


Models for the marrow: A comprehensive review of AI-based cell classification methods and malignancy detection in bone marrow aspirate smears

Tabita Ghete^{1,2}  | Farina Kock³ | Martina Pontones^{1,2} | David Pfrang³ | Max Westphal³ | Henning Höfener³ | Markus Metzler^{1,2,4}

Correspondence: Markus Metzler (markus.metzler@uk-erlangen.de)

Abstract

Given the high prevalence of artificial intelligence (AI) research in medicine, the development of deep learning (DL) algorithms based on image recognition, such as the analysis of bone marrow aspirate (BMA) smears, is rapidly increasing in the field of hematology and oncology. The models are trained to identify the optimal regions of the BMA smear for differential cell count and subsequently detect and classify a number of cell types, which can ultimately be utilized for diagnostic purposes. Moreover, AI is capable of identifying genetic mutations phenotypically. This pipeline has the potential to offer an accurate and rapid preliminary analysis of the bone marrow in the clinical routine. However, the intrinsic complexity of hematological diseases presents several challenges for the automatic morphological assessment. To ensure general applicability across multiple medical centers and to deliver high accuracy on prospective clinical data, AI models would require highly heterogeneous training datasets. This review presents a systematic analysis of models for cell classification and detection of hematological malignancies published in the last 5 years (2019–2024). It provides insight into the challenges and opportunities of these DL-assisted tasks.

INTRODUCTION

The progress of artificial intelligence (AI) is an important milestone in medicine. In particular, it has become an invaluable tool in radiology and digital pathology due to its exceptional image recognition capabilities.^{1–3} Similarly, the evaluation of differential blood counts, one of the most common diagnostic tests in medicine, is essentially based on image recognition of typical physiological and pathological blood cells. This process is now implemented in analyzers for daily clinical routine.⁴

In the diagnosis of bone marrow (BM) disorders, such as acute myeloid leukemia (AML), acute lymphoblastic leukemia (ALL), chronic myeloid leukemia (CML), and myelodysplastic syndrome (MDS), manual microscopic analysis of BM morphology remains the primary diagnostic tool. However, the quantitative and qualitative analysis of BM cells poses a number of additional challenges. In contrast to peripheral blood (PB), bone marrow aspirate (BMA) contains a significantly higher number of cell classes that are not distributed in their physiological, spatially fixed environment as in histology. The selection of representative areas in the context of smear preparation has a significant impact on the result.

Cellular trails behind BM particles with well-spread cells are considered optimal for obtaining cytological detail.⁵ Furthermore, the cell maturation process is continuous, with subtle morphological changes within lineages, which can lead to subjectivity and high inter-observer variability in cell differentiation.⁶ Although the process of manually counting and classifying several hundred nucleated cells is time-consuming, it remains the fastest means of obtaining a basis for determining the most appropriate therapeutic approach. In emergency situations, such as acute promyelocytic leukemia (APL), morphologic detection and correct classification of abnormal promyelocytic blasts is essential for prompt therapeutic decision-making and initiation of treatment. As the availability of trained medical personnel is becoming increasingly limited, this may lead to a bottleneck in routine clinical care.

In recent years, highly sophisticated deep learning (DL) algorithms have been developed and applied for the accurate detection and classification of cells in both PB^{7,8} and BMA smears.^{9,10} A simplified workflow is presented in Figure 1. In leukemia and MDS, where the distinction between physiological and leukemic or dysplastic cells is of paramount importance for risk stratification and therapeutic

¹Department of Pediatrics and Adolescent Medicine, University Hospital Erlangen, Erlangen, Germany

²Bavarian Cancer Research Center (BZKF), Erlangen, Germany

³Computational Pathology, Fraunhofer Institute for Digital Medicine (MEVIS), Bremen, Germany

⁴Comprehensive Cancer Center Erlangen-EMN (CCC ER-EMN), Erlangen, Germany

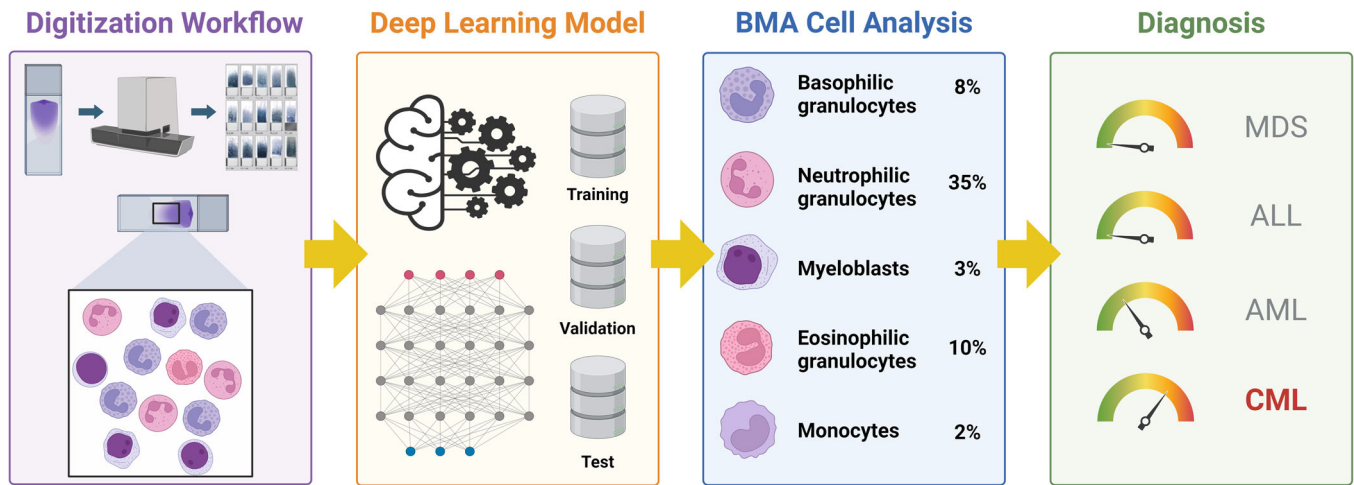


FIGURE 1 Example of an end-to-end pipeline for deep learning-based cell analysis and subsequent diagnosis on a digitized bone marrow aspirate smear from a patient with chronic myeloid leukemia. Created with BioRender.com.

decision-making, AI models show promising performance.^{11–13} In addition, automated end-to-end analysis systems, such as MorphoGo and Scpio Labs Full-Field Bone Marrow Aspirate™ application, include hardware and software components for digitization of BMA smears and subsequent AI-assisted detection and classification of nucleated cells.^{14–18} It is therefore important to gain a deeper understanding of its strengths and limitations as a support mechanism for filtering urgent cases.

Medical devices with AI software must conform to challenging regulatory aspects to ensure the system's transparency and data protection.¹⁹ PB analyzers have set rigorous standards for clinical validation, with an increased focus on the explainability of the implemented neural networks. Moreover, the real-world performance of these devices demonstrates accuracy and reliability in clinical settings. Recently, the Scpio Labs Full-Field Bone Marrow Aspirate™ application received clearance from the Food and Drug Administration (FDA), thus marking an important milestone as the first application for AI-based analysis of BMA smears.

Consequently, the following review presents a systematic compilation of individual steps required for the evaluation of BMA smears with regard to the automatic classification of cells. The aim is to offer an overview of the research that was carried out in the last 5 years (2019–2024). The following terms and their combinations were used to search relevant publications on PubMed: “artificial intelligence,” “machine learning,” “deep learning,” “convolutional neural network,” “bone marrow,” “leukemia,” “myelodysplastic syndrome,” and “cell classification.” Only models trained and used on digitized BMA smears were included.

In addition, we provide an analysis of what is already available and identify where further development is needed. Performance metrics of the reviewed publications are not directly comparable, as the majority of publications implemented their own methods on their own datasets, and few were validated on external datasets. Moreover, the tasks (region selection, cell classification, diagnosis prediction) and statistical methodology varied between publications. If available, 95% confidence intervals (CI) were reported.

Bone marrow morphology in hematological malignancies

While the use of advanced diagnostic tools such as flow cytometry and cytogenetic analysis is increasing, traditional morphologic examination remains the gold standard for diagnosing hematologic disorders.

Leukemia can occur in acute or chronic form. The acute form can be characterized by the presence of $\geq 20\%$ immature cells, called blasts, of myeloid (AML)²⁰ or lymphoid (ALL)²¹ lineage in the BM, which suppresses normal hematopoiesis. In addition, acute leukemias present with a diverse spectrum of immunophenotypes and aberrations in hematopoietic-associated genes, such as mutations in *RUNX1* and *KMT2A*.²²

CML is a myeloproliferative disorder characterized by clonal expansion of myelopoietic cells and the presence of the *BCR::ABL1* oncogene.^{23,24} In contrast to AML, in which the myeloid lineage is typically represented by myeloblasts and mature granulocytes (*hiatus leucaemicus*), the BM morphology of a patient with CML in the chronic phase shows hypercellularity (Figure 2), blast cells below 10%, and a complete spectrum of granulocytic progenitors (left shift), with an increased myeloid to erythroid ratio. CML may progress to a rapidly growing acute leukemia (blast phase), which morphologically resembles ALL or AML, as it presents with an increase ($>20\%$) in lymphoblasts or myeloblasts, respectively.²⁵

MDS is a heterogeneous class of hematologic disorders that affect the BM. With the exception of MDS with excess blasts (MDS-EB), which can present with up to 19% blasts, MDS typically presents with $<5\%$ blasts in BM.²⁰ Cells from erythropoietic, megakaryopoietic, or granulopoietic lineages exhibit morphologic abnormalities, including nuclear hypersegmentation, megaloblastic changes, or asynchronous nuclear and cytoplasmic maturation. The complexity of MDS and the subjective assessment of dysplastic features imply a high inter-observer variability.^{26,27} MDS can progress to AML through the accumulation of myeloid blasts in PB and/or BM. In addition, BM architecture, cellularity, or tissue composition may be altered in patients with hematologic neoplasms. AI models were successfully implemented to estimate cellularity^{28,29} and classify cell lineages in BM biopsies, as well as to predict diagnosis, genetic aberrations, and progression to AML in patients with myelodysplastic/myeloproliferative disorders.³⁰ These models, which can navigate the complex BM morphologic landscape, may provide a valuable complement to BMA smears for an inclusive BM analysis pipeline.

