

1 **Image-based Explainable Artificial Intelligence Accurately Identifies**
2 **Myelodysplastic Neoplasms Beyond Conventional Signs of Dysplasia**

3 Jan-Niklas Eckardt,^{1,2} Ishan Srivastava,^{1,2} Freya Schulze,¹ Susann Winter,¹ Tim Schmittmann,³
4 Sebastian Riechert,^{2,3} Martin M. K. Schneider¹, Lukas Reichel,¹ Miriam Eva Helena Gediga,¹ Katja
5 Sockel,¹ Anas Shekh Sulaiman,¹ Christoph Röllig,¹ Frank Kroschinsky,¹ Anne-Marie Asemissen,⁴
6 Christian Pohlkamp,⁵ Torsten Haferlach,⁵ Martin Bornhäuser,^{1,6,7} Karsten Wendt,^{2,3} and Jan Moritz
7 Middeke^{1,2}

8 ¹ Department of Internal Medicine I, University Hospital Carl Gustav Carus, TUD Dresden University
9 of Technology, Dresden, Germany

10 ² Else Kröner Fresenius Center for Digital Health, TUD Dresden University of Technology, Dresden,
11 Germany

12 ³ Institute of Software and Multimedia Technology, TUD Dresden University of Technology, Dresden,
13 Germany

14 ⁴ Department of Hematology, Oncology and Bone Marrow Transplantation with Section of
15 Pneumology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

16 ⁵ Munich Leukemia Laboratory, Munich, Germany

17 ⁶ German Cancer Consortium (DKTK), Partner Site Dresden, and German Cancer Research Center
18 (DKFZ), Heidelberg, Germany

19 ⁷ National Center for Tumor Diseases Dresden (NCT/UCC), Dresden, Germany

20

21 **Running title: Deep learning detects MDS from bone marrow smears**

22 **Corresponding author:** Jan-Niklas Eckardt, MD, MSc, MHBA; Department of Internal Medicine I,
23 University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Fetscherstraße 74,
24 01307 Dresden, Germany; phone: +49 351 458 11542; e-mail: [jan-niklas.eckardt@uniklinikum-](mailto:jan-niklas.eckardt@uniklinikum-dresden.de)
25 [dresden.de](mailto:jan-niklas.eckardt@uniklinikum-dresden.de)

26 **Key words:** myelodysplastic neoplasms, deep learning, artificial intelligence, bone marrow

27 **Competing Interests:** The authors declare no competing interests.

28 text: 2190 words; abstract: 196 words; figures/tables: 3/3; supplementary figures/tables: 0/1;

29 references: 62

30 **Abstract**

31 Evaluation of bone marrow morphology by experienced hematologists is key in the diagnosis of
32 myeloid neoplasms, especially to detect subtle signs of dysplasia in myelodysplastic neoplasms
33 (MDS). The majority of recently introduced deep learning (DL) models in cytomorphology rely
34 heavily on manually drafted cell-level labels, a time-consuming, laborious process that is prone to
35 substantial inter-observer variability, thereby representing a substantial bottleneck in model
36 development. Instead, we used robust image-level labels for end-to-end DL and trained several state-
37 of-the-art computer vision models on bone marrow smears of 463 patients with MDS, 1301 patients
38 with acute myeloid leukemia (AML), and 236 bone marrow donors. For the binary classifications of
39 MDS vs. donors and MDS vs. AML, we obtained an area-under-the-receiver-operating-characteristic
40 (ROCAUC) of 0.9708 and 0.9945, respectively, in our internal test sets. Results were confirmed in an
41 external validation cohort of 50 MDS patients with corresponding ROCAUC of 0.9823 and 0.98552,
42 respectively. Explainability via occlusion sensitivity mapping showed high network attention on cell
43 nuclei not solely of dysplastic cells. We not only provide a highly accurate model to detect MDS from
44 bone marrow smears, but also underline the capabilities of end-to-end learning to solve the bottleneck
45 of time-consuming cell-level labeling.

46

47 **Introduction**

48 Myelodysplastic neoplasms (MDS) encompass clonal myeloid malignancies that are characterized by
49 ineffective hematopoiesis, cytopenia, myelodysplasia, and recurrent genetic events.¹ The incidence of
50 MDS appears to be underestimated and incidence rates increase dramatically over the age of 70 years
51 (up to an estimated 75:100,000 cases), representing a substantial societal burden in an aging
52 population.²⁻⁴ Although genetic findings are becoming increasingly important according to the new
53 WHO 2022 classification, accurate cytomorphologic evaluation of the bone marrow remains crucial
54 for the initial diagnosis, response assessment, and detection of disease transformation to acute myeloid
55 leukemia (AML).⁵ While counting myeloblasts is rather straightforward, signs of dysplasia are more
56 subtle and their accurate identification requires experienced investigators. Still, detection is often

57 challenging and prone to inter-observer variability, even for seasoned morphologists⁶⁻⁸, and shows
58 discrepancies between site and central review.⁹

59 In general, cytomorphologic evaluation of bone marrow aspirates in hematology remains essentially
60 unchanged over the last decades, as both preparation and evaluation are performed manually,
61 rendering the entire process time- and cost-intensive, as well as dependent on the experience and
62 subjective judgement of the observer.¹⁰⁻¹² With the advent of deep learning (DL) systems for computer
63 vision¹³, a multitude of applications in the healthcare sector have been identified where DL is applied
64 in image-based diagnostics.^{14,15} Convolutional neural nets (CNN), which consist of multiple artificial
65 neurons that are interconnected via convoluted deep layers, are commonly used for computer vision
66 tasks.¹⁶ In cytomorphology, recent studies have utilized neural networks in order to correctly classify
67 peripheral blood and bone marrow cells based on their respective morphology¹⁷⁻²⁸, as well as to
68 accurately identify myeloid malignancies.^{29,30}

69 In this study, we used an end-to-end DL system to accurately differentiate between MDS, AML, and
70 healthy donor bone marrow samples based on image-level labels, without the need for manually
71 labeling cells or dysplastic morphologies.

