ARTICLES
https://doi.org/10.1038/s42256-019-0101-9
nature machine intelligence

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

Christian Matek[1,2], Simone Schwarz[2], Karsten Spiekermann[2,3,4,5,6*] and Carsten Marr[1,5,6*]

Reliable recognition of malignant white blood cells is a key step in the diagnosis of haematologic malignancies such as acute myeloid leukaemia. Microscopic morphological examination of blood cells is usually performed by trained human examiners, making the process tedious, time-consuming and hard to standardize. Here, 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 by comparing to inter- and intra-expert variability. 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: (1) if a given cell has blast character and (2) 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.

Microscopic examination and classification of blood cells is an important cornerstone of haematological diagnostics[1–3]. Specifically, morphological evaluation of leukocytes from peripheral blood or bone marrow samples is one of the initial steps in the diagnosis of haematopoietic malignancies such as acute myeloid leukaemia (AML)[4–6]. In particular, the common French–American–British (FAB) classification of AMLs strongy relies on cytomorphology[7]. Having been part of routine work-up of haematological diagnosis since the nineteenth century, cytomorphological examination of leukocytes has so far defied automation and is usually performed by trained human experts. Cytomorphological classification is thus tedious and time-consuming to produce, suffers from considerable intra- and inter-observer variation[8–10] that is difficult to account for, and is hard to deliver in situations where trained experts are lacking. Furthermore, it is difficult to reliably correlate with the results of other, intrinsically more quantitative diagnostic modalities such as immunophenotyping or molecular genetics[4]. An automated approach would allow for consistent classification of cytomorphologies, removing the intrinsically subjective human element of the process. Reliable, automated differentiation of cell morphology and recognition of malignant cells is also a key prerequisite to allow screening for haematological 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[11–13]. Recently, CNNs have been applied to various medical imaging tasks, including skin cancer recognition[14], evaluation of retinal disorders[15] and the analysis of histological sections[16,17], for example through mitosis detection[18], region of interest detection and analysis[19] or tissue type segmentation[20]. This motivated us to apply CNNs to cytomorphological classification of blood cells, in particular those relevant in AML.

Previous work on leukocyte classification has mainly focused on feature extraction from cytological images[21,22]. In that context, lymphoblastic leukaemias, where the cytomorphology is less diverse than in the myeloid case, have received more attention[23,24]. Providing a sufficient number of labelled images for deep learning methods to work has proven challenging in medical image analysis due to restrictions on access to and the expense of expert time for providing ground truth annotations[25,26]. Therefore, most studies have worked on datasets limited in the number of included patients or annotated single-cell images [27,28]. Hence, applications of CNNs to white blood cell classification have so far been focused on the differentiation of specific subtypes, such as erythroid and myeloid precursors[27], or on the differentiation of non-malignant cell types[29,30].

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

Materials and methods

Dataset selection. 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 haematological malignancies in the same time frame. The study set-up was reviewed by the local ethics committee, and consent was obtained under reference number 17–349.

Dataset digitization and annotation. For all selected blood smears, we followed the workflow depicted in Fig. 1. An area of interest (AOI) comprising approximately 20 mm^2 was selected from a

[1]Institute of Computational Biology, Helmholtz Zentrum München – German Research Center for Environmental Health, Neuherberg, Germany. [2]Laboratory of Leukemia Diagnostics, Department of Medicine III, University Hospital, LMU Munich, Munich, Germany. [3]German Cancer Consortium (DKTK), Heidelberg, Germany. [4]German Cancer Research Center (DKFZ), Heidelberg, Germany. [5]Joint corresponding authors: Karsten Spiekermann, Carsten Marr. [6]These authors jointly supervised this work: Karsten Spiekermann, Carsten Marr. *e-mail: karsten.spiekermann@med.uni-muenchen.de; carsten.marr@helmholtz-muenchen.de

