Remixing Embeddings for Patient Augmentation in Data Scarce Multiple Instance Learning

Muhammed Furkan Dasdelen[1*], Fatih Ozlugedik[1*], Anastasia Litinetskaya[1], Nassir Navab[2,3], Carsten Marr[1,3,4,5,6#], and Ario Sadafi[1,2,3#]

[1] Computational Health Center & Helmholtz AI, Helmholtz Munich, Munich, Germany
[2] Computer Aided Medical Procedures, Technical University of Munich, Munich, Germany
[3] Munich Center for Machine Learning (MCML), Munich, Germany
[4] Department of Medicine III, Ludwig Maximilian University Hospital, Munich, Germany
[5] Department of Physics, Ludwig-Maximilians-Universität München, Munich, Germany
[6] DKTK, German Cancer Consortium, partner site Munich, Germany

Abstract. Data scarcity is a major bottleneck in medical Multiple Instance Learning (MIL), especially for rare diseases and expensive modalities. We introduce a statistically grounded patient augmentation approach that generates realistic patients directly in embedding space. Using Gaussian mixture models to probabilistically cluster pooled instance embeddings from all patients, our method learns disease-specific "recipes"–statistical distributions of instances across unsupervised clusters. New patients are then generated by sampling embeddings from clusters based on learned recipes. Unlike existing methods that require examples from all categories, our method can generate patients offline by remixing pooled embeddings. Generated patients are further selected based on uncertainty quantification to improve MIL performance. We evaluate our method across three clinically relevant scenarios: (i) cross-dataset transfer, where an entirely missing healthy class is generated using statistics from an external cohort; (ii) low-data regimes, where class sizes are extremely limited; and (iii) small-cohort non-image tasks, including single-cell RNA-seq and flow cytometry. Across all experiments, our method improves performance over baseline, often outperforming other bag-mixing strategies. Notably, in the missing-class scenario, a performance comparable to full-dataset training is achieved, demonstrating its potential for rare disease diagnostic and privacy-preserving patient augmentation. The code is available at https://github.com/marrlab/RECIPE.

Keywords: single cell · pathology · flow cytometry · cytology

[*] Equal contribution
[#] Co-corresponding: {carsten.marr,ario.sadafi}@helmholtz-munich.de

1 Introduction

Multiple Instance Learning (MIL) has emerged as a powerful paradigm for weakly supervised learning in medical imaging, where only bag-level labels (e.g., patient diagnoses) are available and instance-level annotations (e.g., individual cell or image patch labels) are impractical to obtain [17]. In MIL, a "bag" comprises a set of instances—such as histopathology patches [13], single cell images [4], or genomic/immune features [10,2]—and the learning objective is to predict the bag label while weighting key instances.

One persistent bottleneck in healthcare MIL is data scarcity, particularly in rare diseases and in diseases whose diagnosis depends on specialized, expensive technologies or invasive data collection. Traditional augmentation techniques—rotations, flips, and color perturbations—operate at the instance level and fail to generate realistic variation at the bag level. We introduce Remixing Embeddings via Clustering for Patient augmEntation (RECIPE), which generates new bags for MIL pipelines in data-scarce settings. Given disease-level labels from internal or external datasets, RECIPE unsupervisedly extracts instance "recipes" needed to form each disease class and enables uncertainty quantification. We showcase our pipeline in three scenarios: (i) when one class has no patients and the recipe is derived from an external dataset, (ii) when only a few patients exist in one class, and (iii) when the dataset is naturally small, as in single-cell RNA-seq or flow cytometry.

2 Related work

Data augmentation improves generalization in machine learning [22,19,14,16], yet its application in MIL is limited. Since whole slide images (WSI) contain largely non-discriminative patches, naive instance-level augmentation often amplifies noise and redundancy. Moreover, in MIL, patch-level augmentation entails significant computational overhead due to the need for feature re-extraction.

