

# Human-level recognition of blast cells in acute myeloid leukemia with convolutional neural networks

Christian Matek<sup>1, 2</sup>, Simone Schwarz<sup>2</sup>, Karsten Spiekermann<sup>\*2</sup>, and Carsten Marr<sup>\*†1</sup>

<sup>1</sup>Institute of Computational Biology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany

<sup>2</sup>Department of Internal Medicine III, University Hospital Munich, Ludwig-Maximilians-Universität München - Campus Großhadern, Munich, Germany

## Abstract

Reliable recognition of malignant white blood cells is a key step in the diagnosis of hematologic malignancies such as Acute Myeloid Leukemia. Microscopic morphological examination of blood cells is usually performed by trained human examiners, making the process tedious, time-consuming and hard to standardise.

We compile an annotated image dataset of over 18,000 white blood cells, use it to train a convolutional neural network for leukocyte classification, and evaluate the network's performance. The network classifies the most important cell types with high accuracy. It also allows us to decide two clinically relevant questions with human-level performance, namely (i) if a given cell has blast character, and (ii) if it belongs to the cell types normally present in non-pathological blood smears.

Our approach holds the potential to be used as a classification aid for examining much larger numbers of cells in a smear than can usually be done by a human expert. This will allow clinicians to recognize malignant cell populations with lower prevalence at an earlier stage of the disease.

## Introduction

Microscopic examination and classification of blood cells is an important cornerstone of hematological diagnostics [1, 2, 3]. Specifically, morphological evaluation of leukocytes from peripheral blood or bone marrow samples is one of the initial steps in the diagnosis of hematopoietic malignancies such as Acute Myeloid Leukemia (AML) [4, 5]. In particular, the common French-American-British (FAB) classification of AMLs strongly relies on cytomorphology [6]. Having been part of routine workup of hematological diagnosis since the 19th century, cytomorphological examination of leukocytes has so far defied automation and is regularly performed by trained human experts. Therefore, cytomorphological classification is tedious and time-consuming to produce, suffering from

---

\*Joint corresponding authors.

†to whom correspondence and offprint requests should be addressed under: Helmholtz Zentrum München German Research Center for Environmental Health Ingolstädter Landstrasse 1, 85764 Neuherberg; Phone + 49 (0) 89 3187 (0), Fax + 49 (0) 89 3187 3322, Email: [carsten.marr@helmholtz-muenchen.de](mailto:carsten.marr@helmholtz-muenchen.de)

considerable intra- and inter-observer variation that is difficult to account for, and hard to deliver in situations where trained experts are lacking. Furthermore, it is difficult to reliably correlate with the result of other, intrinsically more quantitative diagnostic modalities such as immunophenotyping or molecular genetics. Reliable, automated differentiation of cell morphology and recognition of malignant cells is also a key prerequisite to allow screening for hematological neoplasms, potentially enabling their earlier detection and treatment.

As cytomorphological examination is based on evaluating microscopic cell images, it can be formulated as an image classification task. Deep convolutional neural networks (CNNs) have proven very successful in the field of natural image classification [7, 8, 9]. Recently, CNNs have been successfully applied to various medical imaging tasks, including skin cancer recognition [10], evaluation of retinal disorders [11] and the analysis of histological sections [12, 13], for example through mitosis detection [14], region of interest detection and analysis [15] or tissue type segmentation [16]. This motivates us to apply CNNs to cytomorphological classification of blood cells, in particular those relevant in AML.

Previous work on leukocyte classification has mainly been focused on feature extraction from cytological images [17, 18]. In that context, lymphoblastic leukemias, where the cytomorphology is less diverse than in the myeloid case, have received more attention [19, 20]. Providing sufficiently many labelled images for deep learning methods to work has proven challenging in medical image analysis, due to restrictions on access to and expense of expert time for providing ground truth annotations [21, 22]. Therefore, most studies have worked on datasets limited in the number of patients included or individual cytological images classified [23, 24]. Hence, applications of CNNs to white blood cell classification have so far been focused on differentiation of specific subtypes such as erythroid and myeloid precursors [23].

Here, we introduce a database comprising 18,365 individual cell images from 200 individuals, and develop a CNN that is able to classify individual cells from peripheral blood smears and judge for malignancy with high accuracy.

