Self-Supervised Multiple Instance Learning
for Acute Myeloid Leukemia Classification

Salome Kazeminia*
TUM
Helmholtz Munich

salom.kazeminia@helmholtz-muenchen.de

Dragan Bosnacki
Eindhoven University of Technology

d.bosnacki@tue.nl

Max Joosten*
Eindhoven University of Technology
Helmholtz Munich

m.p.h.joosten@student.tue.nl

Carsten Marr
Helmholtz Munich

carsten.marr@helmholtz-muenchen.de

Abstract

Automated disease diagnosis using medical image analysis relies on deep learning, often requiring large labeled datasets for supervised model training. Diseases like Acute Myeloid Leukemia (AML) pose challenges due to scarce and costly annotations on a single-cell level. Multiple Instance Learning (MIL) addresses weakly labeled scenarios but necessitates powerful encoders typically trained with labeled data. In this study, we explore Self-Supervised Learning (SSL) as a pre-training approach for MIL-based AML subtype classification from blood smears, removing the need for labeled data during encoder training. We investigate the three state-of-the-art SSL methods SimCLR, SwAV, and DINO, and compare their performance against supervised pre-training. Our findings show that SSL-pretrained encoders achieve comparable performance, showcasing the potential of SSL in MIL. This breakthrough offers a cost-effective and data-efficient solution, propelling the field of AI-based disease diagnosis.

1. Introduction

The precise classification of Acute Myeloid Leukemia (AML) genetic subtypes is imperative for selecting appropriate treatment modalities and improving patient outcomes. Achieving such accurate classification entails the utilization of genetic testing to discern distinct AML genetic subtypes [21], albeit with inherent cost and complexity. In this context, blood smear microscopy emerges as a compelling alternative method, entailing a scrupulous microscopic examination of stained blood samples to reveal elusive malignant cells [26]. However, the intrinsic rarity of these malignant cells presents a formidable challenge for clinicians undertaking this diagnostic endeavor.

Automated disease diagnosis through image-based approaches has garnered remarkable attention in research and clinical spheres [18, 17, 24, 8, 8]. However, in the context of AML diagnosis, a particular challenge arises due to the availability of weakly labeled data. In this scenario, each diagnosis relies on patient-specific leukocyte images, forming part of an extensive image set that also encompass non-relevant cells. This weak data labeling poses an intricate task for achieving accurate classification, thereby necessitating innovative methodologies to surmount the constraints imposed by the limited availability of precisely labeled information. To surmount this hurdle, Multiple Instance Learning (MIL) emerges as a compelling, weakly supervised method. Hehr et al. elucidates the remarkable performance of MIL by emphasizing class-specific leukocyte images detected through an attention mechanism [11]. The triumph of MIL, however, hinges significantly upon the caliber and efficacy of the representations acquired by its encoder model. When confronted with insufficient training data to effectively train the encoder, pre-training emerges as a viable alternative [11]. However, accessing fully-supervised datasets can be challenging, especially in medical environments where data privacy and scarcity concerns often arise.

