scientific data

OPEN

DATA DESCRIPTOR

A large expert-annotated single-cell peripheral blood dataset for hematological disease diagnostics

Sayedali Shetab Boushehri[1,2,3,9], Salome Kazeminia[1,3,9], Armin Gruber[1,4], Christian Matek[1], Karsten Spiekermann[4,5,6], Christian Pohlkamp[7], Torsten Haferlach[7] & Carsten Marr[1,4,5,8] ✉

Distinguishing cell types in a peripheral blood smear is critical for diagnosing blood diseases, such as leukemia subtypes. Artificial intelligence can assist in automating cell classification. For training robust machine learning algorithms, however, large and well-annotated single-cell datasets are pivotal. Here, we introduce a large, publicly available, annotated peripheral blood dataset comprising >40,000 single-cell images classified into 18 classes by cytomorphology experts from the Munich Leukemia Laboratory, the largest European laboratory for blood disease diagnostics. By making our dataset publicly available, we provide a valuable resource for medical and machine learning researchers and support the development of reliable and clinically relevant diagnostic tools for diagnosing hematological diseases.

Background & Summary

Microscopic examination and classification of blood cells play a crucial role in diagnosing hematological diseases. This process involves evaluating the morphology of leukocytes and is typically performed by human experts trained over years. Like other diagnostic tasks, it is repetitive, time-consuming, and susceptible to intra- and inter-observer variation[1]. One promising solution is the development of automatic single-cell classifiers using machine learning, which can substantially reduce the time and effort required by experts[2]. Deep learning, in particular, has been used for diagnosing hematological diseases from single-cell images in peripheral blood[3–9] and bone marrow[10–12].

As supervised deep learning crucially relies on large amounts of annotated data, a current lack of large datasets creates a bottleneck for improving the accuracy of classifiers[13]. This work presents the largest publicly available, expert-annotated dataset of peripheral blood single-cells, with over 40,000 images. While our dataset is being published here for the first time, it has been used in previous studies[4,5,14–17].

Methods

Ethics declaration. Informed consent was obtained indirectly at the time of routine collection for possible research. All patients in the MLL23 dataset were at least 18 years old. Ethics approval was granted by the Ethics Committee of LMU Munich (reference number 25-0744).

The data acquisition process at the Munich Leukemia Laboratory comprised several steps (see also Hehr *et al.*[4]). Blood samples and smears were collected between 2021 and 2024 from patients with a wide distribution of hematological diagnoses. A patient cohort with blood samples from adult patients who gave informed consent to the use of their data for research purposes was selected. Blood smears were stained using the Pappenheim method and scanned using a fully automated scanning device (Metafer software platform, MetaSystems, Altlussheim, Germany), which was modified in its technical settings for this application. Image acquisition was performed using an automatic autofocus system integrated in the scanning device, without manual focus adjustments. Slides were first scanned with a 10x objective to obtain an overview image. Cell detection was performed

[1]Computational Health Center, Helmholtz Munich – German Research Center for Environmental Health, Neuherberg, Germany. [2]Data & Analytics, Pharmaceutical Research and Early Development (pRED), Roche Innovation Center Munich (RICM), Penzberg, Germany. [3]TUM School of Computation, Information and Technology, Technical University of Munich, Munich, Germany. [4]Department of Medicine III, University Hospital, LMU Munich, Munich, Germany. [5]German Cancer Consortium (DKTK), Heidelberg, Germany. [6]German Cancer Research Center (DKFZ), Heidelberg, Germany. [7]Munich Leukemia Laboratory, Munich, Germany. [8] Munich Center for Machine Learning (MCML), Munich, Germany. [9]These authors contributed equally: Sayedali Shetab Boushehri, Salome Kazeminia. ✉e-mail: carsten.marr@helmholtz-munich.de