72

73 **Methods**

74 **Data sets**

75 We identified 463 MDS patients that have been previously diagnosed and treated at the University
76 Hospital Dresden, Germany. The first control group comprised 1301 AML patients that had been
77 diagnosed and treated under the auspices of the multicenter German Study Alliance Leukemia (SAL)
78 within the following previously reported multicenter trials: AML96³¹ [NCT00180115], AML2003³²
79 [NCT00180102], AML60+³³ [NCT 00180167], and SORAML³⁴ [NCT00893373]. Patients were
80 eligible upon diagnosis of MDS or AML according to the revised WHO/ICC criteria^{5,35}, age
81 ≥ 18 years, and available biomaterial at initial diagnosis including bone marrow smears. The second

82 control group consisted of 236 bone marrow samples from healthy bone marrow donors who
83 underwent allogeneic bone marrow donation at our center as previously reported.³⁶ An additional
84 external validation cohort was obtained from the Munich Leukemia Laboratory (MLL), Munich,
85 Germany, consisting of 50 patients with diagnosed MDS according to the above-mentioned eligibility
86 criteria. Prior to analysis, written informed consent was obtained from all patients and donors
87 according to the revised Declaration of Helsinki.³⁷ All studies were approved by the Institutional
88 Review Board of the TUD Dresden University of Technology (EK 98032010 and EK 289112008).

89

90 **Image digitization**

91 Bone marrow smears (BMS) were prepared from anticoagulated bone marrow according to WHO
92 guidelines.³⁸ Staining of MDS, AML, and donor BMS was performed with the May-Grünwald-Giemsa
93 method.¹¹ Image-level labels were derived from case-level diagnostics, including cytomorphology,
94 histology, cytogenetics and molecular genetics, previously documented for each case during routine
95 diagnostics or as part of the respective clinical trial. Using a Pannoramic 250 FLASH III
96 (3DHISTECH), we obtained high-resolution whole slide images. For every AML patient and bone
97 marrow donor, one image (50x magnification) per whole slide image was obtained using SlideViewer
98 (3DHISTECH). We assumed that subtle signs of dysplasia would not be fully captured in one field of
99 view alone. Therefore, for each MDS patient, we obtained four pictures (50x magnification) of
100 different areas of interest in the BMS.

101

102 **Deep learning**

103 *End-to-end image-level prediction on bone marrow slides*

104 We extended our previously described DL pipeline^{29,30} for binary image-level predictions for the
105 delineation of MDS, AML, and healthy controls. Based on case-level diagnosis, images were labeled
106 with either “MDS”, “AML”, or “healthy donor”. Importantly, no cell-level manual labeling was
107 performed. The pipeline was adapted to evaluate cases in a binary fashion, i.e. MDS vs. AML and

108 MDS vs. healthy donors. Potentially, imbalanced training data can bias a classifier towards the
109 predominant class. Considering the imbalances between the data sets (n=463 samples for MDS with 4
110 images per patient, resulting in 1852 MDS images in total; n=1301 samples for AML with 1 image per
111 patient; n=236 samples per donor with 1 image per donor), we used image augmentation techniques,
112 such as random sized cropping, color shifting and linear transformations, to balance the data sets for
113 each binary classification task. For all binary classifications, a 5-fold internal cross-validation was
114 used, i.e. a train-test-split of 80:20. Cases that were used for model training were strictly separated
115 from cases that were used for testing. In DL, determination of an optimal model cannot be done a
116 priori, but rather has to be evaluated given the specific use case, data set, and model architecture.
117 Hence, we evaluated six recently introduced DL architectures for computer vision including ResNet-
118 18/34/50/101/152³⁹, ResNeXt-50_32x4d/101_32x8d⁴⁰, Wide-ResNet-50/101⁴¹, DenseNet-
119 121/161/169/201⁴², ShuffleNet v2_x0_5/v2_x1_02⁴³, and SqueezeNet v1.1⁴⁴. All DL models were pre-
120 trained on ImageNet data.⁴⁵ The final architecture for each model was determined using automated
121 hyperparameter optimization with the Optuna framework.⁴⁶ DL models were implemented in Python
122 using the PyTorch framework. Computations were performed using the high-performance computing
123 (HPC) cluster of the TUD Dresden University of Technology.

124

125 **Performance evaluation**

126 Recall (*syn.*: sensitivity), precision (*syn.*: positive predictive value), and accuracy were used to
127 evaluate classification performances. Recall is defined as the fraction of all positive predictions among
128 all relevant events and precision is defined as the fraction of true positives among all positive
129 predictions. Further, the area-under-the-curve (AUC) was determined for the receiver-operating-
130 characteristic (ROC). All metrics are reported for each binary classification for the internal test sets as
131 well as for the external validation cohort with 95% confidence intervals.

132

133 **Explainability of classifications via occlusion sensitivity maps**

134 To highlight network attention and thereby identify morphological cues the network used to delineate
135 MDS, AML, and healthy donors, we used occlusion sensitivity maps (OSM). In OSM, random image
136 areas are iteratively blocked from view of the CNN and classification performance is measured. If the
137 blocked image area is highly relevant for accurate image-level classification, model performance will
138 drop accordingly. This process is repeated for the entire image. Thus, image areas that are crucial for
139 accurate predictions are highlighted so that morphologies that prompt the CNN classifier to predict a
140 label can be evaluated and interpreted.