Self-Supervised Learning (SSL) has proved to be a promising pre-training method for Multiple Instance Learning (MIL) classifiers, particularly in data-limited scenarios like medical environments where obtaining extensive labeled datasets is challenging [14, 16]. Unlike supervised pre-training methods that rely on costly annotations, SSL techniques enable the encoder to learn informative representations from abundant unlabeled data without explicit annotations. By leveraging the inherent structure and relationships within the data, SSL empowers the encoder to
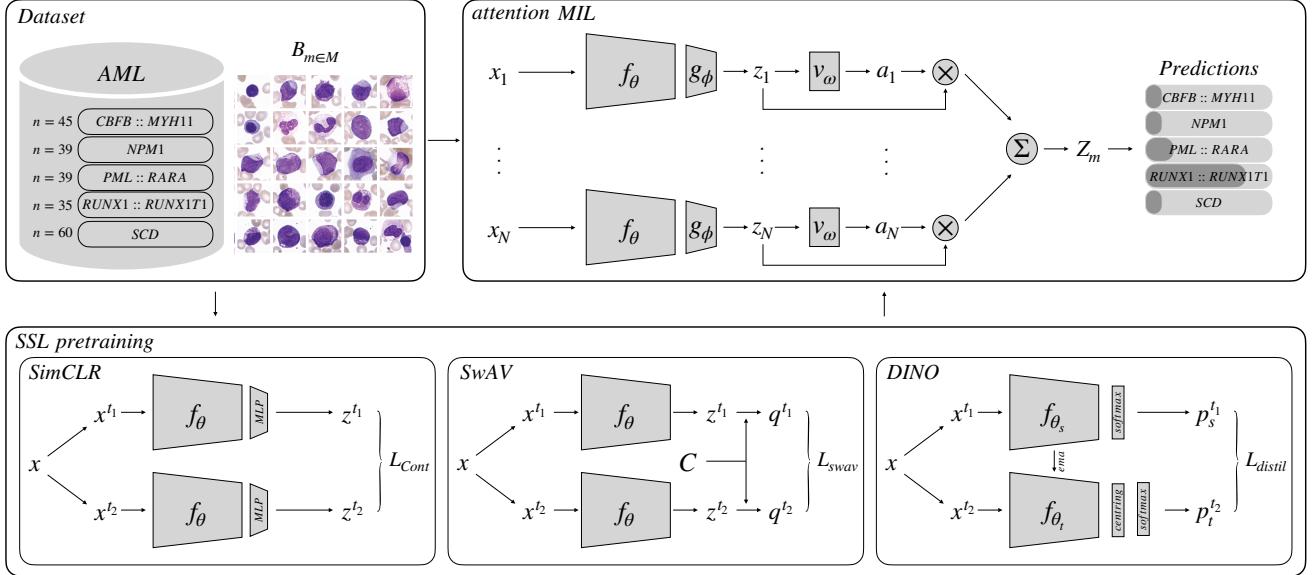
Figure 1: Overview of our data-efficient MIL model for AML genetic subtypes classification. We pre-train an MIL encoder f_θ with one of the state-of-the-art SSL models, i.e., SimCLR, SwAV, and DINO. Then we embed the trained encoder in the attention MIL architecture. For both SSL and MIL the same training dataset is used.

capture useful domain-specific features, enhancing the performance and generalizability of MIL models.

In light of annotated data limitation, we turned to Self-Supervised Learning (SSL) techniques as a compelling solution, enabling the encoder to learn meaningful representations from vast amounts of unlabeled cells in without needing explicit annotations. This strategic shift towards SSL not only mitigates the reliance on fully-labeled datasets but also empowers our model to capture more robust and domain-adaptive features of MIL frameworks [13, 6, 4].

In this work we investigate three SSLmethods that utilize distinct strategies to address classification of AML genetic subtypes within the framework of MIL. Specifically, we conducted experiments using SimCLR, SwAV, and DINO, which represent state-of-the-art performance in contrastive-instance, clustering-based, and self-distillation-based SSL approaches, respectively, and have demonstrated promising performance on widely recognized computer vision benchmark datasets, such as ImageNet [22]. By benchmarking these SSL techniques, we aim to assess their performance and suitability for AML subtype classification in the context of MIL (see Figure 1). This validation highlights the efficacy and promise of leveraging SSL techniques for accurate and data-efficient AML subtype classification within the MIL framework without tedious single-cell annotation tasks.

2. Related Work

2.1. Self-supervised learning models

The primary objective of the SSL method is to acquire meaningful representations from unlabeled data, thereby circumventing the need for explicit annotations. Several SSL models have been introduced, each characterized by ingenious methodologies aimed at harnessing the wealth of information present in large-scale, unlabeled datasets. However, mode collapse is a critical challenge in SSL methods, where learned features collapse into a single point or limited space, hindering model generalization and downstream task performance [12]. Addressing mode collapse is crucial to ensure the effectiveness and robustness of SSL approaches.

Simple framework for contrastive learning of visual representations (SimCLR) [5] employs a contrastive learning framework to learn meaningful and transferable image representations. For each image $x_i \in X$, diverse augmentations $t \in T$ are applied, resulting in augmented views $x_i^{t_1}$ and $x_i^{t_2}$ and an encoder function followed by a projection head maps them to latent representations $z_i^{t_1}, z_i^{t_2}$. The contrastive loss function used in SimCLR aims to maximize the cosine similarity (Sim) between the projection feature vectors of positive pairs (augmented views of the same image), while minimizing the similarity between the projection feature vectors of negative pairs (augmented views of different images, indicated with $\mathbb{1}_{[k \neq i]}$, which

equals 1 if $k \neq i$ and 0 otherwise). Mathematically, the contrastive loss function can be defined as follows:

$$L_{Cont} = -\log \frac{\exp(Sim(z_i^{t_1}, z_i^{t_2})/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(Sim(z_i^{t_1}, z_k^{(t)})/\tau)}. \quad (1)$$