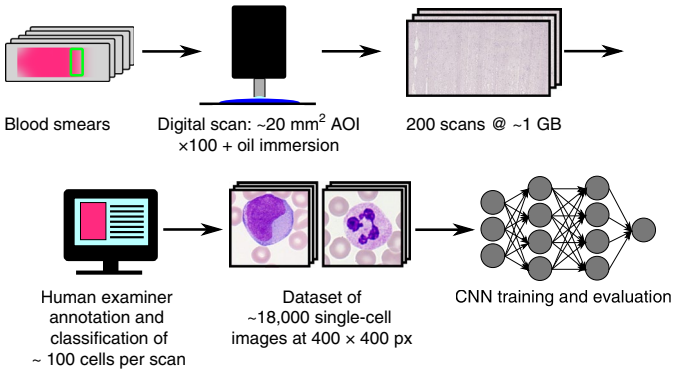

Fig. 1 | Data handling workflow. AOIs from the monolayer region of peripheral blood smears from 100 patients with AML and 100 patients without signs of haematological malignancy were digitized using an oil-immersion microscope at ×100 magnification. After annotation of single cells on the scan using the classification scheme shown in Fig. 2b, a CNN was trained and evaluated.

low-resolution pre-scan. In accordance with standard practice[31], the AOI was chosen from the so-called monolayer region, which is the part of the smear where single cells lie next to each other just without overlapping, and where cytologic features of individual cells can therefore be discerned best. Restricting the scan to the monolayer region hence enables scanning of the diagnostically relevant region of a smear (Supplementary Fig. 1). The precise area of the monolayer scanned reflects a tradeoff between scan time and number of cells imaged. The AOI thus defined was scanned at ×100 optical magnification with oil immersion, using an M8 digital microscope/scanner (Precipoint, Freising/Germany). The resulting digitized data consisted of multi-resolution pyramidal images (approximately 1 GB per scan). A trained examiner experienced in routine cytomorphological diagnostics at Munich University Hospital differentiated physiological and pathological leukocytes present in the 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 FAB subtype M3v[3]. Within each scanned AOI, about 100 leukocytes were chosen manually and annotated (with a self-written annotation software) by the trained examiner using the classification scheme (Fig. 2). This is in agreement with standard practice, where the examiner chooses a representative subsample of cells to differentiate. The examiner was asked to annotate cells that would be considered in a routine setting. Importantly, during the annotation process, the examiner had access to the whole scan when differentiating the leukocytes, allowing comparison of the different cells present. We regard annotations obtained in this way as gold standard annotation, and use them as the ground truth for training and evaluating our network. Based on the annotation, subimage patches of size 400 × 400 pixels (corresponding to approximately 29 µm × 29 µm) around the differentiated cells were extracted from the scan without further filtering or cropping, including background components such as erythrocytes, platelets and cell fragments. Overall, the annotation process resulted in a dataset of 18,365 single-cell images. The full class-wise statistics of this dataset are provided in Supplementary Table 1, and sample images of the most important physiological and pathological classes are shown in Fig. 2a,c.

Morphological classes containing fewer than 10 images were merged with neighbouring classes of the taxonomy into an overarching higher-level class. 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 (Table 1).

To estimate inter-examiner variability, a representative subset of 1,905 single-cell images from all morphological categories was presented to an independent examiner, and annotated for a second time (Supplementary Note 1). During this first re-annotation process, the second examiner only had access to the cropped single-cell images rather than the whole AOI scan, so no comparison between different cells was possible. This is in contrast to the gold standard annotation, which provides the ground truth. For an assessment of intra-examiner variability and self-consistency, we repeated the re-annotation 11 months later with the same human examiner (second re-annotation), establishing good self-agreement (Supplementary Note 2 and Supplementary Fig. 3).

ResNeXt CNN training. For our image classification task, we use a ResNeXt CNN described in ref. [32], which derives from a residual network, and achieved a second rank in the classification task of the ImageNet ILSVRC 2016 competition. Although several versions of residual networks have been shown to be successful in natural image classification, ResNeXt is characterized 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 adapted the ResNeXt CNN using the implementation of ref. [33] for Keras 2.0[34] to input dimensions of 400 × 400 × 3, retained the cardinality hyper-parameter at $C = 32$ as used in ref. [32], and adjusted the final dense layer to our 15-category scheme. Further tuning of hyper-parameters based on our dataset was avoided.

In training the ResNext network, it was found that the training loss stopped decreasing significantly after ~15 epochs. To prevent overfitting, early stopping was used beyond that number of epochs for each fold[35]. Training took a computing time of approximately four days on a Nvidia GeForce GTX Titan X graphics processing unit.

To assess the influence of the choice of a particular network architecture, we trained and tested a sequential model on the same dataset. Overall, the perfomance of that network is only mildly inferior to ResNeXt. We discuss the results of that alternative network topology in more detail in Supplementary Note 5 and Supplementary Fig. 6.

