ROBUSTNESS EVALUATION OF A CONVOLUTIONAL NEURAL NETWORK FOR THE CLASSIFICATION OF SINGLE CELLS IN ACUTE MYELOID LEUKEMIA

Christian Matek & Carsten Marr
Institute of Computational Biology
Helmholtz Zentrum München–German Research Center for Environmental Health
85764 Munich, Germany
{christian.matek,carsten.marr}@helmholtz-muenchen.de

ABSTRACT

Robustness to domain-characteristic perturbations is a key prerequisite for deploying neural networks in a real-world diagnostic setting for morphological classification of cytological images, thus closing the gap between proof of concept and routine use. Here, we present a robustness analysis of a recently published network for the classification of malignant and non-malignant blood cells relevant for the diagnosis of Acute Myeloid Leukemia (AML) with respect to three modes of plausible image corruptions. We find that the network is robust to defocus blurring and JPEG compression, whereas performance deteriorates for changes in brightness. We show that retraining of the network with a corresponding augmentation strategy can ensure robustness also for brightness variation. Our analysis and training strategy paves the way for the application of neural networks in clinical and laboratory settings for rapid, reproducible and reliable single-cell classification.

1 INTRODUCTION

Based on the rapid evolution of modern neural networks, significant progress has been made in recent years at applying machine learning approaches to analysis of light microscopy images from both cytopathology and histopathology (Landau & Pantanowitz, 2019; Ibrahim et al., 2020; Kather et al., 2019). Across a varied set of disease entities, algorithms were developed that attain promising levels of performance, matching human experts in some well-defined subtasks (Hekler et al., 2019; Wei et al., 2019; Matek et al., 2019a). However, evaluating the robustness of these algorithms can be challenging. In particular, medical data are often difficult to share, so that publicly available, high-quality and expert-annotated datasets are scarce. This can complicate the application of algorithms to data from multiple centers, which is a key step in the evaluation of deep learning models in pathology (Schmitt et al., 2021). Furthermore, sample handling and data acquisition protocols may differ significantly between institutions, and are typically optimised towards the habits of human examiners rather than computational analysis. To nevertheless be able to assess the robustness of neural-network based algorithms, it can be useful to simulate domain-characteristic image perturbations, thus trying to cover expected differences between different data sources or batches. Here, we follow this strategy for a robustness evaluation of one specific use case from hematology, namely morphology-based recognition of malignant blast cells, which are abnormal leukocytes characteristic of Acute Myeloid Leukemia (AML), in peripheral blood smears.

Morphologic evalutaion of peripheral blood smears using light microscopy is a key step in the diagnostic workup of hematological diseases (Bain, 2005). For its relative technical simplicity and wide availability, it represents an important cornerstone in the diagnosis and classification of various subtypes of leukemia (Döhner et al., 2017; Bennett et al., 1985). As an image-classification problem, morphologic classification of leukocytes is a potential use case of convolutional neural networks. A recent, neural network-based approach used the ResNeXt architecture to classify while blood cells from peripheral blood into a diagnostically relevant, 15-category classification scheme and recognize blast cells at human-level performance (Matek et al., 2019a). The approximately

18,000 single-cell images used to train the network were obtained from 200 patients from a single center, and handled and digitised following the same workflow using identical equipment. The full labelled image dataset is publicly available from TCIA (Matek et al., 2019b). The system was recently included in an audit, which indicated that the model was robust in the context of a limited set of sample images (Oala et al., 2020).