In this equation, N represents the number of instances in the mini-batch, $z_i^{t_1}, z_i^{t_2}$ are the projection feature vectors of the positive pairs (augmented views of the same image), $Sim(z_i^{t_1}, z_i^{t_2})$ denotes the cosine similarity between these vectors and τ is the temperature parameter, which scales the similarity values. The contrastive loss encourages the projection feature vectors of positive pairs to have high cosine similarity, making them closer in the embedding space, while the similarity between the projection feature vectors of negative pairs is minimized, pushing them further apart. This way, the model learns to create compact clusters for instances belonging to the same image (positive pairs) and maximize the separation between instances from different images (negative pairs), leading to more meaningful and discriminative feature representations.

Swapping Assignments between Views (SwAV) [2] is a clustering-based SSL approach, where visual features are learned by contrasting the cluster assignments of different augmentations $t \in T$ applied to the same image. The encoder network maps these augmented views into encoded feature vectors z^{t_1} and z^{t_2}, respectively. Prototype vectors representing cluster centroids are maintained as c_j for each cluster $j \in J$, where J is the set of all clusters. The objective loss function is formulated as follows

$$L_{swav}(z^{t_1}, z^{t_2}) = l(z^{t_1}, q^{t_2}) + l(z^{t_2}, q^{t_1}) \quad (2)$$

where q^{t_1} and q^{t_2} are abstract information of latent representations z^{t_1} and z^{t_2} correspondingly. The loss function $l(.)$ measures the fit between representations of two augmentation of image as

$$l(z^{t_1}, q^{t_2}) = -\sum_{i \in J} q_{(i)}^{t_2} \log p_{(i)}^{t_1}, \quad (3)$$

where

$$p_{(i)}^{t_1} = \frac{\exp(z^{t \in T} c_i/\tau)}{\sum_{j \in J} \exp(z^{t \in T} c_j/\tau)}. \quad (4)$$

To obtain cluster assignments for the encoded feature vectors, a softmax function with a temperature parameter τ is applied, resulting in cluster assignment probabilities $p_{(k)}^{t_1}$ and $p_{(k)}^{t_2}$. SwAV then swaps these assignments to form new assignment pairs, represented as $(p_{ij}^{t_1}, p_{ik}^{t_2})$ and $(p_{ij}^{t_1}, p_{ik}^{t_1})$, where $k \in J$ represents a different cluster. The prototype vectors c_j are updated using these swapped assignment pairs, aiming to maintain balanced assignments and ensure each feature vector is associated with a single prototype with the highest probability. This fosters a diverse and well-balanced representation of clusters, mitigating the risk of collapsing into a few dominant clusters. SwAV's approach enables the learning of robust visual features without the need for explicit data labeling, improving model generalization performance.

Self-distillation with no labels (DINO) [3] leverages the concept of self-distillation, where a teacher network f_{θ_t} guides the learning of a student network f_{θ_s} in a self-supervised manner. The main idea is to train the student network to mimic the output of the teacher network, which helps the student network learn useful representations without the need for explicit labels. During training, different augmentations of an image $x \in X$ pass through student and teacher networks and a distillation loss encourages the student to approximate the teacher's predictions as

$$\mathcal{L}_{distill} = -\sum_{i=1}^{K} p_t^i \log(p_s^i), \quad (5)$$

where $p^i(x)$ represents the probability of image x falling into soft-class i. These probabilities are calculated as:

$$p_{s \vee t}^i(x) = \frac{\exp(f^i(x^t; \theta_{s \vee t}/\tau)}{\sum_{k=1}^{K} \exp(f^k(x^t; \theta_{s \vee t})/\tau)}, \quad (6)$$

where K is the number of soft-classes referring to a continuous and smooth representation of the distribution of each data and its augmentations. Backpropagation is applied only to the student network, while the teacher network is updated using the exponential moving average of the student's parameters, allowing knowledge transfer from student to teacher. To prevent mode collapse and improve the learning process, DINO introduces two additional techniques: centering and sharpening. Centering is applied to the teacher network and involves subtracting the average of projection vectors from individual vectors to balance predictions. Sharpening uses a lower temperature (τ) in the softmax layer for the teacher network, making the prediction task more challenging for the student network.