To address this, bag-level strategies have been developed to generate new bags by mixing instances. PseMix [11] generates synthetic bags in a mini-batch by combining patches from two WSIs into one mixed bag; the label is mixed in the same proportion. ReMix [21] similarly augments within the training batch, but first summarizes each WSI into a few representative patch groups and then mixes those summaries. While these methods increase diversity, they face critical limitations. First, they operate exclusively online, producing ephemeral synthetic bags during training that cannot be validated or reused. Second, their local mixing scope fails to capture global population statistics. Most critically, they require real examples from every class, rendering them ineffective when an entire class is missing—a frequent challenge when a healthy control or a rare disease class is missing.

We introduce RECIPE, a paradigm for offline, statistically-grounded patient generation. Unlike local mixing, RECIPE models global instance distributions via clustering to learn class-specific "recipes" of disease phenotypes. This enables the creation of persistent, analyzable cohorts. Uniquely, RECIPE supports

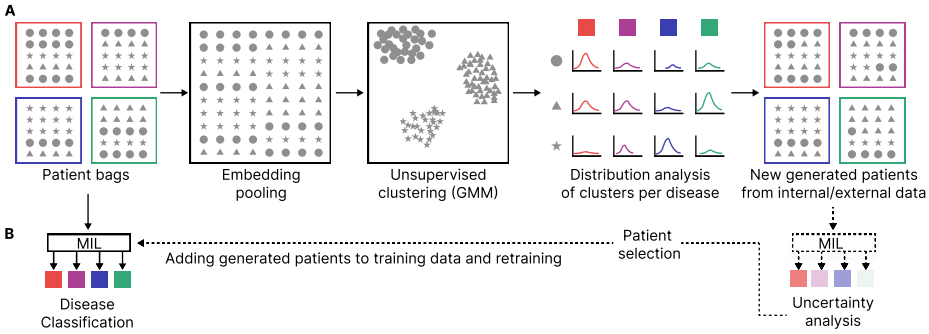

Fig. 1. Remixing embeddings via clustering for patient augmentation (RECIPE). (A) A Gaussian mixture model (GMM) is fitted to pooled instance embeddings to define unsupervised clusters. Class-specific "recipes" are computed as distributions across these clusters (mean ± s.d.). New patients are generated by sampling instances according to these recipes. **(B)** Informative patients are selected by quantifying patient-level uncertainty via Monte Carlo dropout. The model is then retrained using the most uncertain generated samples to improve performance.

cross-dataset transfer, generating missing classes using external statistics without sharing sensitive data. To our knowledge, this is the first MIL framework addressing missing patient classes, with applicability extending to single-cell and cytometry data.

3 Methods

3.1 Dataset and Preprocessing

We evaluated RECIPE on both image (hematologic cytology) and non-image (single-cell RNA-seq, flow cytometry) datasets where limited sample sizes benefit from patient-level augmentation.

AML-Hehr [4] contains single white blood cell (WBC) images from 189 patients, including healthy donors ($n = 60$) and four AML genetic subtypes ($n = 129$; *CBFB::MYH11*, *NPM1*, *PML::RARA*, *RUNX1::RUNX1T1*), collected at Munich Leukemia Laboratory (MLL) (2009–2020).

cAItomorph [1] comprises peripheral blood smear images from 2043 patients across seven hematological conditions and healthy donors, collected at MLL (2021–2022). Images were resized to 224×224, normalized (ImageNet statistics), and embedded into 768-dimensional vectors using the DinoBloom-B [7] hematology foundation model.

Covid-scRNA-seq [18] is a single-cell RNA-seq cohort of 130 donors (647,366 cells) with varying COVID-19 severity. Following [10], we utilized three major classes: healthy ($n = 23$), mild ($n = 23$), and severe ($n = 13$). Cell embeddings were generated using PCA (30-dim) and scVI [12] (50-dim; default parameters with *site* and *sample ID* covariates).

Covid-flow [9] consists of flow cytometry data from COVID-19 patients and healthy controls. Merging mild and moderate classes [2] resulted in 172 healthy, 43 mild/moderate, and 54 severe cases. We used BDC1 and BDC2 panels, each with 29 channels, and applied channel-wise z-score normalization.

All datasets were evaluated using 5-fold cross-validation.

