

Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image dataset

Christian Matek^{1,2}, Sebastian Krappe^{3,4}, Christian Münzenmayer³, Torsten Haferlach⁵, Carsten Marr¹

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

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

3 Image Processing and Medical Engineering Department, Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany

4 Department of Computer Science, University of Koblenz-Landau, Koblenz, Germany

5 MLL Munich Leukemia Laboratory, Munich, Germany

Supplementary Material

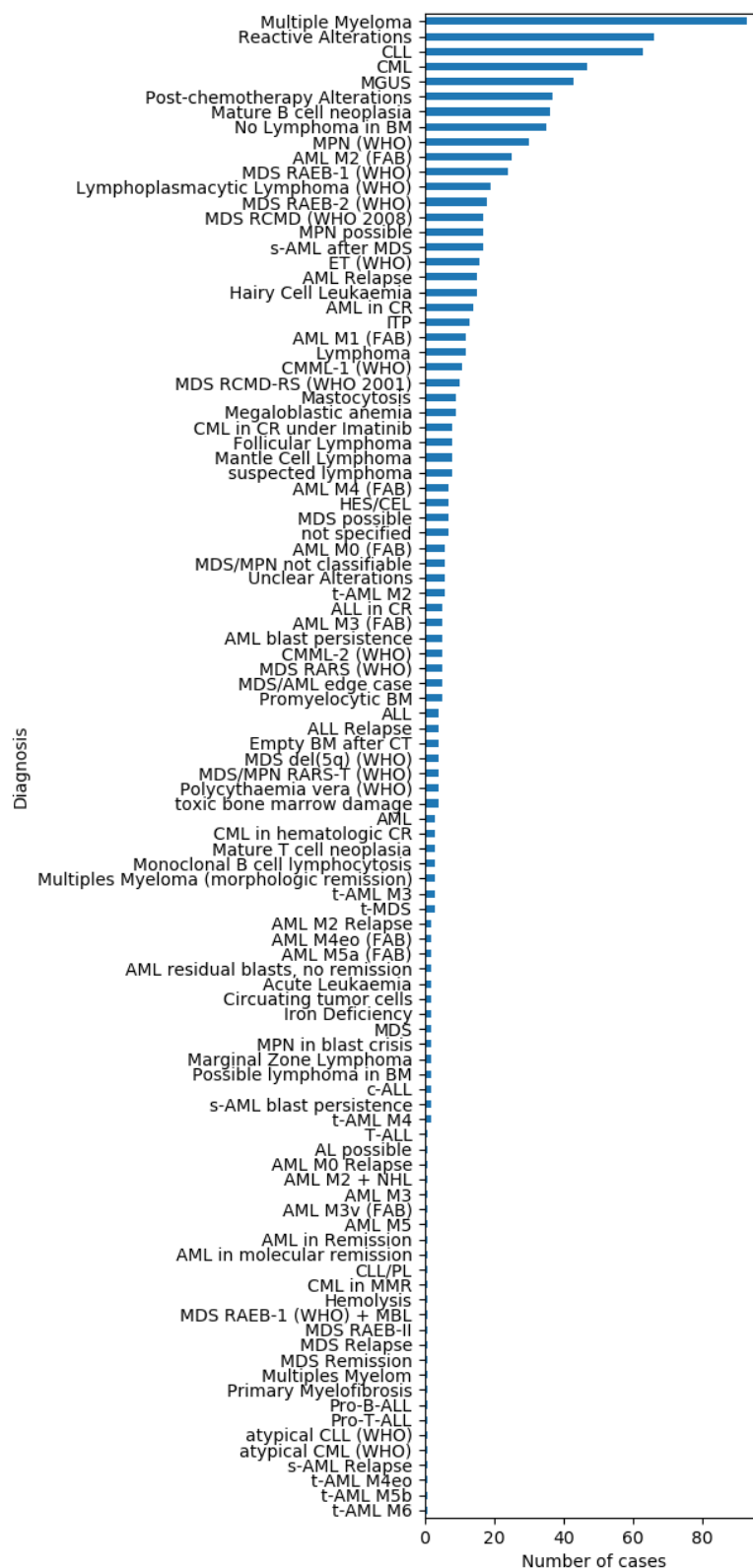


Figure S1: Distribution of diagnoses for the 945 patients included in the study.