3. Overview of the approach

Our framework, similar to the approach proposed in [11], is specifically designed for analyzing microscopic images within a Multiple Instance Learning (MIL) setting.

In this context, bags denoted as $B_1, ..., B_M$ represent sets of blood sample images containing white blood cells, where each individual cell is represented as $x_1, ..., x_N \in B_{m \in M}$. For each bag, we have one expert-annotated label denoted as y_m, indicating the specific subtype of Acute Myeloid Leukemia (AML) associated with the corresponding blood sample.

Notably, the framework is customized to meet the demands of this particular domain, where cells are distributed

along independent spatial locations within the image. This is different from cases where tiling techniques are employed to capture cells [23].

3.1. Feature extraction

An encoder is utilized to extract class-related features from individual cells, which are then mapped to k-dimensional feature vectors denoted as $z_1^k, ..., z_N^k$. The encoder comprises a pre-trained ResNet 18 [10] with frozen weights. For pre-training, the encoder is incorporated into a Self-Supervised Learning (SSL) architecture, where the training data is passed through the network without any labels, enabling the encoder to learn to map similar cells and their augmentations close to each other in the feature space. Additionally, a four-convolution layer network $g(.)$ with learnable parameters ϕ is introduced and trained with the Multiple Instance Learning (MIL) loss. This network takes the feature vectors z_n^k and outputs $z_n^{k'}$ in a lower-dimensional space k' with $k' < k$. The combination of the encoder and the convolutional network allows for effective feature extraction and dimension reduction, facilitating accurate and efficient classification within the MIL framework.

3.2. Attention pooling

A multi-head deep attention network, denoted as v_ω, is equipped with two linear layers to estimate the attention scores a_n^c, for each cell's feature vector $z_n^{k'}$ with respect to each class $c \in C$, where C denotes the set of all classes. The attention scores are computed as

$$a_n^c = v_\omega(z_n^{k'}), \tag{7}$$

Subsequently, the final representation of the bag B_m is calculated as follows:

$$Z_m^c = \sum_{n=1}^{N} a_n^c \cdot z_n^{k'}, \tag{8}$$

where Z_m^c denotes the representation of bag B_m for class c, obtained by combining the attention scores a_n^c with the lower-dimensional feature vectors $z_n^{k'}$.

This representation Z_m^c is then utilized for the final multi-class prediction, and the objective function L_{MIL} is defined as:

$$L_{MIL} = CE(f_{MIL}(Z_m^{k'}; \phi, \omega), y_m), \tag{9}$$

where CE represents the categorical cross-entropy loss and $f(.)$ denotes the classifier head function. The overall MIL framework leverages this mechanism to predict the AML subtype for each blood sample, utilizing both the attention scores and the lower-dimensional feature vectors to make accurate and explainable predictions. This formulation enables the identification of the most contributing cells based on the class-specific attention scores, allowing for cell-level explanation and providing valuable insights into the model's decision-making process, which is particularly crucial for medical applications.

4. Experiments

4.1. Dataset

The AML dataset utilized in this study is publicly available, introduced by Hehr et al. in their work on explainable AI for AML genetic subtype classification [11] (AML dataset). The dataset comprises 242 blood smears from the Munich Leukemia Laboratory (MLL), which were associated with four distinct AML genetic subtypes: APL with "$PML :: RARA$ fusion" ($n = 45$), AML with "$NPM1$" mutation ($n = 39$), AML with "$CBFB :: MYH11$" fusion (without NPM1 mutation) ($n = 39$), and AML with "$RUNX1 :: RUNX1T1$ fusion" ($n = 35$). Additionally, healthy stem cell donors (SCD) were included as controls ($n = 60$). The ground truth labels of AML genetic subtypes was determined through genetic testing.

Each blood smear was stained and scanned, resulting in a total of 218 patients and $101,947$ single-cell images with a resolution of 144×144 per blood smear, corresponding to a size of $24.9 \mu m \times 24.9 \mu m$. Furthermore, $1,983$ single-cell images were expertly annotated by a cytologist, allowing for experiments using single-cell labels. This comprehensive dataset facilitated the evaluation of our MIL-based classification framework for accurate and interpretable AML subtype classification.