3.2 Cluster-based unsupervised sampling

Our pipeline has two parts. In part 1 (Fig. 1A), we pool N instance embeddings, each of dimension D, $\{\mathbf{x}_i\}_{i=1}^{N} \subset \mathbb{R}^D$, from all training patients, and fit a K-component Gaussian mixture model (GMM) using *scikit-learn (v1.5)*:

$$p(\mathbf{x}_i) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,, \tag{1}$$

where $\sum_{k=1}^{K} \pi_k = 1$ and $\boldsymbol{\Sigma}_k$ is diagonal. We maximize the data log-likelihood via Expectation–Maximization (EM).

For each real patient in class c, we count the instances assigned to each cluster k. Aggregating these yields the class–cluster mean $\mu_{c,k}$ and standard deviation $\sigma_{c,k}$, defining a class-specific *recipe*. To generate a patient of class c, we first sample a continuous cluster count $\tilde{n}_{c,k} \sim \mathcal{N}(\mu_{c,k}, \sigma^2_{c,k})$ and discretize it as $n_{c,k} = \max(0, \lceil \tilde{n}_{c,k} \rceil)$. We then draw $n_{c,k}$ embeddings from the pool of instances assigned to cluster k. The sampled embeddings form a new bag, generating new patients by remixing instances according to disease-specific statistics.

3.3 Uncertainty aware patient selection

Part 2 (Fig. 1B) of the pipeline selects the most informative generated patients to improve MIL classification performance and is independent of Part 1. After generating patients, we estimate the predictive uncertainty of the pretrained MIL model (f) on these bags using Monte Carlo (MC) dropout at inference, with dropout applied in the classifier head. For a generated patient bag $\mathbf{x}$ and T stochastic forward passes, we compute the logits

$$\mathbf{z}^{(t)}(\mathbf{x}) = f(\mathbf{x};\, \boldsymbol{\theta}^{(t)}) \in \mathbb{R}^C, \qquad \mathbf{p}^{(t)}(\mathbf{x}) = \mathrm{softmax}\Big(\mathbf{z}^{(t)}(\mathbf{x})\Big)\,, \tag{2}$$

and form the mean predictive distribution. We quantify uncertainty U with the predictive entropy of $\bar{\mathbf{p}}(\mathbf{x})$:

$$\bar{\mathbf{p}}(\mathbf{x}) \;=\; \frac{1}{T}\sum_{t=1}^{T} \mathbf{p}^{(t)}(\mathbf{x}), \qquad U(\mathbf{x}) \;=\; H[\bar{\mathbf{p}}(\mathbf{x})] \;=\; -\sum_{c=1}^{C} \bar{p}_c(\mathbf{x}) \, \log \bar{p}_c(\mathbf{x}). \tag{3}$$

Here, higher entropy indicates more uncertain predictions [15,3]. We also ablate BALD (Bayesian Active Learning by Disagreement) [5] and the standard deviation of the predicted class's maximum logit (Max-STD) across posterior samples [3] as alternative uncertainty metrics (see section 5.4).