We randomly divide the images contained in each class of our dataset into a test set and a training set, with the training set containing about 80% of the images and the test set 20%. For five-fold cross-validation, we perform a random stratified split of the cell images into five 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 (Supplementary Fig. 4). In the strategy pursued here, single-cell images are divided into train and test sets, regardless of the patient from whom they are taken. This allows the splitting of cells of different classes evenly between test and training sets. Alternatively, single-cell images may be split according to the patient from which they are taken, which avoids possible correlations between single-cell images in the training and tests sets due to the fact they were taken from the same smear. We followed this alternative strategy and found similar results for both strategies (Supplementary Fig. 5), as discussed in more detail in Supplementary Note 4.

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°, as well as random horizontal and vertical flips to the single-cell images in our dataset. Using these operations, we augmented the dataset in such a way that each class contained about 10,000 images for training.

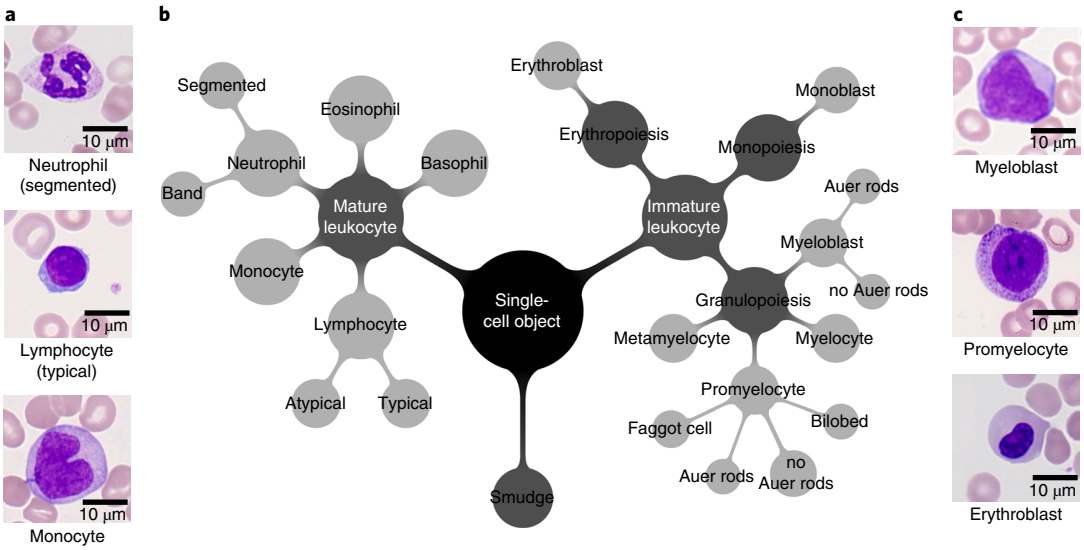

Fig. 2 | Classification of 18,000 single-cell images into an 18-class scheme. a, Sample images of the three most frequent mature cell classes contained in the dataset. **b**, Taxonomy tree of the classification scheme used for annotation of the single-cell images. Cell morphologies present in peripheral blood under physiological conditions are contained in the left branch, while cells normally absent under physiological conditions are contained in the right branch. Smudge cells, that is, remnants of leukocytes destroyed during smear production, 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 (see main text). **c**, Sample images of the three most frequent immature cell classes contained in the dataset.

Table 1 | Class-wise precision and sensitivity of the network, determined by five-fold cross-validation

Class	Precision	Sensitivity	No. of images
Mature leukocytes			
Neutrophil (segmented)	0.99 ± 0.00	0.96 ± 0.01	8,484
Neutrophil (band)	0.25 ± 0.03	0.59 ± 0.16	109
Lymphocyte (typical)	0.96 ± 0.01	0.95 ± 0.02	3,937
Lymphocyte (atypical)	0.20 ± 0.40	0.07 ± 0.13	11
Monocyte	0.90 ± 0.04	0.90 ± 0.05	1,789
Eosinophil	0.95 ± 0.04	0.95 ± 0.01	424
Basophil	0.48 ± 0.16	0.82 ± 0.07	79
Immature leukocytes			
Myeloblast	0.94 ± 0.01	0.94 ± 0.02	3,268
Promyelocyte	0.63 ± 0.16	0.54 ± 0.20	70
Promyelocyte (bilobed)	0.45 ± 0.32	0.41 ± 0.37	18
Myelocyte	0.46 ± 0.19	0.43 ± 0.07	42
Metamyelocyte	0.07 ± 0.13	0.13 ± 0.27	15
Monoblast	0.52 ± 0.30	0.58 ± 0.26	26
Erythroblast	0.75 ± 0.20	0.87 ± 0.09	78
Smudge cell	0.53 ± 0.28	0.77 ± 0.20	15
Total			18,365