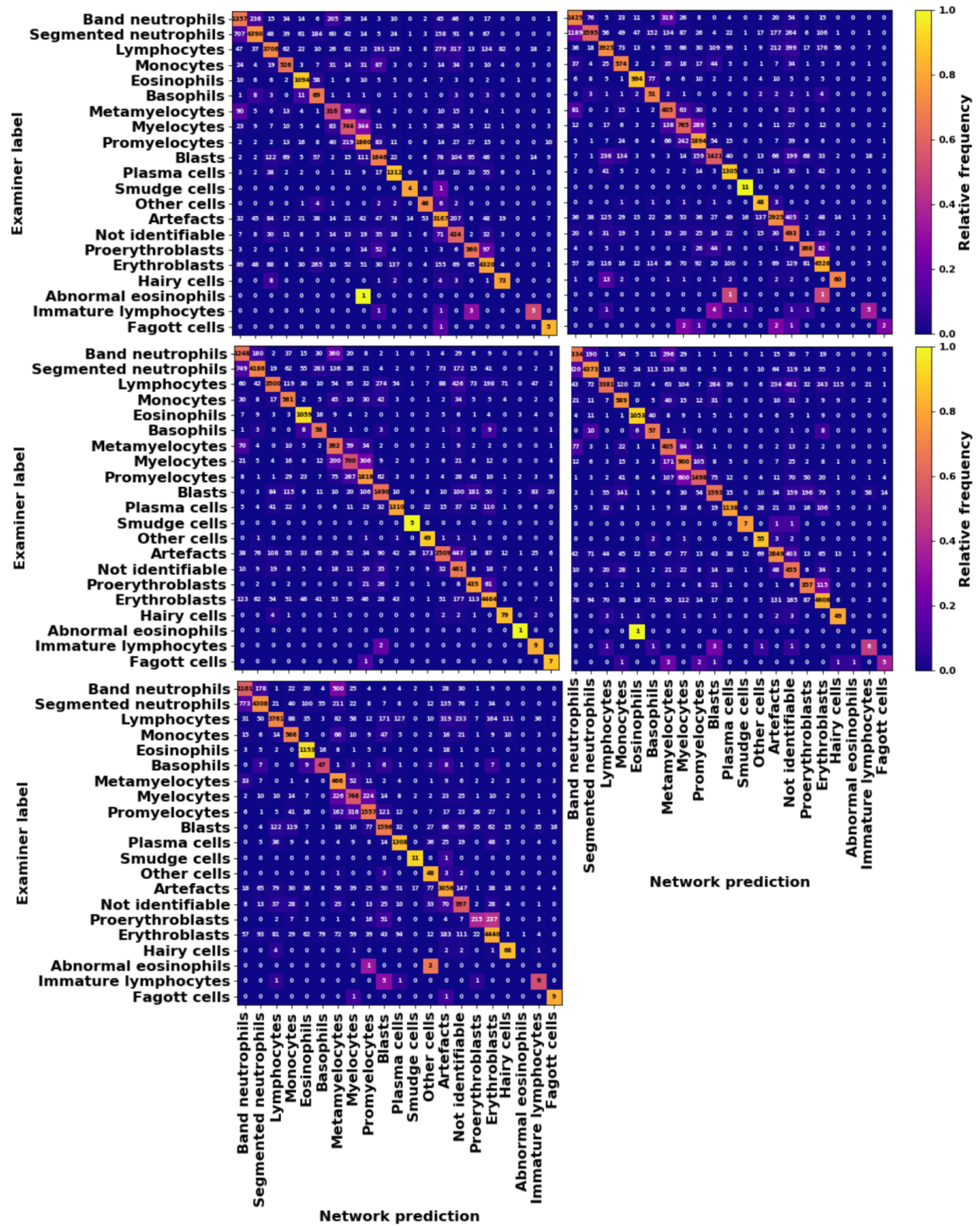


Figure S2: Confusion matrices for the 5 distinct folds trained for the ResNeXt-50 model. In order to account for class imbalance, data is normalized row-wise and coloured accordingly.

