ratio, providing evidence that our bag-level formulation is more robust to the compounding effect of label imbalance in pathology.

3.3.3 Impact of sampling

We found that SC-MIL with random sampling performs better than class balanced sampling in most cases. We hypothesize that this is due to reduced diversity in the feature space as a side effect of oversampling the minority classes or under sampling the majority class when using class-balanced sampling, which ultimately hurts performance by interfering with feature learning [31].

3.3.4 Impact of temperature

We experimented with temperature values of $\tau \in \{0.1, 0.5, 1\}$ and found that the models are generally robust to temperature changes as shown in Table 8. We reason about this through two desirable properties of representations learned through contrastive learning: *uniformity* in the hypersphere, i.e, inter-class separation and *tolerance* to potential positives, i.e., intra-class similarity [30]. The former is favored by low values of temperature while higher values favor the latter. As shown in [30], in problems with a larger number of classes, uniformity is harder to achieve and higher values of temperature harm feature quality. In contrast, we see that for RCC and NSCLC subtyping with 3 and 2 classes respectively, model performance is less sensitive to changes in temperature.

Table 9. Comparison of one-stage vs two-stage training on TCGA-RCC subtyping. All comparisons used SC-MIL-RS training and $\tau = 1$

Ratio	Stage	F1	AUC
1	1	88.67±2.21	98.13±0.67
	2	87.54±2.10	97.66±0.60
5	1	86.30±2.03	96.83±0.66
	2	86.06±2.47	97.34±0.77
10	1	87.42±2.07	96.40±0.94
	2	85.26±2.56	96.01±0.91

3.3.5 Two-stage vs single-stage training

We conducted an ablation by training models in a two-stage manner, with SCL loss in the first stage for feature learning followed by cross-entropy (CE) loss in the second stage. We see that single-stage SC-MIL model (joint SCL and CE training) performs better overall as shown in Table 9. This could be due to incompatible feature learning between SCL and CE stages in two-stage training. Using a smooth curriculum allows a gradual transition from feature learning to classifier learning, leading to superior performance.

4. Conclusion

Label imbalance in pathology is a challenging problem owing to the highly skewed distribution of classes both at dataset and WSI level. We propose SC-MIL, a novel integration of supervised contrastive learning into the MIL framework to tackle this label imbalance problem. Experiments show our bag-level formulation to be superior to patch-level SC-MIL and other baselines across multiple degrees of label imbalance. Moreover, these improvements persist in out-of-distribution pathology datasets. We hope that this improved generalization performance in imbalanced settings drives adoption of ML in real-world clinical applications.

5. Acknowledgement

The authors would like to thank Victoria Mountain and Jacqueline Brosnan-Cashman for their valuable feedback in drafting this work.

References

- [1] Jaime Bosch, Chuhan Chung, Oscar M Carrasco-Zevallos, Stephen A Harrison, Manal F Abdelmalek, Mitchell L Shiffman, Don C Rockey, Zahir Shanis, Dinkar Juyal, Harsha Pokkalla, et al. A machine learning approach to liver histological evaluation predicts clinically significant portal hypertension in nash cirrhosis. *Hepatology*, 74(6):3146–3160, 2021. 1
- [2] Wouter Bulten, Maschenka Balkenhol, Jean-Joël Awoumou Belinga, Américo Brilhante, Aslı Çakır, Lars Egevad, Martin Eklund, Xavier Farré, Katerina Geronatsiou, Vincent Molinié, et al. Artificial intelligence assistance significantly improves gleason grading of prostate biopsies by pathologists. *Modern Pathology*, 34(3):660–671, 2021. 1
- [3] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Mirafior, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 1
- [4] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arcehiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019. 2, 5, 6
- [5] Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16144–16155, June 2022. 1
- [6] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: Rebalanced mixup. In *Computer Vision – ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI*, page 95–110, Berlin, Heidelberg, 2020. Springer-Verlag. 2
- [7] Ozan Ciga, Anne L. Martel, and Tony Xu. Self supervised contrastive learning for digital histopathology. *ArXiv*, abs/2011.13971, 2020. 2
- [8] James A Diao, Jason K Wang, Wan Fung Chui, Victoria Mountain, Sai Chowdary Gullapally, Ramprakash Srinivasan, Richard N Mitchell, Benjamin Glass, Sara Hoffman, Sudha K Rao, et al. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nature communications*, 12(1):1–15, 2021. 1
- [9] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, , and the CAMELYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, 12 2017. 1
- [10] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. 2021. 2
- [11] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 3
- [12] Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and aaditya prakash. Additive MIL: Intrinsically interpretable multiple instance learning for pathology. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 3
- [13] Syed Ashar Javed, Dinkar Juyal, Zahir Shanis, Shreya Chakraborty, Harsha Pokkalla, and Aaditya Prakash. Re-thinking machine learning model evaluation in pathology. *arXiv preprint arXiv:2204.05205*, 2022. 1
- [14] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2021. 2
- [15] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. 4
- [16] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020. 2, 4
- [17] 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*, pages 14318–14328, 2021. 3
- [18] Jin Li, Deepta Rajan, Chintan Shah, Dinkar Juyal, Shreya Chakraborty, Chandan Akiti, Filip Kos, Janani Iyer, Anand Sampat, and Ali Behrooz. Self-training of machine learning models for liver histopathology: Generalization under clinical shifts, 2022. 1
- [19] 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*, pages 3355–3365, 2023. 2
- [20] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 3
- [21] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016. 20th anniversary of the Medical Image Analysis journal (MedIA). 1
- [22] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997. 1

- [23] Carlos Mera, Jose Arrieta, Mauricio Orozco-Alzate, and John Branch. A bag oversampling approach for class imbalance in multiple instance learning. In Alvaro Pardo and Josef Kittler, editors, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 724–731. Cham, 2015. Springer International Publishing. 2
- [24] Ajinkya More. Survey of resampling techniques for improving classification performance in unbalanced datasets. *CoRR*, abs/1608.06048, 2016. 2
- [25] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. Digital pathology and artificial intelligence. *The lancet oncology*, 20(5):e253–e261, 2019. 1
- [26] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4334–4343. PMLR, 10–15 Jul 2018. 2
- [27] 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, 2021. 3
- [28] Eric E Walk. The role of pathologists in the era of personalized medicine. *Archives of pathology & laboratory medicine*, 133(4):605–610, 2009. 1
- [29] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*, 2016. 1
- [30] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2495–2504, 2021. 8
- [31] Peng Wang, K. Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 943–952, 2021. 2, 5, 8
- [32] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013. 5
- [33] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 6
- [34] B. Zhou, Q. Cui, X. Wei, and Z. Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9716–9725, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. 2