IMBALANCED DOMAIN GENERALIZATION FOR ROBUST SINGLE CELL CLASSIFICATION IN HEMATOLOGICAL CYTOMORPHOLOGY

Rao Muhammad Umer[1], **Armin Gruber**[1,2], **Sayedali Shetab Boushehri**[1,3], **Christian Metak**[1], **Carsten Marr**[1]
{christian.metak, carsten.marr}@helmholtz-muenchen.de
Institute of AI for Health, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg 85764, Germany.[1]
Laboratory of Leukemia Diagnostics, Department of Medicine III, University Hospital, LMU Munich, Munich, Germany.[2]
Data Science, Pharmaceutical Research and Early Development Informatics (pREDi), Roche Innovation Center Munich, Germany.[3]

ABSTRACT

Accurate morphological classification of white blood cells (WBCs) is an important step in the diagnosis of leukemia, a disease in which nonfunctional blast cells accumulate in the bone marrow. Recently, deep convolutional neural networks (CNNs) have been successfully used to classify leukocytes by training them on single-cell images from a specific domain. Most CNN models assume that the distributions of the training and test data are similar, i.e., the data are independently and identically distributed. Therefore, they are not robust to different staining procedures, magnifications, resolutions, scanners, or imaging protocols, as well as variations in clinical centers or patient cohorts. In addition, domain-specific data imbalances affect the generalization performance of classifiers. Here, we train a robust CNN for WBC classification by addressing cross-domain data imbalance and domain shifts. To this end, we use two loss functions and demonstrate their effectiveness in out-of-distribution (OOD) generalization. Our approach achieves the best F1 macro score compared to other existing methods and is able to consider rare cell types. This is the first demonstration of imbalanced domain generalization in hematological cytomorphology and paves the way for robust single cell classification methods for the application in laboratories and clinics.

1 INTRODUCTION

Hematology deals with the study of blood, blood-forming tissue, and blood-related diseases. Precise and early diagnosis of a hematologic disorder is crucial for the successful treatment. For decades, microscopic examination and classification of blood cells in stained peripheral blood (PB) and bone marrow (BM) samples have been a key step for the diagnosis of hematological malignancies. Cytomorphologic examination using light microscopy (Walter et al., 2022) remains one of the backbones of hematological diagnostics, often representing the first step in the workup and guiding additional methods such as immunophenotyping, cytogenetics, and molecular genetics. During morphologic examination, blood samples are evaluated microscopically by hematologists and screened for the presence of atypical cells populations that can indicate conditions such as leukemia. Typically, at least 200 cells per sample have to be classified according to current clinical guidelines. Such manual evaluation and classification under the microscope can be tedious, repetitive, and time-consuming. It furthermore relies heavily on trained and experienced staff, and is prone to variabilities due to the human factor i.e., limited intra-/inter-observer reproducibility.

Deep learning (DL) models have shown great potential in solving real-world classification tasks in various areas, such as computer vision (He et al., 2016), natural language processing (Brown et al., 2020), and healthcare including medical imaging (Matek et al., 2019; Wang et al., 2019) and histopathology (Komura & Ishikawa, 2018; Hägele et al., 2020). They are usually developed