141

142 **Results**

143 **End-to-end deep learning accurately delineates MDS from AML and healthy controls**

144 Baseline characteristics of the MDS patient cohort are shown in Table 1. We evaluated six different
145 neural network architectures³⁹⁻⁴⁴ for binary classification tasks iteratively. For the distinction between
146 MDS and healthy donors, we found Densenet-201⁴² to provide the highest classification performance,
147 with an accuracy of 0.97791 and a corresponding ROCAUC of 0.9708 (Table 2; Figure 1A). With
148 respect to delineating MDS from AML, the best results were obtained using the Squeezenet⁴⁴
149 architecture, resulting in an accuracy of 0.98072 and a ROCAUC of 0.9945 (Table 2; Figure 1B).
150 Detailed information on metrics and 95% confidence intervals of the best performing models for each
151 use-case is provided in Table 2. Individual model training on the HPC system for MDS vs. healthy
152 donors and MDS vs. AML took 20 hours each. An external validation set encompassing 50 MDS
153 patients was obtained from the Munich Leukemia Laboratory (MLL). Using our pre-trained models,
154 we achieved an accuracy of 0.9972 with a corresponding ROCAUC of 0.9823 in distinguishing
155 external MDS samples from healthy controls (Table 3, Figure 2A). With respect to delineating
156 external MDS samples from AML, an accuracy of 0.92104 was achieved with a ROCAUC of 0.98552
157 (Table 3, Figure 2B).

158

159 **Explainable predictions via occlusion sensitivity maps**

160 In order to make results interpretable to cytomorphologists, we used OSM that iteratively blocks
161 image areas from neural network evaluation and thus highlights (in red) image areas that are of high
162 importance for accurate predictions. In a proof-of-concept fashion, we found OSM to be cell-specific,
163 indicating that network attention is being focused on cells, rather than background or smudge.
164 Network attention was focused on cells in granulopoiesis and erythropoiesis and on megakaryocytes
165 (Figure 3). Interestingly, neural networks focused not only on signs of dysplasia, but also on cells we
166 deemed morphologically inconspicuous. High attention was given to defined signs of dysplasia
167 involving altered nuclear morphology such as chromatin clumping, dysfunctional segmentation, or
168 double nuclei. However, at times, high network attention was also given to cells with no apparent
169 dysplasia as per conventional definition,^{11,47} while network attention in these cells was also mainly
170 focused on the nucleus, sometimes including the perinuclear zone. This indicates more intricate and
171 subtle morphological alterations unquantifiable by human observers. However, other signs of
172 dysplasia, such as hypogranulation, were disregarded by our model. This could possibly be either due
173 to confidence saturation - meaning the model found enough reasons in a given field of view to
174 confidently predict MDS without paying attention to all apparent signs of dysplasia (defined or not) -
175 or low-ranking signs of dysplasia that were not learned in the training process due to their limited
176 weight in making accurate predictions.

177

178 **Discussion**

179 Using end-to-end DL, we developed a software framework to distinguish between MDS, AML, and
180 healthy controls with very high accuracy based on BMS from 2000 individual patients and bone
181 marrow donors. Importantly, we have demonstrated that information abstraction even in MDS with
182 often subtle morphologies is feasible using end-to-end learning, in contrast to recent studies in
183 hematology that primarily rely on the generation of cell-level labels.^{17-28,48,49} Using the latter approach,
184 a bottom-up system has to be devised where first thousands (usually hundreds of thousands) of labels
185 are required to build a robust classifier, and second individual cell-level predictions have to be
186 aggregated to generate a diagnosis-level prediction. Apart from being obviously time-consuming and

187 cost-ineffective, the generation of cell-level labels, i.e. the ground truth many classifiers in hematology
188 currently are based on, is flawed due to substantial classification biases. For instance, Sasada et al.⁷
189 evaluated divergence in cell classifications on 100,000 hematopoietic cells of 499 MDS patients. Up to
190 eleven experienced observers evaluated each cell image, however, only 55.6% of classifications were
191 found to match and especially low classification overlap was reported for dysplastic morphologies like
192 hypo-granularity and Pseudo-Pelger-Huët anomaly.⁷ This bottleneck and pitfall of cell-level labeling
193 can essentially be bypassed by an end-to-end approach, such as ours. Our ground truth labels are not
194 derived from subjective observer judgment. Instead, they are established by routine diagnostics,
195 including cytomorphology, histology, flow cytometry, cytogenetics, molecular genetics and clinical
196 examination, which provide image-level ground truth labels that are much more robust.

197 With respect to explainability, DL is often referred to as a ‘black box’.⁵⁰ The often elusive decision-
198 making of neural networks substantially hampers interpretability and thus acceptance of DL models in
199 such high-risk applications as cancer diagnostics. Using OSM (among other methods of
200 explainability⁵¹) not only enables internal proof-of-concept, but also provides additional information to
201 the human observer, as novel features that are important for prediction can be investigated that
202 otherwise would elude the human eye. Interestingly, our classifiers showed high attention for nuclei
203 not only of dysplastic cells, but also for cells that we did not deem to be morphologically suspicious
204 for dysplasia. Potentially, this alludes to a digital biomarker in MDS distinct from classical signs of
205 dysplasia. Future work will focus on correlating attention maps with genetic alterations and/or gene
206 expression in MDS. While certain molecular alterations have already been linked to certain
207 morphologies, such as mutated *SF3B1* in MDS with ringsideroblasts^{52,53}, CNNs can potentially be
208 used to identify novel gene-morphology links. For instance, Brück et al.⁵⁴ used CNNs on MDS bone
209 marrow core biopsies to predict mutations of *TET2*, spliceosome genes and monosomy 7. Further,
210 Nagata et al.⁵⁵ previously demonstrated a link between MDS morphology (assessed by pathologists)
211 and genomic profiles. This suggests a starting point for CNNs to link gene alterations with specific
212 morphologies and may potentially lead to an image-based predictor of genetic profiles and
213 consequentially patient risk and outcome.