A deep dive into deep learning

AI has been integrated into a multitude of applications, including healthcare and medicine. To facilitate the interpretation of results

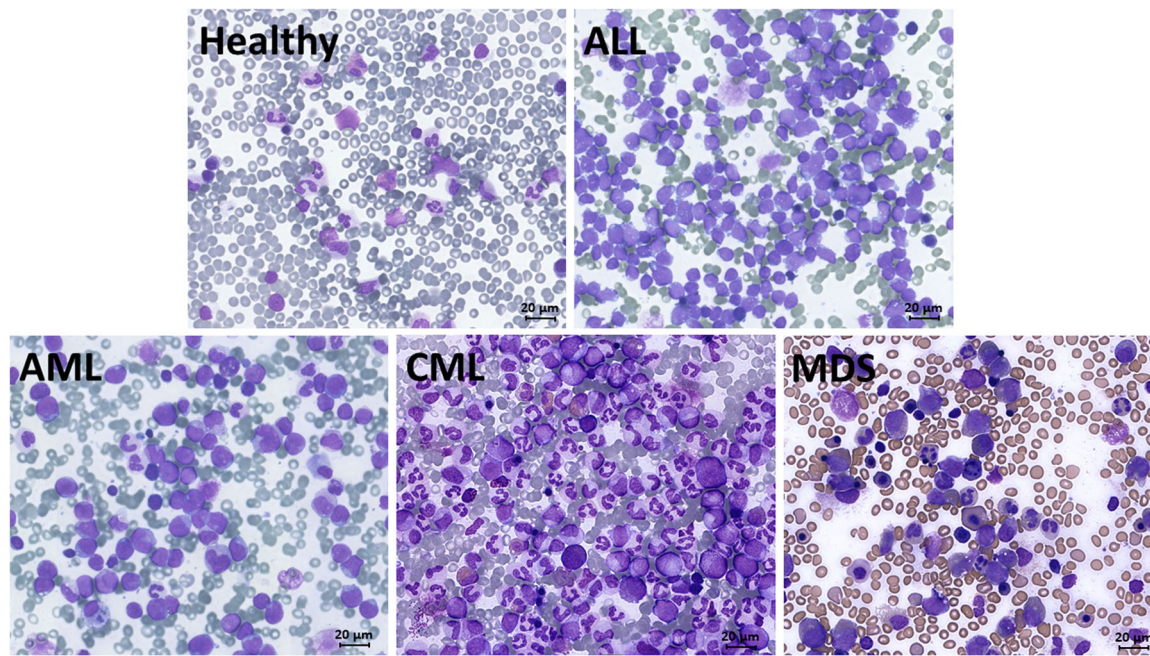


FIGURE 2 Overview of high-level morphology ($\times 400$ magnification) observed in bone marrow aspirate smears from healthy individuals and patients with acute lymphoblastic leukemia, acute myeloid leukemia, chronic myeloid leukemia, and myelodysplastic syndrome.

addressed in the following chapters, we provide a short overview of the terminology and definitions.

Machine learning (ML) is a specific branch of AI that specializes in pattern recognition and the ability of a system to learn about data through the application of supervised or unsupervised methods. While supervised learning is used to predict outcomes using labeled data, unsupervised learning can identify patterns or clusters in unlabeled data. There are different task types in supervised learning (regression, binary classification, multiclass classification, multi-label classification, object detection, segmentation), depending on the type of outcome variable that is predicted. In our context, the most relevant task types are segmentation for region selection, object detection for identifying cells, and multiclass classification for both cell classification and diagnosis prediction. Hereby, each observation is classified into exactly one of multiple target classes (e.g., cell type or disease diagnosis). Popular applicable learning algorithms for such a task are (multinomial) logistic regression, decision trees, random forests, support vector machines, and neural networks, among others. Deep learning is a subset of ML that involves the training of neural networks with multiple layers. The goal of deep neural networks (DNNs) is to model complex patterns in data by optimizing the success of the learning process and the network's self-learning ability.³¹ Convolutional neural networks (CNNs) are a prominent example of DL models that gained considerable traction in the field of image classification and segmentation. Another type of supervised learning is multiple instance learning (MIL), which presents an alternative approach for diagnosis prediction in cases where patient-level annotations are available instead of single-cell annotations. Here, labels are associated with sets of instances as opposed to single instance label.³² Rather than instructing the model to identify all cells as X, we train it to recognize a set of samples, at least one of which is labeled X, and thus categorize the entire set as "positive."

Model evaluation can be based on a variety of different metrics. Commonly used metrics include sensitivity, specificity, accuracy, precision, and F1 score. In the case of multiclass classification, these

metrics can and should be estimated and reported per class, but could also be further aggregated (e.g., mean sensitivity over all classes). The precision of a model is defined as the proportion of positive predictions that are correct. In contrast, the recall of a model is the proportion of positives that a model is able to identify. The F1 score is a metric that represents the harmonic mean of precision and recall and indicates the percentage of correct predictions made by the model. It is thus one possibility to find a balance between precision and recall. Another metric to evaluate the performance of a classification model is the AUC (area under the receiver operator characteristic curve; sometimes also abbreviated as AUROC), as it shows how well the model can distinguish between a healthy person and a patient with leukemia, for example. In contrast to previously mentioned metrics, AUC evaluates a model's performance across multiple thresholds. Therefore, it can only be calculated for prediction models that output a continuous value for class membership. In circumstances where the performance of all classes is of equal importance, it is preferable to employ macro-averaging of the AUC, as opposed to micro-average, which is a global average.

To improve the generalizability of a model and to minimize bias in performance estimates, available data should be split up into train, validation, and test datasets for the different purposes of model training, selection, and testing, respectively. Typically, it is necessary to utilize a large number of training images to achieve a high level of prediction performance. This depends on the task difficulty and in some cases, tens of training images may already yield a satisfactory performance.³³ The advancement of image processing algorithms has led to the development of increasingly sophisticated architectures for the multiclass classification of cell types. In a comparative study, VGG, ResNet, RegNet, and Transformer architectures were tested against the original ResNeXt-50 model proposed by Matek et al.,⁹ using the same dataset of BM cell images. The models performed equally good or better than ResNeXt-50. However, there is no evidence to suggest that complex architectures with a high number of parameters are associated with superior accuracy. In this instance, the

ResNet model with the lowest number of parameters demonstrated the second highest level of accuracy.³⁴

Saliency maps, class activation mapping (CAM), and gradient-weighted CAM (Grad-CAM) are methods that facilitate the interpretation of model predictions and the identification of potential shortcomings, by highlighting which pixels (features) are most relevant for the respective classification. In this way, misclassifications of dysplastic granulocytes can be traced back to a feature focus on cytoplasmic pixels instead of the nucleus, as in correct predictions.³⁵

To enlarge the training image dataset and to prevent overfitting, data augmentation can be implemented.³⁶ Data augmentation involves modifying, for example, the orientation, brightness, or contrast of the training images to create slight variations and thus enhance the variability of the dataset. In contrast, synthetic data involves the artificial creation of new images. These methods are especially important when dealing with high variability between center-dependent staining or scanning methods.

As the capabilities of DL models have steadily increased over the past years, the digitization of tissue samples using whole slide scanners has become increasingly common.³⁷ Nevertheless, the resulting whole slide images (WSIs) may exhibit considerable technical variability between different scanners.³⁸ This, in addition to variations in tissue preparation and staining protocols, can pose a challenge to the general implementation of DL models across multiple datasets.

PIPELINE FOR THE DIGITAL ANALYSIS OF BONE MARROW

Selection of optimal areas in bone marrow aspirate smears

In order to automatically analyze cells in BMA smears, single cells must be detected. Comprehensive reviews of cell detection and/or segmentation models used in PB and BMA smears have already been published,^{39,40} therefore this review focuses on models for region detection and cell classification. Due to the heterogeneous cell density in BMA, selecting the right region for cell classification is imperative. Several studies used DL algorithms to automatically identify optimal regions for cell classification in WSIs of BMA smears.

The three-component ROI-BMC-DNN framework established by Su et al. works by segmenting the BM particle region and sampling the periphery for optimal patches with good cell distribution. The segmentation model achieved recall and precision

of 0.858 and 0.885, respectively.⁴¹ Similarly, Tayebi et al. built a pipeline to identify appropriate or inappropriate region of interest (ROI) tiles for cell classification in BMA smears from patients with a wide range of hematologic disorders, including carcinoma, MDS, and hypo- and hypercellular slides. Using a DenseNet-121 architecture for binary classification between ROI tiles, the model achieved cross-validation accuracy and precision of 0.97 and 0.90, respectively.⁴²

Wang et al. proposed a model where the first layer CNN model detects BM particles and cellular trails at low resolution, and the second layer CW-Net performs cell detection and classification at high resolution. With this particular architecture, BM particles and cell traces were detected with a precision of 1.00 and an accuracy of >0.930.⁴³ Similarly, Lewis et al. implemented a preliminary slide region CNN model classifier to distinguish between optimal regions near BM particles and regions without cells. The model discriminated between four region classes, including “optimal,” “particle,” “hemodiluted,” and “outside,” with an AUC of >0.999.⁴⁴

Distinguishing which areas of the BMA smear are appropriate for differential cell counting is an important step (Figure 3), as it allows for effective cytologic and morphologic analyses. It also limits the potential for diagnostic inconsistencies.⁵ In this context, the models presented here provide a practical and efficient basis for subsequent cell classification.