The model achieves precision and sensitivity above 0.9 on classes for which more than 400 images are available, such as segmented neutrophils, typical lymphocytes and myeloblasts. Large deviations across folds occur for classes with small number of images, for example, metamyelocytes and promyelocytes.

Performance metrics. To quantitatively evaluate the class-wise prediction quality, we calculate the precision, specificity and sensitivity as comparison metrics, which are defined as follows:

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

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

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

Here '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.

Results

Multi-class prediction. We evaluate the classification performance of the trained network by feeding single-cell images through it, and comparing the output prediction with the ground-truth labels assigned by the human examiner in the gold standard annotation. The network outputs a vector of probabilities $\mathbf{P} = (P_1, \ldots, P_i, \ldots, P_{15})$, 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 prediction is the class m with the highest predicted probability P_m.

The class-wise prediction quality of the network is shown in the confusion matrix of Fig. 3a. We note that the network achieves excellent agreement with human examiner annotations for the most common physiological cell types, including segmented neutrophils, typical lymphocytes, monocytes and eosinophils (each above 90% in precision and sensitivity; Table 1). Also, myeloblasts, whose presence in the peripheral blood is common in myeloid leukaemias[4], are recognized with a very high precision and sensitivity of 94% (Table 1). Other classes are more challenging for the network, in particular the intermediate stages of granulopoiesis and