4.2. Training

In our experiments, we conduct 5-fold stratified cross-validation for both the SSL encoder pre-training and the MIL head training. The dataset is divided into 60% for training, 20% for validation, and 20% for testing. To account for the uncertainty introduced by the MIL training procedure, we repeated the training of the MIL head three times with different random seeds.

For the supervised encoder, we adopt a ResNet 34 model pre-trained on an external dataset with $300,000$ annotated single-cell images comprising 23 different classes. The pre-training process in [11] involved random rotation, rescaling, translation, random erasing, and horizontal/vertical flipping, while oversampling was utilized to handle class imbalance. The model was trained for 50 epochs, and the stochastic gradient descent (SGD) optimizer with a learning rate of 5×10^{-4} was employed.

The three SSL methods, SimCLR, SwAV, and DINO, are fine-tuned with specific hyperparameters for encoder pretraining. For **SimCLR**, we use SGD with momentum, learning rate scaling, and cosine annealing during training, along with Large Batch Training of Convolutional Net-

Figure 2: The confusion matrix presents the test fold results of the first run for each pre-training method. While different SSL pre-training methods lead to varying class-wise performance, overall SSL pre-training performs correspondingly to fully supervised pre-training

works (LARC) [25] to improve model convergence. The contrastive loss function is employed with a temperature of 0.1, and a two-layer projection head with dimensions $[512, 512]$ and $[512, 64]$. Random augmentations such as rotation, resize cropping, flipping, color jittering, greyscale, and Gaussian blur are applied to the input images.

Similarly, for **SwAV**, we utilize SGD with momentum, learning rate scaling, and cosine annealing, trained for 500 epochs with a batch size of 512, and LARC was used for convergence enhancement. The projection head comprised a 3-layer MLP with specific dimensions and 300 prototype vectors were updated with the Sinkhorn-Knopp algorithm [7]. Augmentations included multi-crops with global and local views, vertical/horizontal flipping, Colorjitter, and Gaussian blurring.

For **DINO**, we used the AdamW [15] optimizer with momentum and a weight decay scheduler, batch size of 512, and LARC for convergence. Learning rate scaling and cosine annealing is applied, and the teacher model is warmed up with a teacher temperature of 0.04, increasing to 0.07. The momentum of the momentum encoder is adjusted during training. Similar augmentations as SwAV is applied, with 2 global and 8 local crops generated from a single image.

The **MIL** model is fine-tuned for 50 epochs or until the validation loss plateau for 20 epochs. We utilize the SGD optimizer with Nesterov momentum [20] and set the learning rate to 0.015. To ensure stable training in later epochs, the learning rate is gradually reduced using cosine annealing. Back-propagation is performed only after gradients are accumulated for 10 batches. The batch size is determined based on the number of instances associated with each patient, with a maximum of 500 instances per batch. During training, we apply random vertical and horizontal flipping as augmentations to the instances in each bag. To address the class imbalance between the AML subtypes, we employ a higher sampling rate for bags containing the minority class. This approach results in under-sampling of majority class bags and over-sampling of minority class bags, effectively balancing the class distribution and enhancing the model's performance in handling imbalanced data.

We implement both the SSL pre-training and the classifier in PyTorch, utilizing the VISSL self-supervised learning framework [9] for pre-training. All experiments are conducted on a single desktop-class NVIDIA RTX 3090 GPU. To ensure robustness and reliability, we run each pre-trained model three times, and with 5-fold cross-validation, resulting in 15 sets of predictions. The performance metrics are reported as the mean and standard deviation of these prediction sets.

To reduce CPU overhead during training, the augmentation pipelines are implemented using albumentations [1]. Additionally, a combination of mixed precision training and activation checkpointing techniques is employed to accelerate the training process and reduce the memory footprint, ensuring efficient utilization of computational resources.