Challenges and opportunities in DL-assisted cell classification

Due to the higher cell density and presence of different cells at different stages of maturation, automated examination of BM samples is more challenging compared to PB. Nevertheless, several DL models have been successfully implemented for this purpose (Table 1 and Supporting Information S1: Table 1). Figure 4 provides an overview of the most frequently classified cell types in the reviewed publications.

Detection and classification of blast cells is of paramount importance for rapid diagnosis of leukemia. In 2020, Chandradevan et al. used cell images from non-neoplastic BM smears to train a VGG16 cell classification algorithm that was implemented to recognize blast cells in samples from three patients with AML, with a resulting AUC of 0.893. As expected, a subset of blast cells was incorrectly predicted as promyelocytes, highlighting the subtle intralinear maturation process.⁴⁸ The same misclassification was also observed by Lewis et al.⁴⁴ and Wang et al.⁶³

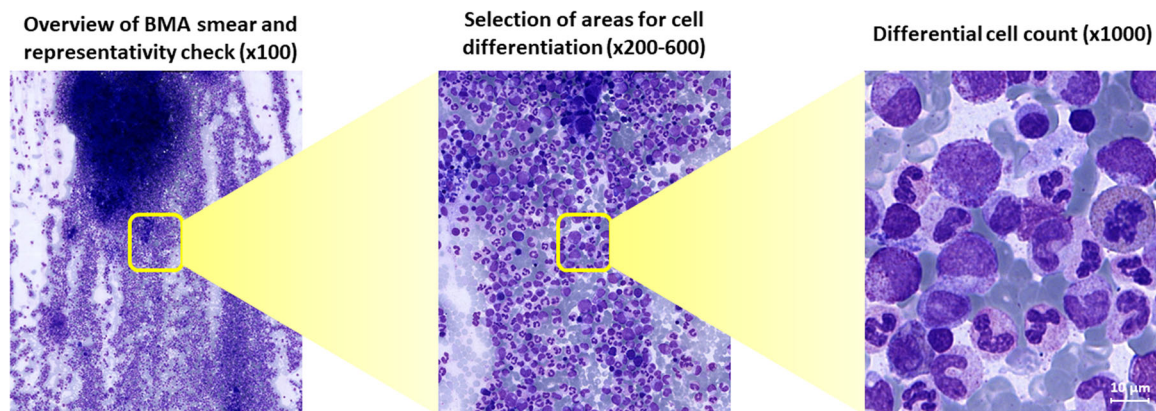


FIGURE 3 Workflow of a hematologist: quality check and establishment of specimen's representativity with $\times 10$ objective (bone marrow particles, cellularity assessment); search of optimal areas for cell classification with $\times 20$ – 60 objective; differential cell count with $\times 100$ objective.

TABLE 1 Summary of studies with AI application on BMA smears including single-cell classification; 95% CI was included, if available. (continued on next page)

Refs.	Year	BM dataset	No. of cell types	Model	Metrics for cell/disease classification	Additional input/output data
[45]	2019	48,037 images of BM samples	5	Support vector machine (SVM)	<ul style="list-style-type: none"> • Mean precision 0.938 • Mean recall 0.875 	No
[46]	2019	150 BMA smears from patients with hematological disease, including ALL, MDS, MPN, MM, AA	11	Path-Aggregation Network	<ul style="list-style-type: none"> • Mean average precision (mAP) 0.659 • Accuracy 0.801 	No
[47]	2019	580 images from 6 noncancerous BMA and 19 cancerous PB samples	8	Decision tree classifier	<ul style="list-style-type: none"> • Sensitivity 0.983 • Specificity 0.994 • Accuracy 0.990 	No
[48]	2020	BMA smears from patients with AML (3), MM (2), and nonneoplastic patients (17)	12	VGG16 convolutional network (CN)	<p>For all cell types (macro AUC):</p> <ul style="list-style-type: none"> • AUC 0.912 (AML) • AUC 0.906 (MM) • AUC 0.982 ± 0.03 (nonneoplastic) 	No
[15]	2020	3000 BMA smears	12	Artificial neural network (ANN) with 27 layers (Morphogo)	<ul style="list-style-type: none"> • Overall accuracy 0.901 (95% CI, 0.898–0.905) 	No
[49]	2020	35 BMA smears from patients with MDS or other hematological disease; 1797 cells	3	Faster R-CNN with ResNet-101 backbone	<ul style="list-style-type: none"> • AUC 0.944 • Sensitivity 0.910 • Specificity 0.977 • Accuracy 0.972 	No
[50]	2020	BMA smears from patients with ALL, AML, MDS, AA, MM, MPD, lymphoma; 17,319 annotated cells	8	BMSNet with YOLO v3 architecture	<ul style="list-style-type: none"> • AUC 0.948 (>5% blasts) • AUC 0.942 (>20% blasts) • Correlation AI/hematologists: >0.845 	No
[51]	2021	609 single-cell images	7	Faster RCNN with Feature Pyramid Network (FPN)	<ul style="list-style-type: none"> • Average precision (AP) 0.744 	No
[9]	2021	BMA smears from 945 patients with several hematological disorders; 171,374 single-cell images	21	ResNeXt-50	<ul style="list-style-type: none"> • Mean precision 0.51 • Mean recall 0.689 	No
[52]	2021	230 BMA images; 8239 single-cell images	8	Cell Recognition Network (CRNet)	<ul style="list-style-type: none"> • AP 0.953 • AR 0.958 	No
[10]	2021	1732 BMA images; 24,165 cells; 2983 cell debris	20	Combination of three ResNet models (ResNext101_32x8d swsl, ResNext50_32x4d swsl and ResNet50)	<p>Classification of countable cells:</p> <ul style="list-style-type: none"> • Precision 0.861 • Accuracy 0.829 • F1 score 0.820 <p>ALL diagnosis:</p> <ul style="list-style-type: none"> • Sensitivity 0.86 • Specificity 0.95 • Accuracy 0.89 	Diagnosis
[53]	2022	Dataset from Matek et al.	20	Probabilistic Siamese network with triplet loss function	<ul style="list-style-type: none"> • Weighted AP 0.93 • AR 0.91 • Accuracy 0.84 • F1 score 0.91 	No

TABLE 1 (Continued)

Refs.	Year	BM dataset	No. of cell types	Model	Metrics for cell/disease classification	Additional input/output data
[43]	2022	BMA smears from 37 patients (8 AML, 2 MDS, 2 CMML, 3 ALL, 6 MM, 1 CML, 12 non-neoplastic) and 3 normal smears	17	Cascade R-CNN	<ul style="list-style-type: none"> Recall 0.842 Accuracy 0.988 	No
[42]	2022	BMA smears from 51 APL, 1,048 non-APL AML, 236 healthy donors	3	Cell classification: Xception CNN Disease classification: Binary ensemble neural nets (ENNs)	<ul style="list-style-type: none"> AUC 0.874 (myeloblast) AUC 0.920 (promyelocyte) AUC 0.836 (Auer rods) APL vs. healthy: macro AUC 0.959 (95% CI, 0.933–0.984) APL vs. non-APL AML: macro AUC 0.858 (95% CI, 0.783–0.932) 	Diagnosis
[54]	2022	7484 images of BM cells	15	Class balance classification method (CBCM) with Resnet152 backbone	<ul style="list-style-type: none"> Precision 0.845 ± 0.002 Sensitivity 0.844 ± 0.007 Specificity 0.993 	No
[55]	2022	12,756 BMA cell images; 20,421 cell images from PB smears	12	C-WGAN-GP (Wasserstein GAN with gradient penalty) and Sequential CNN model	<p>On synthetic dataset:</p> <ul style="list-style-type: none"> Precision 0.969 Recall 0.966 Specificity 0.951 Accuracy 0.970 F1 score 0.960 	No
[35]	2022	BMA smears from 34 MDS, 24 healthy patients; 8065 cells	8	InceptionV3 architecture	<ul style="list-style-type: none"> AUC 0.945–0.996 Sensitivity 0.640–0.900 Specificity 0.948–0.944 Accuracy 0.912–0.993 F1 score 0.643–0.938 	No
[56]	2022	47 BMA smears	17	Semi-supervised learning (SSL) using confirmed self-training (CST)	<ul style="list-style-type: none"> AP 0.977 AR 0.976 Total accuracy 0.976 	No
[42]	2022	250 WSI of BM	19	ROI detection: DenseNet 121 architecture Cell detection and classification: YOLO+AL	<p>ROI detection:</p> <ul style="list-style-type: none"> Precision 0.90 AUC 0.99 Accuracy 0.97 <p>Cell classification:</p> <ul style="list-style-type: none"> AP 0.83 AR 0.75 F1 score 0.78 	No
[57]	2023	1306 single-cell images from the dataset from Matek et al.	7	Integrated fine-tuned DenseNet121 with an attention mechanism	<ul style="list-style-type: none"> Accuracy 0.970 	No
[58]	2023	16,456 annotated cells	19	Detection of BM particles: CNN model Cell detection and classification: CW-Net	<p>For every cell type:</p> <ul style="list-style-type: none"> Recall >0.95 Accuracy >0.99 	No
[59]	2023	1204 BM cell images; 13,059 single-cell images	15	YOLOv7-CTA	<ul style="list-style-type: none"> Precision 0.784 Recall 0.851 	No