## Materials and Methods

We selected peripheral blood smears from 100 patients diagnosed with different subtypes of AML at the Laboratory of Leukemia Diagnostics at Munich University Hospital between 2014 and 2017, and smears from 100 patients found to exhibit no morphological features of hematological malignancies in the same time frame. The study setup was reviewed by the local ethics committee, and consent was obtained under reference number 17-349.

For all selected blood smear images, we followed the workflow depicted schematically in Fig. 1: An area of interest comprising approximately  $20 \text{ mm}^2$  within the monolayer region of the smear was selected from a low-resolution pre-scan, and scanned at 100-fold optical magnification with oil immersion using an M8 digital microscope / scanner (Precipoint GmbH, Freising, Germany). The resulting digitised data consisted of multi-resolution pyramidal images of a size of approximately 1 GB per scanned area of interest. A trained examiner experienced in routine cytomorphological diagnostics at Munich University Hospital differentiated physiological and pathological leukocyte types contained in the microscopic scans into the classification scheme shown in Fig. 2B, which is derived

from standard morphological categories and was refined to take into account subcategories relevant for the morphological classification of AML, such as bilobed Promyelocytes, which are typical of the FAB subtype M3v [3]. Annotation was carried out on a single-cell basis, and approximately 100 cells were differentiated in each smear. Subimage patches of size 400 x 400 pixels (corresponding to approximately  $29\mu\text{m} \times 29\mu\text{m}$ ) around the annotated cells were extracted without further cropping or filtering, including background components such as erythrocytes, platelets or cell fragments. When examining the screened blood smears, the cytologist followed the routine clinical procedure. Overall, 18,365 single-cell images were annotated and cut out of the scan regions. The full class-wise statistics of the dataset are given in Tab. S1, and sample images of the most important physiological and pathological classes are shown in Fig. 2A and C. The database of single-cell images is provided online.

Annotations of single-cell images provide the ground truth for training and evaluation of our network. Morphological classes containing fewer than 10 images were merged with neighbouring classes of the taxonomy. Specifically, myeloblasts with and without Auer rods were merged into a common myeloblast class, and faggot cells and promyelocytes with and without Auer rods were merged into a common promyelocyte class, resulting in 15 classes for training and evaluation. A subset of 1,905 single-cell images from all morphological categories were presented to a second, independent examiner, and annotated for a second time in order to estimate inter-rater variability.

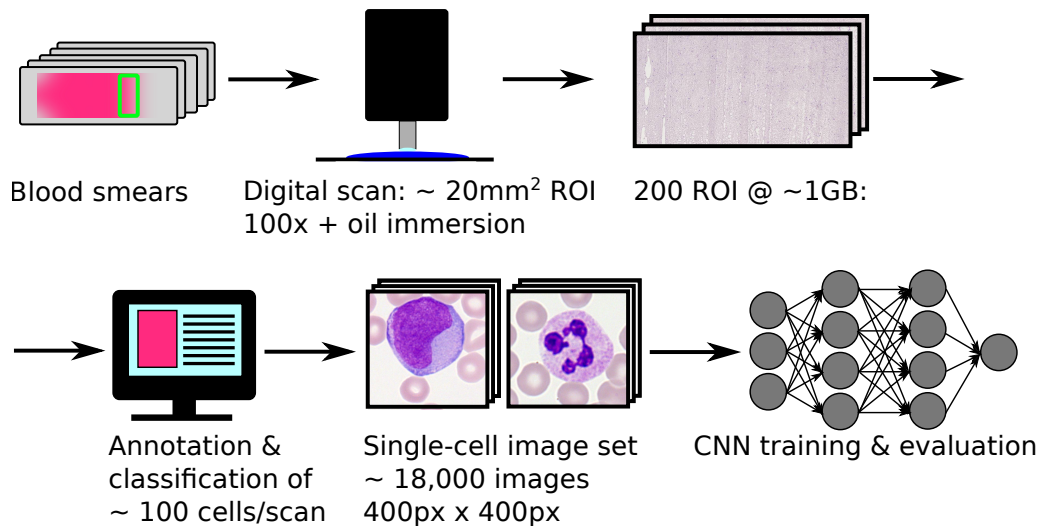
For our image classification task, we used a ResNeXt CNN topology described by Xie *et al.* [25], which derives from a residual network, and achieved a second rank in the classification task of the ImageNet ILSVRC 2016 competition. While several versions of residual networks have been shown to be successful in natural image classification, ResNeXt is characterised by a comparatively small number of free hyper-parameters, and is therefore expected to be a convenient choice, in particular as no networks pre-trained on similar datasets are available. We adopted the network to input image dimensions of 400 x 400 x 3 and retained the cardinality hyper-parameter at  $C = 32$  as used in Ref. [25], avoiding further tuning of hyper-parameters based on our dataset. The ResNeXt network was implemented using the implementation of Ref. [26] for Keras 2.0 [27], with the input size adjusted to accept images of size  $400 \times 400$  pixels, and the final dense layer adapted to our 16-category classification scheme.