4.3. Quantitative results

The MIL classifier, when trained on an SSL pre-trained encoder, exhibits comparable performance to the classifier trained on a supervised encoder in classifying AML subtypes (see Figure 2). The macro-averaged classification metrics, as shown in Table 1, are very similar across the pre-training methods. For instance, the classifier trained on the supervised encoder achieves high classification performance (F1-macro 0.86 ± 0.09), and this level of performance is matched by the SimCLR encoder (F1-macro 0.85 ± 0.04).
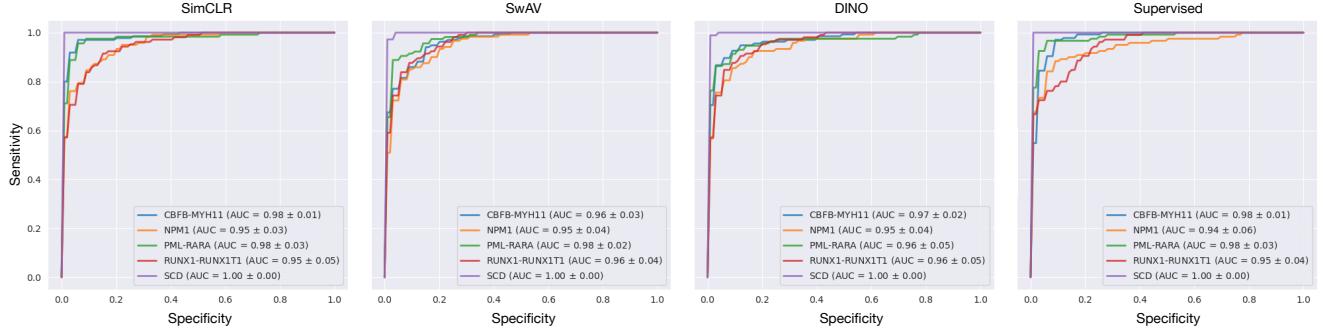
Figure 3: Similar mean ROC curves for MIL model resulting from different pre-training methods. Differences in the AUC score for genetic subtype prediction is within the margin of error.
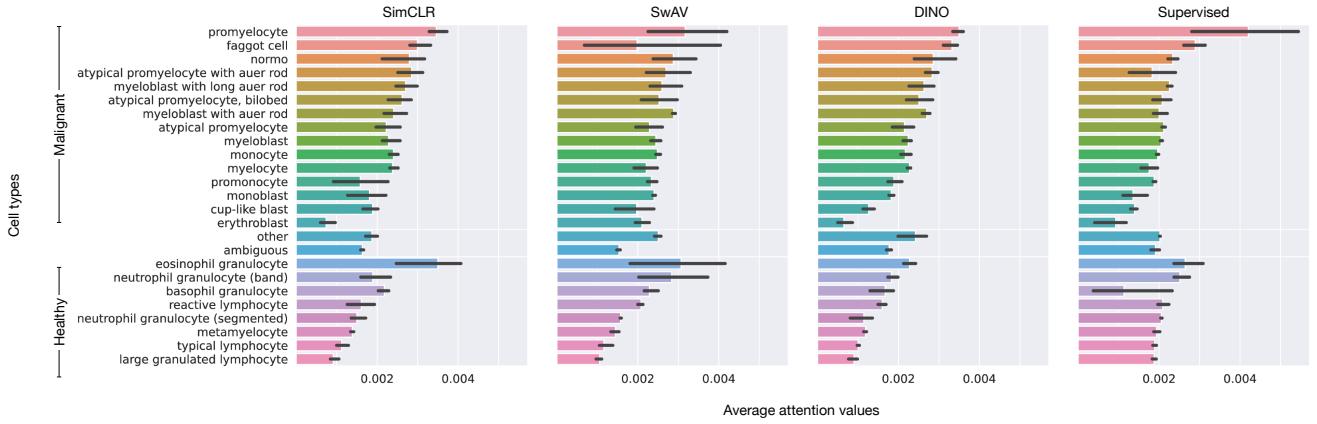

Figure 4: Averaged attention values for all labeled instances per model. Most of cell types with higher attention are malignant, while most lower attention cell types a healthy. SimCLR and DINO pretrained models provide more robust attention scores for malignant cells compared to the supervised pre-training.

Pre-training	F1	ROC AUC	PR AUC
DINO	0.85 ± 0.07	0.97 ± 0.03	0.91 ± 0.08
SimCLR	0.85 ± 0.04	$\mathbf{0.97 \pm 0.02}$	$\mathbf{0.92 \pm 0.05}$
SwAV	0.82 ± 0.06	$\mathbf{0.97 \pm 0.02}$	0.91 ± 0.05
Supervised	$\mathbf{0.86 \pm 0.09}$	0.97 ± 0.03	0.92 ± 0.07

Table 1: Averaged classification performance of MIL pre-trained with different SSL methods vs. supervised learning methods. Results are averaged over 3 runs over 5-fold cross validation and reported in mean ± s.d. format.