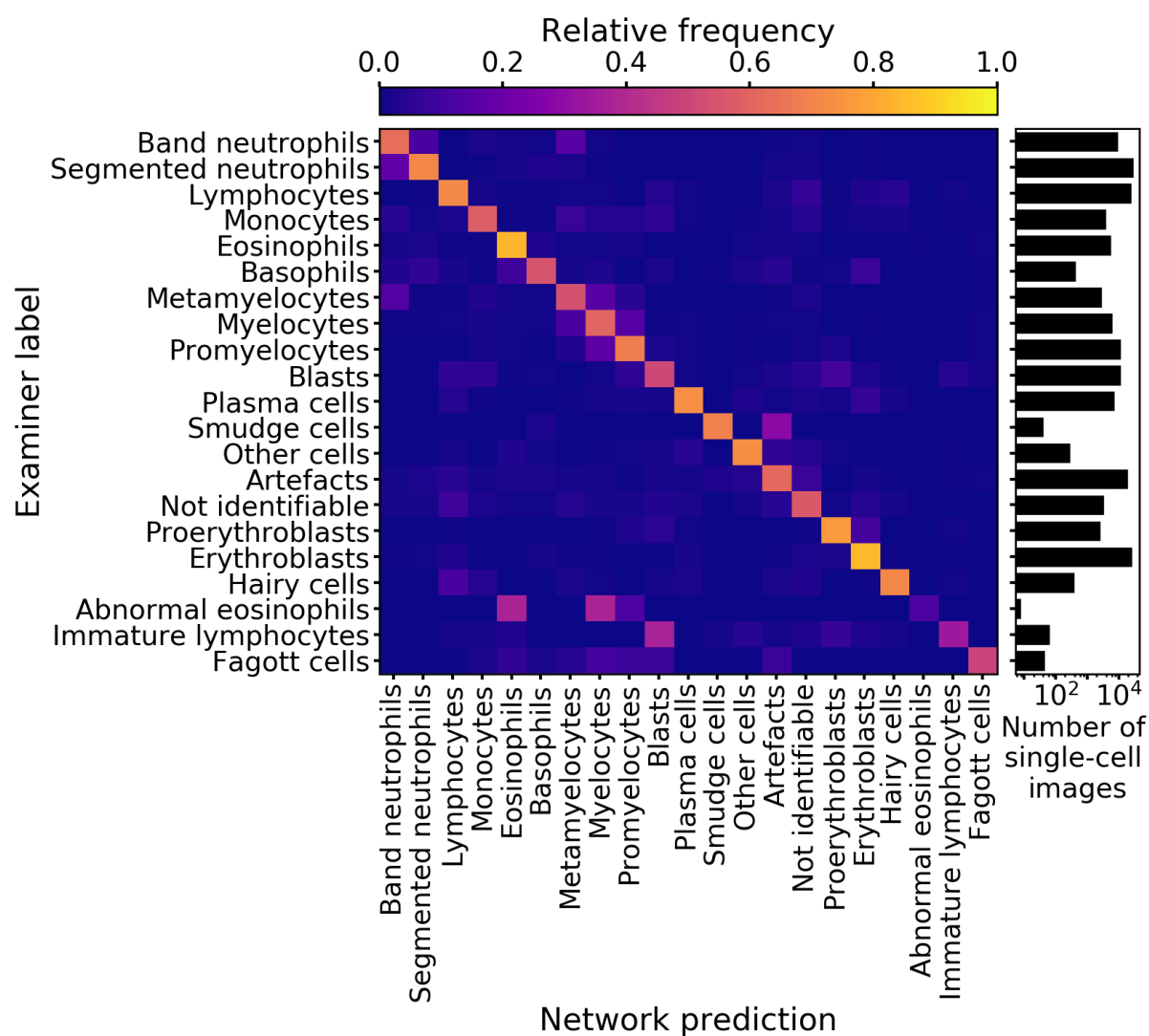


Figure S3: Confusion matrix for the sequential model trained with cross-validation on the same split of the data into 5 folds. To visually correct for class imbalance, the matrix is normalised row-wise. As in Fig. 2 of the main text, cell abundance for the ground-truth is shown in a histogram on a logarithmic scale.