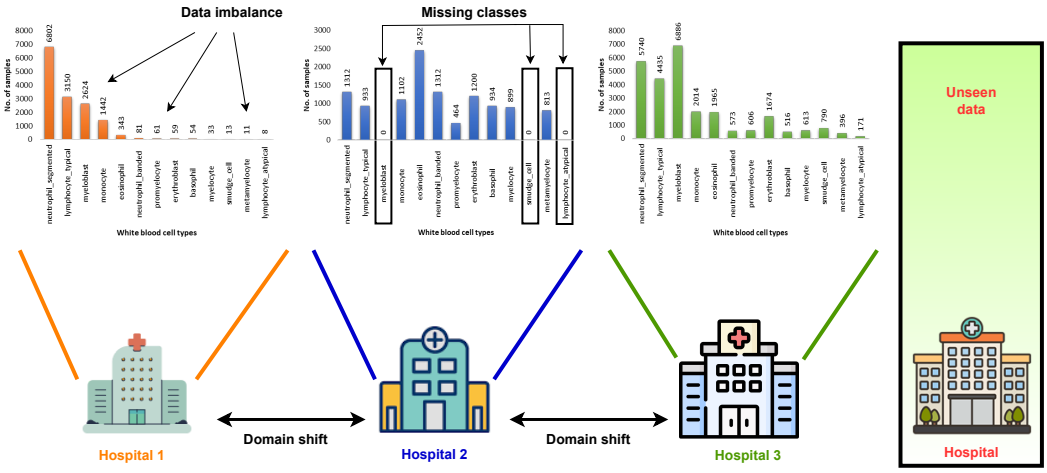


Figure 1: Key challenges for robust classification in an unseen target domain are data imbalance, missing classes, and domain shifts. In our application, hospital 1, 2, and 3 provide single white blood cell datasets from distinct domains with different class distributions.

and tested under the implicit assumption that the train and test data are drawn independently and identically from the same distribution (IID). Generalizing to unseen test domains is natural to humans, but challenging to machines. Domain generalization (DG) aims at training for domain invariant models robust against distribution shifts by utilizing data from distinct domains (Zhou et al., 2022; Wang et al., 2022). However, real-world data from multiple distinct domains often exhibit imbalanced label distributions i.e., a few classes contain a very large number of samples in the one domain, while only a few samples or none in another domain (see Fig. 1). Moreover, heavily imbalanced data distributions are common, and minority class samples in one domain could be abundant in other domains. Therefore, tackling the problem of cross-domain data imbalance is key to develop generalizable and diagnostically reliable models.

Recently, several deep-learning models for the classification of WBCs have been proposed (Sidhom et al., 2021; Cheuque et al., 2022; Eckardt et al., 2022; Salehi et al., 2022; Hehr et al., 2023). In Matek et al. (2019), the authors trained a deep CNN model for the classification of single cell images from peripheral blood smears. The training data used for the model consisted of expert-annotated single cell images from the different subtypes of acute myeloid leukaemia (AML) patients. The images were processed using a ResNeXt-50 model (Xie et al., 2017), which provided high precision and recall for most diagnostically relevant classes, by assigning the class with the highest prediction probability to each image. So far, all existing state-of-the-art (SOTA) DL approaches are based on homogeneous datasets (identical train and test distribution), making the assumption of data balance during training by oversampling the minority classes or some loss re-weighting techniques (Zhou et al., 2022). These datasets tend to be imbalanced not only due to varying cell distributions between different disease subtypes, but also due to varying patient populations and disease prevalences at different centers, making it difficult to predict the performance of a model on under-represented cases. This can result in limited performance and carries the risk of poor diagnostic performance. Additionally, single cell images from different labs can vary in sharpness, brightness, contrast, scale, color, and other properties, thus making it necessary to develop a robust classifier that can confidently classify single cell images regardless of their source domain and preanalytic handling.

Since single cell image data often originate from multiple distinct domains (i.e., hospitals and laboratories), many challenges arise (refer to section-2.1 for more details) for robust classification, as shown in Fig. 1. The first challenge is the data imbalance within and across domains, the second one is the domain shift within and across domains, and the third one is missing class samples in certain domains. Our aim is to train a robust classifier, learning invariant features among multiple distinct domains, with each domain having its own domain shift and imbalanced label distribution problems, and generalizing to an unseen test target domain. To deal with the above problems, we train a robust classifier (as shown in Fig. 2) in an end-to-end fashion by minimizing the whole objective function (1) inspired by the work of Yang et al. (2022). The loss functions operate over the latent features and