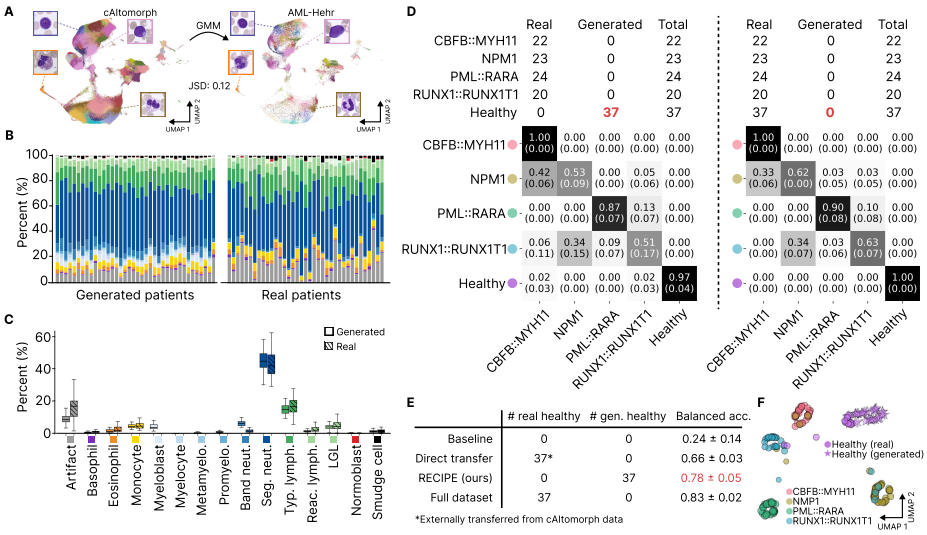

Fig. 2. Generating realistic patients using externally derived recipes. (A) We fit a Gaussian Mixture Model (GMM) on pooled instance embeddings from the cAItomorph data to obtain unsupervised clusters and class-specific *recipes*. We then transfer this clustering structure to AML-Hehr and generate healthy individuals by sampling only from the AML-Hehr embedding pool according to the externally learned healthy recipe. Transferring the GMM yields aligned clusters across datasets: real cell-label distributions within corresponding clusters agree well between cAItomorph and AML-Hehr (JSD= 0.12). **(B)** Per-patient cell-type composition of generated versus real AML-Hehr patients, computed using real cytomorphology labels. **(C)** Cell-type proportions across patients shows that generated patients recapitulate the real cell-composition statistics. **(D, E)** Training with generated patients achieves test-set performance comparable to the full-dataset setting. **(F)** Generated patients align well with real patients in the latent space, as visualized by UMAP.

4 Experiments

Previous patient-level augmentation approaches have mainly targeted already large cohorts (e.g., TCGA), where gains are often marginal [11,21]. In practice, medical datasets are frequently imbalanced, rare classes and healthy controls can be difficult to collect, and some diagnostic tests are invasive or too costly to apply broadly. We therefore evaluate our pipeline in three clinically relevant data-scarce scenarios:

(i) Generating patient recipes from an external dataset. We intentionally removed all healthy controls from the AML-Hehr training split and derived the healthy recipe from healthy controls in the cAItomorph dataset. Using this external recipe, we then generated healthy individuals by sampling only from the AML-Hehr diseased instance embedding pool. We generated the same number of healthy controls as in the original AML-Hehr dataset ($n = 37$). We could not

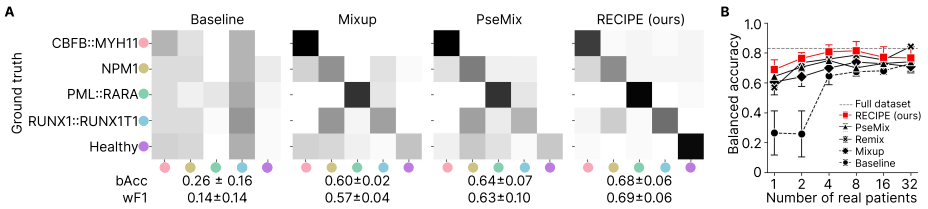

Fig. 3. Augmenting classes with few patients. (A) Performance comparison of augmentation methods when only one real healthy control is available in the training set. **(B)** RECIPE achieves the highest performance for almost all sample sizes.

include baseline augmentation comparisons in this setting because other existing methods assume that at least some real samples are available for every class.

(ii) Generating patients from few samples. We reduced the number of healthy controls in the AML-Hehr training set to $n \in 1, 2, 4, 8, 16, 32$ and generated additional healthy patients to match the original class size. In this setting, the recipe statistics were estimated from the few available healthy samples, while new bags were created by sampling instances from the AML-Hehr embedding pool. We compared our method with other augmentation methods.

(iii) Application in single cell omics/cytometry. We applied our approach to modalities where sample sizes are naturally limited, including single-cell RNA-seq and flow cytometry. In these experiments, we augmented the training set with generated patients corresponding to 30% of the original training data size (see 5.4) and compared with other augmentation methods.

In all experiments, the number of GMM clusters is fixed to $K = 50$ (see 5.4).