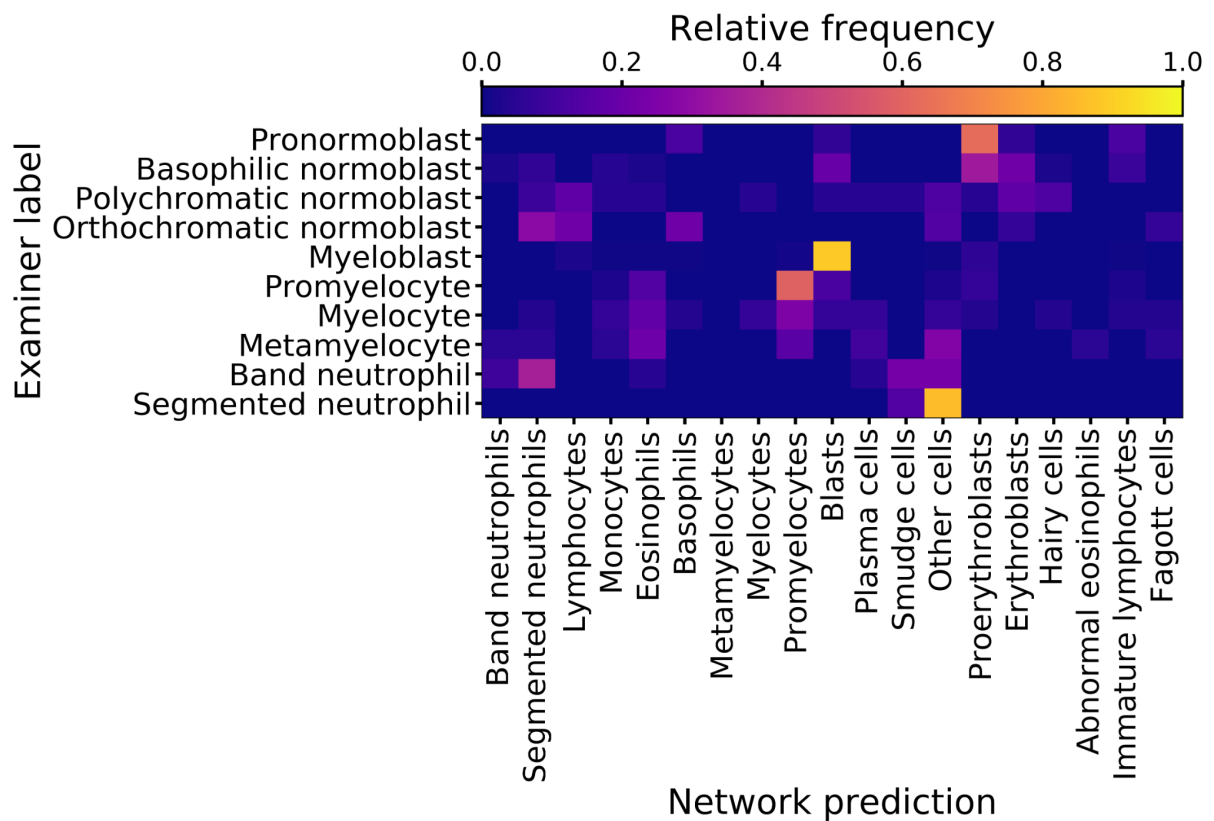


Figure S4: External validation results on the sequential model using the same test data by Choi *et al.* as used in Fig. 5 of the main paper. Out of the 627 external test set images, 207 (corresponding to 33%) were classified as unknown cells or artifacts. Classification performance is shown for the remaining 420 images, with rows normalised. Also the sequential network exhibits fair performance at classifying key cell classes in the myeloid and erythroid lineage, especially Normoblasts and Myeloblasts. However the network is less successful than ResNeXt on mature steps of myelopoiesis.