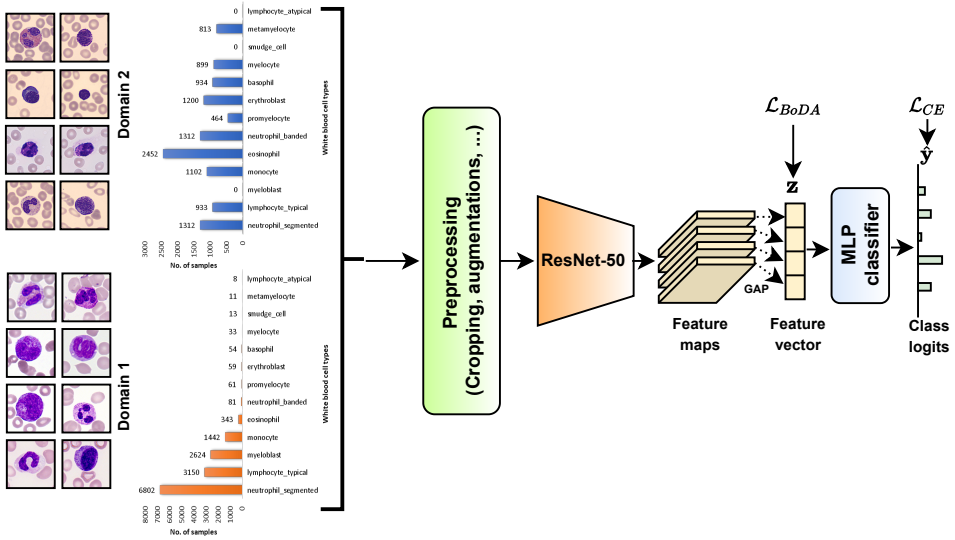


Figure 2: Setup of our robust WBC classification approach: We take the single cell images from two different domains, preprocess with random resized cropping and various data augmentations, feed them into a pre-trained ResNet-50 network, and get the feature maps and the feature vector ($\mathbf{z}$) by using global average pooling (GAP). The feature vectors are fed into the multilayer perceptron (MLP) classifier to get the class logits ($\hat{\mathbf{y}}$). The total loss consists of two terms ($\mathcal{L}_{\text{BoDA}}$, $\mathcal{L}_{\text{CE}}$) and is minimized during the fine-tuning of the network in an end-to-end manner.

output layer respectively, and encourage similarity between features of the same class in different domains, and dissimilarity between features of different classes within and across domains, as well as correctly predicted class labels.

2 METHODOLOGY

In this section, we explain robust classification challenges in detail and then propose a network training scheme to solve the existing problems for our task.

2.1 ROBUST CLASSIFICATION CHALLENGES

2.1.1 DATA IMBALANCE

Imbalanced data, where the number of samples in one class is significantly higher than the number in another class, can cause the learning algorithm to be biased towards the majority class. Data imbalance is an intrinsic problem (Buda et al., 2018; Cao et al., 2019; Yang & Xu, 2020) in real-world data, and it is even more severe in the medical domain. For example, the Matek_19 (Matek et al., 2019) dataset has severe class imbalance, where the majority class has more than 6500 samples, while the minority classes have less than 50 samples (Fig. 1). In the Acevedo_20 (Acevedo et al., 2020) dataset, we have somehow balanced data (Fig. 1), while in the INT_20 dataset, we have a data imbalance problem with a long-tailed distribution (Fig. 1).

2.1.2 DOMAIN SHIFT

In the medical domain, domain shifts can emerge due to different staining protocols, different scanners or acquisition protocols (i.e., background light, focus, etc.), different magnifications / resolutions, and variations in clinical centers or patient cohorts. The data can have distribution shifts even within a domain. For example, all three datasets used here have color variations due to different staining protocols and have different image resolutions due to different scanners (see Fig. 2). Furthermore, in the Acevedo_20 (Acevedo et al., 2020) dataset, as shown with visual image samples in Fig. 2, there are strong color variations within the domain.

Table 1: Imbalanced DG classification results (mean±std) determined by five-fold cross-validation on Acevedo_20 and Matek_19 validation-sets. Our base-line model is ResNet50, pretrained on ImageNet.