4.1 Multiple instance learning architecture and training

Since our pipeline is architecture-agnostic, we used attention-based multiple instance learning (ABMIL) [6]. We also show compatibility with other aggregators such as Transformer [20] and DSMIL [8] (see 5.4). ABMIL uses a 256-dimensional hidden representation with an MLP classifier. We trained ABMIL with batch size 16 and learning rate 1×10^{-5} using AdamW (weight decay $= 0.01$) for up to 150 epochs, with early stopping based on validation loss. Training used a single NVIDIA A100 80GB GPU.

5 Results

5.1 Generating patient recipes from an external dataset

We fit a GMM on cAItomorph and computed healthy-control statistics (*recipes*) from the resulting clusters, without using cell-type labels. We then applied the same GMM to AML-Hehr to assign clusters and generated healthy individuals using the precomputed recipe (Fig. 2A). The clustering structure transferred

Table 1. Balanced accuracy (bAcc) on non-image COVID-19 datasets. Values are mean ± std. Best result in each column is shown in bold.

	scRNA-seq		**Flow cytometry**	
Method	**PCA**	**scVI**	**BDC1**	**BDC2**
Baseline	0.56 ± 0.12	0.58 ± 0.15	0.50 ± 0.05	0.49 ± 0.05
Mixup	0.60 ± 0.14	0.65 ± 0.12	0.57 ± 0.04	0.58 ± 0.04
ReMix	0.54 ± 0.08	0.58 ± 0.10	0.56 ± 0.05	0.55 ± 0.05
PseMix	0.63 ± 0.14	0.62 ± 0.04	0.57 ± 0.05	0.56 ± 0.04
RECIPE (ours)	$\mathbf{0.64 \pm 0.05}$	$\mathbf{0.73 \pm 0.09}$	$\mathbf{0.61} \pm 0.07$	$\mathbf{0.60} \pm 0.03$

well: corresponding clusters in the two datasets showed similar real cell-type compositions, with a low Jensen–Shannon divergence (0.12). Using the original AML-Hehr cell labels for evaluation, the generated individuals closely matched the expected cellular composition of healthy peripheral blood (Fig. 2B,C), with segmented neutrophils comprising ∼60% and typical lymphocytes ∼20%. We generated 37 healthy controls (matching the original cohort size) and trained ABMIL using these cases. On a test set of real patients, training with generated healthy controls achieved performance close to the full dataset (Fig. 2D,E; balanced accuracy of 0.78 ± 0.05 vs. 0.83 ± 0.02) and outperformed directly transferring healthy cases from cAItomorph to AML-Hehr (0.66 ± 0.03).

5.2 Generating patients from few samples

Next, we evaluated our pipeline in a low-data setting where only a few healthy controls are available. Disease recipes were estimated from a limited number of healthy samples $n \in \{1, 2, 4, 8, 16, 32\}$, and healthy individuals were generated from the pooled AML-Hehr embedding set based on these recipes. For a single healthy individual ($n = 1$), the baseline achieved only 0.26 ± 0.16 balanced accuracy (Fig. 3A). MixUp [23] and PseMix improved performance to 0.60 ± 0.02 and 0.64 ± 0.07, respectively, while RECIPE achieved the best result (0.68 ± 0.06). RECIPE also improved sensitivity across classes, particularly for the healthy class (Fig. 3A). This trend was consistent across all sample sizes (Fig. 3B).

5.3 Application in single cell omics/cytometry

Finally, we evaluated RECIPE on real-world non-image modalities using scRNA-seq and flow cytometry (Table 1). On scRNA-seq, RECIPE achieved the best performance for both embedding types, improving balanced accuracy from 0.56 ± 0.12 to 0.64 ± 0.05 with PCA (+14%) and from 0.58 ± 0.15 to 0.73 ± 0.09 with scVI (+26%). On flow cytometry, RECIPE improved performance on both panels (+22%): for BDC1, balanced accuracy increased from 0.50 ± 0.05 to 0.61 ± 0.07, for BDC2, RECIPE increased balanced accuracy from 0.49 ± 0.05 to 0.60 ± 0.03.