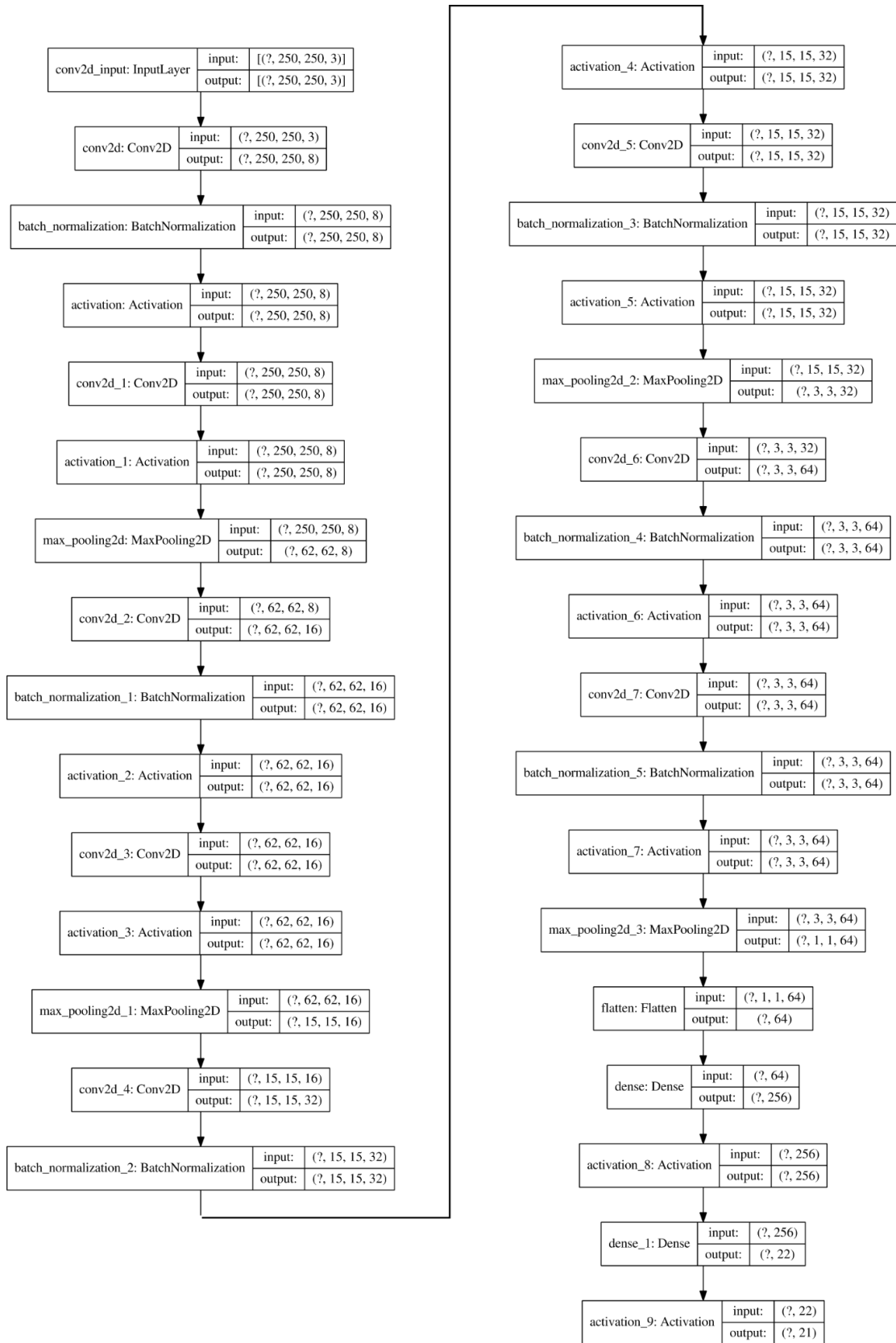


Figure S5: Structure of the sequential model used in this work for comparison with the results obtained using the ResNeXt architecture.