The network was trained for at least 20 epochs, which took a computing time of approximately 4 days on a Nvidia GeForce GTX TITAN X GPU.

We randomly divide the images contained in each class of our dataset in a test- and training group, where the training group contains approximately 80%, and the test group 20% of the images. For 5-fold cross-validation, we performed a stratified split of the cell images into 5 folds, where each fold contains approximately 20% of the images in each class. An individual image is contained in one fold only. Consequently, five different models are trained, each of which uses one fold for testing, and the four remaining folds for training.

In order to cope with the imbalance of cell numbers contained in different classes and take advantage of the rotational invariance of the cell classification problem, we generated additional images by applying random rotational transformations of 0–359 degrees, as well as random horizontal and vertical flips to the single-cell images in our dataset. Using these operations, we augmented the data set in such a way that each class contained approximately 10,000 images for training.

To quantitatively evaluate the class-wise accuracy, we calculate precision, specificity



**Figure 1: Data handling workflow.** Peripheral blood smears from 100 AML patients and 100 patients without signs of hematological malignancy were digitised using an oil-immersion microscope at 100-fold magnification. After annotation of single cells using the classification scheme shown in Fig. 2B, a convolutional neural network was trained and evaluated.

and sensitivity as comparison metrics, which are defined as follows:

$$\text{sensitivity} = \frac{\text{true positive}}{\text{positive}} \quad (1)$$

$$\text{specificity} = \frac{\text{true negative}}{\text{negative}} \quad (2)$$

$$\text{precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}} \quad (3)$$

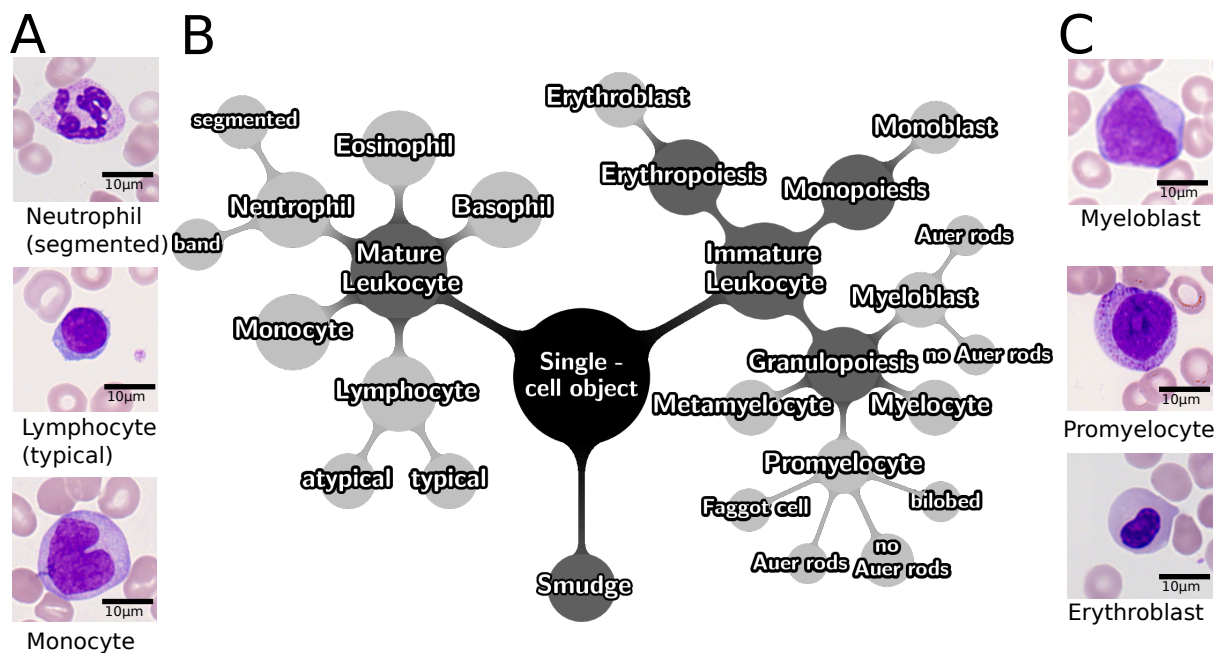
where “true positive” and “true negative” are the number of images correctly ascribed or not ascribed to a given class by the network, respectively, “positive” and “negative” are the overall number of images shown belonging or not belonging to a certain class, and “false positive” is the number of cell images wrongly ascribed to that class.