$RUNX1T1$ fusion compared to the other pre-training methods. The average area under the Receiver Operator Characteristic curve (ROC AUC) and the average area under the Precision-Recall curve (PR AUC) (Figure 3) also align closely with the results from the SimCLR model. Detailed classification results for all combinations of models and classes can be found in Appendix A.

The attention module integrated into the MIL classifier demonstrates its efficacy in focusing on malignant cells. These cells consistently receive higher attention values, enhancing the model's interpretability. By ranking the input images based on the attention values, the model highlights cells crucial for predictions, even in the absence of per-cell labels. Examining the attention patterns of classifiers trained with different pre-trained SSL encoders in Figure 4 reveals that malignant cells consistently obtain higher attention values than healthy cells. Notably, the classifier trained with the SimCLR encoder exhibits a larger gap in

Notably, different pre-training methods exhibit varying classification performances across different classes. For instance, SimCLR demonstrates superior performance in predicting $CBFB :: MYH11$ fusion, while it performs slightly worse in the classification of $RUNX1 ::$
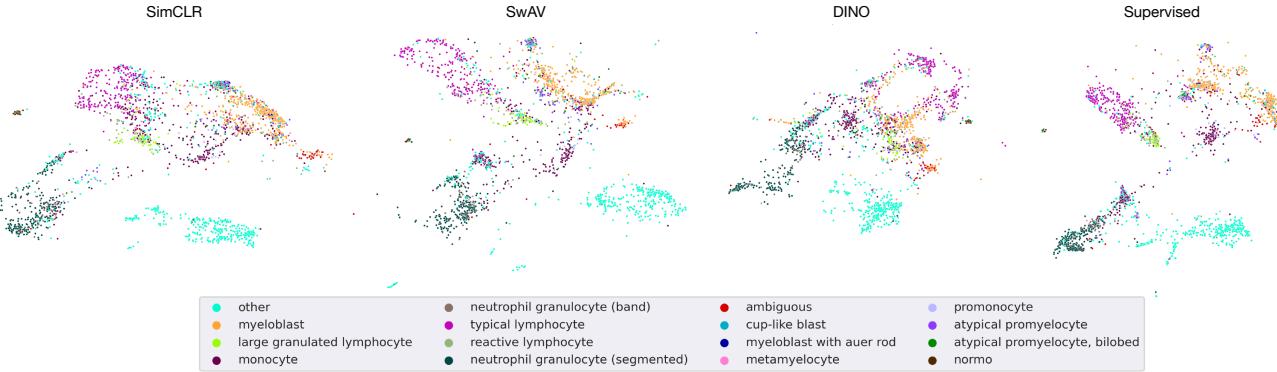
Figure 5: UMAP embedding visualization of encoder features with expert cytologist-assigned cell labels. SSL pre-trained encoder could distinguish cell-types without any label similar to fully supervised pre-trained encoder.

average attention between malignant and healthy cells compared to the classifier trained with the supervised encoder. This observation suggests that pre-trained encoders produce features that are more easily ranked by the classifier, resulting in more accurate ranking of malignant cells. Consequently, the attention mechanism proves to be a valuable asset in identifying crucial cells for AML subtype classification, contributing to the model's improved performance and interpretability.

4.4. Qualitative results

In this work, we delve into the impact of various Self-Supervised Learning (SSL) methods on the encoder's latent space, which plays a crucial role in the Acute Myeloid Leukemia (AML) classification task. To assess the quality of the latent space, we employ UMAP [19], a non-linear dimensionality reduction technique that preserves both local and global data structures. The high-dimensional features extracted from the first cross-validation fold for each encoder (512 dimensions) are projected into a low-dimensional map (2 dimensions) for visualization. We focus on a subset of the data with expert cytologist annotations to gain insights into the latent space's arrangement and structure (see Figure 5). The visualization, based on a subset of expert-annotated cells, demonstrates that all SSL pre-training methods achieve good class separation without labeled data. Interestingly, the ambiguous cells are clearly separated into their own cluster by SSL pre-training, while the supervised encoder tends to associate them more with the myeloblast cluster. This indicates that SSL pre-training effectively generates distinctive features, even for challenging cases, enhancing the encoder's ability to detect such edge cases.