Methods	**F1-micro↑**	**F1-macro↑**
ERM (Vapnik, 1999)	**0.93** ±0.01	0.77 ±0.02
DANN (Ganin et al., 2016)	0.87 ±0.03	0.67 ±0.04
CORAL (current SOTA DG) (Sun & Saenko, 2016)	0.92 ±0.01	0.76 ±0.03
Ours	**0.93** ±0.01	**0.78** ±0.05
Ours$^+$	0.90 ±0.02	0.76 ±0.04

2.1.3 MISSING CLASSES

In a domain i.e., the Acevedo_20 (Acevedo et al., 2020) dataset (see Fig. 1), we have no single cell image data at all for certain classes. In some cases, divergent label distributions across domains can occur, rendering the problem more complex. Training data from different domain enhance classifier performance, where the minority samples or even no samples of the classes from one domain (hospital) are enriched with instances form the other domains.

2.2 NETWORK TRAINING LOSSES

For robust classifier learning, we train a deep CNN network (i.e., ResNet-50 (He et al., 2016)) with standard hyperparameter settings as done in Yang et al. (2022) by minimizing the following loss function:

$$\mathcal{L} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{CE}} + \lambda \mathcal{L}_{\mathrm{BoDA}}. \tag{1}$$

Here, θ are the parameters of the network and λ is a trade-off hyperparameter between the two loss terms.

For the $\mathcal{L}_{\mathrm{CE}}$ loss function, we use the standard cross-entropy loss applied to the output layer of the network ($\hat{\mathbf{y}}$) with the ground-truth label ($\mathbf{y}$). It is defined as:

$$\mathcal{L}_{CE}(\hat{\mathbf{y}}, \mathbf{y}) = -\frac{1}{N} \sum_{n=1}^{N} \mathbf{y}_n \log \hat{\mathbf{y}}_n + (1 - \mathbf{y}_n) \log (1 - \hat{\mathbf{y}}_n) \tag{2}$$

Here, N is the mini-batch size and $\hat{\mathbf{y}}_n$ are the predicted class logits for image n.

For the $\mathcal{L}_{\mathrm{BoDA}}$ loss function, we use Balanced Domain-Class Distribution Alignment (BoDA) (Yang et al., 2022) loss to tackle the data imbalance across domain-class (d_i, c_i) pairs, which is applied to the latent features ($\mathbf{z}$) as:

$$\mathcal{L}_{BoDA}(\mathbf{z}, \boldsymbol{\psi}) = \sum_{\mathbf{z}_i \in \mathcal{Z}} \frac{-1}{|\mathcal{D}| - 1} \sum_{d \in \mathcal{D} \setminus \{d_i\}} \log \frac{\exp\left(-\boldsymbol{w}_{d_i,c_i}^{d,c_i} \hat{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\psi}_{d,c_i})\right)}{\sum_{(d',c') \in \mathcal{M} \setminus \{(d_i,c_i)\}} \exp\left(-\boldsymbol{w}_{d_i,c_i}^{d',c'} \hat{\mathsf{d}}(\mathbf{z}_i, \boldsymbol{\psi}_{d',c'})\right)}. \tag{3}$$

In Eq. (3), the numerator represents positive cross-domain pairs distance $\hat{\mathsf{d}}(.)$ that should be minimized (i.e., attract the same classes) during training, while the denominator represents negative cross-class pairs distance $\hat{\mathsf{d}}(.)$ that should be maximized (i.e., separate different classes) during training. The $\boldsymbol{\psi}$ is mean of the feature vectors of domain-class pairs (d, c). The $\boldsymbol{w}_{d,c}^{d',c'}$ is the calibration parameter, which indicates how much to transfer (d, c) to (d', c') based on their relative sample size. The distance d can be set to the Euclidean distance $\mathsf{d}(\mathbf{z}, \boldsymbol{\psi}_{d,c}) = \sqrt{(\mathbf{z} - \boldsymbol{\psi}_{d,c})^\top (\mathbf{z} - \boldsymbol{\psi}_{d,c})}$, which captures first-order statistics. To match higher-order statistics such as covariance, $\mathsf{d}(\mathbf{z}, \{\boldsymbol{\psi}_{d,c}, \boldsymbol{\Sigma}_{d,c}\}) = 1/N_{d_i,c_i} * \sqrt{(\mathbf{z} - \boldsymbol{\psi}_{d,c})^\top \boldsymbol{\Sigma}_{d,c}^{-1} (\mathbf{z} - \boldsymbol{\psi}_{d,c})}$ can be used, similar to the Mahalanobis distance (De Maesschalck et al., 2000).