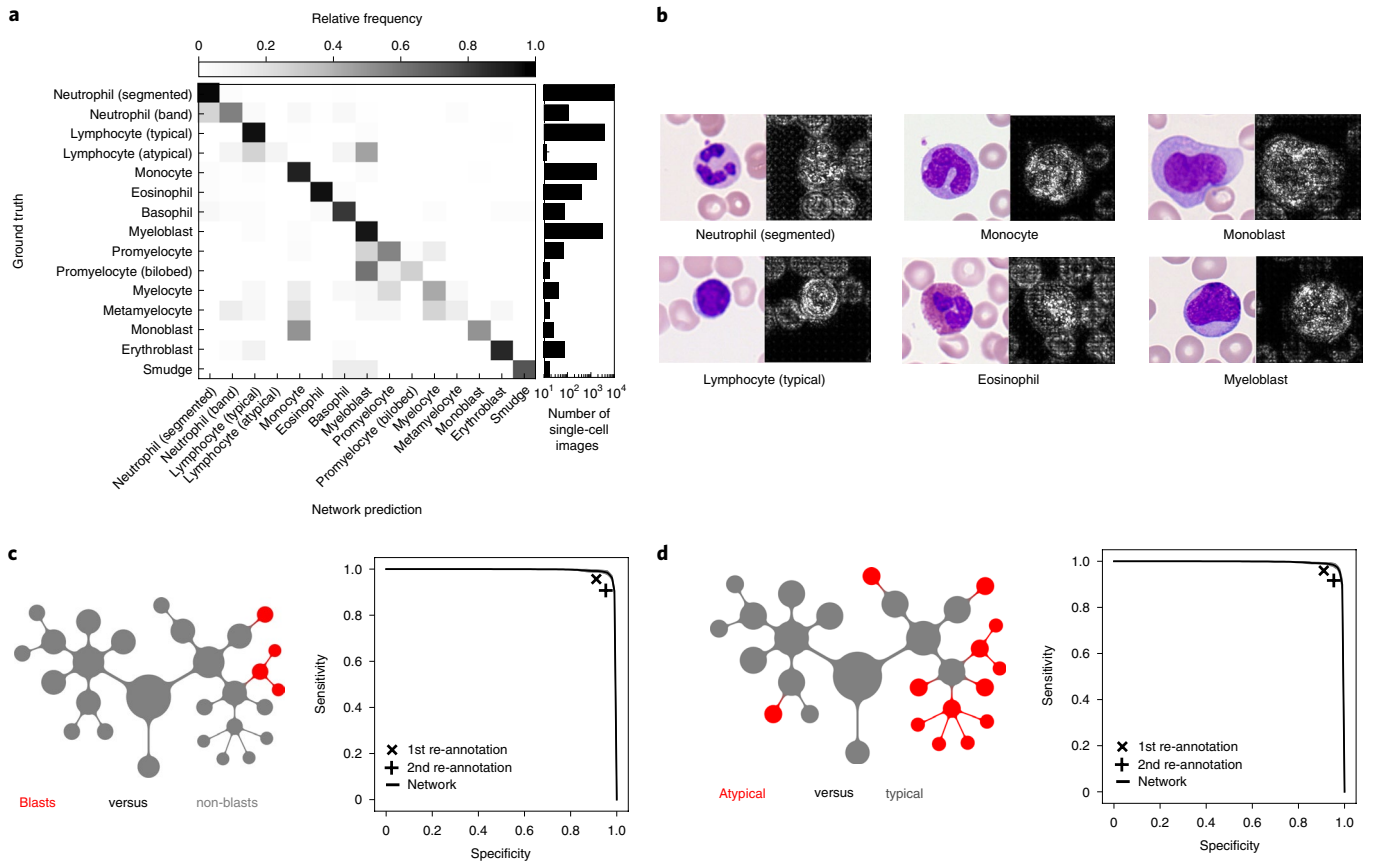

Fig. 3 | Human-level network performance in single-cell classification, pixel-wise attention and binary decision tasks. a, Confusion matrix between network prediction and ground-truth human examiner label, obtained by five-fold cross-validation. To the right of the matrix, the number of available single-cell images is indicated on a logarithmic scale. Very good performance is observed for key cell classes such as myeloblasts. The full statistics of prediction quality are provided in Table 1, and individual results for all five folds are described in Supplementary Note 3 and depicted in Supplementary Fig. 4. **b**, Saliency maps illustrate the gradient of a pixel with respect to the network's loss function. Brighter pixels have a higher influence on the network's classification decision. The maps suggest that the network learns to focus on the leukocyte and map out its internal structures, while giving less weight to background content. **c**, Left: schematic depiction of the taxonomy showing the classes involved in the blast versus non-blast binary decision. Right: receiver operating characteristic (ROC) of a binary decision between cells with and without blast character. The network performs very well with an area under the curve (AUC) of 0.992 ± 0.001. All ROCs are obtained by averaging across five folds. The network attains the performance of two independent re-annotations of single-cell images at different times by a second human examiner (indicated by 'x' and '+'; Supplementary Note 1 and Supplementary Fig. 2). **d**, Left: Schematic depiction of the taxonomy showing the classes involved in the typical versus atypical binary decision. Right: On the binary classification for atypical cells normally absent in non-pathological smears, an AUC of 0.991 ± 0.002 is observed, which lies in the domain of outstanding performance[38]. For this binary question, the network also attains human-level performance, as measured by two independent re-annotations (Supplementary Note 1).

erythropoiesis, and basophils, for which our test and training dataset contains fewer than 100 images. Note that for consecutive steps of granulopoiesis, a precise assignment of individual cells is known to be difficult. Hence, misclassifications between these classes have been considered tolerable[22]. Values of precision and sensitivity for all cell classes obtained by five-fold cross-vaildation are given in Table 1. Due to the varying number of cell types present in smears, the number of test and training images varies by up to two orders of magnitude for different classes (Table 1). In evaluating the model, we refer to class-wise precision and sensitivity, and do not calculate an overall accuracy score, which would be biased towards the classes with a high number of samples[11].

Inter- and intra-rater performance. To relate the network performance to the inter-rater variability encountered among human examiners, we asked another, independent examiner to re-annotate a subset of 1,905 single-cell images containing all subtypes previously annotated. We repeated this re-annotation with a time distance of 11 months in order to assess intra-rater variability (Supplementary Note 2 and Supplementary Fig. 3). The level of agreement between the gold standard annotation and the first and second re-annotation was excellent, as measured by Cohen's kappa, with $\kappa = 0.84$ and $\kappa = 0.87$ for the first and second re-annotations, respectively (Supplementary Note 1). Notably, the CNN prediction and the second annotation show similar patterns of deviation from the gold standard annotation, specifically as far as the classification of atypical lymphocytes and promyelocytes as myeloblasts is concerned (Fig. 3a and Supplementary Fig. 2). This may reflect visual similarities between the instances of these cell types recognized by both the network and human examiners, which intrinsically limit the single-cell annotation process. In these cases, access to the entire scan, as in the gold standard annotation, might be expected to be particularly helpful for single-cell classification.