2 Network robustness evaluation

2.1 Image corruption modes

Here, we systematically examine the robustness of blast classification by the network of Matek et al. (2019a) with respect to corruptions of the test data. To simulate the variability of image data, we use the method of Hendrycks & Dietterich (2019), which introduced 15 distinct image corruption modes at 5 levels of severity. As it was developed for natural image classification, not all corruption types are plausible for single-cell images from peripheral blood smears. We therefore restricted our analysis to defocus blur, JPEG compression and brightness variation, which all represent realistic sources of variability in the domain of digital pathology. Other corruption modes due to sample handling and stain variability are likely to play a role in model performance in addition to the corruption modes analysed here. However, given the relatively large number of smears and single cells included in the dataset (Matek et al., 2019b), we expect these factors to be to some extent reflected in the original data. Examples of the effect of the corruptions studied here on one single-cell image of a blast cell at different levels of severity are shown in Fig. 1.

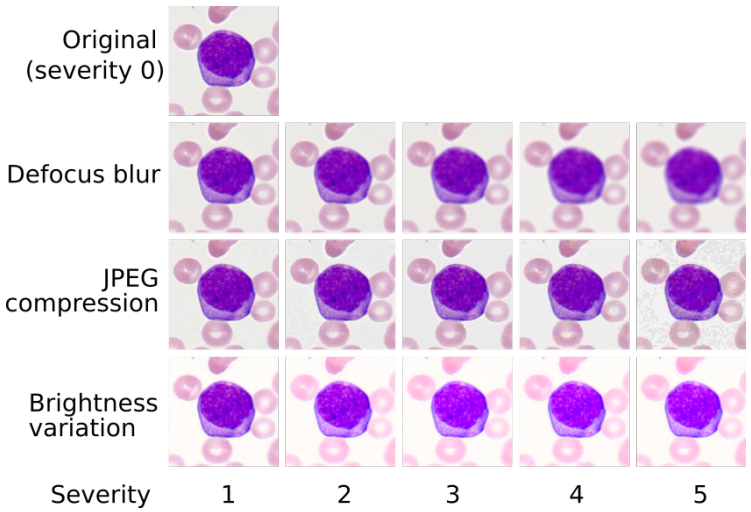

Figure 1: Effect of different levels of severity for three image corruption modes applied to the image of a myeloblast, whose appearance in peripheral blood is characteristic of AML.

2.2 Effect of image corruptions on blast recognition

Evaluation of the network was performed using the corrupted data in the same way as for the original data, following the identical stratified split of the dataset to perform 5-fold cross-validation, as described in Matek et al. (2019a). We visualise the effect of the image corruptions considered here by plotting the receiver operating characteristic (ROC) curve, and calculating the area under the ROC curve (AUC). The effect of different levels of severity on the ROC curves obtained by 5-fold cross-validation is shown in Fig. 2a for the three different modes considered here. For defocus blur and JPEG compression, we observe only minor reductions in blast cell recognition performance, and hence conclude that the model is stable with respect to these corruption modes. In contrast, we observe that for brightness variation, blast recognition performance suffers significantly. AUC scores for different corruption modes and severity levels are summarized in Fig. 3. Note that the inter-fold variability increases for increasing levels of corruption severity, reflecting higher noise levels in network prediction in the presence of corrupted input data.