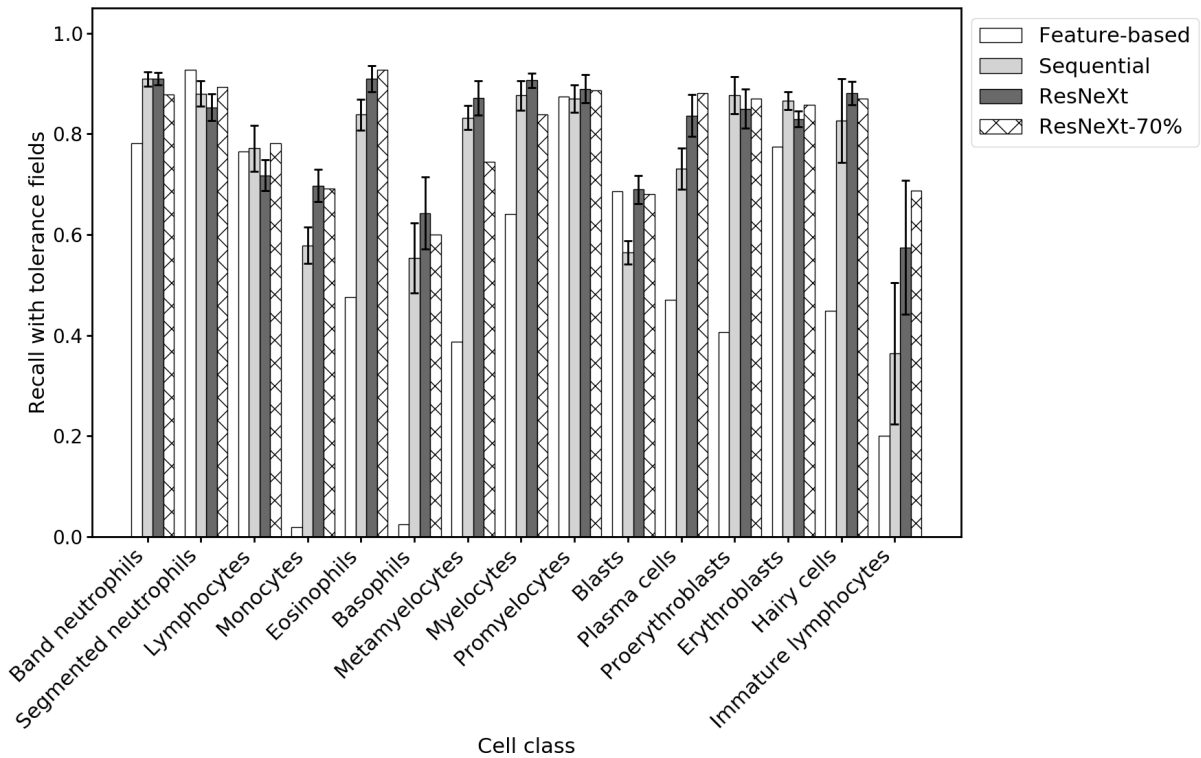


Figure S6: Influence of the train-test-split on classifier recall. In order to assess the dependence of the classifier recall on the precise train-test-split chosen, we re-trained a ResNeXt-50 model using a stratified train-test-split of 70% vs. 30%, as was used in the feature-based approach (cf. cross-hatched columns). For most classes, the results of the 70% vs- 30% split strategy fall within the margin of variation obtained through 5-fold cross-validation from a 80% vs. 20% strategy, indicating that the model results are robust against the variation in train-test split strategy.