Saliency maps. To evaluate if our network focuses on relevant parts of the single-cell images, we calculated saliency maps following the

procedure outlined in ref. [36]. These maps allow visualization of 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 single-cell image. No obvious correlation could be discerned between the saliency map and the result of the classifications, suggesting that both correct and incorrect classifications were obtained by the network by focusing on the single-cell image region containing the leukocyte.

Binary classification decisions. A key clinical question when examining blood cell morphology is whether 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[4]. 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 ROC curve is the result of sweeping t between 0 and 1, and is shown in Fig. 3c. The AUC, which we measure as 0.992 ± 0.001 using five-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 usual criteria of diagnostic test assessment[37,38]. In comparison to the network's performance, the human re-annotator reproduces the cell label provided by the gold standard annotation with a sensitivity of 95.7% and 90.7% and a specificity of 91.1% and 95.2% for the first and second re-annotations, respectively (Fig. 3c). Hence, both re-annotations lie close to, but somewhat below, the network ROC curve, indicating that the network achieves slightly superior performance than the human examiner in classifying single-cell images relative to the ground truth.

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_{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 AUC of 0.991 ± 0.002 when testing for atypicality of a cell, which lies in the domain of outstanding performance[37,38] and compares to a human re-annotation sensitivity of 95.9% and 91.7% and specificity of 91.0% and 95.3% for the first and second re-annotations, respectively (Fig. 3d). Importantly, our network outperforms the human second examiner's result on the subset of 1,905 single-cell images in both clinically relevant binary decision tasks tested here and for both re-annotations (Fig. 3c,d).

Network robustness. We test the robustness of our results by assessing an additional way to divide single-cell images into training and test sets, and by considering a different network architecture for the same classification task perfomed by ResNeXt.

As described in the Materials and methods section, assignment of single-cell images into the five different folds used for training and testing of the model was performed in a random fashion, disregarding from which scan a single-cell image was taken. Hence, different single-cell images from one scan could be present in the training and test sets at the same time. To rule out that this way of splitting the data into training and test sets introduces correlations between different single-cell images taken from the same scan, we performed a patient-wise split of the image data, as detailed in Supplementary Note 4. This mode of assignment is difficult for classes with a very low number of single-cell images, which typically come from only a few scans. Nevertheless, we observe that classification results are very well reproduced in the patient-wise split strategy. Specifically, the network trained on a case-wise basis achieves an AUC of 0.992 for the binary test for the blast character of a single-cell image, and an AUC of 0.986 for the binary test for atypicality. Also, for the full single-cell image classification problem, the network attains a good performance (Supplementary Fig. 5). These results suggest that possible correlations between different cells on a smear can be neglected when training and testing the model.

We also reproduced our results by repeating the full training and five-fold cross-validation on a deep sequential model (Supplementary Note 5, Supplementary Fig. 6 and Supplementary Table 2). Overall, the results appear comparable if slightly inferior to those of ResNeXt, so the results appear to be robust across different specific model architectures.

Discussion

The CNN 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 AMLs. For the most common physiological leukocyte classes as well as for myeloblasts, it attains a precision and sensitivity above 90%, allowing these cells to be identified with a very high accuracy that outperforms other classifiers in the literature[22,27]. The classification predictions can be used to answer clinically relevant binary questions. Blast character and atypicality are of high relevance in practice and can be recognized 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 dataset and the performance of the network on that class (Table 1). To the authors' knowledge, the image set presented in this Article is the largest used so far in the literature. We anticipate that further enlarging the dataset will also improve the network's classification performance for rare cell types. Additionally, a dataset including images obtained using different staining, illumination and scanning equipment would probably further increase the generalizability of the network predictions.

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 of rare leukocyte species in particular[8–10]. Inter- and intra-observer variability of cytomorphologic classification in our dataset was estimated by performing two independent re-annotations of image data. We note that the intra- and inter-observer 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 the scan, unlike the network and the second annotator, who performed a single-cell image classification.