5. Conclusion

We explored various self-supervised learning (SSL) methods for encoder pre-training in the context of acute myeloid leukemia (AML) subtype classification. We compared three state-of-the-art SSL methods: SimCLR, SwAV, and DINO, to a fully-supervised encoder. Our results showed that SSL-based pre-training can yield useful features for AML subclass classification, achieving comparable performance to a fully-supervised approach while using a smaller, unlabeled dataset. SimCLR demonstrated the best performance, likely due to its effective contrastive learning framework. On the other hand, SwAV's clustering-based approach showed less favorable results, possibly due to the difficulty of distinguishing subtle differences in cytology images. Overall, SSL methods have the potential to leverage unlabeled data in medical environments, opening up new possibilities for efficient and effective image classification tasks.

6. ackknowledgement

References

[1] Alexander Buslaev, Vladimir I Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A Kalinin. Albumentations: fast and flexible image augmentations. *Information*, 11(2):125, 2020.

[2] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Ad-*

vances in neural information processing systems, 33:9912–9924, 2020.

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[4] 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*, pages 16144–16155, 2022.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *37th International Conference on Machine Learning, ICML 2020*, PartF168147-3:1575–1585, 2 2020.

[6] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022.

[7] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[8] Jan-Niklas Eckardt, Jan Moritz Middeke, Sebastian Riechert, Tim Schmittmann, Anas Shekh Sulaiman, Michael Kramer, Katja Sockel, Frank Kroschinsky, Ulrich Schuler, Johannes Schetelig, et al. Deep learning detects acute myeloid leukemia and predicts npm1 mutation status from bone marrow smears. *Leukemia*, 36(1):111–118, 2022.

[9] Priya Goyal, Quentin Duval, Jeremy Reizenstein, Matthew Leavitt, Min Xu, Benjamin Lefaudeux, Mannat Singh, Vinicius Reis, Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Ishan Misra. Vissl. https://github.com/facebookresearch/vissl, 2021.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Matthias Hehr, Ario Sadafi, Christian Matek, Peter Lienemann, Christian Pohlkamp, Torsten Haferlach, Karsten Spiekermann, and Carsten Marr. Explainable ai identifies diagnostic cells of genetic aml subtypes. *PLOS Digital Health*, 2(3):e0000187, 2023.

[12] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

[13] Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol. Self-supervised learning in medicine and healthcare. *Nature Biomedical Engineering*, 6(12):1346–1352, 2022.

[14] 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.

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[16] Ming Y. Lu, Bowen Chen, Andrew Zhang, Drew F. K. Williamson, Richard J. Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19764–19775, June 2023.

[17] Christian Matek, Sebastian Krappe, Christian Münzenmayer, Torsten Haferlach, and Carsten Marr. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood, The Journal of the American Society of Hematology*, 138(20):1917–1927, 2021.

[18] Christian Matek, Simone Schwarz, Karsten Spiekermann, and Carsten Marr. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nature Machine Intelligence*, 1(11):538–544, 2019.

[19] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[20] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[21] Elli Papaemmanuil, Moritz Gerstung, Lars Bullinger, Verena I. Gaidzik, Peter Paschka, Nicola D. Roberts, Nicola E. Potter, Michael Heuser, Felicitas Thol, Niccolo Bolli, Gunes Gundem, Peter Van Loo, Inigo Martincorena, Peter Ganly, Laura Mudie, Stuart McLaren, Sarah O'Meara, Keiran Raine, David R. Jones, Jon W. Teague, Adam P. Butler, Mel F. Greaves, Arnold Ganser, Konstanze Döhner, Richard F. Schlenk, Hartmut Döhner, and Peter J. Campbell. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *The New England journal of medicine*, 374(23):2209, 6 2016.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

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

[24] John-William Sidhom, Ingharan J Siddarthan, Bo-Shiun Lai, Adam Luo, Bryan C Hambley, Jennifer Bynum, Amy S Duffield, Michael B Streiff, Alison R Moliterno, Philip Imus, et al. Deep learning for diagnosis of acute promyelocytic leukemia via recognition of genomically imprinted morphologic features. *NPJ precision oncology*, 5(1):38, 2021.

[25] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[26] Jifeng Yu, Peter Y.Z. Jiang, Hao Sun, Xia Zhang, Zhongxing Jiang, Yingmei Li, and Yongping Song. Advances in targeted therapy for acute myeloid leukemia. *Biomarker Research*, 8(1):1–11, 5 2020.