TABLE 1 (Continued)

Refs.	Year	BM dataset	No. of cell types	Model	Metrics for cell/disease classification	Additional input/output data
[60]	2023	41,595 single-cell images from BMA smears of 50 patients with normal BM morphology	23	ResNeXt-50	<ul style="list-style-type: none"> • Mean precision 0.89 • Mean recall 0.89 • Mean AUC 0.99 	No
[44]	2023	10,948 BM regions; 23,609 annotated cells	16	Region classification: CNN model Cell classification: Efficient NetV2L backbone CNN with pretrained ImageNet weights	<ul style="list-style-type: none"> • Mean AUC 0.999 Cell classification (excluding unknown intact and disrupted cells) <ul style="list-style-type: none"> • Mean AUC > 0.95 	No
[16]	2023	508 BMA smears; 385,207 nucleated cells	25	Morphogo	<ul style="list-style-type: none"> • Sensitivity 0.810 • Specificity 0.995 • Accuracy > 0.956 (for every cell type) 	No
[61]	2023	236 healthy BM; 1095 AML, of which 43 APL	6	MILIE (Multiple Instance Learning for Leukocyte Identification)	<ul style="list-style-type: none"> • AUC 0.895 (promyelocytes) • AUC 0.862 (myeloblast) Distinction between AML and APL: <ul style="list-style-type: none"> • AUC 0.99 	Diagnosis
[62]	2023	Dataset from Matek et al.; additional 1363 BM cells from 67 patients with leukemia (CMU dataset)	21	DAGDNet (Dual attention gate denseNet) model	<ul style="list-style-type: none"> • Mean precision 0.881 on the MLL dataset • Mean precision 0.903 on the CMU dataset 	No
[41]	2023	WSI data set containing 120 BM images; 230 manually labeled patches	8	ROI-BMC-DNNNet	Region segmentation: <ul style="list-style-type: none"> • Precision 0.885 • Recall 0.858 Cell classification: <ul style="list-style-type: none"> • Mean precision: 0.950 • Mean recall: 0.946 	No
[63]	2023	728 BM smears; 11,788 BMA images; 131,300 single-cell images; AL subtype prediction: 40 healthy, 40 patients with leukemia	19	MLFL-Net (Multi-Level Feature Learning Network)	Cell classification: <ul style="list-style-type: none"> • Precision 0.756–0.976 • Recall 0.694–0.976 • Total accuracy 0.895 AL type prediction: <ul style="list-style-type: none"> • 92.5% of cases labeled identically between model and experts 	Diagnosis
[17]	2023	333 BMA smears including CML, ET, ITP, AL, and patients with normal BM	4	Morphogo	<ul style="list-style-type: none"> • Precision 0.829 • Recall 0.825 • Sensitivity 0.966 • Specificity 0.897 • Accuracy 0.944 • F1 score 0.827 	No
[34]	2024	Dataset from Matek et al.	21	Regnet_y_32gf	<ul style="list-style-type: none"> • Mean precision 0.787 ± 0.060 • Mean recall: 0.755 ± 0.061 • Mean F1 score: 0.762 ± 0.050 	No

TABLE 1 (Continued)

Refs.	Year	BM dataset	No. of cell types	Model	Metrics for cell/disease classification	Additional input/output data
[64]	2024	BMA smears from 408 patients with AML; >2,000,000 single-cell images	4	Two CNNs based on ResNet18; Cell Filtering Model (CFM) + Genetic Feature Extraction Network (GFEN)	<p>Cell classification (two CNNs):</p> <ul style="list-style-type: none"> • Accuracy 0.82/0.88, and 0.94/0.91, respectively <p>AUC for genetic prediction (temporal validation):</p> <ul style="list-style-type: none"> • 0.64 (95% CI, 0.4–0.84) - ELN 2017 favorable risk • 0.68 (95% CI, 0.35–0.89) - MIRC cytogenetic • 0.70 (95% CI, 0.53–0.87) - NPM1 • 0.72 (95% CI, 0.54–0.87) - FLT3-ITD • 0.91 (95% CI, 0.6–1.00) - CBFB::MYH11 	Diagnosis; 5 genetic categories
[65]	2024	270 BMA smear images	8	Cell Detection and Confirmation Network (CDC-NET)	<ul style="list-style-type: none"> • Precision 0.917 ± 0.031 • Recall 0.785 ± 0.042 • F1 score 0.846 ± 0.038 	No

Abbreviations: AA, aplastic anemia; AL, acute leukemia; ALL, acute lymphoid leukemia; AML, acute myeloid leukemia; AP, average precision; APL, acute promyelocytic leukemia; AR, average recall; AUC, area under the receiver operating characteristic curve; BM, bone marrow; BMA, bone marrow aspirate; CI, confidence interval; CML, chronic myeloid leukemia; CMML, chronic myelomonocytic leukemia; ET, essential thrombocythemia; ITP, immune thrombocytopenia; MDS, myelodysplastic syndrome; MM, multiple myeloma; MPD, myeloproliferative disease; MPN, myeloproliferative neoplasm.

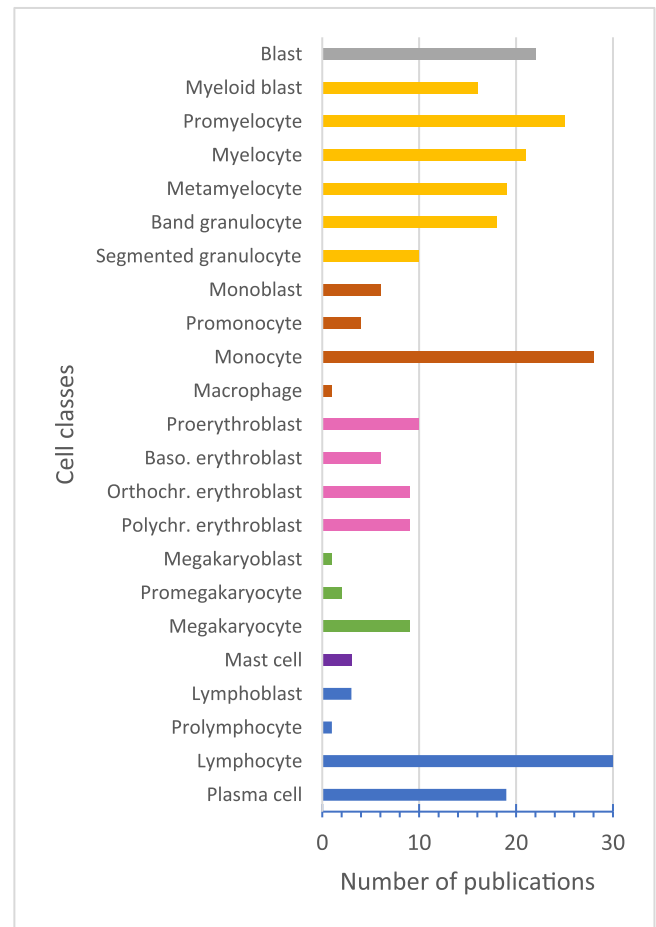


FIGURE 4 Distribution of classifications across cell lineages in the reviewed publications.

By providing a large publicly available image dataset and the classification of 21 cell types, including immature lymphocytes and erythroid progenitor cells, the work of Matek et al. represents a significant advance in the field of automated cell classification in BMA smears. Using the proposed ResNeXt-50 cell classification model, blast cells were correctly identified with a precision and recall of 0.75 and 0.65, respectively.⁹

Several studies were conducted using the dataset from Matek et al. for model training^{34,53,57,62} or evaluation.⁶³ In 2022, a Siamese classification network was proposed; it differs from ResNeXt-50 in that it does not extract features from individual images. Instead, it prioritizes similarity and dissimilarity between images of the same and different classes, respectively. This approach achieved classification precision and recall of blast cells of 0.89.⁵³ Unlike other models, the DAGDNet (Dual attention gates denseNet) architecture proposed by Peng et al. suppresses background signals in single-cell images to minimize their impact on network features. While the model achieved a mean precision of 0.881 on the Matek et al. dataset, the performance of the blast cell classification was lower than that of the Siamese network (precision and recall of 0.852 and 0.834, respectively).⁶² Another DenseNet121 model, fine-tuned with an attention mechanism, achieved an accuracy of 0.97 in classifying cells from the same dataset. However, only seven cell classes were selected, none of which included blast cells.⁵⁷

Automatic differentiation of blasts of different cell lineages, such as myeloblast, monoblast, and lymphoblast, is a valuable tool for distinguishing acute lymphoid and myelocytic leukemia, as well as certain subtypes like myelomonocytic or monoblastic/monocytic leukemia. Wang et al. proposed MLFL-Net, a model comprising a ResNet-50 backbone and supplementary branches designed for the precise classification of 19 BM cell types. For myeloblast and immature lymphocyte, this model achieved precision of 0.893 and 0.949, respectively, and recall of 0.903 and 0.968, respectively. In contrast, precision for the monocytic lineage (monoblast, promonocyte, monocyte) ranged between 0.756 and 0.891, while recall was between 0.694 and 0.733. With the exception of acute myelomonocytic leukemia (one correct vs. four false predictions), the entities were accurately predicted based on the percentual cell distribution. However, the sample size is limited (three cases of acute monoblastic/monocytic leukemia).⁶³