Table 2. One-factor-at-a-time ablation study on Covid-scRNA-seq scVI embeddings (left) and RECIPE performance with various aggregators on Covid-scRNA-seq PCA embeddings (right). The first row denotes default settings; each subsequent row varies a single hyperparameter. Best balanced accuracy is shown in bold. Default: predictive entropy uncertainty, global most uncertain selection, 30% augmentation ratio, and $K = 50$ clusters

Hyperparameter	**bAcc**
Default	$\mathbf{0.73 \pm 0.09}$
Uncertainty: BALD	0.62 ± 0.23
Uncertainty: Max-STD	0.61 ± 0.24
Mixed most uncert.	0.72 ± 0.17
Per-class most uncert.	0.68 ± 0.19
Random certain	0.70 ± 0.16
Most certain	0.53 ± 0.12
Aug. ratio: 5%	0.68 ± 0.16
Aug. ratio: 10%	0.69 ± 0.20
Aug. ratio: 20%	0.70 ± 0.15
Aug. ratio: 60%	0.68 ± 0.21
Aug. ratio: 90%	0.70 ± 0.14
K: 3	0.61 ± 0.16
K: 10	0.60 ± 0.10
K: 30	0.52 ± 0.16
K: 100	0.48 ± 0.17

Architecture	**bAcc**
ABMIL (Baseline)	0.56 ± 0.12
ABMIL (RECIPE)	0.64 ± 0.05
Improvement	+15%
DSMIL (Baseline)	0.45 ± 0.11
DSMIL (RECIPE)	0.58 ± 0.17
Improvement	+29%
Transformer (Baseline)	0.49 ± 0.09
Transformer (RECIPE)	0.67 ± 0.11
Improvement	+38%

5.4 Ablation studies

We conducted a one-factor-at-a-time ablation study (Table 2) to evaluate the impact of key hyperparameters. Predictive entropy proved to be the most effective uncertainty measure, outperforming BALD and Max-STD by over 10%. We also examined the selection scope, which determines how uncertain patients are prioritized: *global* (top uncertain patients across the entire dataset), *per-class* (top uncertain patients within each class), or *mixed*. Global selection achieved the highest balanced accuracy (0.73 ± 0.09), suggesting that prioritizing the most uncertain samples regardless of class labels provides the most information. To demonstrate the benefit of selecting the most uncertain patients, we also ablated augmenting with the most certain patients or randomly selected patients. Selecting the most certain patients yielded the lowest performance, while selecting random patients resulted in intermediate performance. The augmentation ratio showed an optimal peak at 30%; exceeding this threshold did not improve more. For the number of clusters, $K = 50$ provided the best resolution for capturing semantic diversity. Finally, RECIPE is architecture-agnostic, providing significant performance gains—up to 38%—across different aggregators including ABMIL, DSMIL, and Transformer.

6 Conclusion

We propose RECIPE, a patient-level augmentation framework designed to overcome data scarcity in MIL. By deriving probabilistic class "recipes" via unsupervised clustering, our method generates realistic patients by remixing instance embeddings. A key innovation of RECIPE is its ability to generate entirely missing cohorts using externally derived statistics, demonstrating that disease phenotypes can be effectively transferred across datasets without sharing patient data. Furthermore, RECIPE proved highly effective in non-imaging domains, yielding performance gains in data-limited single-cell RNA-seq and flow cytometry tasks.

Acknowledgments. C.M. received funding from the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme (grant agreements 866411, 101113551, and 101213822) and support from the Hightech Agenda Bayern.

Disclosure of Interests. The authors have no competing interest.