214 Our study is limited by several factors. As is the case for most recent studies of computer vision in
215 (hemato-)pathology, our analysis is based on retrospective data. While external validation confirmed
216 high classification accuracy, prospective validation is still warranted. In our study, we differentiated
217 only between AML, MDS, and healthy bone marrow donors in a binary way. Still, some dysplastic
218 morphologies can also be present to a certain degree in non-malignant disorders⁵⁶ such as congenital
219 syndromes⁵⁷, nutritional deficiencies^{58,59}, infectious disease⁶⁰, and drug- or toxin-mediated bone
220 marrow damage^{61,62}. To increase routine applicability of our DL framework, future work will also
221 focus on acquiring image data from reactive and non-neoplastic specimen exhibiting bone marrow
222 dysplasia in order to make our classifier more versatile and applicable in clinical routine.

223 In summary, we have developed a DL framework trained on patient and donor samples, achieving
224 high accuracies in our internal test set and external validation set in distinguishing between MDS,
225 AML, and healthy bone marrow donors.

226

227 **List of Abbreviations**

228 AML – acute myeloid leukemia; BMS – bone marrow smear(s); CNN – convolutional neural net; DL
229 – deep learning; MDS – myelodysplastic neoplasm(s)

230

231 **Acknowledgements**

232 The authors are grateful to the Centre for Information Services and High-Performance Computing of
233 the TUD Dresden University of Technology for providing its facilities for training and execution of
234 deep learning models. This study was funded in part by Novartis Oncology. The funder had no role in
235 conceptualization, design, data collection, analysis, decision to publish, or preparation of the
236 manuscript.

237

238 Author Contributions

239 J-NE and JMM conceptualized the project. J-NE, FS, MEVG, KS, ASS, CR, UP, CP, TH, MB, and
240 JMM provided patient samples. J-NE, FS, MS, LR, and MEHG acquired BMS images. IS, TS, SR,
241 and KW developed computer vision models. All authors analyzed and interpreted the data. J-NE wrote
242 the initial draft. All authors approved the final version of the manuscript and agreed to be accountable
243 for all aspects of the work.

244

245 Competing Interests

246 J-NE, TS, SR, and JMM are co-owners of Cancilico.

247

248 References

- 249 1 Cazzola M. Myelodysplastic Syndromes. *N Engl J Med* 2020; **383**: 1358–1374.
- 250 2 Greenberg PL, Attar E, Bennett JM, Bloomfield CD, Borate U, De Castro CM *et al.*
251 Myelodysplastic syndromes: clinical practice guidelines in oncology. *J Natl Compr Canc*
252 *Netw* 2013; **11**: 838–874.
- 253 3 Neukirchen J, Schoonen WM, Strupp C, Gattermann N, Aul C, Haas R *et al.* Incidence
254 and prevalence of myelodysplastic syndromes: Data from the Düsseldorf MDS-registry.
255 *Leukemia Research* 2011; **35**: 1591–1596.
- 256 4 Cogle CR. Incidence and Burden of the Myelodysplastic Syndromes. *Curr Hematol Malig*
257 *Rep* 2015; **10**: 272–281.
- 258 5 Khoury JD, Solary E, Abla O, Akkari Y, Alaggio R, Apperley JF *et al.* The 5th edition of
259 the World Health Organization Classification of Haematolymphoid Tumours: Myeloid
260 and Histiocytic/Dendritic Neoplasms. *Leukemia* 2022; **36**: 1703–1719.
- 261 6 Goasguen JE, Bennett JM, Bain BJ, Brunning R, Vallespi M-T, Tomonaga M *et al.*
262 Dyserythropoiesis in the diagnosis of the myelodysplastic syndromes and other myeloid
263 neoplasms: problem areas. *Br J Haematol* 2018; **182**: 526–533.
- 264 7 Sasada K, Yamamoto N, Masuda H, Tanaka Y, Ishihara A, Takamatsu Y *et al.* Inter-
265 observer variance and the need for standardization in the morphological classification of
266 myelodysplastic syndrome. *Leuk Res* 2018; **69**: 54–59.
- 267 8 Valent P, Orazi A, Steensma DP, Ebert BL, Haase D, Malcovati L *et al.* Proposed
268 minimal diagnostic criteria for myelodysplastic syndromes (MDS) and potential pre-MDS
269 conditions. *Oncotarget* 2017; **8**: 73483–73500.

- 270 9 Zhang L, Stablein DM, Epling-Burnette P, Harrington AM, Moscinski LC, Kroft S *et al.*
271 Diagnosis of Myelodysplastic Syndromes and Related Conditions: Rates of Discordance
272 between Local and Central Review in the NHLBI MDS Natural History Study. *Blood*
273 2018; **132**: 4370.
- 274 10 Lee S-H, Erber WN, Porwit A, Tomonaga M, Peterson LC, International Council for
275 Standardization In Hematology. ICSH guidelines for the standardization of bone marrow
276 specimens and reports. *Int J Lab Hematol* 2008; **30**: 349–364.
- 277 11 Bain BJ, Clark DM, Wilkins BS. *Bone Marrow Pathology*. John Wiley & Sons, 2019.
- 278 12 Bain BJ, Bailey K. Pitfalls in obtaining and interpreting bone marrow aspirates: to err is
279 human. *J Clin Pathol* 2011; **64**: 373–379.
- 280 13 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional
281 neural networks. In: *Proceedings of the 25th International Conference on Neural*
282 *Information Processing Systems - Volume 1*. Curran Associates Inc.: Red Hook, NY,
283 USA, 2012, pp 1097–1105.
- 284 14 Chan H-P, Samala RK, Hadjiiski LM, Zhou C. Deep Learning in Medical Image
285 Analysis. *Adv Exp Med Biol* 2020; **1213**: 3–21.
- 286 15 Walter W, Haferlach C, Nadarajah N, Schmidts I, Kühn C, Kern W *et al.* How artificial
287 intelligence might disrupt diagnostics in hematology in the near future. *Oncogene* 2021;
288 **40**: 4271–4280.
- 289 16 Dhillon A, Verma GK. Convolutional neural network: a review of models, methodologies
290 and applications to object detection. *Prog Artif Intell* 2020; **9**: 85–112.
- 291 17 Matek C, Krappe S, Münzenmayer C, Haferlach T, Marr C. Highly accurate
292 differentiation of bone marrow cell morphologies using deep neural networks on a large
293 image data set. *Blood* 2021; **138**: 1917–1927.
- 294 18 Choi JW, Ku Y, Yoo BW, Kim J-A, Lee DS, Chai YJ *et al.* White blood cell differential
295 count of maturation stages in bone marrow smear using dual-stage convolutional neural
296 networks. *PLoS One* 2017; **12**: e0189259.
- 297 19 Kainz P, Burgsteiner H, Asslaber M, Ahammer H. Training echo state networks for
298 rotation-invariant bone marrow cell classification. *Neural Comput Appl* 2017; **28**: 1277–
299 1292.
- 300 20 Matek C, Schwarz S, Spiekermann K, Marr C. Human-level recognition of blast cells in
301 acute myeloid leukaemia with convolutional neural networks. *Nat Mach Intell* 2019; **1**:
302 538–544.
- 303 21 Putzu L, Caocci G, Di Ruberto C. Leucocyte classification for leukaemia detection using
304 image processing techniques. *Artif Intell Med* 2014; **62**: 179–191.
- 305 22 Rodellar J, Alférez S, Acevedo A, Molina A, Merino A. Image processing and machine
306 learning in the morphological analysis of blood cells. *Int J Lab Hematol* 2018; **40 Suppl**
307 **1**: 46–53.

