Beyond Labels: Visual Representations for Bone Marrow Cell Morphology Recognition

Shayan Fazeli
Computer Science Department
University of California, Los Angeles
Los Angeles, USA
shayan@cs.ucla.edu

Alireza Samiei
Department of Pathology and Laboratory Medicine
University of California, Los Angeles
Los Angeles, USA
alireza_samiei@icloud.com

Thomas D. Lee
Department of Pathology and Laboratory Medicine
University of California, Los Angeles
Los Angeles, USA
tdlee@mednet.ucla.edu

Majid Sarrafzadeh
Computer Science Department
University of California, Los Angeles
Los Angeles, USA
majid@cs.ucla.edu

Abstract—Analyzing and inspecting bone marrow cell cytomorphology is a critical but highly complex and time-consuming component of hematopathology diagnosis. Recent advancements in artificial intelligence have paved the way for the application of deep learning algorithms to complex medical tasks. Nevertheless, there are many challenges in applying effective learning algorithms to medical image analysis, such as the lack of sufficient and reliably annotated training datasets and the highly class-imbalanced nature of most medical data. Here, we improve on the state-of-the-art methodologies of bone marrow cell recognition by deviating from sole reliance on labeled data and leveraging self-supervision in training our learning models. We investigate our approach’s effectiveness in identifying bone marrow cell types. Our experiments demonstrate significant performance improvements in conducting different bone marrow cell recognition tasks compared to the current state-of-the-art methodologies.

Index Terms—machine learning, eHealth, medical imaging, computational pathology, pathology, computer vision.

I. INTRODUCTION

Bone marrow biopsy is a crucial component of hematopathology. It is used to diagnose neoplastic and non-neoplastic blood disorders. As part of the biopsy, an aspirate slide is created by smearing the bone marrow spicules on a glass slide to spread the cells and allow an expert hematopathologist to perform a morphological examination of individual cells. Specifically, a hematopathologist assesses the cells for possible morphological aberrancies and generates a differential cell count of various cell types. According to the International Council for Standardization in Hematology (ICSH) and the World Health Organization (WHO), the differential cell count is a required component of the diagnostic criteria for many neoplastic and non-neoplastic disorders, such as acute leukemias, myelodysplastic syndromes, and plasma cell dyscrasias [1], [2].

The ICSH recommends that when a precise percentage of an abnormal cell type is required for diagnosis, at least 500 bone marrow cells are to be counted in a minimum of two bone marrow smears to generate a representative percentage of individual cell lines [1]. Nonetheless, generating differential cell counts on individual bone marrow biopsies is a time-consuming task for busy hematopathology services. Given the prevalence and necessity of bone-marrow biopsy, which is carried out over half a million times in the United States every year, proposing automated methodologies can significantly reduce the costs associated with the laborious and error-prone process of manually reviewing the images [3].

Recent machine learning advances in the computer vision domain have allowed researchers to apply deep neural inference pipelines to automate medical image analysis. Nevertheless, there are major challenges associated with applying these pipelines in the medical image analysis field.

First, effective training of deep neural networks often requires access to large, representative datasets with millions of images. However, healthcare datasets, which are commonly generated at large academic hospitals, reflect the patient volume and population-specific to that institution. These datasets are small, in the order of thousands, and are biased towards a specific population of patients who are referred to tertiary academic centers.

Second, expert-annotated datasets are scarce in healthcare. Annotating medical images is labor-intensive, time-consuming, and in some cases, noisy due to inter-annotator disagreement. Many medical images which are prone to inter-annotator variabilities, such as HER2 score staining in pathol-
We propose a novel solution to the bone-marrow cell cytomorphology classification task by leveraging the recent advancements in self-supervised learning. We have investigated the effectiveness of invariance-centered self-supervision techniques in polishing the learned representations in the latent space. We provide empirical findings emphasizing the importance of our solution in generating separable and interpretable features. This is shown via training a shallow linear classifier leveraging fixed versions of our latent representations and evaluating the resulting network on a bone marrow cell cytomorphology benchmark.