## Data and code availability

Code for the network trained in this study and network weights are available at CodeOcean, together with a subset of the single-cell image data used to test the network. Additional image data may be available from the authors upon reasonable request, with permission from Munich University Hospital.

## Results

We evaluate the performance of the trained network by feeding single-cell images through it, and comparing the output prediction with the labels assigned by human examiners. The network outputs a vector of probabilities  $\mathbf{P} = (P_1, \dots, P_i, \dots, P_{15})$ , where the components  $P_i$  are the respective predicted probabilities for the image to belong to class  $i$  out of the 15 overall classes. The network’s image class prediction is then the class  $m$  with the highest corresponding predicted probability  $P_m$ .



**Figure 2: Classification of 18,000 single-cell images into a 18-class scheme.**

(A): Sample images of the three most frequent mature cell classes contained in the dataset. Scale bars correspond to  $10\ \mu m$ . (B): Taxonomy tree of the classification scheme used for annotation of single-cell images. Cell morphologies present under physiological conditions are contained in the left branch, while cells normally absent under physiological conditions are contained in the right branch. Smudge cells are classified in a separate category. To ensure sufficient population of classes, some leaves of the taxonomy were combined into overall classes for training and testing (cf. main text). (C): Sample images of the three most frequent immature cell classes contained in the dataset. Scale bars correspond to  $10\ \mu m$ .

The class-wise prediction accuracy of the network is shown in the confusion matrix of Fig. 3A. We note that the network achieves excellent agreement with human annotations (each above 90%) for the most common physiological cell types, including segmented neutrophils, typical lymphocytes, monocytes, and eosinophils. Also myeloblasts, whose presence in the peripheral blood is common in myeloid leukemias [5], are recognised with high accuracy, yielding a precision and sensitivity of 94% (cf. Tab. 1). Other classes are more challenging for the network, in particular the intermediate stages of granulopoiesis and erythropoiesis, and basophils, where our test and training dataset contains less than 100 images. Values of precision and sensitivity for all cell classes obtained by 5-fold cross-validation are given in Tab. 1. We note that due to the varying number of specific cell types present in smears, the number of test and training images varies by up to two orders of magnitude for different classes. In order to avoid biased evaluation of our classifier, we refer to class-wise precision and sensitivity, and do not evaluate an overall accuracy score, which would be biased towards the classes with a high number of samples [7].

In order to relate the network performance to the inter-rater variability encountered among human examiners, we asked another, independent cytologist to re-annotate a subset of 1,905 single-cell images containing all subtypes previously annotated. The level of agreement between the two annotators is shown in Fig. S1. Notably, the CNN and the second annotator show similar patterns of deviation to the first examiner, specifi-



cally as far as classification of atypical lymphocytes and promyelocytes as myeloblasts is concerned. This may reflect visual similarities between instances of these cell types recognised by both the network and human examiners, which intrinsically limit the single-cell labelling process. Furthermore, the first examiner had access to the whole smear and was hence able to compare to other cells present, whereas the second examiner, just as the network, classified only single-cell images.