The French-American-British (FAB) classification indicates that a high percentage of erythroid precursors, proerythroblasts, is characteristic of acute erythroid leukemia (FAB AML M6), while megakaryoblasts are characteristic of acute megakaryoblastic leukemia (FAB AML M7).²⁰ Of the reviewed publications, the classification of megakaryoblasts and promegakaryocytes has only been included in the model proposed by Zhou et al. The researchers combined three ResNet architectures into an ensemble model and achieved accuracy, AUC, and F1 scores of 0.829, 0.987, and 0.829, respectively, across 20 types of cells. Average precision for the classification of lymphoblasts, megakaryoblasts, promegakaryocytes, and proerythroblasts ranged from 0.8 to 0.9, indicating that the model may be successfully applied in predicting certain subtypes of leukemia. However, 12% of myeloblasts were incorrectly identified as lymphoblasts, which could have implications for the following diagnosis.¹⁰ Notably, the Morphogo system also classifies promegakaryocytes and several forms of megakaryocytes.¹⁷

While some studies group all erythroid maturation stages together as one cell type classification,^{41,50,54} others automate the distinction between multiple stages of erythroblasts (basophilic, polychromatic, orthochromatic).^{10,54,60} Similarly, the eosinophilic granulocyte is predominantly classified as a merged class,^{9,42,63,66} or as an immature versus mature cell,⁶⁰ which is the standard approach in most clinical situations. However, the differentiation of all precursor forms appears to be possible with the assistance of DL-based classification.^{16,59}

An increase in basophils is observed in myeloproliferative diseases and has been identified as a key indicator of CML progression.^{67,68} In this regard, fast and accurate classification of basophils, as reported in several publications,^{48,55,60} may be beneficial in clinical risk assessment.

An important study that paved the way for image feature recognition in BMA smears from patients with MDS was conducted by Mori et al. The study proposed labeling cells according to the degree of neutrophil dysplasia. A Faster R-CNN architecture was trained to distinguish between cells with the following attributes: normal, intermediate, dysplasia, and severe dysplasia using single-cell images of BMA smears from patients with MDS and patients with non-MDS diseases. The model achieved an AUC of 0.944 and an accuracy of 0.972.⁴⁹

In comparison, Lee et al. expanded the feature range and included dysplasia from all three cell lineages (dysgranulopoiesis, dyserythropoiesis, and dysmegakaryopoiesis). The algorithm shows promising results for distinguishing between normal and dysplastic cells in BMA smears, with AUC ranging from 0.945 to 0.996, and F1 score between 0.643 and 0.938.³⁵

The application of DL enables rapid identification of an elevated blast percentage in BM, which may assist in the early detection of

disease progression, such as from MDS to AML or CML in the chronic phase to CML in the blast phase.

Disease prediction models

The diagnostic process for hematological diseases is complex and time-consuming, requiring a significant investment of resources. Nevertheless, the performance of recently employed DL models shows potential for automating a part of this process, specifically with regard to both MDS and leukemia. The relatively monotonous blast cell morphology observed in ALL (Figure 2) allows for accurate prediction of diagnosis based on BMA smears by DL models.^{69,70} Notably, Zhou et al. developed a system for the diagnosis of leukemia, which first identifies and excludes cells that are crushed or uncountable, and then subsequently classifies the remaining leukocytes. The system was able to predict the diagnosis of ALL with a sensitivity of 0.86 and specificity of 0.95.¹⁰ A summary of the publications that have implemented DL for disease recognition without single-cell classifications is presented in Table 2 and Supporting Information S1: Table 2.

The application of AI to distinguish between various morphological subtypes of ALL represents a promising area of research. For example, FAB classification ALL-L1, -L2, and -L3 can be accurately identified in PB smear images.^{73,74} Furthermore, the classification of B-cell or T-cell ALL, which is currently distinguished using flow cytometry and cytogenetic analyses, may be more beneficial in clinical settings, as they are usually linked to different outcomes.⁷⁵

It is crucial to promptly diagnose APL as it is considered a hematological emergency. The unequivocal diagnosis of APL is based on the distinctive morphology of promyelocytic blasts and the presence of Auer rods, in combination with the cytogenetic characteristics of chromosomal translocation t(15;17) and the fusion gene *PML::RARA*.⁷⁶ To the best of our knowledge, Ouyang et al. were the first to apply CNNs for APL detection in BM with an average precision of 0.625.⁷¹ Eckardt et al. demonstrated the ability of DL to distinguish between APL, non-APL AML, and healthy BM in a small dataset using only BMA smear images. The implemented model consisted of a multi-step ML workflow with individual DL models for different binary tasks. Mean AUC for APL versus healthy BM was 0.959 (95% CI, 0.933–0.984), while the mean AUC for APL versus non-APL AML cases was 0.858 (95% CI, 0.783–0.932).¹² Although this method necessitates the use of cell class labels, the detection of APL can also be achieved through the application of annotation-free DL. In a study by Manescu et al., MILLIE (Multiple Instance Learning for Leukocyte Identification) was trained with patient diagnosis-level labels alone, resulting in an average AUC of 0.99 in distinguishing between the same entities.⁶¹

CML is rather a rarity in the field of DL-based disease prediction.⁷⁷ In the study by Huang et al., the diagnosis of three types of leukemia, including AML, ALL, and CML, achieved a prediction accuracy of over 0.95 by incorporating three different frameworks to construct classification models.⁷⁰ The model was trained and tested using a preselected set of images of BMA smears containing multiple cells. Alternatively, CML could be diagnosed in the future using single-cell classification models that can accurately differentiate between multiple maturation stages of myeloid and erythroid cells,^{10,43,56,60} whose ratio is increased in the BM of patients with CML.²⁵

The characteristics of dysplasia in MDS are often subject to interpretation, which can lead to diagnostic difficulty. In comparison to hematologists, the AI model proposed by Wu et al. demonstrated superior performance in cases of MDS with <5% blasts (AUC of 0.929 and 0.948, respectively), but inferior to that of pathologists

TABLE 2 Summary of studies with AI application on BMA smears for disease recognition excluding single-cell classification; 95% CI was included, if available.

Refs.	Year	BM dataset	Model	Metrics for disease classification	Additional input/output data
[69]	2020	BMA images: 90 ALL, 100 MM	DCNN (Dense Convolutional Neural Network)	ALL vs. MM: <ul style="list-style-type: none"> Precision 1.000 Recall 0.940 Specificity 0.952 Accuracy 0.973 F1 score 0.969 	Diagnosis
[70]	2020	1322 BM cell images from 104 subjects (18 healthy, 53 AML, 23 ALL, and 18 CML patients)	DenseNet121	Accuracy: <ul style="list-style-type: none"> 0.90 (healthy BM) 0.99 (AML) 0.97 (ALL) 0.95 (CML) 	Diagnosis
[71]	2021	13,504 BMA images from patients with APL or other disease	Augmented pretrained Mask R-CNN	<ul style="list-style-type: none"> AP 0.625 AR 0.841 	Diagnosis
[72]	2022	1251 patients with AML; 236 healthy BM	AML vs. healthy: Xception CNN; NPM1 mutation: ResNet50	AML vs. healthy: <ul style="list-style-type: none"> Macro AUC 0.970 (95% CI, 0.968–0.972) NPM1 mutation status: <ul style="list-style-type: none"> AUC 0.92 (95% CI, 0.877–0.963) Accuracy 0.86 	Diagnosis; NPM1 mutation status
[41]	2022	Model development, internal validation: 115 BM smears (32 MDS, 26 AA, 57 AML)	ResNet-50	MDS or not MDS: <ul style="list-style-type: none"> AUC 0.985 (95% CI, 0.979–0.991) Sensitivity 0.992 (95% CI, 0.980–1.000) Specificity 0.881 (95% CI, 0.854–0.908) Accuracy 0.914 (95% CI, 0.895–0.934) Distinction MDS, AA, or AML: <ul style="list-style-type: none"> AUC 0.968 (95% CI, 0.960–0.976); Sensitivity 0.857 (95% CI, 0.828–0.886) Specificity 0.967 (95% CI, 0.956–0.978) Accuracy 0.929 (95% CI, 0.916–0.941) 	Diagnosis
[66]	2023	8245 BMA images from 651 patients with AML	AMLnet with EfficientNet backbone	AML vs healthy: <ul style="list-style-type: none"> Accuracy >0.9 Distinction of 9 AML subtypes: <ul style="list-style-type: none"> AUC 0.885 (95% CI, 0.874–0.897) – image level AUC 0.921 (95% CI, 0.915–0.927) – patient level 	Diagnosis

Abbreviations: AA, aplastic anemia; ALL, acute lymphoid leukemia; AML, acute myeloid leukemia; AUC, area under the receiver operating characteristic curve; BM, bone marrow; BMA, bone marrow aspirate; CI, confidence interval; CML, chronic myeloid leukemia; MDS, myelodysplastic syndrome; MM, multiple myeloma.

(AUC of 0.985). Nevertheless, when the blast percentage exceeds 20%, a slight decline in performance was observed (AUC of 0.981 and 0.942, respectively).⁵⁰ A comparable performance was demonstrated by the CNN model proposed by Wang et al., which is capable of recognizing MDS based on images obtained from BMA smears. The model achieved an accuracy of 0.914, AUC of 0.985, and sensitivity of 0.992 in a binary classification to determine whether the patient had MDS or not. A three-way classification model was successfully employed to differentiate between three disorders (aplastic anemia, MDS, or AML) with AUC of 0.968, accuracy of 0.929, and sensitivity of 0.857 (95% CI in Table 2).¹¹

It is crucial to acknowledge that DL models are trained on retrospective data and necessitate a higher degree of generalizability, demonstrated in (prospective) external validation studies, before being employed in the fast-paced clinical setting. Furthermore, cytomorphology represents only one diagnostically relevant aspect, which must often be integrated with molecular and cytogenetic classification.