We investigate the impact of leveraging both supervised signal from the expert annotations as well as the contrast and invariance regularization by optimizing Supervised Contrastive loss.

Based on our findings, we propose that increased reliance on self-supervised learning has the potential to significantly improve performance in medical image classification domains, such as bone marrow cell classification.

We demonstrate that self-supervised learning increases prediction fairness by rendering visual representations more robust to the negative impacts of long-tail label distributions. Therefore, self-supervised learning is particularly effective in improving classification performance on under-represented classes.

We have made our codes and toolkits publicly available to accelerate the introduction and utilization of self-supervised learning algorithms in medicine.

II. RELATED WORKS

The rise of convolutional neural networks (CNNs) in 2012 opened the door for the efficient training of deep inference pipelines on large-scale image recognition datasets [4]. Since then, the performance of CNNs has been significantly improved by the introduction of advanced pipelines such as ResNet and ResNeXt architectures [5], [6]. To this day, most learning architectures are built on supervised learning algorithms, in which annotated data is required to train the network. However, there have been significant advancements in unsupervised learning algorithms and image representation techniques in the recent computer vision literature. Specifically, significant progress has been made in extracting optimal input representation using auto-encoders and contrastive learning, in which the network extracts and learns image representations invariant to various transformations of observations sharing the same identity, such as rotated or altered context views [7]–[13].

Many of these supervised and unsupervised architectures have been applied in the medical field, particularly to image processing tasks in radiology and pathology [14]–[16]. In comparison, the effectiveness of deep learning in histology and histopathology is less investigated. [17] have recently released a publicly available comprehensive bone marrow cell cytomorphology corpus. The authors have also established a highly accurate benchmark for bone marrow cell classification using a supervised deep convolutional neural network.

Here, we present our approach, which further utilizes the semantic context of the cell images in training via self-supervised techniques and improves the current state-of-the-art performance for determining the type of bone marrow cells.
III. METHODS

A. Supervised

We prepared an efficient self-supervised learning inference pipeline to improve the current state-of-the-art and set a new self-supervised learning baseline for the bone marrow cell cytomorphology task. Similar to [17], we have ResNeXt-50 at the core of the supervised pipeline. The pipeline in [17] employs specific color perturbations and augmentations, adopted from [18]. We did not implement these augmentation strategies to limit the confounding effect of such pre/post-processing modules while inspecting our model’s effectiveness. We propose that a stable training procedure with a longer epoch sequence can further improve performance compared to task-specific histopathology pre-processing transformations.

Additionally, [17]’s dataset class distribution is highly imbalanced, reflecting the actual imbalanced distribution of various cell types in the human bone marrow tissue. As such, we used balanced sampling and stratified splitting in our pipeline to deal with this considerable class imbalance. This was particularly beneficial in preventing the under-penalization of images sampled from less frequent categories. In addition, we prepared a supervised baseline performance following the mixup augmentation protocol to further improve the supervised baseline performance and to smooth the decision boundaries [19].

The assumptions and setup for our supervised inference pipeline are as follows: Let \( \mathcal{D} \) represent [17] dataset’s expert image annotations, in which \( \mathcal{D} = \{(x_i, y_i)\}_{i=1}^{\mathcal{D}} \). Each \( x_i \) is an RGB image of dimensions \( H \times W \times 3 \) corresponding to a single bone marrow image from the dataset, with the assumption that magnification and scale are consistent throughout the dataset. \( \mathcal{T}(\cdot) \) is a probabilistic transformation including standardization and flips along horizontal and vertical axes. Given each \( (x_i, y_i) \) sample, class probabilities for all \( j \in \{1, 2, \cdots, C\} \) categories are computed. First, the image is mapped to the semantic space through \( f(\cdot; \Theta) \):

\[
\begin{align*}
  z_i &= f(\mathcal{T}(x_i); \Theta) \quad \text{(Latent Representation)} \\
  i &\in [N] = \{1, 2, \cdots, N\}
\end{align*}
\]