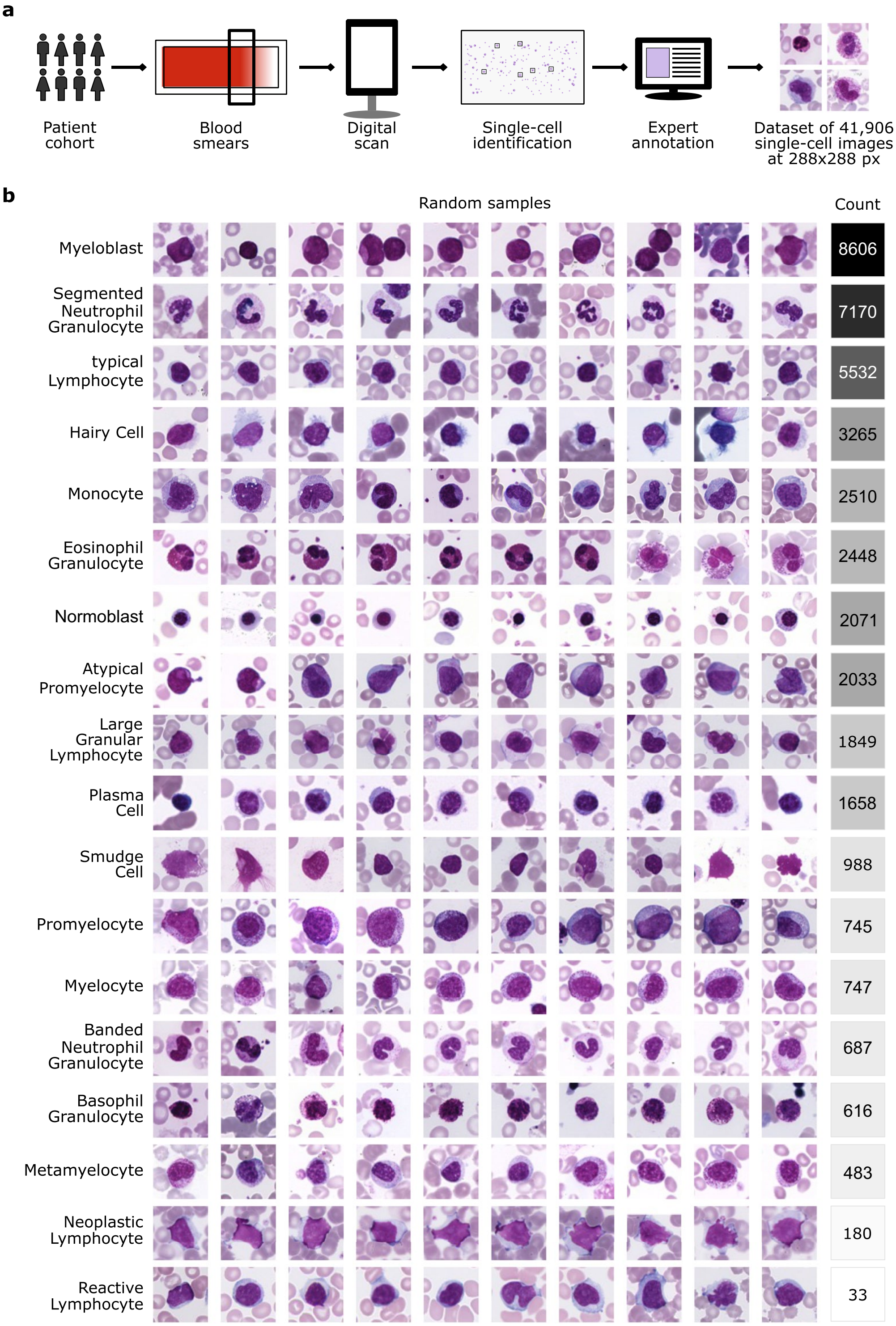

Fig. 1 A fully annotated single-cell peripheral blood dataset. (**a**) Workflow of generating the imaging dataset at the Munich Leukemia Laboratory. (**b**) The MLL23 dataset contains 18 classes with varying numbers of images per class. Ten representative images per class are depicted to provide an overview of the dataset.

using the Metasystems Metafer software. After applying a segmentation threshold and a logarithmic color transformation, stained cells with an object size between 40–800 μm^2 were detected and stored in a gallery. Each gallery image was assigned to a quality level using a classifier to determine cell density and immediate cell neighborhood. High-quality cells identified in the 10x overview images were then re-scanned using a 40x objective. The resulting 41,906 images of single nucleated cells comprise 288 × 288 pixels and 25 μm × 25 μm, corresponding to a resolution of 11.52 pixels per μm. Note that the occasional white bars at the edges of some images result from edge effects when cells are located near boundaries of the scanned field of view. To maintain uniformly sized square images, we padded images with white pixels, matching the background, regardless of horizontal or vertical orientation. Subsequently, five human expert examiners at the Munich Leukemia Laboratory annotated the images, assigning each single cell to one out of 18 classes (Fig. 1a).