- 308 23 Acevedo A, Merino A, Boldú L, Molina Á, Alférez S, Rodellar J. A new convolutional
309 neural network predictive model for the automatic recognition of hypogranulated
310 neutrophils in myelodysplastic syndromes. *Comput Biol Med* 2021; **134**: 104479.
- 311 24 Saraswat M, Arya KV. Automated microscopic image analysis for leukocytes
312 identification: a survey. *Micron* 2014; **65**: 20–33.
- 313 25 Wu Y-Y, Huang T-C, Ye R-H, Fang W-H, Lai S-W, Chang P-Y *et al.* A Hematologist-
314 Level Deep Learning Algorithm (BMSNet) for Assessing the Morphologies of Single
315 Nuclear Balls in Bone Marrow Smears: Algorithm Development. *JMIR Med Inform* 2020;
316 **8**: e15963.
- 317 26 Rezatofghi SH, Soltanian-Zadeh H. Automatic recognition of five types of white blood
318 cells in peripheral blood. *Comput Med Imaging Graph* 2011; **35**: 333–343.
- 319 27 Mori J, Kaji S, Kawai H, Kida S, Tsubokura M, Fukatsu M *et al.* Assessment of dysplasia
320 in bone marrow smear with convolutional neural network. *Sci Rep* 2020; **10**: 14734.
- 321 28 Fu X, Fu M, Li Q, Peng X, Lu J, Fang F *et al.* Morphogo: An Automatic Bone Marrow
322 Cell Classification System on Digital Images Analyzed by Artificial Intelligence. *Acta*
323 *Cytol* 2020; **64**: 588–596.
- 324 29 Eckardt J-N, Middeke JM, Riechert S, Schmittmann T, Sulaiman AS, Kramer M *et al.*
325 Deep learning detects acute myeloid leukemia and predicts NPM1 mutation status from
326 bone marrow smears. *Leukemia* 2022; **36**: 111–118.
- 327 30 Eckardt J-N, Schmittmann T, Riechert S, Kramer M, Sulaiman AS, Sockel K *et al.* Deep
328 learning identifies Acute Promyelocytic Leukemia in bone marrow smears. *BMC Cancer*
329 2022; **22**: 201.
- 330 31 Röllig C, Thiede C, Gramatzki M, Aulitzky W, Bodenstein H, Bornhäuser M *et al.* A
331 novel prognostic model in elderly patients with acute myeloid leukemia: results of 909
332 patients entered into the prospective AML96 trial. *Blood* 2010; **116**: 971–978.
- 333 32 Schaich M, Parmentier S, Kramer M, Illmer T, Stölzel F, Röllig C *et al.* High-dose
334 cytarabine consolidation with or without additional amsacrine and mitoxantrone in acute
335 myeloid leukemia: results of the prospective randomized AML2003 trial. *J Clin Oncol*
336 2013; **31**: 2094–2102.
- 337 33 Röllig C. Intermediate-dose cytarabine plus mitoxantrone versus standard-dose cytarabine
338 plus daunorubicin for acute myeloid leukemia in elderly patients. *Ann Oncol* 2018; **29**:
339 973–978.
- 340 34 Röllig C, Serve H, Hüttmann A, Noppeney R, Müller-Tidow C, Krug U *et al.* Addition of
341 sorafenib versus placebo to standard therapy in patients aged 60 years or younger with
342 newly diagnosed acute myeloid leukaemia (SORAML): a multicentre, phase 2,
343 randomised controlled trial. *Lancet Oncol* 2015; **16**: 1691–1699.
- 344 35 Arber DA, Orazi A, Hasserjian RP, Borowitz MJ, Calvo KR, Kvasnicka H-M *et al.*
345 International Consensus Classification of Myeloid Neoplasms and Acute Leukemias:
346 integrating morphologic, clinical, and genomic data. *Blood* 2022; **140**: 1200–1228.