Table 2: Imbalanced DG classification results (mean±std) determined by five-fold cross-validation on INT_20 testset (unseen domain). Our base-line model is ResNet50, pretrained on ImageNet.

Methods	**F1-micro↑**	**F1-macro↑**
ERM (Vapnik, 1999)	0.64 ±0.03	0.40 ±0.05
DANN (Ganin et al., 2016)	0.59 ±0.07	0.35 ±0.06
CORAL (current SOTA DG) (Sun & Saenko, 2016)	**0.66** ±0.03	0.43 ±0.03
Ours	**0.66** ±0.05	0.43 ±0.06
Ours$^+$	0.59 ±0.09	**0.46** ±0.08

3 Experiments and Results

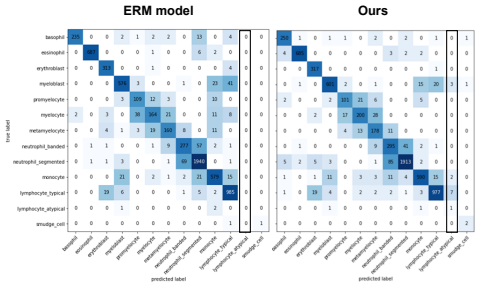
3.1 Datasets and preprocessing

For the robust classification task, we use three datasets of single cell peripheral blood images: Matek_19 (Matek et al., 2019), Acevedo_20 (Acevedo et al., 2020), and an internal data set (INT_20). Appendix A provides detailed statistics of each dataset used in our experiments. We split the data into training (80% of the images) and validation (20% of the images) for the source domains (Acevedo_20 and Matek_19), and testset (20% of the images) for the unseen (OOD) test target domain (INT_20). For a five-fold cross-validation, we perform random stratified splits of the single cell images into five folds.

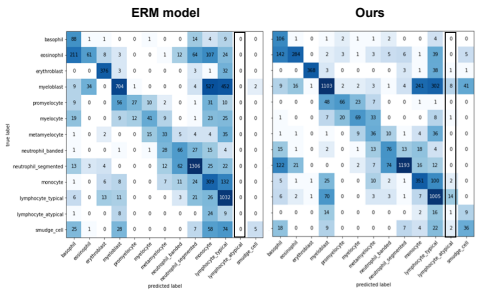
During data preprocessing, we crop all samples to the same size, so that in each dataset, the ratio of cell pixel to background pixel is approximately identical. During training, we apply multiple standard transformations of the images, such as random resize, horizontal / vertical flip, rotation, color-jitter, blur, and grayscale, to make the model more robust to image variations.

3.2 Evaluation Metrics and Results

As the datasets are imbalanced, we aim to achieve a high F1-macro score on unseen test domain (OOD), which gives equal weight to each class regardless of its cardinality. Besides that, we also report the F1-micro score, which gives the weighted average of each class. We report imbalanced DG classification results in Table 1 and Table 2 determined by five-fold cross-validation on the source domains (Acevedo_20 and Matek_19) validation-set and unseen test domain (INT_20). We compare our model to one vanilla training model (2) like ERM (Vapnik, 1999), and two current SOTA DG methods, such as DANN (Ganin et al., 2016) and CORAL (Sun & Saenko, 2016). We train two different variants of our model, one with coupled training of the encoder part (representation, $\mathbf{z}$) and the MLP classifier part, and the other (denoted as Ours$^+$) with decoupled training of the encoder part and the classifier part by using class-balanced sampling. We achieve the best F1-marco score on the INT_20 testset, while the DG method DANN performs worse than the vanilla ERM model due to its assumption of data balance during training. In Fig. 3, we also