Having the latent representation \( z \in \mathbb{R}^{d_{\text{latent}}} \), class probabilities are computed by performing softmax operation on the dot product of the latent representation \( z_i \) and class representations \( (w_j) \) is the \( j \)th column of \( W_{\text{cls}} \), where \( W_{\text{cls}} \) and \( b_{\text{cls}} \) are weights and biases of the last projection layer.

\[
p_i^{(j)} = \frac{\exp(w_i^T \cdot z_i + b_j)}{\sum_{k=1}^{C} \exp(w_i^T \cdot z_i + b_j)}
\]

\( p_i^{(j)} \) gives us the empirical approximation of \( \Pr\{Y_i = j|X_i = x\} \). The objective is then to minimize the following cross-entropy loss:

\[
l_i = -\sum_{j=1}^{C} y_i^{(j)} \cdot \ln p_i^{(j)}
\]

\[
\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} l_i = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} y_i^{(j)} \cdot \ln p_i^{(j)}
\] (4)

One of the known disadvantages of cross-entropy loss is producing uncalibrated probability distributions, leading to high-confidence misclassification of rare classes. We use weighted cross-entropy and balanced sampling to alleviate such effects. Balanced sampling ensures that in each epoch, each example is seen \( N_c \) times. Hence, in Equation 4 we have \( N = C \cdot N_c \).

B. Self-supervised

We trained a visual recognition pipeline following the SwAV strategy to learn visual representations without using any labels or expert annotations [9]. We use ResNet-50 architecture at the core of the model. We demonstrate that upon convergence, it leads to performance comparable to the supervised scheme and even surpasses its performance in most tasks.

Having the dataset \( \mathcal{D} = \{(x_i, y_i)\}_{i=1}^{\mathcal{D}} \), we ignored all of the labels for pretraining and considered \( \mathcal{D}^{\text{unsupervised}} \), which only contains the datapoints.

Since we are not using cross-entropy, the implicit bias due to label imbalance is less impactful on the final training outcome. As we are working with the imbalanced class distribution of the original dataset, we have \( N = |\mathcal{D}| \).

The latent representations are computed using multiple transformed versions of each image according to Equation 1. These representations are then fed to a projection head \( g(\cdot) \). The resulting projected representations of \( m \)th view of each example \( x_i \) is \( u_i^{(m)} \in \mathbb{R}^{d_{\text{proj}}}^{\mathcal{D}} \):

\[
u_i^{(m),(\text{raw})} = g(z_i^{(m)}; \omega) \quad \text{(Projected Representations)}
\]

\[
m \in \{1, 2\} \land i \in [N] = \{1, 2, \cdots, N\}
\]

\[
u_i^{(m)} = \frac{v_i^{(m),(\text{raw})}}{|u_i^{(m),(\text{raw})}|_2} \quad \text{(L2 Normalization)}
\]

A prototype set of \( \Omega = \{v_1, v_2, \cdots, v_{|\Omega|}\} \) is prepared and initialized uniformly on the \( d_{\text{proj}} \)-dimensional unit ball to effectively capture the common attributes of images. For each example, the similarities of the corresponding projection and prototype vectors give a predicted code vector \( c_i \in \mathbb{R}^{|\Omega|} \) such that:

\[
c_i^{(m)}[j] = \text{sim}(u_i^{(m)}, v_j)
\]

The prototypes are also assigned to the examples according to the optimal control protocol and via the Sinkhorn algorithm, leading to pseudo-groundtruth distributions \( \alpha_i \in \mathbb{R}^{|\Omega|} \). The SwAV loss is then computed as below:

\[
l_i = -\frac{1}{M(M-1)} \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} \sum_{j=1}^{|\Omega|} \alpha_i^{(m_1)}[j] \ln c_i^{(m_2)}[j]
\]

(8)
\[ \mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} l_i = -\frac{1}{NM(M-1)} \sum_{i=1}^{N} \sum_{m_1=1}^{M} \sum_{m_2=1}^{M} \sum_{j=1}^{[0]} \omega_i (m_1) [j] \ln c_i (m_2) [j] \] (9)

Intuitively, this loss encourages the examples that are similar but distinct, meaning that they share the same identity, such as different views of a single image, to predict each others' codes. Given that the batch size \( B \) is relatively small for the Sinkhorn algorithm to work effectively, we also have the queue \( Q \) to facilitate better assignments. Once the training is done, the core CNN is frozen, and a linear projection for classification with parameters \((W_{cls}, b_{cls})\) are introduced and trained by optimizing the cross-entropy loss (Equation 4) to perform the task by leveraging these semi-supervised representations.