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 examiners, we used this dataset to train and evaluate a state-of-the-art image classification CNN. The network shows good performance at differentiating the morphological cell types important for recognizing 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 the beginning of relapse. To screen all cells on a scanned blood smear fully automatically, our method may be combined with a segmentation tool that selects single cells, a task for which a large number of algorithms are available[39–42]. In this Article, we tested the model on peripheral blood smears from one laboratory, which were scanned with one type of scanning device. 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 decision support tool for cytological examiners, and might further increase its performance when combined with additional, intrinsically quantitative methods used in the diagnosis of haematological malignancies, such as flow cytometry or molecular genetics.

Data availability

The full single-cell image dataset and corresponding annotations are publicly available at The Cancer Imaging Archive (TCIA): https://doi.org/10.7937/tcia.2019.36f5o9ld[43].

Code availability

Code for the network trained in this study and network weights for one fold are available on CodeOcean, together with a subset of the single-cell image data used to test the network: https://codeocean.com/capsule/9068249/tree/v1[44].

Received: 22 October 2018; Accepted: 3 September 2019;
Published online: 12 November 2019

References

1. Bain, B. J. Diagnosis from the blood smear. *N. Engl. J. Med.* **353**, 498–507 (2005).
2. Tkachuk, D. C. & Hirschmann, J. V. *Wintrobe's Atlas of Clinical Hematology* (Lippincott Raven, 2006).
3. Theml, H., Diem, H. & Haferlach, T. *Color Atlas of Hematology* (Thieme, 2004).
4. Döhner, H. et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* **129**, 424–447 (2017).
5. Swerdlow, S. H. et al. (eds) *WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues* 4th edn (International Agency for Research on Cancer, 2017).
6. Arber, D. A. et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).
7. Bennett, J. M. et al. Proposed revised criteria for the classification of acute myeloid leukemia. A report of the French–American–British Cooperative Group. *Ann. Intern. Med.* **103**, 620–625 (1985).
8. Font, P. et al. Inter-observer variance with the diagnosis of myelodysplastic syndromes (MDS) following the 2008 WHO classification. *Ann. Hematol.* **92**, 19–24 (2013).
9. Font, P. et al. Interobserver variance in myelodysplastic syndromes with less than 5% bone marrow blasts: unilineage vs. multilineage dysplasia and reproducibility of the threshold of 2% blasts. *Ann. Hematol.* **94**, 565–573 (2015).
10. Fuentes-Arderiu, X. & Dot-Bach, D. Measurement uncertainty in manual differential leukocyte counting. *Clin. Chem. Lab. Med.* **47**, 112–115 (2009).
11. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
12. Rawat, W. & Wang, Z. Deep convolutional neural networks for image classification: a comprehensive review. *Neural Comput.* **29**, 2352–2449 (2017).
13. Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision* **115**, 211–252 (2015).
14. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
15. Eulenberg, P. et al. Reconstructing cell cycle and disease progression using deep learning. *Nat. Commun.* **8**, 463 (2017).
16. Janowczyk, A. & Madabhushi, A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J. Pathol. Inform.* **7**, 29 (2016).
17. Fuchs, T. J. & Buhmann, J. M. Computational pathology: challenges and promises for tissue analysis. *Comput. Med. Imaging Graph.* **35**, 515–530 (2011).
18. Albarqouni, S. et al. AggNet: deep learning from crowds for mitosis detection in breast cancer histology images. *IEEE Trans. Med. Imaging* **35**, 1313–1321 (2016).
19. Levenson, R. M., Fornari, A. & Loda, M. Multispectral imaging and pathology: seeing and doing more. *Expert Opin. Med. Diagn.* **2**, 1067–1081 (2008).
20. Gertych, A. et al. Machine learning approaches to analyze histological images of tissues from radical prostatectomies. *Comput. Med. Imaging Graph.* **46**, 197–208 (2015).
21. Bigorra, L., Merino, A., Alférez, S. & Rodellar, J. Feature analysis and automatic identification of leukemic lineage blast cells and reactive lymphoid cells from peripheral blood cell images. *J. Clin. Lab. Anal.* **31**, e22024 (2017).