- 347 36 Parmentier S, Kramer M, Weller S, Schuler U, Ordemann R, Rall G *et al.* Reevaluation of
348 reference values for bone marrow differential counts in 236 healthy bone marrow donors.
349 *Ann Hematol* 2020; **99**: 2723–2729.
- 350 37 World Medical Association. World Medical Association Declaration of Helsinki: ethical
351 principles for medical research involving human subjects. *JAMA* 2013; **310**: 2191–2194.
- 352 38 Swerdlow, SH, Campo, E, Harris, NL, Jaffe, ES, Pileri, SA, Stein, H *et al.* *WHO*
353 *Classification of Tumours of Haematopoietic and Lymphoid Tissues*.
354 [https://publications.iarc.fr/Book-And-Report-Series/Who-Classification-Of-](https://publications.iarc.fr/Book-And-Report-Series/Who-Classification-Of-Tumours/WHO-Classification-Of-Tumours-Of-Haematopoietic-And-Lymphoid-Tissues-2017)
355 [Tumours/WHO-Classification-Of-Tumours-Of-Haematopoietic-And-Lymphoid-Tissues-](https://publications.iarc.fr/Book-And-Report-Series/Who-Classification-Of-Tumours/WHO-Classification-Of-Tumours-Of-Haematopoietic-And-Lymphoid-Tissues-2017)
356 [2017](https://publications.iarc.fr/Book-And-Report-Series/Who-Classification-Of-Tumours/WHO-Classification-Of-Tumours-Of-Haematopoietic-And-Lymphoid-Tissues-2017) (accessed 19 Nov2024).
- 357 39 He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: *2016*
358 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE: Las
359 Vegas, NV, USA, 2016, pp 770–778.
- 360 40 Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated Residual Transformations for Deep
361 Neural Networks. In: *2017 IEEE Conference on Computer Vision and Pattern*
362 *Recognition (CVPR)*. 2017, pp 5987–5995.
- 363 41 Zagoruyko S, Komodakis N. Wide Residual Networks. 2017.
364 doi:10.48550/arXiv.1605.07146.
- 365 42 Huang G, Liu Z, Maaten L van der, Weinberger KQ. Densely Connected Convolutional
366 Networks. 2018. doi:10.48550/arXiv.1608.06993.
- 367 43 Zhang X, Zhou X, Lin M, Sun J. ShuffleNet: An Extremely Efficient Convolutional
368 Neural Network for Mobile Devices. 2017. doi:10.48550/arXiv.1707.01083.
- 369 44 Iandola FN, Han S, Moskewicz MW, Ashraf K, Dally WJ, Keutzer K. SqueezeNet:
370 AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. 2016.
371 doi:10.48550/arXiv.1602.07360.
- 372 45 Deng J. ImageNet: A large-scale hierarchical image database. In: *2009 IEEE Conference*
373 *on Computer Vision and Pattern Recognition* 248–255. 2009
374 doi:10.1109/CVPR.2009.5206848.
- 375 46 Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation
376 Hyperparameter Optimization Framework. In: *Proceedings of the 25th ACM SIGKDD*
377 *International Conference on Knowledge Discovery & Data Mining*. Association for
378 Computing Machinery: New York, NY, USA, 2019, pp 2623–2631.
- 379 47 Bain BJ. Diagnosis from the blood smear. *N Engl J Med* 2005; **353**: 498–507.
- 380 48 Lee N, Jeong S, Park M-J, Song W. Deep learning application of the discrimination of
381 bone marrow aspiration cells in patients with myelodysplastic syndromes. *Sci Rep* 2022;
382 **12**: 18677.
- 383 49 Kimura K, Tabe Y, Ai T, Takehara I, Fukuda H, Takahashi H *et al.* A novel automated
384 image analysis system using deep convolutional neural networks can assist to differentiate
385 MDS and AA. *Sci Rep* 2019; **9**: 13385.

- 386 50 Castelveccchi D. Can we open the black box of AI? *Nature* 2016; **538**: 20–23.
- 387 51 Joshi G, Walambe R, Kotecha K. A Review on Explainability in Multimodal Deep Neural
388 Nets. 2021. doi:10.48550/arXiv.2105.07878.
- 389 52 Malcovati L, Karimi M, Papaemmanuil E, Ambaglio I, Jädersten M, Jansson M *et al.*
390 SF3B1 mutation identifies a distinct subset of myelodysplastic syndrome with ring
391 sideroblasts. *Blood* 2015; **126**: 233–241.
- 392 53 Papaemmanuil E, Cazzola M, Boultonwood J, Malcovati L, Vyas P, Bowen D *et al.* Somatic
393 SF3B1 mutation in myelodysplasia with ring sideroblasts. *N Engl J Med* 2011; **365**:
394 1384–1395.
- 395 54 Brück OE, Lallukka-Brück SE, Hohtari HR, Ianevski A, Ebeling FT, Kovanen PE *et al.*
396 Machine Learning of Bone Marrow Histopathology Identifies Genetic and Clinical
397 Determinants in Patients with MDS. *Blood Cancer Discovery* 2021; **2**: 238–249.
- 398 55 Nagata Y, Zhao R, Awada H, Kerr CM, Mirzaev I, Kongkiatkamon S *et al.* Machine
399 learning demonstrates that somatic mutations imprint invariant morphologic features in
400 myelodysplastic syndromes. *Blood* 2020; **136**: 2249–2262.
- 401 56 Steensma DP. Dysplasia has A differential diagnosis: distinguishing genuine
402 myelodysplastic syndromes (MDS) from mimics, imitators, copycats and impostors. *Curr*
403 *Hematol Malig Rep* 2012; **7**: 310–320.
- 404 57 Iolascon A, Heimpel H, Wahlin A, Tamary H. Congenital dyserythropoietic anemias:
405 molecular insights and diagnostic approach. *Blood* 2013; **122**: 2162–2166.
- 406 58 Huff JD, Keung Y-K, Thakuri M, Beaty MW, Hurd DD, Owen J *et al.* Copper deficiency
407 causes reversible myelodysplasia. *Am J Hematol* 2007; **82**: 625–630.
- 408 59 Batata M, Spray GH, Bolton FG, Higgins G, Wollner L. Blood and bone marrow changes
409 in elderly patients, with special reference to folic acid, vitamin B12, iron, and ascorbic
410 acid. *Br Med J* 1967; **2**: 667–669.
- 411 60 Sheikha A. Dyserythropoiesis in 105 patients with visceral leishmaniasis. *Lab Hematol*
412 2004; **10**: 206–211.
- 413 61 Michot F, Gut J. Alcohol-induced bone marrow damage. A bone marrow study in
414 alcohol-dependent individuals. *Acta Haematol* 1987; **78**: 252–257.
- 415 62 Dusse LMS, Moreira AMB, Vieira LM, Rios DRA, Silva RMM e, Carvalho M das G.
416 Acquired Pelger-Huët: what does it really mean? *Clin Chim Acta* 2010; **411**: 1587–1590.