C. Supervised Contrastive

Cross-entropy loss is the de-facto for training deep inference pipelines. Nevertheless, it suffers from challenges leading to an increased likelihood of non-calibrated predictions over the space of labels [12]. When sufficient expert annotation is available, it is preferred to make the most of both worlds. This means that we would like to use both the expert annotations and at the same time, utilize contrastive techniques in regularizing model consistency for semantically similar examples. To do so, we trained the ResNet50 model using the Supervised Contrastive criterion [20]. The results help shed light on the implicit bias due to the long-tail distribution of labels and the advantages and disadvantages compared to the purely self-supervised approach.

For each image, we apply the probabilistic transformation \( T \) multiple times to obtain \( M \) different views. The latent representations are then computed according to Equation 1. We will then train a head \( g_{\text{supcon}}(\cdot;\omega) \) which gives us projected representations:

\[ u_i^{(m)}(z_i^{(m)};\omega) \text{ (Projected Representations)} \]

\[ m \in \{1, 2\} \land i \in [N] = \{1, 2, \cdots, N\} \] (10)

In each training batch \( B \) with indices \( i_1, \ldots, i_B \), each view \( m \) of example \( x_i \) will have the following positive set:

\[ P(i) = \{i_j | y_{i_j} = y_i\} \] (11)

This means that in a batch of \( B \) items (each item being the projected latent representation of a view of an image), all of the different views of the same image and every view of every image with the same label are considered similar. The following loss is then used as optimization objective:

\[ \mathcal{L} = -\frac{1}{B} \sum_{i=1}^{B} \frac{1}{|P(i)|} \sum_{j \in P(i)} \ln \frac{\exp((u_i^T u_j)/\tau)}{\sum_{k \in [B], k \neq i} \exp((u_i^T u_k)/\tau)} \] (12)

The intuition of this loss is the same as contrastive approaches such as SimCLR and InfoNCE Loss [7]; however, the difference is that the positive set for each example includes all samples with the same label; thus, leveraging the labels while leveraging inter-example similarities to obtain more representative features. Similarly, to perform the final classification head, the projection head \( g_{\text{supcon}}(\cdot;\omega) \) is discarded, and a linear classifier is trained on the frozen representations.

IV. Experiments

Dataset: Here, we utilized [17]'s large bone marrow cell image dataset, which is publicly available, to train our pipelines. The dataset consists of 171374 images (250 x 250 pixels) obtained from 945 patients with a range of reactive and malignant hematological disorders [17]. All images were obtained at high magnification (40X oil immersion objective) and consisted of a bone marrow cell at the center of the patch, while the image periphery in most cases also contains portions of neighboring marrow cells. The annotations for these images correspond to 21 bone marrow cell types shown in Table I.

Supervised: We first experimented with training the proposed pipeline for a short sequence of only 20 epochs and cross-entropy as the optimization objective to compare with the baseline performance in [17], the results of which are reported in Table II. The macro-level average of precision and recall over categories were 0.07 and 0.04 higher than the reported values in the baseline. The macro average F-1 score
of this experiment was 0.07 higher than the baseline\footnote{Please note that the baseline F-1 score reported here is computed from the reported values for mean precision and recall, which might be different from the mean F-1 results, as the baseline in [17] does not report F-1 values.}, further highlighting the potential of our simple modifications to the training pipeline.

Next, we built a more comprehensive supervised pipeline. We designed two 5-fold experiments with the more extended training sequence of 100 epochs to ascertain a full convergence. Table III shows the detailed report of the precision and recall per each class, compared with the state-of-the-art model in [17]. Our models show consistent improvement across most cell classes over the baseline model. Mixup augmentation showed advantages in improving the performance in some of the cell classes; nevertheless, it did not lead to a consistent overall improvement.