References

1. Dasdelen, M.F., Kukuljan, I., Lienemann, P., Ozlugedik, F., Sadafi, A., Hehr, M., Spiekermann, K., Pohlkamp, C., Marr, C.: AI-based hematological malignancy prediction from peripheral blood smears in a large diagnostic laboratory cohort. Leukemia pp. 1–5 (2026). https://doi.org/10.1038/s41375-026-02934-1
2. Ding, Z., Baras, A.: Application and characterization of the multiple instance learning framework in flow cytometry. Scientific Reports (2026). https://doi.org/10.1038/s41598-025-32093-9
3. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: International conference on machine learning. pp. 1183–1192. PMLR (2017)
4. Hehr, M., Sadafi, A., Matek, C., Lienemann, P., Pohlkamp, C., Haferlach, T., Spiekermann, K., Marr, C.: Explainable ai identifies diagnostic cells of genetic aml subtypes. PLOS Digital Health **2**(3), e0000187 (2023). https://doi.org/10.1371/journal.pdig.0000187
5. Houlsby, N., Huszár, F., Ghahramani, Z., Lengyel, M.: Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745 (2011)
6. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2127–2136. PMLR (10–15 Jul 2018), https://proceedings.mlr.press/v80/ilse18a.html
7. Koch, V., Wagner, S.J., Kazeminia, S., Sancar, E., Hehr, M., Schnabel, J.A., Peng, T., Marr, C.: Dinobloom: a foundation model for generalizable cell embeddings in hematology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 520–530. Springer (2024). https://doi.org/10.1007/978-3-031-72390-2_49
8. 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)

9. Liechti, T., Iftikhar, Y., Mangino, M., Beddall, M., Goss, C.W., O'Halloran, J.A., Mudd, P.A., Roederer, M.: Immune phenotypes that are associated with subsequent covid-19 severity inferred from post-recovery samples. Nature communications **13**(1), 7255 (2022). https://doi.org/10.1038/s41467-022-34638-2
10. Litinetskaya, A., Shulman, M., Hediyeh-zadeh, S., Moinfar, A.A., Curion, F., Szałata, A., Omidi, A., Lotfollahi, M., Theis, F.J.: Multimodal weakly supervised learning to identify disease-specific changes in single-cell atlases. bioRxiv pp. 2024–07 (2024). https://doi.org/10.1101/2024.07.29.605625
11. Liu, P., Ji, L., Zhang, X., Ye, F.: Pseudo-bag mixup augmentation for multiple instance learning-based whole slide image classification. IEEE Transactions on Medical Imaging **43**(5), 1841–1852 (2024). https://doi.org/10.1109/TMI.2024.3351213
12. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., Yosef, N.: Deep generative modeling for single-cell transcriptomics. Nature methods **15**(12), 1053–1058 (2018). https://doi.org/10.1038/s41592-018-0229-2
13. 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). https://doi.org/10.1038/s41551-020-00682-w
14. Mumuni, A., Mumuni, F.: Data augmentation: A comprehensive survey of modern approaches. Array **16**, 100258 (2022). https://doi.org/10.1016/j.array.2022.100258
15. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
16. Simard, P., Steinkraus, D., Platt, J.: Best practices for convolutional neural networks applied to visual document analysis. In: Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings. pp. 958–963 (2003). https://doi.org/10.1109/ICDAR.2003.1227801
17. Song, A.H., Jaume, G., Williamson, D.F.K., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering **1**(12), 930–949 (Dec 2023). https://doi.org/10.1038/s44222-023-00096-8
18. Stephenson, E., Reynolds, G., Botting, R.A., Calero-Nieto, F.J., Morgan, M.D., Tuong, Z.K., Bach, K., Sungnak, W., Worlock, K.B., Yoshida, M., et al.: Single-cell multi-omics analysis of the immune response in covid-19. Nature medicine **27**(5), 904–916 (2021). https://doi.org/10.1038/s41591-021-01329-2
19. Vapnik, V.N.: The Nature of Statistical Learning Theory. Information Science and Statistics, Springer, New York, NY, 2 edn. (2000)
20. 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). https://doi.org/10.1016/j.ccell.2023.08.002
21. Yang, J., Chen, H., Zhao, Y., Yang, F., Zhang, Y., He, L., Yao, J.: Remix: A general and efficient framework for multiple instance learning based whole slide image classification (2022). https://doi.org/10.1007/978-3-031-16434-7_4
22. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. CoRR **abs/1611.03530** (2016), http://arxiv.org/abs/1611.03530
23. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=r1Ddp1-Rb