417

418 **Tables**

419 **Table 1. MDS patient characteristics**

Parameter	
N	463
Age in years, median (IQR)	66 (18-89)
Sex, %	
Male	59
Female	41
MDS type (WHO 2022), %	
MDS-5q	10
with <i>SF3B1</i> mutation	0.1
with <i>TP53</i> mutation	0.1
MDS <i>biTP53</i>	0.2
MDS <i>SF3B1</i>	8.8
MDS-LB	29.3
MDS, hypoplastic	2.7
MDS-IB1	15.6
MDS-IB2	20
MDS with fibrosis	1.3
MDS/MPN-RS-T	0.5
CMML-1	1.2
CMML-2	9.8
IPSS-R, %	
Very low risk	6.4
Low risk	21.6
Intermediate risk	38.7
High risk	22.3
Very high risk	10.9
Blood count	
WBC in GPt/l, median (IQR)	3.58 (0.57-91.1)
Hb in g/dl, median (IQR)	9.9 (4.4-15.6)
Plt in GPt/l, median (IQR)	96 (3-1531)
PB blasts in %, median (IQR)	0 (0-15)
BM blasts in %, median (IQR)	5.5 (0-26.0)

420 *BM* bone marrow, *CMML-1/2* chronic myelomonocytic leukemia subgroup 1/2, *Hb* hemoglobin, *MDS*
421 myelodysplastic neoplasm, *MDS biTP53* MDS with biallelic *TP53* inactivation, *MDS-5q* MDS with
422 low blasts and isolated 5q deletion, *MDS-IB1/2* MDS with increased blasts 1/2, *MDS-LB* MDS with
423 low blasts, *MDS/MPN-RS-T* myelodysplastic/myeloproliferative neoplasm with ring sideroblasts and
424 thrombocytosis, *MDS-SF3B1* MDS with low blasts and *SF3B1* mutation, *N* number, *PB* peripheral
425 blood, *Plt* platelet count, *WBC* white blood cell count.

426 **Table 2. Test set performance for binary image-level classifications.**

	MDS vs. healthy donors		MDS vs. AML	
DL architecture	Densenet-201		Squeezenet v1.1	
Accuracy	0.97791 [0.9561 - 0.9977]		0.98072 [0.9686 - 0.9904]	
	MDS	Healthy donors	MDS	AML
Precision	0.9973 [0.9948 - 1.0]	0.8547 [0.7121 - 0.9791]	0.97065 [0.9637 - 0.9967]	0.98118 [0.9565 - 0.9904]
	MDS	Healthy donors	MDS	AML
Recall	0.9775 [0.9507 - 0.9974]	0.9787 [0.9574 - 1.0]	0.98180 [0.9565 - 0.9906]	0.98030 [0.9616 - 0.9968]
ROCAUC	0.9708 [0.9241 - 0.9893]		0.9945 [0.98824 - 0.9984]	

427 Brackets indicate 95% confidence intervals. *AML* acute myeloid leukemia, *DL* deep learning, *MDS*

428 myelodysplastic neoplasm, *ROCAUC* area-under-the-curve of the receiver-operating-characteristic.

429 **Table 3. Model performance on external validation set**

	MDS (MLL cohort) vs. healthy donors		MDS (MLL cohort) vs. AML	
DL architecture	Densenet-201		Squeezenet v1.1	
Accuracy	0.9972 [0.9811 - 0.9963]		0.92104 [0.8905 - 0.9567]	
	MDS	Healthy	MDS	AML
Precision	0.9925 [0.9892 - 1.0]	0.9852 [0.9787 - 0.9957]	0.91418 [0.8880 - 0.9398]	0.94668 [0.8245 - 1.0]
	MDS	Healthy	MDS	AML
Recall	0.9970 [0.9940 - 0.9980]	0.9938 [0.9755 - 1.0]	0.97516 [0.9139 - 1.0]	0.80834 [0.7375 - 0.8667]
ROCAUC	0.9823 [0.9593 - 0.9972]		0.98552 [0.9746 - 0.9951]	

430 Brackets indicate 95% confidence intervals. *AML* acute myeloid leukemia, *MDS* myelodysplastic
431 neoplasm, *MLL* Munich Leukemia Laboratory, *ROCAUC* area-under-the-curve of the receiver-
432 operating-characteristic.

433 **Figure Legends**

434 **Figure 1: Performance of deep learning models for binary classifications delineating MDS,**
435 **AML, and healthy donors.** The receiver-operating characteristic (ROC) with the corresponding area-
436 under-the-curve (AUC) is depicted for the best performing models for each classification task. For
437 MDS vs. healthy donors, best results were achieved with Densenet-201 (A). For MDS vs. AML, best
438 results were achieved with Squeezenet (B). Internal cross-validation was performed with an 80:20
439 split. Individual run performance (Fold 1-5; graphs in light blue, orange, green, red, and purple) as
440 well as aggregate macro average performance (graph in dark blue) are reported. Only testing results
441 are reported.