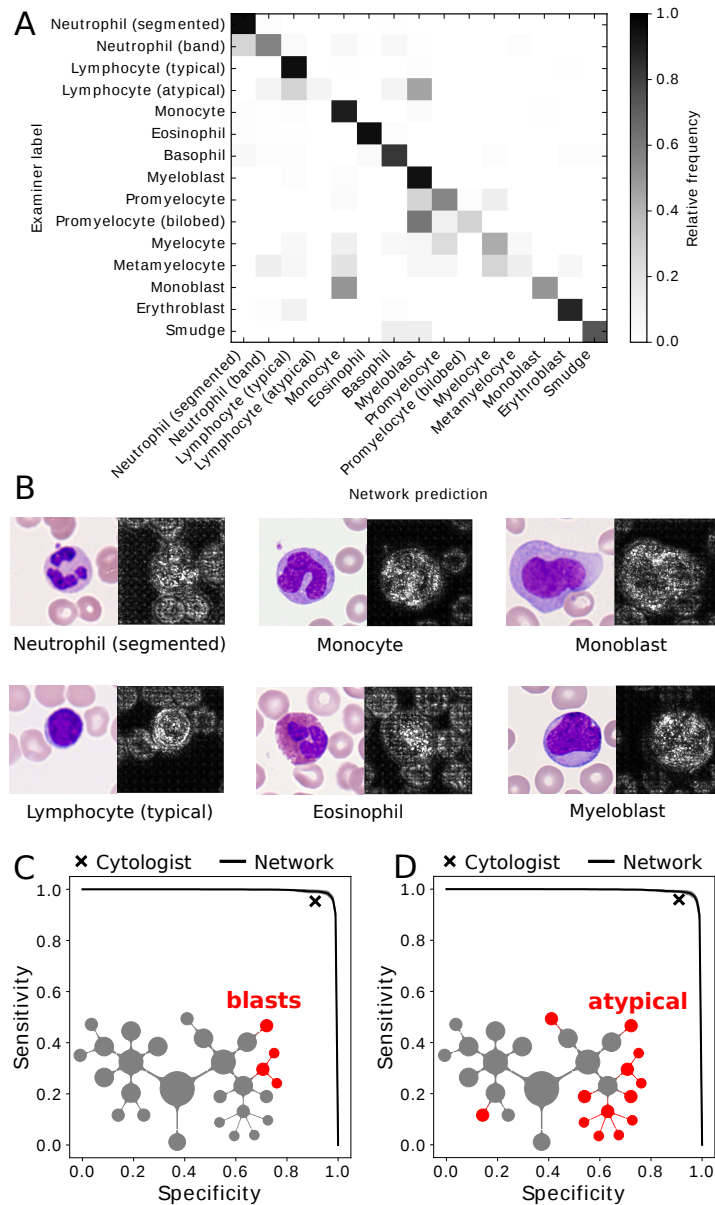
To evaluate if our network fixates on relevant parts of the single-cell images, we calculated saliency maps following the procedure outlined by Simonyan *et al.* [28]. These maps allow visualising how important a given pixel is for the network’s classification decision. Saliency maps for several test images are shown in Fig. 3B, demonstrating that pixels within the leukocyte are most important for the network’s classification decision, suggesting that the network has learned to focus on relevant areas of the input image.

A key clinical question when examining blood cell morphology is if a given cell is a myeloblast or monoblast, as these two cell types are counted as blast equivalents, and are generally required to be present in the peripheral blood for a diagnosis of AML [5]. Using the output of our network, we can determine the probability of a cell to possess blast character,  $P_{\text{blast}} = P_{\text{myeloblast}} + P_{\text{monoblast}}$ , and choose a threshold probability  $t$ , so that the binary prediction of the network is given by  $\hat{y} = P_{\text{blast}} \geq t$ . The receiver operating characteristic (ROC) curve is the result of sweeping  $t$  between 0 and 1, and is shown in Fig. 3C. The area under the curve (AUC), which we measure as  $0.992 \pm 0.001$  using 5-fold cross-validation, shows that our network provides a test of the blast character of a given single-cell image that can be considered outstanding by the criteria of diagnostic test assessment [29]. In comparison to the network’s performance, the second human rater reproduces the cell label provided by the first annotator with a sensitivity of 95.3% and a specificity of 91.1% (cf. Fig. 3C), which lies close but somewhat below the network ROC curve.