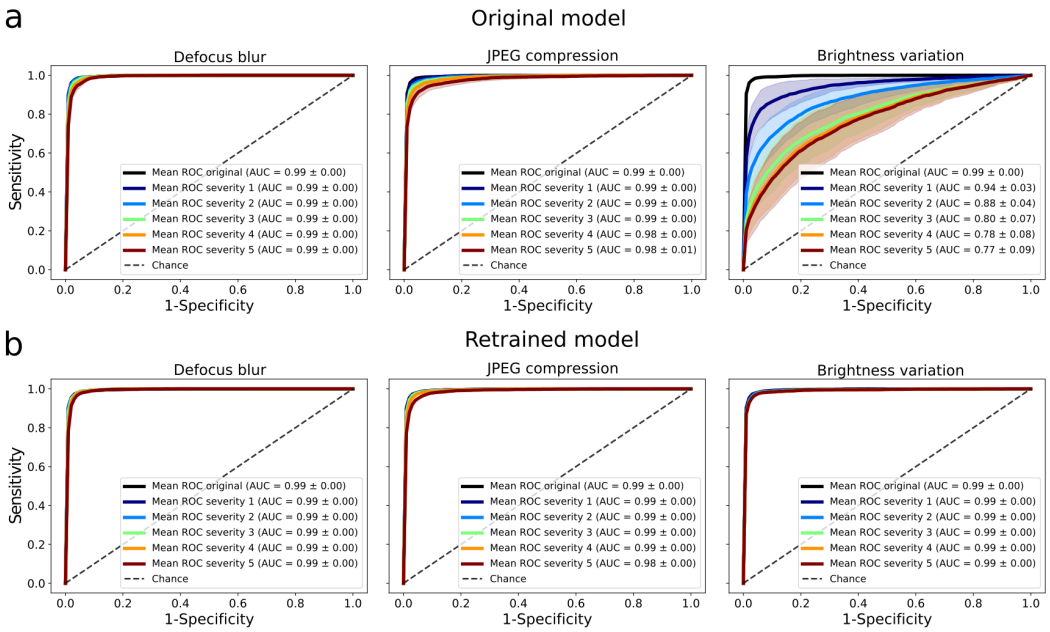

Figure 2: Mean receiver operating characteristic (ROC) curves of the binary classifier for blast recognition at different corruption modes and levels of severity. for the original model (a) and retrained model (b). The binary classifier for blast cells is initially stable for all levels of JPEG compression and defocus blur. Augmentation of the training data leads to robustness also with respect to brightness variation.

The observed differences in robustness of the network with respect to the tested image corruptions may be rationalised by considering the data acquisition procedure followed in compiling the AML dataset (Matek et al., 2019b): The microscope parameters for imaging single cells on a blood smear typically imply a shallow focus depth, thus increasing the likelihood that some cells images are not fully in focus, which is visible in the AML single-cell dataset. Therefore, we hypothesize that appearance of slightly out-of-focus images in the training dataset leads to a model that is robust with respect to defocus blurring.

For JPEG compression, the original model appears equally robust with respect to all levels of corruption severity. While the quality loss with increasing compression level is clearly visible (cf. Fig. 1), only an insignificant performance drop is observed in the network (Fig. 2a). This indicates that lower-quality images suffice to answer the diagnostically relevant question if a cell possesses blast character or not. This may also be true for human observers, although systematic investigations on human diagnostic robustness with respect to image corruption are lacking.

For brightness variation, a decrease of the network's performance level is observed (Fig. 2a, 3a). This may reflect the fact that the single-cell dataset was imaged using one slide scanner (Matek et al., 2019a), so that the brightness value of single-cell images across the dataset is fairly homogeneous.

Overall, assessment of the model under plausible image corruptions shows that performance of the trained network is robust against JPEG compression and defocus blur, but vulnerable to variations in image brightness.

3 Model performance boost by brightness augmentation

Based on the results of the model robustness evaluation, we retrain the network using the same 5-fold stratified split and geometrical augmentation strategy as described in Matek et al. (2019a). Additionally, we randomly perturb the brightness of the training images with an enhancement factor in the range between 0.5 and 1.5, where 1.0 corresponds to the original image. We find that the retrained network retains the performance and robustness level with respect to defocus blur and

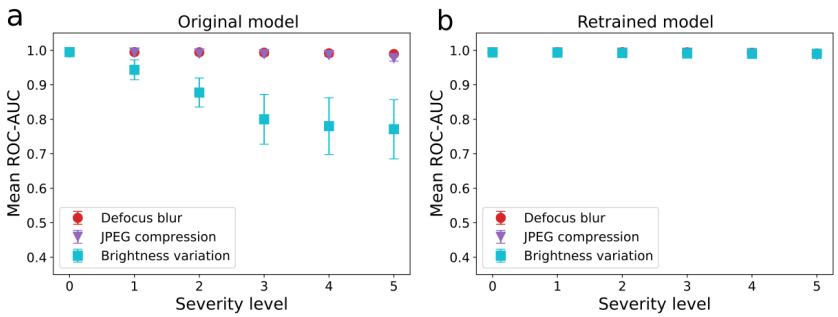

Figure 3: Mean ROC-AUC score obtained by the network on blast cell recognition for different corruption modes and severity levels for the original model (a) and the model retrained with brightness-augmented data (b). A severity level of zero indicates the original, unperturbed test set. All values are shown as mean $\pm$ standard deviations obtained by 5-fold cross-validation.