(a) Source (Matek_19, Acevedo_20) domain val-set



(b) Target (INT_20) domain testset (unseen)

Figure 3: Confusion matrices show an improved classification of cells from the lowly populated **lymphocyte_atypical** class with our method compared to the standard ERM model.

compare the confusion matrix of our model with the ERM method. Our model performs better by recognizing samples from the minority class, while ERM misses samples from the minority class, such as lymphocyte_atypical.

4 CONCLUSION

We develop a robust CNN model for out-of-distribution generalization in hematological cytomorphology that tackles three main challenges: data imbalance, domain shifts, and missing classes. We show how existing pre-trained deep models can be improved for distinct domains by optimizing the loss function in the latent feature space and output logits of the network. Our work shows how biological, epidemiological, and technical variabilities in hematologic single WBC classification can be addressed for training robust AI-based cell classifiers, paving the way for their safe and productive use in a clinical setting.

REFERENCES

Andrea Acevedo, Anna Merino, Santiago Alférez, Ángel Molina, Laura Boldú, and José Rodellar. A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in brief*, 30, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks*, 106:249–259, 2018.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

César Cheuque, Marvin Querales, Roberto León, Rodrigo Salas, and Romina Torres. An efficient multi-level convolutional neural network approach for white blood cells classification. *Diagnostics*, 12(2):248, 2022.

Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000.

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.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Miriam Hägele, Philipp Seegerer, Sebastian Lapuschkin, Michael Bockmayr, Wojciech Samek, Frederick Klauschen, Klaus-Robert Müller, and Alexander Binder. Resolving challenges in deep learning-based analyses of histopathological images using explanation methods. *Scientific reports*, 10(1):1–12, 2020.

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*, pp. 770–778, 2016.

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*, pp. 1–17, 2023.

Daisuke Komura and Shumpei Ishikawa. Machine learning methods for histopathological image analysis. *Computational and structural biotechnology journal*, 16:34–42, 2018.

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.

Raheleh Salehi, Ario Sadafi, Armin Gruber, Peter Lienemann, Nassir Navab, Shadi Albarqouni, and Carsten Marr. Unsupervised cross-domain feature extraction for single blood cell image classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22*, pp. 739–748, 2022.

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.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pp. 443–450. Springer, 2016.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Wencke Walter, Christian Pohlkamp, Manja Meggendorfer, Niroshan Nadarajah, Wolfgang Kern, Claudia Haferlach, and Torsten Haferlach. Artificial intelligence in hematological diagnostics: Game changer or gadget? *Blood Reviews*, pp. 101019, 2022.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Qiwei Wang, Shusheng Bi, Minglei Sun, Yuliang Wang, Di Wang, and Shaobao Yang. Deep learning approach to peripheral leukocyte recognition. *PloS one*, 14(6):e0218808, 2019.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1492–1500, 2017.

Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. *Advances in neural information processing systems*, 33:19290–19301, 2020.

Yuzhe Yang, Hao Wang, and Dina Katabi. On multi-domain long-tailed recognition, imbalanced domain generalization and beyond. In *European Conference on Computer Vision (ECCV)*, 2022.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

A DATASET DETAILS

In this section, we provide detailed information of the three datasets used in our experiments. Table 3 provides statistics and properties of Matek_19, Acevedo_20, and INT_20.

The Matek_19 (Matek et al., 2019) dataset contains 14681 single cell images, divided into 13 classes, each image with a size of $400 \times 400 \times 3$ pixels corresponding to 29×29 micrometers. The resolution is 13.8 pixels per micron.