Another clinically important, binary decision on individual white blood cells is whether a given cell belongs to one of the typical cell types, present in peripheral blood under normal circumstances, or to atypical cell types that occur in pathological situations, namely myeloblasts, monoblasts, myelocytes, metamyelocytes, promyelocytes, erythroblasts and atypical lymphocytes. As in the test for blast character, we determine the overall probability  $P_{\text{atypical}}$  for a given cell to belong to one of these groups by adding the output probabilities of all atypical cell classes, and defining a threshold probability  $t$  for the atypicality test to be positive. Again, the network yields an outstanding test of the atypicality of a cell in a given image, with an AUC of  $0.991 \pm 0.002$ , which compares to a human second-rater sensitivity of 95.9% and specificity of 91.0% (Fig. 3D). Importantly, our network outperforms the cytologist who re-annotated the subset of 1,905 single-cell images in both clinically relevant binary decision tasks tested here (cf. Fig. 3C and D).

## Discussion

The convolutional neural network presented in this study shows Outstanding performance at identifying the most important morphological white blood cell types present in non-pathological blood, as well as the key pathological cell types in Acute Myeloid Leukemias. For the most common physiological leukocyte classes as well as for myeloblasts, it attains precision and sensitivity values above 90%, allowing these cells to be identified with a very high accuracy which outperforms other classifiers in the literature [23, 18]. The



**Figure 3: Human-level network performance in single-cell classification, pixel-wise attention and binary decision tasks.**

(A): Class-wise agreement between network prediction and ground-truth human examiner label. Outstanding classification performance is observed for key cell classes such as myeloblasts. The full, class-wise statistics of prediction quality is given in Tab. 1. (B): Saliency maps illustrate the gradient of a pixel with respect to the networks’s loss function. Brighter pixels have a higher influence on the network’s classification decision. Maps suggest that the network has learned to focus on the leukocyte and map out its internal structures, while giving less weight to background content. (C): ROC curve of binary decision between cells with and without blast character. The network performs very well with an AUC of  $0.992 \pm 0.001$  (obtained by training  $n = 5$  networks for cross-validation, each tested on a separate 20% of overall image data), and attains the performance of a second human annotator shown by 'x'. Inset: Schematic depiction of cell class taxonomy, where blast equivalent classes are coloured red. (D): On the binary classification for atypical cells normally absent in non-pathological smears, outstanding performance is observed with an AUC of  $0.991 \pm 0.002$ . Also for this binary question, the network attains human-level performance. Inset: Schematic depiction of cell class taxonomy, with atypical cell classes coloured red.

	Class	Precision	Sensitivity	Number of images
Mature leukocytes	Neutrophil (segmented)	$0.99 \pm 0.00$	$0.96 \pm 0.01$	8,484
	Neutrophil (band)	$0.25 \pm 0.03$	$0.59 \pm 0.16$	109
	Lymphocyte (typical)	$0.96 \pm 0.01$	$0.95 \pm 0.02$	3,937
	Lymphocyte (atypical)	$0.20 \pm 0.4$	$0.07 \pm 0.13$	11
	Monocyte	$0.90 \pm 0.04$	$0.90 \pm 0.05$	1,789
	Eosinophil	$0.95 \pm 0.04$	$0.95 \pm 0.01$	424
	Basophil	$0.48 \pm 0.16$	$0.82 \pm 0.07$	79
Immature leukocytes	Myeloblast	$0.94 \pm 0.01$	$0.94 \pm 0.02$	3,268
	Promyelocyte	$0.63 \pm 0.16$	$0.54 \pm 0.20$	70
	Promyelocyte (bilobed)	$0.45 \pm 0.32$	$0.41 \pm 0.37$	18
	Myelocyte	$0.46 \pm 0.19$	$0.43 \pm 0.07$	42
	Metamyelocyte	$0.07 \pm 0.13$	$0.13 \pm 0.27$	15
	Monoblast	$0.52 \pm 0.30$	$0.58 \pm 0.26$	26
	Erythroblast	$0.75 \pm 0.20$	$0.87 \pm 0.09$	78
Smudge cell	$0.53 \pm 0.28$	$0.77 \pm 0.20$	15	
Total				18,365