We reduced the dataset to 41,621 cells by deleting duplicate images. Some duplicate images also had differing labels, corresponding to indecisive borderline cases. Note that some cells are depicted in two or more images, but with differing focus or cropping. Also, dysplastic cells were excluded from the dataset to ensure clarity in cell type classification.

In the group of lymphoid cells, there are mature 'typical lymphocytes' (number of single-cell images = 5,532) and 'atypical lymphocytes' like plasma cells (1,658), 'large granular lymphocytes' (1,849), 'reactive lymphocytes' (33), 'hairy cells' (3,265) and other 'neoplastic lymphocytes' (180), as well as 'smudge cells' (988). In comparison, the group of myeloid cells is divided into mature cells like band 'neutrophil granulocytes' (687), 'segmented neutrophil granulocytes' (7,170), 'eosinophil granulocytes' (2,448), 'basophil granulocytes' (616), 'monocytes' (2510), and immature cells like 'myeloblasts' (8,606), 'metamyelocytes' (483), 'promyelocytes' (745), 'myelocytes' (747), and 'atypical promyelocytes' (2,033). Lastly, 'normoblasts' (2071) are also present in the dataset. The cell types occur with specific frequencies in the peripheral blood in healthy and pathological patients. Due to the Munich Leukemia Laboratory's focus on hematologic neoplasms, the dataset is inherently imbalanced in terms of the number of images per class. For instance, it contains over 8,000 myeloblasts but only 33 reactive lymphocytes (Fig. 1b).

Technical Validation

All data in the MLL23 dataset originate from routine diagnostics at the Munich Leukemia Laboratory (MLL), one of Europe's largest reference centers for hematologic malignancies. As part of the standard diagnostic workflow, all cytological preparation and image acquisition is subject to stringent internal quality control and external benchmarking, including regular participation in inter-laboratory ring trials and accreditation processes. Each image was labeled by one of five expert examiners at MLL, assigning single cells to one of 18 morphologically defined classes.

A limitation of the MLL23 dataset is the natural rarity of certain cell types in peripheral blood samples. Because these minority cell types occur infrequently under both normal and pathological conditions, we cannot increase their representation during data collection. This biological constraint directly results in class imbalance, which reflects real-world distributions but poses challenges for training machine learning models on this dataset.

Data availability

The dataset is available at https://doi.org/10.5281/zenodo.14277609. It comprises 18 ZIP files, each named after a specific cell type (e.g., basophil.zip). Each ZIP file contains high-quality TIFF images of individual cells belonging to the corresponding class, with file names following a consistent format that includes the class name and a unique identifier (e.g., basophil_0001.TIF).

Code availability

No custom code was used in this study. All analyses were performed without the need for proprietary or bespoke software.

Received: 16 February 2025; Accepted: 30 October 2025;
Published online: 11 November 2025