JPEG compression corruptions of the original model, while additionally gaining robustness towards input brightness variations (cf. Fig. 2b and 3b). For a full analysis of leukocytes in peripheral blood smears, binary single-cell image classification has to be extended into a finer-grained classification scheme covering the 15 morphological classes of the single-cell dataset. When evaluated on multiclass classification performance on uncorrupted data, the retrained network shows the same level of performance as the original network, as can be observed by comparing both models' confusion matrices (cf. Fig. 4).

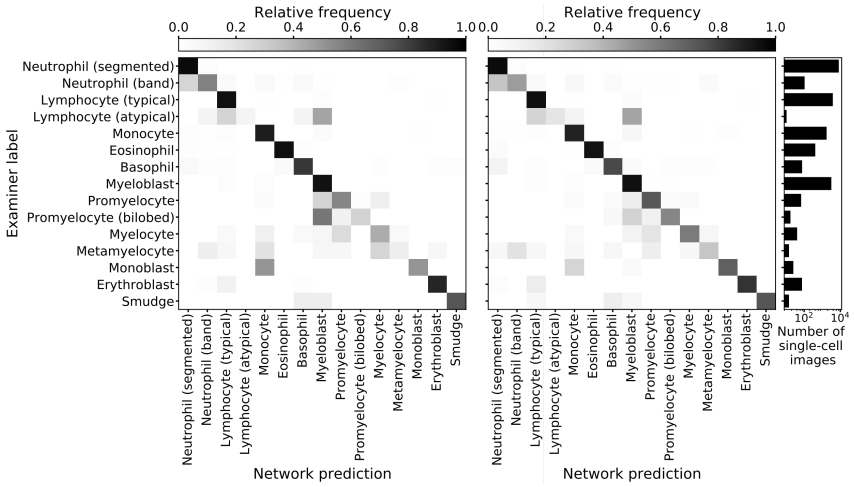

Figure 4: 15 cell class confusion matrix of the original model (left panel) and the retrained model (right panel), evaluated uncorrupted data using 5-fold cross-validation. Comparison of both matrices shows that both models have a very similar multiclass classification behaviour. Both matrices are normalized row-wise.

4 Conclusions

We have used a recently proposed and widely applied benchmark for natural image corruptions to systematically investigate the robustness of a neural network for white blood cell classification with respect to changes in focus, brightness and data compression, which are plausible in the domain of light microscopy. We find that a model trained on uncorrupted data is robust with respect to defocussing and JPEG compression. Robustness with respect to brightness variations can be achieved by retraining the model on data transformed by a random brightness enhancement factor. The retrained model retains the performance level of the original model both for binary and multiclass classifica-

tion. Our work shows the utility of simulated data corruptions to assess the robustness of an image classification model and identify factors that may hamper its generalizability. Moreover, we show how models can be retrained to achieve robust and reliable performance for expected image variability. Thus, our work is a first step to close the proof of concept to production gap for AI applications in cytopathology.

ACKNOWLEDGMENTS

Christian Matek and Carsten Marr received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 866411) and acknowledge support from the Helmholtz Innovation and Translation fund.

REFERENCES

B. J. Bain. Diagnosis from the blood smear. *The New England journal of medicine*, 353:498–507, 2005. ISSN 1533-4406. doi: 10.1056/NEJMra043442.

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, 1985. ISSN 0003-4819.

