Imbalanced Domain Generalization for Robust Single Cell Classification in Hematological Cytomorphology

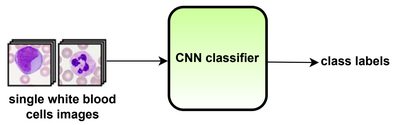
Rao Muhammad Umer, Armin Gruber, Sayedali Shetab Boushehri, Christian Matek, and Carsten Marr

Institute of AI for Health, Helmholtz Munich, Germany.

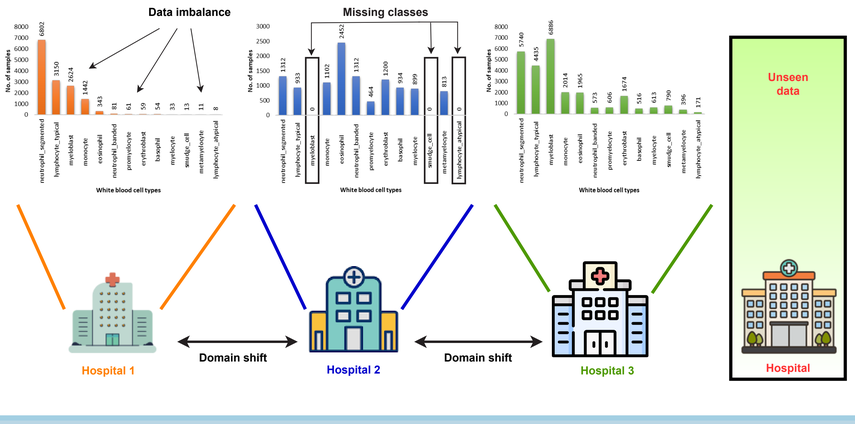
Contact: raoumer.github.io

Problem Formulation

Problem: Cytomorphology as a image classification problem.

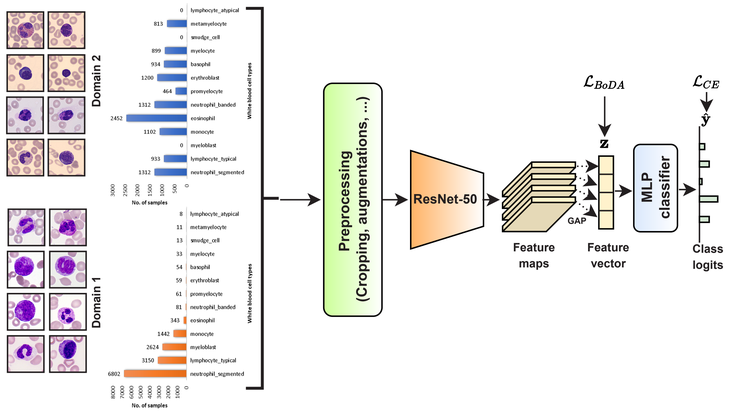


Goal: Train a robust classifier for unseen target data of white blood cell classification by addressing cross-domain data imbalance and domain shifts.



Network Training

Approach:



Training Loss:

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

Standard cross-entropy (CE) applied to output layer:

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

Balanced Domain-Class Distribution Alignment (BoDA) loss to tackle the data imbalance across domain-class pairs, which is applied to the latent features:

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

Positive cross-domain pairs

Negative cross-class pairs

Data and Experimental Results

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

Dataset	# classes	Image size	Image resolution	Single cell images
`Matek_19` [1]	13	$400 \times 400 \times 3$	$29.0\,\mu m \times 29.0\,\mu m$ $= 13.8$ pixels/micron	14681
`Acevedo_20` [2]	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

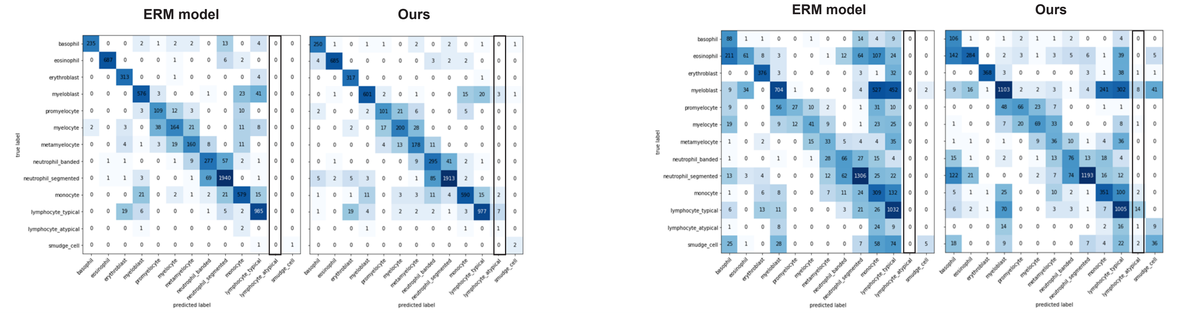
Quantitative Results:

- Imbalanced domain generalization classification results (mean±std) determined by five-fold cross-validation on Acevedo_20 & Matek_19 validation sets and INT_20 testset (unseen domain). Our base-line model is ResNet50, pretrained on ImageNet.

Methods	F1-micro↑	F1-macro↑
ERM [3]	**0.93** ±0.01	0.77 ±0.02
DANN [4]	0.87 ±0.03	0.67 ±0.04
CORAL (current SOTA DG) [5]	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

Methods	F1-micro↑	F1-macro↑
ERM [3]	0.64 ±0.03	0.40 ±0.05
DANN [4]	0.59 ±0.07	0.35 ±0.06
CORAL (current SOTA DG) [5]	**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

- Confusion matrices show an improved classification of cells from the lowly populated **lymphocyte_atypical** class with our method compared to the standard ERM model on validation and test set, respectively.



Conclusions

- We develop a robust CNN model for out-of-distribution generalization in hematological cytomorphology classification that tackles three main challenges: data imbalance, domain shifts, and missing classes.
- Our work shows how biological, epidemiological, and technical variabilities in hematologic single white blood cell classification can be addressed for training classifiers.

References

[1] C. Matek, S. Schwarz, K. Spiekermann, and C. Marr, "Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks," *Nature Machine Intelligence*, 2019.

[2] A. Acevedo, A. Merino, S. Alférez, Á. Molina, L. Boldú, and J. Rodellar, "A dataset of microscopic peripheral blood cell images for development of automatic recognition systems," *Data in brief*.

[3] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, 1999.

[4] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, 2016.

[5] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *ECCV*, 2016.