References

1. Fuchs, T. J. & Buhmann, J. M. Computational pathology: challenges and promises for tissue analysis. *Comput. Med. Imaging Graph.* **35**, 515–530 (2011).
2. Walter, W. *et al.* Artificial intelligence in hematological diagnostics: Game changer or gadget? *Blood Rev.* **58**, 101019 (2023).
3. Matek, C., Schwarz, S., Spiekermann, K. & Marr, C. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nat Mach Intell* **1**, 538–544 (2019).
4. Hehr, M. *et al.* Explainable AI identifies diagnostic cells of genetic AML subtypes. *PLOS Digit Health* **2**, e0000187 (2023).
5. Salehi, R. *et al.* Unsupervised Cross-Domain Feature Extraction for Single Blood Cell Image Classification. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022* 739–748, https://doi.org/10.1007/978-3-031-16437-8_71 (Springer Nature Switzerland, Cham, 2022).
6. Sadafi, A. *et al.* Pixel-Level Explanation of Multiple Instance Learning Models in Biomedical Single Cell Images. in *International Conference on Information Processing in Medical Imaging – IPMI 2023* 170–182, https://doi.org/10.1007/978-3-031-34048-2_14 (Springer Nature Switzerland, Cham, 2023).
7. Pohlkamp, C. *et al.* Machine Learning (ML) Can Successfully Support Microscopic Differential Counts of Peripheral Blood Smears in a High Throughput Hematology Laboratory. *Blood* **136**, 45–46 (2020).
8. Sidhom, J. W. *et al.* Deep learning for distinguishing morphological features of acute Promyelocytic Leukemia. *Blood* **136**, 10–12 (2020).
9. Acevedo, A., Alférez, S., Merino, A., Puigví, L. & Rodellar, J. Recognition of peripheral blood cell images using convolutional neural networks. *Comput. Methods Programs Biomed.* **180**, 105020 (2019).
10. Matek, C., Krappe, S., Münzenmayer, C., Haferlach, T. & Marr, C. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood* **138**, 1917–1927 (2021).
11. Eckardt, J. N. *et al.* Deep learning detects acute myeloid leukemia and predicts NPM1 mutation status from bone marrow smears. *Leukemia* **36**, 111–118 (2022).
12. Eckardt, J. N. *et al.* Deep learning identifies Acute Promyelocytic Leukemia in bone marrow smears. *BMC Cancer* **22**, 201 (2022).
13. Shetab Boushehri, S., Qasim, A. B., Waibel, D., Schmich, F. & Marr, C. Systematic Comparison of Incomplete-Supervision Approaches for Biomedical Image Classification. in *Artificial Neural Networks and Machine Learning – ICANN 2022* 355–365, https://doi.org/10.1007/978-3-031-15919-0_30 (Springer International Publishing, 2022).

14. Umer, R. M., Gruber, A., Boushehri, S. S., Metak, C. & Marr, C. Imbalanced Domain Generalization for Robust Single Cell Classification in Hematological Cytomorphology. *ICLR 2023 Workshop on Domain Generalization* (2023).
15. Deutges, M., Sadafi, A., Navab, N. & Marr, C. Neural cellular automata for lightweight, robust and explainable classification of white blood cell images. in *Lecture Notes in Computer Science* 693–702, https://doi.org/10.1007/978-3-031-72384-1_65 (Springer Nature Switzerland, Cham, 2024).
16. Koch, V. *et al*. DinoBloom: A foundation model for generalizable cell embeddings in hematology. in *International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024* 520–530, https://doi.org/10.1007/978-3-031-72390-2_49 (Springer Nature Switzerland, Cham, 2024).
17. Sadafi, A. *et al*. A continual learning approach for cross-domain white blood cell classification. *MICCAI Workshop on Domain Adaptation and Representation Transfer*, https://doi.org/10.1007/978-3-031-45857-6_14 (Springer Nature Switzerland, Cham, 2023).

Acknowledgements

The authors thank Xudong Sun, Matthias Hehr, Sophia J. Wagner, Valentin Koch, and Matteo Wohlrapp (all from Munich) for fruitful discussions. SSB has received funding from F. Hoffmann-la Roche LTD (no grant number is applicable). SSB and SK are supported by the Helmholtz Association under the joint research school 'Munich School for Data Science - MUDS.' AG has received funding from the LMU Munich Faculty of Medicine. CM has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 866411 & 101113551 & 101213822) and acknowledges support from the Hightech Agenda Bayern.

Author contributions

Ch.M. conceived the project idea with C.M. S.S.B. and A.G. performed the data cleaning, wrote the manuscript, and designed the figures with C.M. C.M. supervised the study with K.S. S.K. helped with the manuscript consistency and edits. C.P. and T.H. performed main data collection, annotation, and pseudonymization.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Carsten Marr.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.