22. Krappe, S., Wittenberg, T., Haferlach, T. & Munzenmayer, C. 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 hematopoesis. *Proc. SPIE* **9785**, 97853C (2016).
23. Scotti, F. Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. In *Computational Intelligence for Measurement Systems and Applications (CIMSA)* 96–101 (IEEE, 2005).
24. Mohapatra, S., Patra, D. & Satpathy, S. An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. *Neural Comput. Appl.* **24**, 1887–1904 (2014).
25. Greenspan, H., van Ginneken, B. & Summers, R. M. Deep learning in medical imaging: overview and future promise of an exciting new technique. *IEEE Trans. Med. Imaging* **35**, 1153–1159 (2016).
26. Shen, D., Wu, G. & Suk, H. Deep learning in medical image analysis. *Ann. Rev. Biomed. Eng.* **19**, 221–248 (2017).
27. Choi, J. W. et al. White blood cell differential count of maturation stages in bone marrow smear using dual-stage convolutional neural networks. *PLoS One* **12**, e0189259 (2017).
28. Kainz, P., Burgsteiner, H., Asslaber, M. & Ahammer, H. Training echo state networks for rotation-invariant bone marrow cell classification. *Neural Comput. Appl.* **28**, 1277–1292 (2017).
29. Su, M.-C., Cheng, C.-Y. & Wang, P.-C. A neural-network-based approach to white blood cell classification. *Sci. World J.* **2014**, 796371 (2014).
30. Macawile, M. J., Quiñones, V. V., Ballado, A., Cruz, J. D. & Caya, M. V. White blood cell classification and counting using convolutional neural network. In *2018 3rd International Conference on Control and Robotics Engineering (ICCRE)* 259–263 (IEEE, 2018).
31. Keohane, E. M., Smith, L. & Walenga, J. M. *Rodak's Hematology—Clinical Principles and Applications* 5th edn (Elsevier, 2016).
32. Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 5987–5995 (IEEE, 2017).
33. Dietz, M. ResNeXt implementation for Keras. *GitHub Gist* https://gist.githubusercontent.com/mjdietzx/ (2017).
34. Chollet, F. et al. Keras 2.0. *Keras* https://keras.io (2017).
35. Bychkov, D. et al. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Sci. Rep.* **8**, 3395 (2018).
36. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: visualising image classification models and saliency maps. Preprint at https://arxiv.org/abs/1312.6034 (2013).
37. Mandrekar, J. N. Receiver operating characteristic curve in diagnostic test assessment. *J. Thorac. Oncol.* **5**, 1315–1316 (2010).
38. Hosmer, D. & Lemeshow, S. *Applied Logistic Regression* 2nd edn (Wiley, 2000).
39. Xing, F. & Yang, L. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. *IEEE Rev. Biomed. Eng.* **9**, 234–263 (2016).
40. Cuevas, E. et al. White blood cell segmentation by circle detection using electromagnetism-like optimization. *Comput. Math. Methods Med.* **2013**, 395071 (2013).
41. Alomari, Y. M., Abdullah, S. N. H. S., Azma, R. Z. & Omar, K. Automatic detection and quantification of WBCs and RBCs using iterative structured circle detection algorithm. *Comput. Math. Methods Med.* **2014**, 979302 (2014).
42. He, K., Gkioxari, G., Dollár, P. & Girshick, R. Mask R-CNN. In *Proceedings of the International Conference on Computer Vision (ICCV)* 2980–2988 (IEEE, 2017).
43. Matek, C., Schwarz, S., Spiekermann, K. & Marr, C. A single-cell morphological dataset of leukocytes from AML patients and non-malignant controls (AML-Cytomorphology_LMU). *TCAI* https://doi.org/10.7937/tcia.2019.36f5o9ld (2019).
44. Matek, C., Schwarz, S., Spiekermann, K. & Marr, C. A neural network for classifying leukocyte images from blood smears. *CodeOcean* https://codeocean.com/capsule/9068249/tree/v1 (2019).

Acknowledgements

We thank N. Chlis for comments on the manuscript, K. Metzeler for helpful discussions and A. Holzäpfel for contributions to the annotation task. This work was supported by the German Research Foundation DFG within the Collaborative Research Center SFB 1243. C. Matek acknowledges support from Deutsche José Carreras-Leukämie Stiftung.

Author contributions

C. Matek, C. Marr and K.S. conceived the initial idea. C. Matek selected the cohort, digitized blood smears, wrote annotation software, and trained and evaluated the network. S.S. contributed to selecting the cohort and annotated the image data. C. Matek, C. Marr and K.S. interpreted data and wrote the paper. All authors approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s42256-019-0101-9.

Correspondence and requests for materials should be addressed to K.S. or C.M.

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.