STREAMLINING GENETIC PROFILING

Although cytomorphological findings may reflect genetic abnormalities, such as AML with t(8;21) presenting with blast cells with an indented nucleus,⁷⁸ genetic profiling in AML is crucial for risk stratification and therapeutic decision-making. However, conventional methods can require days or weeks to acquire, which may result in delays in the implementation of targeted therapeutic strategies. Kockwelp et al. developed a pipeline to predict favorable or high-risk (HR) genetic abnormalities in AML based solely on morphological characteristics derived from BM samples obtained at the time of diagnosis. This approach markedly accelerates the genetic analysis process. The five profiles included “ELN 2017 favorable risk,” “MRC cytogenetic,” “*NPM1* mutations,” “*FLT3*-ITD mutations,” and “*CBFB::MYH11*,” and could be predicted with AUC of 0.64, 0.68, 0.70, 0.72, and 0.91, respectively (95% CI in Table 1).⁶⁴

NPM1 mutated AML cases have been associated with cup-like nuclei in blast cells,⁷⁹ but it is noteworthy that the proposed DL model from Eckardt et al. focused on novel distinct morphological features for mutation prediction. A multi-step workflow incorporating several DL models was employed to accurately predict the *NPM1* mutation status based on BM cytomorphology, with an accuracy of 0.86 (AUC 0.92; 95% CI, 0.877–0.963). Furthermore, mutated *NPM1* was identified based on condensed chromatin and perinuclear lightening zones in myeloblasts, while wild-type *NPM1* was associated with prominent nucleoli.⁷²

Application of AI-based prediction models for identifying genetic abnormalities associated with AML could facilitate faster diagnosis and risk assessment, thereby offering a promising direction for future investigations and applications in other forms of leukemia with yet unidentified cytological characteristics, such as AML with *RUNX1* mutation,⁷⁸ which is associated with poor prognosis and short overall survival.⁸⁰

THE FUTURE OF AI-ASSISTED HEMATOPATHOLOGY

The utilization of AI in the diagnosis of leukemia based on morphological analysis of BMA smears offers a multitude of advantages over traditional methods. First, the rapid analysis of thousands of cells surpasses the quantity of manually counted cells, thereby reducing statistical uncertainty regarding the diagnosis. Recently, DeepHeme was developed as a cell classification model that was trained using single-cell images from both pediatric and adult patients with normal BM morphology. The model achieved a mean AUC of 0.99 (precision

0.98; recall 0.89) across 23 distinct cell classes. In competition with hematopathologists, DeepHeme classified 25 images per cell class in 0.36 s with a mean precision and recall of 0.9, while the same number of cells were classified in three hours by the hematopathologists with lower performance (mean precision 0.78; mean recall 0.76).⁶⁰ In this format, the diagnostic process may be streamlined by providing rapid preliminary assessments.

Second, ML models are capable of identifying patterns in vast amounts of data that may be difficult for clinicians to detect. A comprehensive study of 1,079 patients with myelodysplastic and myeloproliferative neoplasms, conducted by Nagata et al., revealed correlations between morphological profiles and genetic abnormalities. The large cohort was clustered into five distinct morphological profiles, including both HR and low-risk (LR) MDS. In addition, six and eight specific genetic profiles were identified for HR and LR, respectively. Interestingly, 77% of patients with HR MDS were classified in one morphological profile, while all LR patients were distributed among the remaining four groups. In addition, a total of 52 associations between morphology and genotype were identified, some of which were novel associations. For instance, the correlation between mutated *STAG2* and *SRSF2* with myeloid dysplasia, and mutated *ASXL1* with megakaryocytic dysplasia, were identified.⁸¹

This study could pave the way for the discovery of additional potential pathognomonic relationships in other disease entities through the use of AI, such as a transition from CML to blast phase, or further elucidate the genetic abnormalities associated with leukemic predisposition.

Furthermore, ML models can predict remission and overall survival in patients with AML^{82,83} and MDS⁸⁴ using clinical parameters alone. The models proposed by Eckardt et al. demonstrated the ability to predict complete remission with AUC ranging from 0.77 to 0.86. Additionally, the models exhibited AUC between 0.63 and 0.74 for the 2-year overall survival.⁸² Although the selected predictive features were already known from previous studies, the utilization of ML represents a significant advancement in the discovery of new biomarkers for the prediction of remission and survival. Similarly, Didi et al. trained neural networks with 52 diagnostic variables and achieved an accuracy of over 0.62 for predicting the overall survival of patients with AML.⁸³ Furthermore, MDS could be predicted one year prior to diagnosis with an AUC of 0.87 without cytogenetics or blast cell percentage as input data. This demonstrated that AI can identify individuals at risk without the need for information from invasive procedures such as BM biopsies.⁸⁴ Using image markers from BMA smears, ML predicts the risk of relapse in patients who have undergone hematopoietic cell transplantation.⁸⁵ The models could be employed as a decision-support system for risk stratification in routine clinical settings. While automatic cell classification can be a valuable tool, there are some limitations regarding the input data. When the model is trained on images previously annotated by experts, the subjective assessment of cell morphology may be easily transferred to the model as a biased interpretation. In addition, it is inevitable that certain cells will be manually categorized into the “all-inclusive” class of unknown or unidentifiable cells. This may be attributed to technical factors, such as the use of unfocused images or contrast that is too high, or to morphological considerations, including the difficulty in distinguishing between different maturation stages. These cells present a challenge to the model due to the high degree of heterogeneity within the class, and it is therefore expected that they will be predicted with low accuracy.^{9,44}

A high quantity of data is typically associated with enhanced classification accuracy, particularly in the case of rare or morphologically heterogeneous cells, such as reactive lymphocytes. The presence of a class imbalance in BM specimens is a persistent challenge in the training of cell classification models, given that certain cell classes are physiologically unevenly distributed, and can lead

to poor performance on minority classes. In such cases, it is recommended that metrics for each cell class be reported. Furthermore, macro averages may be used instead of micro averages, as the macro average treats all classes as equally weighted. Guo et al. proposed a solution that harmonizes data within 15 cell classes using a three-component class balance classification method (CBCM) and improves overall accuracy (0.895 vs. 0.909). This method includes data pre-processing, fine-tuning with pre-trained models, and class-balanced focal loss. The model achieved a precision of 0.845 and a specificity of 0.993.⁵⁴ An alternative approach to addressing the issue of low numbers of cell images was proposed by Hazra et al., who combined different databases and generated synthetic single-cell images. A total of 12 cell classes were classified with an accuracy of 0.97 using the WGAN-GP with an additional classifier. Compared to the performance on the original dataset, the model achieved higher accuracy, specificity, and sensitivity on the balanced dataset (>0.84 vs. >0.95).⁵⁵

To facilitate the prospective clinical integration of AI models, quality control steps, which may be second nature to medical personnel during the microscopic evaluation of a BMA smear, must be integrated into model development. These checkpoints include the detection of BM particles as indicative of a representative sample,⁴³ estimations of cellularity,²⁹ and detection and correction of unfocused images.⁸⁶ It should be noted that CNN models are not always universally deployable, as samples must be uniformly preprocessed and analyzed. Semi-automated staining devices, like CellaVision RAL® StainBox, reduce staining variability and may improve model performance. Variations in aspirate thickness and the presence of BM particles may change the sample focus point in the scanning process, resulting in unfocused areas during the scanning process. These areas can be automatically detected⁸⁶ and re-scanned at varying focal planes along the vertical (z)-axis to combine into a single composite image. Translating these quality control steps into quantifiable metrics not only increases the robustness of the model but also offers clinicians a more transparent pipeline.

The initial deployment of AI models in clinical settings would be as a preliminary tool, to be verified by clinicians, with the final objective of establishing a fully automated end-to-end pipeline, such as that exemplified by automated PB analyzers. In a modular interface, the user is presented with preselected areas of the sample, images of labeled single cells, a report including absolute and relative numbers of cell types, as well as a predicted diagnosis based on the distribution of cell types. Moreover, the user has the option to validate or correct the cell label. For experienced users, a module with saliency or occlusion maps offers insight into the classification decision.

Hematological diseases are inherently variable. This could provide an opportunity for national and international cooperation among multiple medical institutions to establish a centralized database of BMA smears from diverse malignancies, which could be employed to train generalized DL models for future clinical support. The concept of federated learning makes this possible with decentralized data processing, effectively eliminating the potential for data leakage or unauthorized access. Another significant advancement toward this goal was achieved recently with foundation models for cell classification that were trained on 380,000 publicly available cell images from PB and BM.⁸⁷ These complex models rely on self-supervised learning and are rapidly becoming the preferred model choice due to their capacity to be trained on a diverse range of data types, as opposed to task-specific data, and be adapted for a multitude of downstream applications. Furthermore, they can be implemented effectively on previously unseen data⁸⁸ and applied in the development of interactive AI assistants for pathology questions.⁸⁹ Such systems could potentially be adapted for use in hematology applications in the future.