442

443 **Figure 2: External validation of deep learning models for binary classifications delineating**
444 **MDS, AML, and healthy donors.** The receiver-operating characteristic (ROC) with the
445 corresponding area-under-the-curve (AUC) is depicted for the best performing models for the binary
446 classifications MDS (MLL) vs. healthy donors (A) and MDS (MLL) vs. AML (B). Individual run
447 performance (Fold 1-5; graphs in light blue, orange, green, red, and purple) as well as aggregate macro
448 average performance (graph in dark blue) are reported.

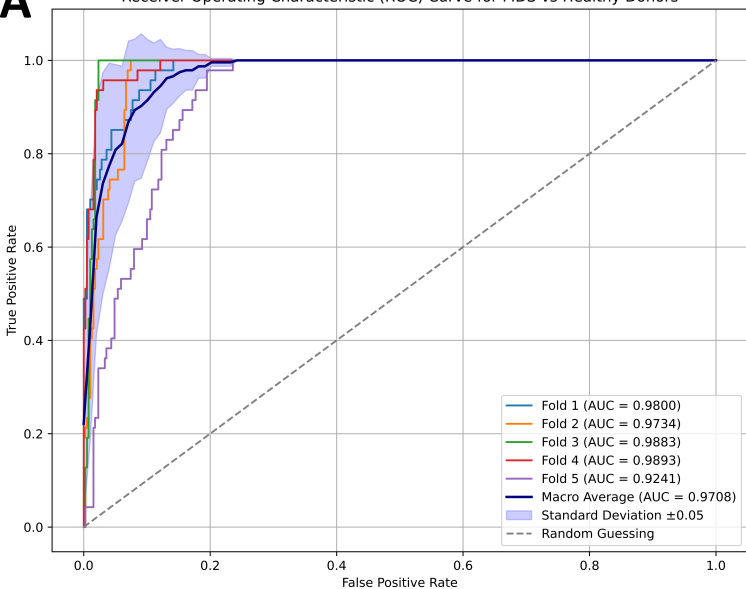
449

450 **Figure 3: Occlusion Sensitivity Mapping (OSM) highlights network attention for explainable**
451 **output interpretation.** OSM iteratively blocks image areas from being evaluated by the deep learning
452 network. If an image area is highly important for classification, the network's performance will thus
453 drop substantially in the given iteration. Image areas that are of high importance for correct
454 classification can thereby be highlighted (high attention shown in red). A standard field of view of
455 bone marrow smears from MDS patients is shown in A, C, and E. The corresponding OSM is
456 displayed in B, D, and F, respectively. First, in a proof-of-concept fashion, the network focuses its
457 attention on cells and specifically on nuclei. It does not consider background, noise or smudge as
458 important for classification. Second, high attention is directed at erythropoietic and granulopoietic
459 cells as well as megakaryocytes.

Figure 1

A

Receiver Operating Characteristic (ROC) Curve for MDS vs Healthy Donors



B

Receiver Operating Characteristic (ROC) Curve for MDS vs AML

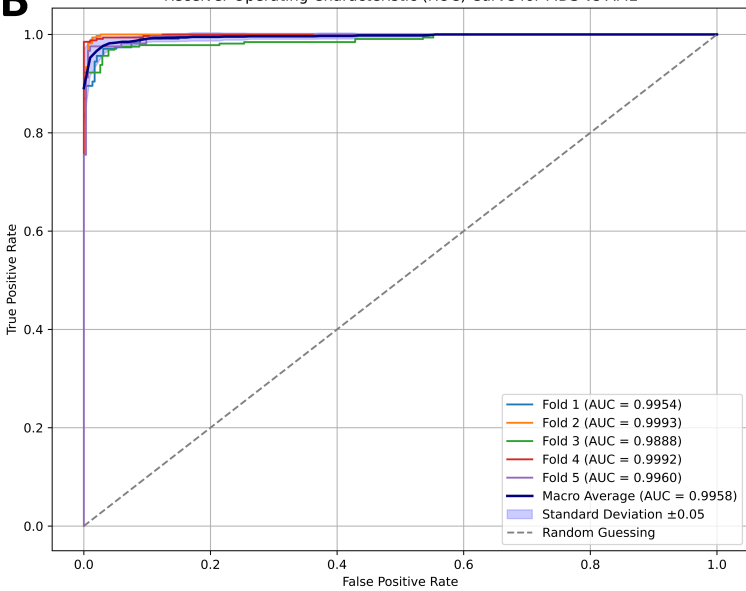
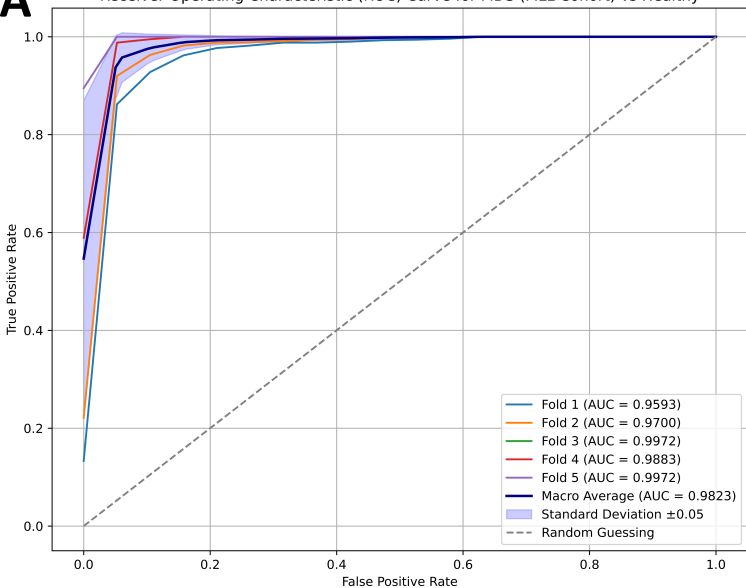


Figure 2

A

Receiver Operating Characteristic (ROC) Curve for MDS (MLL Cohort) vs Healthy



B

Receiver Operating Characteristic (ROC) Curve for MDS (MLL) vs AML