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, 2017. ISSN 1528-0020. doi: 10.1182/blood-2016-08-733196.

A. Hekler, J. S. Utikal, A. H. Enk, C. Berking, J. Klode, D. Schadendorf, P. Jansen, C. Franklin, T. Holland-Letz, D. Krahl, C. von Kalle, S. Fröhling, and T. J. Brinker. Pathologist-level classification of histopathological melanoma images with deep neural networks. *European Journal of Cancer*, 115:79–83, 2019. ISSN 0959-8049. doi: https://doi.org/10.1016/j.ejca.2019.04.021. URL https://www.sciencedirect.com/science/article/pii/S0959804919302758.

D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. ICLR conference paper (arXiv:1903.12261), 2019.

A. Ibrahim, P. Gamble, R. Jaroensri, M. M. Abdelsamea, C.H. Mermel, P.-H. C. Chen, and E. A. Rakha. Artificial intelligence in digital breast pathology: Techniques and applications. *The Breast*, 49:267–273, 2020. ISSN 0960-9776. doi: https://doi.org/10.1016/j.breast.2019.12.007. URL https://www.sciencedirect.com/science/article/pii/S0960977619312147.

J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber, L. Jansen, C. C. Reyes-Aldasoro, I. Zörnig, D. Jäger, H. Brenner, J. Chang-Claude, M. Hoffmeister, and N. Halama. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16(1): 1–22, 01 2019. doi: 10.1371/journal.pmed.1002730. URL https://doi.org/10.1371/journal.pmed.1002730.

M. S. Landau and L. Pantanowitz. Artificial intelligence in cytopathology: a review of the literature and overview of commercial landscape. *Journal of the American Society of Cytopathology*, 8(4): 230–241, 2019. ISSN 2213-2945. doi: https://doi.org/10.1016/j.jasc.2019.03.003. URL https://www.sciencedirect.com/science/article/pii/S2213294519300067.

C. Matek, S. Schwarz, K. Spiekermann, and C. Marr. Human-level recognition of blast cells in acute myeloid leukemia with convolutional neural networks. *Nature Machine Intelligence*, 1: 538—-544, 2019a.

C. Matek, S. Schwarz, K. Spiekermann, and C. Marr. A single-cell morphological dataset of leukocytes from aml patients and non-malignant controls [data set]. The Cancer Imaging Archive, 2019b. URL https://doi.org/10.7937/tcia.2019.36f5o9ld.

L. Oala, J. Fehr, G. Jaramillo-Gutierrez, L. Gilli, P. Balachandran, A. Werneck Leite, D. Xie Li, G. Nobis, E. Alejandro Munoz Alvarado, F. Kherif, C. Matek, A. Shroff, B. Sanguinetti, and T. Wiegand. Ml4h quality assessment: From paper to practice. *Machine Learning for Health / Proceedings of Machine Learning Research (accepted)*, 2020.

M. Schmitt, R. C. Maron, A. Hekler, A. Stenzinger, A. Hauschild, M. Weichenthal, M. Tiemann, D. Krahl, H. Kutzner, J. S. Utikal, S. Haferkamp, J. N. Kather, F. Klauschen, E. Krieghoff-Henning, S. Fröhling, C. von Kalle, and T. J. Brinker. Hidden variables in deep learning digital pathology and their potential to cause batch effects: Prediction model study. *J Med Internet Res*, 23(2):e23436, Feb 2021. ISSN 1438-8871. doi: 10.2196/23436. URL https://www.jmir.org/2021/2/e23436.

J. W. Wei, L. J. Tafe, Y. A. Linnik, L. J. Vaickus, N. Tomita, and S. Hassanpour. Pathologist-level classification of histologic patterns on resected lung adenocarcinoma slides with deep neural networks. *Scientific Reports*, 9(1):3358, Mar 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-40041-7. URL https://doi.org/10.1038/s41598-019-40041-7.