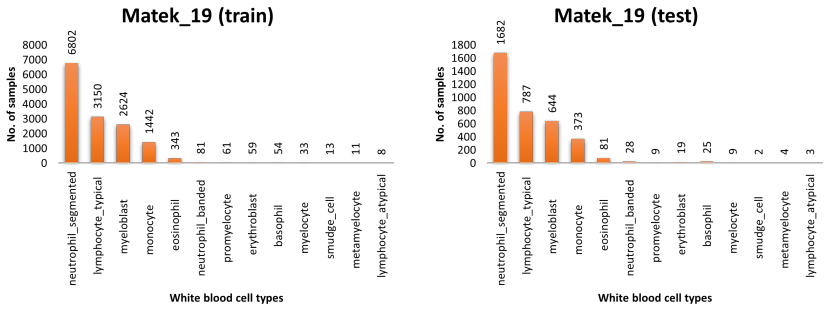
The Acevedo_20 (Acevedo et al., 2020) dataset contains 11421 single cell images, divided into 10 classes, each image of $360 \times 363 \times 3$ pixels, corresponding to 36×36.3 micrometers. The resolution is 10 pixels per micron.

INT_20 is an internal dataset (currently not publicly available), which contains 26379 single cell images, divided into 13 classes, each image of size $288 \times 288 \times 3$ corresponding to 25×25 micrometers, so the resolution is 11.52 pixels per micron.

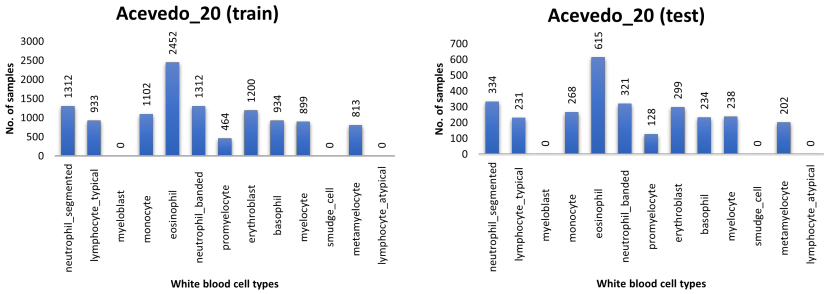
Table 3: Statistics and properties of the three datasets used in our experiments.

Dataset	**# classes**	**Image size**	**Image resolution**	**# data samples**
`Matek_19`	13	$400 \times 400 \times 3$	$29.0\ \mu m \times 29.0\ \mu m$ $= 13.8$ pixels/micron	14681
`Acevedo_20`	10	$360 \times 363 \times 3$	$36.0\ \mu m \times 36.3\ \mu m$ $= 10$ pixels/micron	11421
`INT_20`	13	$288 \times 288 \times 3$	$25.0\ \mu m \times 25.0\ \mu m$ $= 11.52$ pixels/micron	26379

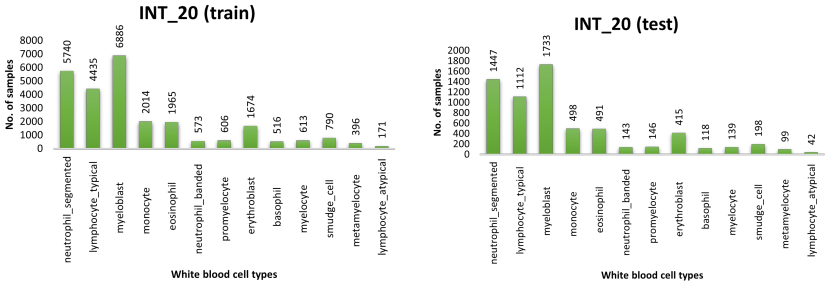
Fig. 4 shows the sample distribution of each class in the trainset (80% of the single cell images) and the testset (20% of the single cell images) for all three datasets. We see highly imbalanced data in Matek_19, missing samples in Acevedo_20 domain, and a long-tail distribution in INT_20.



(a) Matek_19 train and test set



(b) Acevedo_20 train and test set



(c) INT_20 train and test set

Figure 4: Train and test set class distributions of the three single cell datasets.