In conclusion, it is becoming more and more realistic to expect the implementation of DL models in clinical practice. The

successful application of AI in predicting genetic abnormalities from morphological characteristics^{64,72} provides compelling evidence of the transformative impact of AI in hematopathology. The strengths of AI may be employed in a complementary manner to the expertise of hematologists as an auxiliary diagnostic tool in clinical routine, offering time-saving assistance for the assessment of BMA smears. Moreover, the availability of readily accessible end-to-end systems for the automatic classification of cells¹⁶ could prove invaluable in the education and training of future medical professionals.

ACKNOWLEDGMENTS

Open Access funding enabled and organized by Projekt DEAL.

AUTHOR CONTRIBUTIONS

Tabita Ghete: Conception and design, research, data analysis, writing—original draft, and writing—review and editing. **Farina Kock:** Research, data analysis, writing—review and editing. **Martina Pontones, David Pfrang, Max Westphal, Henning Höfener:** Research and editing. **Markus Metzler:** Conception and design, writing—original draft, and writing—review and editing.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

FUNDING

The research presented herein is supported by a grant from the German Federal Ministry of Education and Research (FKZ: O31L0262B; BMDeep) to M. Metzler; Grant to M. Metzler: “Schornsteinfeger helfen krebserkrankten Kindern e.V.” (Dörfles-Esbach, Germany).

ORCID

Tabita Ghete  <http://orcid.org/0009-0005-8602-3020>

SUPPORTING INFORMATION

Additional supporting information can be found in the online version of this article.

REFERENCES

1. Shamir SB, Sasson AL, Margolies LR, Mendelson DS. New frontiers in breast cancer imaging: the rise of AI. *Bioengineering*. 2024;11(5):451.
2. Pinto-Coelho L. How artificial intelligence is shaping medical imaging technology: a survey of innovations and applications. *Bioengineering*. 2023;10(12):1435.
3. Hijazi A, Bifulco C, Baldin P, Galon J. Digital pathology for better clinical practice. *Cancers*. 2024;16(9):1686.
4. Kratz A, Bengtsson HI, Casey JE, et al. Performance evaluation of the CellaVision DM96 system: WBC differentials by automated digital image analysis supported by an artificial neural network. *Am J Clin Pathol*. 2005;124(5):770-781.
5. Lee SH, Erber WN, Porwit A, Tomonaga M, Peterson LC. ICSH guidelines for the standardization of bone marrow specimens and reports. *Int J Lab Hematol*. 2008;30(5):349-364.
6. Fuentes-Arderiu X, Dot-Bach D. Measurement uncertainty in manual differential leukocyte counting. *Clin Chem Lab Med*. 2009; 47(1):112-115.

7. Hehr M, Sadafi A, Matek C, et al. Explainable AI identifies diagnostic cells of genetic AML subtypes. *PLOS Digital Health*. 2023;2(3):e0000187.
8. Matek C, Schwarz S, Spiekermann K, Marr C. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nat Mach Intell*. 2019;1(11):538-544.
9. Matek C, Krappe S, Münzenmayer C, Haferlach T, Marr C. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. *Blood*. 2021;138(20):1917-1927.
10. Zhou M, Wu K, Yu L, et al. Development and evaluation of a leukemia diagnosis system using deep learning in real clinical scenarios. *Front Pediatr*. 2021;9:693676.
11. Wang M, Dong C, Gao Y, Li J, Han M, Wang L. A deep learning model for the automatic recognition of aplastic anemia, myelodysplastic syndromes, and acute myeloid leukemia based on bone marrow smear. *Front Oncol*. 2022;12:844978.
12. Eckardt JN, Schmittmann T, Riechert S, et al. Deep learning identifies acute promyelocytic leukemia in bone marrow smears. *BMC Cancer*. 2022;22(1):201.
13. Patel H, Shah H, Patel G, Patel A. Hematologic cancer diagnosis and classification using machine and deep learning: state-of-the-art techniques and emerging research directives. *Artif Intell Med*. 2024;152:102883.
14. Fu X, Fu M, Li Q, et al. Morphogo: an automatic bone marrow cell classification system on digital images analyzed by artificial intelligence. *Acta Cytol*. 2020;64(6):588-596.
15. Jin H, Fu X, Cao X, et al. Developing and preliminary validating an automatic cell classification system for bone marrow smears: a pilot study. *J Med Syst*. 2020;44(10):184.
16. Lv Z, Cao X, Jin X, Xu S, Deng H. High-accuracy morphological identification of bone marrow cells using deep learning-based Morphogo system. *Sci Rep*. 2023;13(1):13364.
17. Wang X, Wang Y, Qi C, et al. The application of morphogo in the detection of megakaryocytes from bone marrow digital images with convolutional neural networks. *Technol Cancer Res Treat*. 2023;22:15330338221150069.
18. Bagg A, Raess P, Rund D, et al. Performance evaluation study of a novel digital microscopy system for the quantitative analysis of bone marrow aspirates. *Blood*. 2021;138:4000.
19. Zanca F, Brusasco C, Pesapane F, Kwade Z, Beckers R, Avanzo M. Regulatory aspects of the use of artificial intelligence medical software. *Semin Radiat Oncol*. 2022;32(4):432-441.
20. Vardiman JW, Harris NL, Brunning RD. The World Health Organization (WHO) classification of the myeloid neoplasms. *Blood*. 2002;100(7):2292-2302.
21. Brown PA, Shah B, Advani A, et al. Acute lymphoblastic leukemia, version 2.2021, NCCN clinical practice guidelines in oncology. *J Natl Compr Cancer Netw*. 2021;19(9):1079-1109.
22. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*. 2016;127(20):2391-2405.
23. Heisterkamp N, Groffen J. Molecular insights into the Philadelphia translocation. *Hematol Pathol*. 1991;5(1):1-10.
24. Melo JV. The molecular biology of chronic myeloid leukaemia. *Leukemia*. 1996;10(5):751-756.
25. Suttorp M, Millot F, Sembill S, Deutsch H, Metzler M. Definition, epidemiology, pathophysiology, and essential criteria for diagnosis of pediatric chronic myeloid leukemia. *Cancers*. 2021;13(4):798.
26. Font P, Loscertales J, Benavente C, et al. Inter-observer variance with the diagnosis of myelodysplastic syndromes (MDS) following the 2008 WHO classification. *Ann Hematol*. 2013;92(1):19-24.
27. Naqvi K, Jabbour E, Bueso-Ramos C, et al. Implications of discrepancy in morphologic diagnosis of myelodysplastic syndrome between referral and tertiary care centers. *Blood*. 2011;118(17):4690-4693.
28. Hatayama Y, Endo Y, Kojima N, et al. Construction of an automatic quantification method for bone marrow cellularity using image analysis software. *Yonago Acta Med*. 2023;66(2):322-325.
29. Nielsen FS, Pedersen MJ, Olsen MV, Larsen MS, Røge R, Jørgensen AS. Automatic bone marrow cellularity estimation in H&E stained whole slide images. *Cytometry Part A*. 2019;95(10):1066-1074.
30. van Eekelen L, Pinckaers H, van den Brand M, Hebeda KM, Litjens G. Using deep learning for quantification of cellularity and cell lineages in bone marrow biopsies and comparison to normal age-related variation. *Pathology*. 2022;54(3):318-327.
31. Shouval R, Fein JA, Savani B, Mohty M, Nagler A. Machine learning and artificial intelligence in haematology. *Br J Haematol*. 2021;192(2):239-250.
32. Gadermayr M, Tschuchnig M. Multiple instance learning for digital pathology: a review of the state-of-the-art, limitations & future potential. *Comput Med Imaging Graph*. 2024;112:102337.
33. Schouten JPE, Matek C, Jacobs LFP, Buck MC, Bošnački D, Marr C. Tens of images can suffice to train neural networks for malignant leukocyte detection. *Sci Rep*. 2021;11(1):7995.
34. Glüge S, Balabanov S, Koelzer VH, Ott T. Evaluation of deep learning training strategies for the classification of bone marrow cell images. *Comput Methods Programs Biomed*. 2024;243:107924.
35. Lee N, Jeong S, Park MJ, Song W. Deep learning application of the discrimination of bone marrow aspiration cells in patients with myelodysplastic syndromes. *Sci Rep*. 2022;12(1):18677.
36. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6(1):60.
37. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nat Rev Clin Oncol*. 2019;16(11):703-715.
38. Chu M-L, Ge XYM, Eastham J, et al. Assessment of color reproducibility and mitigation of color variation in whole slide image scanners for toxicologic pathology. *Toxicol Pathol*. 2023;51(6):313-328.
39. Anilkumar KK, Manoj VJ, Sagi TM. A survey on image segmentation of blood and bone marrow smear images with emphasis to automated detection of Leukemia. *Biocybern Biomed Eng*. 2020;40(4):1406-1420.
40. Saleem S, Amin J, Sharif M, Mallah GA, Kadry S, Gandomi AH. Leukemia segmentation and classification: a comprehensive survey. *Comput Biol Med*. 2022;150:106028.
41. Su J, Wang Y, Zhang J, et al. ROI-BMC-DNNNet: an efficient automatic analysis model of whole-slide scanned bone marrow aspirate images for the diagnosis of hematological disorders. *Biomed Signal Process Control*. 2023;86:105243.
42. Tayebi RM, Mu Y, Dehkharghanian T, et al. Automated bone marrow cytology using deep learning to generate a histogram of cell types. *Commun Med*. 2022;2:45.
43. Wang CW, Huang SC, Lee YC, Shen YJ, Meng SI, Gaol JL. Deep learning for bone marrow cell detection and classification on whole-slide images. *Med Image Anal*. 2022;75:102270.
44. Lewis JE, Shebelut CW, Drumheller BR, et al. An automated pipeline for differential cell counts on whole-slide bone marrow aspirate smears. *Mod Pathol*. 2023;36(2):100003.
45. Liu H, Cao H, Song E. Bone marrow cells detection: a technique for the microscopic image analysis. *J Med Syst*. 2019;43(4):82.
46. Yu T-C, Chou W-C, Yeh C-Y, et al. Automatic bone marrow cell identification and classification by deep neural network. *Blood*. 2019;134:2084.
47. Ghane N, Vard A, Talebi A, Nematollahy P. Classification of chronic myeloid leukemia cell subtypes based on microscopic image analysis. *EXCLI J*. 2019;18:382-404.
48. Chandradevan R, Aljudi AA, Drumheller BR, et al. Machine-based detection and classification for bone marrow aspirate differential counts: initial development focusing on nonneoplastic cells. *Lab Invest*. 2020;100(1):98-109.