**Table 1: Class-wise precision and sensitivity of the network, determined by 5-fold cross-validation.**

The model achieves precision and sensitivity above 0.9 on classes for which more than 100 images are available, such as segmented neutrophils, typical lymphocytes and myeloblasts. Large deviations across folds occur for classes with small sample number, e.g. metamyelocytes and promyelocytes.

classification predictions can be used to answer clinically relevant binary questions. Blast character and atypicality are of high relevance in practice and can be provided with very high confidence by the network.

As expected for our data-driven classification method, a correlation can be observed between the number of images available for a specific class in our data set and the performance of the network on that class (cf. Tab. 1). Although to the authors' knowledge, the image set presented in this paper is the largest used so far in the literature, we anticipate that further enlarging the dataset would improve the network's classification performance also for rare cell types.

Sources of disagreement between the network and the ground truth are linked to the inter-rater variability of the cytomorphologic examination, which is known to limit the reproducibility particularly of rare leukocyte species [30]. We have estimated inter-observer variability of cytomorphologic classification in our dataset by re-annotation of single-cell images. We note that the inter-rater variability is particularly high for myeloblasts, reflecting to some degree the polymorphic nature of that cell class. Furthermore, the examiner providing the ground-truth labels for single-cell images had access to the whole scan and was therefore able to compare the morphologies of cells present in a smear. In contrast, both the second annotator and the network had to make a classification decision based solely on a single-cell image, without the ability to compare to other cells from the same patient.

We have compiled a dataset of 18,365 single-cell images of different cell morphologies relevant in the diagnosis of AML from the peripheral blood smears of 200 individuals. After annotation by human experts, we used this dataset to train and evaluate a state-of-the-art image classification convolutional neural network. The network shows good performance at differentiating morphological cell types important for recognising malignancy in peripheral blood smears. For the diagnostically relevant binary questions if a



given cell is considered to have blast character or to be atypical, the network achieves outstanding accuracies with an ROC AUC of approximately 0.99 in both cases. Given this level of performance and the fact that our method is scalable and fast, our algorithm can be used to quickly evaluate thousands of cells on a blood smear scan, helping cytologists to find suspicious cells more readily. This might be particularly useful in situations where the number of malignant cells is expected to be small, such as in the early stages of the disease or beginning relapse. In our present paper, we tested the model on peripheral blood smears from one lab scanned with one type of scanning device. In order to evaluate the model's performance in a realistic routine setting in more detail, further validation is required using data from different sources and disease classes. However, given the variability already present in a dataset compiled from 200 independently stained smears from different patients, we expect the variability to be represented reasonably well in our dataset.

Our method holds the potential to act as a rapid pre-screening and quantitatively informed decision tool for cytological examiners, and might further increase its performance when combined with additional, intrinsically quantitative methods used in the diagnosis of hematological malignancies, such as flow cytometry or molecular genetics.

## Acknowledgements

We thank Nikos Chlis for comments on the manuscript. This work was supported by German Research Foundation grant SFB 1243. Christian Matek gratefully acknowledges support from Deutsche José Carreras-Leukämie Stiftung.

## References

- [1] B. J. Bain. Diagnosis from the blood smear. *The New England journal of medicine*, 353:498–507, August 2005.
- [2] D. C. Tkachuk and J. V. Hirschmann. *Wintrobe's Atlas of Clinical Hematology*. Lippincott Raven, 2006.
- [3] H. Thöml, H. Diem, and T. Haferlach. *Color Atlas of Hematology*. Thieme, 2004.
- [4] M. F. Fey and C. Buske. Acute myeloblastic leukaemias in adult patients: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. 24:vi138–vi143, 2013.
- [5] H. Döhner, E. Estey, D. Grimwade, S. Amadori, F. R. Appelbaum, T. Büchner, H. Dombret, B. L. Ebert, P. Fenaux, R. A. Larson, R. L. Levine, F. Lo-Coco, T. Naoe, D. Niederwieser, G. J. Ossenkoppele, M. Sanz, J. Sierra, M. S. Tallman, H. Tien, A. H. Wei, B. Löwenberg, and C. D. Bloomfield. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*, 129:424–447, January 2017.
- [6] J. M. Bennett, D. Catovsky, M. T. Daniel, G. Flandrin, D. A. Galton, H. R. Gralnick, and C. Sultan. Proposed revised criteria for the classification of Acute Myeloid Leukemia. A report of the French-American-British Cooperative Group. *Annals of internal medicine*, 103:620–625, October 1985.