Supplementary Tables

| Class | Precision _{tolerant} | Recall _{tolerant} | Precision _{strict} | Recall _{strict} | Images |
|-----------------------|-------------------------------|----------------------------|-----------------------------|--------------------------|--------|
| Band neutrophils | 0.93 ± 0.01 | 0.91 ± 0.01 | 0.50 ± 0.04 | 0.63 ± 0.08 | 9968 |
| Segmented neutrophils | 0.95 ± 0.00 | 0.88 ± 0.03 | 0.90 ± 0.02 | 0.71 ± 0.04 | 29424 |
| Lymphocytes | 0.85 ± 0.02 | 0.77 ± 0.05 | 0.85 ± 0.02 | 0.72 ± 0.05 | 26242 |
| Monocytes | 0.51 ± 0.04 | 0.58 ± 0.04 | 0.51 ± 0.04 | 0.58 ± 0.04 | 4040 |
| Eosinophils | 0.76 ± 0.08 | 0.84 ± 0.03 | 0.76 ± 0.08 | 0.84 ± 0.03 | 5883 |
| Basophils | 0.10 ± 0.04 | 0.55 ± 0.07 | 0.10 ± 0.04 | 0.55 ± 0.07 | 441 |
| Metamyelocytes | 0.65 ± 0.04 | 0.83 ± 0.02 | 0.28 ± 0.04 | 0.54 ± 0.08 | 3055 |
| Myelocytes | 0.81 ± 0.03 | 0.88 ± 0.03 | 0.50 ± 0.04 | 0.60 ± 0.04 | 6557 |
| Promyelocytes | 0.92 ± 0.02 | 0.87 ± 0.03 | 0.76 ± 0.04 | 0.68 ± 0.04 | 11994 |
| Blasts | 0.75 ± 0.02 | 0.56 ± 0.02 | 0.72 ± 0.02 | 0.51 ± 0.03 | 11973 |
| Plasma cells | 0.79 ± 0.05 | 0.73 ± 0.04 | 0.79 ± 0.05 | 0.73 ± 0.04 | 7629 |
| Smudge cells | 0.05 ± 0.02 | 0.71 ± 0.16 | 0.05 ± 0.02 | 0.71 ± 0.16 | 42 |
| Other cells | 0.16 ± 0.07 | 0.73 ± 0.12 | 0.16 ± 0.07 | 0.73 ± 0.12 | 294 |
| Artefacts | 0.86 ± 0.03 | 0.62 ± 0.04 | 0.86 ± 0.03 | 0.62 ± 0.04 | 19630 |
| Not identifiable | 0.26 ± 0.04 | 0.57 ± 0.04 | 0.26 ± 0.04 | 0.57 ± 0.04 | 3538 |
| Proerythroblasts | 0.58 ± 0.08 | 0.88 ± 0.04 | 0.46 ± 0.05 | 0.77 ± 0.02 | 2740 |
| Erythroblasts | 0.90 ± 0.02 | 0.87 ± 0.02 | 0.89 ± 0.02 | 0.85 ± 0.02 | 27395 |
| Hairy cells | 0.71 ± 0.04 | 0.83 ± 0.08 | 0.15 ± 0.03 | 0.71 ± 0.09 | 409 |
| Abnormal eosinophils | 0.04 ± 0.08 | 0.20 ± 0.40 | 0.04 ± 0.08 | 0.20 ± 0.40 | 8 |
| Immature lymphocytes | 0.25 ± 0.05 | 0.36 ± 0.14 | 0.02 ± 0.01 | 0.34 ± 0.11 | 65 |
| Fagott cells | 0.03 ± 0.01 | 0.50 ± 0.10 | 0.03 ± 0.01 | 0.50 ± 0.10 | 47 |

Table S1: Full class-wise performance evaluation of the sequential network trained for this study using 5-fold cross-validation. Both the strict and tolerant evaluation strategies are shown. For most classes, performance values are somewhat below the corresponding values of the ResNeXt architecture, which is unsurprising given the significantly lower number of trainable parameters.