49. Mori J, Kaji S, Kawai H, et al. Assessment of dysplasia in bone marrow smear with convolutional neural network. *Sci Rep.* 2020;10(1):14734.
50. Wu YY, Huang TC, Ye RH, et al. A Hematologist-Level Deep Learning Algorithm (BMSNet) for assessing the morphologies of single nuclear balls in bone marrow smears: algorithm development. *JMIR Med Inform.* 2020;8(4):e15963.
51. Wang D, Hwang M, Jiang WC, Ding K, Chang HC, Hwang KS. A deep learning method for counting white blood cells in bone marrow images. *BMC Bioinformatics.* 2021;22(5):94.
52. Su J, Han J, Song J. A benchmark bone marrow aspirate smear dataset and a multi-scale cell detection model for the diagnosis of hematological disorders. *Comput Med Imaging Graph.* 2021;90:101912.
53. Ananthkrishnan B, Shaik A, Akhouri S, Garg P, Gadag V, Kavitha MS. Automated bone marrow cell classification for haematological disease diagnosis using siamese neural network. *Diagnostics (Basel).* 2022;13(1):112.
54. Guo L, Huang P, Huang D, et al. A classification method to classify bone marrow cells with class imbalance problem. *Biomed Signal Process Control.* 2022;72:103296.
55. Hazra D, Byun YC, Kim WJ. Enhancing classification of cells procured from bone marrow aspirate smears using generative adversarial networks and sequential convolutional neural network. *Comput Methods Programs Biomed.* 2022;224:107019.
56. Nakamura I, Ida H, Yabuta M, et al. Evaluation of two semi-supervised learning methods and their combination for automatic classification of bone marrow cells. *Sci Rep.* 2022;12(1):16736.
57. Alshahrani H, Sharma G, Anand V, et al. An intelligent attention-based transfer learning model for accurate differentiation of bone marrow stains to diagnose hematological disorder. *Life.* 2023;13(10):2091.
58. Wang CW, Huang SC, Khalil MA, Hong DZ, Meng SI, Lee YC. CW-NET for multitype cell detection and classification in bone marrow examination and mitotic figure examination. *Bioinformatics.* 2023;39(6):btad344.
59. Cheng Z, Li Y. Improved YOLOv7 algorithm for detecting bone marrow cells. *Sensors.* 2023;23(17):7640.
60. Goldgof GM, Sun S, Van Cleave J, et al. DeepHeme: a generalizable, bone marrow classifier with hematopathologist-level performance. *bioRxiv.* 2023. doi:10.1101/2023.02.20.528987.
61. Manescu P, Narayanan P, Bendkowski C, et al. Detection of acute promyelocytic leukemia in peripheral blood and bone marrow with annotation-free deep learning. *Sci Rep.* 2023;13(1):2562.
62. Peng K, Peng Y, Liao H, Yang Z, Feng W. Automated bone marrow cell classification through dual attention gates dense neural networks. *J Cancer Res Clin Oncol.* 2023;149(19):16971-16981.
63. Wang W, Luo M, Guo P, Wei Y, Tan Y, Shi H. Artificial intelligence-assisted diagnosis of hematologic diseases based on bone marrow smears using deep neural networks. *Comput Methods Programs Biomed.* 2023;231:107343.
64. Kockwelp J, Thiele S, Bartsch J, et al. Deep learning predicts therapy-relevant genetics in acute myeloid leukemia from Papanheim-stained bone marrow smears. *Blood Adv.* 2024;8(1):70-79.
65. Su, J, Liu Y, Zhang J, Han J, Song J. CDC-NET: a cell detection and confirmation network of bone marrow aspirate images for the aided diagnosis of AML. *Med Biol Eng Comput.* 2024. 62(2): p. 575-589.
66. Yu Z, Li J, Wen X, et al. AMLnet, A deep-learning pipeline for the differential diagnosis of acute myeloid leukemia from bone marrow smears. *J Hematol Oncol.* 2023;16(1):27.
67. Valent P, Horny HP, Arock M. The underestimated role of basophils in Ph(+) chronic myeloid leukaemia. *Eur J Clin Invest.* 2018;48(10):e13000.
68. Denburg J, Wilson W, Bienenstock J. Basophil production in myeloproliferative disorders: increases during acute blastic transformation of chronic myeloid leukemia. *Blood.* 1982;60(1):113-120.
69. Kumar D, Jain N, Khurana A, et al. Automatic detection of white blood cancer from bone marrow microscopic images using convolutional neural networks. *IEEE Access.* 2020;8:142521-142531.
70. Huang F, Guang P, Li F, Liu X, Zhang W, Huang W. AML, ALL, and CML classification and diagnosis based on bone marrow cell morphology combined with convolutional neural network: a STARD compliant diagnosis research. *Medicine.* 2020;99(45):e23154.
71. Ouyang N, Wang W, Ma L, et al. Diagnosing acute promyelocytic leukemia by using convolutional neural network. *Clin Chim Acta.* 2021;512:1-6.
72. Eckardt JN, Middeke JM, Riechert S, et al. Deep learning detects acute myeloid leukemia and predicts NPM1 mutation status from bone marrow smears. *Leukemia.* 2022;36(1):111-118.
73. Shafique S, Tehsin S. Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks. *Technol Cancer Res Treat.* 2018;17:1533033818802789.
74. Morteza M, Samadzadehaghdam N, Kermani S, Talebi A. Enhanced recognition of acute lymphoblastic leukemia cells in microscopic images based on feature reduction using principle component analysis. *Front Biomed Technol.* 2015;2(3):128-136.
75. Teachey DT, Pui CH. Comparative features and outcomes between paediatric T-cell and B-cell acute lymphoblastic leukaemia. *Lancet Oncol.* 2019;20(3):e142-e154.
76. Grignani F, Ferrucci PF, Testa U, et al. The acute promyelocytic leukemia-specific PML-RAR α fusion protein inhibits differentiation and promotes survival of myeloid precursor cells. *Cell.* 1993;74(3):423-431.
77. Salah HT, Muhsen IN, Salama ME, Owaidah T, Hashmi SK. Machine learning applications in the diagnosis of leukemia: Current trends and future directions. *Int J Lab Hematol.* 2019;41(6):717-725.
78. Bain BJ, Béné MC. Morphological and immunophenotypic clues to the WHO categories of acute myeloid leukaemia. *Acta Haematol.* 2019;141(4):232-244.
79. Park BG, Chi HS, Jang S, et al. Association of cup-like nuclei in blasts with FLT3 and NPM1 mutations in acute myeloid leukemia. *Ann Hematol.* 2013;92(4):451-457.
80. Greif PA, Konstandin NP, Metzeler KH, et al. RUNX1 mutations in cytogenetically normal acute myeloid leukemia are associated with a poor prognosis and up-regulation of lymphoid genes. *Haematologica.* 2012;97(12):1909-1915.
81. Nagata Y, Zhao R, Awada H, et al. Machine learning demonstrates that somatic mutations imprint invariant morphologic features in myelodysplastic syndromes. *Blood.* 2020;136(20):2249-2262.
82. Eckardt JN, Röllig C, Metzeler K, et al. Prediction of complete remission and survival in acute myeloid leukemia using supervised machine learning. *Haematologica.* 2023;108(3):690-704.
83. Didi I, Alliot JM, Dumas PY, et al. Artificial intelligence-based prediction models for acute myeloid leukemia using real-life data: a DATAML registry study. *Leuk Res.* 2024;136:107437.
84. Radhachandran A, Garikipati A, Iqbal Z, et al. A machine learning approach to predicting risk of myelodysplastic syndrome. *Leuk Res.* 2021;109:106639.
85. Arabyarmohammadi S, Leo P, Viswanathan VS, et al. Machine learning to predict risk of relapse using cytologic image markers in patients with acute myeloid leukemia postthematopoietic cell transplantation. *JCO Clin Cancer Inform.* 2022;6:e2100156.
86. Senaras C, Niazi MKK, Lozanski G, Gurcan MN. DeepFocus: detection of out-of-focus regions in whole slide digital images using deep learning. *PLoS One.* 2018;13(10):e0205387.
87. Koch V, Wagner SJ, Kazemina S. DinoBloom: a foundation model for generalizable cell embeddings in hematology. In: *Medical image computing and computer assisted intervention - MICCAI 2024: lecture notes in computer science*, vol. 15012. Springer Nature Switzerland; 2024:520-530.
88. Schäfer R, Nicke T, Höfener H, et al. Overcoming data scarcity in biomedical imaging with a foundational multi-task model. *Nat Comput Sci.* 2024;4(7):495-509.
89. Lu MY, Chen B, Williamson DFK, et al. A multimodal generative AI Copilot for human pathology. *Nature.* 2024;634:466-473.