- [7] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.
- [8] W. Rawat and Z. Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29:2352–2449, September 2017.
- [9] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [10] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, February 2017.
- [11] P. Eulenberg, N. Köhler, T. Blasi, A. Filby, A. E. Carpenter, P. Rees, F. J. Theis, and F. A. Wolf. Reconstructing cell cycle and disease progression using deep learning. *Nature communications*, 8:463, September 2017.
- [12] A. Janowczyk and A. Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7:29, 2016.
- [13] T. J. Fuchs and J. M. Buhmann. Computational pathology: challenges and promises for tissue analysis. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 35:515–530, 2011.
- [14] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab. Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE transactions on medical imaging*, 35(5):1313–1321, 2016.
- [15] R. M. Levenson, A. Fornari, and M. Loda. Multispectral imaging and pathology: seeing and doing more. *Expert opinion on medical diagnostics*, 2:1067–1081, 2008.
- [16] A. Gertych, N. Ing, Z. Ma, T. J. Fuchs, S. Salman, S. Mohanty, S. Bhele, A. Velásquez-Vacca, M. B. Amin, and B. S. Knudsen. Machine learning approaches to analyze histological images of tissues from radical prostatectomies. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 46:197–208, 2015.
- [17] L. Bigorra, A. Merino, S. Alférez, and J. Rodellar. Feature analysis and automatic identification of leukemic lineage blast cells and reactive lymphoid cells from peripheral blood cell images. *Journal of clinical laboratory analysis*, 31(2), 2017.
- [18] S. Krappe, T. Wittenberg, T. Haferlach, and C. Münzenmayer. Automated morphological analysis of bone marrow cells in microscopic images for diagnosis of leukemia: nucleus-plasma separation and cell classification using a hierarchical tree model of hematopoiesis. *Bildverarbeitung für die Medizin 2016 : Algorithmen - Systeme - Anwendungen; Proceedings des Workshops vom 13. bis 15. März 2016 in Berlin*, 2016.
- [19] F. Scotti. Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. In *Computational Intelligence for Measurement Systems and Applications, 2005. CIMSA. 2005 IEEE International Conference on*, pages 96–101. IEEE, 2005.

- [20] S. Mohapatra, D. Patra, and S. Satpathy. An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. *Neural Computing and Applications*, 24(7-8):1887–1904, 2014.
- [21] H. Greenspan, B. van Ginneken, and R. M. Summers. Deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.
- [22] D. Shen, G. Wu, and H. Suk. Deep Learning in Medical Image Analysis. *Annual review of biomedical engineering*, 19:221–248, June 2017.
- [23] J. W. Choi, Y. Ku, B. W. Yoo, J.-A. Kim, D. S. Lee, Y. J. Chai, H.-J. Kong, and H. C. Kim. White blood cell differential count of maturation stages in bone marrow smear using dual-stage convolutional neural networks. *PloS one*, 12:e0189259, 2017.
- [24] P. Kainz, H. Burgsteiner, M. Asslaber, and H. Ahammer. Training echo state networks for rotation-invariant bone marrow cell classification. *Neural Computing and Applications*, 28(6):1277–1292, 2017.
- [25] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks.
- [26] M. Dietz. ResNeXt implementation for Keras. <https://gist.github.com/mjdietzx/>, 2017.
- [27] F. Chollet et al. Keras 2.0. <https://keras.io>, 2017.
- [28] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.
- [29] J. N. Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 2010.
- [30] X. Fuentes-Arderiu and D. Dot-Bach. Measurement uncertainty in manual differential leukocyte counting. *Clinical chemistry and laboratory medicine*, 47:112–115, 2009.