**Self-Supervised:** We next turned our attention to a self-supervised learning approach to classify cells in [17]’s bone marrow dataset. One of the significant advantages of using a self-supervised learning approach in this highly class-imbalanced dataset is its robustness to the long-tail distribution of labels, which is known to be a likely cause of performance degradation in imbalanced datasets [21].

Our configuration for the self-supervised experiment is as follows:

- The probabilistic transformation $T(\cdot)$ that consists of usual transforms, such as color jitter in the HSV color space and random flipping.
- $f(\cdot; \Theta)$ is ResNet-50 up until the final pooling layer, returning $z_i \in \mathbb{R}^{2048}$. 
- $g(\cdot; \omega)$ is a linear layer projecting $z_i$ to a subspace of $\mathbb{R}^{128}$.
- The training is carried out for 200 epochs, with a queue (of size 1280) being considered starting from epoch 15 and prototypes $\Omega$ frozen for the first 2 epochs.
- The cardinality of our prototype set is $|\Omega| = 1000$.
- TheSinkhorn algorithm is carried out for 3 iterations, with an epsilon of $\epsilon = 0.03$.
- The batch size was 256, and the model was trained on 4 NVIDIA Quadro RTX-5000 GPUs.

**Supervised Contrastive:** Lastly, following our fully unsupervised experiments, we took a hybrid approach of optimizing the supervised contrastive loss to demonstrate that one can leverage the expert annotations while benefiting from contrastive learning’s higher level of robustness to dataset challenges. We performed the experiment with the temperature hyperparameter of $\tau = 0.2$. At the core, our $f(\cdot; \Theta)$ was a ResNet-50 architecture truncated at the last pooling output. The projection head $g_{\text{supcon}}(\cdot; \omega)$ was mapping to a subspace of $\mathbb{R}^{128}$. The model was trained with the batch size of 256 on 4 NVIDIA Quadro RTX-5000 GPUs.

The results of both supervised contrastive and self-supervised methods, shown in Table IV, demonstrate the advantages of our approach. Furthermore, the very best results come from the completely unsupervised approach.

| Class         | Abbreviation | Support |
|---------------|-------------|---------|
| Abnormal eosinophils | ABE | 8       |
| Artefacts     | ART | 19630   |
| Band Neutrophils | NGB | 9608   |
| Basophils     | BAS | 441    |
| Blasts        | BLA | 11973  |
| Eosinophils   | EOS | 5883   |
| Erythroblasts | EBO | 27395  |
| Faggot cells  | FGC | 47     |
| Hairy cells   | HAC | 409    |
| Immature lymphocytes | LLY | 65     |
| Lymphocytes   | LYT | 26242  |
| Metamyelocytes | MMZ | 3085   |
| Monocytes     | MON | 4040   |
| Myelocytes    | MYB | 6557   |
| Not identifiable | NIF | 3538   |
| Other cells   | OTH | 294    |
| Plasma cells  | PLM | 7629   |
| Proerythroblasts | PEB | 2740  |
| Promyelocytes | PMO | 11994  |
| Smudge cells  | KSC | 42     |

| Class         | F1-Score (avg) |
|---------------|---------------|
| Abnormal eosinophils | 0.04          |
| Artefacts     | 0.78          |
| Basophils     | 0.23          |
| Blasts        | 0.70          |
| Erythroblasts | 0.85          |
| Eosinophils   | 0.88          |
| Faggot cells  | 0.27          |
| Hairy cells   | 0.49          |
| Immature lymphocytes | 0.14        |
| Lymphocytes   | 0.79          |
| Metamyelocytes | 0.41        |
| Monocytes     | 0.63          |
| Myelocytes    | 0.55          |
| Band Neutrophils | 0.59        |
| Segmented neutrophils | 0.80      |
| Not identifiable | 0.38        |
| Other cells   | 0.35          |
| Proerythroblasts | 0.60        |
| Plasma cells  | 0.82          |
| Promyelocytes | 0.74          |

Avg: 0.53 0.62
supervised contrastive setting. Our empirical findings showed considerable improvement in this domain.
